

**B2W CHALLENGE**

Price Analysis

# Problem

In the highly competitive Market, constantly the companies have to find ways to adjust or optimize the relationship between Price and Demand.

# Goal

The main objective is to create a model to predict the quantity sold for each product given a prescribed price. Along with the statistical model, we need metrics, relationships and descriptions of these data in order to understand the sales behavior.

# Methodology

To complete this challenge, I performed the following steps, according some steps to the CRISP-DM methodology:

- Data Understanding

- Data Preparation

- Modeling

- Evaluation

# Data Understanding

The purpose of this step is to better understand the behavior of each product sold. For this, I applied several methods to find out more about imported data.

Following I will show some descriptive statistics of the analyzed products.

## Analysis of Quantity Sold by Products

The following table shows some measurements of the quantity sold by product, as a minimum, maximum, mean, median, standard deviation and count.

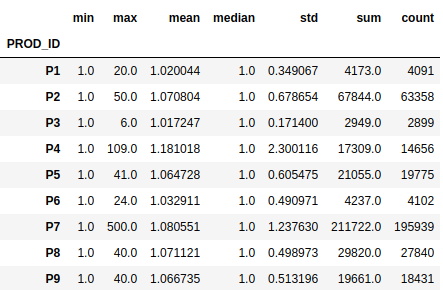


Figure 1

By looking at the table above, you can see very large values in certain measurements as an example the product P7 that has the highest quantity of products sold in relation to the others. In second is the product P4 with maximum value of 109 quantities sold. Although its maximum value is much lower than the value of the product P7, the value of its standard deviation is much higher. This means that the variation in the quantity sold is further from the mean value.

For a better understanding, I will show below the data plotted in the chart.

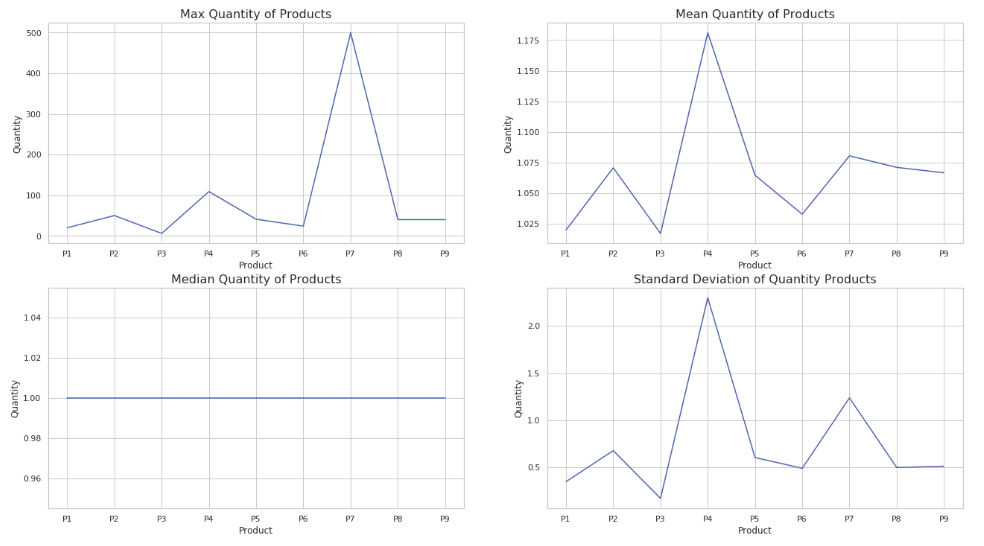


Figure 2

## Analysis of Revenue by Products

As well as in the analysis of quantity sold, I will show some measurements in the table below.

