Taller 1 grupal

Título de máximo 12 palabras.

Rafael Eduardo Torregroza Vidal , Andres Felipe Useche

JUAN CARLOS GARCIA DIAZ

*Estudiante de Ciencia de Datos*

*Profesor, Departamento de Ingeniería Industrial*

*Pontificia Universidad Javeriana, Bogotá, Colombia*

**GUIA TALLERES ANALÍTICA I BASADOS EN CRISP-DM**

**BUSINESS UNDERSTANDING**

**Determine Business objectives**

Background: Tenemos una empresa de ropa y joyería la cual vende por catálogo y el cuerpo directivo espera mediante el análisis si es posible modelar la demanda de la ropa de mujer en cada zona donde están ubicados mediante los recursos que se asignan a cada una de estas.

Business goal: Objetivos de negocio

1. Analizar la relación entre las variables independientes y la variable objetivo ropamujer, con el fin de identificar los factores que influyen significativamente en la demanda o comportamiento de compra de ropa para mujer.
2. Desarrollar un modelo predictivo que estime con precisión los valores de ropamujer a partir de las variables independientes disponibles, con el propósito de mejorar la toma de decisiones estratégicas en campañas de marketing y distribución.

Business success criteria: Como criterio de rendimiento tenemos es RMSE obtenido de los datos predichos de los reales.

**Determine Data mining goals**

Data mining goal: Objetivos de Análisis

1. Desarrollar un modelo de regresión que permita explicar y predecir el comportamiento de la variable ropamujer a partir de las variables independientes disponibles, mediante una ecuación matemática que refleje de forma significativa las relaciones existentes en los datos.

Data mining success criteria: Como criterios

1. R^2
2. AIC

**DATA UNDERSTANDING**

**Dimensiones de las bases de datos:**

Base de datos Test: Cuenta con 12 columnas con 1600 registros.

Base de datos Train: Cuenta con 11 Columnas(Ya que la columna ropamujer es la variable a predecir), con 400 registros.

**Tipos de datos de las variables:**

-Int: idloc - edadloc - correo - paginas - telefono

-núm: impresa - ropamujer

-chr: Tamamer

**Descripción de las variables:**

ropamujer: ventas de prendas para mujer en millones de pesos durante el año.

correo: número de catálogos enviados durante el año.

páginas: número de páginas del catálogo.

teléfono: número promedio de líneas para llamada abiertas para pedidos.

impresa: cantidad gastada en publicidad impresa.

servicio: número de representantes del servicio al cliente.

idmercado: tipo de mercado. Solo se listan números de clases o tipos de mercado que se desea atacar. No se proveerá información adicional.

tamañomer: tamaño potencial del mercado, proyectado de acuerdo a cifras del área de marketing.

Idloc: ID de la tienda.

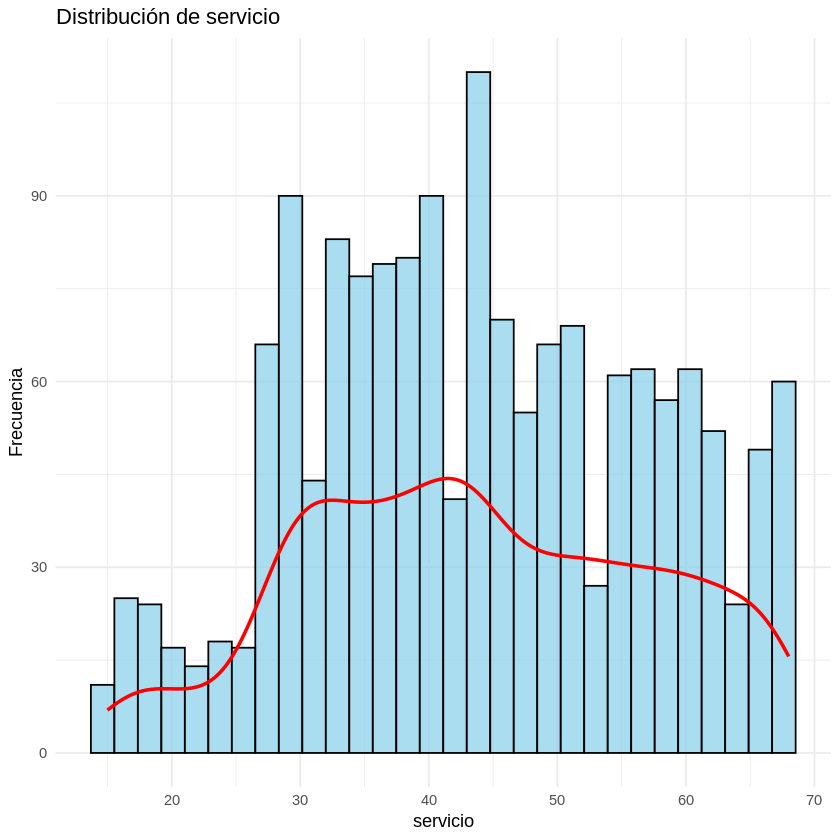
Edadloc: años transcurridos desde la llegada a la zona.

