SDSC3013 | INTRODUCTION TO SOCIAL MEDIA ANALYTICS

BRAZILIAN E-COMMERCE ANALYSIS

Presented to you by:
LESMANA Matthew Elbert (57319566)
MIJAATA Xavier Felixan (57195713)
ELEAZAR Christopher Otniel (57210339)
BUDIMAN Darren Nathanael (56672641)

TABLE OF CONTENT

- INTRODUCTION
- DATASET
- DATASET PREPROCESSING
- DATA ANALYSIS
- SENTIMENT ANALYSIS
- MACHINE LEARNING
- HAPPINESS ANALYSIS
- CONCLUSION

INTRODUCTION

01

- E-commerce is evolving rapidly, driven by recent advancements in technology.
- It is hoped that this analysis will provide insights useful to accommodate the rapidly changing digital marketplace.

02

DATASET

OLIST AND 3 COLLABORATORS · UPDATED 3 YEARS AGO

3309 | New Notebook | Download

Brazilian E-Commerce Public Dataset by Olist

100,000 Orders with product, customer and reviews info

Data Card | Code (541) | Discussion (58) | Suggestions (0)

About Dataset

Brazilian E-Commerce Public Dataset by Olist

Welcome! This is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows 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.

This is real commercial data, it has been anonymised, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

Join it With the Marketing Funnel by Olist

We have also released a Marketing Funnel Dataset. You may join both datasets and see an order from Marketing perspective now!

Instructions on joining are available on this Kernel.

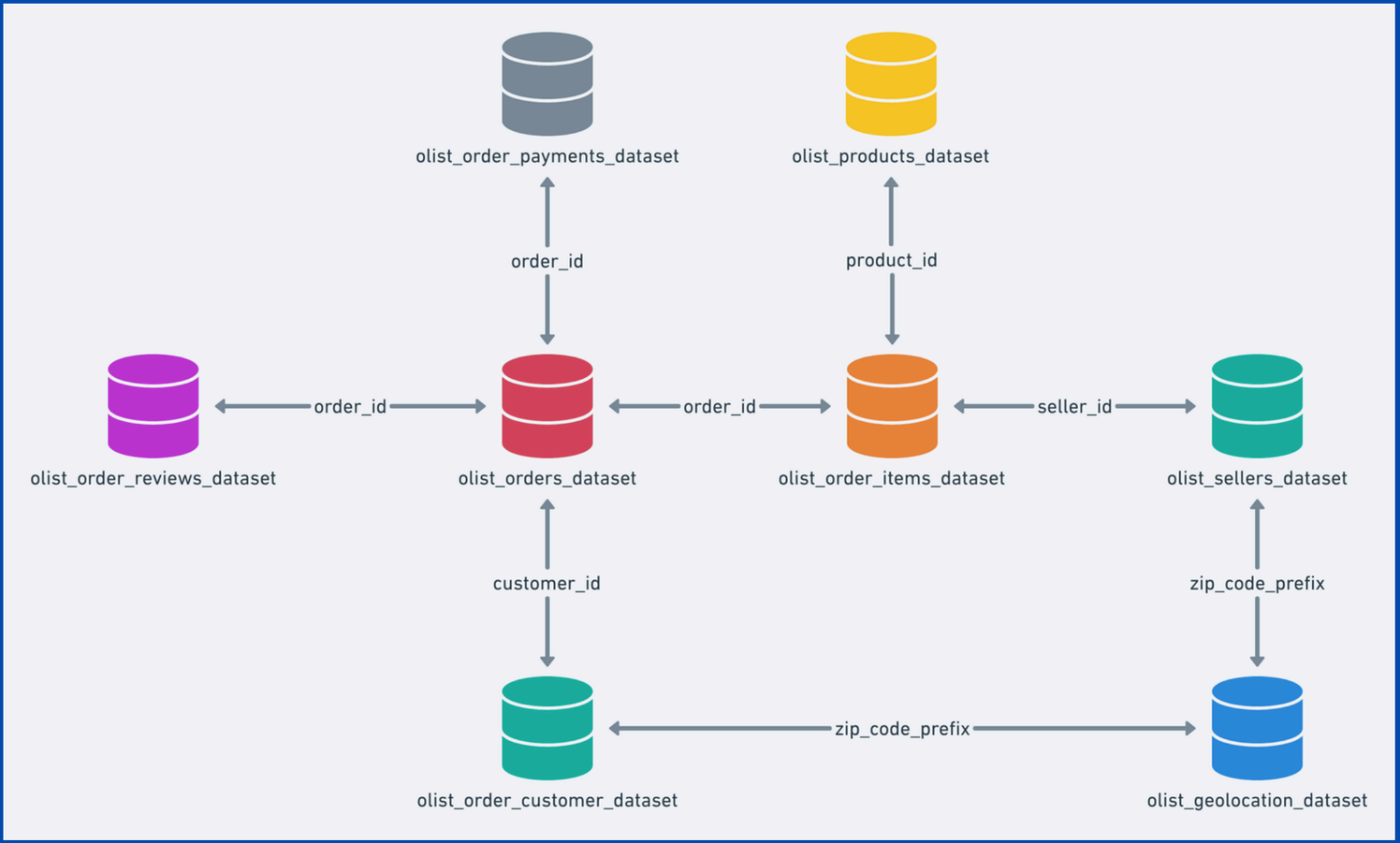
Usability
10.00

License
CC BY-NC-SA 4.0

Expected update frequency
Never

Tags
- Business
- Data Visualization
- Exploratory Data Analysis
- NLP

DATASET



- The dataset was provided by a Brazilian e-commerce, Olist Store, which originally has information on 100,000 orders from 2016 to 2018 made at multiple marketplaces in Brazil.
- Includes 9 different CSV files, containing information on order status, price, products information, etc.
- After preprocessing, we have 49,679 rows and 41 features.

03

DATA PREPROCESSING

03 DATA CLEANING

- Preprocessing steps:
  - Convert time-related columns to DateTime format.
  - Create new columns for hours, months, years, and dates for deeper analysis.
  - Add and join columns to increase features and extract more insights.
- Data focus:
  - Time range for EDA: 2017 to August 2019.
  - Excluding irrelevant data before and after this range to support analysis.

03 TRANSLATION

- Since the dataset is in Portugese, translation is needed to convert it to English for further analysis.
- The Google Translate API was used from 2 different libraries which were deep_translator and googletrans to avoid rate limits.

review_comment_message
Recebi bem antes do prazo estipulado.
Parabéns lojas lannister adorei comprar pela I...
aparelho eficiente. no site a marca do aparelh...
Mas um pouco ,travando...pelo valor ta Boa.\r\n
Vendedor confiável, produto ok e entrega antes...

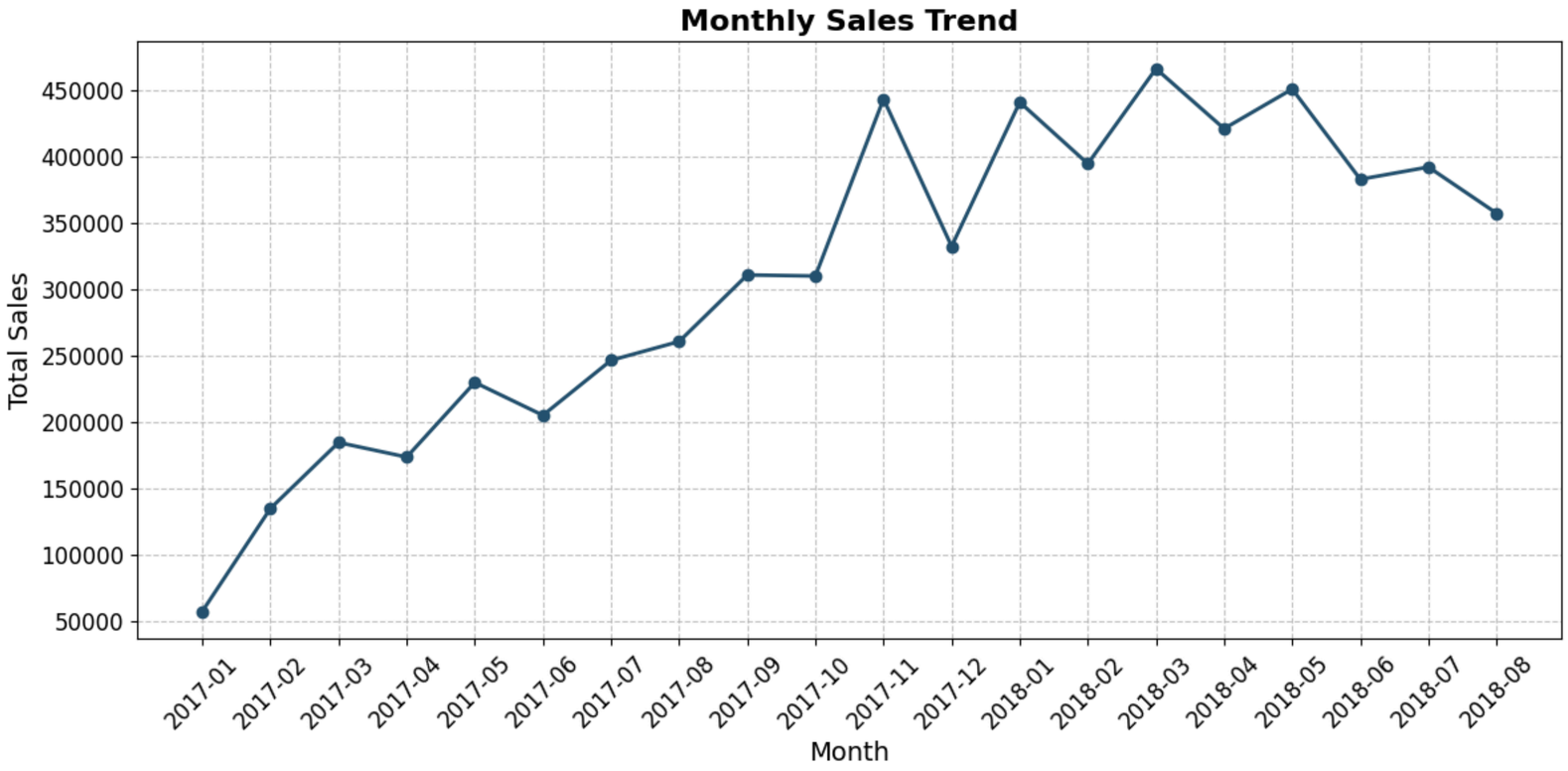
translated_review_comment_message
I received it well before the stipulated deadl...
Congratulations Lannister Stores, I loved buyi...
efficient device. on the website the brand of ...
But a little bit, freezing... for the price it...
Reliable seller, ok product and delivery befor...

