Data analysis of Olist e-commerce store

Tahani Khanom, Piam Emrul Hasan, Wasique Mohammad, Jaki Fayek Alvi Rahman

Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia.

Abstract— The Olist Store is an online platform where people can post or look for buying phones, fashionable items such as clothes and shoes or electronic devices. This paper presents an overview of The datasets of the olist store are used to measure many things including needs of customers, predicting trends and customer's behavior depending on many factors.

Keywords— Sales, Python, Sentiment analysis, Market basket analysis, Apriori, Linear regression.

I. Introduction

This project is related to a Brazilian ecommerce public dataset of orders which is made at Olist Store. The founder of Olist is Tiago Dalvi, the aim of creating it was to help the shopkeepers to reach huge and best marketplaces nationally and internationally. Products from the merchants are shipped directly to the customers through Olist's logistics partners.

It has become one of the Top 3 largest department stores inside Brazil's largest marketplaces which makes $1.8 Million in revenue annually. The dataset consists of 100k orders information from the year 2016 to 2018, orders made at different marketplaces in Brazil.

The orders can be viewed from various dimensions with the features which includes: price, order status, payment and performance to customer location, product attributes and at last reviews written by customers.

The project consists of 9 Tables with datasets. .Many research has been already conducted with this public dataset , still many business questions yet to be answered from the dataset.The purpose of this research is to understand the ecommerce domain better by analysing this datasets.

II. Related Works

This project is to compare the customer behavior to help the shopkeepers to reach huge and best marketplaces nationally and internationally. E-commerce is changing the way of business. It helps to manage the customer better, allows to discover new plans for marketing, expand the range of products, and work more efficiently. A key

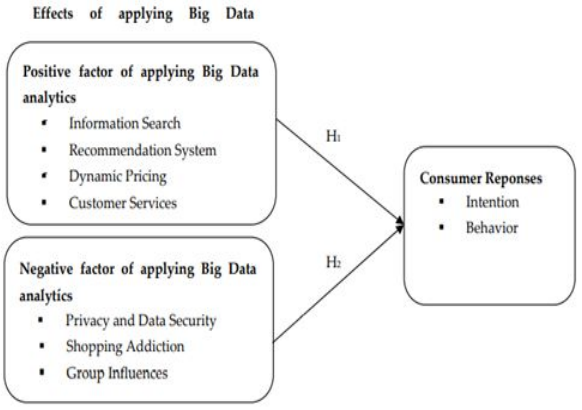
an enabler of this change is the widespread use of increasingly sophisticated data mining tools. The term 'data mining' describes the procedure of analyzing a company's internal data for customer profiling and targeting.

Accordingly, the following hypothesis is suggested: Hypothesis 1 (H1). Positive factor of applying Big Data analytics is positively associated with customers' responses. Chevalier and Mayzlin recommended that consumers' product reviews have an effect on the c purchasing behavior of customers.

Prices are charged according to customers' location, product, or time is known as Dynamic pricing which is often referred to in the economic terms as individual-level price differentiation. This strategy has become very common with Internet marketing as relying on it increases. These dynamic pricing is to increase the seller's profit by charging consumers with the highest prices as the consumers are ready to pay by being manipulated.

Fig. 1 Research Model

Big Data analytics is used by Amazon to store what the



customers have added inside their virtual shopping cart. These items are used in recently viewed products or take a purchasing action in the past. This technique is known as

item to item collaborative filtering. Another application is virtual presence which enables online shoppers to interact with shopping experience.
Here is the research model:

III. Data Description

The data set was acquired from the website of Kaggle which provides complimentary datasets which can be accessed and used for further research. We choose Data analysis of Olist e-commerce stores . The dataset includes data for the 2016 to 2018. The dataset includes information of 100k orders made at multiple marketplaces in Brazil. Based on the dataset, the following variables are listed: order status, price, payment, performance to customer location, product attributes and reviews written by customers

IV. Experimental Setup

This part of the paper explains the details of steps, as the correct products to sell is not always an easy task. There are many factors that drive whether a product is going to succeed or not and so, one split decision could be the difference between huge success and equivalent failure. So analysis on e-commerce data works by predicting the changes of a customer's purchase behavior due to qualitative products, seasonal sales and delivery performances and many other factors. The research will provide efficient monitoring analysis that will help retailers to predict potential buying impulses and capitalize on trends by maintaining the sales growth rate.

