ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

DETERMINACIÓN DE PRECIOS DE CALZADO DE UNA EMPRESA DE CONSUMO MASIVO A SUS CANALES DE DISTRIBUCIÓN A TRAVÉS DE REGRESIÓN LINEAL MÚLTIPLE Y LA DETERMINACIÓN DE LA DEMANDA A TRAVÉS DE MODELO AUTOARIMA

Profesor

Mario Salvador González Rodríguez

Autor

Jheremy Josué Ron Guevara

2024

RESUMEN

El presente documento tiene la finalidad de brindar estrategias en cuanto a precios para una empresa comercializadora de varias categorías ubicada en Cuenca, que tiene sus operaciones a nivel nacional, la empresa se ha visto en la necesidad de recurrir a este análisis debido a la variabilidad de los precios en sus canales de distribución y por tipo de producto.

En cuanto a la variabilidad de precios que existe en los canales de distribución, únicamente se acordó la revisión de skus dirigidos a calzado de playa, donde se llevó a cabo un análisis exploratorio para determinar las variables que se encuentran relacionadas con el precio, una depuración de datos y finalmente un análisis a través de regresión lineal múltiple en Python, utilizando OLS (Mínimos Cuadrados Ordinarios) y GLSAR (Mínimos cuadrados generalizados con errores autorregresivos) para determinar los precios con los que se va a distribuir .

Se ha observado que la empresa no tiene mayor disparidad en sus precios y las proyecciones presentan limitaciones, debido a que sus precios tienden a mantenerse dentro de un rango de acuerdo a sus descuentos y por lo tanto sus precios son controlados en sus canales de distribución.

Se evaluó que el modelo es funcional, debido a la metodología seguida paso a paso y a pesar de no generar mayores resultados, se podría aplicar en otros ámbitos de la distribuidora y posiblemente generar mayor valor agregado.

ABSTRACT

This Capstone aims to implement exploratory analysis techniques, database purification, linear regression through OLS (ordinary least squares) and GLSAR (generalized least squares with autoregressive errors), in order to determine the prices at which They are going to be distributed.

The applied methodology involves a detailed analysis of the variables that affect pricing determination, such as the type of distributor, the units purchased, shoe size, color and type of product. Using multiple linear regression, the aim is to discover the interactions between these variables and their influence on the pricing strategy, with the aim that the company can optimize its decisions to pressure distribution prices and increase its competitiveness in the market.

The use of comprehensive and advanced techniques such as OLS and GLSAR, supported by the use of Python, is crucial to handle the particularities of the data, including possible autocorrelations in the errors, residuals and the statistical relevance of the data used. These tools allow the development of a more precise and reliable model for price prediction, which contributes to more efficient decision-making based on the price structure applied to the different distribution channels.

ÍNDICE DEL CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Una comercializadora ubicada en Cuenca, mantiene sus operaciones de distribución en todo el país donde cuenta con varias categorías de productos (electrodomésticos, ropa, calzado, alimentos, bebidas no alcohólicas, entre otros.) donde para propósitos de este documento se contará con las bases de datos, información y análisis dirigido al calzado de playa.

Para poder compararse, la empresa analiza sus categorías frente a las diferentes industrias en el Ecuador, donde en el caso del calzado, la comercialización no ha tenido un gran impacto en los últimos 10 años, puesto que ha crecido en un 7% (SRI, 2024), y se ha mantenido entre los 130 millones de dólares hasta los 160 millones. Incluso analizando el periodo pre – pandemia, esta industria decreció en – 1% y su mayor crecimiento se ha dado en los últimos años después de la pandemia.

La comercializadora busca trabajar con nuevas metodologías dentro de sus categorías con el fin de conocer, como se distribuye su producto, el comportamiento del público al que está dirigido y los factores que intervienen para la determinación de precios, más allá de sus costos de producción y el margen deseado.

Para la introducción de nuevas metodologías se optará por emplear regresión lineal múltiple a través de Python con modelos OLS y GLSAR para la determinación de precios de calzado de playa, para de esta manera, evaluar el comportamiento de precios, dependencia de factores y posteriormente ampliar el uso de esta metodología a sus otras categorías. Al final se empleará un modelo de Autoarima, simple con el fin de conocer la proyección del precio de

una manera resumida, sin embargo ya conociendo los factores de los que depende el producto.

REVISIÓN DE LITERATURA

La determinación de precios de productos, es algo que las empresas han venido empleando desde hace varios años aplicando diferentes metodologías, como es caso clásico financiero de determinación de precios a partir de los costos fijos y variables de la producción o servicio, donde determinan el margen de ganancia que desean ganar; seguido de la siguiente metodología que es la determinación de precios a través de los precios fijados por la competencia, donde buscan diferenciarse o bien buscan competir por precios más bajos (Kotler, 2024); un modelo reciente es el modelo de fijación de precios según la percepción del consumidor o también llamado Van Westendorp (Guerrero, 2021), donde el precio se fija de acuerdo a lo que el consumidor percibe como caro o barato o si le brinda valor agregado o no; Y finalmente el enfoque se basa en el modelo de determinación de precios a través de regresión lineal, donde se empleará la información histórica de la empresa y se podrá evaluar los factores que inciden con la decisión de fijación de precios dinámica debido a los descuentos que la empresa pueda brindar y que se encuentran en manos de sus trabajadores.

Para evaluar la efectividad del uso de regresión lineal múltiple, se evidencia un caso similar que se empleó en la revista geográfica VALPSO en el 2015 con la determinación de precios de suelo (Ortiz, Arias, Da Silva, & Cardozo, 2015), este estudio tiene como fuente Papers de asignación de precios de suelo determinados por los factores de las localidades residenciales. Dentro de este estudio logra separar la información la información que aporta de manera relevante al objeto de estudio y logra determinar la proporcionalidad en la que los precios pueden llegar a crecer si uno de los factores se ve afectado, que .

Otro caso similar es la determinación de precios de electrodomésticos mediante regresión lineal para Marcimex en el 2012 (Reino Vivanco Andrés Agustin, 2012), donde a través de los precios históricos tanto de la empresa como de la competencia, logra determinar la tendencia de los precios de Marcimex y analiza detenidamente la relevancia estadística de los factores que afectan al precio.

Finalmente para el modelo de valor agregado basado en Autoarima; existen algunos estudios con base en ARIMA como es el caso de pronóstico de exportación de productos no petroleros (Mario, 2024), donde evalúa como varía las exportaciones a través de la series de tiempo, utilizando datos pasados para hacer predicciones futuras (Autorregresión) (AR), ajusta errores anteriores que diferencia a la serie para hacerla más estacionaria (I) y elimina las tendencias para hacer lo más estable posible con su media móvil (MA). (IBM, 2021)

A continuación se mostrará la matriz de referencias donde se ha utilizado modelo de regresión lineal múltiple y modelo ARIMA para la predicción del precio y la demanda:

Tabla 1 Matriz de referencias

N.	Referencia	Tipo de Datos Utilizados	Metodología	Resultados	Implicaciones gerenciales
1	Li Ye, C. B. F. (2023, 4 septiembre). Aplicación de modelo ARIMA para el pronóstico de exportación de flores del Ecuador. http://repositorio.ucsg.edu.ec/handle/3317/21884 (Li Ye, 2023)	Datos del banco central del ecuador Datos Estructurados	ARIMA	Precisión en determinación de unidades exportadas FOB	Mejora en la capacidad de toma de decisiones sobre las exportaciones de flores
2	Herrera Granda, D. E. (2019). Predicción de demanda eléctrica mediante la aplicación de modelos ARIMA y SARIMA en lenguaje de programación R – caso de estudio en la Empresa Eléctrica Quito. 122 hojas. Quito : EPN. (Herrera, 2019)	Datos de empresa eléctrica quito Datos estructurados	Arima y Sarima	Precisión en la energía, potencia activa y reactiva	Mejora en la capacidad de toma de decisiones de distribución eléctrica y proyectando la distribución promedio activa y reactiva.

3	Hernández Navarro, O, Velásquez Henao, J y Dyner Rezonzew, I. (2005). *Modelos arima y estructural de la serie de precios promedio de los contratos en el mercado mayorista de energía eléctrica en colombia.* Universidad Nacional de Colombia Sede Manizales. (Hernández Navarro, 2005)	Datos de empresa eléctrica de Colombia Datos estructurados	Arima	Precisión de determinación de precios por contratos en Colombia	Mejor determinación de precios en contratos por parque eléctrico a mayoristas
4	Del Pilar, S. C. G. (2023, 3 enero). *Gasto público social y pobreza en Ecuador, periodo 2007 - 2020*. https://dspace.unl.edu.ec/jspui/handle/123456789/25993 (Sarmiento Castillo, 2023)	Datos de Banco central del Ecuador Datos estructurados	Regresión Lineal Múltiple	Precisión para determinar los factores que afectan al crecimiento del PIB	Concluye con políticas monetarias y productivas para cada variable de la ecuación del método del gasto.
5	Manosalvas Pazmiño, D. T. (2018). Análisis del efecto del gasto público en la calidad de la educación para los países miembros y asociados de la OCDE, 2015. 39 hojas, Quito : EPN. (Manosalvas Pazmiño, 2018)	Datos del Banco Central del Ecuador Datos Estructurados	Regresión Lineal Múltiple	Precisión para determinar los factores que afectan directamente que inciden en el gasto público	Los resultados arrojan políticas positivas con relación positiva entre inversión en educación pública y calidad de educación pública
6	García, S. L., Arguello, A., Parra, R., & Pincay Pilay, M. (2019). Factores que influyen en el pH del agua mediante la aplicación de modelos de regresión lineal. INNOVA Research Journal, 4(2), 59-71. (García, 2019)https://doi.org/10.33890/innova.v4.n2.2019.909	Datos recopilados en campo Datos Semiestructurados	Regresión Líneal Múltiple	Precisión para determinar los factores que afectan el PH del agua en el río Chimbo	Arroja resultados de cuidado del agua en el río chimbo en la provincia de Bolivar
7	Rodolfo, A. R. L., & De -- Universidad del Bío-Bío Concepción 2020, T. E. G. (2020). *Evaluación predictiva del modelo ARIMA optimizado con fuerza bruta operacional aplicado en el precio del cobre y el índice bursátil Dow Jones 2011-2019*. http://repobib.ubiobio.cl/jspui/handle/123456789/3606 (Améstica Rivas, 2020)	Datos recopilados de bolsa de valores Datos Estructurados	Arima	Precisión para la predicción del precio en cobre e índice bursátil Dow Jones	Mejor administración de cartera de inversiones, evaluación de precios futuros

Fuente: Elaboración del autor

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

El objeto de estudio de este Captsone es el desarrollo de un modelo de regresión lineal múltiple para la determinación de precios en la línea de calzado para playa, ver su dependencia del precio hacia los distintos factores y evaluar nuevas ideas o propuestas dirigido a la empresa comercializadora de Cuenca.

Adicionalmente se empleará un modelo simple de Autoarima o contando así con un modelo de predicción de unidades adicional a la regresión lineal.

Actualmente los datos de las empresas de consumo masivo pueden ir desde 500.000 registros en adelante, lo que dificulta el manejo de la información con herramientas como Excel y la manipulación de los datos a través de Power BI; el límite en cuanto al uso de celdas en Excel se encuentra hasta 1.05 Millones de registros, y la comercializadora excede estos datos al tener cerca de 1.7 Millones hasta 2.0 Millones de registros por categoría. Lo que al momento de tratar de depurar la información con la que se va a trabajar, incurre en consumo excesivo de recursos físicos como memoria del procesador de las computadoras o también el tiempo de espera del procesamiento superior a 30 minutos por cambio de depuración de datos. Donde actualmente ya se cuenta con herramientas para estas actividades como lo es Python o R Studio.

Las metodologías de predicción de datos se encuentran arraigadas con la cultura organizacional de la empresa y la tradición de obedecer a históricos, ha limitado la explotación del potencial de la empresa en cuanto a aumentar la cuota de mercado o la atracción de nuevos clientes. Actualmente la predicción se basa en el crecimiento promedio de los últimos meses vs el crecimiento promedio de los años anteriores por categoría, sin añadir factores como lo es canal de distribución, color del SKU, talla del SKU o las unidades compradas.

