****

**Big Data Analytics**

**CSCI 4341, Sem 1-2021/2022**

**Lecturer: DR. SHAHRYAR WANI**

**Prepared by:**

|  |  |
| --- | --- |
| **MD NAJMUL HUDA** | **1627521** |

**Due Date: 23.01.2022**

**Project Title:**

Data analysis and visualization for Brazilian e-commerce and sales forecasting using Random Forest Regression model.

**Abstract:**

The expansion of e-commerce and digital marketing in recent years has resulted in a massive amount of opinionated data. Enterprises would get insight into the data by analyzing it, allowing them to make better business decisions. The goal of this project is to examine public statistics in order to gain a better understanding of the e-commerce industry. We'll be working on the Brazilian E-Commerce Public Dataset by Olist, which we found on Kaggle. We'll work on issues such as determining the most popular products in the store in order to develop a better business model, the most popular categories in terms of sales, seller and customer frequency by state, trending of a growing e-commerce business, the correlation of delivery days and customer satisfaction, and predicting the sum of the sales value for the coming weeks. This study might be used to do sentiment analysis on commodity appraisals of purchased items on e-commerce sites.

**Introduction:**

E-businesses can no longer separate themselves from their online competition by just selling things online. Companies are looking for new strategies to outperform their competition now that many more have an internet presence. This research is primarily aimed at analysing one E-commerce dataset (Brazilian E-Commerce Public Dataset by Olist), in which product values and popularity will be assessed by comparing sales history. According to the study, certain actions to do will be suggested in order to boost the company's popularity by projecting next year's sales conditions using time series analysis in terms of dataset history such as product status, most sold product history, and so on. On the other hand, understanding consumer feedback is critical to a company's success. Analyzing the reviews aids in determining the customer's various tastes, likes, and dislikes, among other things. The information gathered may then be utilised to improve customer service and satisfaction. Many studies have been undertaken utilising public datasets, but real-time datasets have yet to provide answers to many business concerns. The goal of this study is to gain a better understanding of the e-commerce area by examining public datasets. By comparing the product status history thus far, an overall depth analysis will be performed. The project's major goal is to build a solution that will help the company increase its popularity. As a result, this study will provide a method for them to boost their sales.

**Research Questions:**

* Which cities has the large number of customers?
* Which category has the most popularity in terms of sale?
* Is ecommerce business in Brazil Is growing day by day?
* Is delivery time most responsible for good scores?
* How happy customers are with their products, services, and their overall experience?
* What would be the sum of the sales value for the coming weeks?

**Research Objective:**

* To find the most sold products for developing a better business model.
* To find the most popular category in terms of sale.
* To analysis the growth rate of sale in different state.
* To analyze time series data with a regression model for forecasting next week sales condition.

**Research Significance:**

This study may be used to do sentiment analysis on commodity evaluations of purchased items on e-commerce platforms. Analyzing the sentiment trend of consumer evaluations can assist firms on e-commerce platforms enhance service quality and customer happiness by providing a reference for other consumers. We investigated the impact of associated aspects on model performance, such as the size of the thesaurus, the length of the input phrase, and the number of model iterations, and performed experiments to improve our model.

**Literature Review:**

Sentiment analysis-based Fuzzy decision support model for e-commerce item comparison. To describe online reviews, the model uses probability multivalued neutrosophic linguistic numbers (PMVNLNs). By considering neutral information and reluctance in text reviews, it overcomes the limitations of previous models. The suggested characterization technique, unlike existing item rating algorithms, incorporates positive, neutral, and negative information in text evaluations and reflects review reticence. Because of the characteristics of item comparison difficulties in e-commerce, the model proposes QUALIFLEX. Using the integrated regret theory-QUALIFLEX, it addresses customers' constrained rational behavior, unlike previous models. [1]

Consumer behavior and expectations vary across geography as multiple factors affect the purchase, re-purchase or return of a product, ranging from the product features, inventory, logistics and customer support [2, 3, 4]. In Brazil, the e-commerce market is challenging due to customer uncertainty in the security of payments, fulfilment of deliveries, and high cross border taxes, all providing the advantage to the local retailers [5, 6, 7]. Many studies found that RF performs the best amongst other algorithms. Gender classification on micro-blogging sites were studied by classifying the emoticons, textual information using natural language processing, and emotional punctuations [8]. In this scenario, the RF outperforms NV, AdaBoost (AB) and Support Vector Machine (SVM) with the highest F1-score of the prediction. Classification algorithms were used to predict the shopping platform which users will use the next time they make a purchase by analyzing temporal, user profile, demographics and loyalty features using RF, NV, SVM, and Long Short-Term Memory Network (LSTM) [9]. Again, RF obtained the highest accuracy, precision, recall, and F1-score. A study of repeat buyer prediction to identify buyers with the potential to purchase more products was carried using GB, RF, and XGBoost using transaction data, transaction history and sample promotion information [10].

