Final Project

Data Science Bootcamp

9th Group

Mentor:
Abdullah Ghifari

Wafa Hasnaghina
Yeni Rahmatika Maizar

TOC

Overview

Problem to solve

Project objectives

Data Exploration

Process

Insight

Overview

Planning a celebration is a balancing act of preparing just enough food to go around without being stuck eating the same leftovers for the next week. The key is anticipating how many guests will come. Grupo Bimbo must weigh similar considerations as it strives to meet daily consumer demand for fresh bakery products on the shelves of over 1 million stores along its 45,000 routes across Mexico.

Currently, daily inventory calculations are performed by direct delivery sales employees who must single-handedly predict the forces of supply, demand, and hunger based on their personal experiences with each store. With some breads carrying a one week shelf life, the acceptable margin for error is small.

Problems to solve

In this competition, Grupo Bimbo invites Kagglers to **develop a model to accurately forecast inventory demand based on historical sales data**. Doing so will make sure consumers of its over 100 bakery products aren't staring at empty shelves, while also reducing the amount spent on refunds to store owners with surplus product unfit for sale.

Project objective

To make a machine learning model that good enough for helping Grupo Bimbo in forecast (view, manage, decide) their inventory demand.

The Business

01

Grupo Bimbo requires **Data Demand** on its transactions, so they can provide supply without wasting the inventory.

The Data

02

Do we understand the business process? **Yes**.

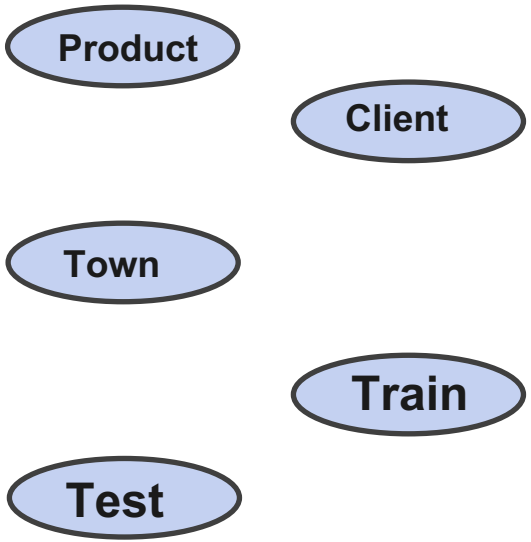
Now, it’s time to explore the data!

Data Exploration

The labels of the data using Spanish

1. Semana — **Week number** (From Thursday to Wednesday)
2. Agencia_ID — **Sales Depot ID**
3. Canal_ID — **Sales Channel ID**
4. Ruta_SAK — **Route ID** (Several routes = Sales Depot)
5. Cliente_ID — **Client ID**
6. NombreCliente — **Client name**
7. Producto_ID — **Product ID**
8. NombreProducto — **Product Name**
9. Venta_uni_hoy — **Sales unit this week** (integer)
10. Venta_hoy — **Sales this week** (unit: pesos)
11. Dev_uni_proxima — **Returns unit next week** (integer)
12. Dev_proxima — **Returns next week** (unit: pesos)
13. Demanda_uni_equil — **Adjusted Demand** (integer) (This is the target you will predict)

What data are available?



Number of Data

The number of **Train** data is **67000 rows** without **columns Id.**

The number of **Test** data is **33000 rows** without **columns Demanda_uni_equil** (Demand unit next week)**.**

What are some things that affect Demand?