PLANTEAMIENTO DEL PROBLEMA

Actualmente la empresa ha venido manejando su información en herramientas destinadas para analítica descriptiva como lo es Excel, Tableau o Power BI, estas herramientas se encuentran limitadas al momento de querer realizar predicciones o encontrar patrones y comportamientos debido a que se realiza de manera manual, donde incluso para encontrar relevancia estadística entre los datos, han tenido que ahondar en la exploración de datos a nivel de SKU, Producto y Canal; Aquí se puede visualizar una oportunidad de mejora, al poder generar analítica predictiva con herramientas más robustas y dirigidas al manejo de datos como lo es Python. Adicionalmente, la idea de este estudio es que migre la metodología a otras categorías que maneja la empresa.

Por otra parte, la determinación de precios no ha tenido mayor variación en los últimos años (en cuanto a la categoría de distribución de calzado) donde claramente la industria no ha tenido un crecimiento significativo; sin embargo, la empresa para poder ser competitiva en el mercado ha decidido asignar descuentos a nivel de distribuidor, donde han perdido el seguimiento del precio actual de sus SKU’s; a través del uso de regresión lineal múltiple y autoarima, se puede evaluar la situación actual del precio, brindar guía en cuanto a los factores que determinan el precio y la predicción del precio en el futuro con base a históricos.

OBJETIVO GENERAL

Implementar analítica predictiva a la categoría de calzado de una empresa comercializadora, optimizando la precisión en las proyecciones de precio a través de regresión lineal múltiple y la demanda a través de modelo ARIMA.

OBJETIVOS ESPECÍFICOS

1. Implementar un modelo de determinación de precios de regresión lineal múltiple que identifique las variables más influyentes que inciden en él y se pueda realizar un forecast de precios.

2. Implementar un modelo de determinación de demanda ARIMA que evalúe el comportamiento y la estacionalidad de los productos, que adicionalmente proyecte hacia los siguientes 4 meses.

3. Evaluar la factibilidad de ampliar estos modelos de predicción a las diferentes categorías que la empresa maneja con el fin de mejorar su precisión a la hora de proyectar precios y demanda.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

Se determinó el uso de regresión lineal múltiple debido a que se encuentra atada al conocimiento de factores que afectan directamente al precio y la determinación del precio en un futuro; se optará por este modelo debido a la simplicidad de la metodología y poder llegar de manera objetiva y con metodologías ya conocidas por los directores, lo que permitirá abordar otras metodologías para otras líneas.

Adicionalmente se desarrollará un modelo simple de autoarima con el fin de evaluar la estacionalidad de los precios de los SKUs y brindar un panorama amplio en cuanto a la dependencia del precio a sus factores y de las temporadas del año.

Normalmente estos modelos de determinación de precios suelen utilizar modelos de regresión lineal simple, múltiple o también modelos ARIMA, por cuanto se detallarán los beneficios de los usos de estos modelos a continuación:

1.- Regresión Lineal Múltiple por OLS (Mínimos Cuadrados Ordinarios):

Se utilizará este modelo debido a su simplicidad, ya que es fácil de comprender y es un modelo interpretativo claro de la relación entre las variables independientes y la variable dependiente y como interactúan estas entre sí. Es un modelo óptimo que tiene años de trayectoria y en muchos casos obtiene la aproximación para entender el modelo de los datos.

Sin embargo tiene inconvenientes con valores atípicos, que en este caso fueron depurados desde un inicio y requiere que estrictamente se cumpla otros supuestos como homocedasticidad y control en residuos, ya que al no contar con estas estimaciones, estas pueden ser inexactas.

Si existe autocorrelación durante un periodo de tiempo debido a la estacionalidad de los datos, puede llegar a generar errores, a esto se lo conoce como multicolinealidad.

2.- Regresión Lineal Múltiple por GLSAR (Mínimos cuadrados generalizados con errores autorregresivos):

Se utilizará este modelo como comprobación debido a que al ser más robusto, limita los errores o residuos que el modelo arroja, lo que normalmente suele ocurrir en series temporales como se explicó arriba; en este caso el modelo evita la autocorrelación entre los datos estacionarios.

Al ser un modelo robusto de control de autocorrelación, modela las relaciones temporales, que en este caso puede llegar a estar atado con los precios y las ventas.

El proyecto se apoyará en el modelo GLSAR una vez realizado el modelo OLS, debido a que existe dificultad de interpretación de sus datos y podríamos llegar a sesgar el modelo al elegir variables incorrectas o al discriminar las variables

3.- Regresión Múltiple por WLS (Mínimos Cuadrados Ponderados):

En este caso utilizaremos el modelo WLS, como modelo de apoyo para evaluar los modelos GLSAR y OLS ya que este modelo permite corregir la heterocedasticidad.

Este modelo se ajusta al asignar pesos a las observaciones lo que ajusta al resto de modelos haciendo que las estimaciones sean más precisas.

De igual manera se utilizará este modelo como apoyo una vez realizado el modelo clásico OLS y el modelo GLSAR .

4.- Modelo AUTOARIMA:

Primero se dará a conocer que es el modelo ARIMA, el modelo ARIMA se diferencia en 3 metodologías aplicadas las cuales son:

AR – Sub modelo de autorregresión: dependencia de “p” valores, donde hace referencia al número de valores pasados en los que el modelo usa para hacer predicciones.

I – Integridad: dependencia de “d” valores, donde hace referencia a la remoción de componentes no estacionarios.

MA – Movilidad promedio o base móvil: Dependencia de “q” valores, donde hace referencia al número de errores pasados que el modelo ARIMA puedo haber tenido al momento de realizar predicciones.

¿Porqué se llama AUTOARIMA el modelo a utilizar? ¿En qué se diferencia con ARIMA?

Para poder realizar el modelo ARIMA, es imprescindible ir acompañando al modelo mientras se va generando y se debe ir interpretando los datos con el fin de encontrar los parámetros óptimos que definan el modelo, lo que probablemente al existir una mala interpretación de los resultados dirigidos a “p”, “d” y “q”, pueda llegar a afectar de manera significativa al desempeño del modelo ARIMA.

Lo que realiza el modelo **AUTOARIMA** es que automáticamente se definan los parámetros de “p”, “d” y “q” de manera óptima y que el modelo arroje predicciones o forecast de precios acertados.

Prácticamente AUTOARIMA resume el paso a paso de ARIMA sin descuidar la relevancia estadística y dando resultados acertados.

En cuanto al paso a paso a seguir del modelo, primero se iniciará con regresión lineal múltiple.

1.- Carga de bibliotecas a utilizar, archivo e identificación de encabezados:

.

Ilustración 1 Bibliotecas implementadas

```
In [197]: import numpy as np
          import pandas as pd
          import matplotlib.pyplot as plt
          import seaborn as sns

          import warnings
          warnings.filterwarnings("ignore")

In [198]: import statsmodels.stats.api as sms
          import statsmodels.api as sm
          from statsmodels.formula.api import ols
          from statsmodels.compat import lzip
          import statsmodels.api as sm
          from statsmodels.regression.linear_model import GLSAR
```

Fuente: Elaboración del autor.

- Utilizaremos las librerías de **Pandas y Numpy** para la manipulación de los datos.

- **Matplotlib.pyplot** utilizaremos para generar gráficos y poder interpretar.

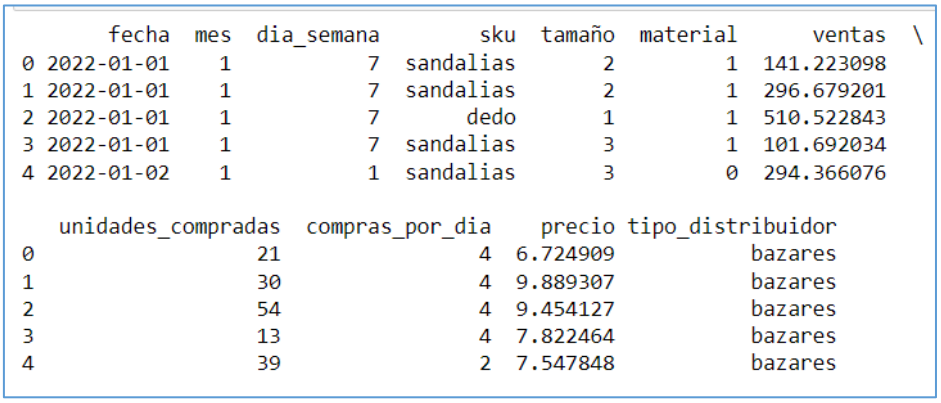
- La librería de **statsmodel** sirve principalmente para la generación del modelo como tal de regresión lineal, aquí se alojan tanto los modelos de **OLS(mínimos cuadrados ordinarios)** como **GLSAR (Mínimos cuadrados generalizados con errores autorregresivos)**.

- Estas son las bibliotecas a utilizar en este modelo y el modelo AUTOARIMA se explicará más adelante.

- Es importante mencionar que la base de datos tuvo un proceso de enmascaramiento de la información y la alteración de las cantidades de precio y cantidades fueron multiplicados por factores, por políticas de la empresa.

2.- Análisis exploratorio de datos:

Ilustración 2 Detalle Data Frame



	fecha	mes	dia_semana	sku	tamaño	material	ventas	\
0	2022-01-01	1	7	sandalias	2	1	141.223098	
1	2022-01-01	1	7	sandalias	2	1	296.679201	
2	2022-01-01	1	7	dedo	1	1	510.522843	
3	2022-01-01	1	7	sandalias	3	1	101.692034	
4	2022-01-02	1	1	sandalias	3	0	294.366076	

	unidades_compradas	compras_por_dia	precio	tipo_distribuidor
0	21	4	6.724909	bazares
1	30	4	9.889307	bazares
2	54	4	9.454127	bazares
3	13	4	7.822464	bazares
4	39	2	7.547848	bazares

Fuente: Elaboración del autor

- En primer lugar, definimos de manera objetiva los encabezados con los que contamos.

Tabla 2 Identificación de columnas

Encabezado	**Formato**	**Descripción**
Fecha	Fecha	Muestra cuando se realizó el registro de la información
Mes	Número	Mes de la fecha del registro
Dia_semana	Número	Muestra el día de la semana , correspondiente a la fecha (del 1 al 7 siendo lunes=”1” y domingo= “7”).
Sku	Texto	Corresponde al ítem en el registro, contamos con 3 skus, sandalias, crocs y “de dedo”, que corresponde al calzado de playa.
Tamaño	Número	corresponde a la talla del sku donde 1=”small”, 2=”medium” y 3= “large”.
Material	Número	Corresponde a la procedencia del material, donde corresponde al mismo

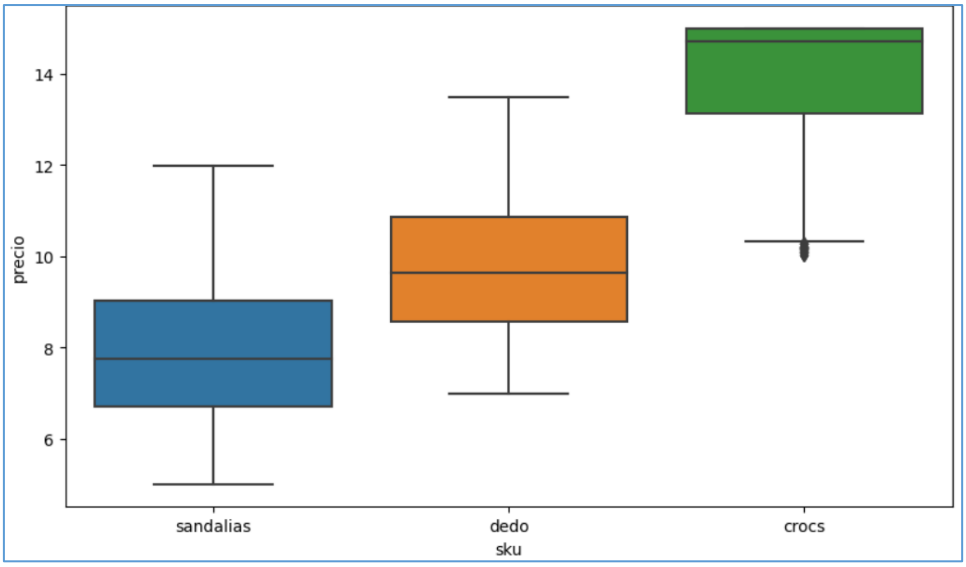
		material pero de diferentes países de proveniencia
Ventas	Número con decimales	Corresponde a la multiplicación entre la información alojada en Unidades compradas * Precio.
Unidades_compradas	Número	Corresponde a la cantidad comprada de pares de zapatos
Compras_por_dia	Número	Corresponde al número de compras realizadas por el distribuidor en el día
Precio	Número con decimales	Corresponde al precio fijado por el vendedor hacia el canal de distribución, donde únicamente se coloca el precio final, omitiendo el descuento realizado
Tipo_distribuidor	Texto	Corresponde a los tipos de distribuidores en el canal donde encontramos: “Bazares”, “Tiendas de Barrio”, “Stands de playa”, “Centros deportivos” y “Charoleros”

Fuente: Elaboración del autor

Una vez identificados los datos con los que vamos a trabajar, procedemos a realizar análisis exploratorio a través de evaluar los encabezados, los datos con los que contamos dentro de los mismo y cruces de variables para evaluar la significatividad de los datos a primera vista; esto se realiza con el fin de ir ajustando el modelo desde un inicio, también se puede aplicar un modelo de Random_Forest o árboles de decisión para evaluar los datos más significativos; sin embargo por motivos del modelo se aplicará el procedimiento normal de la metodología

- Dado que vamos a determinar que la variable dependiendo es el precio, enfocaremos nuestros análisis exploratorios hacia esa variable.