**Methodology (Estimated Plan):**

The analysis of e-commerce might be narrated with certain principles and materials as follows:

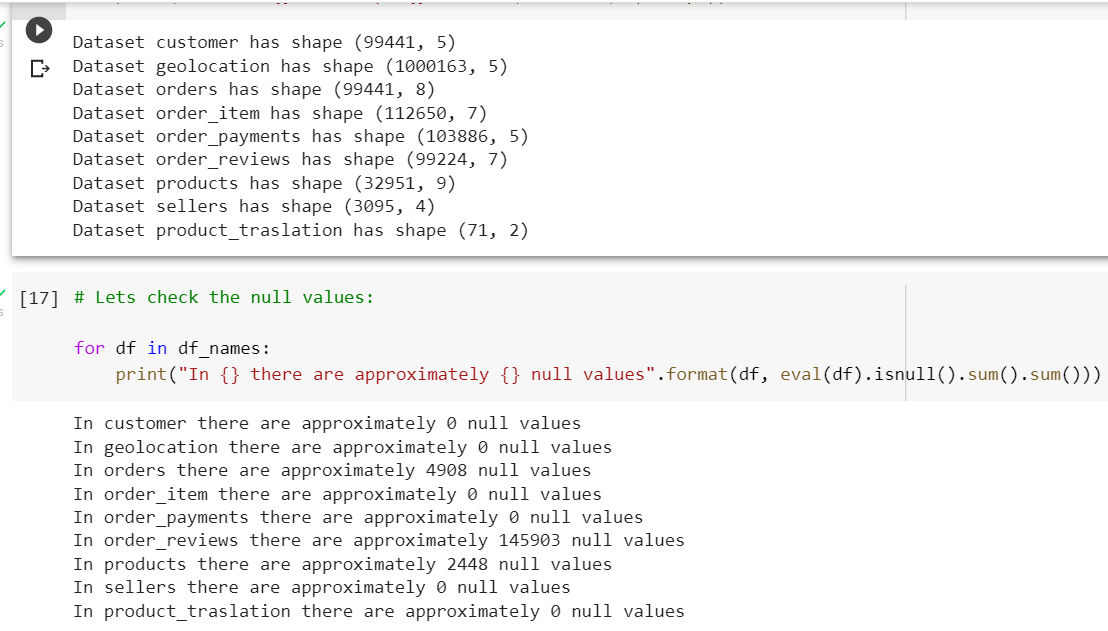
* **Data sets:** The data sets will be used from the website of Kaggle. We will use Pakistan's Largest E-Commerce dataset:

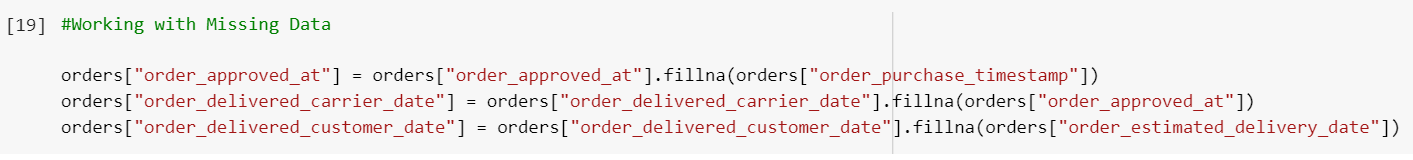
<https://www.kaggle.com/olistbr/brazilian-ecommerce>