Venta_uni_hoy

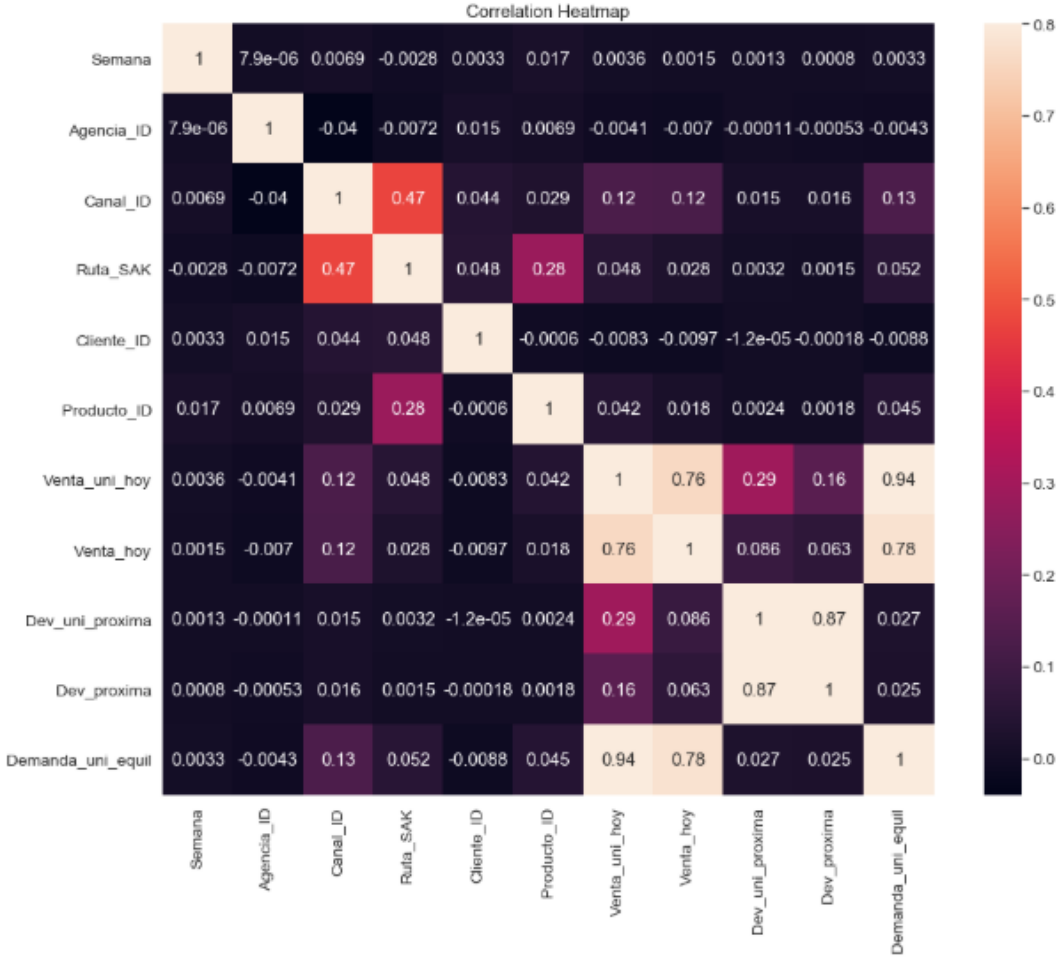
(Sales unit this Week)

Dev_uni_proxima

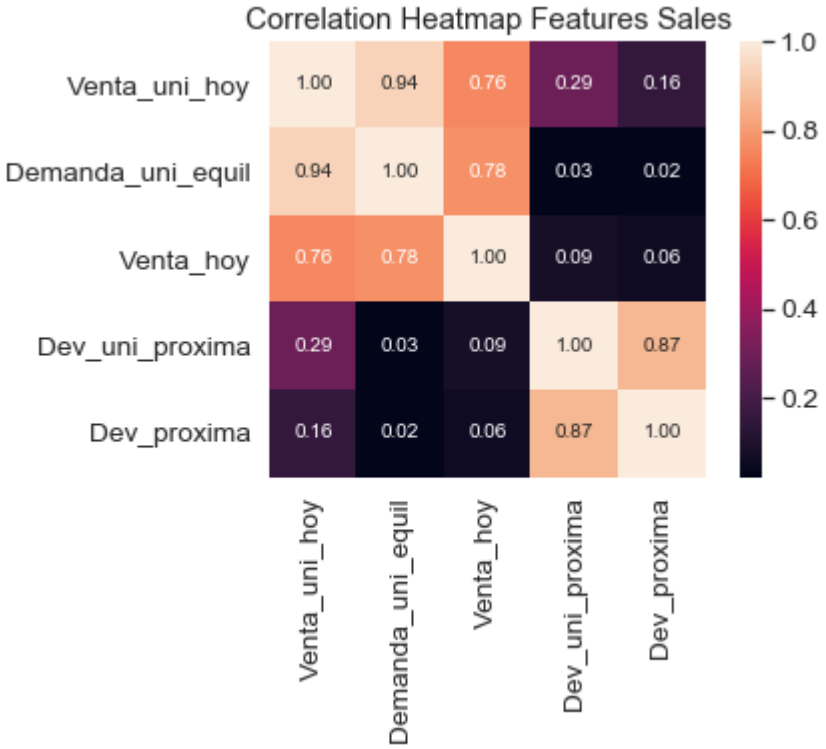
(Return unit next Week)