04

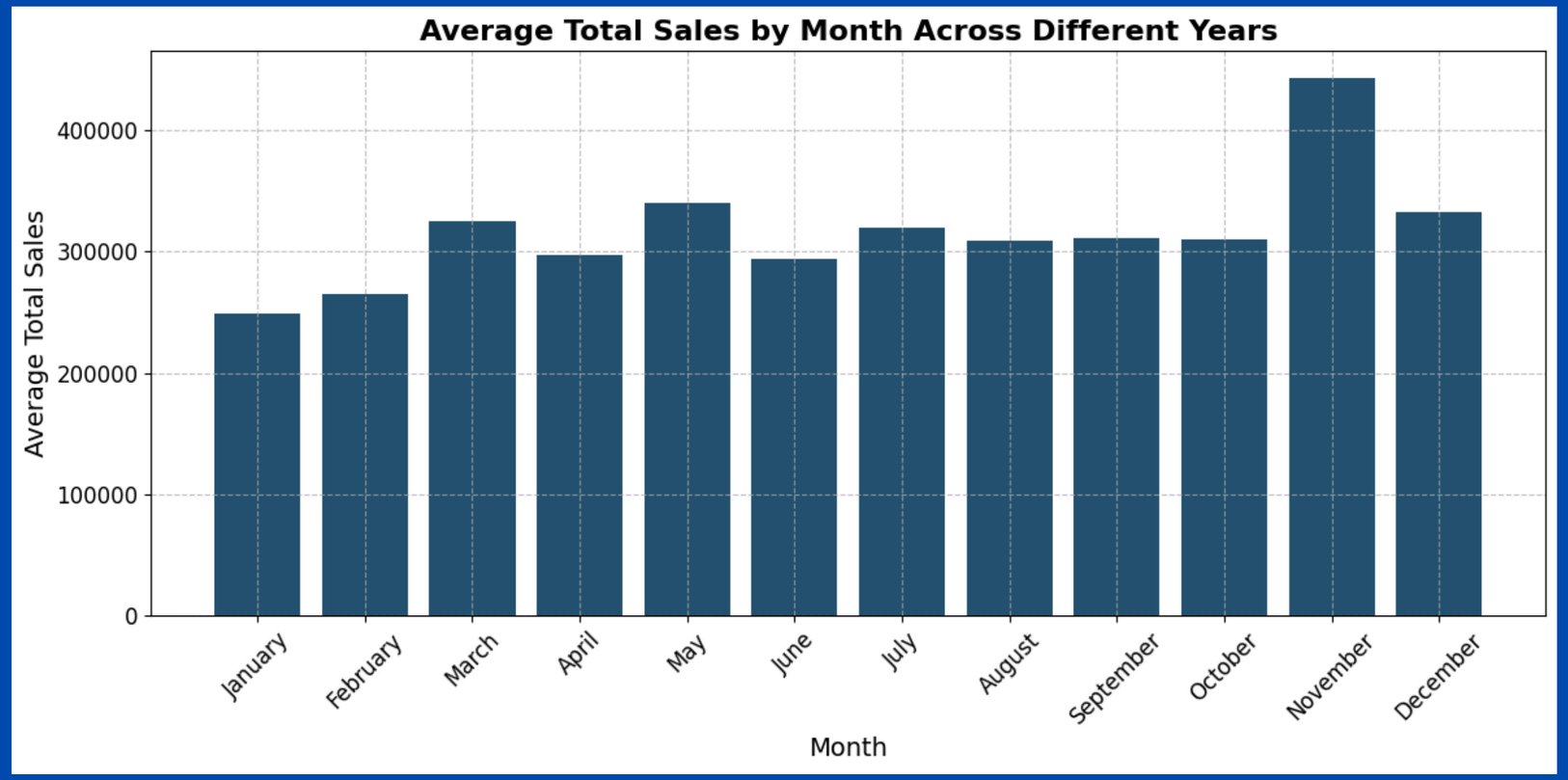
DATA ANALYSIS

Exploratory Data Analysis

MONTHLY SALES TREND

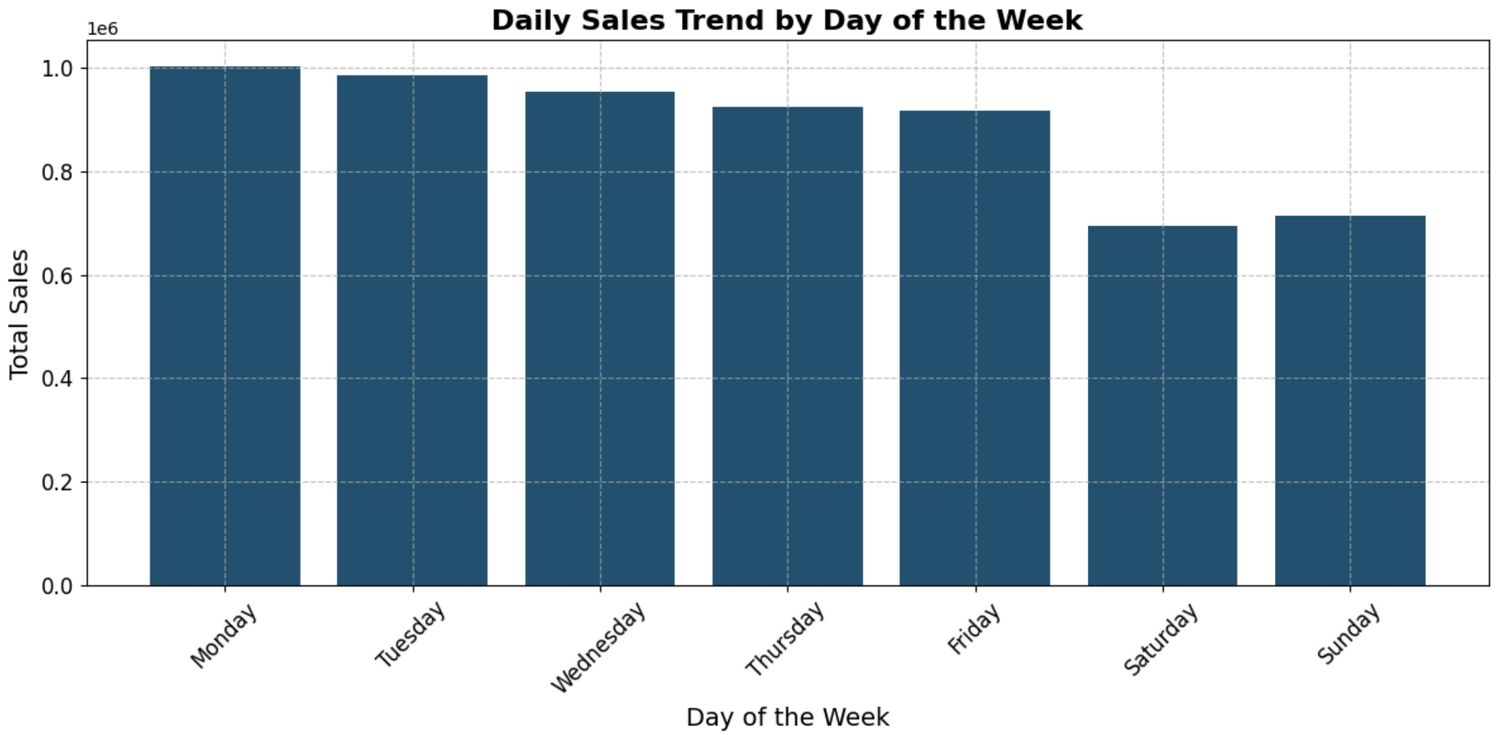


- A notable increase from January 2017 to November 2017
- Downward trend starting from February 2018



AVERAGE SALES PER MONTH

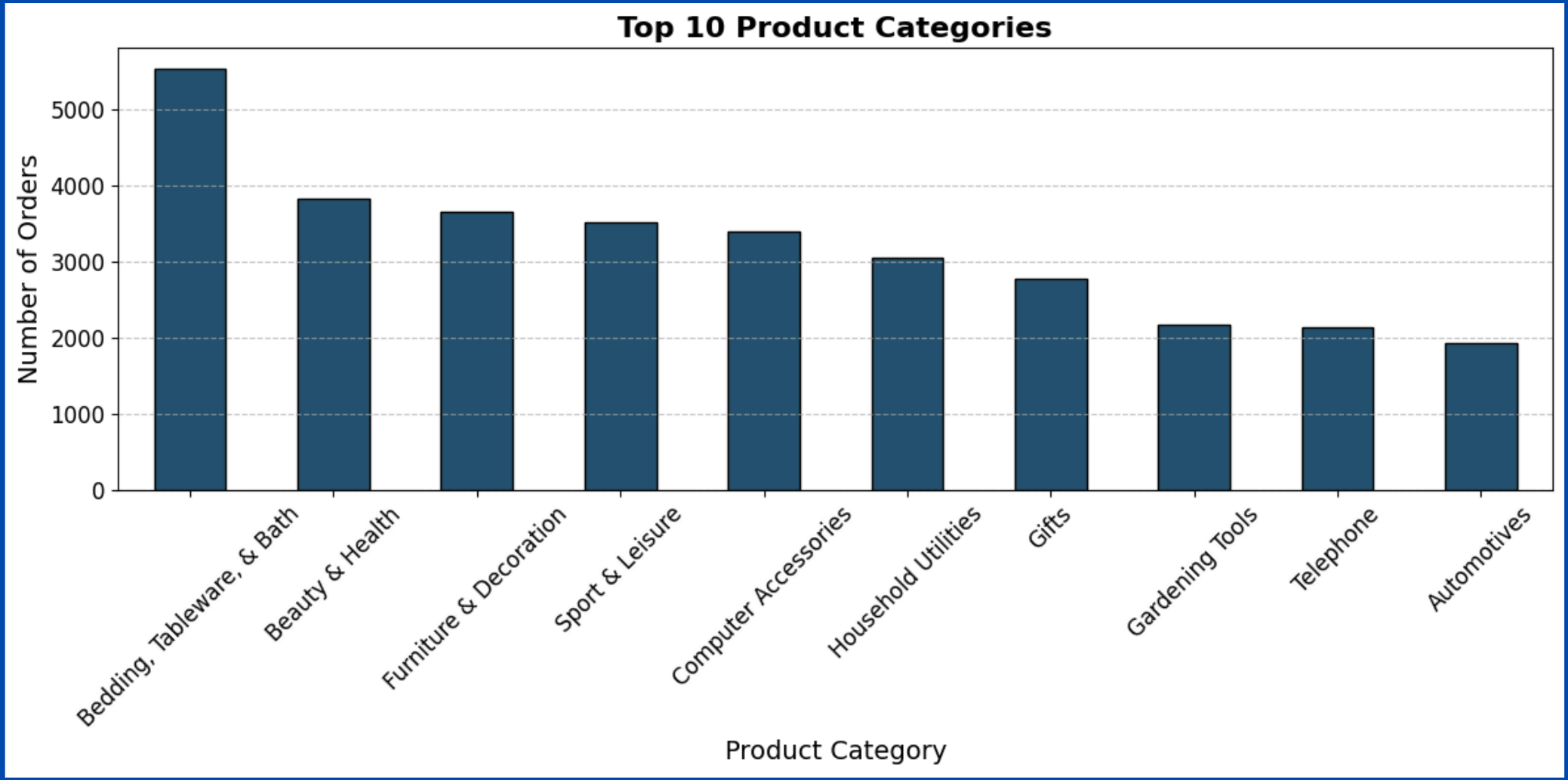
- Average total sales remain relatively consistent throughout the year, with a slight increase in the months of November and December.