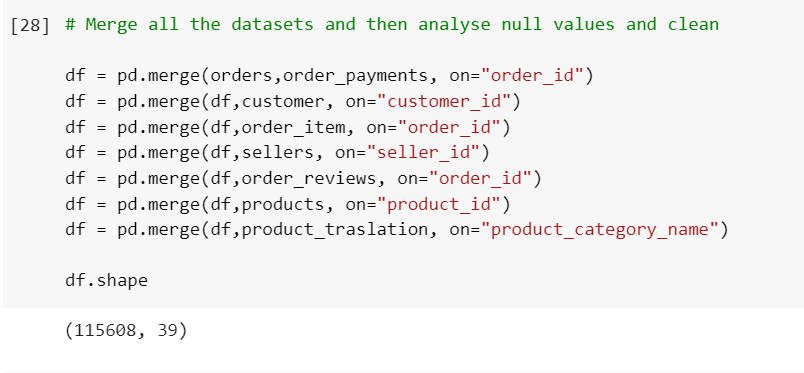
* The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.
* **Tools:** There are some mandatory tools that are necessary in order to do the implementation. The tools that are going to be used for our project are: Jupyter Notebook, Google Colab.
* **Libraries:** In this project, there are some important mentions of python libraries which are as follows: pandas, numpy, matplotlib and other plot functions in order to visualize data.
* **Algorithm:** In this model Random Forest Regression machine algorithms will be applied. It is a supervised learning algorithm which uses the approach of ensemble learning for classification and regression.

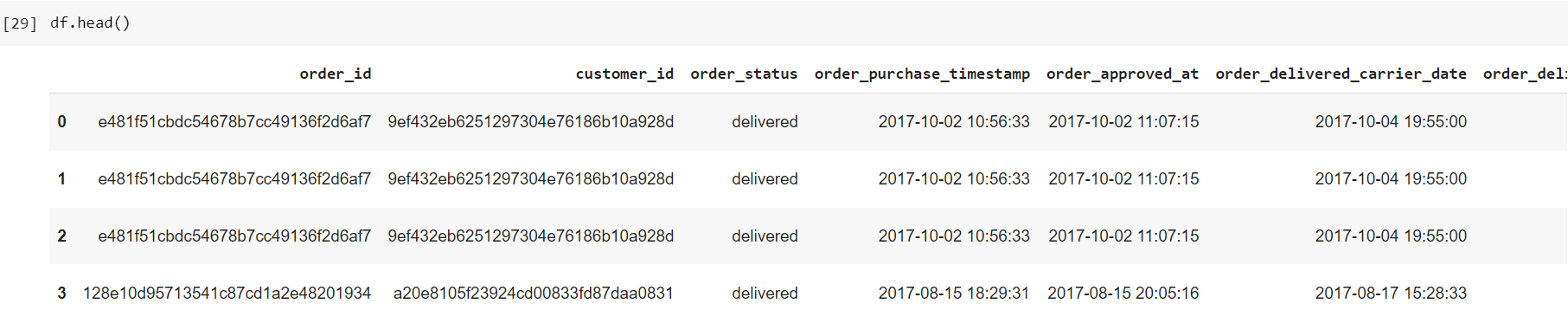
**Data Cleaning:**

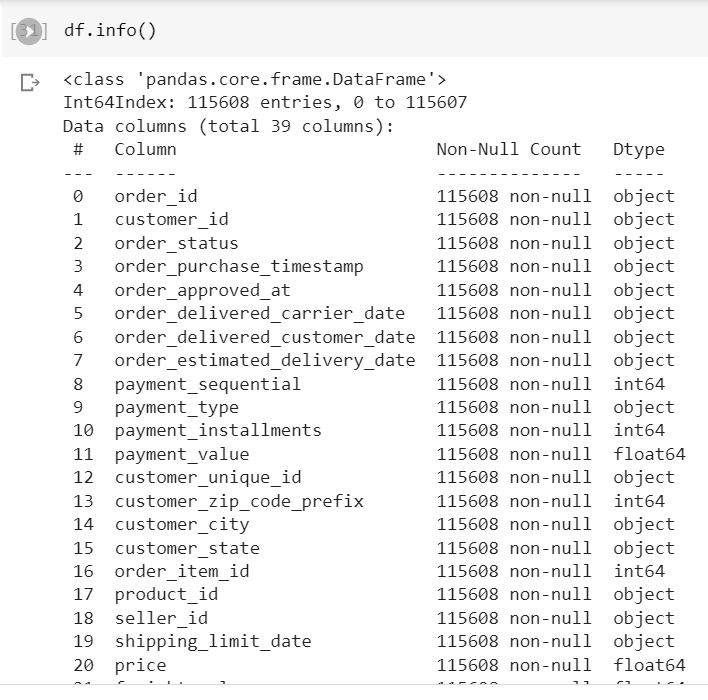
The datasets to be utilized are first reviewed and, if required, cleansed. Several data cleaning procedures were used to guarantee that the datasets were suitable for testing. To begin, the dataset is loaded into Google collab directly from Kaggle. For our project. Raw data will have missing values, columns containing information that we do not want, and some of them may be duplicated. I tried to eliminate the rows with missing values, So, I used the python command. isnull().sum() to see the null values than I remove missing values from the dataset and marge all file into one. I only want the completed version of the dataset with no missing values since we want to achieve an accurate result.