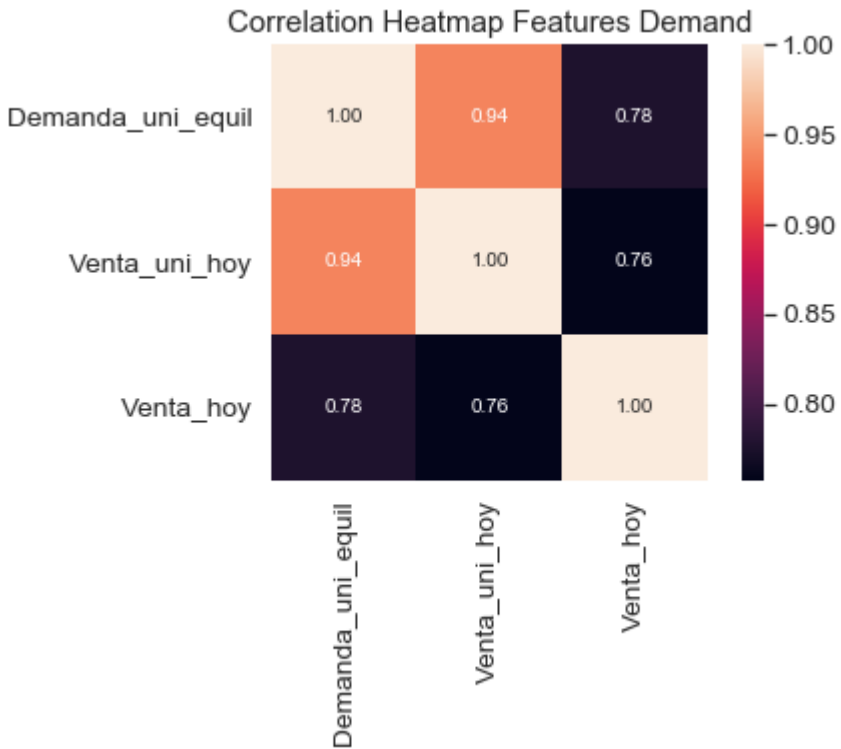
Correlation

Before we move forward to analyze the data, first we decided to see the correlation between features in data.



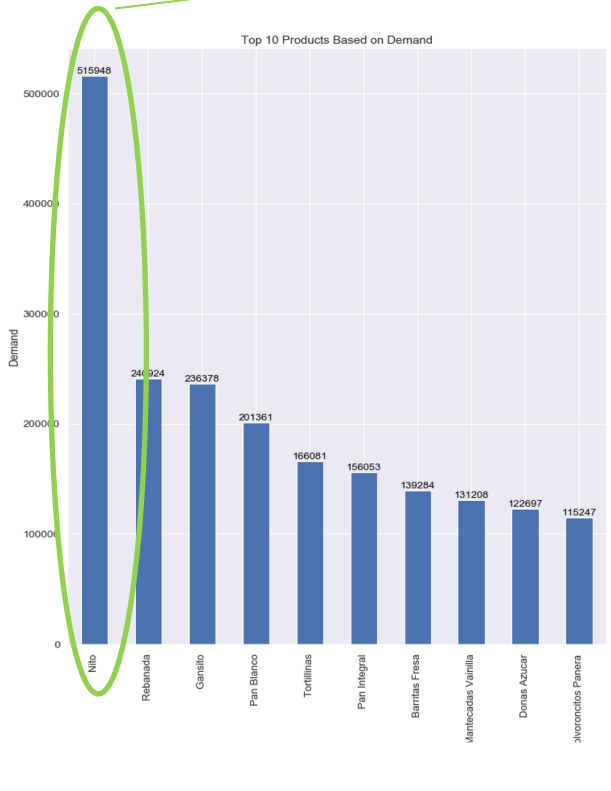
Features

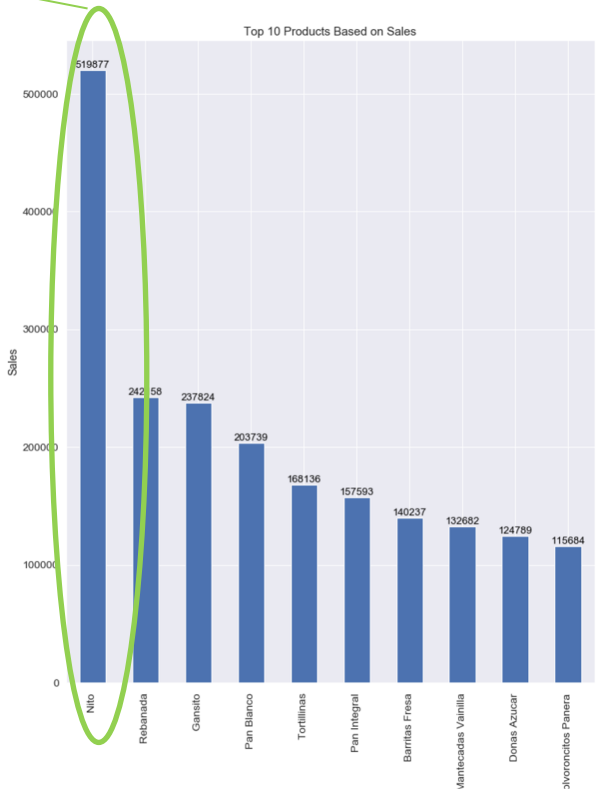

Correlation Heatmap Features Sales
Venta_uni_hoy
Demanda_uni_equil
Venta_hoy
Dev_uni_proxima
Dev_proxima
1.00 0.94 0.76 0.29 0.16
0.94 1.00 0.78 0.03 0.02
0.76 0.78 1.00 0.09 0.06
0.29 0.03 0.09 1.00 0.87
0.16 0.02 0.06 0.87 1.00
1.0
0.8
0.6
0.4
0.2
Venta_uni_hoy
Demanda_uni_equil
Venta_hoy
Dev_uni_proxima
Dev_proxima


