



Análisis exploratorio de datos

Retail data



Introducción: Corporación favorita

Empresa multinacional latinoamericana con base en Ecuador que está centrada en la industria del retail.

Se buscará predecir 15 días de la venta de 33 familias de productos en los 54 locales de la empresa en Ecuador.

El dataset es de una competencia de Kaggle sobre time-series forecasting

El potencial impacto de un modelo de predicción con alto nivel de precisión en la industria podría significar una reducción en la pérdida de comida gracias a la optimización de stock.

Ecuador es un país el cual tiene una economía que depende fuertemente del petróleo por lo que se incluirá esta información al análisis.



Framework de resolución

1. problema: Cuántas ventas vamos a tener para las 33 familias de productos en los 54 comercios?
2. Data: ¿Qué información tenemos para resolver el problema?
3. ¿Qué dice la data? Hay tendencias y estacionalidad?

Extra:

4. Elección de un modelo de predicción y ajuste del mismo

Data

The **training** data, comprising time series of features store_nbr, family, and onpromotion as well as the sales.

Store data including city, state, type, and cluster.

Daily oil price. Includes values during both the train and test data timeframes.

Holidays and Events, with metadata.

Transactions with dates and store number.

Oil with daily price of oil during the timeframe of the dataset.

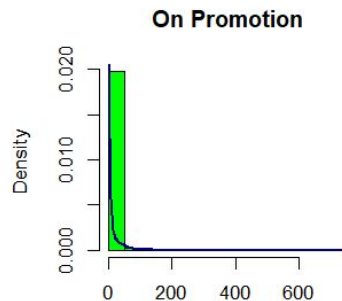
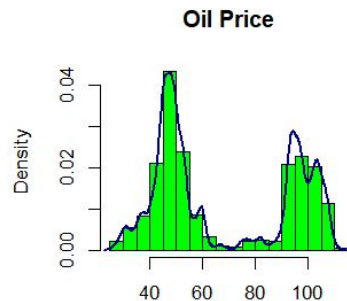
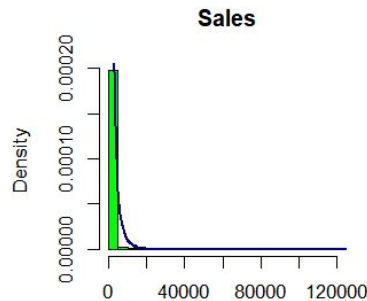
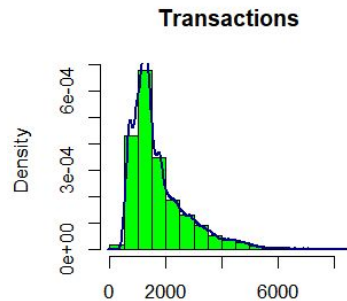
Missing values y Outliers

Los cambios en el precio del barril de crudo afectan gravemente a la economía de Ecuador y, por consecuencia, podrían afectar a nuestro modelo de predicción por eso es importante tener esta información.

- Es la única tabla que presenta Na 's. Alrededor del 3%
- Interpolación lineal como método de imputación de na's