****

****

****

****

****

**Result Analysis:**

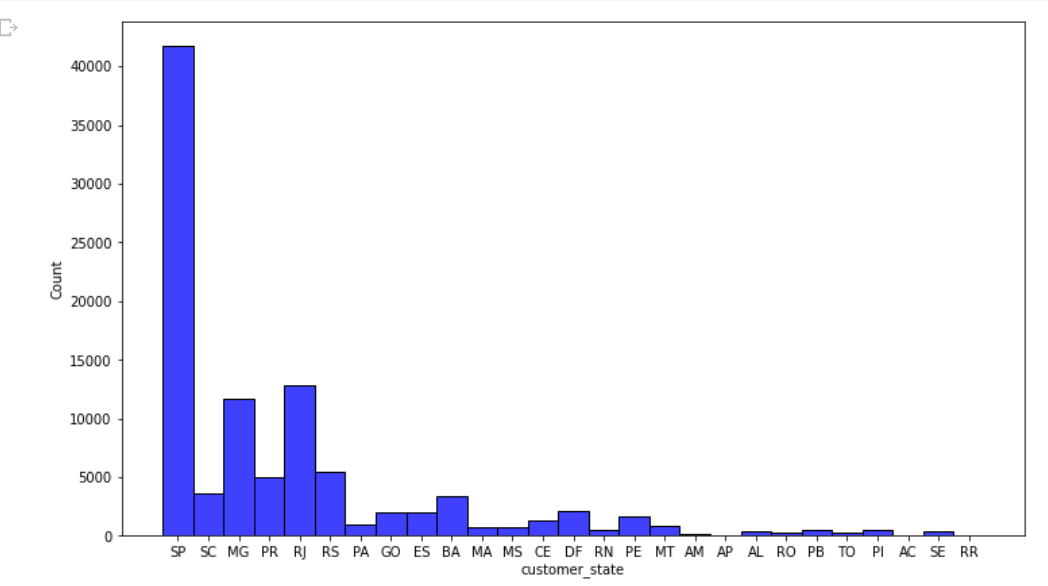
****

Figure 1 Customer Frequency by State

From above histogram, we can see that most numbers of customer are from SP (Sao Paul) state,followed by Rj (Rio de Janeiro) state.

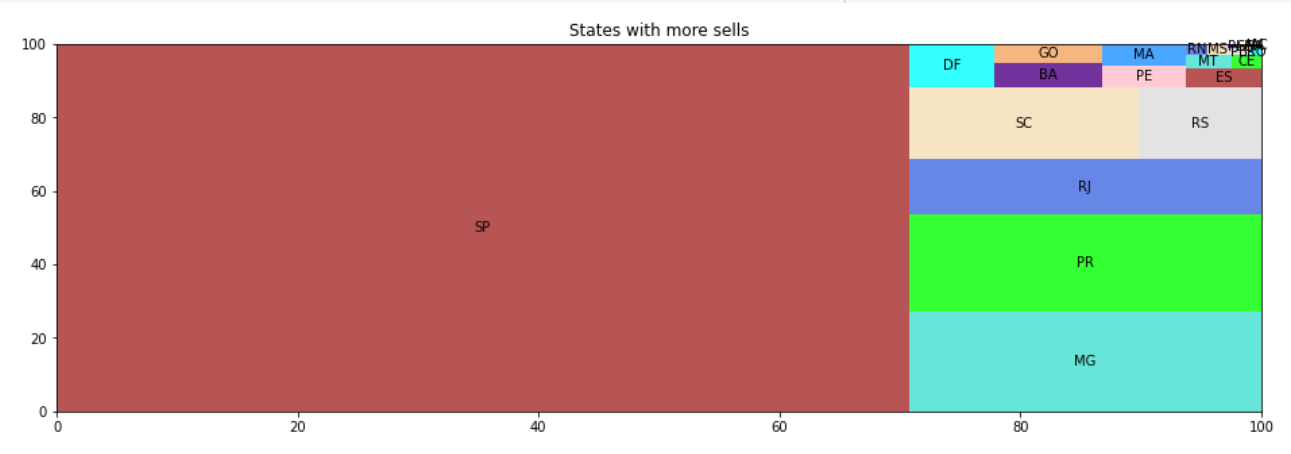
****

Figure 2 Seller Frequency by State

we can see that most numbers of seller are also from SP (Sao Paul) state. This is an expected behavior since in the Southeast region we have a much higher number of customers.