Correlation Heatmap Features Demand
Demanda_uni_equil
Venta_uni_hoy
Venta_hoy
1.00 0.94 0.78
0.94 1.00 0.76
0.78 0.76 1.00
1.00
0.95
0.90
0.85
0.80
Demanda_uni_equil
Venta_uni_hoy
Venta_hoy

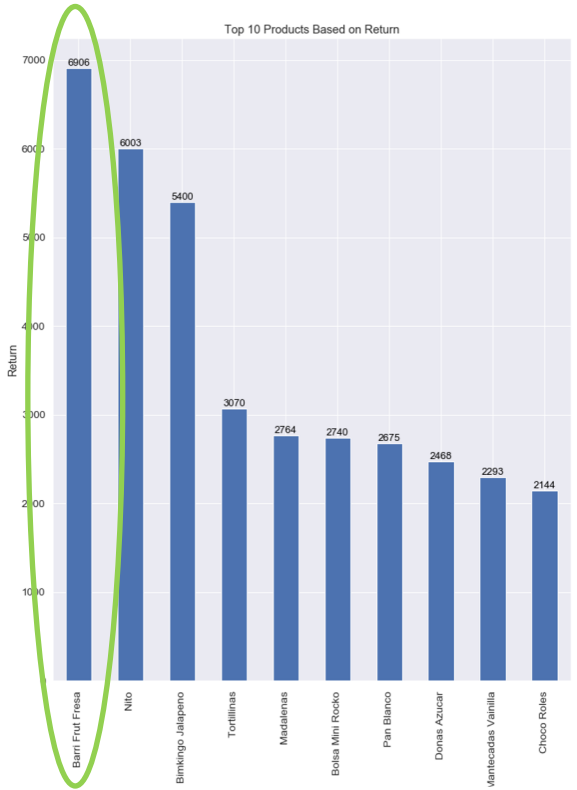
Product

Demand and Sales
Peak : Nito

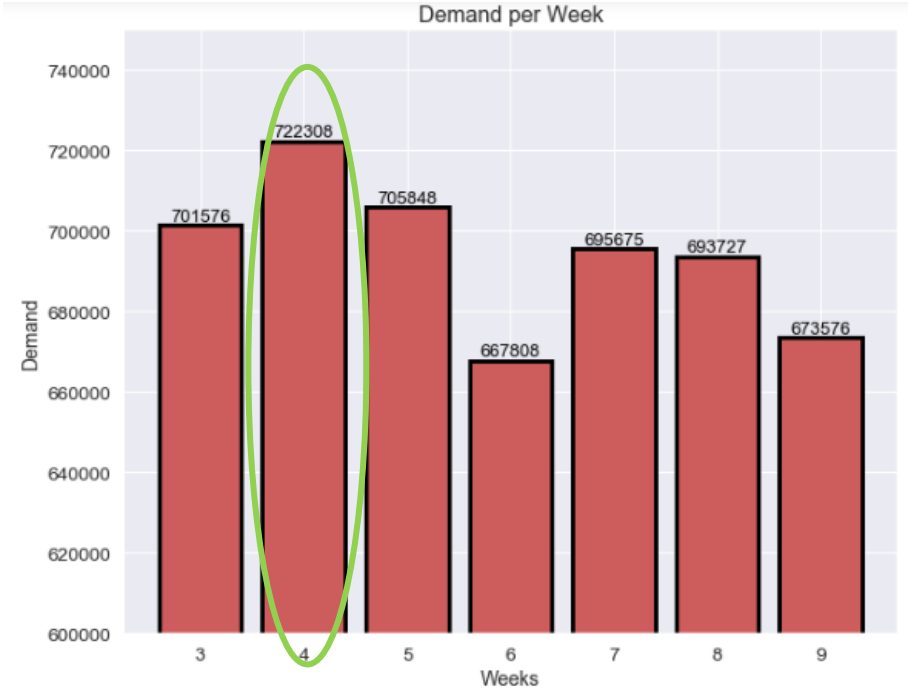




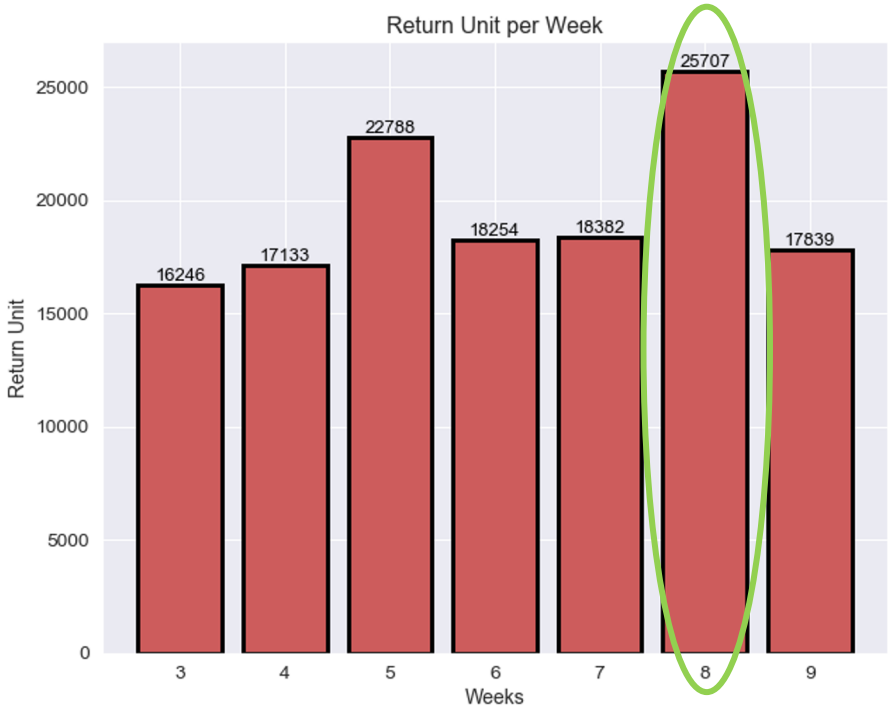
Product vs Return
Peak : Barri Frut Fresa