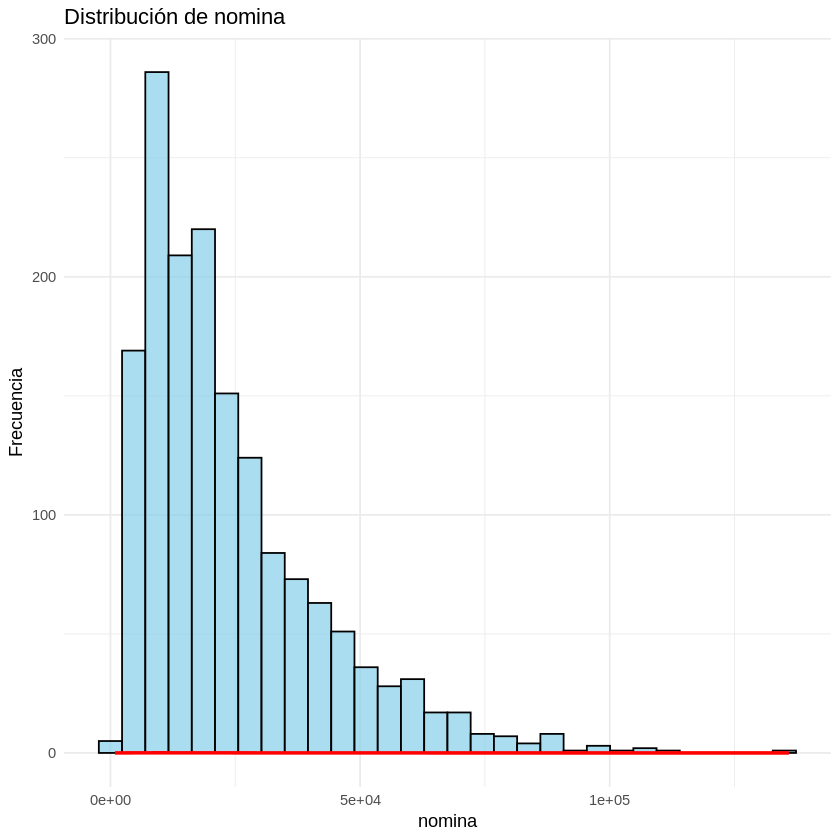
promo (promoción): tipo de promoción que se llevó a cabo durante el año. De nuevo, información descriptiva sobre estas promociones es reservada.

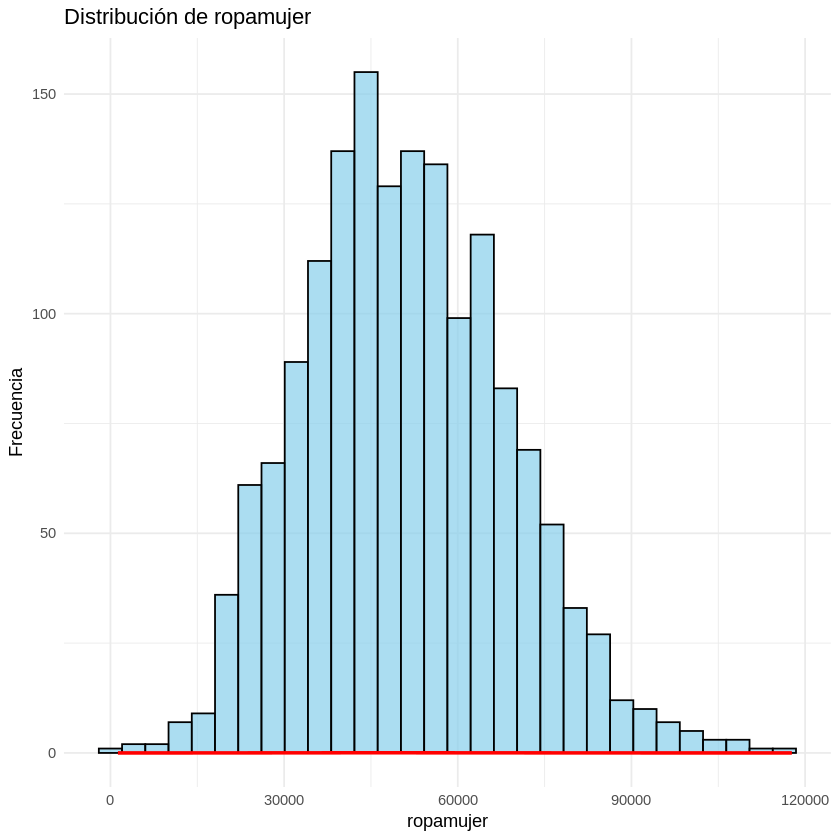
nomina: valor total de la nómina durante el año.