Which cities has the large number of customers?

**Chart

Description automatically generated**

Figure 3 Cities with most Customer

We see that in sao paulo, rio de janeiro and bele city have the highest number of customers.

Which category has the most popularity in terms of sale?

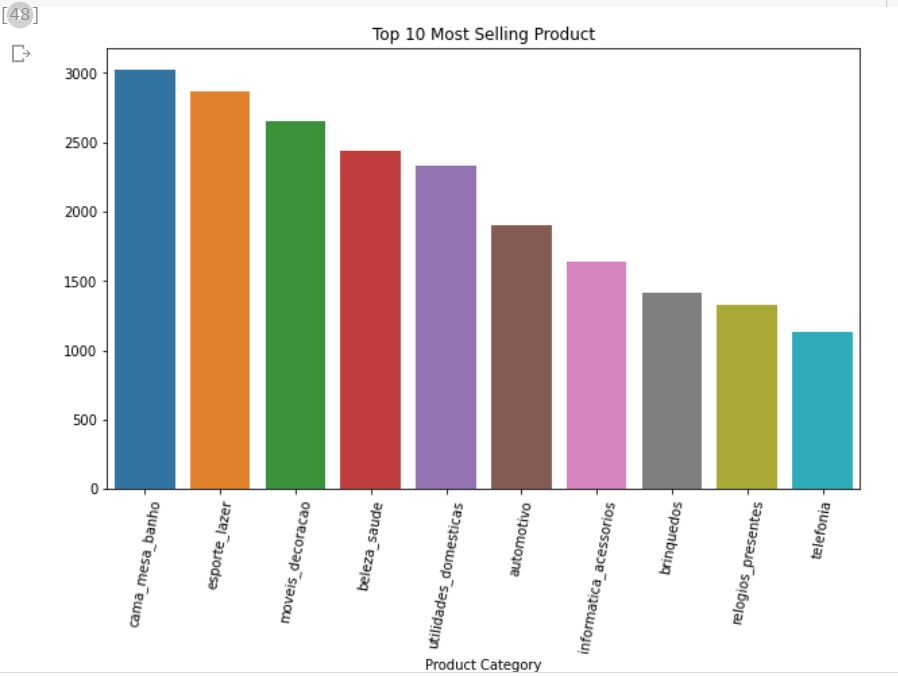
****

Figure 4 Most Selling Product

The bar graph shows that most selling product are cama\_mesa\_banho followed by esporte\_lazer , moveis\_decoracao and beleza\_saude.

Is ecommerce business in Brazil Is growing day by day?

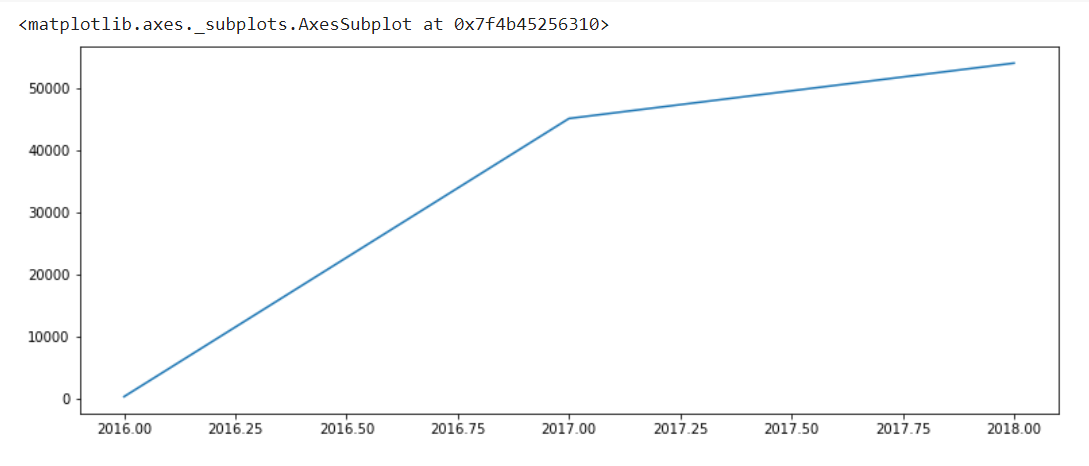
****

Figure 5 Growing rate of E-commerce business In Brazil

From the plot its clearly shows that E-commerce business In Brazil is growing rapidly.

