Final Project – Data Science

Study of dataset “Women Clothing Ecommerce Sales Data”

EDA - Time Series Analysis

Name: Rafael Uréndez Valderrama

Center: IT Academy (Barcelona Activa)

Date: March 2023

1. Presentation of the chosen dataset

For this project I've chosen the dataset "Women Clothing Ecommerce Sales Data", that has been extracted from the website https://www.kaggle.com/.

The main reason I have chosen this dataset is to explore a new topic that has not been covered during the data science course, **Time Series**. And within time series, some typology that is related to the current job on retailing.

This type of analysis of temporal structures is very important for making product purchase forecasts, minimizing stock as much as possible, resizing the company's structure according to future projections, etc. In conclusion, advise business owners on making decisions based on data.

On the other hand, there are times when the data we have is not sufficient, and we must find solutions to expand this data and reach more relevant conclusions for the company.

2. General characteristics

The dataset has been shared on Kaggle by the owner of a women's clothing e-commerce platform, with the aim of allowing the Data Science community to work with it.

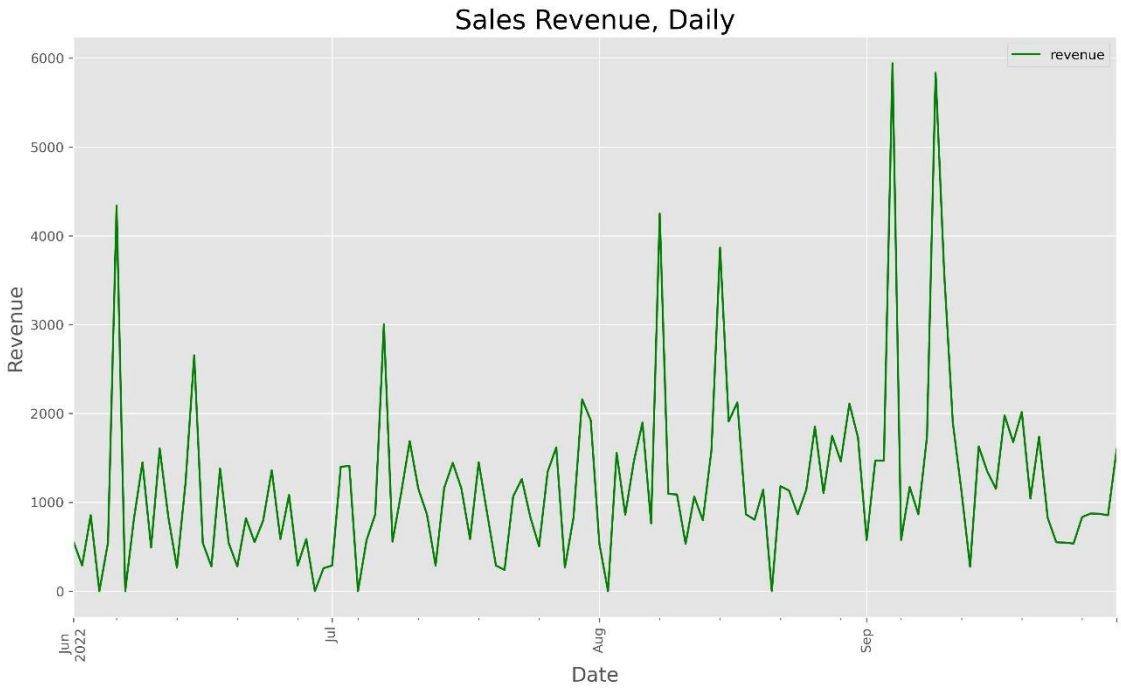
With the rise of data science, more and more companies are using their data to analyze the performance of a brand, a product launch, user behavior in online purchases among other things. The structure of the stored data allows for the creation of different mathematical models and the extraction of conclusions. In particular, online commerce has a great advantage in being able to generate and store a large amount of data to work with.

In this case, online clothing sales, as compared to in-store sales, allow for significant savings in logistics, and there are many companies that are increasingly betting on online commerce at the expense of physical store sales.

The data provided by the owner is limited. The period only covers 4 months, from June to September 2022. It contains only one csv file with the data for each issued ticket or invoice. We are not given any additional information that would allow us to work towards a more robust model. We also do not know the target audience or the country in which this data is based.

The data is related to women's clothing, with reference indicators, size, color, and unit cost.