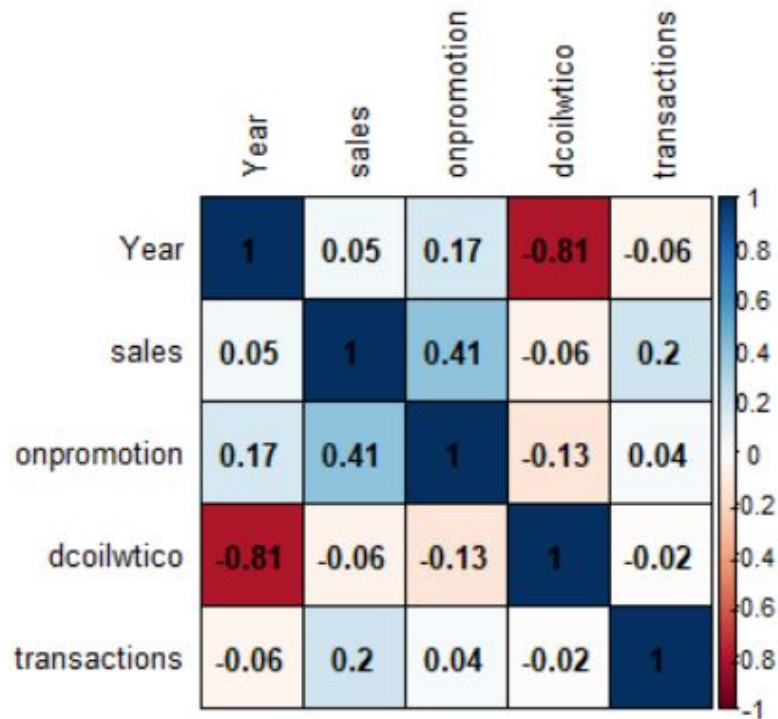
Missing values y Outliers



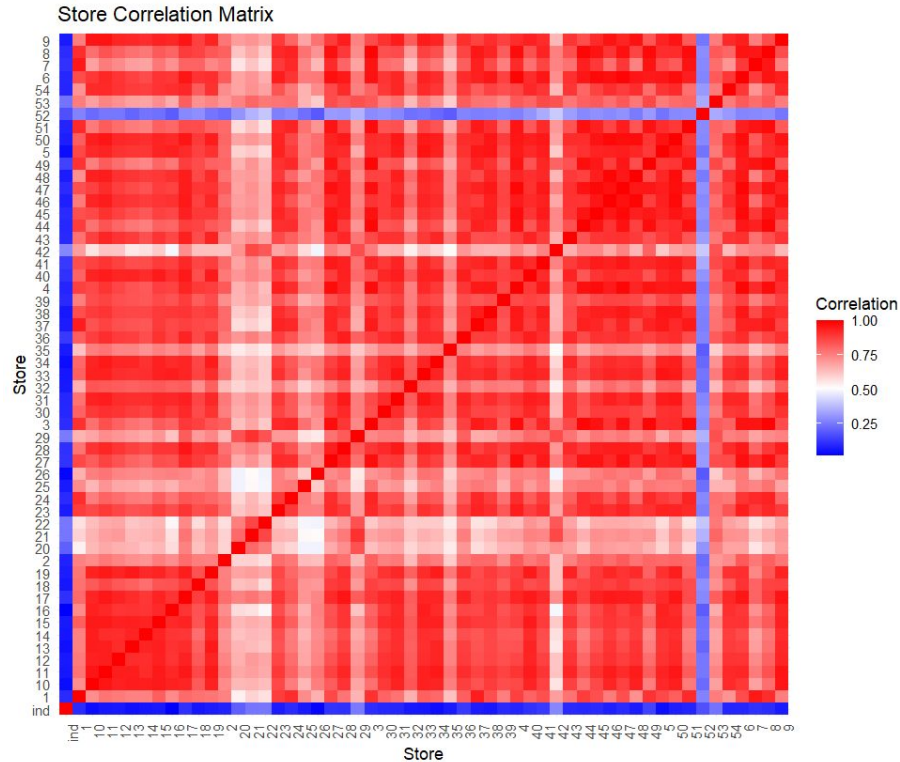
```
> hist_function(df1$transactions, 'Transactions')
      Min. 1st Qu. Median 3rd Qu. Max.
[1,]    5    1046   1395   2081 8359
> hist_function(df1$sales, 'Sales')
      Min. 1st Qu. Median 3rd Qu. Max.
[1,]    0     0    11 196.011 124717
> hist_function(df1$dcoilwtico, 'Oil Price')
      Min. 1st Qu. Median 3rd Qu. Max.
[1,] 26.19   46.41   53.43   95.81 110.62
> hist_function(df1$onpromotion, 'On Promotion')
      Min. 1st Qu. Median 3rd Qu. Max.
[1,]    0     0     0     0    741
```

Se encontró que la tienda con el mayor número de productos en promoción está ubicada en Manta, Manabí, y que esto resulta en un aumento significativo en las ventas y transacciones. Los datos parecen ser legítimos y están correlacionados positivamente en la matriz de correlación.

