Modelo de regresión lineal generalizados para el análisis de la distribución en gastos incurridos en el sector minorista de alimentos a nivel nacional

Felipe Grimaldos Osorio y Karen Adriana Acero Alvarez

Trabajo de grado para optar por el título de Ingeniero Industrial

Director

Henry Lamos Diaz

PhD en Física-Matemática

Codirector

David Esteban Puentes Garzón

MA. en Ingeniería industrial

Universidad Industrial de Santander

Facultad de Ingeniería Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2021

DEDICATORIA

A *María* y *Bernardo*, mis padres por su amor incondicional, respaldo, apoyo durante cada etapa de mi vida, por educarme en valores, por sus palabras de aliento en cada dificultad aun en la distancia y por siempre estar ahí para recordarme que tengo las capacidades para lograr lo que desee.

A *Carolina*, mi hermana por su apoyo emocional, por cuidarme, ser mi confidente y compartir conmigo.

Karen Adriana Acero Alvarez

A mi madre por su gran amor, esfuerzo, dedicación y la formación que me dio a lo largo de mi vida.

A mi padre por acompañarme, apoyarme y brindarme tanta ayuda a lo largo de este camino.

A *Nico* por sus consejos, ser mi guía y enviarme cariño desde la distancia.

A *Akira* por estar ahí siempre que perdía la calma y ser mi fiel hermanita.

A mi familia pues esto es para ustedes.

Felipe Grimaldos Osorio

AGRADECIMIENTOS

Dios, por darme salud, la sabiduría y la oportunidad de cursar mis estudios profesionales.

A *David Esteban Puentes*, por su acompañamiento, paciencia, disposición y guía durante cada etapa del proyecto.

A mi compañero *Felipe Grimaldos* por su paciencia y colaboración.

A mi familia que siempre ha estado presente en cada etapa de una u otra manera.

A mis amigos por estar presentes durante mi desarrollo personal.

Karen Adriana Acero Alvarez

A *David Esteban Puentes*, por instruirnos con su conocimiento y por la confianza brindada.

A mi familia por todo su amor, apoyo y esfuerzo para culminar este camino.

A mi compañera de proyecto *Karen Adriana Acero* por su paciencia y colaboración.

A todos mis compañeros que hicieron parte de mi crecimiento personal.

Felipe Grimaldos Osorio

Contenido

Pág.

Lista de Figuras

Pág.

Lista de Tablas

Pág.

Lista de Apéndices

(Ver apéndices adjuntos y pueden ser consultados en la base de datos de la Biblioteca UIS)

Resumen

Título: Modelo de regresión lineal generalizados para el análisis de la distribución en gastos incurridos en el sector minorista de alimentos a nivel nacional*

Autores: Felipe Grimaldos Osorio
Karen Adriana Acero Alvarez**

Palabras clave: Análisis multivariado, Análisis de componentes principales (ACP), Conglomerados, Distribución de gastos, Gastos, Modelo lineal generalizado, Python, Sector minorista, Ventas.

Descripción:

El sector minorista en Colombia ha contribuido durante los últimos años al crecimiento del Producto Interno Bruto (PIB) aun así es el sector que más bajas ha asumido debido a su entorno competitivo y a la exigencia de los clientes, por esta razón existe la necesidad de continuar trabajando para obtener un desempeño económico favorable dentro del sector mediante la correcta administración, y esto en parte se alcanza haciendo un adecuado control de los gastos, pues de ello depende la rentabilidad de la empresa; una correcta gestión se logra en parte mediante el análisis de cada uno de los gastos en los que incurre la empresa, permitiéndole así tomar decisiones correctas, ofrecer bienes y servicios de calidad y ganar competitividad.

El objetivo principal de esta investigación es la realización de un modelo lineal generalizado mediante el cual se logra llevar a cabo un pronóstico de gastos futuros de empresas minoristas a partir de datos históricos recogidos por el Departamento Administrativo Nacional de Estadística (DANE) en su Encuesta Anual de Comercio (EAC). El modelo se desarrolla en su mayoría a partir del uso del lenguaje de programación Python, se realizan análisis descriptivos para las variables de gasto, se hace uso de análisis multivariado para la obtención del modelo y de herramientas estadísticas que facilitan llevar a cabo un análisis de los gastos en los que incurren las empresas del sector minorista de alimentos. Los resultados señalan: las relaciones entre grupos de variables de gasto, los gastos cuya administración es crucial para la rentabilidad del sector, una caracterización del comportamiento de los gastos en empresas grandes y pequeñas tanto para actividades económicas de categorías especializadas como no especializadas logrando hacer observaciones clave como que las empresas con surtido no especializado tienen gastos promedios más altos a nivel general y requieren mayor número de empleados; y se encuentra finalmente mediante la validación que tan bien predice el modelo los gastos del sector minorista para el año 2018.

---

* Proyecto de grado
** Facultad de ingeniería Físico Mecánicas. Escuela de Estudios Industriales y Empresariales. Programa de Ingeniería Industrial. Director PhD. Henry Lamos Diaz

Abstract

Title: Generalized linear regression model for the analysis of the distribution of expenses incurred in the food retail sector at national level.*

Authors: Felipe Grimaldos Osorio

Karen Adriana Acero Álvarez**

Keywords: Principal Component Analysis (PCA), Cluster, Expenses, Generalized Linear Model, Python, Retail Sector, Sales.

Description:

The retail sector in Colombia has contributed in recent years to the growth of the Gross National Product (PIB), even so, it is the sector that has taken the lowest losses due to its competitive environment and the demands of customers, for this reason, there is a need to continue working to obtain a favorable economic performance within the sector through proper administration, and this is partly achieved by making an adequate control of expenses, since the profitability of the company depends on it; a correct management which is achieved by analyzing each of the expenses incurred by retailers, thus allowing them to make correct decisions, offer quality goods and services and gain competitiveness.

The main objective of this research is the realization of a generalized linear model through which a forecast of future expenses of retail companies is made based on historical data collected by the National Administrative Department of Statistics (DANE) in its Annual Trade Survey (EAC). The model is developed mostly from the use of Python programming language, descriptive analyses are carried out for the expense variables, multivariate analysis is used to obtain the model and statistical tools are used to facilitate the analysis of the expenses incurred by companies in the food retail sector. The results point out: the relationships between groups of expenditure variables, the expenses whose management is crucial for the profitability of the sector, a characterization of the behavior of expenses in large and small companies for both specialized and non-specialized economic activities, making key observations such as that companies with non-specialized assortments have higher average expenses at the general level and require a greater number of employees; and finally, it is found through the validation that the model predicts the expenses of the retail sector so well for the year 2018.

---

* Proyecto de grado

** Facultad de ingeniería Físico Mecánicas. Escuela de Estudios Industriales y Empresariales. Programa de Ingeniería Industrial. Director PhD. Henry Lamos Diaz

Introducción

"El comercio es un importante motor del crecimiento que genera empleo, reduce la pobreza y multiplica las oportunidades económicas" (Banco Mundial, 2019, pág. 44). Según estudio sobre el panorama del comercio minorista, "Colombia se posiciona como el mercado más atractivo para el desarrollo de retail en la región" (Jones Lang LaSalle, 2020, pág. 2); el (Departamento Administrativo Nacional de Estadística (DANE), 2020) reveló que para el año 2019 se obtuvo el Producto Interno Bruto (PIB) más alto desde el 2014 y dentro de las actividades que más contribuyeron al crecimiento del PIB en 2019 se encuentran el comercio al por mayor y al por menor, el cual aportó 0,9 puntos porcentuales a la variación anual. Según (Cruz, Machuca, & Figueroa, 2017), Está claro que todas las organizaciones tienen ingresos y costos, cualquiera que fuese el bien o servicio ofrecido, por lo que los gerentes deben entender la manera en que se comportan, o correrán el riesgo de perder el control; además, de que el cálculo de los costos es una de las actividades clave para la toma de decisiones de una empresa, por lo que los trabajos de investigación realizados en relación con este sector pueden llegar a ser oportunos, debido a que contribuyen a la compresión de la dinámica empresarial en el país y son insumo clave para apoyar la toma de decisiones.

Hasta la fecha se han abordado de distintas maneras los temas relacionados con medir eficiencia y desempeño de las organizaciones, la mayoría de ellas hacen relación con temas como comportamiento del cliente, rendimientos de inversiones, desempeño económico y ventas; para otros sectores también se han propuesto distintos modelos en donde se analiza el impacto de la gestión de los costos en la rentabilidad y el rendimiento de las empresas; para todo ello se han utilizado herramientas como datos de tipo panel, análisis econométrico, análisis envolvente de datos y algoritmos heurísticos.

El análisis cuantitativo aporta a una mejor toma de decisiones mediante procesos de información y conocimiento para la creación de valor y competitividad. Las razones financieras, permiten un análisis de la dimensión óptima de la empresa y la calidad de la gestión empresarial para controlar y comparar a la empresa con el entorno competitivo, por lo que determinar relaciones existentes entre distintos rubros mediante su interpretación permite obtener información acerca del desempeño de la empresa y su postura para el futuro cercano (Flores, Gómez, Briones, & Cervantes, 2013).

Considerando la repercusión que tienen este tipo de análisis es pertinente realizar un estudio para comprender cómo se está realizando la gestión de gastos del sector minorista de alimentos a nivel nacional, haciendo uso de herramientas estadísticas con el objetivo de analizar distintas variables de gasto, conocer la relación entre las mismas, para luego establecer un modelo de regresión lineal generalizado que de ser posible nos permita finalmente predecir para años posteriores valores de ventas totales al año teniendo en cuenta sus arrendamientos de inmueble, propaganda y publicidad, transporte, ratios salariales, entre otros en los que normalmente incurren las empresas de este sector.

Tabla 1.

Cumplimiento de objetivos

Objetivo	Cumplimiento
Realizar una revisión de la literatura a partir de diferentes fuentes de información sobre el comportamiento y evolución de los gastos en empresas del sector minorista de alimentos.	Capítulo 5, apéndice A, B
Determinar los principales gastos incurridos por las empresas de sectores minoristas con surtido compuesto por alimentos a nivel nacional.	Capítulo 7, apéndices C, D, E
Establecer relaciones entre los gastos, niveles de gasto y las variables más representativas de desempeño en la gestión empresarial como ventas y valor agregado.	Capítulo 7, apéndices F,G, H, K
Evaluar y comparar el desempeño de la de regresión lineal mediante métricas de bondad de ajuste.	Capítulo 7, apéndices F, G, H
Validar el modelo para datos históricos del año 2018 para las empresas.	Capítulo 8, apéndices I, J

1. Planteamiento del problema

El sector minorista dentro de toda una cadena de abastecimiento es el encargado de hacer llegar al consumidor o usuario final un bien de consumo; dentro de todos los sectores de la economía, el de alimentos es de los sectores más importantes debido a su aporte económico a un país, a pesar de esto, los establecimientos comerciales dedicados a la venta de bienes de la canasta familiar se exponen constantemente a reestructuraciones en sus procesos, situación que los lleva a ser el sector que más bajas tiene que asumir en todo el panorama comercial (Schiller, 1988).

El entorno empresarial es extremadamente competitivo hoy día, los consumidores son más exigentes que nunca, por lo que el análisis de la eficiencia se ha convertido en un tema fundamental para el sector, la eficiencia favorece la gestión y juega un papel importante en el control y manejo de empresas minoristas (Assaf, Barros, & Sellers-Rubio, 2011). Además, si una empresa desea obtener un mayor flujo de caja bruto año tras año es de crucial importancia que se dé una correcta gestión de los gastos (Vargas, Rojas, & Fino, 2019).

Dependiendo del capital, los ingresos y costos en operación, las empresas minoristas deben ir cambiando sus operaciones adecuándose a las exigencias del mercado, cambios que se realizan a partir de información recopilada del entorno, el mercado, la economía del país, las cifras internas evaluadas y la experiencia del encargado de la toma de decisiones (Cruz B. P., 2017).

De tal manera que el acceso a grandes cantidades de datos se ha convertido en una herramienta esencial debido a que está permitiendo un análisis más completo y profundo en los negocios, esto abre paso hacia la mejora de la eficiencia y efectividad en el sector minorista (Almohri et al., 2019).

Es por eso por lo que un modelo de regresión lineal generalizado haciendo uso de las grandes cantidades de datos que se generan en una empresa y utilizando de manera específica los gastos incurridos por empresas minoristas con surtido compuesto por alimentos, podrían llegar a brindar un panorama más amplio debido a que la gestión de gastos dentro de las empresas.

2. Objetivos

2.1 Objetivo General

Construcción de un modelo de regresión lineal generalizado para la distribución de los gastos incurridos por empresas de los sectores minoristas con surtido compuesto por alimentos a nivel nacional.

2.2 Objetivos Específicos

Realizar una revisión de la literatura a partir de diferentes fuentes de información sobre el comportamiento y evolución de los gastos en empresas del sector minorista de alimentos.

Determinar los principales gastos incurridos por las empresas de sectores minoristas con surtido compuesto por alimentos a nivel nacional.

Establecer relaciones entre los gastos, niveles de gasto y las variables más representativas de desempeño en la gestión empresarial como ventas y valor agregado.

Evaluar y comparar el desempeño de la de regresión lineal mediante métricas de bondad de ajuste.

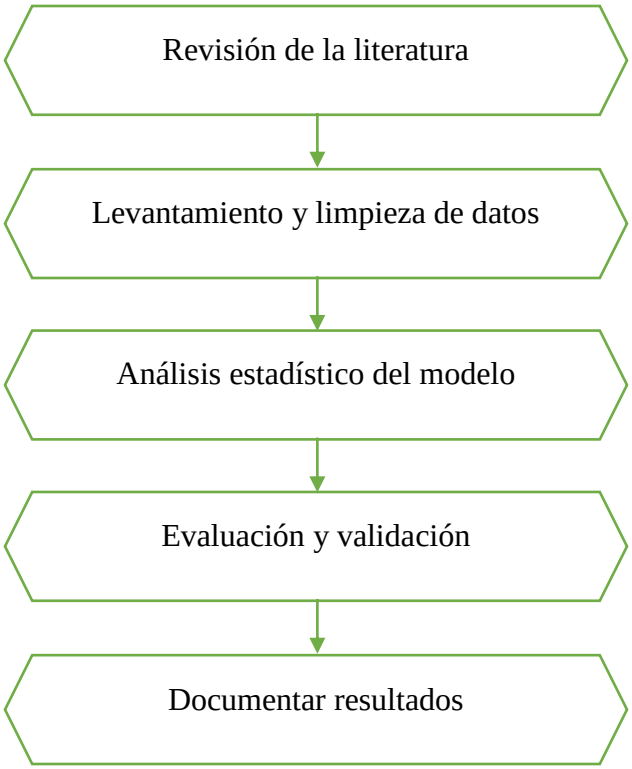
Validar el modelo para datos históricos del año 2018 para las empresas.

3. Metodología

A lo largo de este capítulo se explicará cómo se divide el cuerpo del trabajo, el contenido del mismo y se presenta una estructura de fases las cuales dan cumplimiento a los diferentes objetivos planteados en esta investigación, a continuación, se muestra en la *figura 1* la metodología que se llevará a cabo.