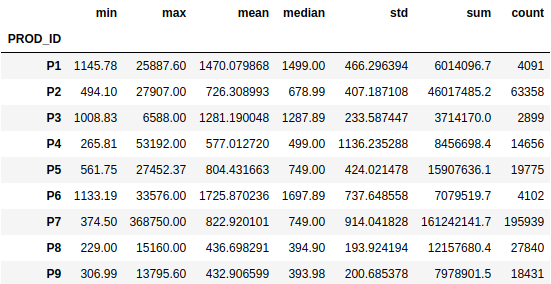
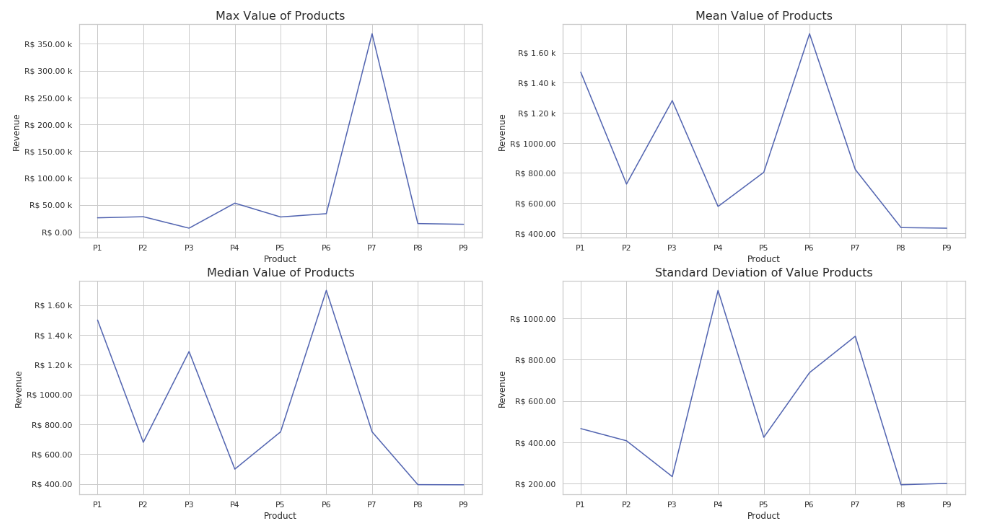


Figure 3

The value of 500 quantities sold of product P7 in an order, reflects in the data presented in the table above. The product P7 has the same behavior as in the previous table.

For a better understanding, I will show below the data plotted in the chart.



In this scenario it’s possible to observe several behaviors. The first chart (Max Value) shows us the product P7 as the top 1 in the maximum value. But on the next chart we can see a completely scenario, the mean value of product P7 is significantly lower.

The product P6 has the highest mean value in related to other products. To explain this behavior we can see the sum and count columns. You can see that it has a high revenue value and a low frequency of sales per order.

## Daily Revenue

O gráfico abaixo apresenta o total das vendas realizadas de 01/01/2015 até 31/10/2015.