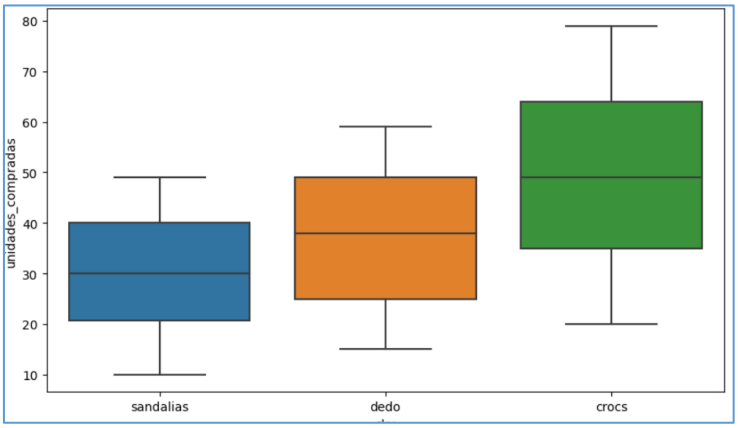
Ilustración 3 Boxplots de análisis exploratorio



Fuente: Elaboración del autor

- Se observa que existen precios definidos por SKU, sin embargo lo que determina el diagrama de cajas es que existen variabilidad en sus precios dentro de un rango determinado con una media clara, en el caso de crocs la media se encuentra colocada por encima de la caja lo que explica que el precio se acerca a 15 USD. Se tomará esta variable para participar en el modelo.

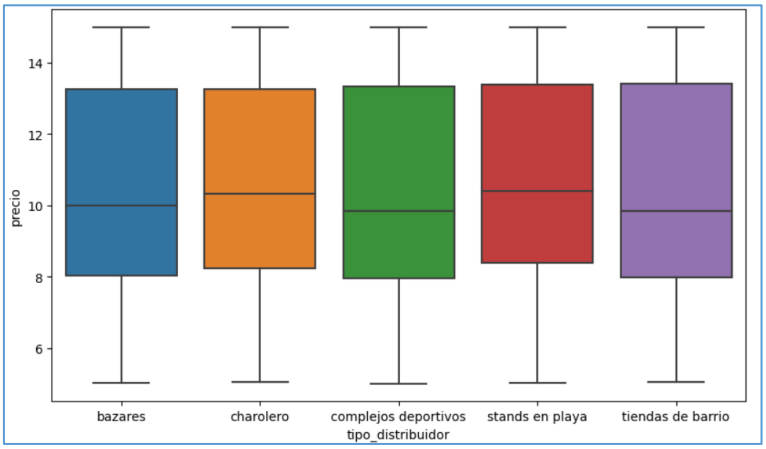
Ilustración 4 Boxplot de unidades compradas



Fuente: Elaboración del autor

- Las sandalias se encuentran en el rango de compra entre 20 y 40 unidades, las sandalias “dedo” van desde las 25 unidades hasta las 50 unidades y las crocs están desde las 30 unidades hasta las 70 unidades aproximadamente; esto explica el comportamiento de compra de combos y pacas, ya que normalmente se usa un mix de productos para brindar variedad a los canales de distribución. Utilizaremos estas variables.

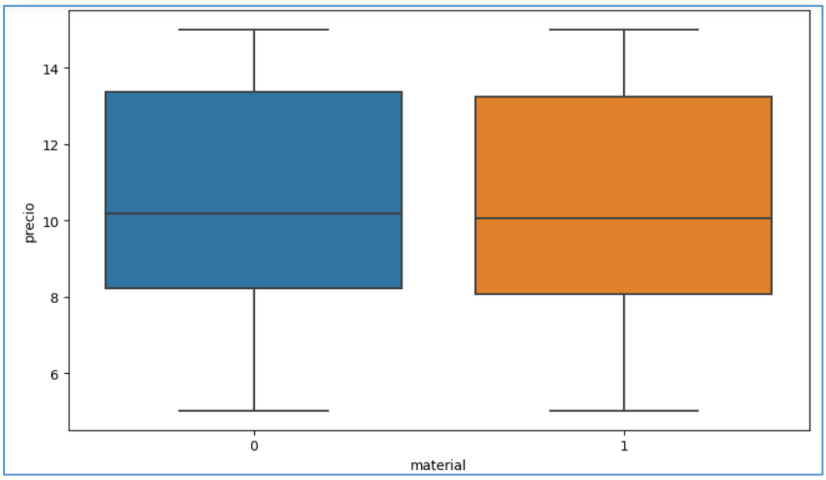
Ilustración 5 Boxplor de tipo de distribuidor con relación al precio



Fuente: Elaboración del autor

- Los precios de distribuidor no varían entre ellos debido a que la comercializadora maneja precios estandarizados hace varios años atrás, sin embargo los descuentos si varían por canal de distribución, en este caso lo que son stands de playa y charoleros tienen precios más elevados y su media también se encuentra elevada con respecto a los otros tipos de distribuidores. Tomaremos estas variables para el análisis, sin embargo las dummificaremos con el fin de tener más variables para definir la dependencia del modelo.

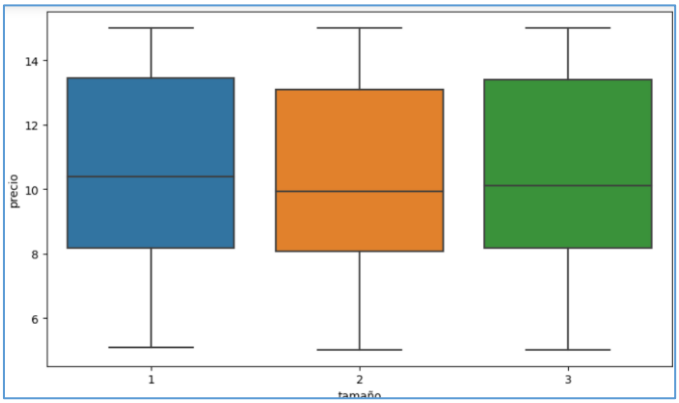
Ilustración 6 Boxplot de precio vs. material



Fuente: Elaboración del autor

- Como se comentaba anteriormente, el material es el mismo sin embargo lo único que cambia es el país de procedencia, por lo cuanto lo retiraremos del modelo para evitar errores.

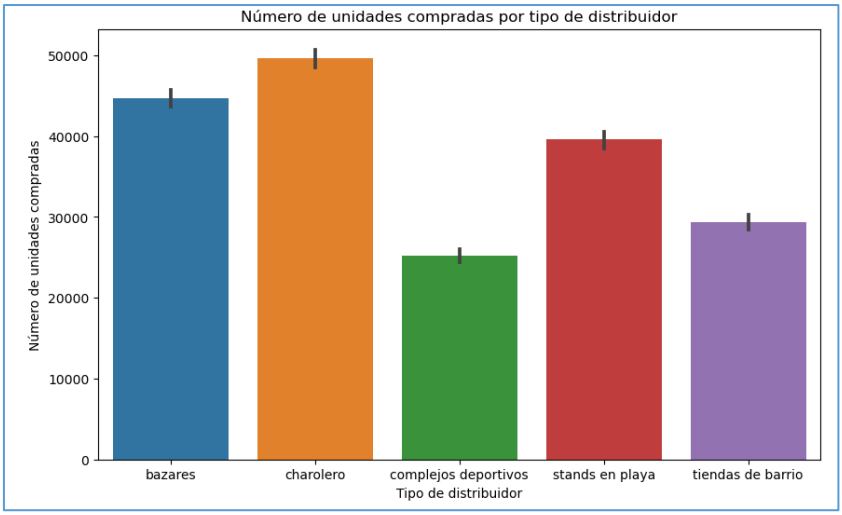
Ilustración 7 Boxplot de precio vs tamaño



Fuente: Elaboración del autor.

- Los precios no varían de acuerdo al tamaño debido a que los moldes y los materiales que se utilizan son los mismos para los 3 modelos y los precios se encuentran estandarizados. Adicionalmente, la fábrica reutiliza los desperdicios, manteniendo de esta manera un precio estable.
- Se puede observar que en todas las gráficas existen rangos de precios limitados, esto se debe a que existe un precio tope en cuanto a los productos y sobre ese se trabaja en un descuento.

Ilustración 8 Barras de unidades por tipo de distribuidor



Fuente: Elaboración del autor

- Se puede apreciar que el tipo de distribuidor que más compra son charoleros, seguidos de bazares, stands en playa, tiendas de barrio y complejos deportivos; esto hace sentido a la cantidad vendida entre los distribuidores y su giro del negocio.

3.- Depuración de bases de datos

- A continuación procedemos a buscar vacíos, información errónea, máximos y mínimos.

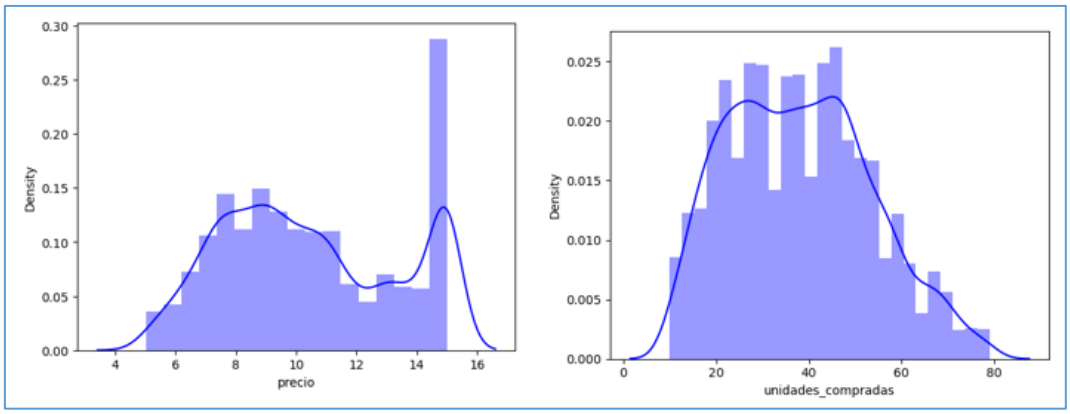
Ilustración 9 Dataframe previo a depuración

```
Out[15]: fecha                datetime64[ns]
         mes                          int64
         dia_semana                   int64
         sku                         object
         tamaño                       int64
         material                     int64
         ventas                     float64
         unidades_compradas           int64
         compras_por_dia              int64
         precio                     float64
         tipo_distribuidor           object
         dtype: object
```

Fuente: Elaboración del autor

Lo que se observa en el dataframe es principalmente que no existe mayor dispersión en los datos y los rangos intercuartiles guardan datos normales.
La base de datos en cuanto al tipo de data que se va a analizar está estandarizada por cuanto debemos transformar todo a números más que intervenir con transformaciones de texto.