A. Algorithm – Linear regression

Linear regression algorithms will be used to predict the future sales growth rate. By using a linear algorithm we will train our model where it will try to find the correlation between previous sales rate and months.With the trained model ,predictive analysis of sales growth will be conducted.

Apriori algorithm

To determine frequently bought item sets association mining rules are used. There are three metrics for measuring association and they are support, confidence and lift. Support of an item is defined by the number of transactions containing the item divided by the total number of transactions. Confidence is the probability of an item being bought when another item is bought.

B. Performance Evaluation

To perform tasks on massive amounts of data, Apriori algorithm is used. Apriori algorithm can generate association rules from the frequently bought item sets. For an online marketing platform, delivery services play a vital role for customer satisfaction so we will analyse the delivery service's efficiency.
After analysing all these aspects, a visual representation of our research findings will be shown to make the research finding understandable to people from all domains .

C. Understanding dataset

Dataset name : Brazilian E-Commerce Public Dataset by Olist
This dataset is provided by Olist where data is divided into multiple tables for better understanding. We had to collect the data from a single source (Kaggle). The names of tables are listed below:
Kaggle (9 tables) namely: "olist_customer_dataset", "olist_geolocation_dataset", "olist_order_items_dataset", "olist_order_payment_dataset", "olist_order_reviews_dataset", "olist_order_dataset", "olist products dataset", "olist sellers dataset", "product_category_name_translation".

Fig. 2 Data Schema

D. Data Cleaning

The datasets were first examined before it was used as it should be cleaned if needed. There are multiple steps of data cleaning to make sure that the datasets are suitable for using and analysis.

First of all, we explored the dataset and cleaned the dataset to gather important features from it (feature engineering) for analysing our research questions. In online selling platforms, customers' interests towards the products usually change depending on the season. For this reason we searched for the product buying trend first then we analysed which products are to be kept in the stock depending on the customer demand.

Figure below shows the data cleaning steps taken, being done using it.

Fig.3 Data Cleaning

IV. Analysis of Results

We have analysed the payment methods which were widely used by the customers and negative or positive review's effect on product sale.

We have also analysed which product has been sold the most from the dataset. From a business perspective it is necessary to know the future sales growth rate so that appropriately the initiatives can be taken to prevent loss.

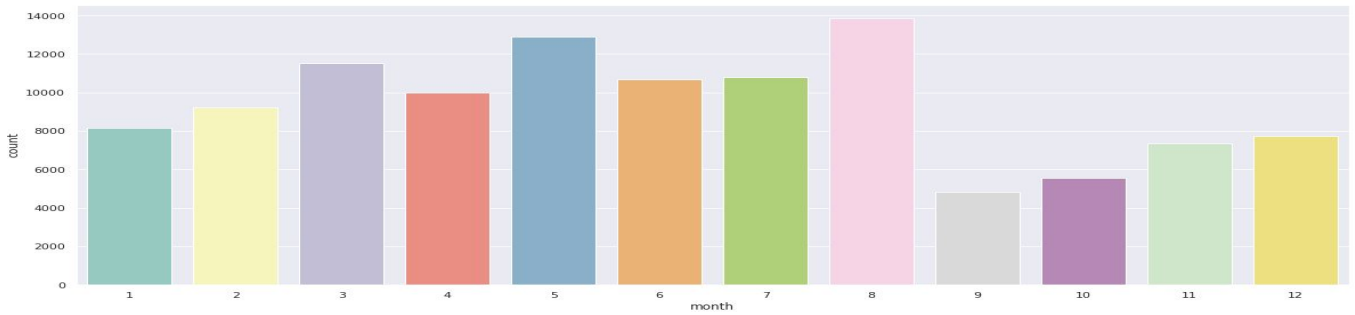


Fig. 4 sales trend

Based on fig 4, there is a significant trend between the month and the sales. The overall sales are at pick on the month of August, followed by the months May and March. From the month of September we can notice a trend of sales increasing from October until March.