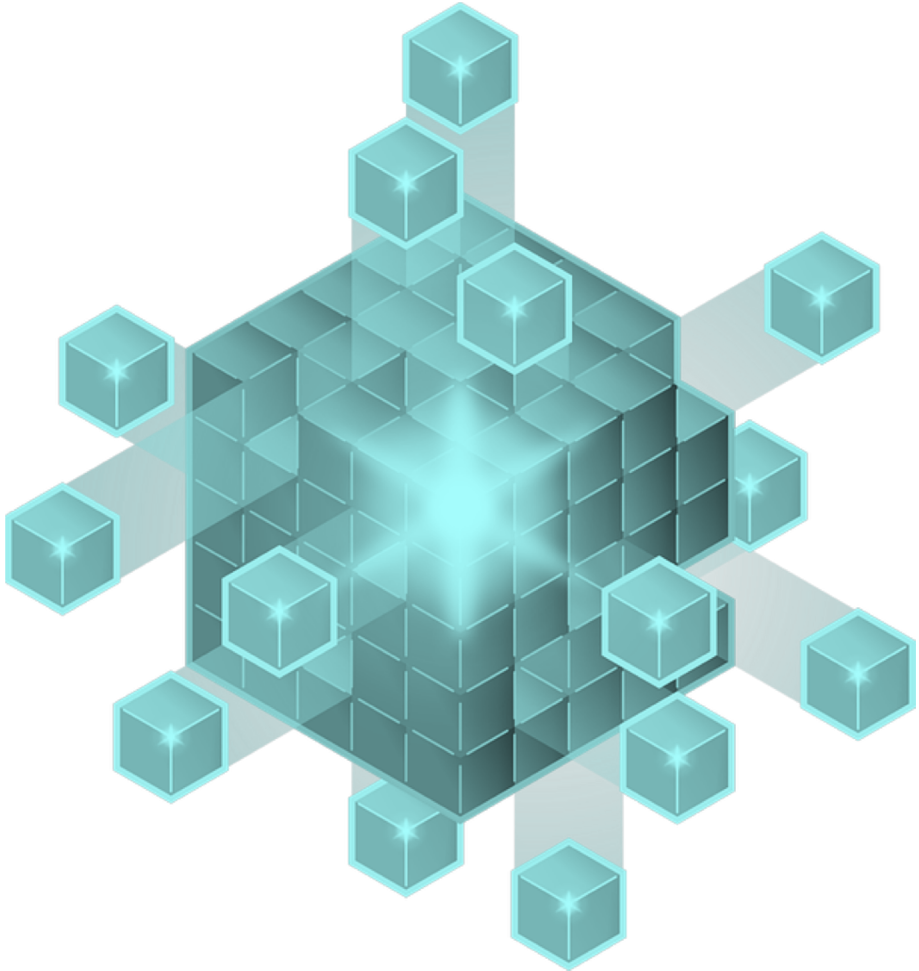
Based on Weeks



Based on Weeks



Process

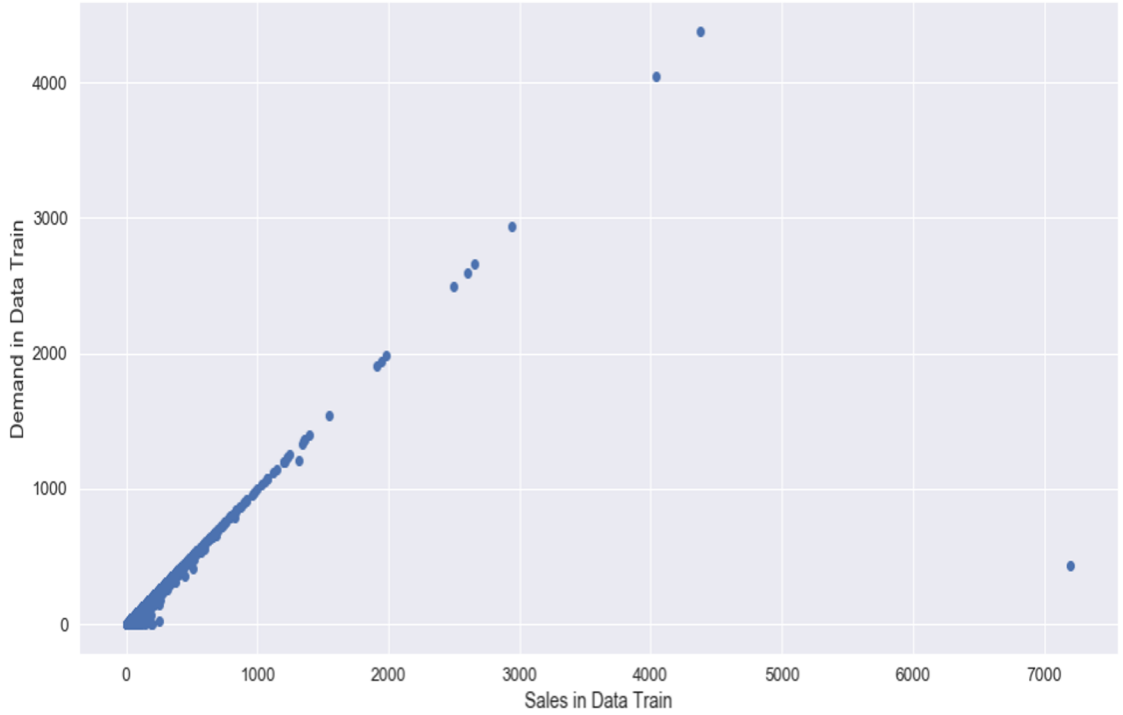


Pre-Processing

- Are there missing values in the data?
