# 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 está 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.

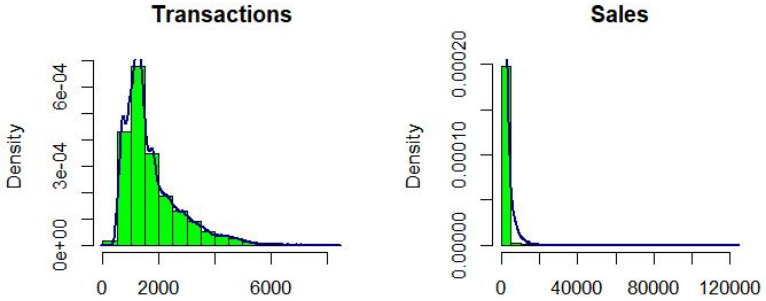
**Oil** with daily price of oil during the timeframe of the dataset.

# 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 está información.

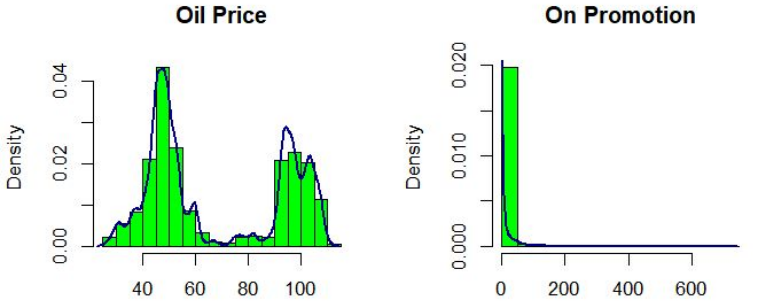
- Es la única tabla que presenta Na 's. Alrededor del 3%
- Interpolación lineal como método de imputación de na’s