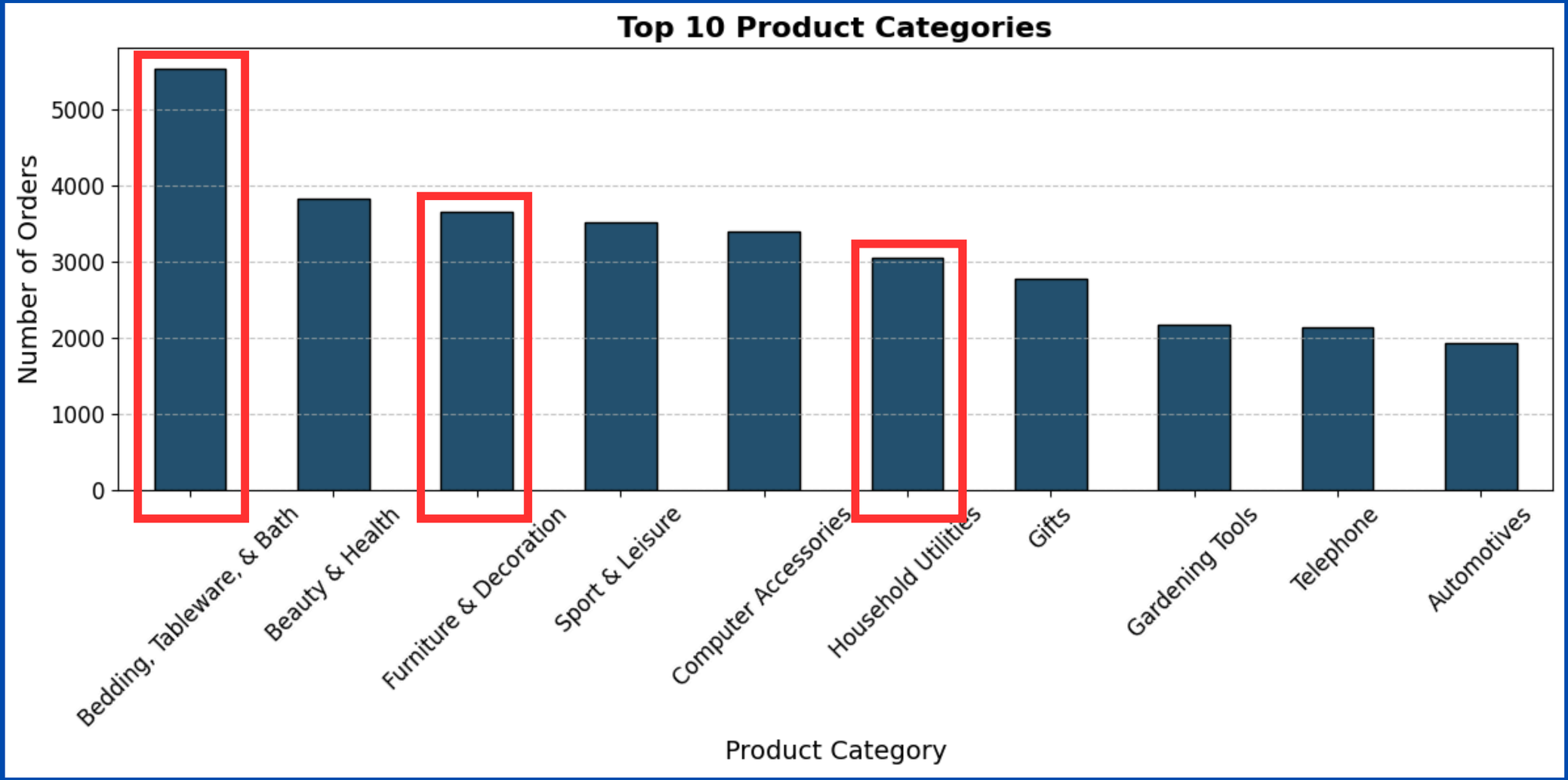
- Sales are relatively consistent throughout the week with slight variations
- Most Customer prefer to purchase items on weekdays rather than weekends.