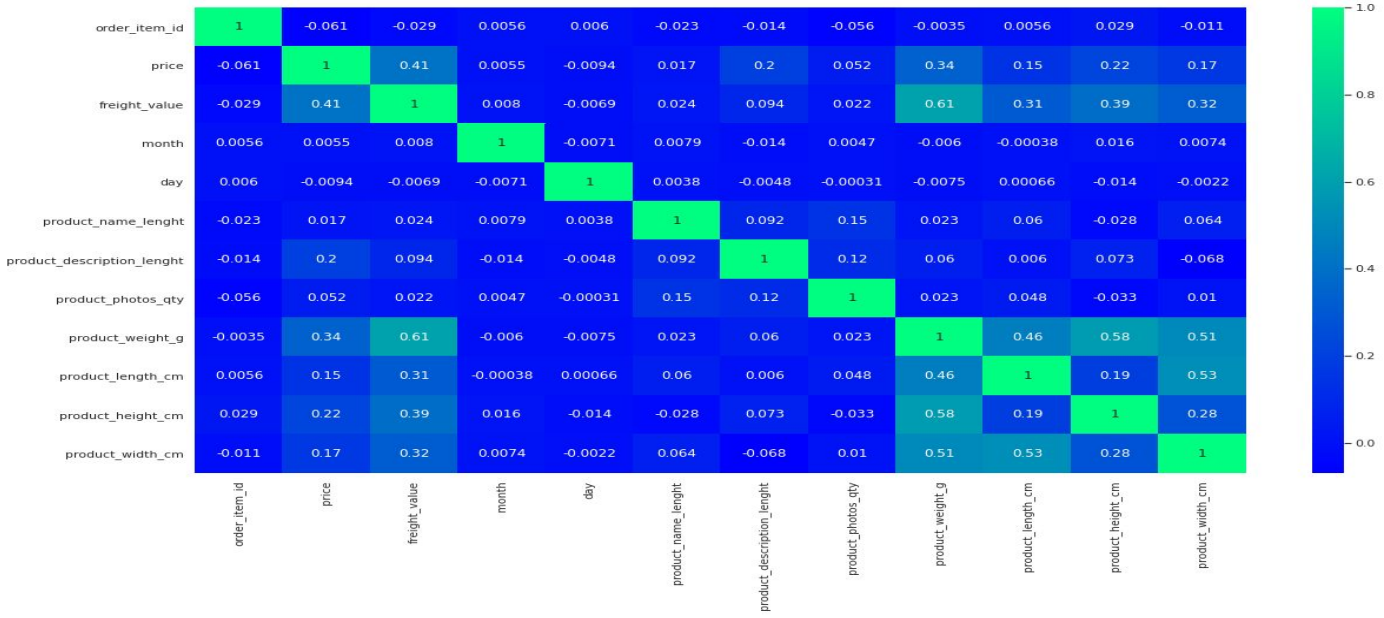


Fig 5. Future sales growth rate

Based on fig 5, it shows the linear relationship among the data. 1.0 indicates the highest relationship and 0.0 indicates the lowest

```
#implementation of algorithoms is taken from sklearn documentation
from sklearn import linear_model
from sklearn.metrics import mean_squared_error
import sklearn.metrics as sm

X = prod_sales[['month','year']].astype(int)
Y = prod_sales['price'].astype(float)
X_train, X_test, y_train, y_test = train_test_split(X, Y, random_state=1)
X_test
regr = linear_model.LinearRegression()
regr.fit(X_train, y_train)
y_pred = regr.predict(X_test)
y_pred
m=mean_squared_error(y_test, y_pred)

accuracy=regr.score(X_test,y_test)
accuracy
print("R2 score =", round(sm.r2_score(y_test, y_pred), 2))
```

```
R2 score = -0.76
```

Fig 6. Linear regression

Based on fig 6, we tried to use linear regression but can not use LinearRegression on this purpose because variables are not linearly related .
We can use the Prophet algorithm , a facebook developed algorithm to predict or forecast future observation. In our case, we predicted the future sales. Below is the figure with the prophet algorithm.