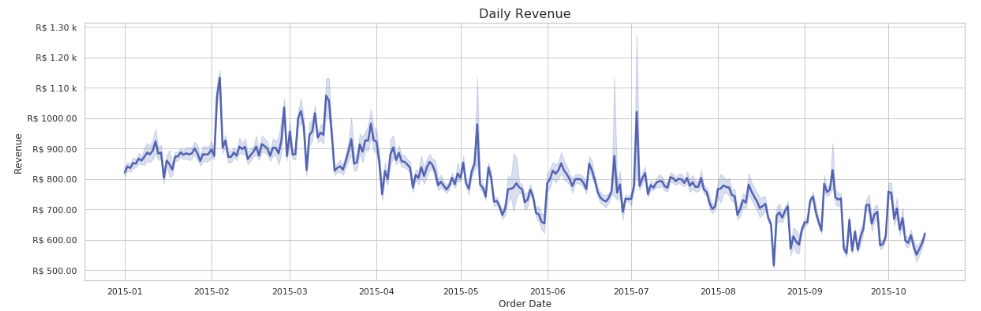


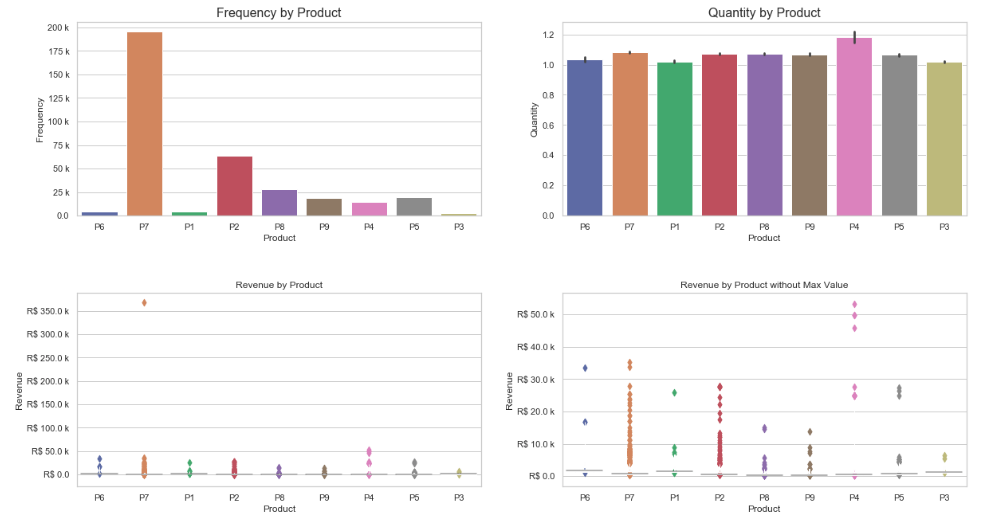
Figure 4

As vendas durante este período de 10 meses possuem diversas variações em dias específicos. Ao comparar essas variações, pode-se observar que grande parte delas são picos crescentes no seu total vendido, poucas delas são frequentes no gráfico, mesmo tendendo ao decréscimo ao longo do tempo.

Outro fator importante a ser observado são os posicionamentos de cada aumento de pico, muito deles ocorrem no início dos meses.

## Analysis of Products

O objetivo a ser alcançado no plot do dados abaixo, é entender o comportamento de cada produto, analisando a sua frequência, quantidade e receita das vendas.



**Frequency by Product**

Neste plot observamos que os produtos 7 e 2 possuem uma frequência maior em cada vendas no geral (não sendo considerado as suas respectivas quantidades vendidas).

**Quantity by Product**

Neste gráfico analisamos a quantidade total vendida de cada produto em cada venda. Apesar dos produtos 7 e 2 terem mais frequências em cada venda realizada, os dados nos indicam que o produto 4 possui uma maior quantidade vendida. Mesmo que ele não tenha tanta frequência em muitas vendas, os clientes compram este produto em maior quantidade do que os outros produtos.

**Revenue by Product**

Assim como no gráfico anterior o produto 7 continua tendo um destaque no valor da receita comparado aos outros produtos. Aqui visualizamos claramente que existe uma venda onde o valor do seu produto é muito superior aos outros, tornando assim essa diferença e a distância visualizada no gráfico.

**Revenue by Product without Max Value**

Neste plot retiramos o valor da receita significativa do produto 7 para analisarmos melhor o comportamento dos outros produtos. Ao retirarmos podemos observar em um primeiro momento que o produto 4 possui um valor de receita significativo logo após o produto 7.

Mesmo que o produto 4 esteja destacado em relação aos outros no gráfico, podemos observar que essa diferença depende da quantidade dos produtos vendidos em cada venda, pois os seus valores possuem uma distância entre eles.

Já para o produto 7 vemos uma concentração muito próxima dos pontos, deixando a reta dos seus valores mais homogêneos do que o produto 4, se não tivermos uma quantidade significativa para o produto, o seu valor de receite tende a diminuir.

O produto 2 também apresenta o mesmo comportamento do produto 7 sem o outlier, uma reta onde os seus pontos (valores) são mais próximos uns dos outros, tendo assim uma maior conscistência.

# Data Preparation

Esta estapa é responsável pela preparação dos dados, onde o dataframe será totalmente estruturado e projetado para que possam ser melhor aproveitado no modelo.

Os dados apresentados abaixo, são referentes ao produto P1, porém todos seguem a mesma estrutura.