# Missing values y Outliers









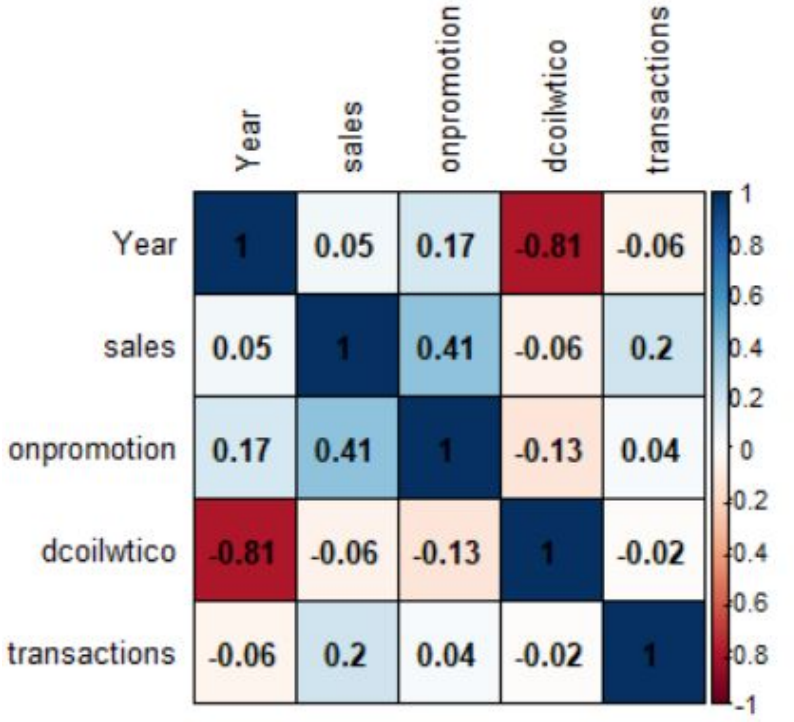




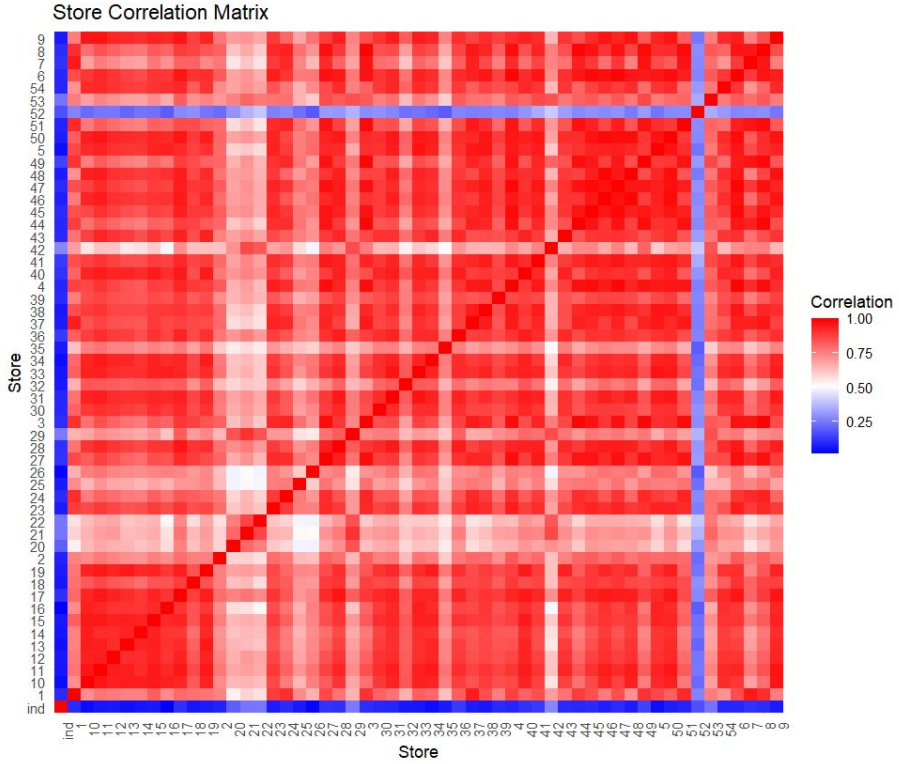
```
> hist_function(df1$transactions, 'Transactions')
     Min. 1st Qu. Median 3rd Qu. Max.
[1,]    5    1046   1395    2081 8359
> hist_function(df1$sales, 'Sales')
     Min. 1st Qu. Median 3rd Qu.   Max.
[1,]    0       0     11 196.011 124717
> hist_function(df1$dcoilwtico, 'Oil Price')
      Min.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



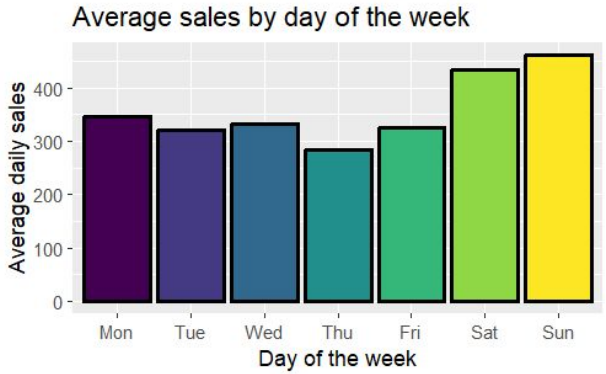