Ilustración 10 Distribución de las observaciones

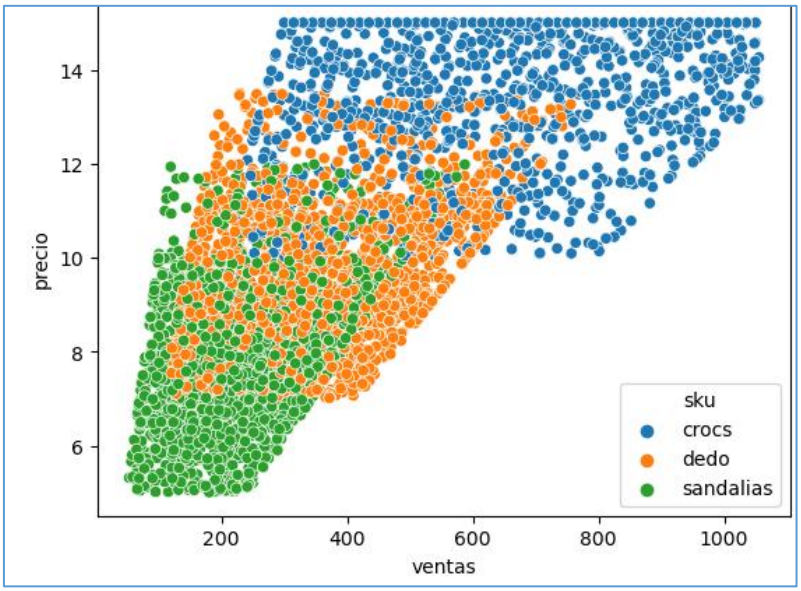


Fuente: Elaboración del autor

- Se procede a depurar máximos y mínimos de las variables numéricas y se procede a colocarlos en sus rangos intercuartiles.

Una vez arreglada la información volvemos a analizar.

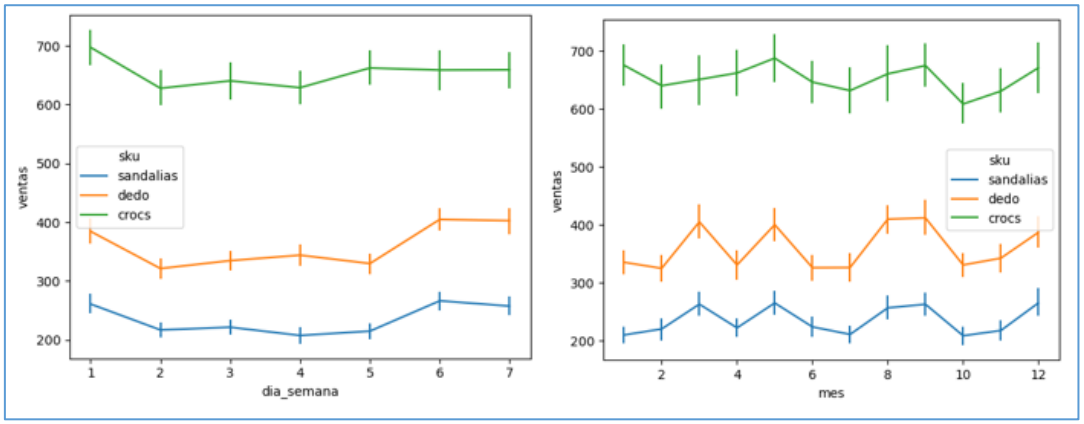
Ilustración 11 Gráfico de dispersión de precio vs. ventas



Fuente: Elaboración del autor

Como se mencionaba anteriormente, los skus tienen un precio límite y a partir de estos se colocan descuentos, donde claramente podemos observar los precios por SKU'S en forma de cono.

Ilustración 12 Ventas vs. mes o día de la semana



Fuente: Elaboración del autor

Se observa estacionalidad en las compras en cuanto a los días de la semana siendo el día lunes, sábado y domingo los días con mayor compra. Esto hace

sentido con el giro del negocio y la distribución que se brinda a los clientes. De igual manera con los meses, podemos observar la estacionalidad donde existe mayor compra de estos productos, donde corresponde al mes de marzo, mayo, agosto y septiembre y finalmente diciembre; que son los meses por lo general donde se encuentra la mayor cantidad de vacaciones registradas históricamente.

Ilustración 13 Matriz de correlación

	mes	dia_semana	tamaño	material	ventas	unidades_compradas	compras_por_dia	precio
mes	1.000000	0.008341	-0.002684	-0.023619	0.002513	-0.026454	-0.021371	0.055382
dia_semana	0.008341	1.000000	0.033015	-0.009222	0.018342	-0.013014	-0.006888	0.069690
tamaño	-0.002684	0.033015	1.000000	0.005620	-0.009491	-0.003178	-0.016071	-0.014694
material	-0.023619	-0.009222	0.005620	1.000000	-0.009148	-0.002901	0.000935	-0.016965
ventas	0.002513	0.018342	-0.009491	-0.009148	1.000000	0.883470	0.008239	0.725010
unidades_compradas	-0.026454	-0.013014	-0.003178	-0.002901	0.883470	1.000000	-0.000330	0.362172
compras_por_dia	-0.021371	-0.006888	-0.016071	0.000935	0.008239	-0.000330	1.000000	0.023741
precio	0.055382	0.069690	-0.014694	-0.016965	0.725010	0.362172	0.023741	1.000000

Fuente: Elaboración del autor

Se puede observar que las variables con mayor correlación son, unidades compradas, ventas y precio, donde se retirarán del modelo debido al que modelo debe entrenarse sin variables directamente relacionadas, evitando la homocedasticidad; y permitiendo que el modelo se desarrolle de manera adecuada.

Como se observa, el tener pocas variables puede llegar a limitar el modelo por lo que procedemos a dummificar las variables, en este caso este término hace referencia al tratar las respuestas como variables dicotómicas, es decir en vez de SKU: Sandalias, dedo y crocs, procedería de la siguiente manera: Sandalias: 1 – sí y 0 – no; y lo mismo con las columnas que se desea dummificar, es el caso de tipo_distribuidor y SKU. En el siguiente gráfico se puede verificar el cambio.

Ilustración 14 Dataframe dummificado

ño	material	ventas	unidades_compradas	compras_por_dia	precio	log_precio	bazares	charolero	complejos deportivos	stands en playa	tiendas de barrio	crocs	dedo	sandalias
2	1	141.223098	21	4	6.724909	1.905818	1	0	0	0	0	0	0	1
2	1	296.679201	30	4	9.889307	2.291454	1	0	0	0	0	0	0	1
1	1	510.522843	54	4	9.454127	2.246451	1	0	0	0	0	0	1	0
3	1	101.692034	13	4	7.822464	2.057000	1	0	0	0	0	0	0	1
3	0	294.366076	39	2	7.547848	2.021263	1	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	0	222.357513	20	3	11.117876	2.408554	0	0	0	0	1	0	0	1
2	1	153.883455	16	4	9.617716	2.263607	0	0	0	0	1	0	0	1
3	1	191.157314	23	2	8.311188	2.117603	0	0	0	0	1	0	0	1
1	1	558.494110	43	5	12.988235	2.564044	0	0	0	0	1	0	1	0
2	1	257.041031	26	2	9.886193	2.291139	0	0	0	0	1	0	0	1

Fuente: Elaboración del autor

Una vez analizadas las variables, procedemos con la generación del modelo con base a OLS:

Ilustración 15 Regresión lineal con modelo OLS

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 precio   R-squared:                       0.741
Model:                            OLS   Adj. R-squared:                  0.741
Method:                 Least Squares   F-statistic:                     1120.
Date:                Sun, 20 Oct 2024   Prob (F-statistic):               0.00
Time:                        21:32:41   Log-Likelihood:                -8508.1
No. Observations:                4707   AIC:                         1.704e+04
Df Residuals:                    4694   BIC:                         1.713e+04
Df Model:                          12
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                  6.4382      0.083     77.304      0.000       6.275       6.602
mes                        0.0503      0.006      8.036      0.000       0.038       0.063
dia_semana                 0.1164      0.011     10.667      0.000       0.095       0.138
tamaño                    -0.0095      0.026     -0.360      0.719      -0.061       0.042
material                  -0.0997      0.043     -2.314      0.021      -0.184      -0.015
unidades_compradas        -0.0063      0.002     -3.987      0.000      -0.009      -0.003
compras_por_dia            0.0340      0.019      1.799      0.072      -0.003       0.071
charolero                  1.3204      0.042     31.143      0.000       1.237       1.403
stands_en_playa            1.3449      0.046     29.382      0.000       1.255       1.435
tiendas_de_barrio          1.2819      0.051     25.139      0.000       1.182       1.382
complejos_deportivos       1.2093      0.054     22.365      0.000       1.103       1.315
bazares                    1.2818      0.043     29.656      0.000       1.197       1.367
dedo                       1.3565      0.041     33.016      0.000       1.276       1.437
sandalias                 -0.4923      0.039    -12.713      0.000      -0.568      -0.416
crocs                      5.5740      0.048    116.376      0.000       5.480       5.668
==============================================================================
Omnibus:                       28.631   Durbin-Watson:                   1.256
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               21.102
Skew:                          -0.050   Prob(JB):                     2.62e-05
Kurtosis:                       2.688   Cond. No.                     1.60e+17
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.23e-28. This might indicate that there are
```

Fuente: Elaboración del autor

Se ejecuta el modelo de primera mano con el fin de observar la relevancia estadística de los datos y la significatividad de los datos relacionados al precio.

Al tener un R2 superior al 70%, se puede observar que existe relación entre los datos analizados, de la misma manera un F_statistic elevado significa que el modelo evaluado como un todo es significativo en su conjunto.

Lo que podemos observar es que los P valores deben estar entre 0.000 y 0.050, lo que no sucede con tamaño y compras por día, por lo que procedemos a evaluar el factor de inflación de la varianza o también conocido como análisis **VIF**:

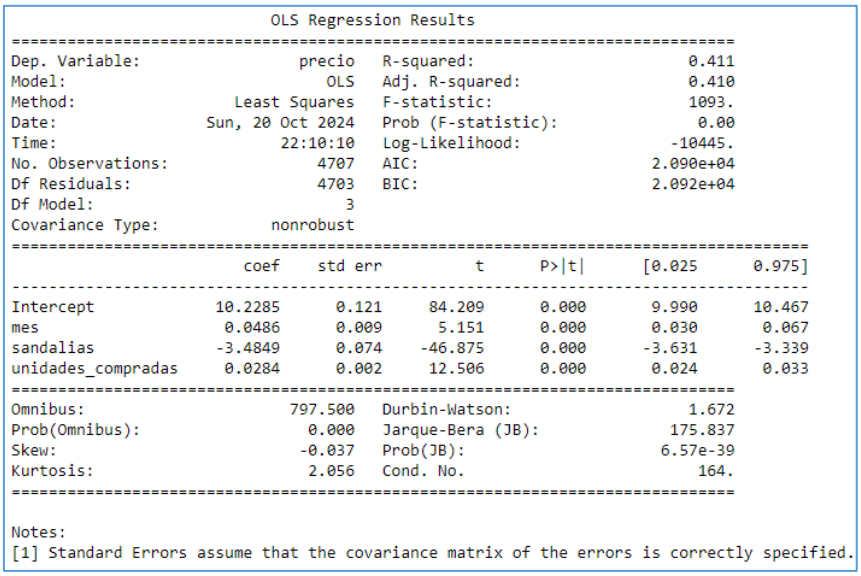
Ilustración 16 identificación de VIF

	feature	VIF
0	bazares	inf
1	charolero	inf
2	complejos_deportivos	inf
3	compras_por_dia	1.003825
4	crocs	inf
5	dedo	inf
6	dia_semana	1.027125
7	material	1.002446
8	mes	1.016177
9	precio	3.863769
10	sandalias	inf
11	stands_en_playa	inf
12	tamaño	1.001907
13	tiendas_de_barrio	inf
14	unidades_compradas	1.277033

Fuente: Elaboración del autor

Se procede a retirar las variables del modelo que tienen p valores elevados y multicolinealidad quedando así el modelo de la siguiente manera:

Ilustración 17 Resultados finales OLS



OLS Regression Results

Dep. Variable:	precio	R-squared:	0.411
Model:	OLS	Adj. R-squared:	0.410
Method:	Least Squares	F-statistic:	1093.
Date:	Sun, 20 Oct 2024	Prob (F-statistic):	0.00
Time:	22:10:10	Log-Likelihood:	-10445.
No. Observations:	4707	AIC:	2.090e+04
Df Residuals:	4703	BIC:	2.092e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.2285	0.121	84.209	0.000	9.990	10.467
mes	0.0486	0.009	5.151	0.000	0.030	0.067
sandalias	-3.4849	0.074	-46.875	0.000	-3.631	-3.339
unidades_compradas	0.0284	0.002	12.506	0.000	0.024	0.033

Omnibus:	797.500	Durbin-Watson:	1.672
Prob(Omnibus):	0.000	Jarque-Bera (JB):	175.837
Skew:	-0.037	Prob(JB):	6.57e-39
Kurtosis:	2.056	Cond. No.	164.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fuente: Elaboración del autor

Existe un problema en cuanto a generar los datos por regresión lineal múltiple ya que únicamente se quedan variables como sandalias, mes y unidades_compradas; por cuanto se debe volver a analizar todos los pasos recorridos.

En este caso al momento de evaluar se observa que existe multicolinealidad en los datos, esto se hace evidente principalmente en las columnas que se llegó a dumificar, ya que es un efecto que la dumificación desemboca en una multicolinealidad en los datos.

Se verificó también con la metodología de codificar las variables a través de “Label_encoder”, sin embargo no tuvo éxito alguno ya que lo que queremos evaluar son los precios que se pueden llegar a generar; por SKU y por tipo de distribuidor; algo que al ser tratado como un todo, el resultado es contraproducente.

En este caso, al verificar las variables determinamos lo siguiente:

- R2 en los primeros modelos es elevado lo que significa que existe relevancia en los datos

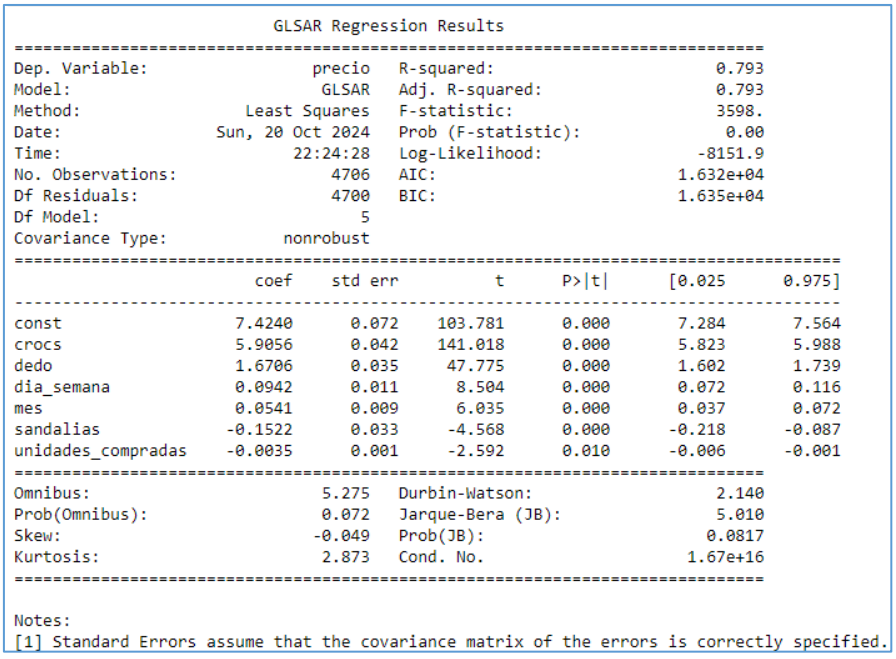
- VIF = Infinito, se evidencia que existe multicolinealidad en los datos,

- No existen variables que lleguen a trabajar para el modelo lo que radica en un problema de proyección.

Una vez analizado esto, podemos decir que se debe recorrer a otros modelos más robustos, ya que la multicolinealidad, la heterocedasticidad y la falta de variables que puedan predecir el modelo, se deben principalmente a que los datos son estacionarios y que la autocorrelación que se está generando se debe principalmente a ese motivo.

Por cuanto recorremos a los otros dos modelos **GLSAR** y **WLS** donde se obtiene lo siguiente:

Ilustración 18 Regresión con modelo GLSAR



GLSAR Regression Results

Dep. Variable:	precio	R-squared:	0.793
Model:	GLSAR	Adj. R-squared:	0.793
Method:	Least Squares	F-statistic:	3598.
Date:	Sun, 20 Oct 2024	Prob (F-statistic):	0.00
Time:	22:24:28	Log-Likelihood:	-8151.9
No. Observations:	4706	AIC:	1.632e+04
Df Residuals:	4700	BIC:	1.635e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.4240	0.072	103.781	0.000	7.284	7.564
crocs	5.9056	0.042	141.018	0.000	5.823	5.988
dedo	1.6706	0.035	47.775	0.000	1.602	1.739
dia_semana	0.0942	0.011	8.504	0.000	0.072	0.116
mes	0.0541	0.009	6.035	0.000	0.037	0.072
sandalias	-0.1522	0.033	-4.568	0.000	-0.218	-0.087
unidades_compradas	-0.0035	0.001	-2.592	0.010	-0.006	-0.001

Omnibus:	5.275	Durbin-Watson:	2.140
Prob(Omnibus):	0.072	Jarque-Bera (JB):	5.010
Skew:	-0.049	Prob(JB):	0.0817
Kurtosis:	2.873	Cond. No.	1.67e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fuente: Elaboración del autor

El modelo de GLSAR, como se comentaba al inicio de la metodología arroja un R2 elevado, siendo este 0.79 y un F-statistic superior al obtenido en OLS,

adicionalmente que se observa que existen más variables que participan en el modelo y permite la flexibilidad al momento de generarlo.

De esta manera confirmamos que existe autocorrelación entre los datos debido a que son datos estacionarios y adicionalmente que al momento de dummificar las variables se generó la multicolinealidad.

Ilustración 19 Regresión con modelo WLS

WLS Regression Results

Dep. Variable:	precio	R-squared:	0.999
Model:	WLS	Adj. R-squared:	0.999
Method:	Least Squares	F-statistic:	1.331e+06
Date:	Sun, 20 Oct 2024	Prob (F-statistic):	0.00
Time:	22:43:07	Log-Likelihood:	-9357.8
No. Observations:	4707	AIC:	1.873e+04
Df Residuals:	4701	BIC:	1.877e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.9224	0.006	2349.054	0.000	13.911	13.934
dedo	-4.2854	0.003	-1264.990	0.000	-4.292	-4.279
dia_semana	0.0060	0.001	4.837	0.000	0.004	0.008
material	0.1086	0.003	38.827	0.000	0.103	0.114
mes	-0.0107	0.001	-20.863	0.000	-0.012	-0.010
sandalias	-6.0838	0.004	-1627.318	0.000	-6.091	-6.076

Omnibus:	4112.283	Durbin-Watson:	1.822
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8466038.024
Skew:	-2.805	Prob(JB):	0.00
Kurtosis:	210.690	Cond. No.	55.9

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified

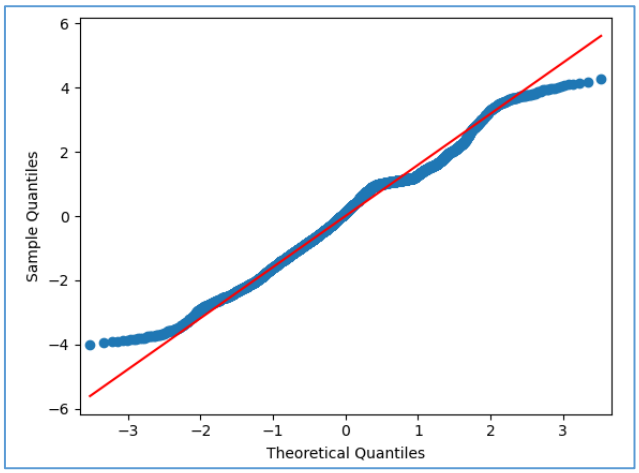
Fuente: Elaboración del autor

Al ejecutar el modelo a través de WLS, se observa que tenemos menos variables, lo que puede llegar a limitar la predicción de precios por la dependencia de sus variables. De igual manera, se observa que se confirma que existe heterocedasticidad en los datos y no permite que el modelo OLS se desempeñe con normalidad.

Una vez realizado esto se procede con las comprobaciones finales donde se comparará estos 3 modelos.

Modelo WLS

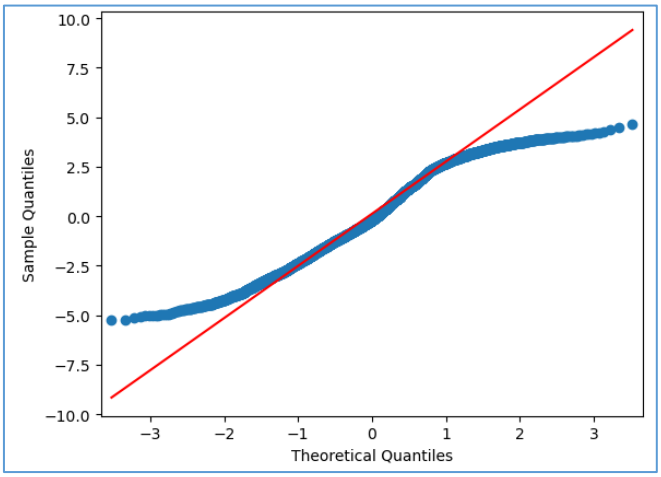
Ilustración 20 Residuos modelo WLS



Fuente: Elaboración del autor

Modelo OLS

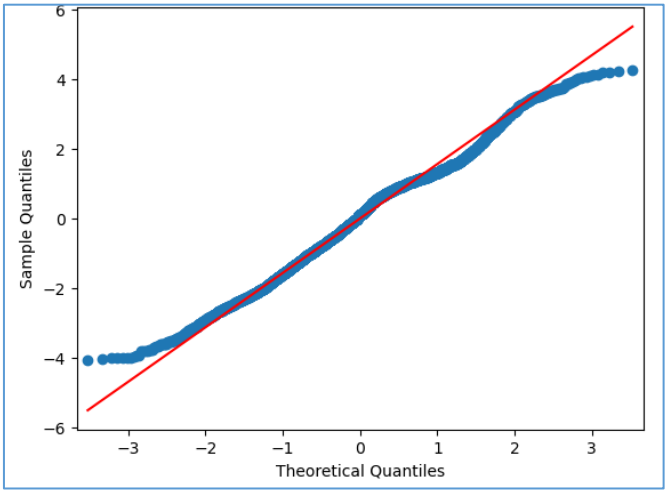
Ilustración 21 Residuos modelo OLS



Fuente: Elaboración del autor

Modelo GLSAR

Ilustración 22 Residuos modelo GLSAR



Fuente: Elaboración del autor

En cuanto a la normalidad de los residuos se puede observar un mejor desempeño de los modelos WLS y principalmente del modelo GLSAR.

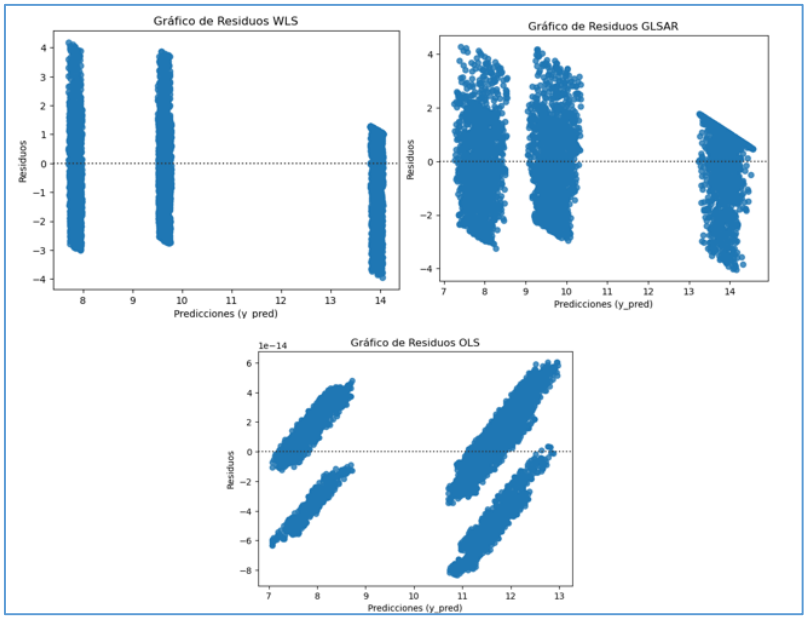
Ilustración 23 Relevancia estadística de modelos

```
In [239]: nombres = ["Jarque-Bera", "Chi^2 two-tail prob.", "Skew", "Kurtosis"]
          jarque_bera = sms.jarque_bera(results_4.resid)
          lzip(nombres, jarque_bera)

Out[239]: [('Jarque-Bera', 175.83749138397673),
           ('Chi^2 two-tail prob.', 6.567103940478575e-39),
           ('Skew', -0.0367547414316382),
           ('Kurtosis', 2.0559901327859404)]

In [238]: nombres = ["Jarque-Bera", "Chi^2 two-tail prob.", "Skew", "Kurtosis"]
          jarque_bera = sms.jarque_bera(results_wls.resid)
          lzip(nombres, jarque_bera)