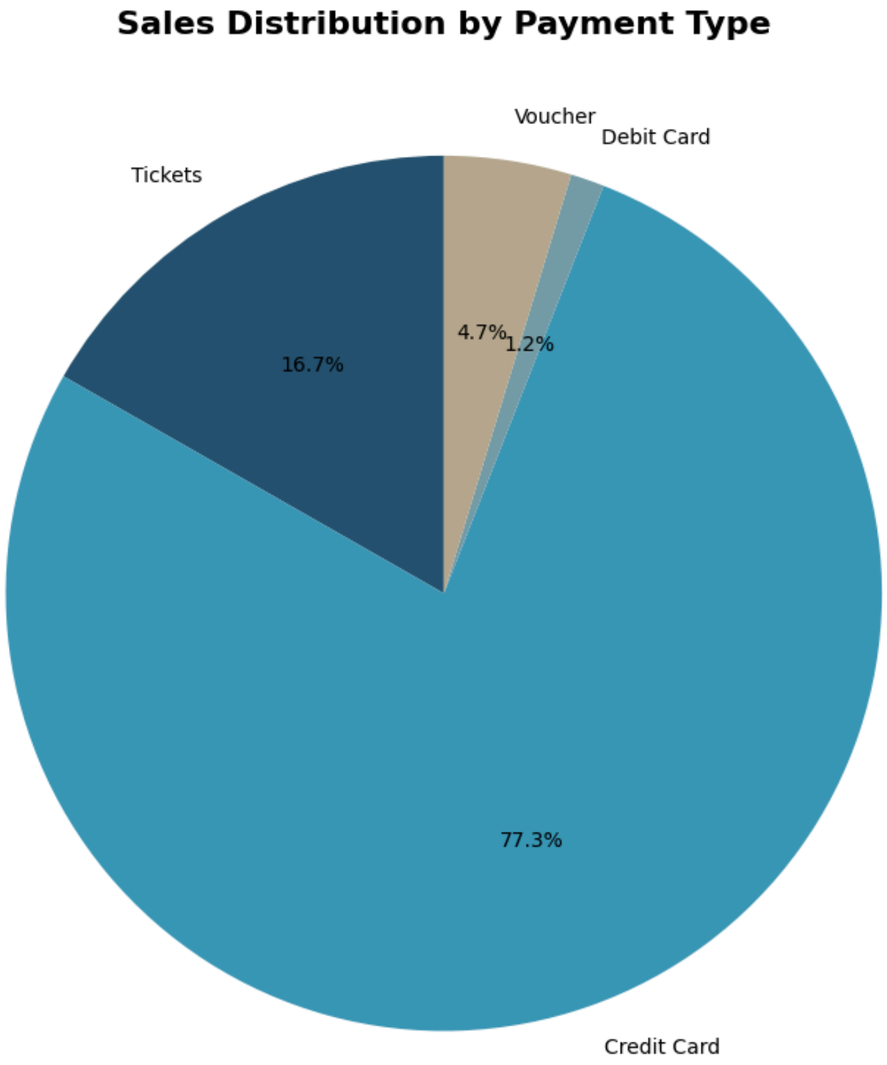
AVERAGE SALES PER DAY OF WEEK

TOTAL SALES BY PRODUCT CATEGORY

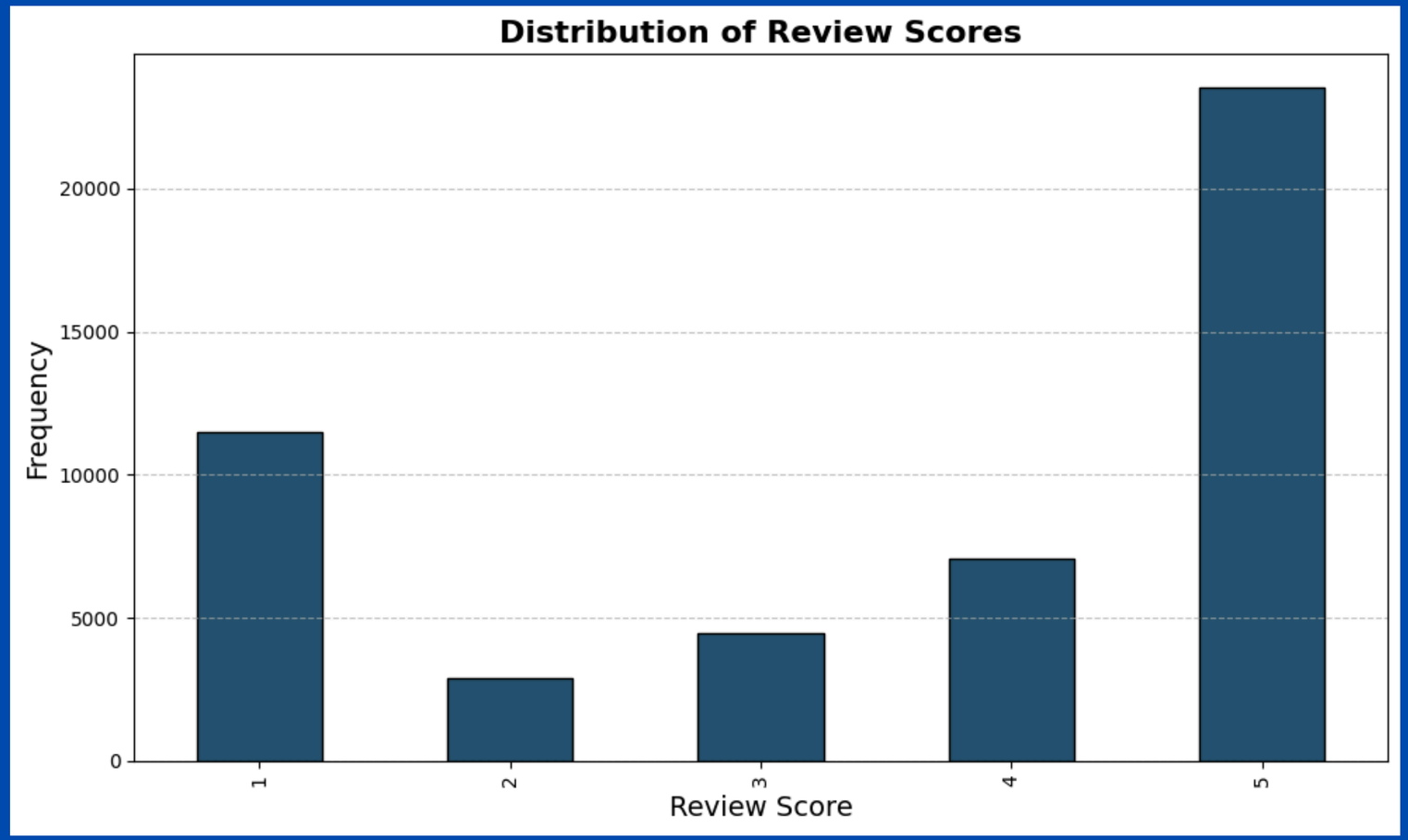


TOTAL SALES BY PRODUCT CATEGORY



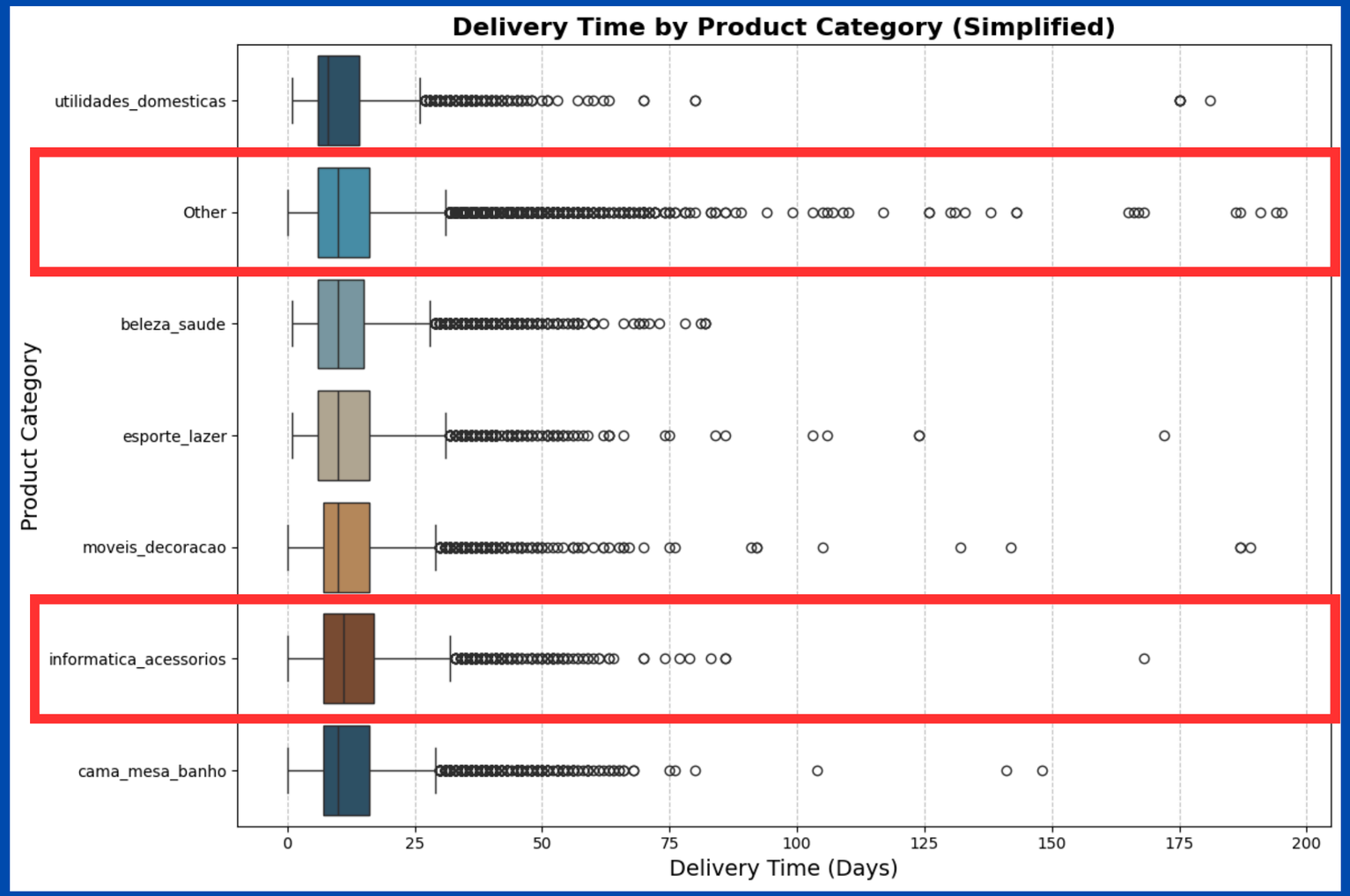


PROPORTION OF TOTAL SALES BY PAYMENT TYPE



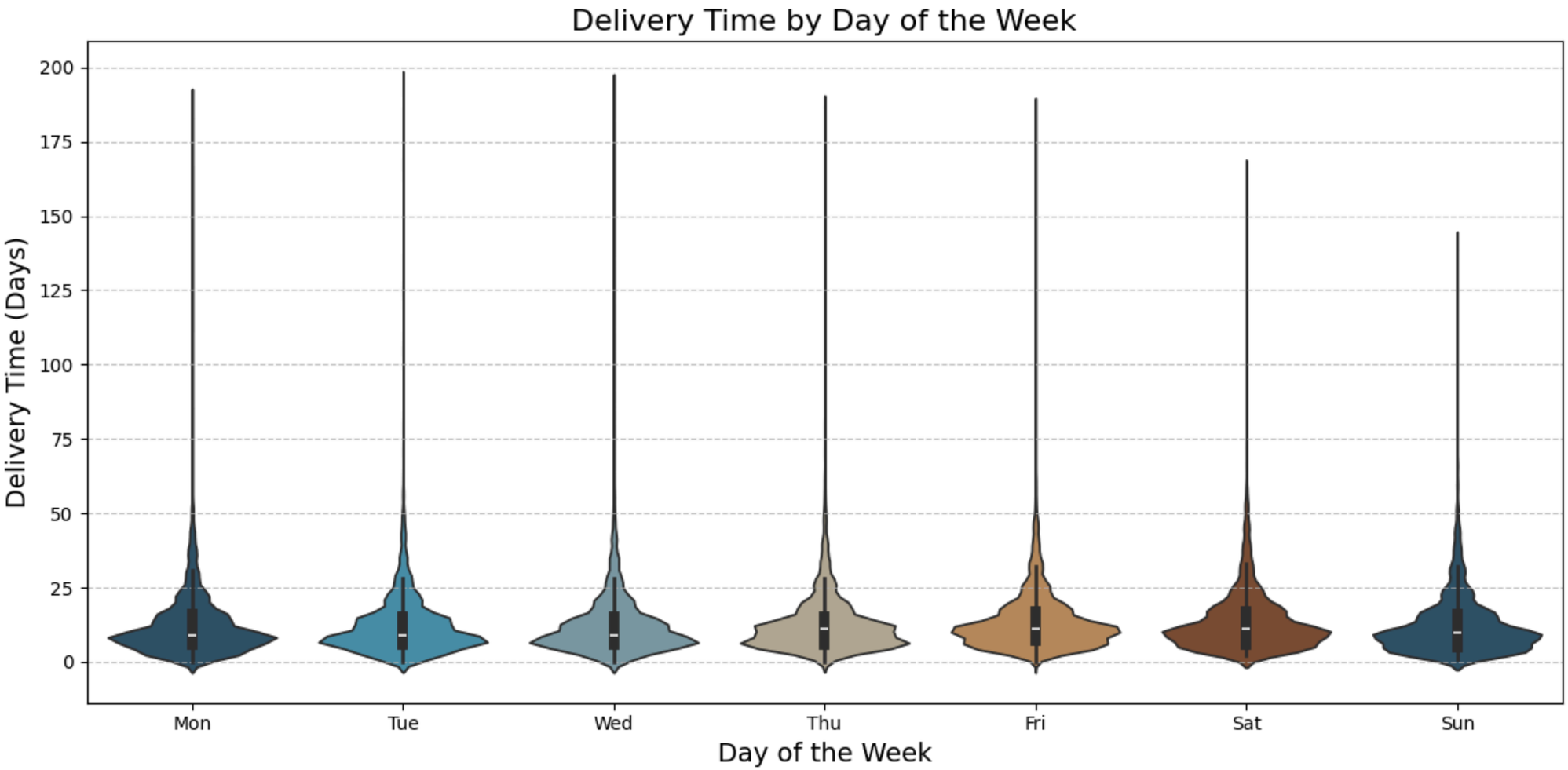
DISTRIBUTION OF REVIEW SCORES (HISTOGRAM)