# Stores y ventas

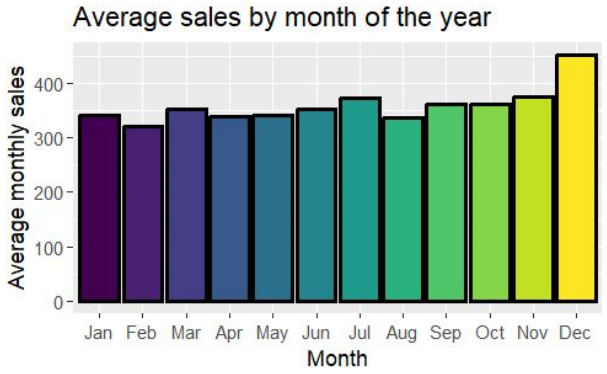


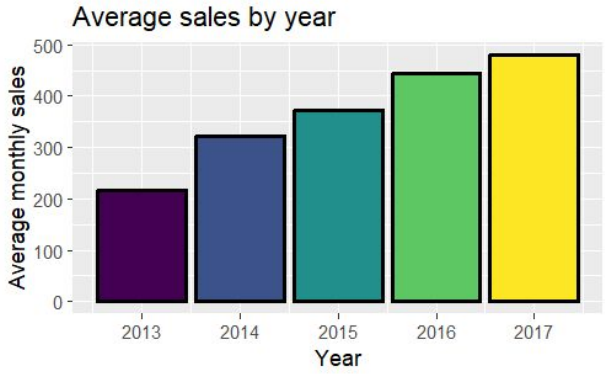
# Deep dive

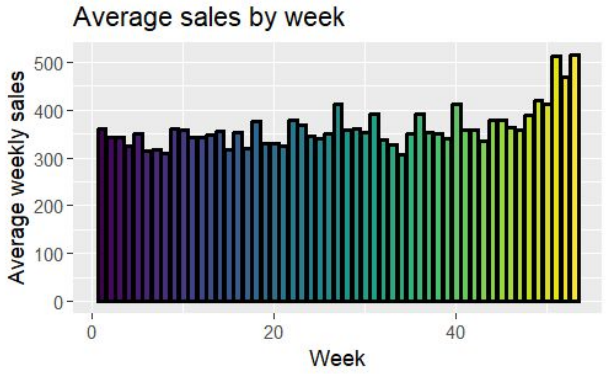

Daily total sales of the stores
20
21
22
29
35
42
52
Total Sales
Total Sales
120000
90000
60000
30000
0
30000
20000
10000
0
20000
15000
10000
5000
0
50000
40000
30000
20000
10000
0
25000
20000
15000
10000
5000
0
Apr 2013
Oct 2013
Apr 2014
Oct 2014
Apr 2015
Oct 2015
Apr 2016
Oct 2016
Apr 2017
Oct 2017
Date