Correlaciones



Stores y ventas

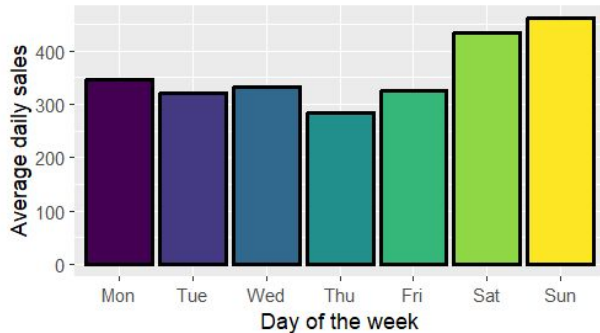


Deep dive

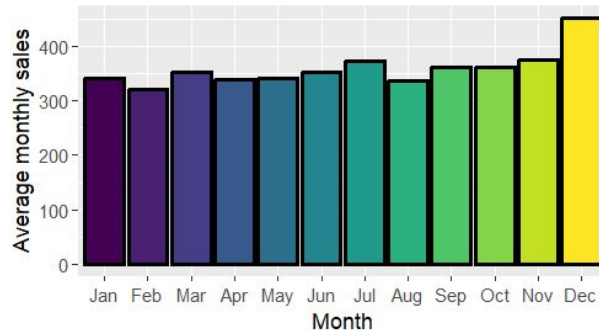


Ventas - Tendencias y estacionalidad

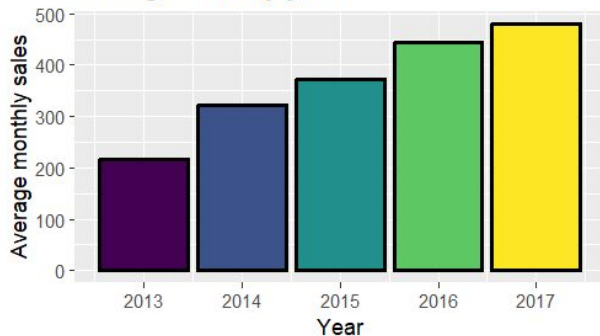
Average sales by day of the week