Figura 1

Metodología

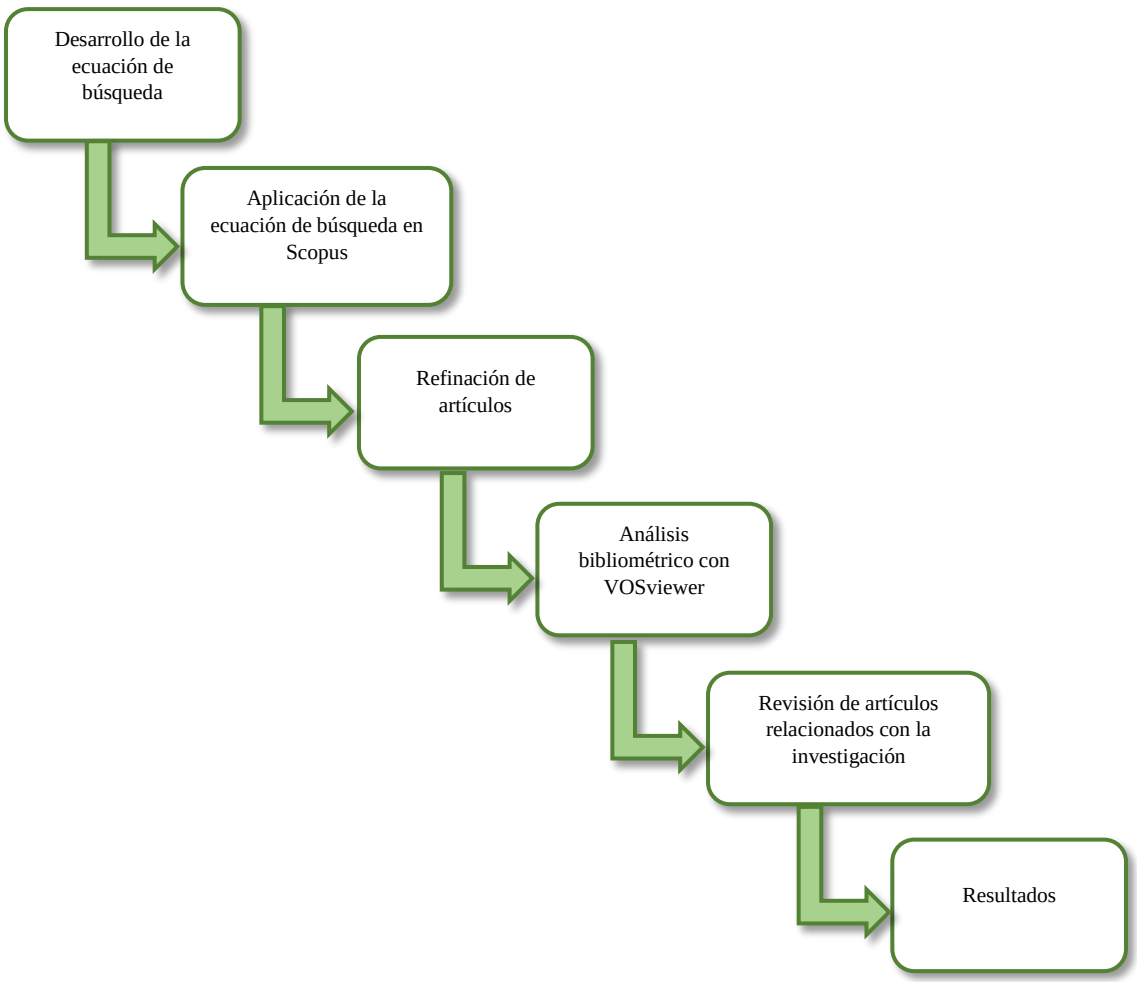


3.1 Fase 1: Definición del problema y revisión sistemática de literatura

En esta primera fase como parte del desarrollo de la investigación se buscaron artículos científicos los cuales permiten analizar los trabajos que se han realizado en el tema. La revisión de literatura tiene como objetivo realizar un análisis para entender la problemática de la distribución de gastos en empresas del sector minorista de alimentos y su manejo a través de los años, logrando un acercamiento a la dinámica que tienen los gastos en distintas organizaciones, además, identificando los principales métodos de estudio de dichas distribuciones y sus resultados. En la *figura 2* se presenta la secuencia de pasos adoptados para el cumplimiento de la fase.

Figura 2

Pasos de la revisión de literatura



La ecuación de búsqueda fue generada sintetizando una serie de palabras claves con las tres grandes temáticas directamente involucradas en el proyecto, las cuales son: modelos estadísticos, distribución de gastos y el sector del comercio al por menor compuesto por alimentos; inicialmente se desarrolló una ecuación por medio de las palabras clave “Linear regression”, “Statistical models”, “Expenses structure”, “Expenses analysis”, “Retail sector” y “ Food retail” palabras clave que en su mayoría se derivan del título del proyecto. Luego de un acercamiento a la temática, se identificaron prácticas y sinónimos ampliamente utilizados en este tipo de estudios, tales como, “econometrics analysis”, “multivariable analysis”, “Costs management”, “Cost distribution”, “business management”, “Food market”, “Retail district”, entre otros.

Con base en lo mencionado anteriormente se generó una ecuación preliminar la cual entregó un total de 401 artículos, los cuales fueron refinados a un número cercano a los 60 artículos, pero dentro de los resultados, aunque estaba directamente relacionado con el sector minorista y la temática general, entregaba pocos resultados relacionados al retail de alimentos. Lo cual afectaría posteriormente la investigación.

Figura 3

Ecuación de búsqueda preliminar

ALL ("econometric methodologies” OR “linear regression” OR “generalized regression model” OR statistical OR “econometric model” OR econometric) AND ALL ("management cost” OR “cost forecast” OR “financial indicator” OR “business management” OR “cost structure" OR “cost distribution") AND ALL ("food retail” OR “retail companies” OR “retail sector” OR “retail district")

Posteriormente se agregaron términos en el último apartado directamente relacionados con el mercado de alimentos y se eliminaron palabras que no se vieran directamente relacionadas con

dicha temática, además, se especificó una búsqueda por título, resumen y palabras clave, generando la siguiente ecuación.

Figura 4

Ecuación de búsqueda final

```
ALL ( "econometri* methodolog*" OR "linea* regressio*" OR "statistical model*" OR
"generalized regressio* model*" OR "statistical model*" OR "econometri* model*"
OR econometric* OR "multivariable* analysis*" ) AND ALL ( "management* cost*"
OR "financial* indicator*" OR "business* management*" OR "cost* structure*" OR
"cost* forecast*" OR "expens* forecast*" OR "expens* analysis*" OR "cost*
distribution*" OR "expens* structur*" ) AND TITLE-ABS-KEY ( "food* retail*" OR
"retail* sector*" OR "food* peddle*" OR "food* market*" OR "food dispense*" )
```

Esta última ecuación entregó un total de 47 artículos los cuales en su mayoría estaban directamente relacionados con los tres grandes grupos temáticos del proyecto y se tomó como ecuación final para el desarrollo del análisis de literatura. La recolección se realizó por medio de la base de datos SCOPUS[1], a la cual se tuvo acceso a través de los recursos electrónicos de la Universidad Industrial de Santander, la totalidad de artículos disponibles se encuentran en el **Apéndice A.**

Posteriormente con el fin de valorar la actividad científica por medio de indicadores de actividad y para conocer la incidencia del tema de investigación, se realizó un análisis bibliométrico por medio del software gratuito VOSviewer[2] en donde se ingresó la totalidad de artículos encontrados por medio de la ecuación de búsqueda, teniendo en cuenta para el análisis

[1] Base de datos bibliográfica de resúmenes y citas de artículos de revistas científicas.
[2] Herramienta de software para analizar y visualizar la literatura científica desarrollada por Nees Jan van Eck y Ludo Waltman del Centro de Estudios de Ciencia y Tecnología (CWTS), de la Universidad de Leiden.

los países que más han contribuido en el desarrollo de la temática, trabajos entre autores y la densidad de trabajos a través de los años, se puede acceder al análisis por medio del **Apéndice B.**

Al contar con los artículos refinados y analizados se procede a hacer lectura de los mismos y sintetizar la información obtenida, dando como resultado los extractos referentes a los modelos de regresión aplicados al sector minorista enfocado en alimentos, su aplicabilidad y métodos de solución. Con base a esto se realizó formalmente la revisión sistemática de literatura que se encuentra en el **capítulo 5**, dando cumplimiento al primer objetivo.

3.2 Fase 2: Levantamiento y limpieza de datos

En esta fase una vez comprendida la fundamentación teórica y las características de los modelos matemáticos, se procede a hacer un levantamiento y limpieza de los datos requeridos para la implementación de dicho modelo y los respectivos análisis previos a su elaboración, todo esto, por medio de herramientas estadísticas. En primera instancia se recurre al Departamento Administrativo Nacional de Estadística (DANE) para obtener los microdatos de la Encuesta Anual de Comercio (EAC) cuyo objetivo es conocer la estructura y el comportamiento económico del sector comercio a nivel nacional, y se selecciona la actividad económica con la que se trabajara haciendo uso de la Clasificación Industrial Internacional Uniforme (CIIU) siendo las empresas con códigos 472 y 4711 (DANE, 2020, págs. 105-106) ambos definidos como empresas enfocadas en comercio al por menor de alimentos tanto para establecimientos distribuidores de alimentos especializados como no especializados respectivamente.

Una vez seleccionadas y descargadas las correspondientes encuestas anuales de comercio y además teniendo en cuenta la necesidad de comprobar posteriormente el modelo resultante por

medio de métricas de bondad de ajuste, se hace una selección de los datos de los años 2014 a 2017 y se deja para verificación los del año 2018.

Tras la selección de las fuentes de datos se procede a elaborar una matriz que consolide la totalidad de los datos, para de esta manera proceder con la selección de variables que de acuerdo a la previa revisión de literatura son relevantes para definir los gastos más representativos en las empresas del sector minorista de alimentos.

Se procede a realizar una limpieza de datos haciendo uso de diferentes métodos estadísticas para la identificación de valores que sean atípicos, faltantes, repetidos, entre otros, **Apéndice C**. Una vez obtenida una matriz limpia se procede a realizar una imputación de dichos datos inválidos y una normalización o estandarización de los mismos para cada variable por medio del lenguaje de programación Python, entregando como resultado de la fase dos una matriz de datos claros y concisos año por año, a la cual se le realizarán análisis estadísticos posteriormente, dicha matriz se encuentra en el **Apéndice D.**

3.3 Fase 3: Análisis estadístico del modelo

A lo largo de esta fase se procede a realizar diferentes análisis estadísticos que ayuden a clasificar las variables y empresas bajo diferentes criterios, se analizarán qué tipos de empresas se encuentran dentro de la totalidad de observaciones, cuáles variables se encuentran correlacionadas y finalmente se realizará un modelo que explique el comportamiento de los gastos y permita hacer predicciones de ventas teniendo en cuenta las variables de gasto del sector; para llevar esto a cabo se realizarán los siguientes procedimientos:

- Estadísticas descriptivas globales.
- Coeficientes de correlación que permite medir el grado de asociación entre variables.

- Análisis de conglomerados o cluster para clasificación de las empresas del sector según sus niveles de gasto y su subdivisión interna en empresas con surtido en alimentos especializado y no especializado.
- Análisis de componentes principales (PCA), con el fin de clasificar las variables que representan la mayor variabilidad y además reducción del número de variables.
- Pruebas de bondad de ajuste para establecer la distribución que mejor se ajusta a las variables seleccionadas.
- Diseño y construcción del modelo de regresión lineal múltiple para predicción de ventas, según los gastos incurridos por las empresas.
- Calcular el desempeño del modelo planteado haciendo uso de métricas como: bondad de ajuste, error medio absoluto y/o raíz cuadrada media del error

Para la realización de los análisis se construye la base de datos de trabajo en Excel generada en la fase inmediatamente anterior, con la cual se realiza la totalidad de procedimientos enunciados anteriormente en el lenguaje de programación Python.

3.4 Fase 4: Evaluación y validación

Se realizará una evaluación y comparación del desempeño del modelo para los datos históricos del año 2018 incluyendo pruebas de bondad de ajuste y error cuadrado medio.

3.5 Fase 5: Documentar resultados

Elaborar un artículo de carácter publicable a partir de la investigación realizada y los resultados obtenidos con las herramientas utilizadas, **Apéndice L.**

4. Marco teórico

A continuación, se presenta una serie de definiciones en aras de mejorar la comprensión de los análisis que se presentan posteriormente.

4.1 Gestión

"Actividades coordinadas para dirigir y controlar una organización" (Icontec, 2015, pág. 15).Es el proceso mediante el cual se formulan objetivos, se miden los resultados obtenidos y se toman acciones pertinentes para la mejora continua de los resultados.

Este término hace referencia a la administración de recursos, sea dentro de una institución estatal o privada, para alcanzar los objetivos propuestos por la misma. Para ello uno o más individuos dirigen proyectos laborales de otras personas para poder mejorar los resultados, que de otra manera no podrían ser obtenidos. (Duran Vasco & Zambrano Loor, 2016).

Gestión es la acción y el efecto de gestionar y administrar, de una forma más específica, una gestión es la diligencia, entendida como el trámite necesario para conseguir algo o resolver un asunto, habitualmente de carácter administrativo o que conlleva documentación; es también un conjunto de acciones u operaciones relacionadas con la administración y dirección de una organización. (Duran Vasco & Zambrano Loor, 2016).

4.1.1 Gestión empresarial

La gestión empresarial hace referencia a las medidas y estrategias llevadas a cabo con la finalidad de que la empresa sea viable económicamente. La misma tiene en cuenta infinidad de factores, desde lo financiero, pasando por lo productivo hasta lo logístico. La gestión empresarial es una de las principales virtudes de un hombre de negocios, engloba a las distintas competencias que se deben tener para cubrir distintos flancos de una determinada actividad comercial en el contexto de una economía de mercado. (Mena, 2012).

4.1.2 Gestión de gastos

Es un sistema de información para predeterminar, registrar, acumular, distribuir, controlar, analizar, interpretar e informar de los costos de producción, distribución, administración y financiamiento, para el uso interno de los directivos de la empresa en el desarrollo de las funciones de planeación, control y toma de decisiones (Moreno & Torres García, 2015).

4.2 Gastos de Operación

Según (Medina & Elvis Vásquez Coloma, 2011), Representan todos aquellos gastos ocasionados por las funciones de compras, ventas y administración del negocio en general. Los estados de resultados muestran generalmente tres categorías de gastos de operación.

4.2.1 Gastos de Venta

Comprenden los gastos relacionados directamente con la venta y la entrega de mercancías, ejemplos de éstos son: los gastos de publicidad, gastos de entrega como salarios, gasolina, depreciación del equipo de reparto, gastos del edificio destinado a ventas, sueldos a los gerentes de ventas, gastos de la oficina de ventas, sueldos a vendedores, gastos de embarques, transportación sobre ventas, gastos de viaje de los vendedores, etcétera.

4.2.2 Gastos generales y administrativos

Comprenden los gastos de supervisión y administración en general, los de llevar los registros y el control contable, gastos de correspondencia, compras, etcétera. Algunos ejemplos son los honorarios de auditoría y contabilidad, gastos de crédito y cobranzas, depreciación del equipo y mobiliario de oficina, gastos de edificio y oficinas de la administración, nómina de oficina, artículos de escritorio, papelería y correo, teléfono y telégrafo, etcétera.

4.2.3 Gastos financieros

Comprenden los gastos en que incurre un negocio debido al uso de fondos externos (pasivo) para financiar sus activos. Este renglón incluye los intereses, la amortización del descuento en emisión de obligaciones, las comisiones, etc. Aquí comienza a apreciarse el destino de las utilidades logradas con los activos.

4.3 Sectores económicos

"La actividad económica del país está dividida en sectores económicos. Su división se realiza de acuerdo con los procesos de producción que ocurren al interior de cada uno de ellos" (Banco de la República | Colombia, s.f.), de estos sectores se derivan actividades económicas independientes y especializadas como el sector de comercio.

4.3.1 Sector de comercio

Este hace parte del sector terciario de la economía, e incluye comercio al por mayor, minorista, centros comerciales, cámaras de comercio, San Andresito, plazas de mercado y, en general, a todos aquellos que se relacionan con la actividad de comercio de diversos productos a nivel nacional o internacional (Banco de la República | Colombia, s.f.).

4.3.1.1 Minorista. "El comercio detallista o minorista es el último eslabón de la distribución comercial, es el intermediario que se dedica a la venta de productos, bienes o servicios a los consumidores o usuarios finales" (Garcia, 2003, pág. 23).