- Are there duplicates data?
- What about the distribution of data?

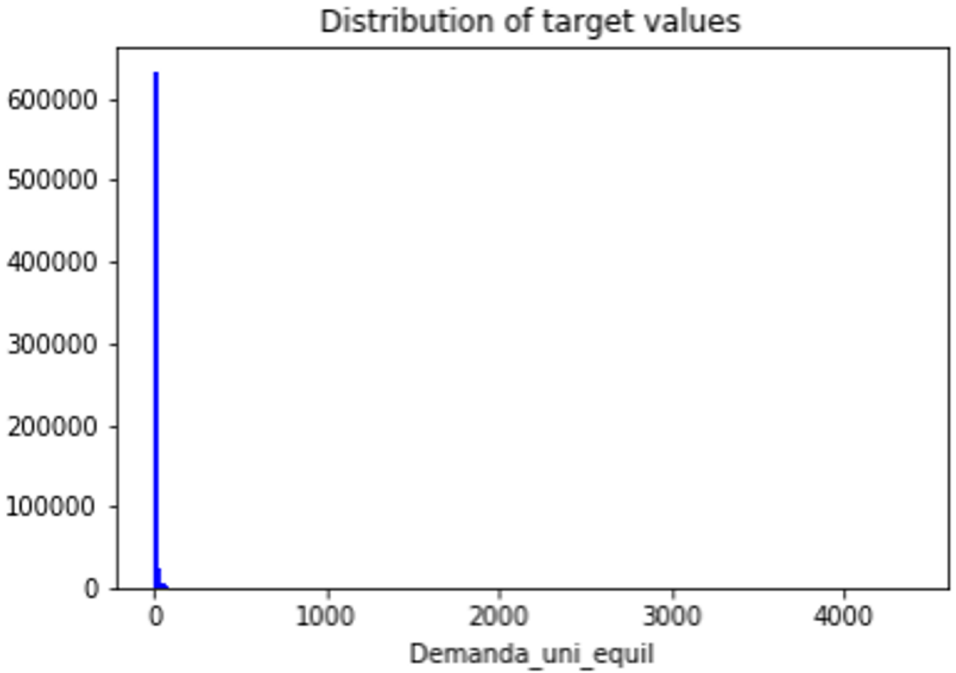
Data Distribution

Data distribution between sales and demand.