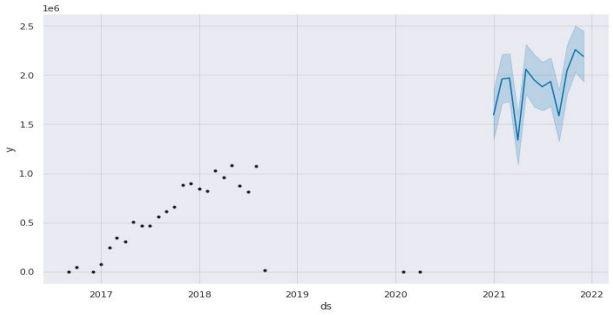


Fig 7. Future sales

Based on Fig 7., it shows the sales of the whole year (2021), with some fluctuations.

We have found out the revenue depending on the categories of the product. The highest revenue product is at the top. Beleza_saude gives the highest revenue Shown below in fig 8.

```
#sales by most sales money
top_sales = pd.DataFrame(q5.groupby('product_category_name').sum()['price'])

top_sales.sort_values(by=['price'], inplace=True, ascending=False)

top_sales = top_sales.head(10)

top_sales
```

product_category_name	price
beleza_saude	1258681.34
relogios_presentes	1205005.68
cama_mesa_banho	1036988.68
esporte_lazer	988048.97
informatica_acessorios	911954.32
moveis_decoracao	729762.49
cool_stuff	635290.85
utilidades_domesticas	632248.66
automotivo	592720.11
ferramentas_jardim	485256.46

Fig 8.highest revenue

Products which are sold in large quantities are shown in the table in fig 9.

```
top_sales_by_count = pd.DataFrame(q5.groupby('product_category_name').count()['order_id'])
top_sales_by_count.sort_values(by=['order_id'], inplace=True, ascending=False)
top_sales_by_count = top_sales_by_count.head(10)
top_sales_by_count
```

product_category_name	order_id
cama_mesa_banho	11115
beleza_saude	9670
esporte_lazer	8641
moveis_decoracao	8334
informatica_acessorios	7827
utilidades_domesticas	6964
relogios_presentes	5991
telefonia	4545
ferramentas_jardim	4347
automotivo	4235

Fig 9. Highest sale product

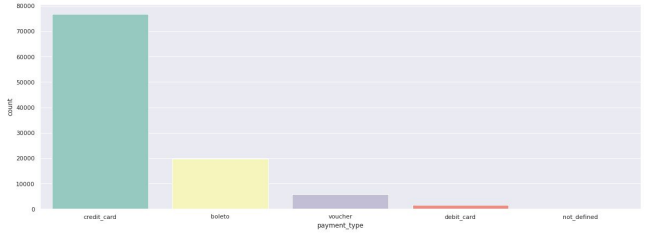


fig 10. Popular payment method

Based on fig 10, credit_card is the most popular payment method among the customers and the second most popular is boleto which counts 20000. Debit_card is least used.