4.3.1.2 Clasificación Industrial Internacional Uniforme (CIIU). "Clasificación uniforme de las actividades económicas productivas. Su propósito principal es ofrecer un conjunto de categorías de actividades económicas que se pueda utilizar para la reunión y presentación de estadísticas de acuerdo con esas actividades" (Departamento Administrativo Nacional de Estadística (DANE), 2020, pág. 7), dentro de la clasificación se encuentran los grupos 472 y 4711,

utilizados en el trabajo de investigación y cuya actividad es el comercio al por menor en establecimientos (Departamento Administrativo Nacional de Estadística (DANE), 2020):

- Clase 4711: Comercio al por menor en establecimientos no especializados con surtido compuesto principalmente por alimentos, bebidas (alcohólicas y no alcohólicas) o tabaco, estos establecimientos se encuentran en los supermercados, cooperativas de consumidores y otros establecimientos similares.
- Clase 472: Comercio al por menor en establecimientos especializados de: frutas y verduras, leche, productos lácteos y huevos, productos cárnicos, pescados y productos del mar, bebidas y productos de tabaco y otros productos alimenticios n.c.p (no clasificados previamente).

4.4 Distribución de probabilidad

La distribución de probabilidad es una función que revela los posibles resultados de un experimento y asigna a cada evento una probabilidad de ocurrencia en el futuro; las distribuciones de probabilidad se caracterizan porque: sus resultados se encuentran en el rango de 0 a 1, además estos resultados son eventos mutuamente excluyentes y la suma de las probabilidades de los diversos eventos es 1 (Lind, Marchal, & Wathen, 2012).

4.5 Prueba de bondad de ajuste

(Maydeu Olivares & García Forero, 2010) definen la prueba de bondad de ajuste como la herramienta utilizada en los modelos estadísticos que describe que tan bien encaja en un conjunto de observaciones, los índices de la prueba muestran la discrepancia entre los valores observados y los esperados de un modelo estadístico, estos índices pertenecen a distribuciones de muestreo conocidas generalmente obtenidas de métodos asintóticos utilizadas en la prueba de hipótesis

estadísticas, dejando ver a través de estos la fuente del desajuste en modelos que no se ajustan bien.

4.6 Modelo estadístico

Este hace parte de los métodos estadísticos requeridos en distintas disciplinas para llevar a cabo una representación formal de los datos presentes en un sistema real, con el que se pretende realizar un análisis e interpretación para posteriormente hacer predicciones y contribuir a su control (Korner-Nievergelt, y otros, 2015).

4.6.1 Modelo Lineal

4.6.1.1 Modelo lineal generalizado (GLM). Según (Montgomery, 2004) un modelo lineal generalizado es básicamente un modelo de regresión, sin embargo, el modelo GLM difiere del modelo de regresión ordinario en dos aspectos importantes: (i) La distribución de la respuesta se elige a partir de la familia exponencial. Por lo tanto, la distribución de la respuesta no necesita ser normal o cercana a lo normal y puede ser explícitamente no normal. (ii) Una transformación de la media de la respuesta está relacionada linealmente con las variables explicativas (Jong & Heller, 2008).

Un modelo de regresión está constituido por un componente aleatorio (lo que se ha llamado generalmente el término del error) y una función determinista de los factores del diseño (las x). Un modelo de regresión lineal de la teoría normal estándar se escribe:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (1)$$

Donde se supone que el término del error ϵ tiene una distribución normal con media cero y varianza constante, y la media de la variable respuesta (y) es:

$$E(y) = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = x'\beta \quad (2)$$

A la parte **x´β** de la *ecuación 2* se le llama predictor lineal. El modelo lineal generalizado contiene la ecuación 1 como un caso especial.

En un modelo lineal generalizado, la variable de respuesta puede tener cualquier distribución que sea un miembro de la familia exponencial. Esta familia incluye las distribuciones normales, de Poisson, binomial, exponencial y gamma, *tabla 2*, por lo que la familia exponencial es una colección rica y flexible de distribuciones aplicables en muchas distribuciones experimentales. Además, la relación entre la media de la respuesta μ y el predictor lineal **x´ β** se determina por una función de enlace.

$$g(\mu) = x'\beta \quad (3)$$

Tabla 2

Modelos y links usados en la modelización de GLM

Modelos	Links	
Normal	identidad	μ
Binomial	inverso	1/μ
Poisson	inverso cuadrático	$1/\mu^2$
Gamma	raíz cuadrada	$\sqrt{\mu}$
Gaussiano inverso	exponencial	$(\mu + c_1)^{c_2}$
tipo	log	$\log(\mu)$

logit	$\log\left(\frac{\mu}{1-\mu}\right)$
loglog	$\log(-\log(\mu))$
probit	$\Phi^{-1}(\mu)$

Nota: Adaptado de *Modelos lineales generalizados* (p.7), por M.A. Martínez & J.M. Socuéllamos, 2001, Universidad Miguel Hernández.

El modelo de regresión que presenta la respuesta media está dado por la *ecuación 4:*

$$E\ (y) = \ \mu = g^{-1}(x'\beta) \tag{4}$$

Por ejemplo, a la función de enlace que lleva al modelo de regresión lineal ordinario en la *ecuación 1* se le llama enlace identidad debido a que $\mu = \mathbf{g}(-\mathbf{1})(\mathbf{x}'\,\boldsymbol{\beta}) = \mathbf{x}'\,\boldsymbol{\beta}$**.** Como otro ejemplo, el enlace log (logarítmico)

$$\ln(\mu) = x'\beta \tag{5}$$

Produce el modelo

$$\mu = e^{x'\beta} \tag{6}$$

El enlace logarítmico se usa con frecuencia con datos de conteos (respuesta de Poisson) y con respuestas continuas que presentan una distribución que tiene una cola larga a la derecha (la distribución exponencial o gamma). Otra función de enlace importante que se usa con datos binomiales es el enlace logit.

$$\ln\left(\frac{\mu}{1-\mu}\right) = x'\beta \tag{7}$$

Esta elección de la función de enlace lleva al modelo.

$$\mu = \frac{1}{1 + e^{x'\beta}} \tag{8}$$

Hay muchas elecciones posibles de la función de enlace, pero debe ser siempre monótona y diferenciable, observe asimismo que, en un modelo lineal generalizado, la varianza de la variable de respuesta no tiene que ser una constante; puede ser una función de la media (y de las variables predictores a través de la función de enlace. Por ejemplo, si la respuesta de Poisson, la varianza de la respuesta es exactamente igual a la media.

4.6.2 Análisis multivariado

Consisten en mediciones u observaciones para p variables o características. **P>1**, asociadas a cada uno de **n** individuos en un punto dado del tiempo. Estas **n** observaciones en **p** variables se puede representar mediante un arreglo rectangular.

O lo que es lo mismo, como la matriz de datos, **X = (Xij),** de orden **n x p**, con **i = 1. 2. …, n. j = 1, 2,**…, **p**, donde **Xij** representa el valor de la variable **j** en el individuo **i**. En la matriz de datos **X,** cada fila representa una observación multivariada.

Las técnicas estadísticas multivariadas constituyen una herramienta esencial para la investigación que se realiza en las diversas disciplinas científicas. La biología, la agronomía, la economía, la ingeniería, la medicina, la psicología, y la demografía son una de las que se benefician de esas técnicas. Entre las técnicas multivariadas básicas se encuentran: análisis de componentes principales, análisis factorial correlación canónica, análisis discriminante, análisis multivariado de varianza y regresión logística. (Rodriguez, 1998).

En el presente trabajo se hace uso del análisis de clúster para identificar patrones relevantes de los datos asociados a la investigación.

4.6.2.1 Análisis de Clúster. Es una técnica estadística multivariante que consiste en conformar grupos homogéneos basados en la idea de que los datos contienen características similares entre ellos, se puede aplicar para agrupar observaciones y/o variables de un conjunto de datos *(Peña, 2002)*. El análisis se realiza para dividir los elementos de tal manera que cada elemento pertenezca a un solo grupo y quede clasificado, y además se espera que cada grupo sea homogéneo internamente *(Peña, 2002)*. Se realiza una partición de datos mediante distintos métodos haciendo uso la matriz de datos, o si se recurren a algoritmos jerárquicos se hace uso de la matriz de distancias entre elementos, usualmente la construcción de los grupos se realiza de manera jerárquica, esta consiste en que los datos se ordenan por niveles de forma tal que los niveles superiores contienen a los niveles inferiores; finalmente se logra la disminución de variables o de observaciones del problema a analizar *(Peña, 2002)*.

Dentro de los de los métodos clásicos de partición se encuentra el algoritmo de k-medias, el objetivo de este es dividir los datos en un número *k* de grupos previamente fijado, para esto según (Peña, 2002) (1) se seleccionan *k* puntos como centro de los grupos iniciales pudiéndose realizar de tres maneras, ya sea: asignándoles de manera aleatoria, tomando como centros los *k* puntos más distantes entre sí, o construyéndose a partir de previa información, una vez establecidos los *k* centros (2) se calculan las distancias euclídeas de cada elemento al centro de los *k* grupos y estos elementos son entonces asignados al grupo más próximo, esto se realiza secuencialmente de tal mera que a asignar un elemento a un grupo se recalculen nuevamente las coordenadas y la media de este, finalmente (3) se define un criterio de optimalidad al asignar cada elemento a otro grupo y se revisa si mejora el criterio o no, para dar así por terminado el algoritmo.

Otra técnica multivariante que se usa en la investigación es el Análisis de Componentes Principales (ACP) cuya función es la reducción de la dimensión de una matriz de datos perdiendo

la menor cantidad de información posible, mediante la construcción de un nuevo conjunto reducido de variables no correlacionadas conservando un alto porcentaje de la varianza generalizada de las variables originales, dando paso a un análisis exploratorio de la información y posteriormente a la construcción de modelos predictivos.

4.6.2.2 Análisis de Componentes Principales (ACP). El análisis de componentes principales (**ACP**) trata de explicar la estructura de las varianzas y covarianzas de un conjunto de variables **Xi** mediante unas cuantas combinaciones lineales entre ellas, llamadas componentes principales. Estos componentes principales no están correlacionados entre sí, y cada uno maximiza su varianza. El **ACP** aspira a reducir o simplificar los datos y facilitar su análisis e interpretación.

4.6.2.2.1 Definición y determinación de los componentes principales. Supongamos que las variables aleatorias **x1, …, xp** poseen una distribución p-dimensional cualquiera, con vector de medias **μ** y matriz de varianza y covarianza**.** De esta distribución se extrae una muestra de **n** observaciones independientes **x1, x2, …, xn,** representada por la matriz de datos **x = (Xij),** para i **= 1. 2. …, n. j = 1, 2,** …, **p,** donde:

xi´ $= (\boldsymbol{xi1} \;\; \boldsymbol{x12} \;\; ... \;\; \boldsymbol{xij} \;\; ... \;\; \boldsymbol{xip})$

= Vector fila **i**- ésimo de la matriz x.

= observación multivariada **i**- ésima (**i = 1. 2. …, n).**

Donde se calcula el vector de medias muéstrales *ecuación 9*:

$$\boldsymbol{X} = \begin{matrix} \boldsymbol{X1} \\ \boldsymbol{X2} \\ . \\ . \end{matrix} = \begin{matrix} \frac{\sum_{i=1}^{n} xi1}{\boldsymbol{n}} \\ . \\ . \\ . \end{matrix}$$

$$\boldsymbol{Xp} \qquad \frac{\sum_{i=1}^{n} xip}{\boldsymbol{n}} \tag{9}$$

Luego matriz de varianzas y covarianzas muestrales *ecuación 10*:

$$\boldsymbol{S} = (\boldsymbol{sjk})$$

Donde,

$$\boldsymbol{Sjk} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{xij} - \boldsymbol{Xj})(\boldsymbol{xik} - \boldsymbol{Xk})}{\boldsymbol{n-1}} \tag{10}$$

Y la matriz de correlaciones muestrales *ecuación 11:*

$$\boldsymbol{R} = (\boldsymbol{rjk})$$

Donde,

$$\boldsymbol{rjk} = \frac{\boldsymbol{sjk}}{\sqrt{\mathbf{sjj}}\sqrt{\boldsymbol{skk}}} \tag{11}$$

Para precisar, consideremos las p combinaciones lineales siguientes, cada una evaluada sobre los *n* individuos de la muestra

$$\boldsymbol{y1} = \boldsymbol{a11x1} + \boldsymbol{a21x2} + \cdots + \boldsymbol{ap1xp} = \boldsymbol{a'1X}$$

$$\boldsymbol{y2} = \boldsymbol{a12x1} + \boldsymbol{a22x2} + \cdots + \boldsymbol{ap2xp} = \boldsymbol{a'2X}$$

.

$$\boldsymbol{yi} = \boldsymbol{a1ix1} + \boldsymbol{a2ix2} + \ldots + \boldsymbol{apixp} = \boldsymbol{a'iX}$$

.

$$\boldsymbol{yp} = \boldsymbol{a1px1} + \boldsymbol{a2px2} + \cdots + \boldsymbol{appxp} = \boldsymbol{a'pX}$$

Los componentes principales muestrales son combinaciones lineales **y1, y2, …, yp,** como las anteriores, pero que tienen vectores de coeficientes **a´1** muy particulares, tales que**:**

a. **Y1, y2, …, yp,** no están correlacionadas entre sí.

b. **Var (y1), Var (y2), …, Var (yp)** alcanzan los calores posible más grandes.

c. Los vectores coeficientes **a´1** son tales que **a´i*ai = 1**

El primer componente principal maestral **y 1** maximiza su varianza muestral **var (y1) = a1´Sa1** sujeto a **a1´*a1 = 1.** El segundo componente principal **y2** maximiza **var (y2) = a´2Sa2** sujeto a **a2*a2=1** y covarianza muestral **cov (y1, y2) = 0.** El tercer componente **y3** maximiza **var (y3) = a´3Sa3** sujeto a **a3´*a3 = 1, cov (y1, y2) = 0** y **cov (y2, y3) = 0.** Sucesivamente hasta el i-ésimo componente. (Rodriguez, 1998).

5. Revisión sistemática de la literatura

A continuación, se presenta la revisión de literatura realizada para llevar a cabo la investigación y dar cumplimiento al primer objetivo de la misma, la información que aquí se encuentra se tiene en cuenta para los análisis posteriores.

La identificación de las fuentes de variación a nivel de empresa en la rentabilidad es un tema de investigación importante en economía, gestión estratégica y contabilidad y finanzas. Además, en economía industrial el papel de los recursos internos específicos de la empresa son determinantes de variaciones directas en la rentabilidad y adquieren una importancia mayor, primordial a la hora de la planeación estratégica y el enfoque de nuevas políticas de gastos empresariales. (Goddard, Tavakoli, & Wilson, 2013)

Los modelos utilizados a través de los años en la investigación de las distribuciones de gastos en las empresas, han tenido como objetivo el mejoramiento de la manufactura y las capacidades estratégicas. Según (Goddard, Tavakoli, & Wilson, 2013) la implementación de investigaciones empíricas econométricas con metodologías tipo panel es uno de los grandes avances en el sector de la manufactura y la prestación de servicios.

(Sellers-Rubio & Mas-Ruiz, 2006), miden la eficiencia económica del sector minorista sobre una muestra de 100 cadenas de supermercados en España entre 1995-2001, utilizando análisis envolvente de datos no parametrizados (DEA), presentando como resultado un alto grado de ineficiencia en el subsector de los supermercados, debido a que las empresas consideradas podrían obtener más producción utilizando los mismos recursos, esto puede ser debido a que solo se analiza a uno de los actores de las cadena de distribución; este análisis favorece la gestión de los productores debido a que les permite identificar minoristas que utilizan eficientemente los recursos para llevar sus productos al mercado, estableciendo la eficiencia como un criterio para la elección de relaciones verticales en al canal de distribución, y a nivel horizontal permite posteriormente realizar análisis estratégico de "benchmarking".

Según el trabajo de (Barros & Sellers-Rubio, 2008) para estimar eficiencia se han desarrollado trabajos en el pasado en donde se han propuesto modelos paramétricos y no paramétricos, aunque ninguna técnica domina sobre otra, han encontrado dentro de la literatura que la mayoría de trabajos aplican análisis envolvente de datos (DEA) como medida de eficiencia no paramétrica. En su investigación proponen un marco para la evaluación comparativa de los minoristas en España entre los años de 2001 a 2004, utilizando un enfoque de frontera econométrica de costos estocásticos con la función de costo de frontera general propuesta por (Aigner, Lovell, & Schmidt, 1977) y (Meeusen & van Den Broeck, 1977) para medir la rentabilidad, encontrando entonces que: las empresas presentan altos niveles de ineficiencia en costes, el modelo de frontera aleatoria describe mejor a los minoristas españoles que los modelos de frontera homogéneos.