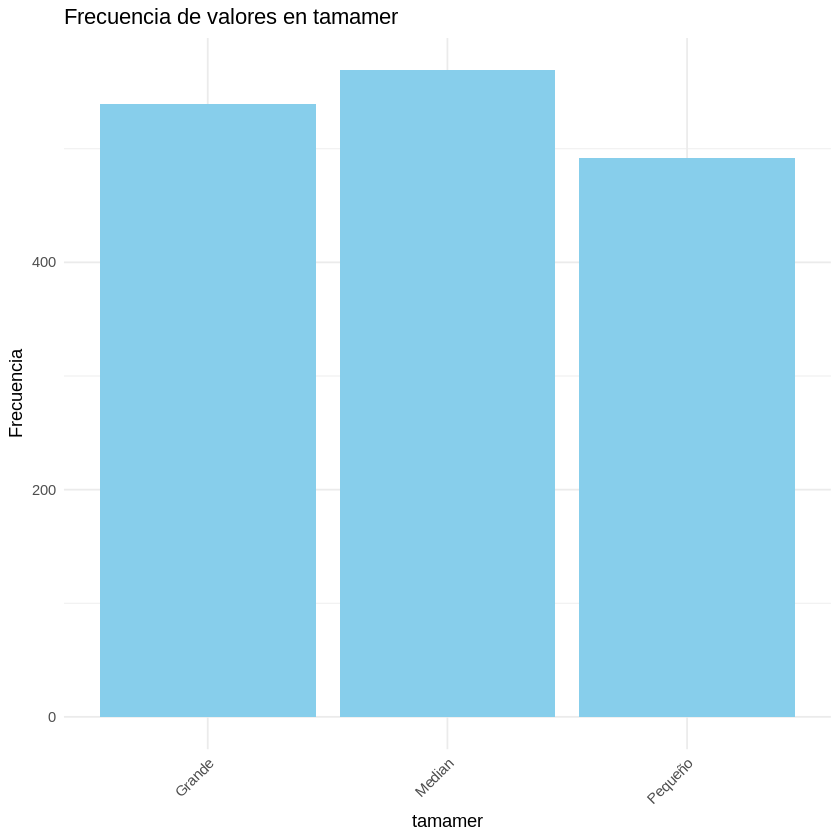
Algunas Gráficas sobre el análisis descriptivo

Estas gráficas corresponden únicamente a las variables de la base de datos train\_data.