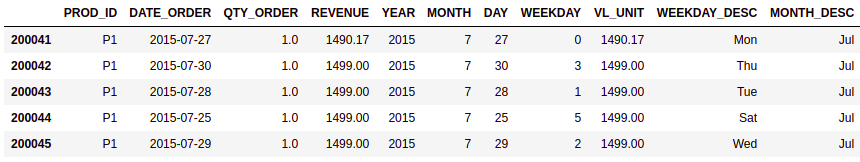


Figure 5

Como os modelos não são realizam muito bem a predição com dados categóricos, foi necessária a transformação dos dados. Na Figure 6 é mostrado o resultado da transformação dos dados categóricos para dummy.

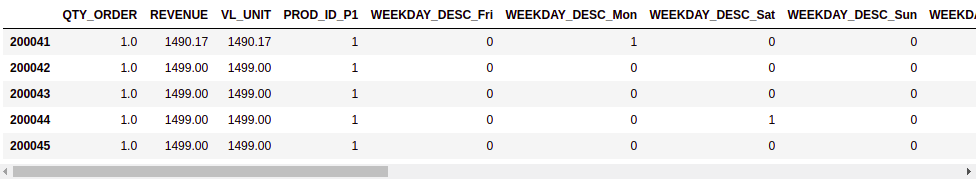


Figure 6

Após a transformação dos dados para cada produto, a análise da correlação dessas features em relação a variável respsota **QTY\_ORDER** faz-se necessário para saber quais delas podem ser utilizadas na modelagem do modelo. Segue abaixo o gráfico de heatmap dos produto P1.

**P1**

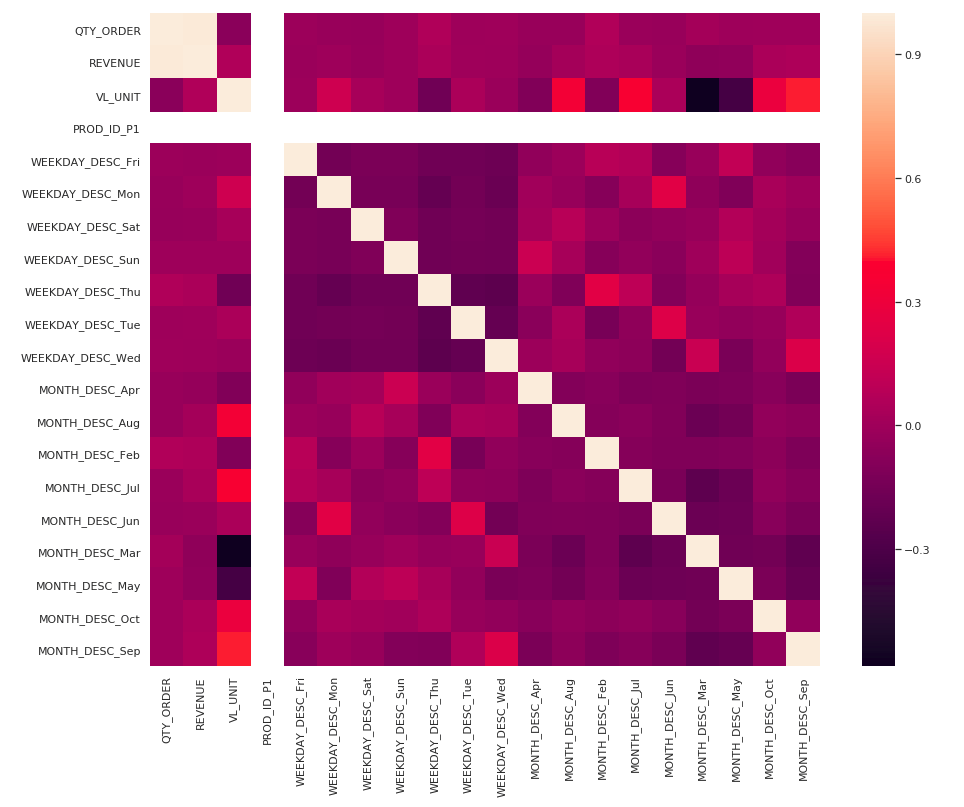


Figure 7 – Heatmap P1

Este gráfico está indicando que poucas features possuem correlação com a variável resposta **QTY\_ORDER**, muitas variáveis estão entre 0.0 e 0.5.