CORRELATION HEATMAP

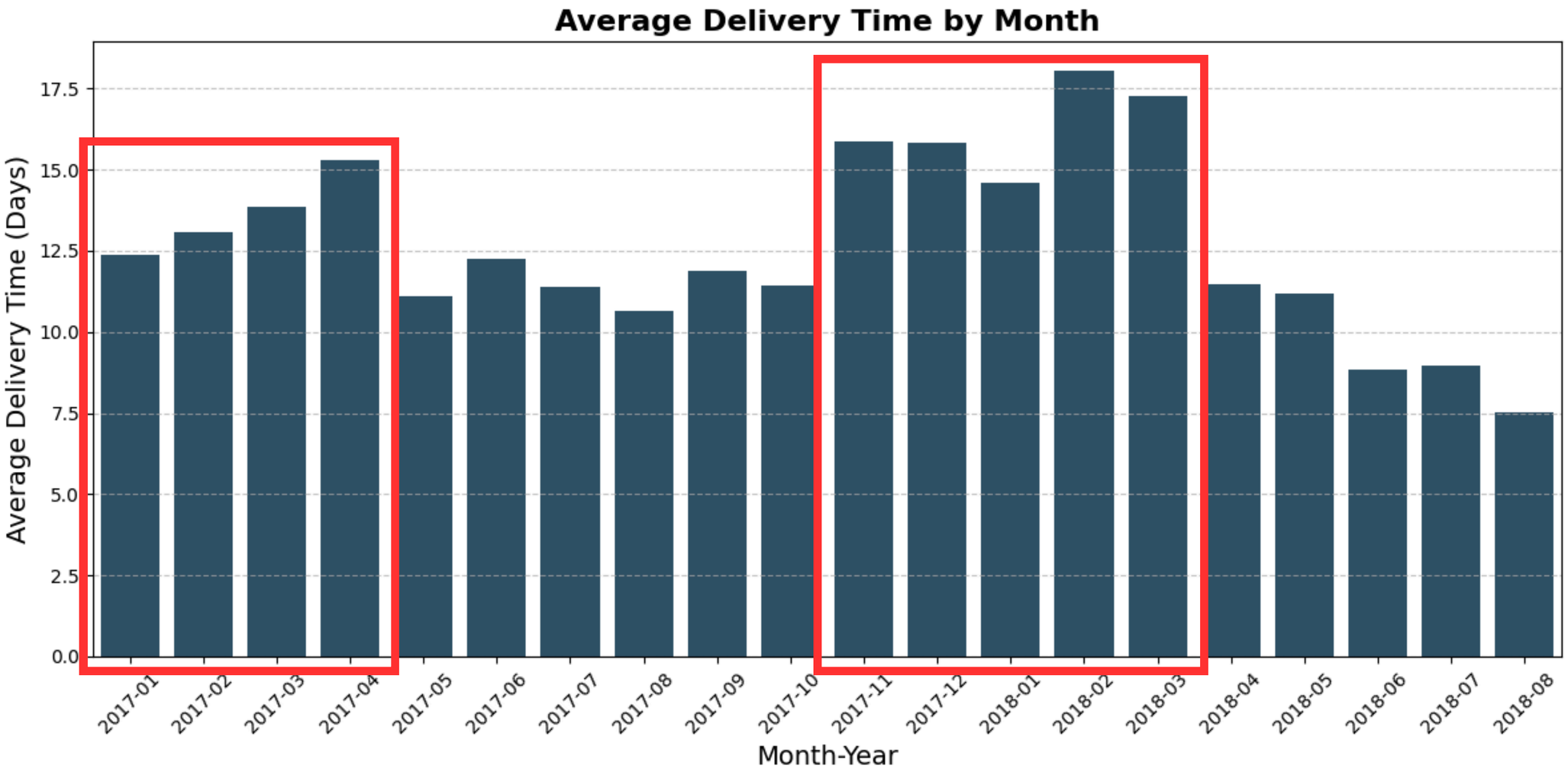


DELIVERY TIME BY PRODUCT CATEGORY

DELIVERY TIME BY DAY OF THE WEEK



AVERAGE DELIVERY TIME BY MONTH



05

SENTIMENT ANALYSIS

Natural Language Processing , Text Preprocessing, and Sentiment Analysis

05 NLP

- Before doing the sentiment analysis, we need to do some preprocessing.
- Selecting the necessary columns for NLP analysis.

```
nlp_df = df[['review_comment_title', 'review_comment_message', 'product_category_name']]
```

	review_comment_title	review_comment_message	product_category_name
0	NaN	Não testei o produto ainda, mas ele veio corre...	utilidades_domesticas
1	NaN	Não testei o produto ainda, mas ele veio corre...	utilidades_domesticas
2	NaN	Não testei o produto ainda, mas ele veio corre...	utilidades_domesticas
3	Muito boa a loja	Muito bom o produto.	perfumaria
4	NaN	O produto foi exatamente o que eu esperava e e...	pet_shop
...	...	...	...
49674	NaN	Ele não é um mini cajon, é um shaker, ou seja,...	instrumentos_musicais
49675	NaN	So uma peça que veio rachado mas tudo bem rs	bebes
49676	NaN	Foi entregue antes do prazo.	eletrodomesticos_2
49677	NaN	Foi entregue somente 1. Quero saber do outro p...	informatica_acessorios
49678	NaN	Foi entregue somente 1. Quero saber do outro p...	informatica_acessorios

05 NLP

- The necessary columns are then cleaned from NaN values and duplicates in the review comment message.

	review_comment_title	review_comment_message
0	NaN	Não testei o produto ainda, mas ele veio corre...
1	NaN	Não testei o produto ainda, mas ele veio corre...
2	NaN	Não testei o produto ainda, mas ele veio corre...
3	Muito boa a loja	Muito bom o produto.
4	NaN	O produto foi exatamente o que eu esperava e e...

	review_comment_title	review_comment_message	product_category_name
0	Muito boa a loja	Muito bom o produto.	perfumaria
1	Nota dez	Obrigado pela atenção. Lojas Iannister perfeit...	perfumaria
2	Ótimo	Sempre vou comprar aqui pois é a melhor parabéns	papelaria
3	NÃO RECOMENDO!!!!	Até o momento não recebi o produto Protetor De...	bebes
4	Enganadores	Foi entregue um produto divergente do que eu h...	climatizacao

05 TEXT PREPROCESSING

- Next is the text preprocessing by doing some cleaning, removing stopwords, and tokenizing the comments.

	review_comment_message	review_comment_message_clean	review_comment_message_tokens
0	Muito bom o produto.	bom produto	[bom, produto]
1	Obrigado pela atenção. Lojas lannister perfeit...	obrigado atenção lojas lannister perfeito tudo	[obrigado, atenção, lojas, lannister, perfeito...
2	Sempre vou comprar aqui pois é a melhor parabéns	sempre vou comprar aqui pois melhor parabéns	[sempre, vou, comprar, aqui, pois, melhor, par...
3	Até o momento não recebi o produto Protetor De...	momento recebi produto protetor colchão berço ...	[momento, recebi, produto, protetor, colchão, ...
4	Foi entregue um produto divergente do que eu h...	entregue produto divergente havia comprado alé...	[entregue, produto, divergente, havia, comprad...

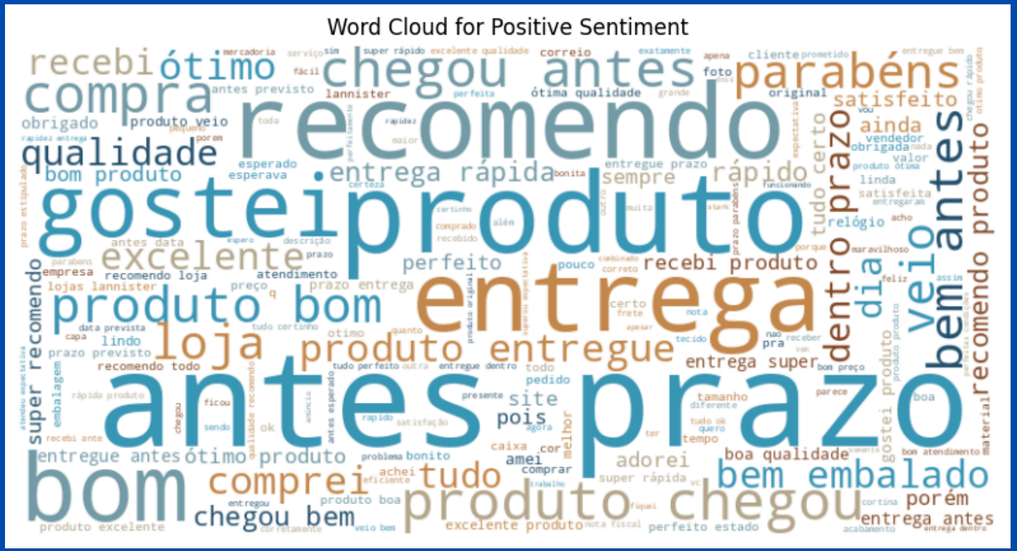
05 SENTIMENT ANALYSIS

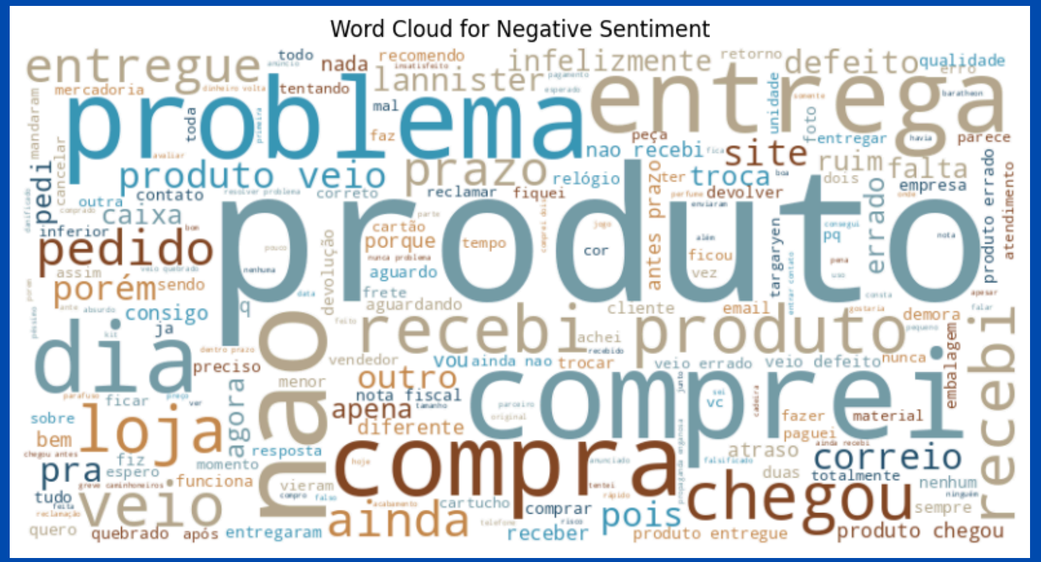
- LeIA library as a sentiment analysis tool for Brazilian Portuguese text.
- Sentiment Analysis classification scores:
  - Positive >= 0.05
  - Negative <= -0.05
  - Neutral is in between