Out[238]: [('Jarque-Bera', 16.527252322934395),
           ('Chi^2 two-tail prob.', 0.00025772275046092 5),
           ('Skew', -0.06039162363576395),
           ('Kurtosis', 2.7360295847174707)]

In [240]: nombres = ["Jarque-Bera", "Chi^2 two-tail prob.", "Skew", "Kurtosis"]
          jarque_bera = sms.jarque_bera(results_pw.resid)
          lzip(nombres, jarque_bera)

Out[240]: [('Jarque-Bera', 19.32129006464552),
           ('Chi^2 two-tail prob.', 6.374339211967695e-05),
           ('Skew', -0.0625487291524133),
           ('Kurtosis', 2.712135782182562)]
```

Fuente: Elaboración del autor

De igual manera los modelos de WLS y GLSAR tienen un mejor desempeño en los indicadores de relevancia estadística.

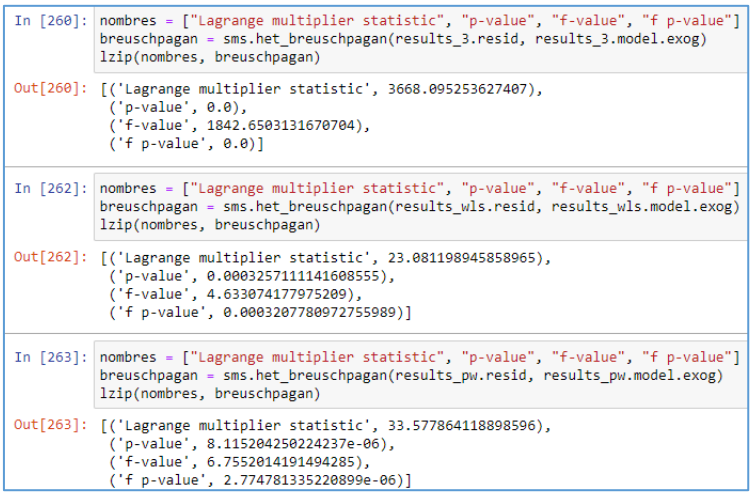
Ilustración 24 Dispersión de residuos de acuerdo a los modelos



Fuente: Elaboración del autor

Se podría considerar a partir de la gráfica anterior que el modelo OLS es más optimo, sin embargo lo que se observa es que existen 3 categorías latentes en la información que como se había estado mencionando anteriormente, y la respuesta radica a que se está tratando de predecir el precio de 3 tipos de sku diferentes; por cuanto considerando todo lo anterior, el modelo de WLS y GLSAR siguen siendo los modelos más óptimos para llevarse a cabo.

Ilustración 25 Heterocedasticidad en los modelos



```
In [260]: nombres = ["Lagrange multiplier statistic", "p-value", "f-value", "f p-value"]
          breuschpagan = sms.het_breuschpagan(results_3.resid, results_3.model.exog)
          lzip(nombres, breuschpagan)

Out[260]: [('Lagrange multiplier statistic', 3668.095253627407),
           ('p-value', 0.0),
           ('f-value', 1842.6503131670704),
           ('f p-value', 0.0)]

In [262]: nombres = ["Lagrange multiplier statistic", "p-value", "f-value", "f p-value"]
          breuschpagan = sms.het_breuschpagan(results_wls.resid, results_wls.model.exog)
          lzip(nombres, breuschpagan)

Out[262]: [('Lagrange multiplier statistic', 23.081198945858965),
           ('p-value', 0.0003257111141608555),
           ('f-value', 4.633074177975209),
           ('f p-value', 0.0003207780972755989)]

In [263]: nombres = ["Lagrange multiplier statistic", "p-value", "f-value", "f p-value"]
          breuschpagan = sms.het_breuschpagan(results_pw.resid, results_pw.model.exog)
          lzip(nombres, breuschpagan)

Out[263]: [('Lagrange multiplier statistic', 33.577864118898596),
           ('p-value', 8.115204250224237e-06),
           ('f-value', 6.7552014191494285),
           ('f p-value', 2.774781335220899e-06)]
```

Fuente: Elaboración del autor

A pesar de que existe heterocedasticidad en los modelos de WLS y GLSAR, estos siguen siendo lo más óptimos en términos de adaptabilidad ya que tienen p valores cercanos a cero, lo que demuestra que son significativos.

En cuanto a la predicción de precios, procedemos a evaluar los 3 modelos:

Predicción con OLS: 0   6.820624
dtype: float64

Esto corresponde a la determinación del precio por el modelo OLS, lo cual no es del todo acertado ya que no se acerca a los precios definidos por los SKUS.

Predicción con WLS: 0   3.687067
dtype: float64

Esta es la predicción con el modelo WLS, lo cual también no es acertado en cuanto a los rangos de precios establecidos.

Predicción con GLSAR: 0   13.227997
dtype: float64

Se utilizará el modelo de GLSAR para la determinación del precio ya que es el más acertado y ha tenido el mejor desempeño en cada una de las pruebas.

Predicción de precios con modelo GLSAR para CROCS: 13.47

Predicción de precios con modelo GLSAR para Sandalias de dedo: 9.23

Predicción de precios con modelo GLSAR para Sandalias: 7.41

Para un mejor entendimiento de como se obtuvo estos resultados procederemos con la explicación de la fórmula de manera matemática:

Ilustración 26 Modelo GLSAR para fórmula matemática

GLSAR Regression Results

Dep. Variable:	precio	R-squared:	0.793
Model:	GLSAR	Adj. R-squared:	0.793
Method:	Least Squares	F-statistic:	3598.
Date:	Sun, 27 Oct 2024	Prob (F-statistic):	0.00
Time:	22:36:58	Log-Likelihood:	-8151.9
No. Observations:	4706	AIC:	1.632e+04
Df Residuals:	4700	BIC:	1.635e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.4240	0.072	103.781	0.000	7.284	7.564
crocs	5.9056	0.042	141.018	0.000	5.823	5.988
dedo	1.6706	0.035	47.775	0.000	1.602	1.739
dia_semana	0.0942	0.011	8.504	0.000	0.072	0.116
mes	0.0541	0.009	6.035	0.000	0.037	0.072
sandalias	-0.1522	0.033	-4.568	0.000	-0.218	-0.087
unidades_compradas	-0.0035	0.001	-2.592	0.010	-0.006	-0.001

Omnibus:	5.275	Durbin-Watson:	2.140
Prob(Omnibus):	0.072	Jarque-Bera (JB):	5.010
Skew:	-0.049	Prob(JB):	0.0817
Kurtosis:	2.873	Cond. No.	1.67e+16

Fuente: Elaboración del autor

A continuación se explicará la fórmula de regresión lineal múltiple de manera manual con el fin de identificar lo más relevante para el precio.

Fórmula:

Precio

$=7.4240+5.9056\times crocs+1.6706\times dedo+0.0942\times dia\_semana+0.0541\times mes+0.5120\times sandalias-0.0035\times unidades\_compradas$

Explicación de la fórmula:

El Intercepto es (constante): 7.4240

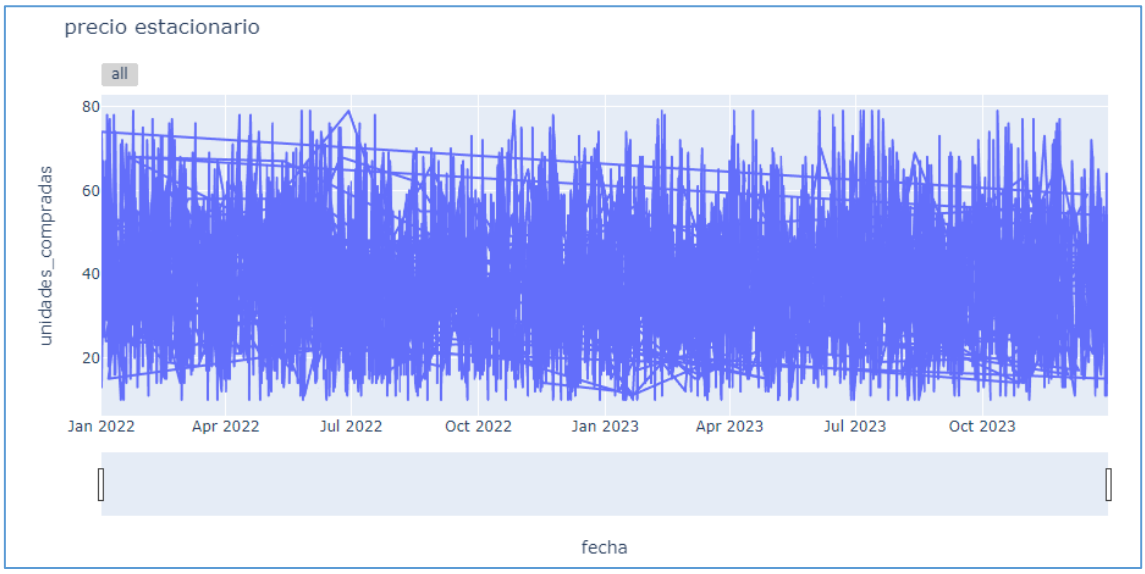
Coeficientes:

- **Crocs:** 5.9056
    - Al ser una variable dicotómica (de sí o no), cuando se establece la predicción de crocs, se sumará automáticamente 5.9 USD al precio (dejando las demás unidades cerradas).
- **Dedo:** 1.6706
    - De igual manera con Sandalias de dedo, una vez aplicada esta variable, el precio se incrementaría en 1.67 más la constante.
- **Sandalias:** 0.5120
    - De igual manera con sandalias con 0.51 USD, siendo este el precio más bajo a añadir.
- **Día de la semana:** 0.0942
    - Dependiendo del día de la semana este puede agregar un valor en 0.0942 debido a la estacionalidad de los precios que se logró visualizar en el análisis exploratorio.
- **Mes:** 0.0541
    - Se hace referencia al mismo caso con día de la semana donde dependiendo del día de la semana este agrega cierto valor.
- **Unidades compradas :** -0.0035
    - Cada unidad comprada adicional reduce el precio en 0.00035 USD esto hace referencia justamente al principio del descuento mencionado en el texto, los descuentos son negociados directamente con el distribuidor.

Modelo AUTOARIMA

Ingresamos el modelo y evaluamos visualmente la información de primera vista para encontrar patrones.

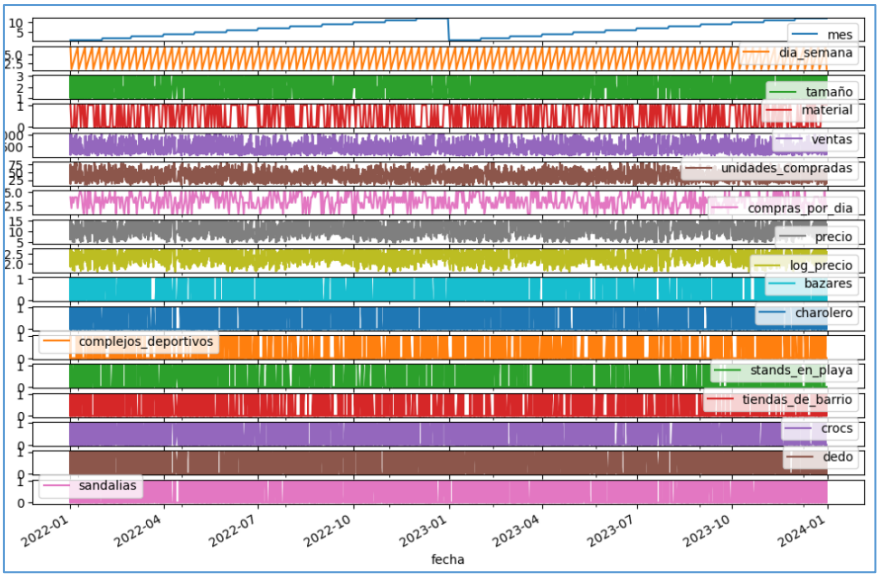
Ilustración 27 Precio estacionario



Fuente: Elaboración del autor

En este caso podemos observar el precio por las unidades compradas, al estar visualizado de manera diaria, puede que exista un sesgo en las unidades acumuladas.

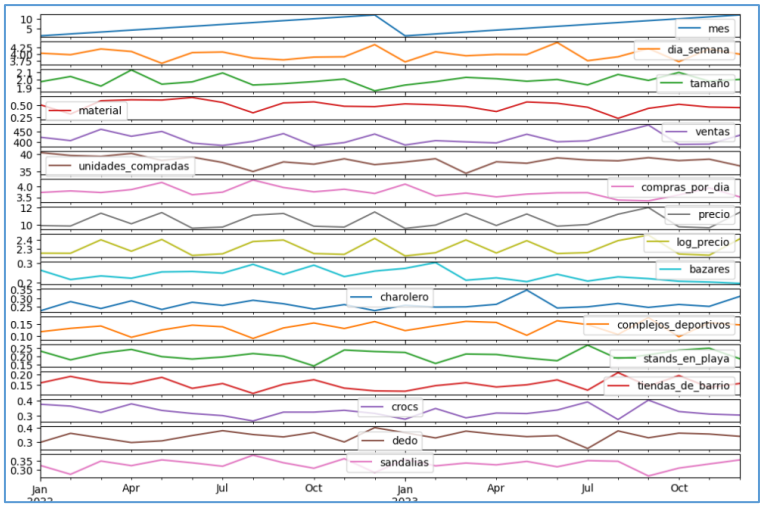
Ilustración 28 fecha vs variables



Fuente: Elaboración del autor

Analizamos las variables y el comportamiento durante el año donde en el lado izquierdo podemos visualizar el rango de respuestas y la columna que se está analizando; previo a realizar un "*resampling*" o re-muestreo aplicado a series temporales con el fin de conseguir patrones más específicos.

Ilustración 29 resampleo de las series temporales

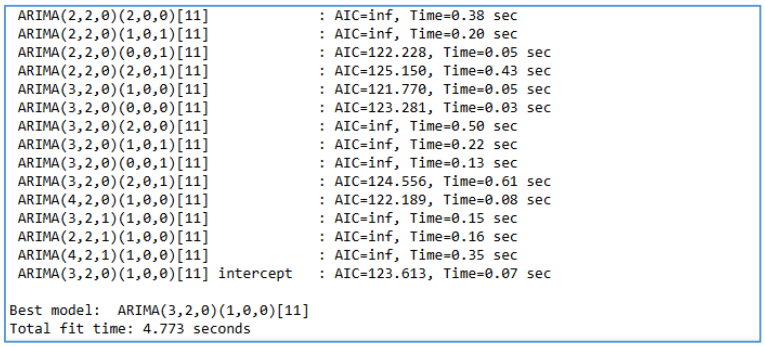


Fuente: Elaboración del autor

Una vez aplicado re-muestreo se puede visualizar tendencias claras de la información por ejemplo los meses picos de compra que son marzo, septiembre, noviembre y diciembre y se conservará estas tendencias para aplicar el modelo autoARIMA.

Procedemos con la aplicación del modelo y definimos la base de pruebas y la base de entrenamiento.

Ilustración 30 Modelo autoarima