**P2**

Assim como no produto P1, a estrutura do dataframe do produto P2 segue o mesmo padrão do P1. O que difere neste momento são as correlações entre as variáveis. Podemos analisar o heatmap plotado abaixo.

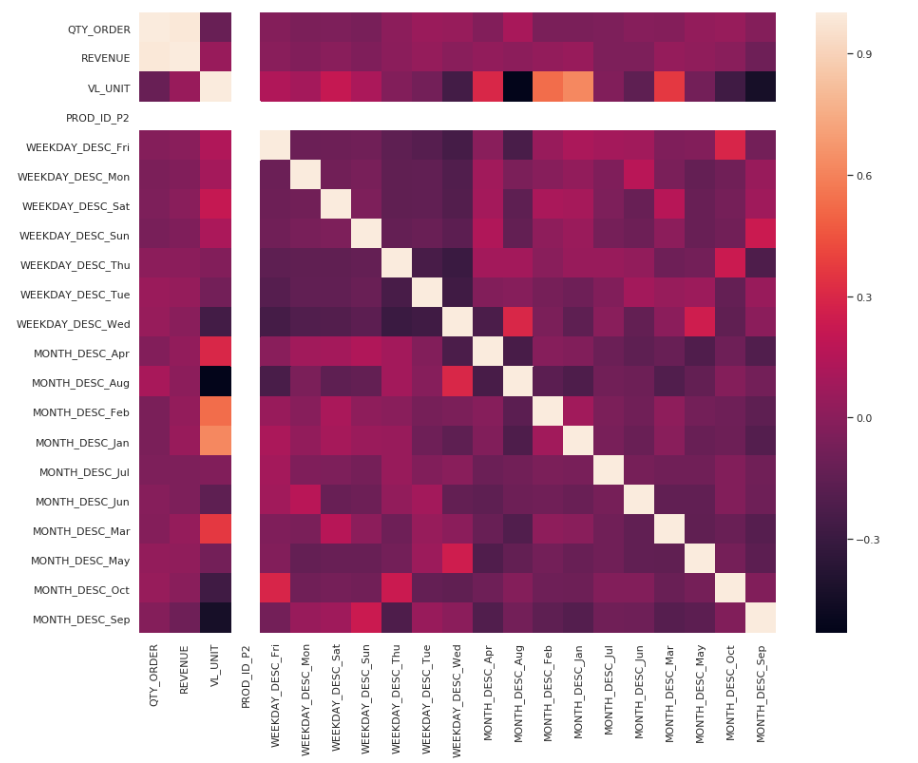


Figure 8 – Heatmap P2

Somente olhando para o gráfico, é possível analisar um comportamento semelhante ao gráfico do produto P1, porém no produto P2 existem mais áreas escuras indicando que existem mais grupos abaixo de -0.3.