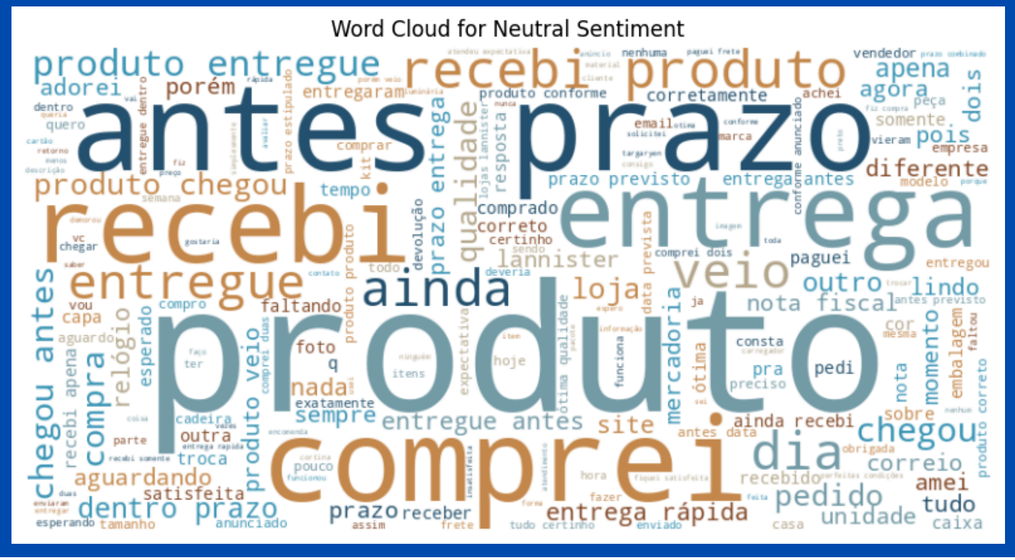
```python
def classify_sentiment(df, column_name='review_comment_message_clean'):
    # Vectorized function to get sentiment classification
    def get_sentiment_classification(text):
        scores = analyzer.polarity_scores(text)
        if scores['compound'] >= 0.05:
            return 'Positive'
        elif scores['compound'] <= -0.05:
            return 'Negative'
        else:
            return 'Neutral'
```

	review_comment_message_clean	review_comment_message_clean_sentiment
0	bom produto	Positive
1	obrigado atenção lojas lannister perfeito tudo	Positive
2	sempre vou comprar aqui pois melhor parabéns	Positive
3	momento recebi produto protetor colchão berço ...	Negative
4	entregue produto divergente havia comprado alé...	Neutral
5	produto veio bem lacrado todo original	Positive
6	prazo entrega excelente produto cuidadosamente...	Positive
7	recebi apenas 1 unidade solicitada deveriam 2 ...	Negative
8	entrega antes data marcada excelente	Positive
9	somente mala pequena bordo avião 360º viagem s...	Neutral
10	passei email solicitando suporte fazer produto...	Neutral