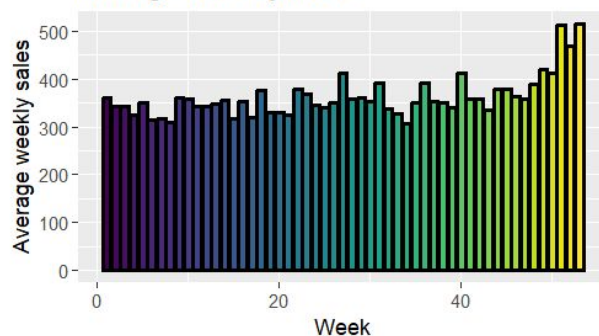
Average sales by month of the year



Average sales by year

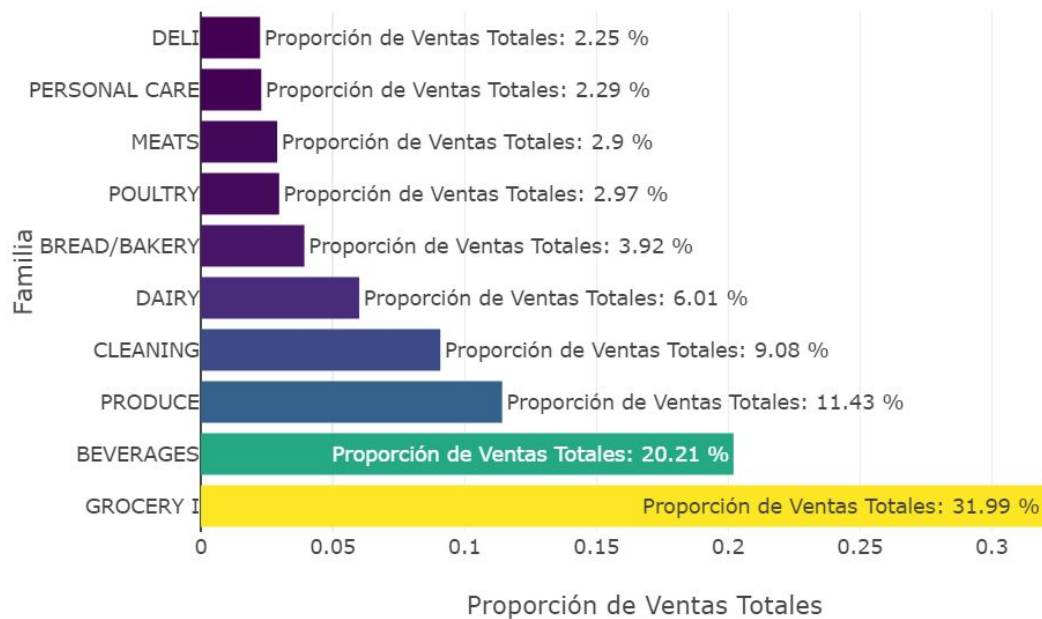


Average sales by week

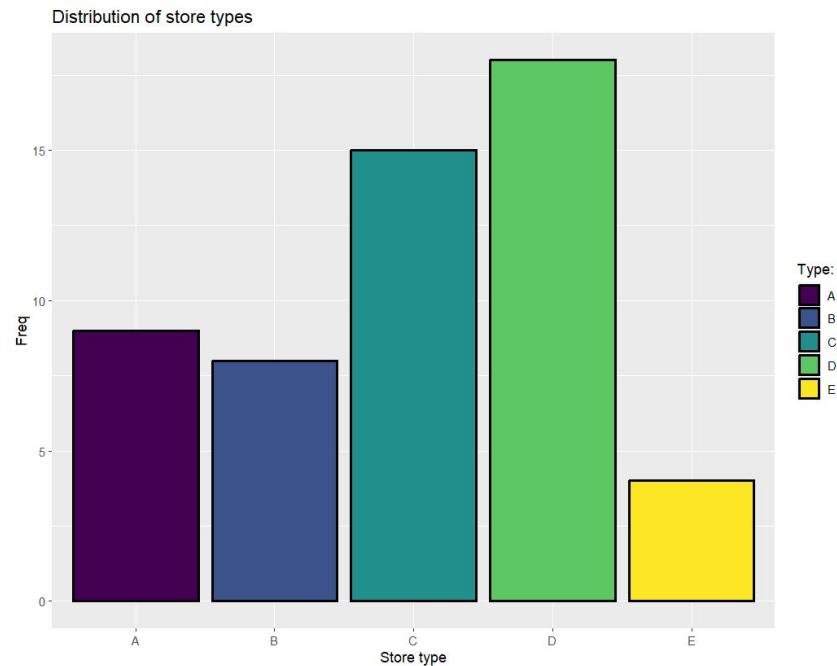
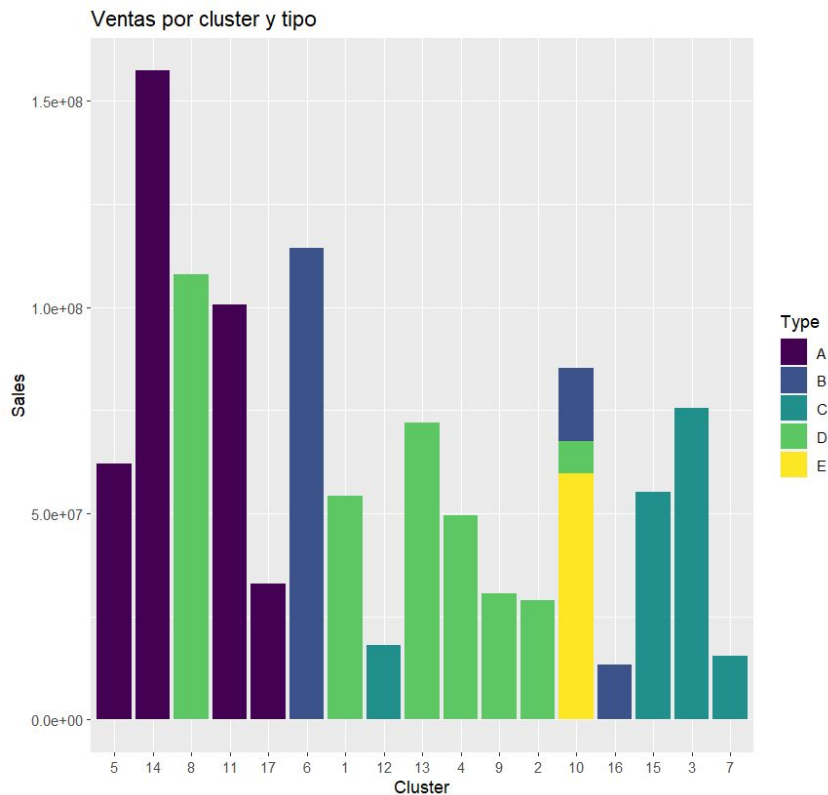