# Ventas - Tendencias y estacionalidad

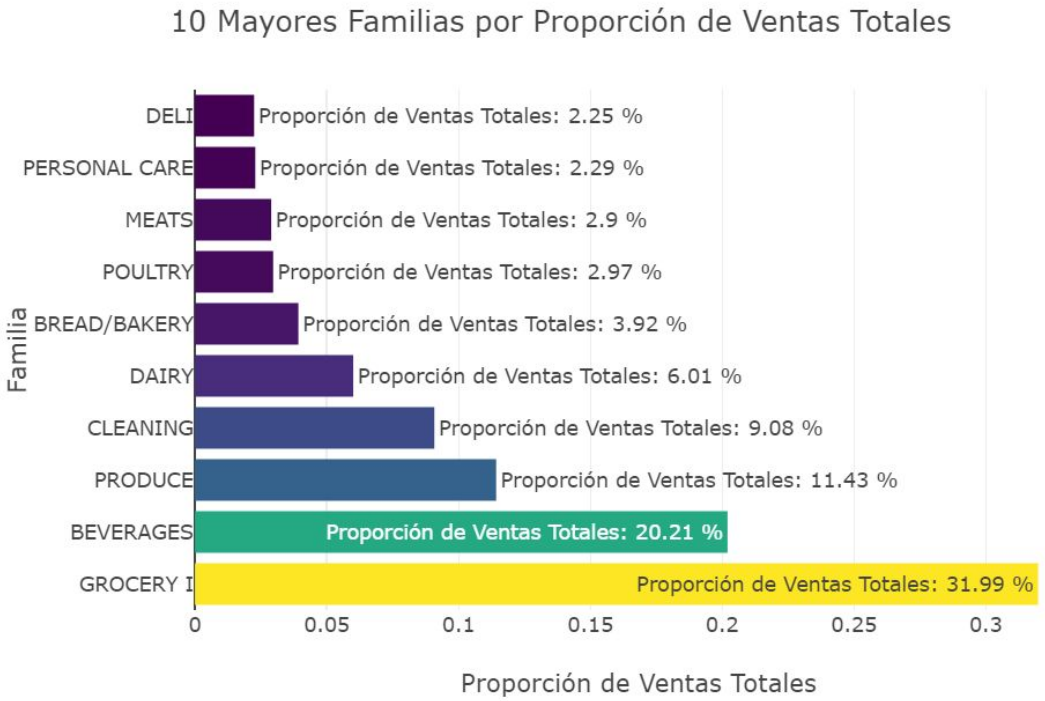


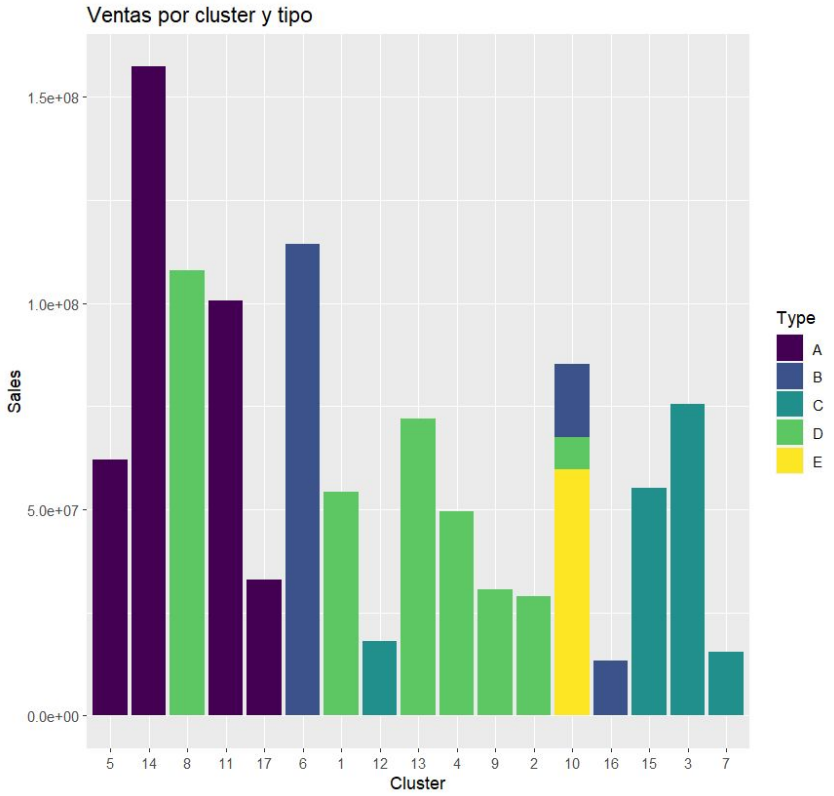



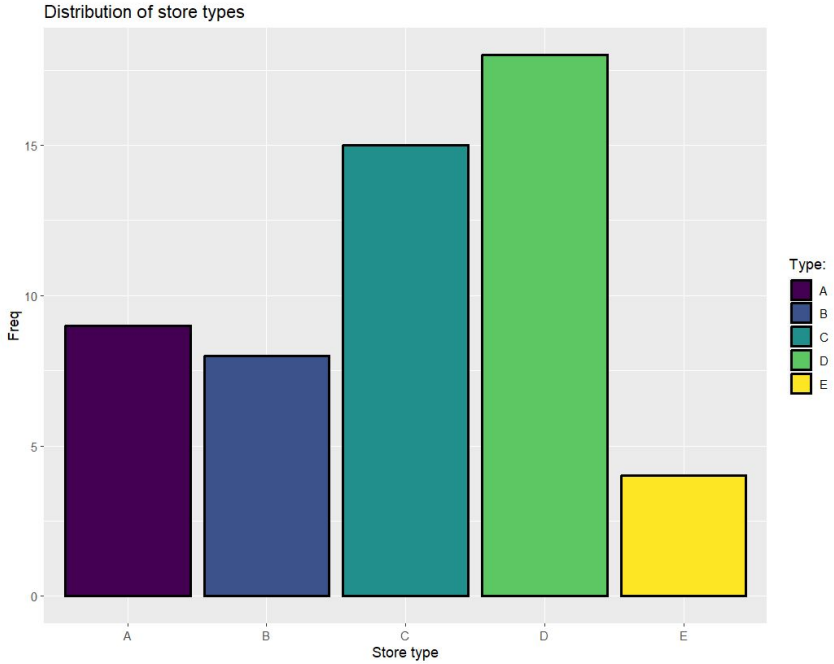



# Ventas

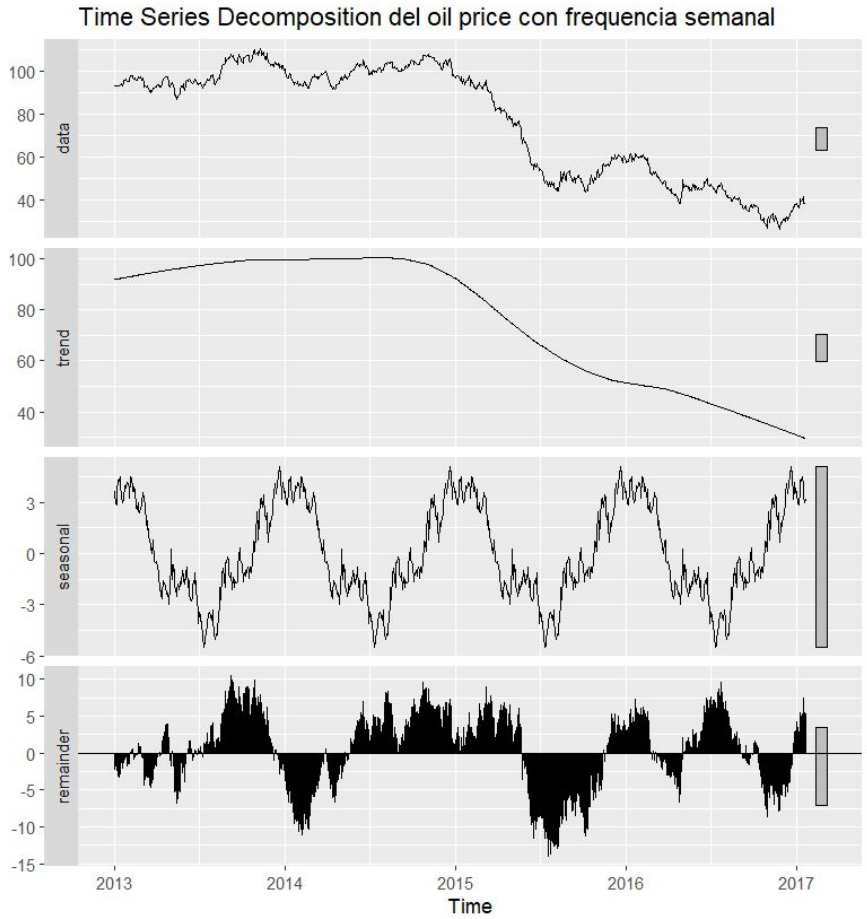

10 Mayores Familias por Proporción de Ventas Totales
DELI
PERSONAL CARE
MEATS
POULTRY
BREAD/BAKERY
DAIRY
CLEANING
PRODUCE
BEVERAGES
GROCERY I
Proporción de Ventas Totales: 2.25 %
Proporción de Ventas Totales: 2.29 %
Proporción de Ventas Totales: 2.9 %
Proporción de Ventas Totales: 2.97 %
Proporción de Ventas Totales: 3.92 %