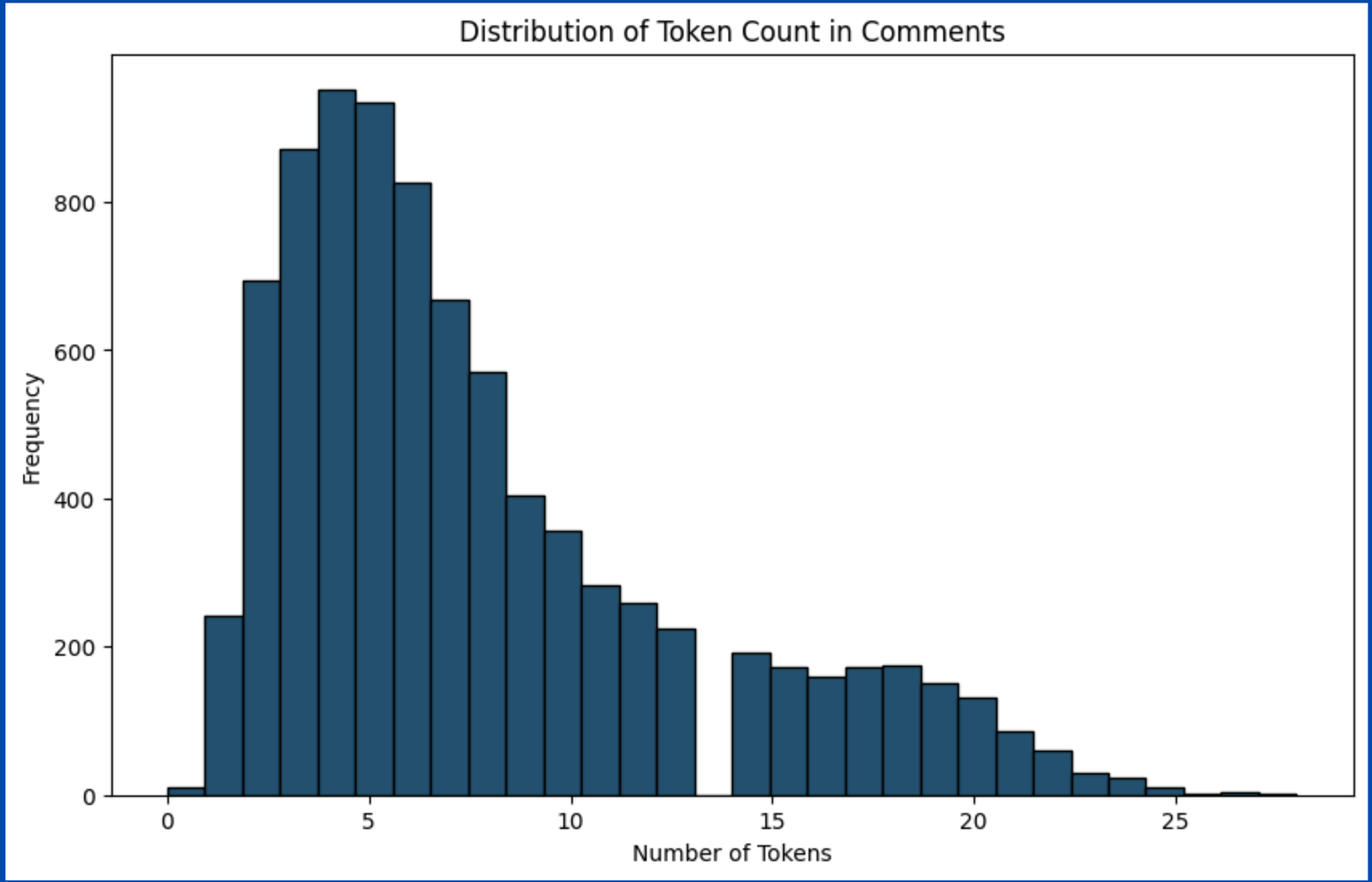
WORD CLOUD FOR EVERY SENTIMENT



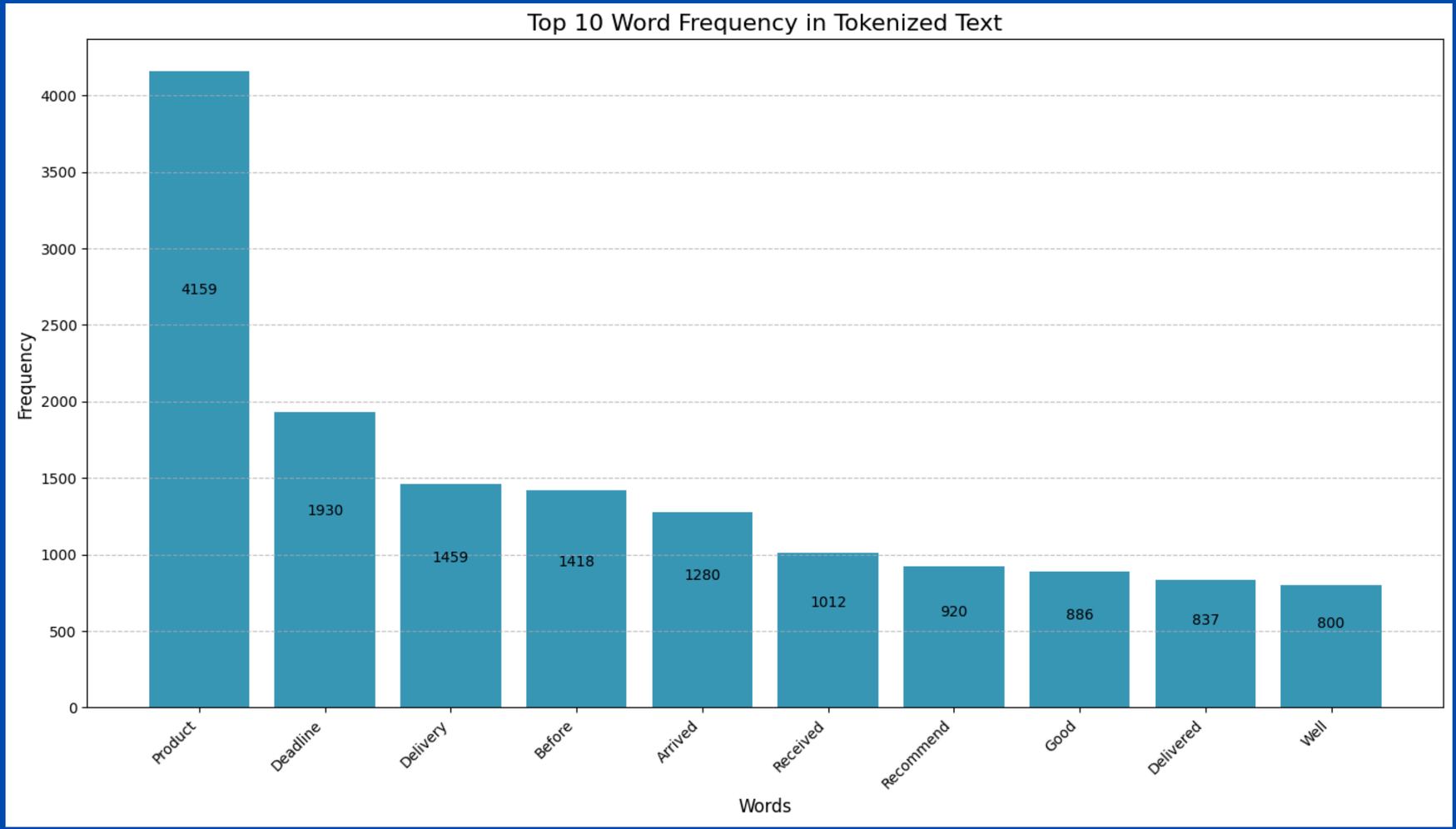




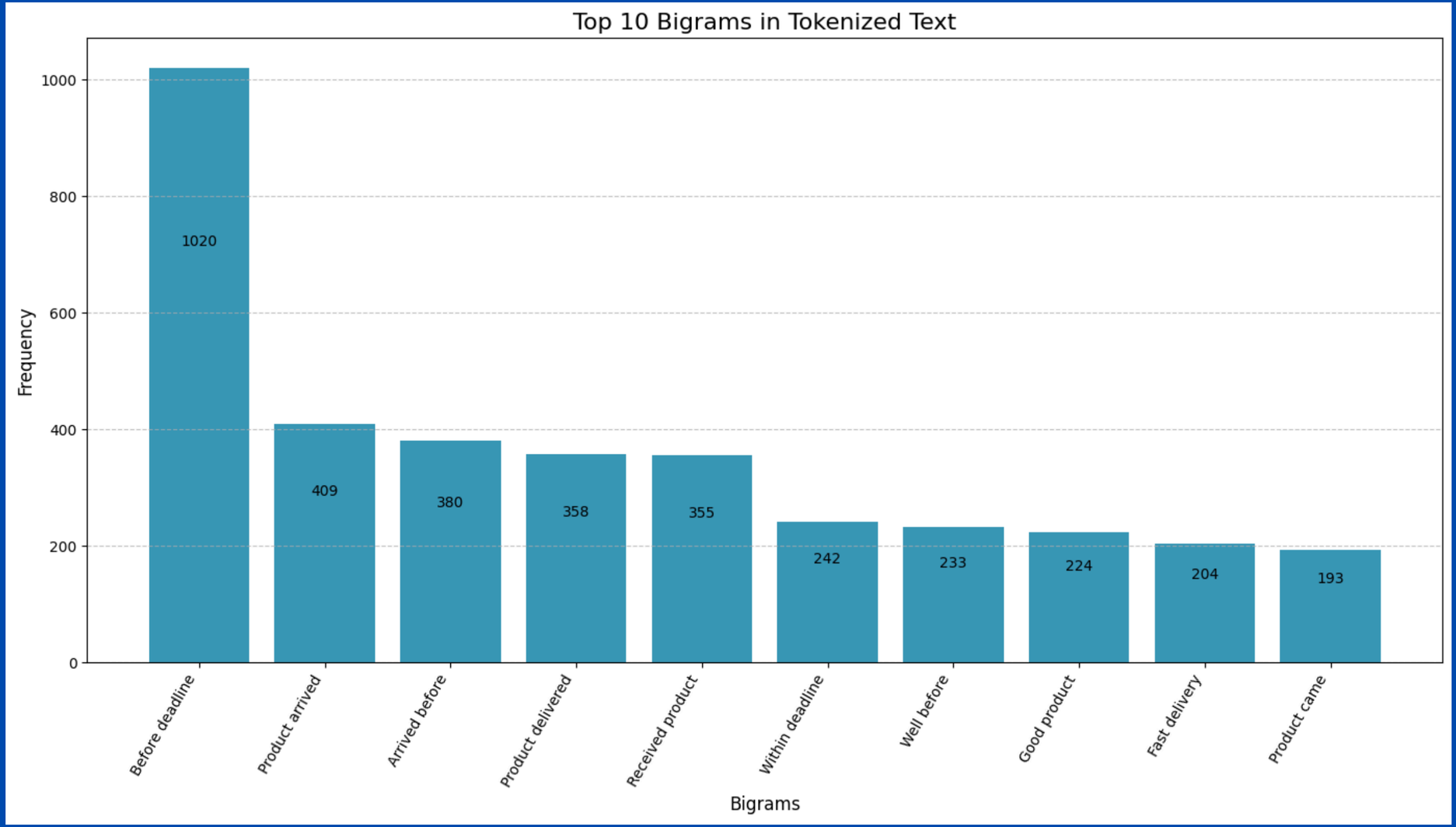
DISTRIBUTION OF TOKEN COUNT



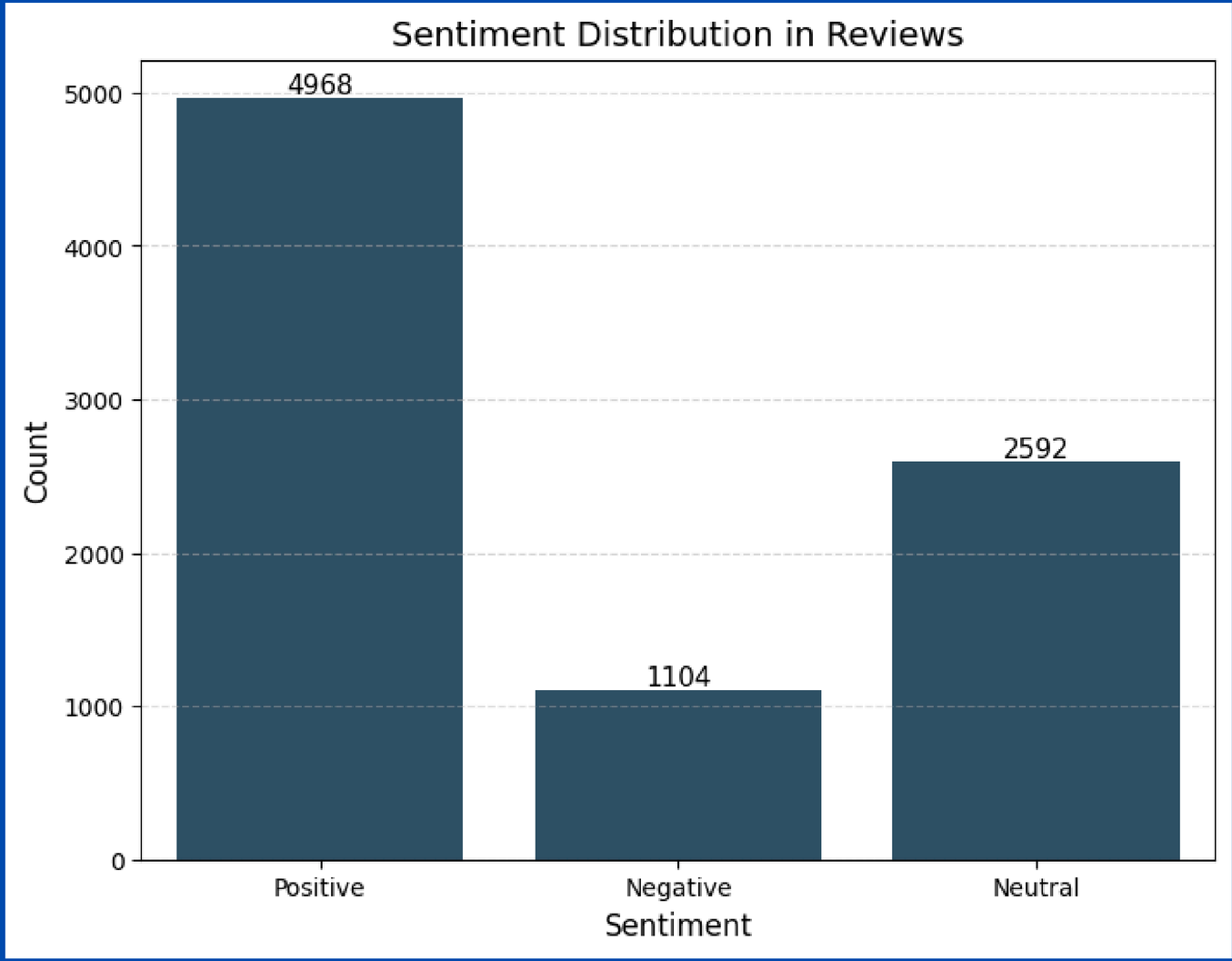
TOP 10 WORD FREQUENCY



TOP 10 BIGRAM FREQUENCY



SENTIMENT DISTRIBUTION



06

MACHINE LEARNING

Vectorization and Machine Learning

06 VECTORIZATION

- Before doing the machine learning the data went through some more preprocessing (Separate Features and Labels, Train Split, and Feature Encoding).
- Then, we do vectorization with Tfidf to convert raw text into numerical vectors for ML algorithms. (Highlight important terms, Efficient numerical representation, Balance between local and global relevance).

```python
# Vectorization with Tfidf
# Vectorize text using TF-IDF
vectorizer = TfidfVectorizer(max_features=5000) # Optional limit for the number of features

# Transform texts into TF-IDF vectors
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)

# Convert to dense arrays (if necessary for some models)
X_train_dense = X_train_tfidf.toarray()
X_test_dense = X_test_tfidf.toarray()

# Viewing
vectorizer
```

```
TfidfVectorizer
TfidfVectorizer(max_features=5000)
```

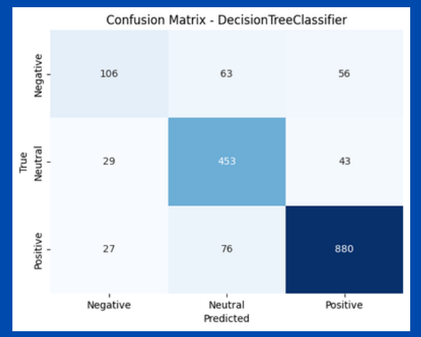
06 MACHINE LEARNING

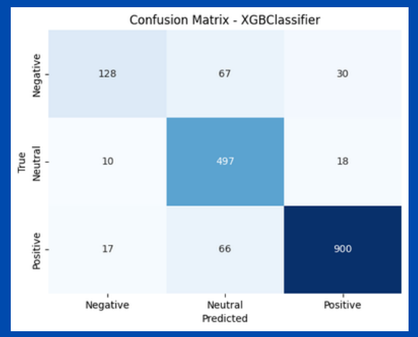
- Model Training and Accuracy to classify review sentiments and predict trends.
- II Models initialized.