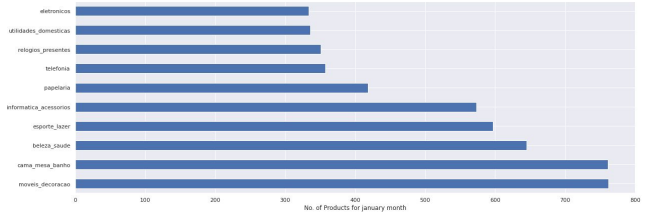


Fig 11. Top product of January

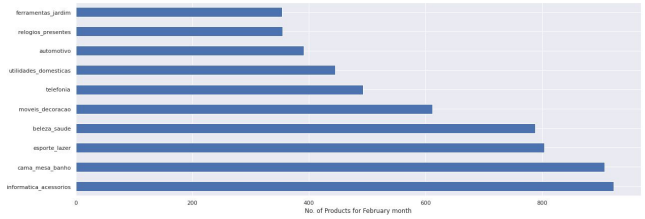


Fig 15. Top product of May

Fig 12. Top product of February

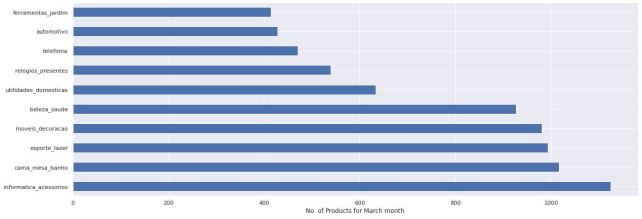


Fig 13. Top product of March

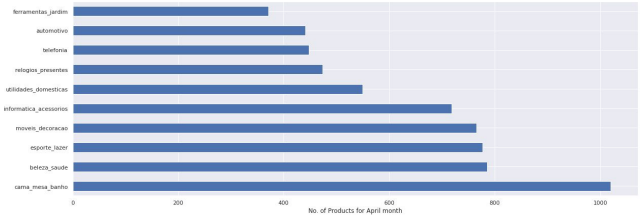
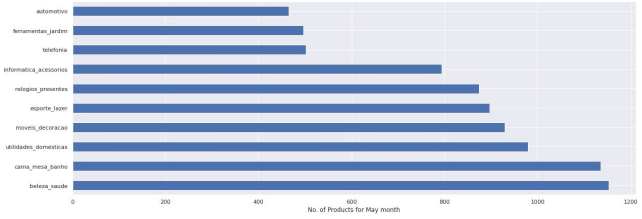