**P3**

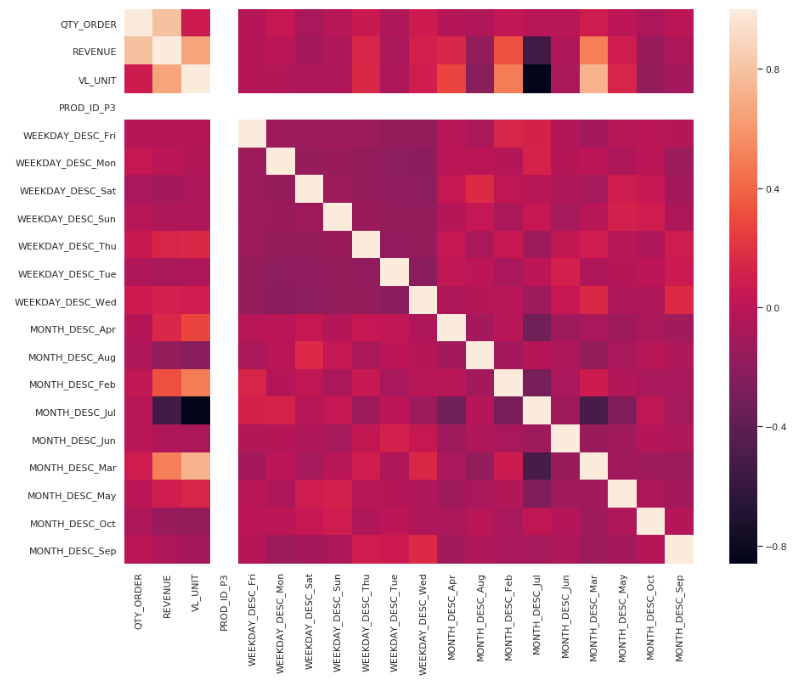


Figure 9 – Heatmap P3

**P4**

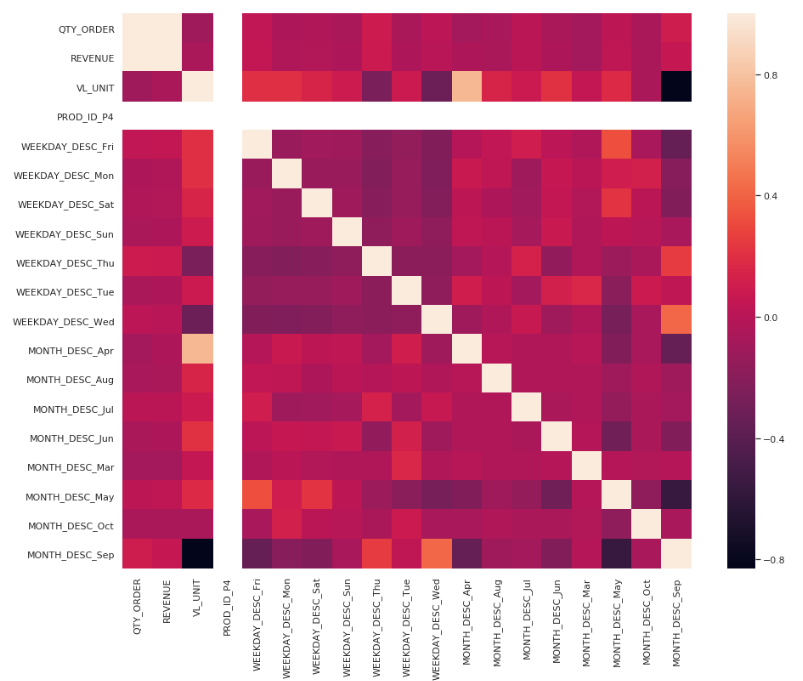


Figure 10 – Heatmap P4

**P5**

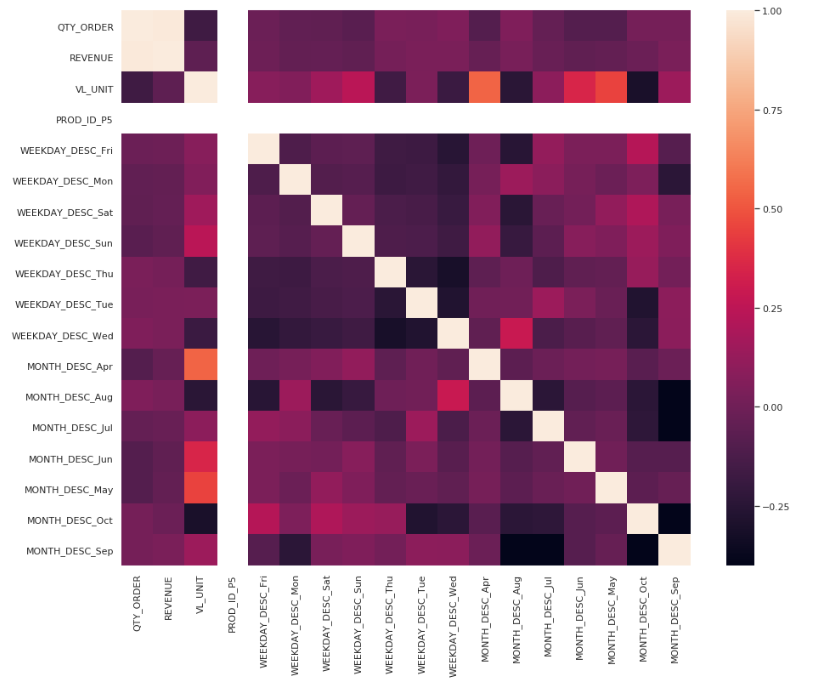