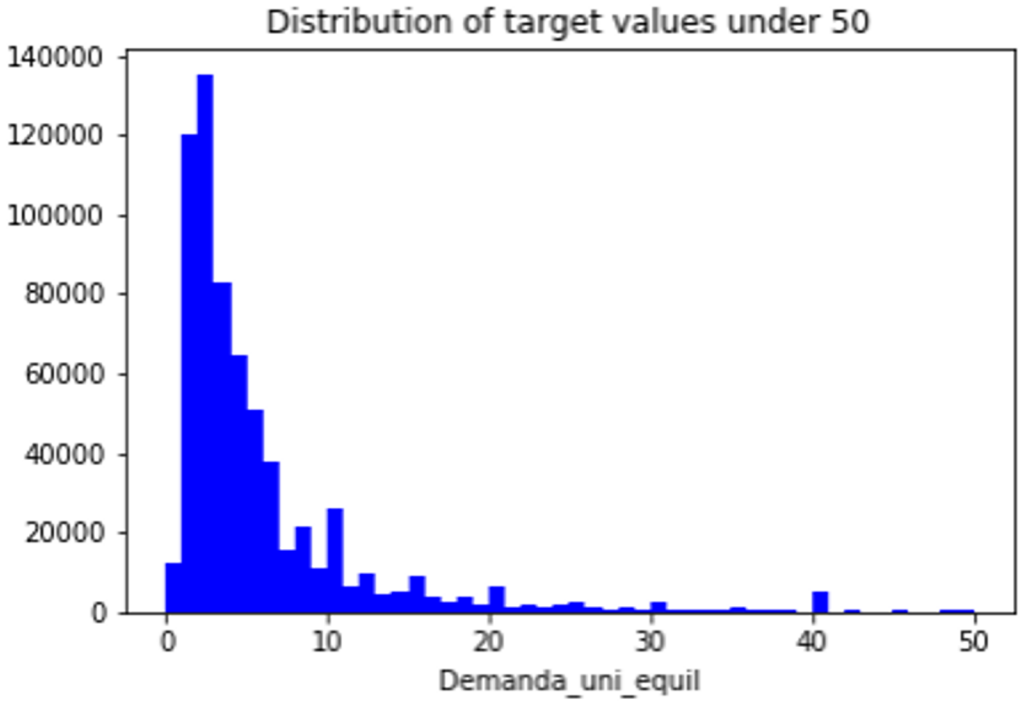
Data Distribution

If we look at the graph on the side, we can clearly see that there is an uneven distribution of data



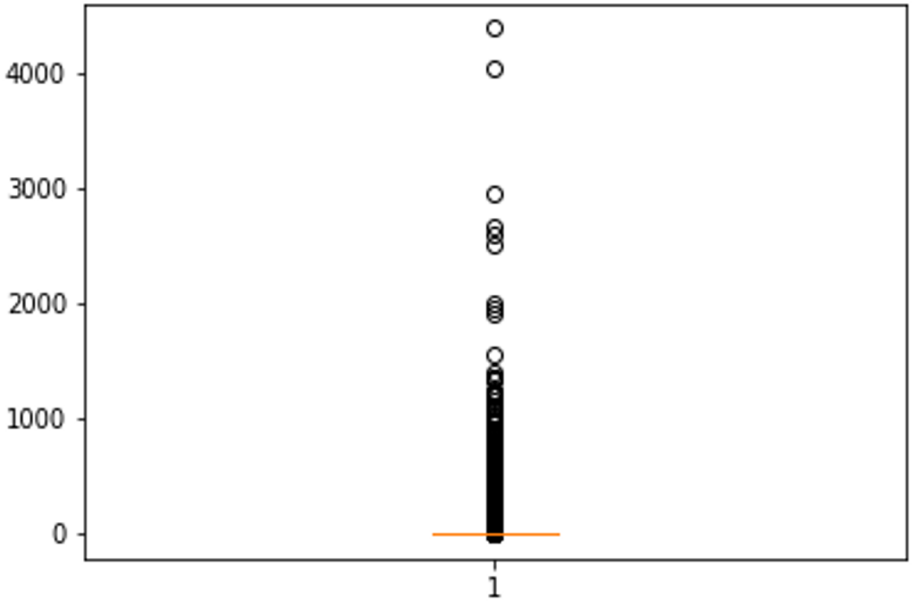
Data Distribution

The data is **skew to the right**, which means that **mean > median.**



Data Distribution

When we see in Boxplot, there are a lot of outlier out there. And we should get rid of those.



Process

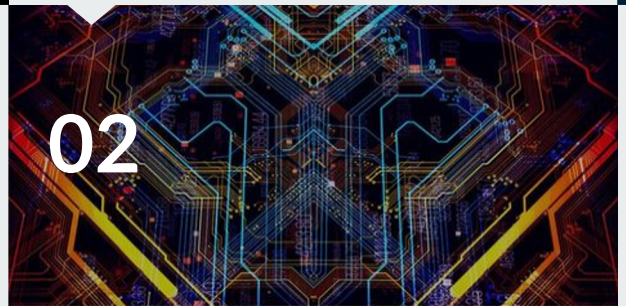
Found the Skew

Venta_uni_hoy (Sales)	78.2
Demanda_uni_equil (Demand)	47.9
Dev_uni_proxima (Return)	559.5

Cut the outlier

We cut the outliers because outliers interrupt existing data. We use **Z-Score**.

Z Score =
(Observation - Mean) / Standard deviation



Do the Test

Enter “clean” data into the test and make predictions with several models.

What models do we use?

- Multiple Linear Regression
- XGBRegressor
- Ridge
- Lasso

K-Fold Cross Validation

```
kf = KFold(n_splits=10,shuffle=False)
model = LinearRegression()
result = model_selection.cross_val_score(model, xtrain, ytrain, cv=kf)
print("Accuracy: %.3f%% (%.3f%%)" % (result.mean()*100.0, result.std()*100.0))
```

```
Accuracy: 95.149% (12.087%)
```

RMSE Value

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	**wafahuu**			0.92921	5	6m

Your Best Entry ↑

Your submission scored 0.92921, which is an improvement of your previous score of 0.94023. Great job!