Por su parte, (Goddard, Tavakoli, & Wilson, 2013) elaboraron un modelo econométrico empírico el cual buscó determinar las principales causas de disminuciones en las ganancias de

empresas de sector minorista europeo, donde definieron como variable dependiente la rentabilidad de la empresa e independientes cinco variables con una organización de datos tipo panel para su posterior procesamiento. Las variables o factores independientes (determinantes) se encuentran el tamaño de la empresa, la cantidad total en ventas, la liquidez y el know how o estructura empresarial. Todos estos factores fueron calculados teniendo en cuenta una serie de características para su clasificación debido a que no todas estas variables fueron cuantitativas. Entre los resultados obtenidos identificaron una relación negativa entre el tamaño de las empresas y su rentabilidad, esto probablemente debido a los gastos que requieren empresas de tamaños mayores, tanto en costos administrativos como en dificultad en la gerencia de este, por el contrario, la relación entre la cuota de mercado y la rentabilidad es consistentemente positiva. Esto puede reflejar la tendencia de muchos fabricantes europeos para participar en costosas estrategias de construcción exceso de capacidad, publicidad y promoción, e innovación para ganar cuota de mercado y desalentar la nueva competencia.

Por otro lado, (Janda & Rausser, 2013) en su Investigación presentan un análisis con base empírica de los factores que influyen rentabilidad en microempresas rurales en el este de Europa, por medio de una modelo de estadística multivariable, donde toman como variables principales la edad de la empresa, su ubicación dividida en provincias y el número de empleados como variables independientes, frente a rentabilidad vista como el ROA (Return On Assets) donde toman estas condiciones individualmente para cada empresa y finalmente encuentran una relación negativa frente a la rentabilidad y el tamaño de la empresa donde resulta ser más conveniente en términos de ganancia las empresas con menos número de empleados ubicadas al tiempo en la provincia con mayor costo de arrendamiento o del terreno.

En la literatura se repite el uso del ROA (Return On Assets) por sus siglas en inglés, lo que se traduce como el retorno sobre los activos el cuales un indicador de rentabilidad que muestra el nivel de eficiencia con el cual se manejan los activos de una empresa, visto de otra forma es la diferencia entre a utilidad neta generada antes de impuestos por la empresa sobre el total de activos promedio de la misma. (Pinelo, 2020).

A su vez, (Gaur & Saravanan, 2015) manejan un modelo estadístico para la identificación de buenas prácticas administrativas, especialmente en la generación de costos debidos al manejo de inventarios, el cual es identificado por ellos como el mayor activo de las pequeñas y medianas empresas. Además de ser el mayor activo físico, este afecta directamente la rentabilidad de la empresa debido a su peso en el cálculo del ROA, en su investigación tienen en cuenta otras variables tales como las ventas y el tamaño de la empresa el cual es directamente proporcional al número de empleados de la misma; entre sus hallazgos se encuentran que el tamaño del inventario es mayor para empresas con mayor tamaño y por tanto si el nivel de ventas no era suficientemente elevado, los costos generados por el inventario se ven traducidos en baja rentabilidad.

La identificación de fuentes de ventajas competitivas en empresas ha sido uno de los principales puntos de investigación, debido al beneficio que generan estratégicamente en la administración y gestión de las organizaciones, basados en esto, (Barney, 1991) realizó una investigación literaria donde encuentra que los recursos de una empresa, ya sea sus activos, capacidades, gestión de gastos, administración, capital humano, atributos, información, entre otros; son fuentes de posibles ventajas competitivas para cada empresa en específico, donde una buena gestión se ve reflejada en rentabilidad y ventas.

Por su parte, (Butigan & Benić, 2017) definen las diferencias entre la rentabilidad de una empresa a otra como una característica transitoria del comportamiento de la empresa; plantean

mediante un modelo de estadística multivariada con variable dependiente el ROA, identificar como el pertenecer a una alianza estratégica entre empresas de un sector, las capacidades de la empresa, su manejo financiero y el comportamiento del sector minorista en general de la empresa, siendo como variables de control, se ve afectada la rentabilidad de las empresas.

(Britchenko, Monte, Kryvovyazyuk, & Kryvoviaziuk, 2018) trabajan sobre la comparación de desempeño y eficiencia de empresas en Portugal y Ucrania, identificando factores que expliquen estos dos asuntos en empresas del sector industrial de la economía realizando un análisis descriptivo e inferencial de la actividad económica de la empresa y sobre estos se aplicaron regresiones multivariadas, método de mínimos cuadrados ordinarios, análisis de regresión múltiple y de componentes principales; haciendo uso de variables como lo son: la relación de la rotación de activos y retorno de activos, tamaño de la empresa, la gestión del capital de trabajo, la solvencia, el margen de deuda, gastos, ventas, entre otras; finalmente encontrando que en ambos países las empresas son eficientes, pero presentan mayor eficiencia en Ucrania, en cuanto a los sectores aunque no hay diferencia significativa, el sector que presenta eficiencia ligeramente superior es la del papel y ligeramente inferior en la de construcción; las empresas no son tan rentables pero presentan un desempeño promedio positivo y ligeramente mayor en Ucrania.

(Almohri, Chinnam, & Colosimo, 2019) estudiaron cuáles eran los factores que impulsan el desempeño de los concesionarios automotrices en comparación con concesionarios automotrices similares, teniendo en cuenta variables como ventas y ganancias y haciendo uso del análisis envolvente de datos (DEA), modelos de mezcla finita (FMM) para segmentar y agrupar tiendas mediante análisis de clúster, mezcla finita de regresiones y modelos de mezcla con restricciones para que todos los datos de un concesionario se agrupen dentro del mismo modelo evitando así que queden distribuidos en distintos modelos; proponen una solución de algoritmo heurístico

basado en el “aprendizaje competitivo” para la mezcla finita de regresiones de estructura de grupo, denominado Modelo de mezcla con aprendizaje competitivo (MMCL), empleado para abordar el problema de agrupamiento de concesionarios de automóviles y la gestión del rendimiento, además realizan una formulación multiobjetivo para mejorar la rentabilidad mientras se controlan métricas de desempeño que cumplan con las expectativas de las partes interesadas. Los métodos son validados mediante experimentos sintéticos, así como datos de un estudio de caso de una red de concesionarios, dando como resultado precisión y eficacia de las metodologías utilizadas.

(Busu, Vargas, & Gherasim, 2020) Realizan un análisis de los factores internos de una empresa que influyen en el desempeño económico del sector y desean demostrar que el número de empleados se relaciona con el desempeño de la organización, hacen uso de análisis econométrico e intentan estimar cuáles de los factores exógenos que son: activos corrientes, activos fijos y número de empleados, que tienen mayor impacto sobre el desempeño económico. Para probar las hipótesis planteadas utilizan el modelo de regresión lineal múltiple con datos transversales. Los resultados demuestran que estas tres variables exógenas son factores significativos del beneficio neto en las empresas del sector.

En la revisión no se encontraron artículos de investigación que hayan tenido como foco de estudio el sector minorista en Colombia, además se encuentra de manera repetida el análisis de rentabilidad como tema fundamental para evaluar el desempeño de una empresa, dentro de los cuales se toma en cuenta principalmente el capital y los ingresos, prestando menos atención al comportamiento de los gastos operativos, los cuales se considera en el presente trabajo que pueden influir en forma determinante en el desempeño de una organización; de aquí la importancia de realizar un análisis que no solamente evalué los gastos operativos de manera selectiva como se

evidencia en la literatura, si no que mediante un modelo estadístico se estudien las relaciones presentes entre cada uno de los gastos presentes en las empresas del sector retail en Colombia.

6. Levantamiento y limpieza de datos

El levantamiento de los datos se hace vía electrónica tomando matrices de datos elaboradas por el DANE las cuales se encuentran con acceso abierto al público, de donde son descargadas y agrupadas con información entre el año 2015 y 2017, con 64 variables diferentes y variables que en su mayoría son de gasto. Parte del trabajo consiste en seleccionar únicamente las identificadas previamente en la revisión de literatura y aquellas que a consideración podrían ser igualmente relevantes para el análisis empresarial.

Tras llevar a cabo la eliminación de dichas variables no representativas se obtiene como resultado una matriz de 3630 datos por 10 variables que serán directamente analizadas y algunas se tienen en cuenta para el modelo de regresión final, dichas variables se encuentran en la *tabla 3*, algunas de ellas son el arriendo de cada organización , gastos en servicios internos y externos, costos en transporte, remuneración devengada por empleados y el número de empleados, niveles de inventario y costos de fabricación, todas estas variables independientes y una única variable dependiente “ventas”; todos estos datos siendo totales anuales.

Tabla 3

Descripción de variables

VARIABLE	DESCRIPCIÓN
CORRELA	Correlativa, es el código CIIU correspondiente a la actividad económica
ARRIENDO	Gasto en Arrendamiento de bienes inmuebles y muebles
SERV_INT	Servicios internos, gasto que comprende los montos de gas, aseo y vigilancia, energía eléctrica, comunicaciones y otros servicios públicos
PUBLICI	Gasto en Propaganda y publicidad
TRANSP	Gatos en empaque, embalaje, transporte, fletes y acarreos
SERV_EXT	Servicios externos, gastos en seguros, mantenimiento y reparaciones
TOTAL_REM	Total remuneración, gastos en sueldos y prestaciones sociales
VENTA	Ventas causadas en el año
TOT_PERSO	Personal total
INV_PRO	Inventario promedio
CTO	Costo de la mercancía vendida

Seguido de la selección de variables se lleva a cabo la imputación de datos por medio de filtros y fórmulas simples desarrolladas en Excel y en Python, en donde se eliminan valores de cero interpretables como errores de digitación, valores excesivamente altos en comparación con todos los demás, se elimina todo valor no numérico y la totalidad de datos especificados como anormales fueron reemplazados por una casilla vacía, la finalidad de este procedimiento es hacer una clarificación de la totalidad de datos y lograr tener la menor cantidad de datos atípicos posibles

para los respectivos análisis; la matriz donde se aprecia los datos que fueron reemplazados se puede apreciar en el **apéndice C.**

Todos estos valores nombrados anteriormente los cuales se representan por celdas Nan en Python se remplazan o imputan por la media del valor de cada variable para la totalidad de empresas, además, se garantiza la coherencia entre datos y se reduce la redundancia de estos por medio de una normalización, este procedimiento se desarrolla aplicando la *ecuación 12* a cada columna de datos. Como resultado se obtiene un dataframe[3] de datos homogéneos, sin datos faltantes o atípicos que conlleven a una mala interpretación en los análisis posteriores y a su vez con todas sus observaciones escaladas entre 0 y 1 debido a que se trabaja una normalización por valores mínimos y máximos, Ver en **Apéndice E.**

$$I_i = \frac{(X_i - X_{min})}{(X_{max} - X_{min})} \qquad (12)$$

7. Análisis estadístico del modelo

En esta etapa se formulan diferentes análisis, pruebas y modelos matemáticos, con los cuales se logra definir con claridad las características de las variables de gasto en el sector minorista con surtido compuesto de alimentos, cada análisis realizado está sujeto a modificaciones durante la formulación y recopilación de resultados, con el propósito de ajustar de la manera más

[3] Permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables.

uniforme los resultados y poder sintetizar una clasificación y definición de la dinámica de los gastos en el sector con claridad.

A partir de esta fase todos los procesos, cálculos y pruebas se llevan a cabo con el lenguaje de programación Python y la gran gama de herramientas que brinda cada una de sus bibliotecas.

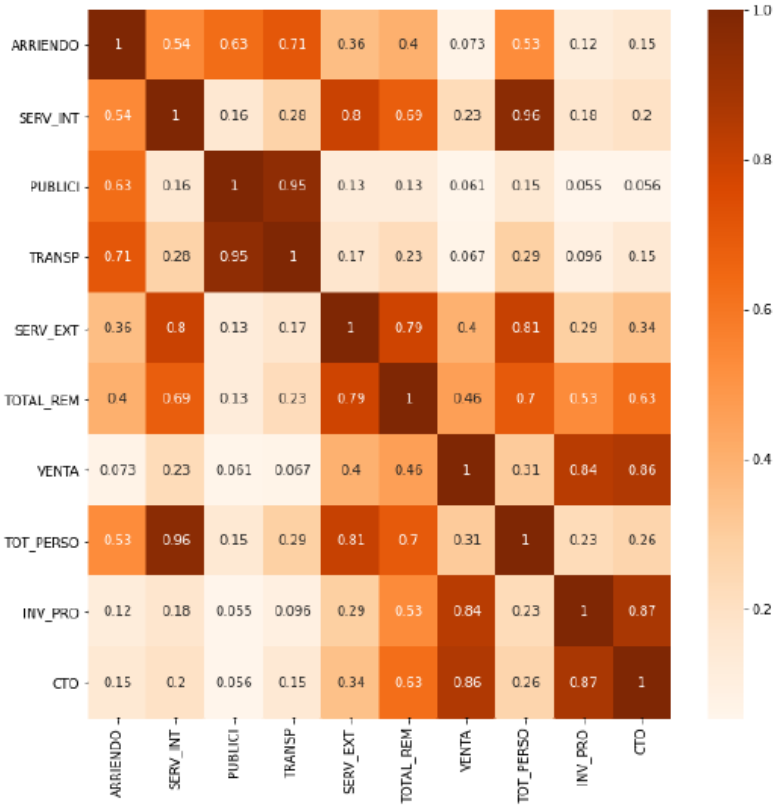
7.1 Análisis de conglomerados o clustering

Con los datos previamente tratados se procede a calcular la matriz de correlación haciendo uso de la librería Matplotlib[4] se realiza un mapa de calor (*figura 5*), que muestra en ambos ejes las diferentes variables contrapuestas, señalando la correlación entre cada par y aumentando la intensidad de su tonalidad anaranjada de acuerdo al nivel de la misma, siendo blanco el nivel más bajo de asociación y marrón el más alto, con la que se busca identificar patrones de comportamiento, este procedimiento se encuentra en el **Apéndice E.**

[4] Biblioteca para la generación de gráficos a partir de datos contenidos en listas o arrays en el lenguaje de programación Python y su extensión matemática NumPy.

Figura 5

Mapa de Calor



A partir de la revisión de literatura previa se espera encontrar fuertes correlaciones entre los gastos de: arriendo, total de personal y fletes, debido a que usualmente una empresa con superficie amplia o múltiples franquicias requerirá mayor cantidad de empleados y a su vez una flota para distribución de alimentos amplia, dentro del mapa de calor se encuentra un coeficiente de correlación fuerte entre las variables de arriendo-transporte del 0.71.

Se presenta a su vez una agrupación de correlaciones fuertes entre el total de personal, el total de remuneraciones, los servicios internos y los servicios externos; la justificación del alto

nivel de correlación del total de personal con el total de remuneraciones causadas por la empresa resulta ser sencilla debido a que toda organización entre mayor sea su número de empleados mayores gastos devengará para cumplir con pagos en nómina, además de esto, los servicios externos son gastos que representan reparaciones técnicas tanto del personal de la empresa como las contrataciones externas temporales que no solo aumentan los gastos en remuneraciones sino que también aumentan el total de personal temporalmente y cuando una empresa cuenta con un mayor nivel de personal y un mayor tamaño en superficie aumentan el consumo de servicios internos entre los cuales se encuentran los servicios básicos.

En cuanto a la publicidad se produce un comportamiento particular respecto a las otras variables debido a que esta se encuentra únicamente correlacionada con la variable transporte y esto puede ser interpretado como una correlación espuria, además, contrario a lo esperado y a lo encontrado en la literatura las ventas no aumentan conforme con la inversión en publicidad.

Al hablar de costos se puede ver que se encuentra altamente correlacionado al inventario promedio manejado por la empresa, la literatura lo confirma al indicar que uno de los principales costos manejados por las organizaciones es el inventario, debido a malos manejos del mismo y además que estos costos disminuyen conforme la empresa adquiere experiencia y un mayor tamaño.