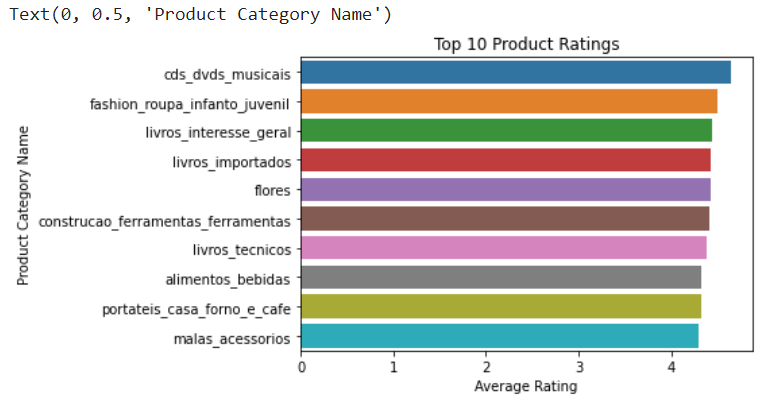
****

Figure 6 Category with highest average ratings

Music, dvd, and cds category have the highest average ratings. After that infant's fashion clothes come.

Is delivery time most responsible for good scores?

****

Figure 7 Correlation between days elapsed for delivery and days late/anticipated and reviews.

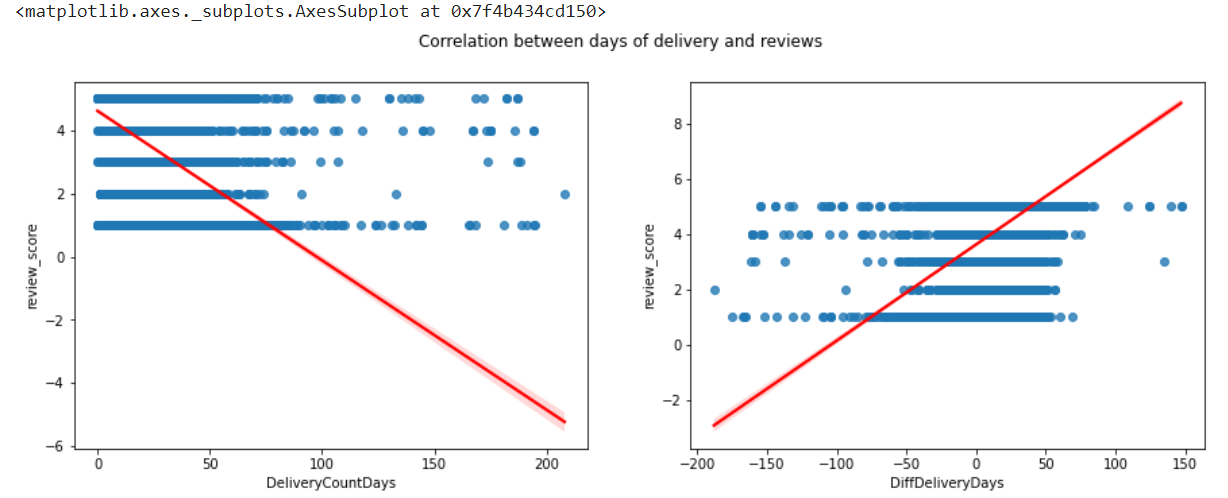
****

Figure 8 Correlation between days elapsed for delivery and days late/anticipated and reviews.

Correlation between days elapsed for delivery and days late/anticipated and reviews. Clearly, we notice that earlier and faster deliveries tend to receive better reviews and, on the contrary, delayed or late deliveries tend to receive worse ratings.

How happy customers are with their products, services, and their overall experience?