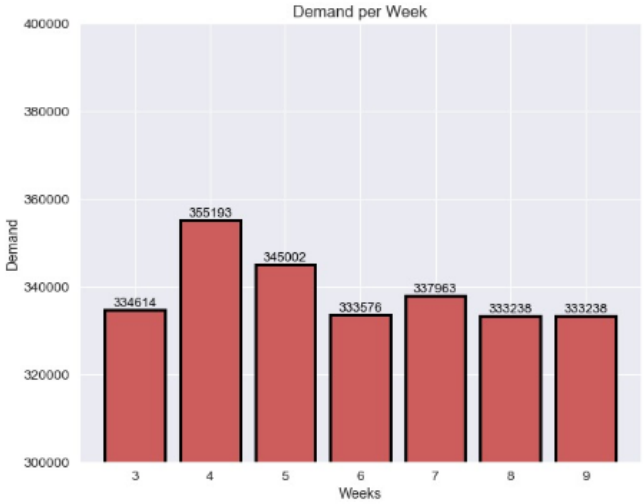
Tweet this!

Insights

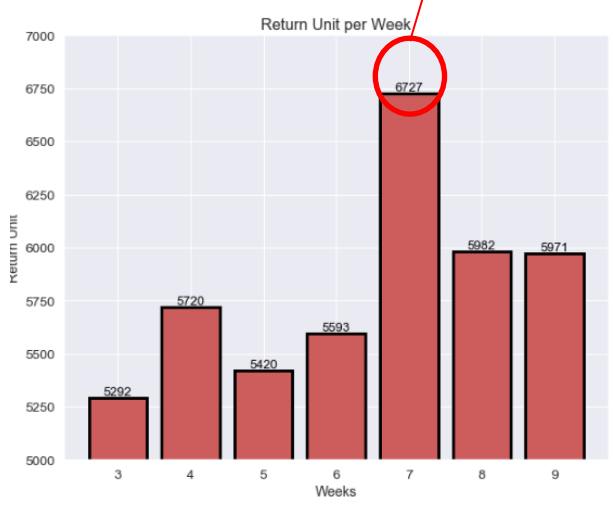
Insight

Peak:
Week 4
Sales Unit : 359,162

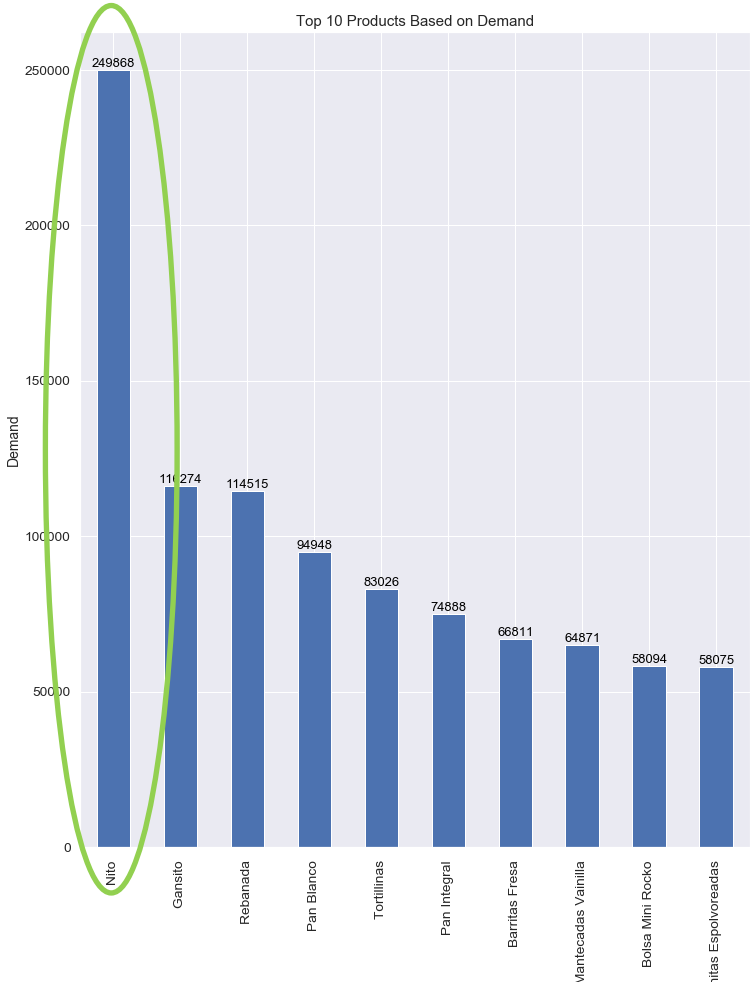
Peak:
Week 4
Demand : 355,193



Peak:
Week 7
Return : 6,727 Unit



Top 10 Products Based on Demand



Thank you.