This is the graphical representation of the revenue:



3. Definition of variables

The dataset has 8 columns and 527 rows, this is the structure of the file:

```
RangeIndex: 527 entries, 0 to 526
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   order_id    527 non-null    int64
 1   order_date  527 non-null    object
 2   sku         527 non-null    object
 3   color       527 non-null    object
 4   size        490 non-null    object
 5   unit_price  527 non-null    int64
 6   quantity    527 non-null    int64
 7   revenue     527 non-null    int64
dtypes: int64(4), object(4)
```

From the variables provided, we can see that we have order_id as the index. The variable order_data is imported as an object, representing time; the variables unit_price, quantity, and revenue are numeric. The variables SKU, color, and size are categorical variables.

The variable order_date, which is currently of object type, should be changed to a time format and made the index of our database. To study the revenue variable, as we have 527 records, we need to resample the data and sum the revenue day-values to get the total daily revenue for the 4 months (122 days/rows).

Here is the link to the dataset, which has been sourced from the Kaggle platform: https://www.kaggle.com/datasets/shilongzhuang/-women-clothing-ecommerce-sales-data

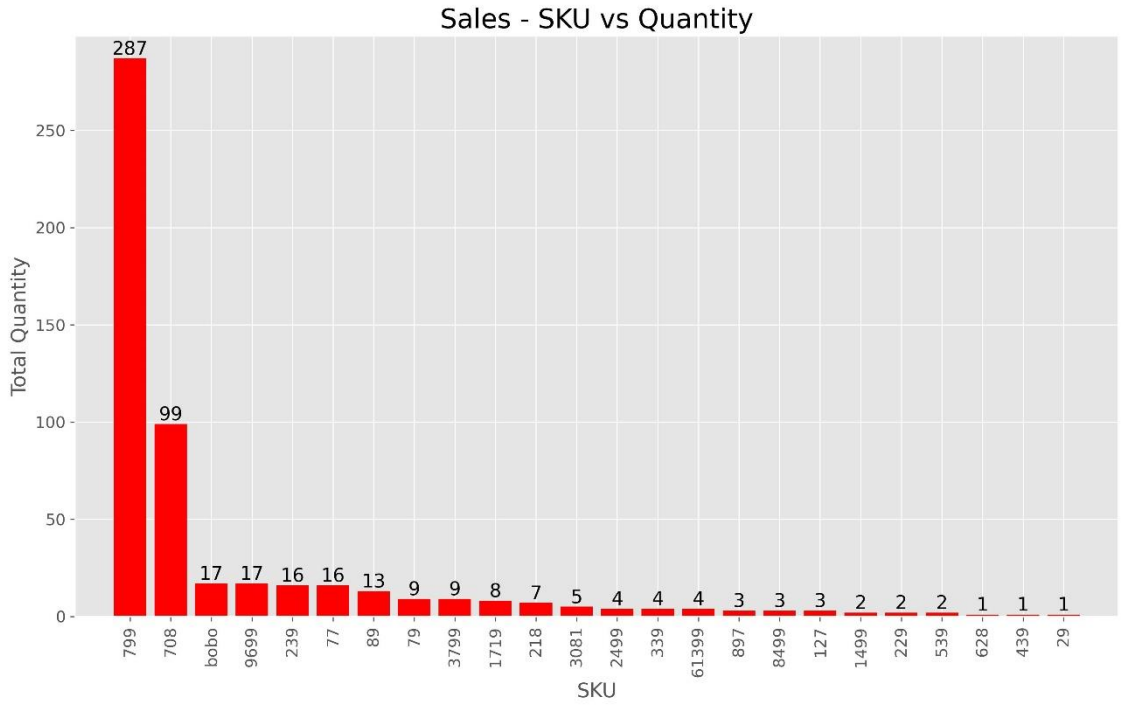
4. Presentation of objectives

As mentioned earlier, the idea is to communicate, through the data, to the company's decision-makers, actions to consider in order to identify opportunities for improvement within the company.