Las ventas se encuentran asociadas con un nivel alto de 0,86 y 0,84 a los costos y el inventario promedio respectivamente, lo que se debe al costo de fabricación el cual aumenta

proporcionalmente a medida que lo hacen las ventas y a su vez se incrementa el nivel de inventario necesario para cumplir con las demandas.

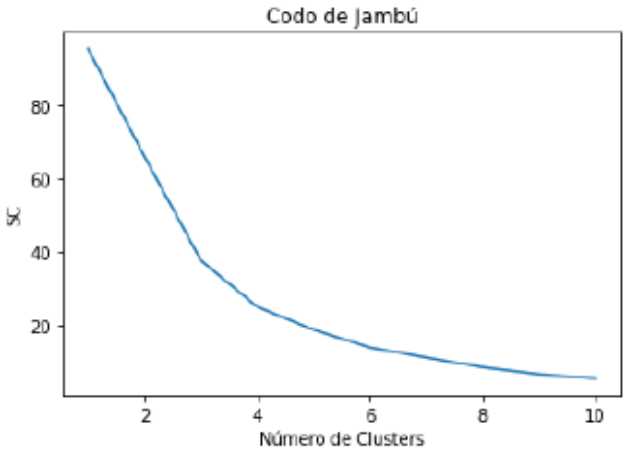
Se presenta una correlación considerable entre costos y el total de remuneraciones, algo que posiblemente esté relacionado con la totalidad del personal y las remuneraciones que devengan, siendo este uno de los gastos más elevados en este grupo de empresas.

Es importante ver cómo muchas de las relaciones encontradas entre algunas variables están en concordancia con lo esperado o con los resultados que arrojan otras investigaciones en el tema, este análisis permite tomar como posibles otras asociaciones que no se esperaban y puedan estar siendo claves en el comportamiento del sector de la economía que se está tratando en esta investigación.

Como siguiente medida una vez graficado el mapa de calor y luego de tener una visión global de las relaciones entre las variables se procede con el análisis por conglomerados aplicando el método de k-means soportado por Python, para su adecuado uso es importante conocer la cantidad óptima de agrupaciones a formar y esto es posible elaborando una gráfica por el método del codo de la librería Matplotlib, la *figura 6* muestra en el eje x el índice que corresponde a la cantidad de agrupaciones y en el eje y su correspondiente valor respecto a qué tan similares son los individuos en cada cluster, de donde se toma el valor *n* del número de cluster identificando el punto de inflexión o de mayor sedimentación en el codo de Jambú.

Figura 6

Codo de Jambú



Al contar con la cantidad de conglomerados a trabajar se procede a hacer uso del método k-means, obteniendo como resultado una solución de 5 grupos heterogéneos entre sí. El método k-means le asigna un grupo específico asociado a cada una de las observaciones o empresas en el dataframe, los cuales se introducen al crear una nueva variable a la que se le asigna el nombre "grupos", este procedimiento se puede ser consultado en el **Apéndice E.**

Para el análisis se realiza un agrupamiento con ayuda de la función grupby[5] de la librería NumPy[6], para conocer el número de empresas presentes en cada conglomerado y en cada actividad comercial (comercio al por menor para alimentos especializados y para no especializados) (*figura 7*). Los cluster número 1,2 y 3 no cuentan con un número de datos representativo, contando estas con entre 7 y 20 empresas máximo.

[5] Instrucción que divide las filas de resultados en grupos, según sus valores en una o varias columnas.
[6] Biblioteca para el lenguaje de programación Python que da soporte para crear vectores y matrices grandes multidimensionales.

Figura 7

Agrupamiento para la totalidad de empresas por actividad económica y por cluster

Cluster	CORRELA	ARRIENDO	SERV_INT	PUBLICI	TRANSP	SERV_EXT	TOTAL_REM	VENTA	TOT_PERSO	INV_PRO	CTO
0	472	2924	2924	2924	2924	2924	2924	2924	2924	2924	2924
	4711	595	595	595	595	595	595	595	595	595	595
1	472	10	10	10	10	10	10	10	10	10	10
	4711	11	11	11	11	11	11	11	11	11	11
2	4711	12	12	12	12	12	12	12	12	12	12
3	472	3	3	3	3	3	3	3	3	3	3
	4711	4	4	4	4	4	4	4	4	4	4
4	472	87	87	87	87	87	87	87	87	87	87
	4711	24	24	24	24	24	24	24	24	24	24

Se elaboran gráficas boxplot[7] por medio de Matplotlib, se calculan medidas de estadística descriptiva, se revisan las correlaciones entre variables y se desarrolla un análisis para identificar qué tipo de organizaciones están contenidas dentro de cada cluster teniendo en cuenta su actividad comercial, ver **Apéndice E.**

Una observación inicial simple señala una clara diferencia en la totalidad de observaciones que se encuentran en cada cluster, donde el cluster 0 carga con el 96% del total de datos, lo que representa 3519 empresas de las cuales es importante notar que en su mayoría son organizaciones con surtido de alimentos especializados con una cantidad de 3011 y tan solo 508 con surtido no especializado; esto como un indicio de la distribución de empresas en el sector.

[7] Diagrama de caja

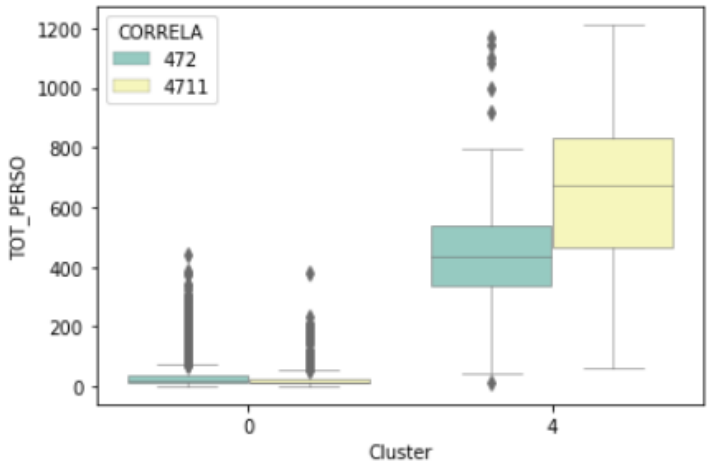
A nivel general el cluster 4 presenta una cantidad promedio de gastos muy superior al cluster 0 pero contiene una cantidad de empresas mucho menor, lo que puede indicar en una primera instancia que los dos conglomerados pueden estar siendo principalmente diferenciados por el tamaño de las organizaciones que representan. Al hacer uso de medidas de estadística descriptiva se encuentra que el valor promedio en arriendo para el primer cluster es de $293.394[8] al año frente a $2´206.262[8] al año para el segundo cluster de empresas lo que se ve traducido en 7,52 veces más el arriendo del cluster 0 y una diferencia inicial en tamaño de superficies bastante importante.

Un indicador de tamaño en una organización se ve representado por el total de remuneraciones pagadas a empleados anualmente y por supuesto el número de empleados que es manejado por dichas empresas, en el cluster 0 el promedio de empleados para empresas comercializadoras de alimentos tanto especializadas como no especializadas es de 64 empleados con un total de remuneraciones promedio de $525.641[8] al año, lo que es aproximadamente 17,31 veces menos que los 526 empleados que maneja el cluster 4 con $8’705.341[8] al año en remuneraciones pagadas, esta notable diferencia se puede ser apreciada con mayor claridad en la *figura 8* donde se presentan 4 diagramas boxplot o diagramas de caja y bigotes, cada uno de estos representa a la totalidad de empleados según el conglomerado en el eje *y*, y subdividiéndolo al mismo tiempo en los dos tipos de surtido o actividad económica, marcando en azul y amarillo, el rango de valores en el cual se encuentran los datos de cada grupo y la media de los mismo por una línea central.

[8] Valor en miles de pesos

Figura 8

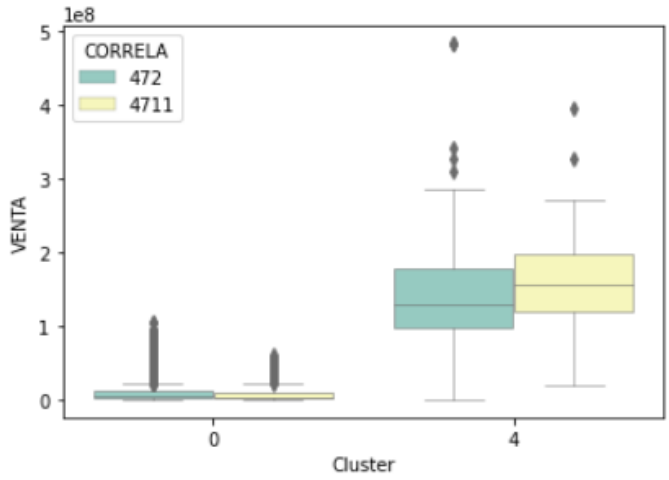
Boxplot para la totalidad de empleados por cluster



Una característica que genera una clara distinción entre estos conglomerados se encuentra en las ventas totales al año, con una diferencia del cluster 0 respecto al cluster 4 de 19,23 veces la cantidad de dinero recibido por ventas, con valores promedio de $8’330.002[8] contra $160’160.998[8]; esto evidencia junto con las variables nombradas anteriormente una clara diferencia en lo que respecta a variables que indican el tamaño de una empresa (*figura 9).*

Figura 9

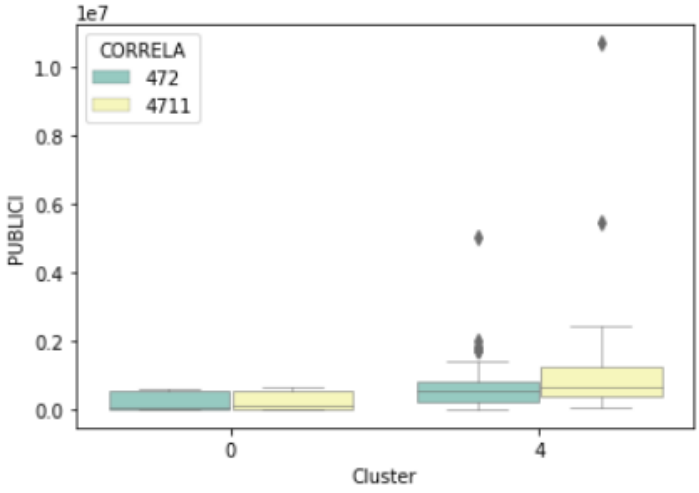
Boxplot para la totalidad de las ventas



La publicidad es la única variable de gasto en la que los conglomerados tienden a estar cerca, a partir del promedio se encuentra que el gasto en el cluster 4 con respecto al cluster 0 es 3,92 veces mayor, esto debido a que las empresas de menor tamaño sin importar si son especializadas o no especializadas presentan una mayor inversión frente a las empresas de mayor tamaño contenidas en cluster 4 esto se evidencia con claridad en la *Figura 10.* En promedio las empresas pequeñas o cluster 0 están invirtiendo casi el mismo monto en publicidad que lo gastado en arrendamiento para instalaciones, lo que viene siendo entre 252.079^{8} y 293.394^{8}.

Figura 10

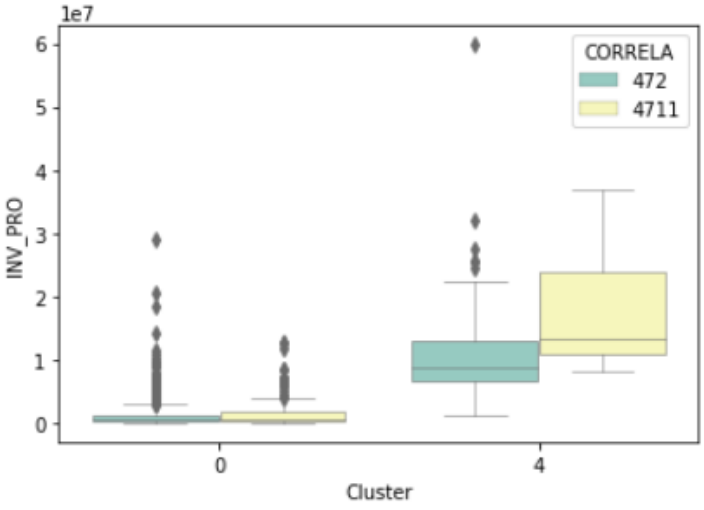
Boxplot para publicidad



A continuación al analizar los tipos de empresas que conforman los cluster, siendo estas, empresas minoristas con surtido de alimentos especializado y con surtido de alimentos no especializado; se encuentran diferencias en el manejo que le es dado a los inventarios, debido a que se aprecia cómo las empresas con **surtido especializado** tienen un menor promedio de gastos en inventario frente a las empresas con **surtido no especializado**, pues a partir de las cifras promedio encontradas para las empresas **especializadas** presentes en el cluster 0 con ($893.387[8]) y el cluster 4 ($11´036.822[8]), y para las empresas **no especializadas** del cluster 0 con ($1’104.454[8]) y del cluster 4 ($17’036.822[8]); se evidencia un mejor manejo de inventarios para las empresas con **surtido especializado**; se aprecia con mayor claridad en la *Figura 11*.

Figura 11

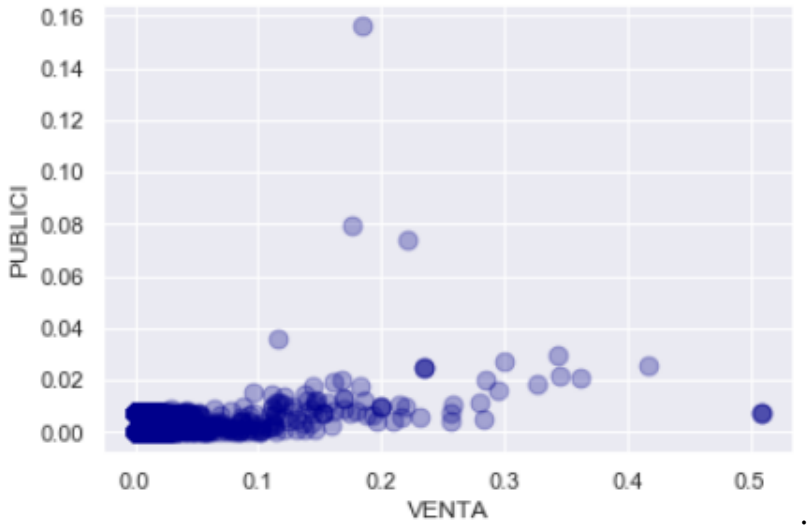
Boxplot para inventario promedio



En la variable publicidad analizada desde el tipo de surtido que manejan las empresas se puede resaltar que presenta valores bajos tanto para las empresas con surtido especializado como no especializado, lo especial de esto es que los niveles de venta promedio para ambas son muy similares pero la inversión en publicidad pasa de ser mayor cuando se trata de empresas en el cluster 0 como se indicó anteriormente al permanecer con cuantías bajas en el cluster 4, (*figura 12*), es decir, el gasto en publicidad para empresas especializadas es de $ 625.630[8] y para no especializadas es $ 1’349.761[8], donde las empresas especializadas invierten menos en publicidad que las empresas con surtido no especializado.

Figura 12

Comportamiento de la publicidad frente a las ventas



En cuanto al número de empleados se puede identificar a las empresas con surtido especializado como empresas que en un inicio requieren de más empleados cuando pertenecen al cluster 0, pero cuando son empresas consideradas grandes con surtido especializado el número de empleados que requieren frente a las empresas con surtido no especializado es bastante inferior con cantidades de 466 contra 635 respectivamente, aun generando un promedio total de ventas muy cercano (*figura 12*).

A nivel general las empresas con surtido no especializado presentan mayores gastos para la gran mayoría de variables sin importar si son de un cluster u otro y sin generar más ventas en una proporción considerable que validen estos gastos mayores, lo que puede indicar que las empresas con surtido no especializado tienen un nivel de rentabilidad menor.