Fig 14. Top product of April



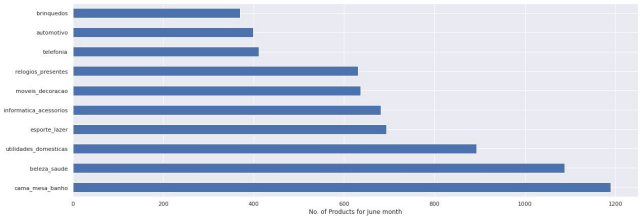


Fig 16. Top product of June

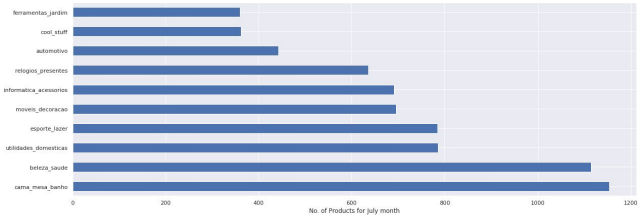


Fig 17. Top product of July

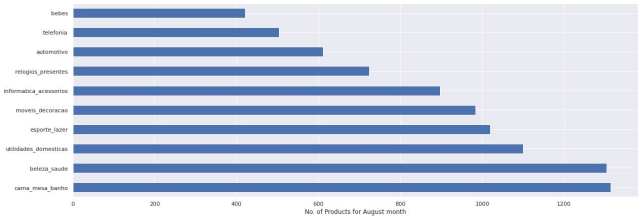


Fig 18. Top product of August

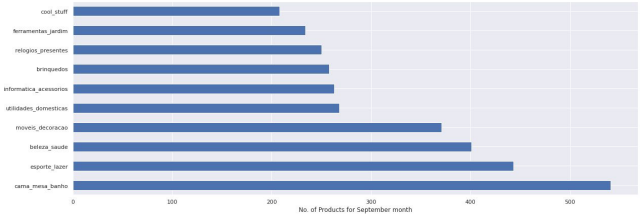


Fig 19.. Top product of September

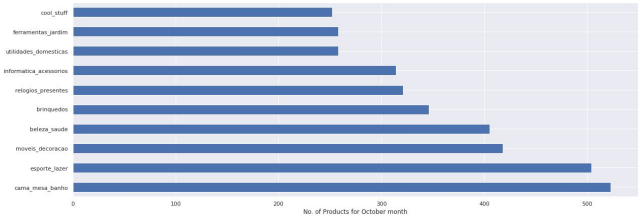


Fig 20. Top product of October

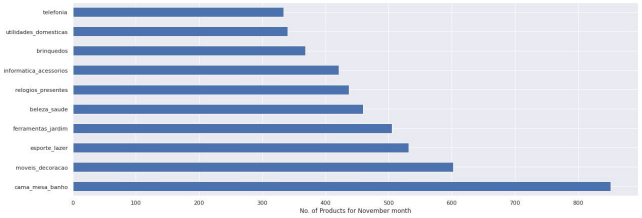


Fig 21. Top product of November

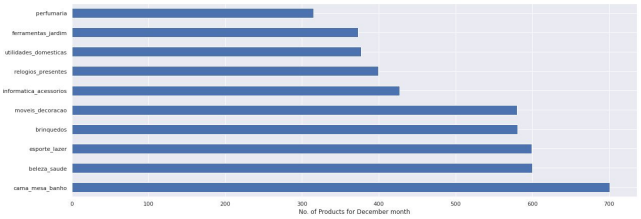


Fig 22. Top product of December

For market basket analysis or to identify frequently bought item sets we have used the Apriori algorithm. We have grouped items by order and generated association rules for item sets. The following table represents the partial table of the highest confidence item sets.

	Item 1	Item 2	confidence
35	5b8a5a9417210b1b84b67b9a7aefb935	e5ae72c62ebfa708624f5029d609b160	1.000000
34	f4f67ccaece962d013a4e1d7dc3a61f7	4fcb3d9a5f4871e8362dfedbdb02b064	0.944444
33	4fcb3d9a5f4871e8362dfedbdb02b064	f4f67ccaece962d013a4e1d7dc3a61f7	0.894737
32	18486698933fbb64af6c0a255f7dd64c	dbb67791e405873b259e4656bf971246	0.875000
29	060cb19345d90064d1015407193c233d	98d61056e0568ba048e5d78038790e77	0.857143
31	5d790355cbeded0cd60e25cbc4c527a2	5fc3e6a4b52b0c414458104ed4037f1c	0.857143
30	5fc3e6a4b52b0c414458104ed4037f1c	5d790355cbeded0cd60e25cbc4c527a2	0.857143
28	e6b314a2236c162ede1a879f1075430f	ad4b5def91ac7c575dbdf65b5be311f4	0.750000
27	3ce943997ff85cad84ec6770b35d6bcd	b7d94dc0640c7025dc8e3b46b52d8239	0.714286
26	35afc973633aaeb6b877ff57b2793310	99a4788cb24856965c36a24e339b6058	0.707317
25	ee57070aa3b24a06fdd0e02efd2d757d	0d85c435fd60b277ffb9e9b0f88f927a	0.666667
24	4d0ec1e9b95fb62f9a1fbe21808bf3b1	9ad75bd7267e5c724cb42c71ac56ca72	0.666667
23	ad0a798e7941f3a5a2fb8139cb62ad78	9463446971569 47d846d27fe0d503033	0.666667
22	36f60d45225e60c7da4558b070ce4b60	e53e557d5a159f5aa2c5e995dfdf244b	0.666667
21	ad4b5def91ac7c575dbdf65b5be311f4	e6b314a2236c162ede1a879f1075430f	0.666667
19	a50acd33ba7a8da8e9db65094fa990a4	dfb97c88e066dc22165f31648efe1312	0.625000
20	b7d94dc0640c7025dc8e3b46b52d8239	3ce943997ff85cad84ec6770b35d6bcd	0.625000
18	e53e557d5a159f5aa2c5e995dfdf244b	36f60d45225e60c7da4558b070ce4b60	0.607143
16	e5ae72c62ebfa708624f5029d609b160	5b8a5a9417210b1b84b67b9a7aefb935	0.600000
15	98d61056e0568ba048e5d78038790e77	060cb19345d90064d1015407193c233d	0.600000

Figure 23.