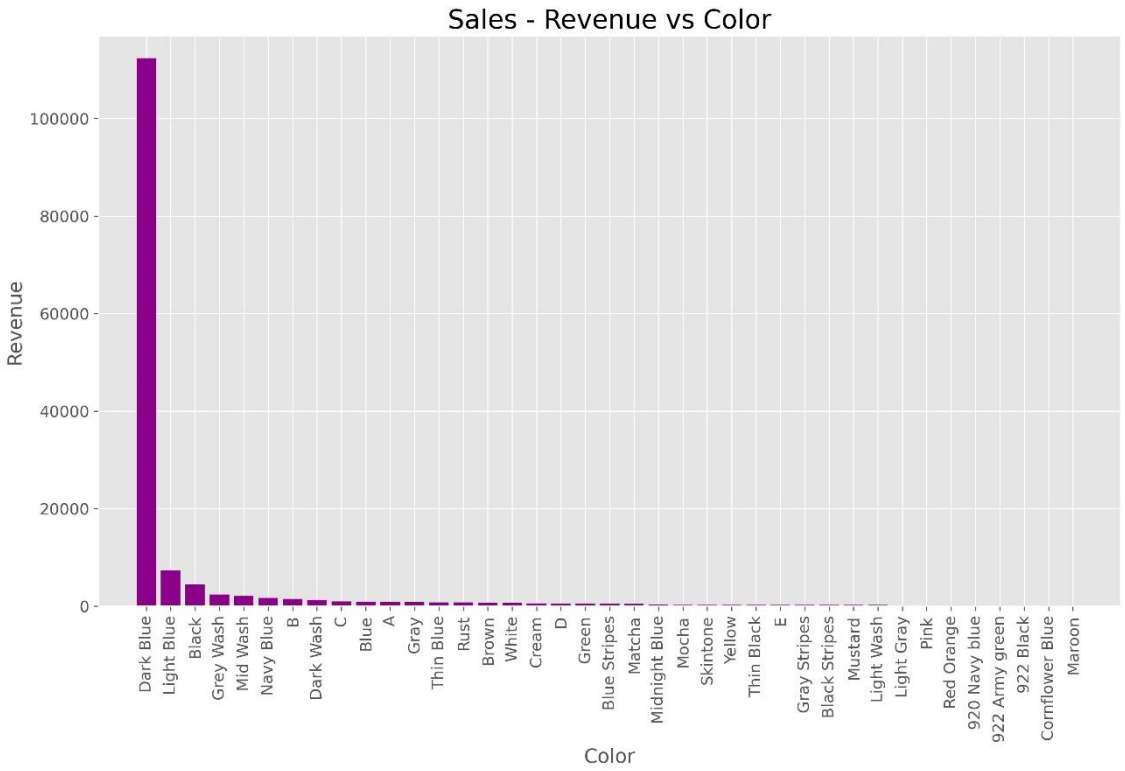
There are two primary objectives. First, using exploratory data analysis (EDA), provide insights into the sales of SKUs, including aspects such as size, color, etc. Second, develop a model to accurately forecast the revenue for the upcoming month.

4.1. EDA analysis info

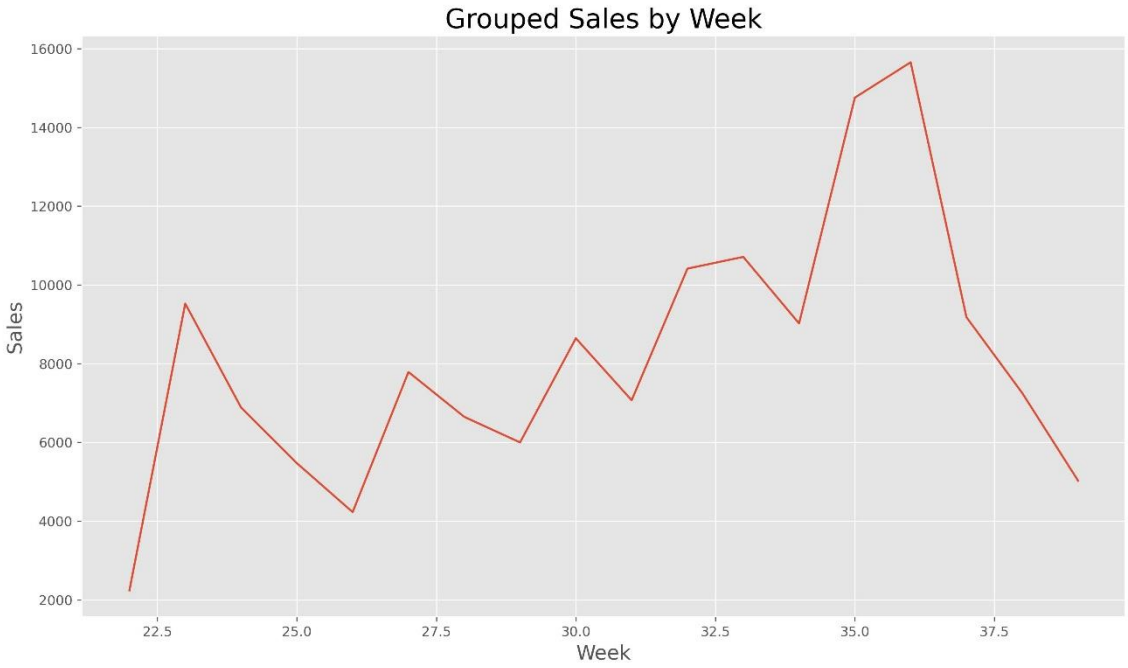
Based on sales, among other questions, show showcase the best and worst-selling SKU items.:



View sales depending on color sold:

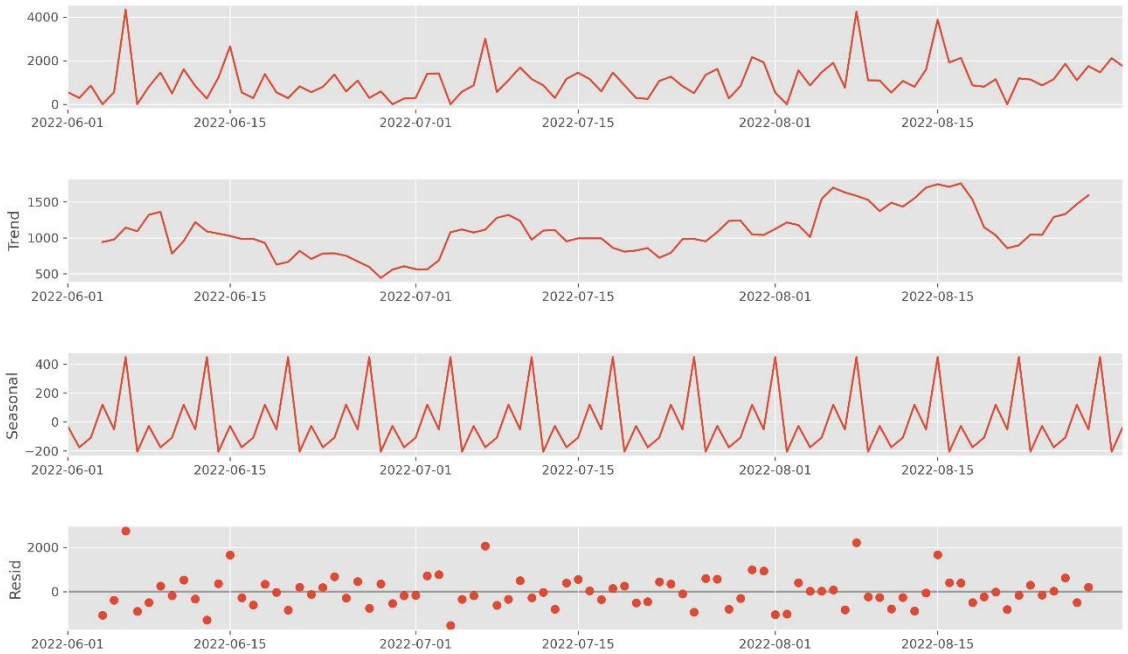


Showing with days are better for selling, and better planning email marketing actions:

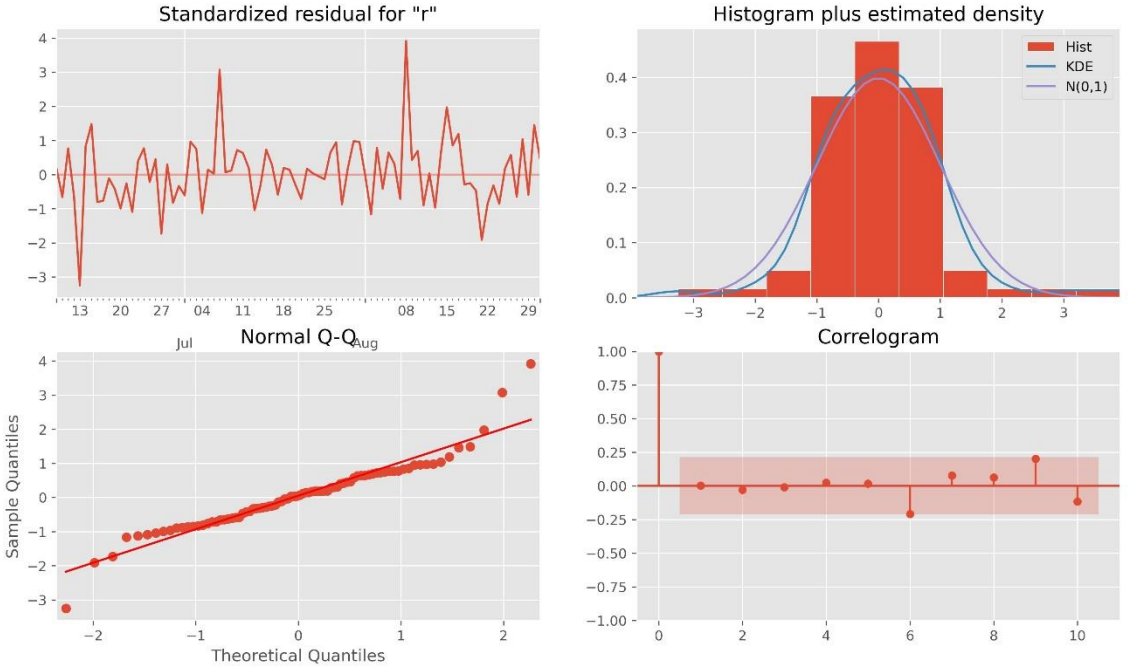


4.2. Forecasting (using SARIMAX modeling)

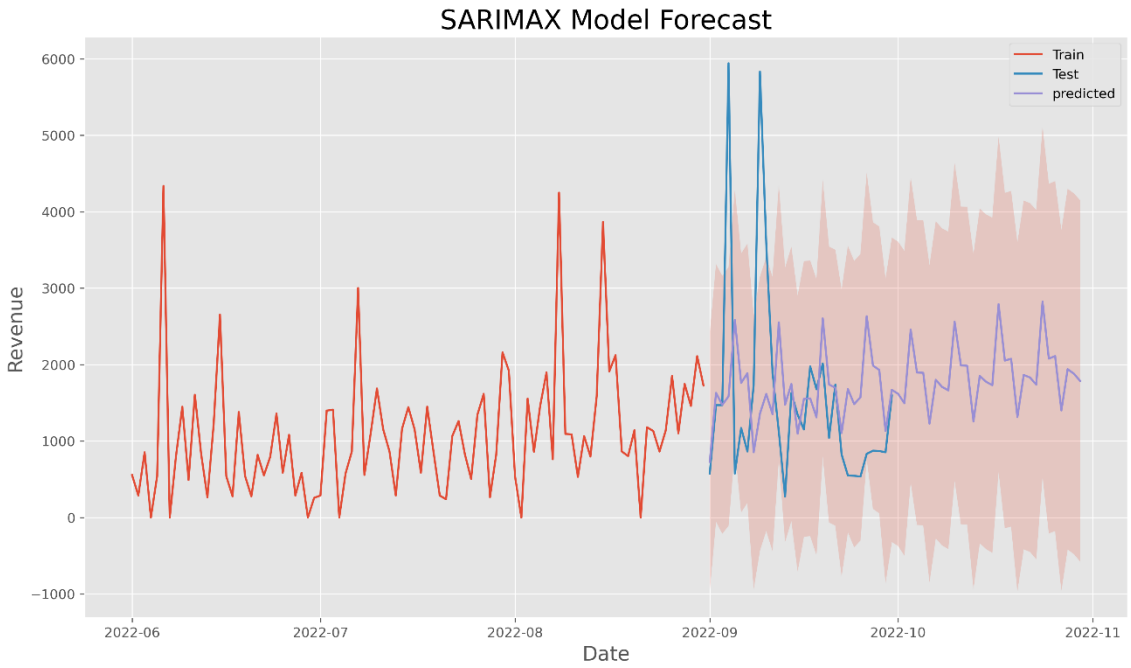
Using the SARIMAX modeling, study the data in order to obtain a good model if possible. This includes the seasonal decomposition of the data, where we can examine all the components, such as trend, seasonal, and residual:



On checking data is important once obtained the order of the model check the plot diagnostic:



And finally get the forecast:



Discuss the outcomes achieved after fine-tuning the parameters and refining the model.

In this instance, it appears that there is no significant seasonal component, and the residual component carries substantial weight in determining the final results.

The results obtained with the SARIMAX forecasting model have not been as good as expected. Even with outlier detection, the model has not improved. In the absence of more data from the e-commerce owner, more time should be invested in considering other prediction methods.