7.2 Análisis de componentes principales (PCA)

Se realiza análisis de componentes principales debido a que las variables de entrada no presentan independencia, mediante el análisis se logra reducir las dimensiones del modelo y; de este análisis resultan ‘componentes principales’ linealmente no correlacionadas para así construir el modelo de regresión; primero se encuentra la media de los datos y se busca la dirección con mayor varianza, ver **Apéndice E**; los datos se encuentran previamente normalizados tras pasar por el análisis de cluster y se halla la matriz de covarianza asociada a estos datos utilizando la función ‘cov’ de Numpy con la matriz de datos transpuesta, esta matriz de covarianza se requiere para realizar una descomposición con la finalidad de obtener los valores que maximizan la varianza expresados en autovalores y autovectores, estos últimos contienen la información de cada nueva componente; se inicia el análisis con 9 variables independientes y la variable respuesta ‘VENTA’, con 3630 registros de cada variable.

Con el fin de obtener los valores propios (autovalores) los cuales proporcionan el número de componentes óptimos a los que se pueden reducir las variables y los vectores propios (autovectores) de cada nueva componente, se utiliza el módulo ‘linalg’ de Numpy, la función ‘eigh’ y la matriz de covarianza que retorna una matriz simétrica de autovectores y autovalores.

Tabla 4

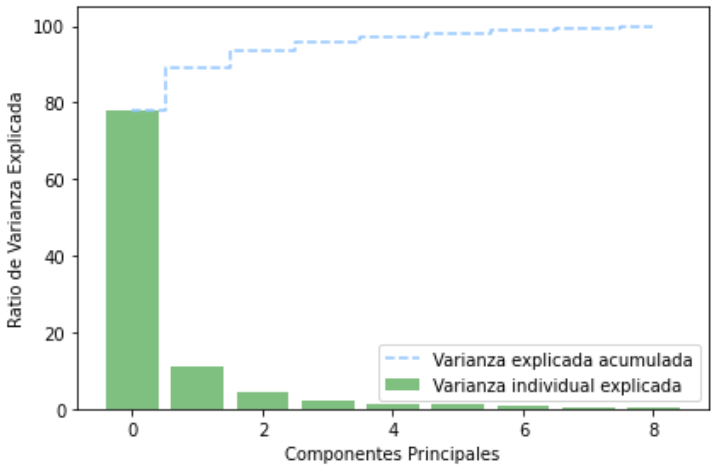
Porcentaje acumulado de autovalores

COMPONENTE	AUTOVALOR	%	% ACUMULADO
1	0,002191337	78,2%	78,2%
2	0,000309361	11,0%	89,2%
3	0,000123104	4,4%	93,6%
4	5,94E-05	2,1%	95,7%
5	3,69E-05	1,3%	97,0%
6	2,90E-05	1,0%	98,1%
7	2,81E-05	1,0%	99,1%
8	1,45E-05	0,5%	99,6%
9	1,19E-05	0,4%	100,0%
TOTAL	0,002803697		

A partir de los valores propios se halla la proporción de variabilidad explicada por cada componente, buscando que la varianza que explique cada componente sea la máxima posible, se ordenan de mayor a menor magnitud de su autovalor (*tabla 4*), también, a partir diagrama de barras *(figura 13)* que muestra la varianza explicada, se observa que con dos componentes principales se logra explicar más del 80% de la varianza de los datos, debido a que el primer componente explica el 78,2% y el segundo el 11%.

Figura 13

Porcentaje de varianza explicada por cada componente



Se conservan entonces dos componentes, los pesos de cada una de las variables de las dos componentes están dados por las magnitudes de los vectores propios o autovectores presentes en la *tabla 5*, se observa entonces que el componente principal 1 está explicado por variables de gasto que están relacionadas con costos de las empresas minoristas como: costo de la mercancía vendida (CTO = 0,555504), inventario promedio (INV_PRO = 0,529401) y por las variables: total de personal, servicios externos y el total de remuneraciones con pesos de entre 0,356 y 0,329.

La componente principal 2 se encuentra explicada en su mayoría por gastos en inventario promedio (INV_PRO = 0,838916) y por otras variables como: los costos de mercancía vendida, el total de personal y los servicios externos, variables que en orden descendente aportan peso sobre la componente 2; encontrando que estas variables también pesan sobre la componente 1 lo cual podría estar indicando que son variables importantes en el sector del comercio al por menor de alimentos, lo que se comprobará luego al momento de llevar a cabo el análisis final para la distribución de gastos.

De manera particular se evidencia que el gasto en publicidad tanto para la componente 1 (PUBLICI= 0,036008) como para la componente 2 (PUBLICI= 0,011758) es el gasto que menos peso tiene sobre las componentes, esto se corrobora mediante la matriz de correlación entre variables y componentes, de tener una débil correlación para con las componentes será considerada la idea de eliminar este gasto del modelo, respaldado además por lo encontrado en análisis realizado para el cluster en donde se halló que de la media de gasto general de las diferentes organizaciones la publicidad contiene uno de los valores medios menos relevantes o más bajos de entre todos los gastos.

Tabla 5

Autovectores de cada componente

VARIABLE	CP1	CP2
ARRIENDO	-0,085813	-0,0499301
SERV_INT	-0,177035	-0,112615
PUBLICI	-0,036008	-0,011758
TRANSP	-0,137803	-0,099939
SERV_EXT	-0,341615	-0,281
TOTAL_REM	-0,329328	-0,262099
TOT_PERSO	-0,356372	-0,261235
INV_PRO	-0,529401	0,838916
CTO	-0,555504	-0,23456

Se halla entonces la correlación entre variables y componentes extraídos, si se encuentra una correlación débil o muy débil entre ambas componentes y variable, lo ideal es eliminar la variable del estudio pues esta no estaría representada por ninguna componente; haciendo uso del módulo 'sqrt' de Numpy, de los autovalores y autovectores y de la matriz de covarianza se hallan las correlaciones, en la (*figura 14*) se encuentra para la primer componente la variable publicidad ('PUBLICI'= -0,34) contiene una débil correlación, así mismo para la componente 2 ('PUBLICI'= -0,042) que presenta una correlación muy débil por lo que se decide excluir esta variable del modelo. Todos los procedimientos previamente enunciados se encuentran en el **Apéndice E.**

Figura 14

Correlación entre componentes y variables

Correlación entre las variables originales
y los componenetes extraídos:

	Comp 1	Comp 2
ARRIENDO	-0.595073	-0.130093
SERV_INT	-0.873545	-0.208785
PUBLICI	-0.339941	-0.0417093
TRANSP	-0.729578	-0.198805
SERV_EXT	-0.878082	-0.271383
TOTAL_REM	-0.881189	-0.263503
TOT_PERSO	-0.925331	-0.254861
INV_PRO	-0.858302	0.511036
CTO	-0.937791	-0.148782

7.3 Análisis de bondad de ajuste de las distribuciones.

Una vez realizado el análisis de componentes principales es importante conocer qué tipo de distribución de probabilidad están siguiendo los datos que luego serán de utilidad al momento de establecer el modelo, por lo que se realizan pruebas de bondad de ajuste.

Los gastos elegidos para este análisis deben presentar una baja correlación entre sí, de tal manera que no se nublen los resultados a causa de asociaciones fuertes entre variables, para determinar cuáles variables usar durante la ejecución de los modelos de regresión se procede a realizar pruebas de hipótesis con un nivel de significancia al 0,05 y para establecer cuáles gastos presentan un coeficiente de correlación de Spearman significante entre cada par de variables.

$H_0: p - valor > 0{,}05$ Las variables no están correlacionadas (se acepta la hipótesis)

$H_0: p - valor \leq 0{,}05$ Las variables están correlacionadas (Rechaza la hipótesis).

Como se puede apreciar en el **Apéndice H**, existe una correlación significativa entre los pares de variables: total de remuneración con servicios externos y total de personal, transporte con publicidad, a su vez costos con inventario promedio, de la misma manera se presenta correlación significativa para servicios internos con el total de personal y finalmente los servicios externos con los servicios internos y el total de remuneraciones. De esta manera se decide eliminar en consecuencia a estas correlaciones una variable por cada par que presente una asociación la cual pueda afectar los resultados de la regresión y se procede a hacer la ejecución de los primeros dos modelos con tan solo 4 variables: arriendo, transporte, total personal y costo de la mercancía vendida. Además, debido a la gran cantidad de observaciones se hace un muestreo aleatorio simple tomando 347 datos y se procede a determinar la distribución de probabilidad de la variable independiente ‘VENTA’ mediante un análisis de regresión múltiple.

Para conocer si los cambios en estos gastos están o no asociados con los cambios en las ventas a partir del nivel de significancia establecido en 0,05 se plantea la siguiente hipótesis:

H_0: $valor\ p > 0{,}05$ Los cambios en la variable no están asociados con los cambios en las ventas (se acepta la hipótesis).

H_1: $valor\ p \leq 0{,}05$ Los cambios en la variable están asociados con los cambios en las ventas (se rechaza la hipótesis nula).

A partir del análisis de varianza (anova) (*tabla 6*) se concluye que se no rechaza la hipótesis nula para la variable transporte (p-valor = 0,056), lo que indica que las variaciones del gasto en transporte no difieren significativamente sobre las ventas, mientras que el resto de las variables como: arriendo, total personal y costo de la mercancía vendida, mostradas en el anova presentan cambios asociados con la variable ventas, debido a que su p-valor es menor que 0,05.

Tabla 6

Análisis de Varianza (ANOVA)

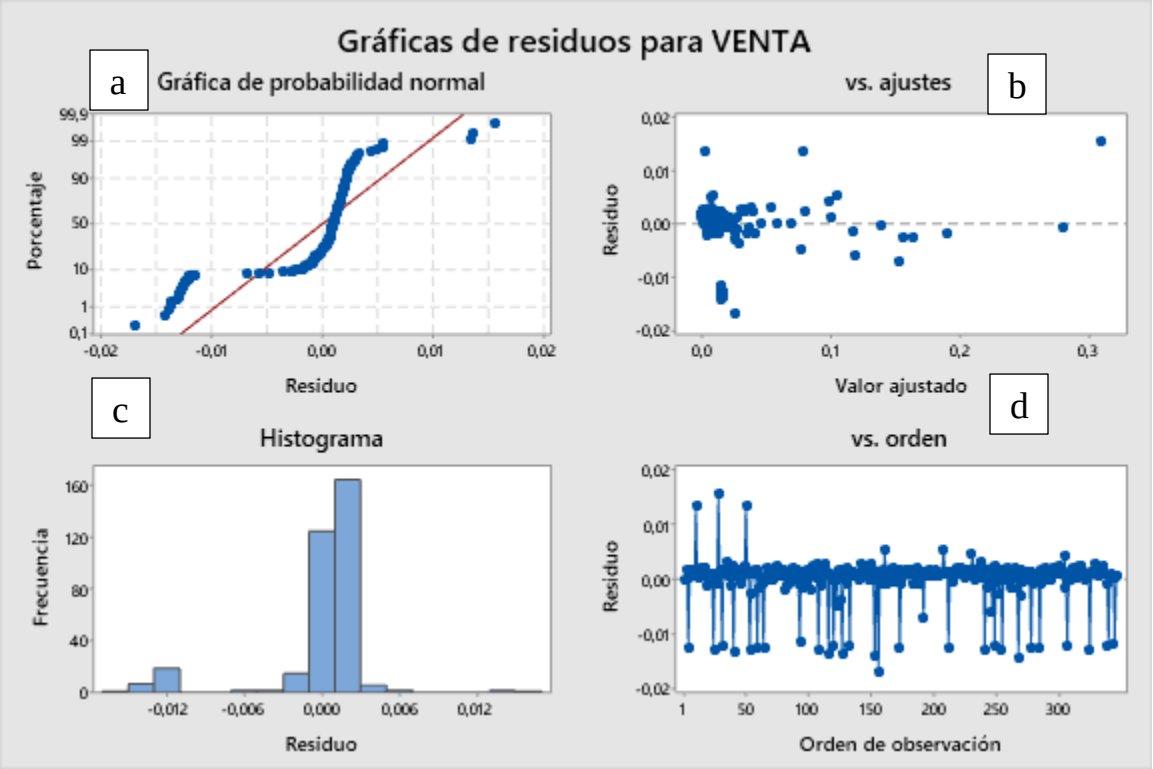
Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	4	0,372807	0,093202	5336,25	0,000
ARRIENDO	1	0,000160	0,000160	9,16	0,003
TRANSP	1	0,000064	0,000064	3,69	0,056*
TOT_PERSO	1	0,002267	0,002267	129,80	0,000
CTO	1	0,033420	0,033420	1913,48	0,000
Error	342	0,005973	0,000017		
Falta de ajuste	308	0,005964	0,000019	73,41	0,000
Error puro	34	0,000009	0,000000		
Total	346	0,378780			

*las variables presentan un p-valor mayor al nivel de significancia (0,05)

El análisis revela un R-cuadrado de 98,42% (**Apéndice K**) por lo que se entiende que la mayoría de los gastos presentes en la regresión explican gran parte de la variabilidad de las ventas en torno a su media, para verificar esto se revisan las gráficas de residuos (*figura 15*), en la gráfica de probabilidad normal (*figura 15a*) se observa que los residuos no se distribuyen de forma normal y se presentan algunos valores atípicos, en la gráfica de residuo vs. orden (*figura 15d*) se observa que los residuos se ubican aleatoriamente alrededor de la línea central, no parecen seguir una tendencia clara y en cuanto a la gráfica de residuos vs. ajuste (*figura 15b*) los puntos no parecen estar dispersos de manera aleatoria y presentan heterocedasticidad, por lo que se concluye que los datos no siguen una distribución normal.

Figura 15

Gráficas de residuos para ventas



Para conocer entonces qué tipo de distribución pueden estar siguiendo los datos se realizan pruebas de bondad de ajuste a la variable respuesta (‘VENTA’) con la herramienta de ‘Identificación de la distribución individual’ con un nivel de confianza del 95% (**apéndice K)**, a partir de las gráficas de cada una de las distribuciones se evidencia que las ventas no siguen una distribución normal (*figura 16*), la distribución que puede estar siguiendo esta variable es una log-logística de 3 parámetros (*figura 17*) aun así no se puede concluir esto debido a que los datos presentan una curvatura y se presentan algunos valores que podrían ser atípicos.