```
 ARIMA(2,2,0)(2,0,0)[11]             : AIC=inf, Time=0.38 sec
 ARIMA(2,2,0)(1,0,1)[11]             : AIC=inf, Time=0.20 sec
 ARIMA(2,2,0)(0,0,1)[11]             : AIC=122.228, Time=0.05 sec
 ARIMA(2,2,0)(2,0,1)[11]             : AIC=125.150, Time=0.43 sec
 ARIMA(3,2,0)(1,0,0)[11]             : AIC=121.770, Time=0.05 sec
 ARIMA(3,2,0)(0,0,0)[11]             : AIC=123.281, Time=0.03 sec
 ARIMA(3,2,0)(2,0,0)[11]             : AIC=inf, Time=0.50 sec
 ARIMA(3,2,0)(1,0,1)[11]             : AIC=inf, Time=0.22 sec
 ARIMA(3,2,0)(0,0,1)[11]             : AIC=inf, Time=0.13 sec
 ARIMA(3,2,0)(2,0,1)[11]             : AIC=124.556, Time=0.61 sec
 ARIMA(4,2,0)(1,0,0)[11]             : AIC=122.189, Time=0.08 sec
 ARIMA(3,2,1)(1,0,0)[11]             : AIC=inf, Time=0.15 sec
 ARIMA(2,2,1)(1,0,0)[11]             : AIC=inf, Time=0.16 sec
 ARIMA(4,2,1)(1,0,0)[11]             : AIC=inf, Time=0.35 sec
 ARIMA(3,2,0)(1,0,0)[11] intercept   : AIC=123.613, Time=0.07 sec