Figure 11 – Heatmap P5

**P6**

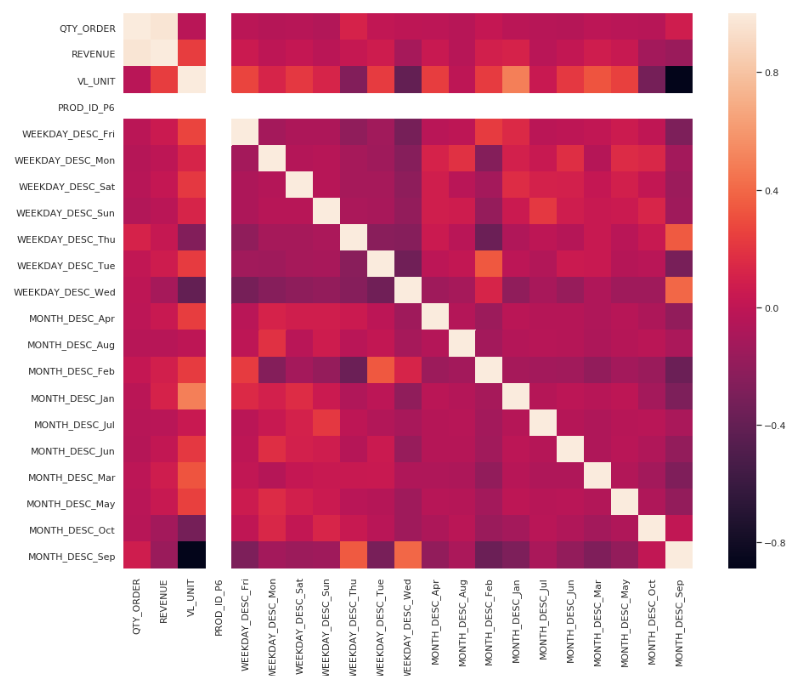


Figure 12 – Heatmap P6

**P7**

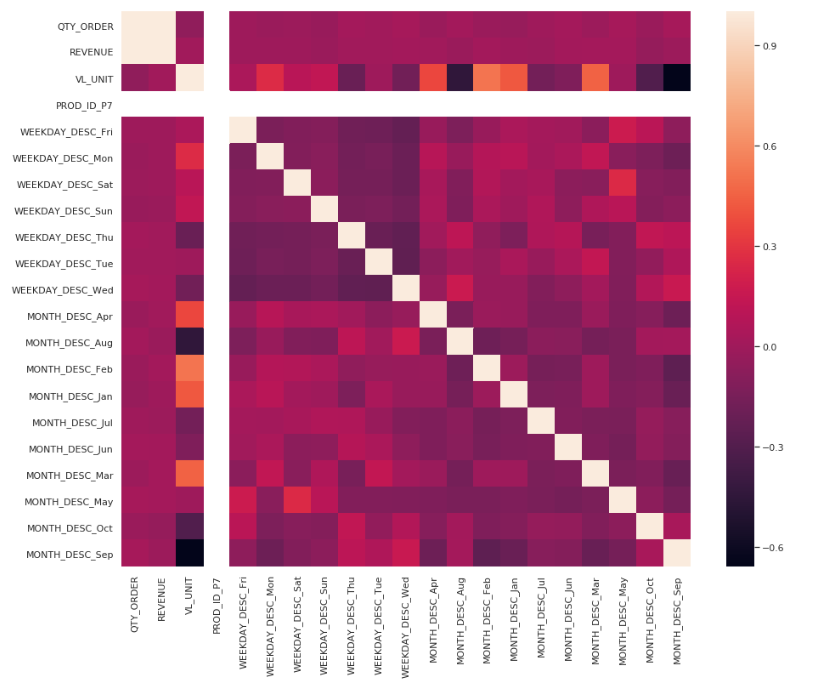


Figure 13 – Heatmap P7