These rules can be used to suggest/recommend products to buyers when they buy certain products.

For customer review analysis, there were a significant amount of NaN values in the review dataset, so after dropping the rows containing NaN values we got the necessary reviews needed for the analysis.

```
9                          recomendo
15                   Super recomendo
19          Não chegou meu produto
22                             Ótimo
34                       Muito bom.
                           ...
99967                              👍
99971              muito bom produto
99972     Não foi entregue o pedido
99974                 OTIMA EMBALAGEM
99975                  Foto enganosa
```

Figure 24. Review Comment Titles

There were also stopwords (i.e as, ele, elas, fosse)to consider when normalizing texts. After removing the stopwords we got ourselves clean review comments.

```
comments #clean comments

'estao parabens sempre chega muiiita antecedencia obrigada',
'ok recomendo',
'produto chegou pc n conseguiu reconhecer portas usb',
'produto boa qualidade chegou antes prazo prometido',
'pedido balde pe blocos montar un r cada n entregue vendido entregue targaryen
'entrega r pida parab ns',
'comprei tres pacotes cinco folhas cada papel transfer tecido escuro so recebi
'faltando apenas produto recebi hoje obrigada tudo certo att elenice',
'bolsa t rmica al m linda super espa osa produto acordo anunciado entrega super
'demorou pra entrega',
'super rapido entrega chegou antes data',
'entrega dividida duas n comunicado loja cheguei pensar s haviam enviado parte
'n funciona n faz sincronismo',
'recomendo todos clientes',
'gostei',
'otima',
'gostei jogo banho estampa florida bordado tudo bonito timo produto',
'chegou apenas pe nota garantia constam duas joias',
'n recebi produto consta sistema recebi al m pagar caro frete',
```

Figure 25. Clean Comments

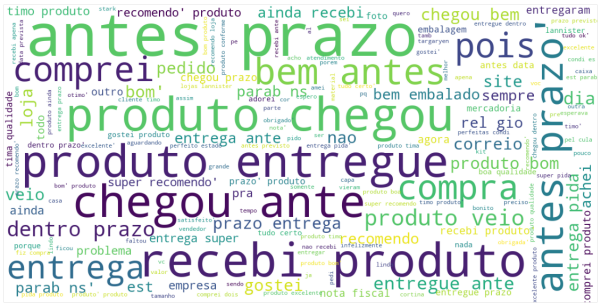


Figure 26. Comment Word Cloud

This word cloud is much more sophisticated than the one before. But it doesn't mean much if we do not understand Portuguese. So we will now translate the comments to English for better understanding.

	Trigrams	frequency	english_translation
0	chegou antes prazo	1112	arrived before deadline
1	bem antes prazo	708	well before term
2	entregue antes prazo	608	delivered before deadline
3	produto chegou antes	448	product arrived before
4	entrega antes prazo	433	delivery before deadline
5	chegou bem antes	417	arrived well before
6	produto entregue antes	374	product delivered before
7	entrega super pida	297	super delivery ask
8	antes prazo previsto	287	before expected deadline
9	ainda recebi produto	261	I still received product
10	produto tima qualidade	257	great quality product
11	produto boa qualidade	234	good quality product
12	entregue dentro prazo	230	delivered on time
13	produto entregue prazo	228	product delivered on time
14	antes prazo produto	227	before product term
15	antes data prevista	218	before expected date

Figure 27. English Translation of the reviews

In Fig. 27, the translation from Portuguese - English was done using Python's Google Translation Library for better understanding of the reviews as it won't mean much if one does not understand Portuguese. Looking through the dataframe, the unhappy comments of displeased customers were more understandable. These comments include 'I did not receive the product', 'I want my money back', 'so far nothing'. They are mostly complaints about getting incomplete delivery or receiving defective/wrong products. By far, these were the major complaints of unhappy customers but we have also come across a significant amount of satisfactory comments among the reviews.