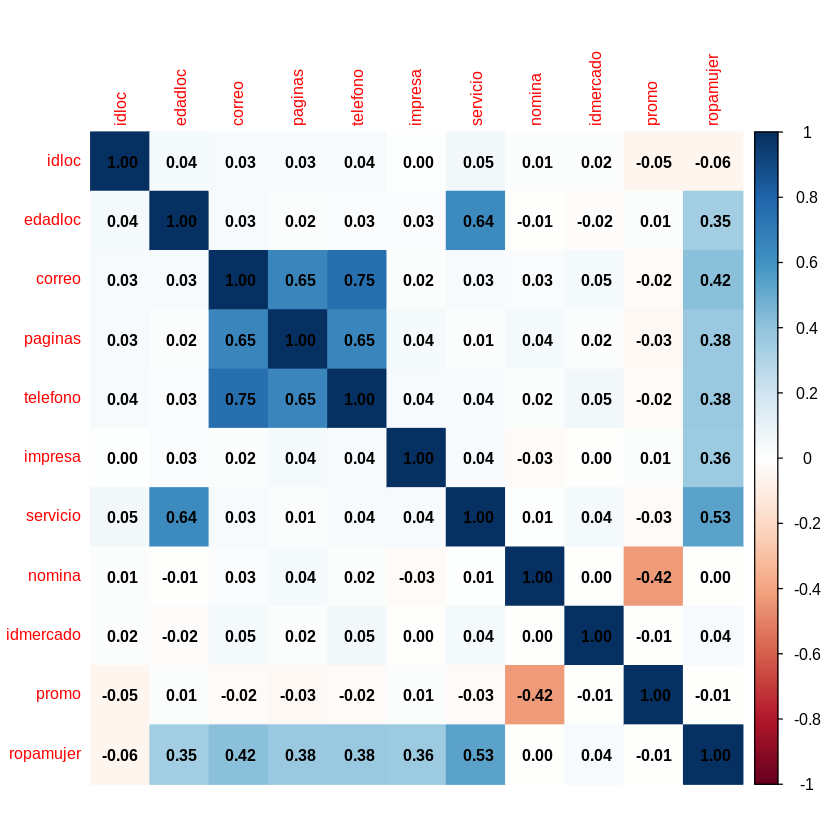








Matriz de correlaciones.



**DATA PREPARATION**

Verify Data Quality: Informar de posibles datos atípicos (tanto posibles como imposibles) y perdidos en las variables, con potenciales explicaciones sobre su aparición.

Valores nulos: Ninguna variable cuenta con valores nulos

Valores atípicos:

Variable dependiente ropa mujer: Se encontraron 9 datos atípicos

Variables independientes: Se encontraron 186 datos atípicos

El AIC es mejor para el modelo que únicamente elimina los outliers de la variable dependiente, por lo tanto se continúa el análisis con este supuesto.

**DATA PREPARATION**

Para la limpieza de datos solamente se eliminaron ciertas filas que contienen valores atípicos los cuales estaban por fuera de rango intercuartílico, no se encontraron datos nulos, faltantes, diferencia en la digitación con los nominales ni los numéricos, finalmente los datos de tipo char se cambiaron por factores para ajustarse a la regresión y se eliminó la columna de Idloc dado que no aporta al análisis siendo una columna de identificación.

**MODELING & EVALUATION**

Las técnicas a emplear son regresión lineal, regresión no lineal y stepwise, el modelo stepwise se usará en lo posible en cada modelo para verificar si las variables usadas son las mejores y entre la regresión lineal y la no lineal se usará aquella que provea un AIC menor.

Inicialmente se hará una regresión con todas las variables de dataset y con el stepwise se filtraran las variables para usar en el siguiente modelo, después de tener las variables a utilizar se aplicará no linealidad variable por variable para verificar que ninguna de las variables finales se desempeña mejor de esa manera, como paso siguiente se hará otro modelo con todas las variables pero esta vez con la comprobación de no linealidad para cada una junto con el stepwise, finalmente se comparan ambos modelos para verificar cuáles variables tienen en común y cuál de los dos da mejores resultados.

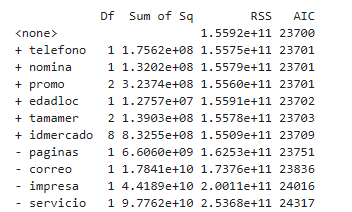
Para el modelo inicial tenemos



con un R^2 = 0.58

AIC=27330

con un modelo final de Step



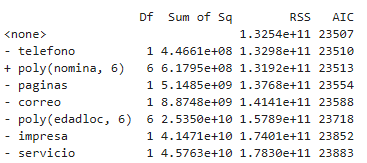
Para la siguiente parte de las pruebas tenemos el modelo



con R^2=0.64

AIC=27125

con un modelo final de Step



Para el último modelo tenemos



con R^2=0.64

AIC=27113

El modelo final siendo uno donde la relación de la variable edadloc es no lineal se modeló usando gam en vez de el poly en lm dado que provee resultados ligeramente mejores según el AIC así mismo que el modelo no lineal usando bs con Knots =(0.10,0.50,0.90)

también se probaron





junto con varios valores para el poly de edadloc de los cuales ninguno dio mejores resultados.

**DEPLOYMENT**

* Finalmente teniendo en cuenta los resultados obtenidos, se le recomienda a la empresa aumentar los recursos de: impresa, paginas, correo, servicio y teléfono si desean aumentar las ventas en ropa de mujer en esa región después de todo tienen una relación directa donde aumentar estos recursos aumenta las ventas, esto principalmente para correo que tiene el mayor efecto, seguido de impresa y teléfono aparte de que con un RMSE de [**10908**](https://www.kaggle.com/competitions/machine-learning-pregrado-2025/submissions#) y un R^2 de 0.64 podemos decir que se explica el comportamiento de las ventas según estas variables.