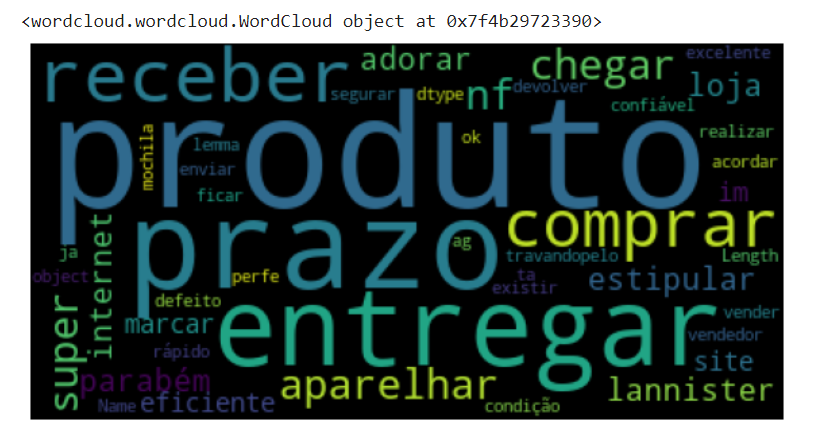


Figure 9 customer satisfaction Visualization

From review messages and the review score we can see what the customers are liking and disliking .In the positive evaluations, we noticed a tendency to praise delivery before the deadline, product quality and packaging**.**

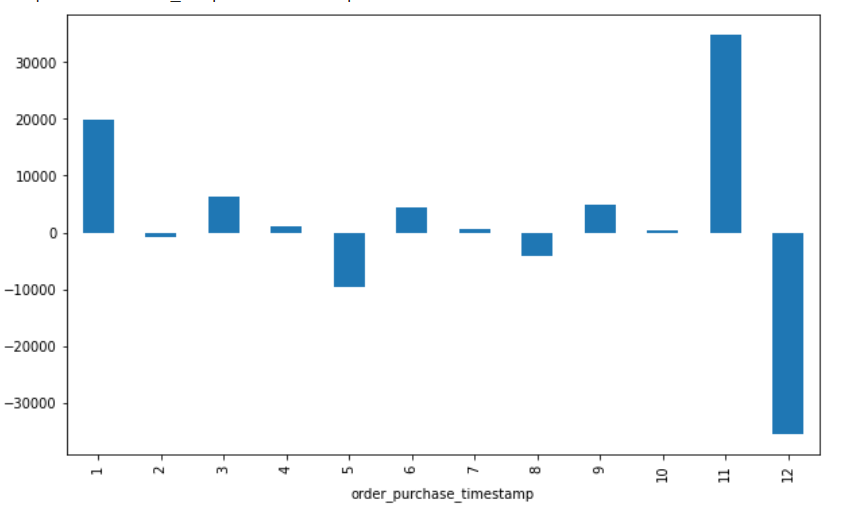
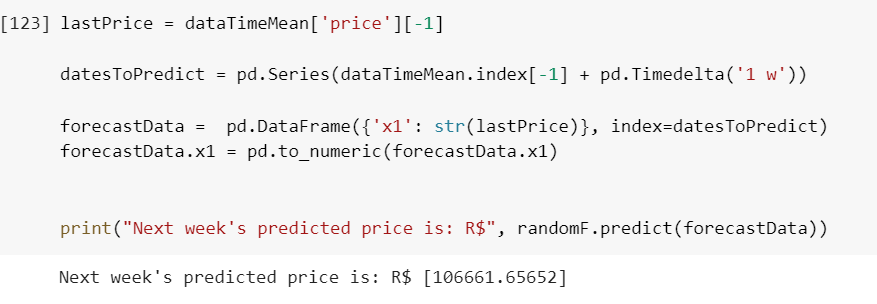
****

Figure 10 best and less selling month

Decomposing the data, we noticed a strong downward trend in values according to seasonality, at the end of each year, in December, probably after an up caused by the Black Friday.

What would be the sum of the sales value for the coming weeks?