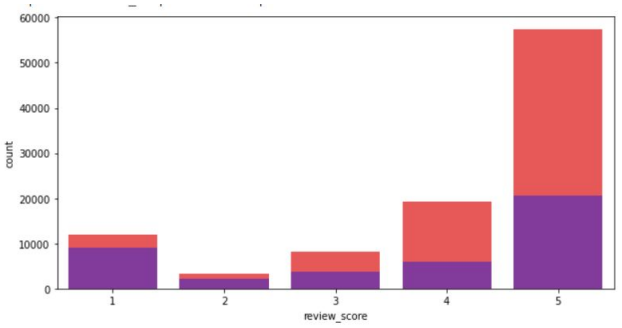


Figure 28. Customer Review Score

In the above countplot, we can notice that about 40% of the 5-star reviewers gave reviews on their purchase while 60% of them did not. Again for 1-star reviewers almost 80% of them gave reviews. So we can conclude that customers tend to give reviews when they are displeased with their purchases.

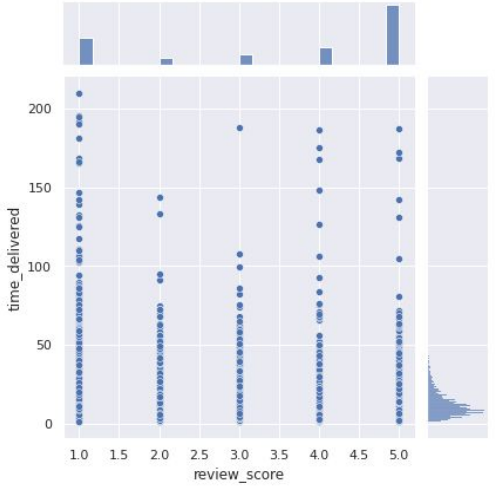


Figure 29. Review score on delivery time

From this plot, we can see that there are more 1-start reviews than 5-stars while considering delivery time. So the longer it takes to deliver the lower the review scores would be.

VII. Conclusions

In conclusion, in this work we have answered the 8 data science questions of different types as proposed. The analysis of the dataset has given us insight into various aspects of the business process and customer sentiments on their purchases which can be utilized to adjust or improve business strategy.

Acknowledgment

This project was completed under the course Big Data Analytics (CSC 3303), a course which is under the Department of Computer Science in Kulliyyah of Information and Communication Technology,

International Islamic University Malaysia, Kuala Lumpur. The authors of the research would like to thank Dr. Sharyar Wani for helping and assisting in the progress for making this paper a success.

References

Apriori: Association Rule Mining In-depth Explanation and Python Implementation | by Chonyy | Towards Data Science. (n.d.). Retrieved January 18, 2021, from https://towardsdatascience.com/apriori-association-rule-mining-explanation-and-python-implementation-290b42afdfc6

apriori-python · PyPI. (n.d.). Retrieved January 18, 2021, from https://pypi.org/project/apriori-python/

TF - IDF for Bigrams & Trigrams - GeeksforGeeks. (n.d.). Retrieved January 18, 2021, from https://www.geeksforgeeks.org/tf-idf-for-bigrams-trigrams/

Time Series Forecasting With Prophet in Python - AnalyticsWeek. (n.d.). Retrieved January 18, 2021, from https://analyticsweek.com/content/time-series-forecasting-with-prophet-in-python/

google-trans-new · PyPI. (n.d.). Retrieved January 18, 2021, from https://pypi.org/project/google-trans-new/

Data Analysis using Python - Sales Analysis. (n.d.). Retrieved January 18, 2021, from https://www.storiesondata.com/post/data-analysis-using-python-sales-analysis