**P8**

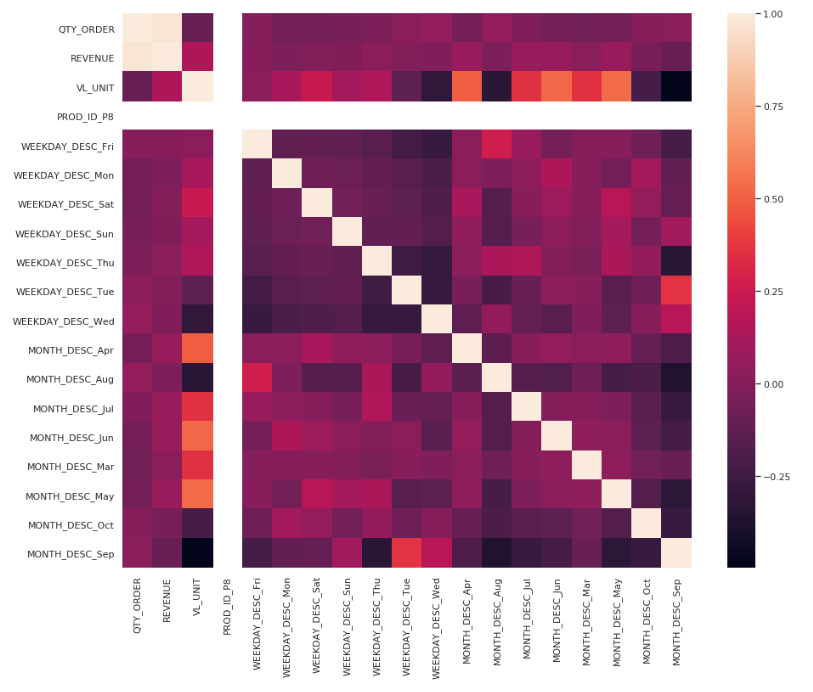


Figure 14 – Heatmap P8

P9

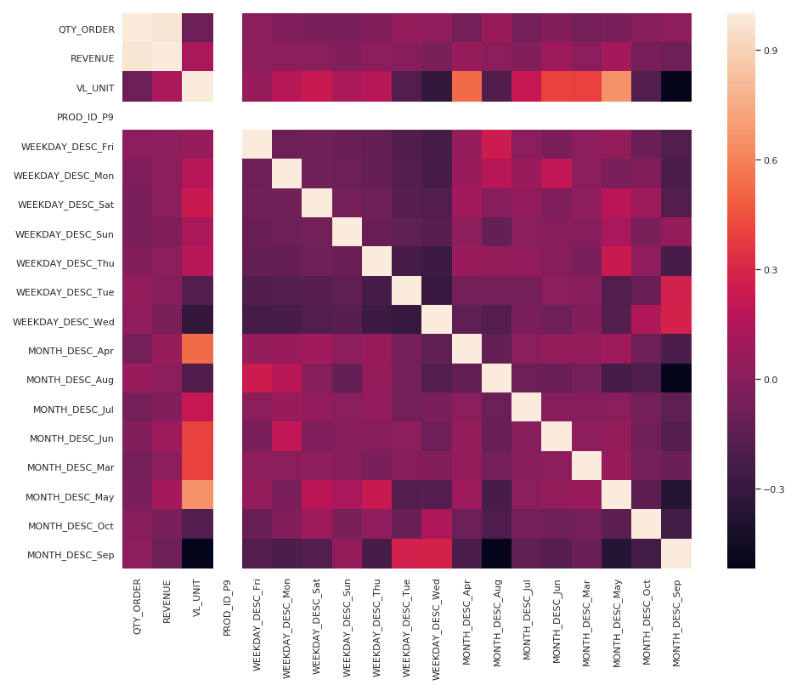


Figure 15 – Heatmap P9

# Modeling

P1

P2

P3

P4

P5

P6

P7

P8

P9

–