Best model:  ARIMA(3,2,0)(1,0,0)[11]
Total fit time: 4.773 seconds
```

Fuente: Elaboración del autor

Ejecutamos el modelo arima y la idea es tener en la estacionalidad = 12 meses, sin embargo eso quiere decir que debería tener más de 24 observaciones, pero el modelo nada más tiene 24 observaciones por cuanto reduciremos este factor a 11 meses, aludiendo que el modelo tiene una periodicidad de 11 tiempos, esto únicamente con fines prácticos, en caso de existir más registros se debe configurar con estacionalidad mensual = 12.

Aquí en el script anterior menciona que el modelo ARIMA Óptimo es el 3.2.0, por lo cual tomará de referencia ese modelo.

Los parámetros se especificarán a continuación:

final_df_crocs'unidades_compradas':

Este es el conjunto de datos que contiene la serie temporal a modelar. En este caso, final_df_crocs'unidades_compradas' representa la columna con los datos de "unidades_compradas" a lo largo del tiempo.
m=11:

Representa el número de períodos en cada ciclo estacional. Es la frecuencia estacional, es decir, el número de observaciones que completan un ciclo estacional (en este caso, cada 11 períodos).
seasonal=True:

Indica que el modelo debe considerar la estacionalidad. Al establecerlo en True, el modelo buscará componentes estacionales de la serie temporal.
start_p=0 y start_q=0:

Estos son los valores iniciales de los órdenes p (autoregresivo) y q (media móvil) para el proceso de búsqueda del mejor modelo. La función auto_arima comenzará a probar modelos ARIMA a partir de estos valores y luego incrementará los órdenes para encontrar el mejor ajuste.
max_order=4:

Especifica el valor máximo para la suma de los órdenes p + d + q + P + D + Q. Este parámetro limita la complejidad del modelo, asegurando que la combinación total de órdenes no exceda 4. Esto ayuda a evitar modelos excesivamente complejos.
test='adf':

Indica la prueba de estacionariedad a aplicar para decidir el orden de diferenciación (d). En este caso, usa la prueba de Dickey-Fuller Aumentada (ADF) para verificar si la serie es estacionaria. Si la serie no es estacionaria, el modelo aumentará el valor de d para diferenciar la serie hasta hacerla estacionaria.
error_action='ignore':

Especifica cómo manejar los errores durante el proceso de ajuste del modelo. ignore significa que, si ocurre un error al ajustar un modelo en particular, auto_arima lo ignorará y pasará a intentar otro modelo en lugar de detenerse.
suppress_warnings=True:

Este parámetro silencia las advertencias generadas durante el ajuste del modelo. Es útil para limpiar el output si no deseas ver mensajes de advertencia en el proceso.

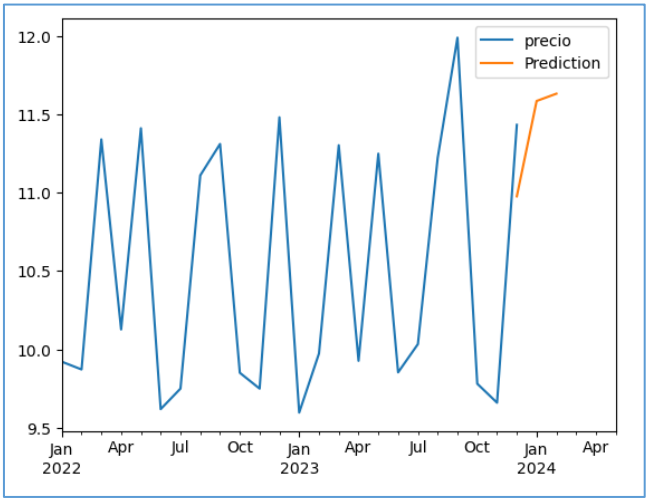
stepwise=True:

Activar la búsqueda por pasos o stepwise. Esta es una técnica de optimización que prueba modelos en un proceso iterativo para encontrar el mejor de manera más rápida y eficiente, en lugar de probar todas las combinaciones posibles.
trace=True:

Al establecer trace=True, se muestra en pantalla el progreso del ajuste del modelo, mostrando cada combinación probada y sus resultados de ajuste. Esto es útil para entender el proceso y ver qué modelos se están considerando.

Tomamos como base de entrenamiento el total de modelo, mientras que para la base de pruebas optaremos por utilizar la información de los últimos 4 meses.

Ilustración 31 Evaluación del predictor Arima



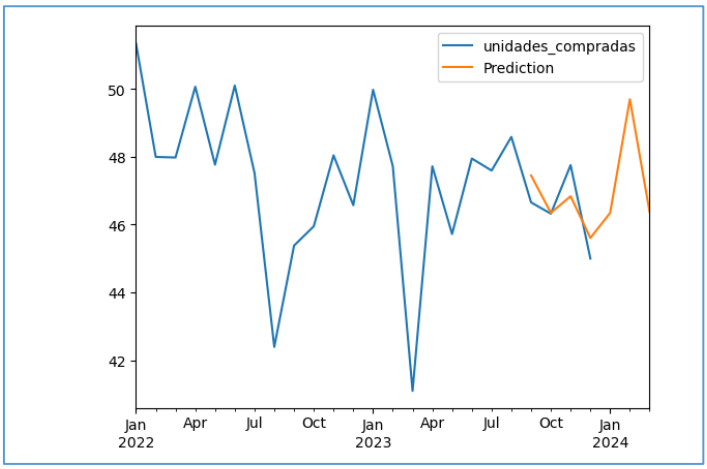
Fuente: Elaboración del autor

En la gráfica se puede observar que el modelo ha adoptado los comportamientos estacionarios del precio y se proyectó la variación del precio para los próximos 4 meses.

Dado que existen 3 SKUs en la base de datos, procederemos con la implementación de un modelo ARIMA por cada SKU:

Modelo ARIMA para CROCS:

Ilustración 32 Predicción para Crocs

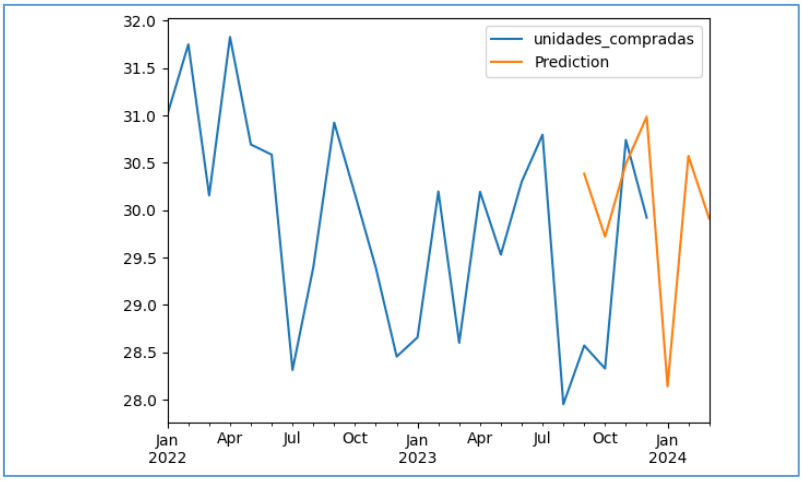


Fuente: Elaboración del autor

Como se observa el modelo a pesar de faltar un periodo para predecir la estacionalidad anual , ha tomado la mejor referencia y las predicciones de las unidades se encuentran dentro de las adecuadas, por tanto la empresa debe tener 50 unidades de crocs disponibles para el mes de enero 2024

Modelo ARIMA para Sandalias:

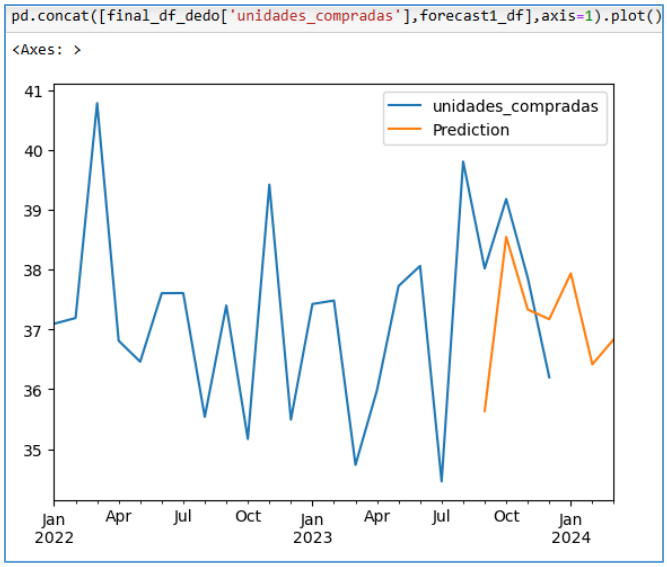
Ilustración 33 Predicción para Sandalias



Fuente: Elaboración del autor

Como se observa el modelo a pesar de faltar un periodo para predecir la estacionalidad anual , ha tomado la mejor referencia y las predicciones de las unidades se encuentran dentro de las adecuadas, por tanto la empresa debe tener 31 unidades de sandalias para diciembre y 28 unidades de sandalias disponibles para el mes de enero 2024

Ilustración 34 Predicción para sandalias de dedo



Fuente: Elaboración del autor.

Como se observa el modelo a pesar de faltar un periodo para predecir la estacionalidad anual , ha tomado la mejor referencia y las predicciones de las unidades se encuentran dentro de las adecuadas, por tanto la empresa debe tener 37 unidades para el mes de enero, en febrero 38 y en febrero 36 en sandalias de dedo.

RESULTADOS

Resultados puntuales del modelo de regresión lineal múltiple:

El modelo elegido es GLSAR por su desempeño en todas las pruebas y supuestos por sobre los otros modelos OLS y WLS, donde se ha logrado proyectar el precio de los 3 skus los cuales son:

- Predicción de precios con GLSAR para **CROCS**: 13.47 USD
- Predicción de precios con GLSAR para **Sandalias de dedo**: 9.23 USD
- Predicción de precios con GLSAR para **Sandalias**: 7.41 USD

La empresa no contaba con una proyección de precios debido a que consta que existe un precio límite de 15 USD y bajo ese precio se empiezan a mover descuentos, sin embargo podemos afirmar los precios anteriormente mencionados como proyección.

Fórmula:

Precio

$$=7.4240+5.9056\times crocs+1.6706\times dedo+0.0942\times dia\_semana+0.0541\times mes+0.5120\times sandalias-0.0035\times unidades\_compradas$$

Explicación de la fórmula:

El Intercepto es (constante): 7.4240

Coeficientes:

- **Crocs:** 5.9056
    - Al ser una variable dicotómica (de sí o no), cuando se establece la predicción de crocs, se sumará automáticamente 5.9 USD al precio (dejando las demás unidades cerradas).
- **Dedo:** 1.6706
    - De igual manera con Sandalias de dedo, una vez aplicada esta variable, el precio se incrementaría en 1.67 más la constante.
- **Sandalias:** 0.5120
    - De igual manera con sandalias con 0.51 USD, siendo este el precio más bajo a añadir.
- **Día de la semana:** 0.0942

    - Dependiendo del día de la semana este puede agregar un valor en 0.0942 debido a la estacionalidad de los precios que se logró visualizar en el análisis exploratorio.
- **Mes:** 0.0541
    - Se hace referencia al mismo caso con día de la semana donde dependiendo del día de la semana este agrega cierto valor.
- **Unidades compradas :** -0.0035
    - Cada unidad comprada adicional reduce el precio en 0.00035 USD esto hace referencia justamente al principio del descuento mencionado en el texto, los descuentos son negociados directamente con el distribuidor.

Resultados puntuales del modelo AUTOARIMA:

El modelo es funcional a pesar de no contar con más periodos para la predicción, forzando a que la temporalidad sea menor a 12 meses es decir m<12 periodos, esto distorsionó ligeramente el modelo, sin embargo en la proyección se pudo observar resultados aceptables, donde con mayor cantidad de registros se puede corregir este desempeño.

En cuanto a predicción de cantidades se obtuvo lo siguiente

- En **Crocs**, la empresa debe tener 50 unidades disponibles para el mes de enero 2024 y entre 46 y 48 para los meses posteriores.
- En **Sandalias**, entre enero y abril la empresa debe tener disponibles entre 28 y 31 unidades.
- En **Sandalias de dedo**, la empresa debe tener 37 unidades para el mes de enero, en febrero 38 y en febrero 36 en sandalias de dedo

Se ha resuelto los 2 principales objetivos del captsone, por cuanto se puede afirmar que el modelo es aplicable también a otras líneas de negocio, por su facilidad de implementación y los resultados que arrojan los modelos.

DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN

Resultados de regresión lineal múltiple:

Se puede resaltar que Crocs tiene mayor incidencia en el precio y más cuando este se vende a medida que avanza la semana y el mes, para esto se debe evaluar la necesidad comercial, si se desea enfocar en disponibilidad o rentabilidad.

- En el caso que se desee enfocar en rentabilidad, se podría dividir la distribución en 3 secciones de la semana por ejemplo: lunes a martes distribuir únicamente Sandalias debido a que estas muestran una menor rentabilidad, de miércoles a viernes distribuir sandalias de dedo debido a que es el segundo SKU con un mejor precio y finalmente sábado y domingo distribuir Crocs ya que estos se venden en mayor cantidad entre 40 y 50 unidades y el precio se eleva los últimos días de la semana. Con todo esto obteniendo la mejor rentabilidad posible
- En el caso de que se desee enfocar en disponibilidad, se debería continuar con el modelo de distribución que se ha venido llevando a cabo debido a que es complicado forzar a los distribuidores a comprar un solo SKU en días específicos, lo que adicionalmente podría ser contraproducente para la empresa incurriendo en una disminución en su rotación de inventarios.

Resultados de AUTOARIMA:

Se observa que no existe mayor variación en cuanto a las unidades requeridas mes a mes, rondando estas entre 40 y 50 unidades por cuanto la rotación de este producto como se había comentado es recurrente a sus unidades de distribución. Tal vez en las otras categorías pueda resaltar el modelo como tal, pero en lo que corresponde a sandalias no.

La idea de esta implementación era evaluar la factibilidad de la implementación a través de los resultados obtenidos, por lo que podemos afirmar que se ha tenido buenos insumos para la toma de decisiones empresariales, por cuanto se podría afirmar que es reditable la inversión en tiempo para cada línea de producto o productos específicos.

CONCLUSIONES Y RECOMENDACIONES

Concluyendo con el objetivo principal, podemos afirmar que se implementó analítica predictiva para la categoría de calzado de playa, ayudando de esta manera a determinar los precios con mayor precisión y su dependencia y su proyección en unidades de ventas

Dentro del modelo de regresión lineal se pudo observar la dependencia de los precios del calzado con la estacionalidad, siendo Crocs más rentable y sandalias menos rentable, adicionalmente que se validó que la estacionalidad incide en el precio y que a mayor avanzada la semana o el mes, más rentabilidad puede existir.

Para el modelo de determinación de la demanda o ARIMA, se pudo validar que el modelo responde favorablemente a la información limitada ingresada, prediciendo las unidades requeridas para los próximos 4 meses, adicionalmente

que se podría observar que existe una mayor demanda de Crocs vs el resto de SKU’s

Finalmente la idea de este captsone era validar la implementación de modelos analítica predictiva a una categoría, lo cual se cumplió a cabalidad y efectivamente se puede aplicar para otras líneas de producto.

En cuanto a las recomendaciones, principalmente se recomienda que la información sea con un detalle diario de mínimo 3 años con el fin de obtener mejores resultados en el modelo de proyección de la demanda.

Se recomienda que las bases de datos tengan amplitud en cuanto a columnas debido a que para efectos del ejercicio se retiraron características del producto que pudieron haber sido significativas

Finalmente se recomienda ampliar el captsone a otras categorías de la empresa con el fin de brindar mayores oportunidades de mejora e insumos para la toma de decisiones.

REFERENCIAS

Améstica Rivas, L. R. (1 de 06 de 2020). *Evaluación predictiva del modelo ARIMA optimizado con fuerza bruta operacional aplicado en el precio del cobre y el índice bursátil Dow Jones 2011-2019*. Obtenido de Evaluación predictiva del modelo ARIMA optimizado con fuerza bruta operacional aplicado en el precio del cobre y el índice bursátil Dow Jones 2011-2019: http://repobib.ubiobio.cl/jspui/handle/123456789/3606

García, S. L. (6 de 05 de 2019). *Factores que influyen en el pH del agua mediante la aplicación de modelos de regresión lineal*. Obtenido de Factores que influyen en el pH del agua mediante la aplicación de modelos de regresión lineal: https://repositorio.uide.edu.ec/handle/37000/3798

Guerrero, M. J. (09 de 11 de 2021). *Cómo calcular el precio con el modelo Van Westendorp*. Obtenido de Cómo calcular el precio con el modelo Van

Westendorp: https://www.minderest.com/es/blog/modelo-pricing-van-westendorp?form=MG0AV3

Hernández Navarro, O. O. (01 de 06 de 2005). *Modelos arima y estructural de la serie de precios promedio de los contratos en el mercado mayorista de energía eléctrica en colombia.* Obtenido de Modelos arima y estructural de la serie de precios promedio de los contratos en el mercado mayorista de energía eléctrica en colombia.: https://repositorio.unal.edu.co/handle/unal/36361

Herrera, D. (5 de 07 de 2019). *Predicción de demanda eléctrica mediante la aplicación de modelos ARIMA y SARIMA en lenguaje de programación R – caso de estudio en la Empresa Eléctrica Quito*. Obtenido de Predicción de demanda eléctrica mediante la aplicación de modelos ARIMA y SARIMA en lenguaje de programación R – caso de estudio en la Empresa Eléctrica Quito: https://bibdigital.epn.edu.ec/handle/15000/20350

IBM. (19 de 08 de 2021). *ARIMA*. Obtenido de ARIMA: https://www.ibm.com/docs/es/spss-modeler/saas?topic=series-arima

Kotler, P. (16 de 9 de 2024). *Cómo se fijan los precios: estrategias clave para el mercado*. Obtenido de Cómo se fijan los precios: estrategias clave para el mercado: Cómo se fijan los precios: estrategias clave para el mercado

Li Ye, K. K. (4 de 07 de 2023). *Aplicación de modelo ARIMA para el pronóstico de exportación de flores del Ecuador.* Obtenido de Aplicación de modelo ARIMA para el pronóstico de exportación de flores del Ecuador.: http://repositorio.ucsg.edu.ec/handle/3317/21884

Manosalvas Pazmiño, D. T. (21 de 11 de 2018). *Análisis del efecto del gasto público en la calidad de la educación para los países miembros y asociados de la OCDE, 2015*. Obtenido de Análisis del efecto del gasto público en la calidad de la educación para los países miembros y asociados de la OCDE, 2015: https://bibdigital.epn.edu.ec/handle/15000/19865?mode=full

Mario, P. (01 de 06 de 2024). *MODELO PREDICTIVO EN LAS EXPORTACIONES NO PETROLERAS DEL ECUADOR* . Obtenido de MODELO PREDICTIVO EN LAS EXPORTACIONES NO PETROLERAS DEL ECUADOR : chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://repositorio.uteq.edu.ec/server/api/core/bitstreams/d829f52f-37d9-48bb-9ecb-bdc8c090e50f/content

Ortiz, R. G., Arias, F. C., Da Silva, C. J., & Cardozo, O. D. (01 de 05 de 2015). *Análisis espacial del precio del suelo con modelos de regresión lineal múltiple (MRLM) y Sistemas de Información Geográfica (SIG), Resistencia (Argentina)*. Obtenido de Análisis espacial del precio del suelo con modelos de regresión lineal múltiple (MRLM) y Sistemas de Información Geográfica (SIG), Resistencia (Argentina): https://ri.conicet.gov.ar/handle/11336/37475

Reino Vivanco Andrés Agustin, T. V. (1 de 06 de 2012). *Universidad Politécnica Salesiana*. Obtenido de Universidad Politécnica Salesiana: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://dspace.ups.edu.ec/bitstream/123456789/3313/1/UPS-CT002541.pdf

Sarmiento Castillo, G. d. (3 de 01 de 2023). *Gasto público social y pobreza en Ecuador, periodo 2007 - 2020*. Obtenido de Gasto público social y pobreza en Ecuador,

periodo 2007 - 2020:
https://dspace.unl.edu.ec/jspui/handle/123456789/25993
SRI. (01 de 10 de 2024). *SAIKU SRI*. Obtenido de SAIKU SRI:
https://srienlinea.sri.gob.ec/saiku-ui/