Ventas

10 Mayores Familias por Proporción de Ventas Totales

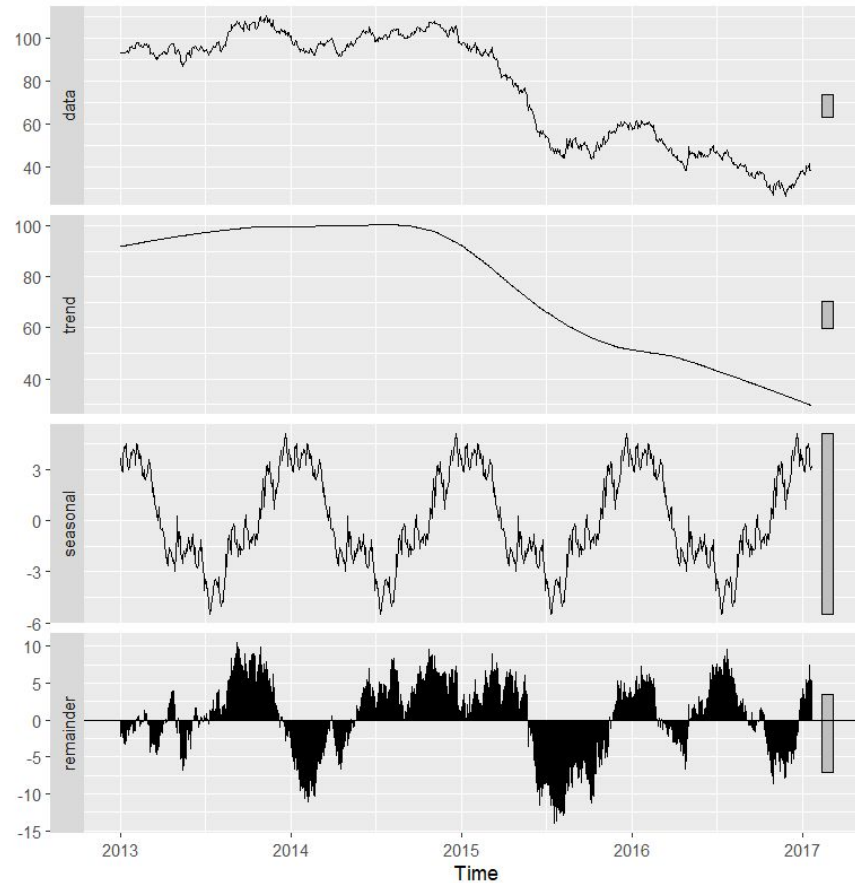


Stores, ventas y tipo



Time series decomposition

Time Series Decomposition del oil price con frecuencia semanal



Time series decomposition

TSD de las ventas promedio diarias con fequencia mensual