Proporción de Ventas Totales: 6.01 %
Proporción de Ventas Totales: 9.08 %
Proporción de Ventas Totales: 11.43 %
Proporción de Ventas Totales: 20.21 %
Proporción de Ventas Totales: 31.99 %
Familia
0
0.05
0.1
0.15
0.2
0.25
0.3
Proporción de Ventas Totales

# Stores, ventas y tipo

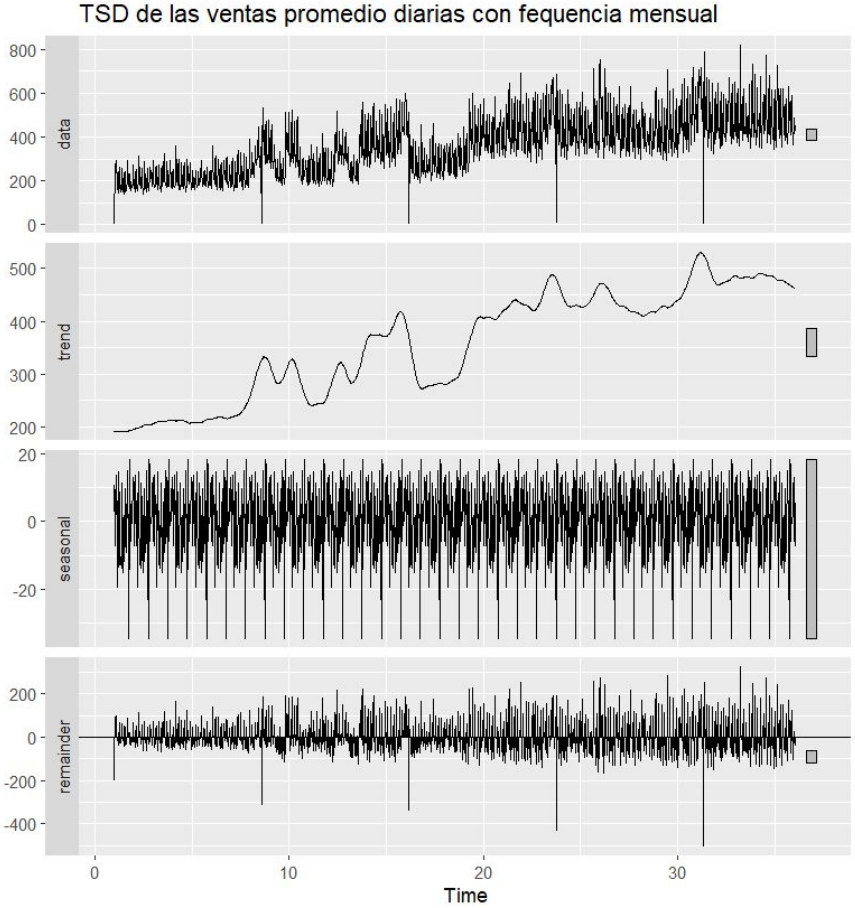


# Time series decomposition



# Time series decomposition



# Muchas gracias!

Fuente de datos:
https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data