Figura 16

Gráfica de probabilidad normal

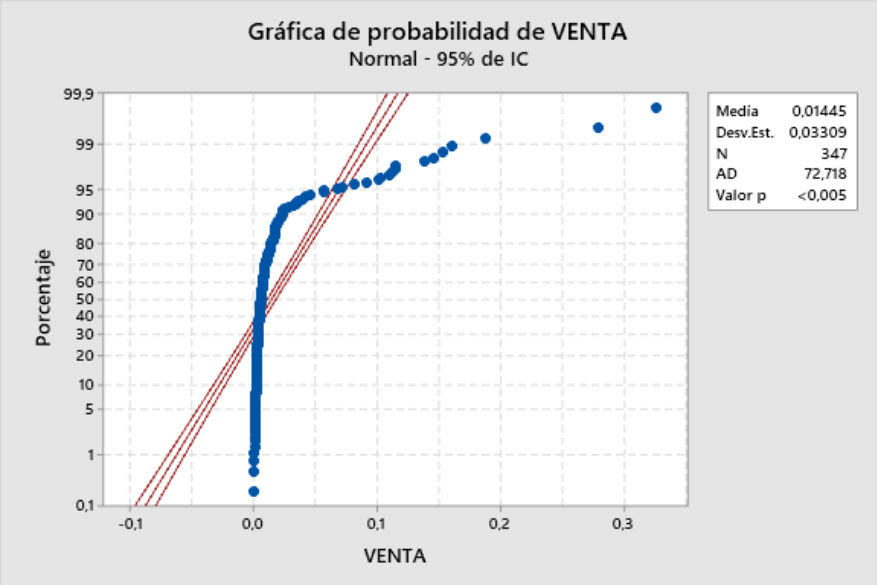
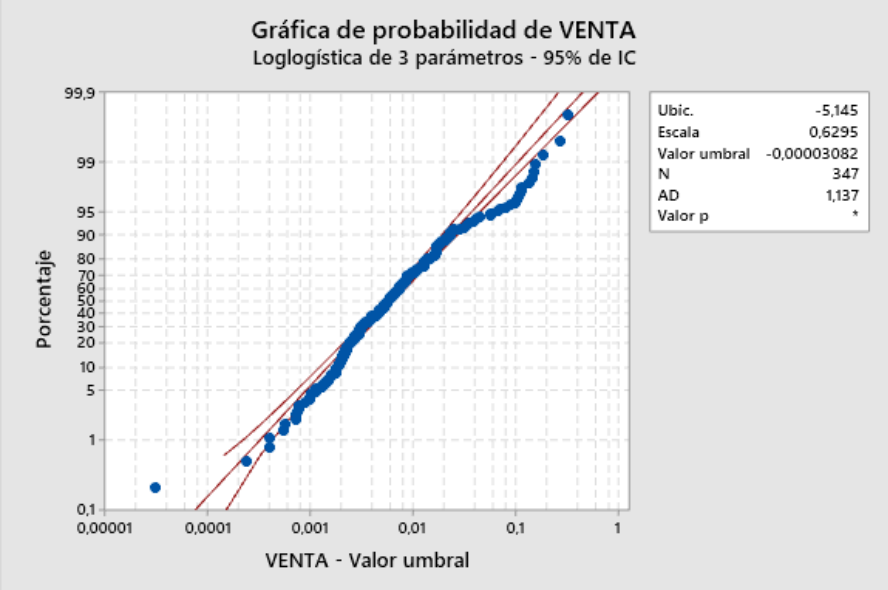


Figura 17

Gráfica de probabilidad log-logística de 3 parámetros



7.4 Formulación y ejecución del modelo de regresión

Una vez realizada una revisión de los tipos de distribución y encontrando que no hay una distribución que defina con claridad el dataframe como la distribución normal o Gamma, se plantea la ejecución del modelo por medio de una transformación logarítmica de la variable dependiente 'VENTA', de tal manera que se hace una multiplicación por logaritmo al valor original para las ventas y se vuelve a realizar una normalización de la misma, de esta forma se procede con la formulación del modelo, donde se planea su ejecución tanto para las componentes principales como para las variables en cuestión que cuya correlación no sea significativa, la transformación se encuentra en el **Apéndice E.**

Inicialmente se plantean diferentes modelos lineales generalizados por medio de la librería Statsmodels[9] para diferentes tipos de distribución sin hacer uso de la transformación logarítmica, con los cuales se busca encontrar pruebas que permitan definir de qué manera se encuentran repartidos los datos, teniendo en cuenta que tan bien explican las variables independientes a la variable respuesta por medio del coeficiente de determinación y el error medio arrojado por los modelos, de lo cual se obtiene como resultado errores altos, por tal motivo se procede a hacer una transformación a la variable respuesta.

Se usan de dos métodos de formulación de modelos que trabajan por mínimos cuadrados, de la librería Statsmodels (OLS Regression) y de Sklearn[10] (Linear Regression) los cuales deben arrojar resultados similares y de esta manera permitir probar la concordancia de los modelos para después probar su ajuste. Inicialmente se cargan 2 bases de datos diferentes para así generar finalmente 4 modelos.

[9] Paquete de Python que permite a los usuarios explorar datos, estimar modelos estadísticos y realizar pruebas estadísticas.

[10]Biblioteca para aprendizaje automático de software libre para el lenguaje de programación Python.Incluye varios algoritmos de clasificación, regresión y análisis de grupos

El primer par de modelos a desarrollar contiene únicamente una selección de las variables trabajadas durante todo el análisis previamente ejecutado. El segundo par de modelos se llevó a cabo con las dos componentes principales que describen más del 80% de la variabilidad de los datos. Ambas bases de datos se encuentran en los **Apéndices F y G.**

Una vez formulado y ejecutado el primer modelo el cual contiene variables no correlacionadas se procede a realizar: análisis de coeficientes, significancia y ajuste haciendo la predicción de los valores de las ventas y enfrentándolas a los valores reales para cada observación por medio de la prueba de R cuadrado de la librería Sklearn.metrics de Python y el error cuadrado medio. Este proceso fue ejecutado para cada uno de los modelos generados.

Es importante señalar que los análisis de coeficientes y p-valor realizados se llevan a cabo principalmente con el modelo de la librería Statsmodels (OLS Regression) debido a que presenta una disposición de datos más ordenada y cuenta con los cálculos para determinar la significancia de cada variable en el modelo, sin embargo, para denotar la igualdad de los pares de modelos los cálculos para el coeficiente de correlación y error medio cuadrado se encontrara en el código por medio de la librería Sklearn (LinearRegression), los modelos realizados se encuentran conjunto en el **Apéndice H.**

Figura 18

Modelo de mínimos cuadrados ordinarios por Statsmodel para 4 variables independientes

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  VENTA   R-squared:                       0.449
Model:                            OLS   Adj. R-squared:                  0.449
Method:                 Least Squares   F-statistic:                     739.6
Date:                Sat, 09 Jan 2021   Prob (F-statistic):               0.00
Time:                        20:00:44   Log-Likelihood:                 3806.5
No. Observations:                3630   AIC:                            -7603.
Df Residuals:                    3625   BIC:                            -7572.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4664      0.002    281.320      0.000       0.463       0.470
0             -2.6142      0.264     -9.912      0.000      -3.131      -2.097
1             -1.4120      0.230     -6.127      0.000      -1.864      -0.960
2              3.2227      0.176     18.320      0.000       2.878       3.568
3              1.4115      0.103     13.683      0.000       1.209       1.614
==============================================================================
Omnibus:                      521.151   Durbin-Watson:                   1.772
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1010.973
Skew:                          -0.892   Prob(JB):                    2.95e-220
Kurtosis:                       4.871   Cond. No.                         197.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
Error:  0.08479059575374294
R2:  0.4493663625901838
```

El primer modelo elaborado revela en los p-valores para cada uno de los coeficientes que son menores a el nivel de significancia establecido (0,05) ver en **Apéndice H.** Se realiza un análisis de las coeficientes (*figura 18*) cuyos resultados aprecian también en la *ecuación 13*, se encuentra lo siguiente: se espera una disminución de las ventas en relación con el arriendo (-2,6142), los gastos en esta variable dependen de la superficie que ocupan y de la ubicación, pues para el mercado minorista es estratégico contar con una excelente ubicación debido a que este gasto podría tener afectación en las utilidades y en el flujo de caja, el gasto en transporte con respecto a las ventas presenta una disminución (-1,4120), el DANE no establece a qué tipo de transporte refiere

la variable, se desconoce si se trata del transporte que debe pagarse a la hora de hacer un pedido o del pago al momento de realizar envío del producto al cliente, en todo caso el transporte puede acarrear una disminución en las ventas debido a que en relación con que deban pagar por un pedido de abastecimiento más alto causa que el producto deba ofrecerse a un valor mayor lo cual podría generar menos entradas por ventas, así mismo cuando los costos de transporte se elevan y no se realiza una planeación previa para contar con suficiente pedido se corre el riesgo de desabastecimiento del producto lo que se traduce en menos disponibilidad para las ventas, y en relación con el hecho de pagar los envíos para hacerlo llegar al cliente puede que sean los minoristas quienes corren con los gastos de embalaje y además pagan fletes de envío, lo que no disminuiría las ventas pero si registraría una entrada por ventas menor a la esperada, esto debido posiblemente a una errónea planeación en los costos por parte del área encargada.

$$Ventas = -2{,}6142(Arriendo) - 1{,}4120(Transporte) + 3{,}2227(Tot.personal) + 1{,}4115(Costo\ de\ mercancía\ vendida) + 0{,}4664 \quad (13)$$

En cuanto al total del personal el cual contiene un coeficiente de (4,3255), aporta al aumento en ventas, pues como se ha encontrado en la revisión de literatura realizada previamente el factor recurso humano aporta a la rentabilidad de las empresas minoristas, contar una óptima cantidad de personal disponible dentro del mercado minorista le permite al personal enfocarse en tareas específicas, hacerlas correctamente, además contar con el personal necesario para atender la demanda ocasiona una mejor atención al cliente debido a que mejora su experiencia de compra.

Las ventas netas y el costo de la mercancía vendida son posiblemente los principales indicadores que generan un alto impacto en la rentabilidad del mercado minorista, pues permiten evaluar de manera significativa los costos de la mercancía vendida, determinando el margen bruto de ganancia lo que da como resultado un aporte positivo del costo de la mercancía vendida para con las ventas (1,4525).

Figura 19

Modelo de mínimos cuadrados ordinarios por Statsmodel para componentes del PCA (2)

GLS Regression Results

Dep. Variable:	VENTA	R-squared:	0.384
Model:	GLS	Adj. R-squared:	0.384
Method:	Least Squares	F-statistic:	1131.
Date:	Fri, 18 Dec 2020	Prob (F-statistic):	0.00
Time:	14:35:44	Log-Likelihood:	3603.1
No. Observations:	3630	AIC:	-7200.
Df Residuals:	3627	BIC:	-7182.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4956	0.001	332.831	0.000	0.493	0.499
0	1.6672	0.036	46.184	0.000	1.596	1.738
1	-1.0372	0.091	-11.344	0.000	-1.217	-0.858

Omnibus:	503.771	Durbin-Watson:	1.692
Prob(Omnibus):	0.000	Jarque-Bera (JB):	835.187
Skew:	-0.934	Prob(JB):	4.38e-182
Kurtosis:	4.426	Cond. No.	61.4

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Error: 0.08967707968225035
R2: 0.38407152847810244

Se procede a realizar el análisis para los modelos generados con ayuda de las dos componentes principales, ver **apéndice H**. En cuanto a la componente 1 la cual ocupa el 78,2% de la varianza explicada al contar con un valor menor a 0,05 es una variable significativa en el modelo;

dicha variable contiene a su vez un coeficiente igual a 1,6672 (*figura 19*), como se encontró en el cuadro de correlación (*tabla 5*) esta componente tiene relación fuerte e inversa con variables como el costo de la mercancía vendida (-0,938) y el total del personal (-0,925), variables que como se explicó anteriormente pueden aportar de manera positiva a las ventas de los minoristas; esto señala el peso de la variable en la *ecuación 14* generada por el modelo, se ve de la siguiente manera:

$$Ventas = 1{,}6672(Componente\ 1) - 1{,}0372(Componente\ 2) + 0{,}4956 \qquad (14)$$

Finalmente se obtiene que la componente número 2 la cual contenía el 11% de la variabilidad según el PCA es una variable significativa que tiene un coeficiente igual a -1,0372 (*figura 19*), como se había encontrado antes en el análisis de los componentes principales dentro de la componente las variables que tienen más peso sobre la componente y que igualmente presentan una correlación más alta que el resto son los gastos en inventario promedio y servicios externos, la disminución que presenta esta coeficiente puede deberse a la afectación que presenta cada una de las variables sobre la variable respuesta como ya se analizó anteriormente.

Es fundamental realizar un análisis de regresión por medio de coeficientes para las variables que no presentan correlaciones altas y para los componentes principales resultantes del PCA debido a que aportan a la comprensión del comportamiento de los principales gastos en el sector minorista en Colombia a nivel nacional.

Por otro lado al contrastar ambos pares de modelos se puede apreciar como al implementar variables que no presenten correlación omite parte del análisis de gastos que aunque estén muy asociados y se espera se comporten de la misma manera que sus variables pares valdría la pena tener en cuenta variables como el inventario promedio, pues según la literatura esta contiene una

relación importante con las ventas, la administración del inventario depende de la naturaleza de la organización, en este caso en la industria minorista el inventario se convierte en un gasto necesario para la empresa, contar con un óptimo nivel en el inventario se traduce en tener bienes de forma precautelar teniendo en cuenta que la mayoría de los productos son perecederos, generan un costo de mantenimiento y además de considerar algunos factores que podrían afectar las ventas a la hora de tener inventario, como ciclos de orden, demoras en reabastecimiento y cantidad de artículos a pedir; garantizan la capacidad de abastecer la demanda y vender.

El total de remuneraciones devengado por una empresa tiene un efecto en las ventas, lo que puede significar en ámbitos de aplicación un acuerdo de la empresa que permita contemplar el valor mensual de cada trabajador generando un contrato entre las partes que ayuden llenar un portafolio de nóminas en el cual se consideren compensatorios remunerados o retribución en dinero tanto para la empresa como para el cliente interno (empleados) y el cliente externo, otra causa de la relación logra estar explicada por la alta rotación que se puede estar dando en el sector minorista o por la incorrecta planificación en la fuerza laboral.

8. Evaluación y validación

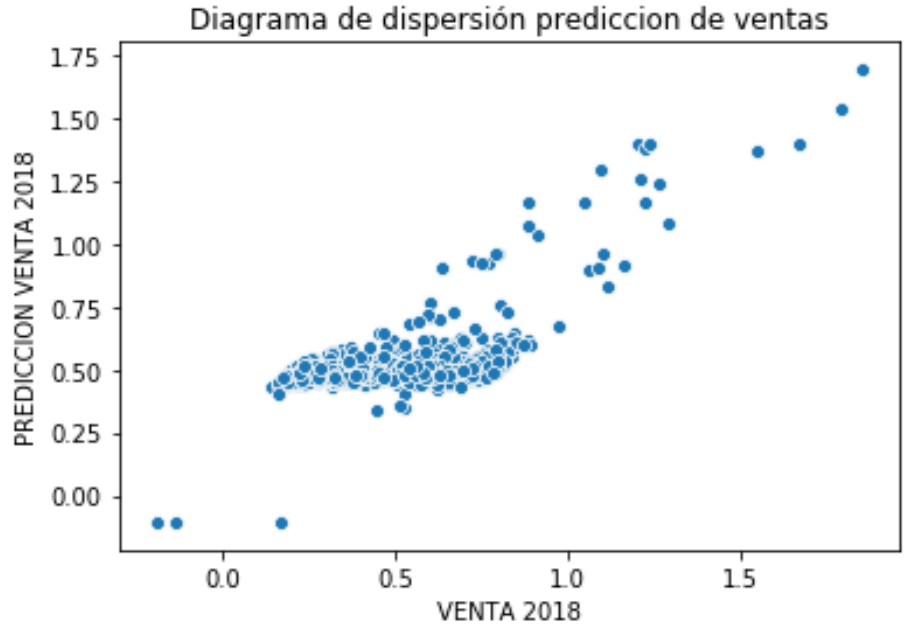
Como última fase y luego de la evaluación de los resultados de la regresión lineal múltiple y del análisis de cada una de las variables de gastos, se lleva a cabo una validación del modelo con los datos entregados por el DANE para el año 2018 como fue planteado en la metodología, a dichos datos se les realiza el mismo tratamiento que a las 3630 observaciones de los 4 años evaluados en el modelo de regresión los cuales fueron: limpieza, imputación, transformación y normalización, la matriz de datos tratados se puede apreciar en el **Apéndice I.**

Se busca determinar cómo se desempeñaría el modelo como un posible método de predicción de ventas para años posteriores, esto se lleva a cabo inicialmente haciendo una transformación logarítmica a la variable respuesta ventas para que de esta manera se ajuste con mayor efectividad en el modelo ya ejecutado previamente; una vez tratados los datos e insertada la matriz de datos se procede a definir las variables independientes y la variable respuesta dentro de Python.

Se calcula una predicción sobre el modelo ya planteado haciendo uso de las variables independientes o gastos para el año 2018, lo cual entrega un arreglo de 808 valores de venta u observaciones, estas predicciones son enfrentadas a los datos originales, *figura 20* y se analiza por medio de métricas de bondad de ajuste, tanto por R cuadrado como el error cuadrado medio, este proceso se encuentra en el **Apéndice J.**

Figura 20

Diagrama de dispersión predicción de ventas



Como resultado se obtiene (*figura 21*) un valor de Error cuadrado medio = 0,1746 lo que frente al Error = 0,0847 presentado por la evaluación del modelo para la totalidad de los datos de los 4 años anteriores representa una disminución en la precisión del estimador. Además, entrega un valor para prueba de R cuadrado = 0,3524 lo que se ve traducido como un 35,2% de ajuste para los datos pronosticados por el modelo frente al valor real de la variable dependiente ventas.

Figura 21

Validación de datos, modelo final para el año 2018

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  VENTA   R-squared:                       0.449
Model:                            OLS   Adj. R-squared:                  0.449
Method:                 Least Squares   F-statistic:                     739.6
Date:                Tue, 12 Jan 2021   Prob (F-statistic):               0.00
Time:                        11:03:48   Log-Likelihood:                 3806.5
No. Observations:                3630   AIC:                            -7603.
Df Residuals:                    3625   BIC:                            -7572.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4664      0.002    281.320      0.000       0.463       0.470
0             -2.6142      0.264     -9.912      0.000      -3.131      -2.097
1             -1.4120      0.230     -6.127      0.000      -1.864      -0.960
2              3.2227      0.176     18.320      0.000       2.878       3.568
3              1.4115      0.103     13.683      0.000       1.209       1.614
==============================================================================
Omnibus:                      521.151   Durbin-Watson:                   1.772
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1010.973
Skew:                          -0.892   Prob(JB):                    2.95e-220
Kurtosis:                       4.871   Cond. No.                         197.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
RESULTADOS DE LA PREDICCION PARA AÑOS 2014-2017
Error:  0.08479059575374294
R2:  0.4493663625901838

RESULTADOS DE LA PREDICCION PARA EL AÑO 2018 CON EL MODELO DE REGRESION FINAL
Error:  0.1746620848707464
R^2:  0.3521854400926139
```

Al presentarse una disminución del porcentaje de ajuste inicial del modelo de 44,94%, al ajuste para los datos del año 2018 de 35,2% se concluyen que la precisión del modelo para

predicción de valores de venta es menor pero igualmente cercana a la original, resultado que es de esperarse debido a que al ingresar datos no conocidos por el modelo este generara un error en la predicción mayor al inicial, esto se puede deber a variadas situaciones, entre las que se encuentra un notorio cambio en la media para los valores de costo de mercancía vendida o (CTO) para el año 2018, a su vez, se identifica una disminución en la media para los costos de transporte, siendo estas dos variables significantes en el modelo y con un peso que puede generar importantes variaciones y cuyas media en estos valores no presentan una disminución continua a través de los años.