****

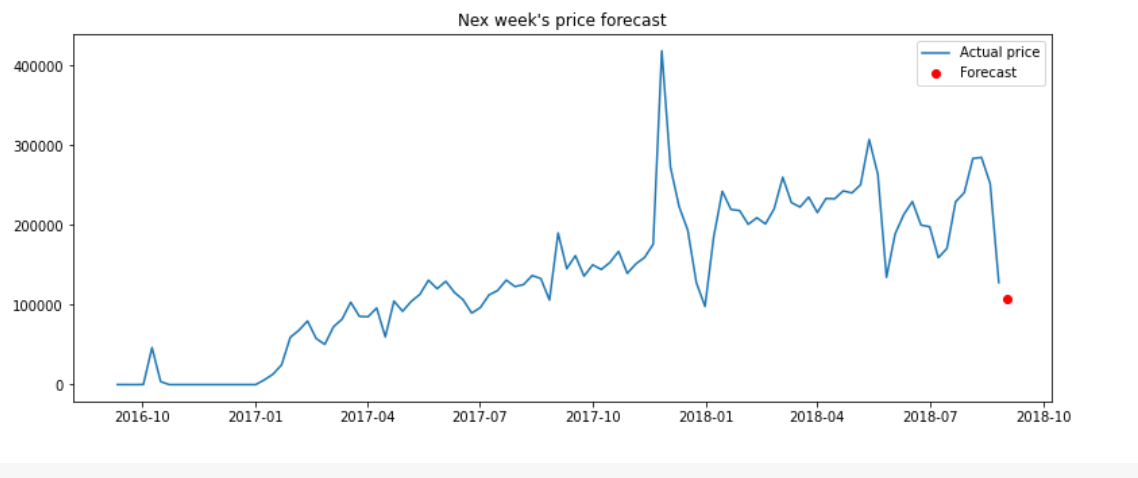
****

Figure 11 Price Forecast

Using Random Forest Regression, the prediction shows that next week product avg price and our prediction almost same to the actual price. So we could forecast revenue for the coming weeks, supporting the team that will set the goals.

**Reference:**

1. L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in IEEE Access, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.
2. P. Ji, H. -Y. Zhang and J. -Q. Wang, "A Fuzzy Decision Support Model With Sentiment Analysis for Items Comparison in e-Commerce: The Case Study of http://PConline.com," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49, no. 10, pp. 1993-2004, Oct. 2019, doi: 10.1109/TSMC.2018.2875163.
3. S. Rose, N. Hair and M. Clark, "Online Customer Experience: A Review of the Business-to-Consumer Online Purchase Context", International Journal of Management Reviews, vol. 13, no. 1, pp. 24-39, 2011, available: 10.1111/j.1468-2370.2010.00280.x.
4. D. Nguyen, S. de Leeuw and W. Dullaert, "Consumer Behaviour and Order Fulfilment in Online Retailing: A Systematic Review", International Journal of Management Reviews, vol. 20, no. 2, pp. 255-276, 2016, available: 10.1111/ijmr.12129.
5. H. Ceribeli, H. Tamashiro and E. Merlo, " Online flow and e-satisfaction in high involvement purchasing processes", BASE, vol. 14, no. 1, p. 16-29, 2017, available: 10.4013/base.2017.141.02.
6. Unido.org, 2020. [Online]. Available: https://www.unido.org/sites/default/files/2017-10/WP\_14.pdf. [Accessed: 8 March 2020].
7. "Brazil – Lucrative but Challenging E-commerce Industry", EOS Intelligence - Powering Informed Decision-Making, 2020. [Online]. Available: https://www.eos-intelligence.com/perspectives/consumer-goods-retail/brazil-lucrative-but-challenging-e-commerce-industry. [Accessed: 8 March 2020].
8. "Brazil Commercial Guide | International Trade Administration", Export.gov, 2020. [Online]. Available: https://www.export.gov/article?id=Brazil-e-Commerce. [Accessed: 8 March 2020].
9. Y. Yu and T. Yao, “Gender Classification of Chinese Weibo Users” in Proc. 2017 International Conference on E-commerce, E-Business and E-Government (ICEEG 2017), New York, NY, USA, 2017, pp. 5–8, doi:https://doi.org/10.1145/3108421.3108423
10. H. Huang, B. Zhao, H. Zhao, Z. Zhuang, Z. Wang, X. Yao, X. Wang, H. Jin, and X. Fu. “A Cross-Platform Consumer Behavior Analysis of Large-Scale Mobile Shopping Data” in Proc. 2018 World Wide Web Conference (WWW ’18), Republic and Canton of Geneva, CHE, 2018, pp. 1785–1794, doi:https://doi.org/10.1145/3178876.3186169
11. T. Charanasomboon and W.Viyanon. “A Comparative Study of Repeat Buyer Prediction: Kaggle Acquired Value Shopper Case Study” in Proc. 2019 2nd International Conference on Information Science and Systems (ICISS 2019), New York, NY, USA, 2019, pp. 306–310, doi:https://doi.org/10.1145/3322645.3322681