Finalmente no se puede tomar como un modelo definitivo para hacer pronósticos concretos o permisibles en la aplicación en la industria minorista para valores de ventas debido a que aunque los valores resultantes son congruentes con lo esperado el nivel de ajuste no es muy alto, se debe resaltar la importancia que tiene las bases de datos iniciales para el desarrollo de modelos estadísticos predictivos, debido a que el alto número de valores atípicos, nulos o inválidos que se presenta en las mismas afectan la inocuidad de los resultados.

9. Conclusiones

En el presente trabajo, se presenta el desarrollo de modelos y diferentes herramientas estadísticas que permite entender detalladamente las dinámicas de un sector de la economía, de tal forma que la información pueda ser usada e interpretada para facilitar la toma estratégica de las decisiones por parte de las personas y organizaciones.

Una herramienta útil es el mapa de calor que permite visualizar las correlaciones entre las diferentes variables de interés para un gerente, específicamente en este trabajo son los gastos, el total de personal, el total de remuneraciones y los servicios tanto internos como externos con los cuales se construyen nuevas métricas todo plasmado en un mismo tablero.

Con base en el análisis de conglomerados se establece una tipología entre las organizaciones consideradas pequeñas a las empresas de mayor tamaño, tanto en niveles de ingresos como de gasto. La publicidad deja de ser un gasto significativo conforme la empresa aumenta considerablemente su tamaño y niveles de venta. En cuanto a los tipos de surtido en las empresas del sector se identifica que las empresas con surtido de alimentos no especializados requieren de una mayor cantidad de trabajadores que las empresas no especializadas, además, tienen mayores gastos a nivel general. Finalmente se evidencia cómo las empresas de menor tamaño tienen un mayor control sobre su inventario y más específicamente las empresas con surtido de alimentos especializados frente a las empresas no especializadas.

Teniendo en cuenta el análisis componentes principales que describen los gastos a tener en cuenta o los que tienen mayor peso para las finanzas en las organizaciones, donde se denota una importancia especial al costo producción de la mercancía vendida y al manejo del inventario, los cuales tienen correlación directa pues al aumentar las ventas aumenta proporcionalmente la cantidad de costos de producción y con un aumento en la producción se generan mayores

necesidades al momento del manejo de los inventarios, por tal motivo se identifica como un punto a tener en cuenta en toda organización.

Adicionalmente se puede ver como los gastos en el total de personal, el número de empleados y servicios externos componen un grupo más de gastos estrechamente relacionados y a su vez identificado como un grupo de interés, donde es sabido que a mayor número de empleados mayor cantidad de dinero se debe remunerar por parte de la empresa y de la misma forma la contratación de servicios externos para mantenimiento y reparación de maquinaria o equipo genera un aumento temporal en el número de empleados y costos.

Se identifica a la publicidad como el único gasto que no es realmente representativo o determinante en las finanzas de una compañía, sin embargo, esta presenta un monto considerable cuando las empresas son de menor tamaño, pues las grandes empresas en proporción presentan un gasto en publicidad mínimo.

Se determina que existe una disminución de las ventas en relación con el arriendo lo que se debe ver traducido como un esfuerzo adicional por parte de las organizaciones en el sector por adquirir instalaciones con precios ventajosos y ubicaciones estratégicas, además de esto, se identifica una relación negativa en los resultados de las empresas pertenecientes al sector minorista de alimentos con los costos en transporte o distribución y las remuneraciones por servicios técnicos.

El análisis de la literatura y los resultados obtenido de la modelación nos permite afirmar que el total de personal manejado por las empresas juega un rol importante, al ser el recurso humano que aporta a la rentabilidad en este tipo de organizaciones, pues permite enfocar fuerza laboral en las áreas de la empresa con mayores requerimientos y para atender la demanda variable.

En resumen, no es posible hacer una validación al modelo de regresión final para su uso en la industria en el cálculo de ventas totales ya que el pronóstico generado por el mismo aunque es concordante no presenta un nivel de efectividad lo suficientemente alto a la hora de tomar decisiones estratégicas, esto se debe a la relevancia que tienen las bases de datos iniciales para el desarrollo de este tipo de modelación predictiva, ya que una alta cantidad de valores atípicos o inválidos afectan directamente la empleabilidad de los resultados finales.

10. Recomendaciones

Una vez finalizado este trabajo de investigación se considera indagar sobre otros aspectos relacionados con el análisis de datos y se propone para futuras investigaciones, trabajar con datos de la encuesta de años más recientes además de incluir otras variables presentes en esta, debido a la pérdida de información que se puede estar dando de años pasados o a los cambios en los mecanismos de recopilación de la información.

Se recomienda hacer una recopilación de datos de fuentes de información distintas a las del DANE, así mismo realizar investigaciones similares a otros sectores importantes de la economía nacional, para llevar a cabo una comparativa.

Debido a la evolución del mercado el comercio electrónico ha comenzado a tomar fuerza durante el último año por lo que se recomienda para futuras investigaciones relacionadas con la rentabilidad minorista incluir variables que estén relacionadas con la Tecnología de la Información y de la Comunicación (TIC) debido a que el aumento del uso de esta tecnología ha comenzado a impactar la rentabilidad de las empresas.

Finalmente se recomienda, realizar el análisis de las variables teniendo en cuenta índices del desempeño de la economía como: el crecimiento del PIB, el desempleo, la inflación, entre otros.

Referencias Bibliográficas

Aigner, D., Lovell, C. A., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 21-37.

Almohri, H., Chinnam, R. B., & Colosimo, M. (2019). Data-driven analytics for benchmarking and optimizing the performance of automotive dealerships. *International Journal of Production Economics*, 69-80.

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2008). Estadística para administración y economía. México: Cengage Learning.

Assaf, A. G., Barros, C., & Sellers-Rubio, R. (2011). Efficiency determinants in retail stores: A bayesian framework. *Omega*, 283-292.

Banco de la República | Colombia. (n.d.). *Sectores económicos*. Retrieved from Banrepcultural: https://enciclopedia.banrepcultural.org/index.php?title=Sectores_econ%C3%B3micos#:~:text=Sector%20de%20comercio%3A%20Hace%20parte,productos%20a%20nivel%20nacional%20o

Banco Mundial. (2019). *Informe anual 2019. Poner fin a la pobreza, invertir para generar oportunidades.* Grupo Banco Mundial.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 99-120.

Barros, C. P., & Sellers-Rubio, R. (2008). Analysing cost efficiency in spanish retailers with a random frontier model. *International Journal of Retail and Distribution Management*, 883-900.

Britchenko, I., Monte, A. P., Kryvovyazyuk, I., & Kryvoviaziuk, L. (2018). The comparison of efficiency and performance of portuguese and ukrainian enterprises. *Ikonomicheski Izsledvania*, 87-108.

Busu, M., Vargas, M. V., & Gherasim, I. A. (2020). An analysis of the economic performances of the retail companies in romania. *Management and Marketing*, 125-133.

Butigan, N., & Benić, Đ. (2017). The Impact of Membership in Strategic Alliances on the Profitability of Firms in the Retail Sector. *Croatian Economic Survey*, 47-82.

Cruz, B. P. (2017). *Uso de Big Data para la toma de decisiones acordes a la estrategia empresarial en el sector retail.*

Cruz, B., Machuca, F., & Figueroa, D. (2017). Principales decisiones relativas a los costos y gastos, para enfrentar escenarios de desaceleración económica en una empresa del retail. *Revista De Investigación Aplicada En Ciencias Empresariales*, 79-102.

DANE, D. A. (2020). *Clasificación Industrial Internacional Uniforme de todas las actividades económicas.* Retrieved from DANE: https://www.dane.gov.co/files/sen/nomenclatura/ciiu/CIIU_Rev_4_AC2020.pdf

De la Garza García, J., Morales Serrano, B. N., & González Cavazos, B. A. (2013). *Análisis Estadístico Multivariante. Un enfoque teórico y práctico.* México: Mc.Graw Hill.

Departamento Administrativo Nacional de Estadística (DANE). (2020). *Clasificación Industrial Internacional Uniforme de todas las actividades económicas, CIIU, Rev.4 adaptada para Colombia.* Retrieved from DANE, Departamento Administrativo Nacional de Estadística: https://www.dane.gov.co/files/sen/nomenclatura/ciiu/CIIU_Rev_4_AC2020.pdf

Departamento Administrativo Nacional de Estadística (DANE). (2020). *Clasificación Industrial Internacional Uniforme de todas las actividades económicas: Revisión 4 Adaptada para Colombia CIIU Rev. 4 A.C. (2020).*

Departamento Administrativo Nacional de Estadística (DANE). (2020). *Producto Interno Bruto (PIB).* Bogotá D.C.

Duran Vasco, M., & Zambrano Loor, J. (2016). Current Considerations on Business Management. Manta, Ecuador. Retrieved from https://dominiodelasciencias.com/ojs/index.php/es/article/view/276/328

Flores, M., Gómez, D., Briones, J. B., & Cervantes, G. P. (2013). Rentabilidad y competitividad en la PYME. *Ciencias Administrativas*, 80-86.

Garcia, J. C. (2003). La gestión moderna del comercio minorista: El enfoque práctico de las tiendas de éxito. ESIC.

Gaur, V., & Saravanan, K. (2015). The Effects of Firm Size and Sales Growth Rate on Inventory Turnover Performance in the U.S. Retail Sector. *Retail Supply Chain Management*, 25-52.

Goddard, J., Tavakoli, M., & Wilson, J. (2013). Determinants of profitability in European manufacturing and services: evidence from a dynamic. *Applied Financial Economics*, 1269-1282.

Gómez, M. C., & Rangel Suárez, A. M. (2014). Diagnóstico de la estructura financiera de las PyMES del sector de confecciones del area metropolitana de Bucaramanga. Bucaramanga, Colombia.

Guerra, C., & Fernandez, L. (2003). Criterios para la selección de modelos estadísticos en la investigación científica. *Revista Cubana de Ciencia Agrícola*.

Icontec, I. (2015). *Norma NTC 9000:2000*. Retrieved from Instituto Colombiano de Normas Técnicas y Certificación: https://www.ramajudicial.gov.co/documents/5454330/14491339/d2.+NTC+ISO+9000-2015.pdf/ccb4b35c-ee63-44b5-ba1e-7459f8714031

Janda, K., & Rausser, G. (2013). Determinants of Profitability of Polish Rural Micro-Enterprises at the Time of EU Accession. *Eastern European Countryside*, 177-217.

Jones Lang LaSalle. (2020). *Reporte Retail Colombia 2019.* Retrieved from JLL: https://grupomapa.co/wp-content/uploads/2020/02/jll-retail-report-colombia-2020.pdf

Jong, P., & Heller, G. Z. (2008). Generalized Linear Models for Insurance Data. New York: Cambridge University.

Korner-Nievergelt, F., Roth, T., von Felten, S., Guélat, J., Almasi, B., & Korner-Nievergelt, P. (2015). *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and Stan.* London, U.K.: Academic Press.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2012). *Estadística aplicada a los negocios y la economía.* México: McGraw-Hill.

Maydeu Olivares, A., & García Forero, C. (2010). Goodness-of-Fit Testing. *International Encyclopedia of Education*, 190-196.

Mayoral, A. M., & Javier Morales Socuéllamos. (2001). *Modelos lineales generalizados.* España.

Medina, M. E., & Elvis Vásquez Coloma. (2011). Análisis de los gastos operativos y su incidencia en la rentabilidad del supermercado SUPERSKANDINAVO Cía. Ltda., para el segundo semestre del año 2010. Ambato, Ecuador. Retrieved from http://repositorio.uta.edu.ec/bitstream/123456789/1800/1/TA0110.pdf

Meeusen, W., & van Den Broeck, J. (1977). Efficiency estimation from cobb-douglas production functions with composed error. *International Economic Review*, 435-444.

Mena, R. G. (2012). La gestión empresarial y el desarrollo económico nacional. (U. P. Dauphine, Ed.) Montréal. Retrieved from https://www.erudit.org/fr/revues/mi/2012-v16-n4-mi0366/1013157ar.pdf

Montgomery, D. C. (2004). *Diseño y análisis de experimentos.* México: LIMUSA.

Moreno, M. S., & Torres García, E. J. (2015). Gestión de costos. Retrieved from https://www.academia.edu/19384170/GESTION_DE_COSTOS

Moreno, M., López, E., & González, N. (2012). La importancia de la contabilidad de costos. Sonora, México.

Peña, D. (2002). *Análisis de Datos Multivariantes.* Madrid: Mc.Graw Hill.

Pinelo, A. M. (2020). Análisis de ROA, ROE y ROI. *Contadores y Empresas*.

Rodriguez, O. H. (1998). *Temas de análisis estadístico multivariado.* Editorial de la Universidad de Costa Rica. Retrieved from https://books.google.com.co/books?id=g-IT184TSS4C&pg=PA8&dq=analisis+estadistico&hl=es&sa=X&ved=2ahUKEwiaovmNvN3qAhXIY98KHUrXAU8Q6AEwBXoECAMQAg#v=onepage&q&f=true

Salazar, C., & Castillo, S. D. (2017). *Fundamentos Básicos de Estadística.* Quito: Sin editorial.

Sánchez, D. J. (2016). Diseño e implementación de una estructura de costos para la empresa "Colaciones El Manjar". Bucaramanga, Colombia.

Schiller, R. (1988). Retail decentralization. A property view. *The Geographical Journal*, 17-19.

Sellers-Rubio, R., & Mas-Ruiz, F. (2006). Economic efficiency in supermarkets: Evidences in spain. *International Journal of Retail and Distribution Management*, 155-171.

Vargas, R. D., Rojas, Y. O., & Fino, M. M. (2019). *Dinámica financiera de las empresas del sector retail y su relación con la macroeconomía colombiana en los últimos años.* Retrieved from Finanzas Y Comercio Internacional: https://ciencia.lasalle.edu.co/finanzas_comercio/542

Walpole, R. E., Myers, R. H., & Myers, S. L. (2012). Probabilidad y estadística para ingeniería y ciencias. México: Pearson.