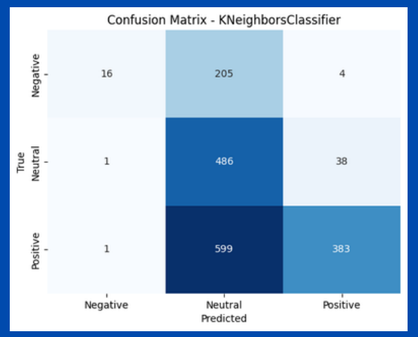
```
# Initialize models
models = {
    "GaussianNB": GaussianNB(),
    "DecisionTreeClassifier": DecisionTreeClassifier(random_state=42),
    "RandomForestClassifier": RandomForestClassifier(n_estimators=100, random_state=42),
    "LogisticRegression": LogisticRegression(random_state=50, max_iter=1000),
    "AdaBoostClassifier": AdaBoostClassifier(random_state=45),
    "XGBClassifier": XGBClassifier(random_state=42),
    "LGBMClassifier": LGBMClassifier(boosting_type='gbdt', bagging_fraction=0.9,
                                     learning_rate=0.05, feature_fraction=0.9, bagging_freq=50,
                                     verbosity=-1, verbose=50),
    "KNeighborsClassifier": KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2),
    "GradientBoostingClassifier": GradientBoostingClassifier(random_state=42),
    "SVC": SVC(kernel='linear', random_state=42, probability=True),
    "CatBoostClassifier": CatBoostClassifier(verbose=0, random_state=42)
}
```

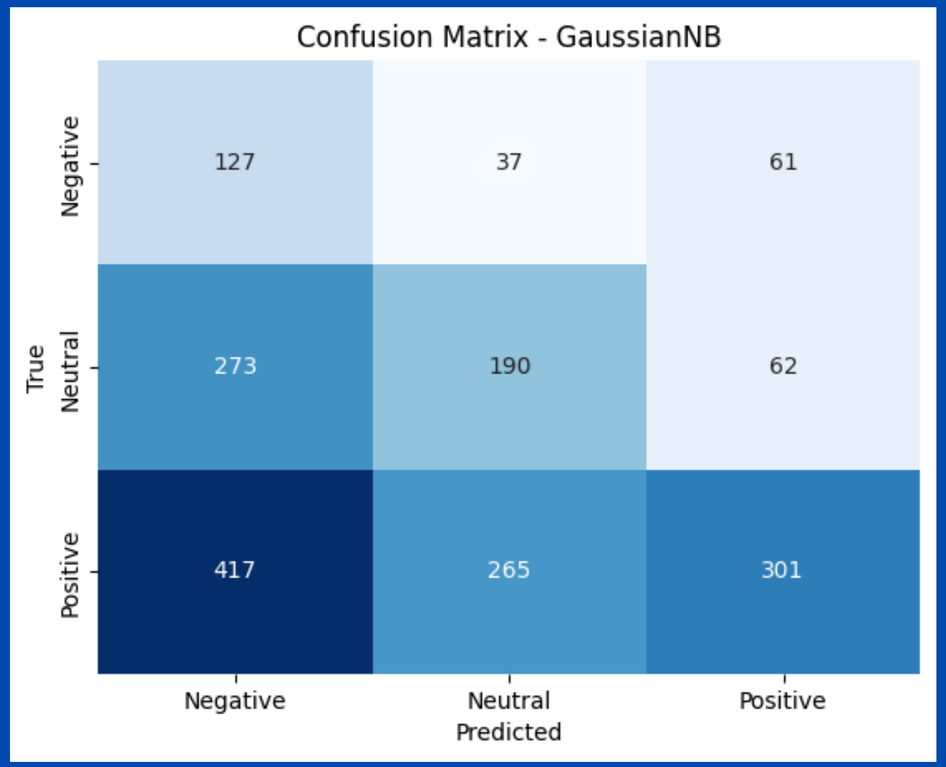
```
CPU times: user 24min 42s, sys: 14.2 s, total: 24min 56s
Wall time: 22min 34s
```

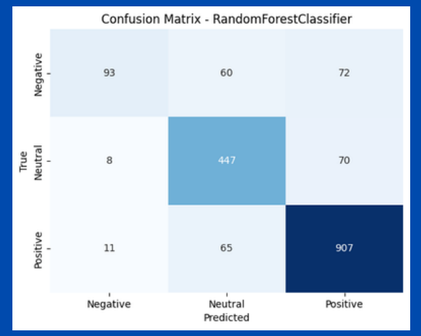
CONFUSION MATRIX

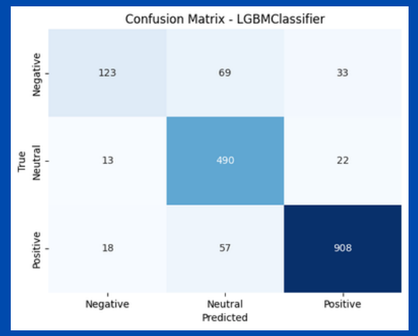


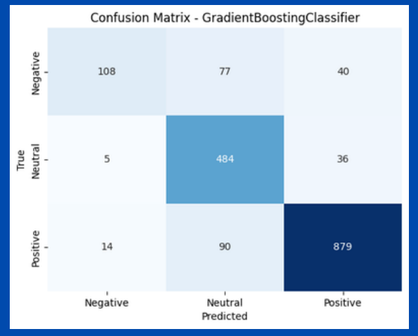


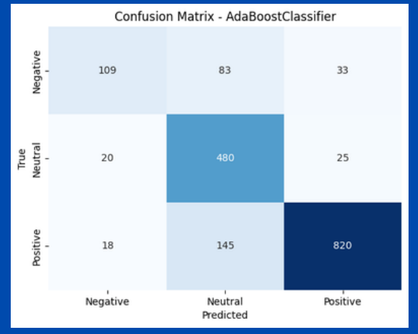


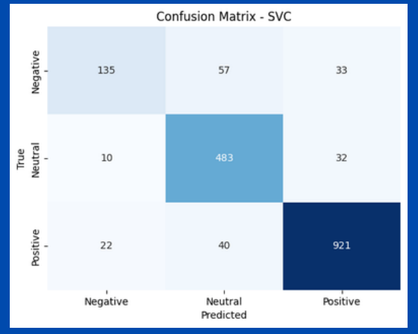


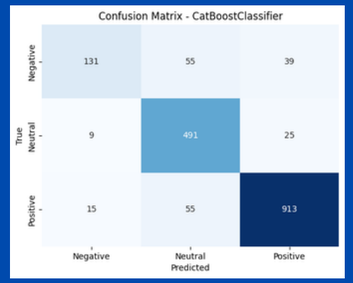




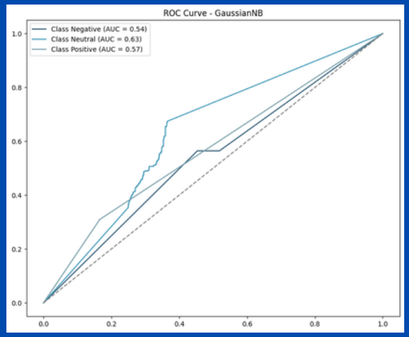


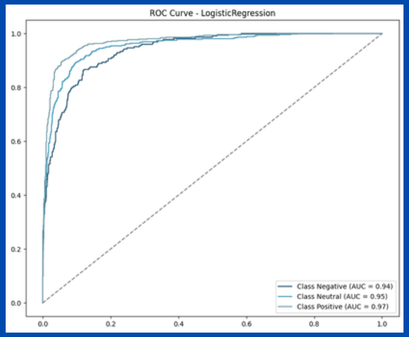


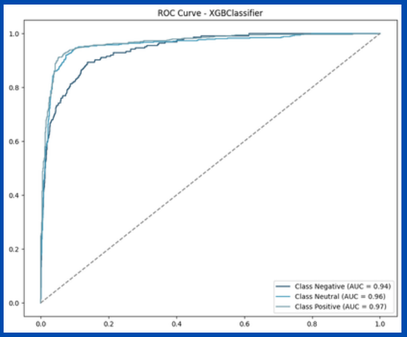


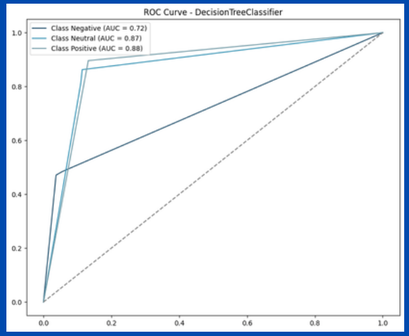


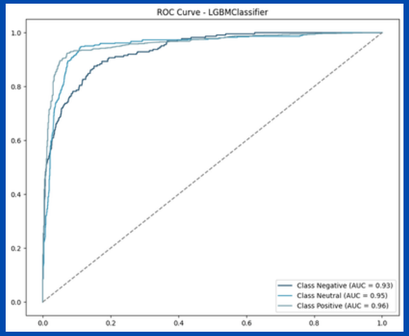
ROC CURVES

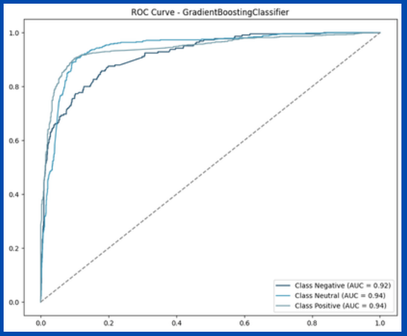

ROC Curve - GaussianNB
Class Negative (AUC = 0.54)
Class Neutral (AUC = 0.63)
Class Positive (AUC = 0.57)

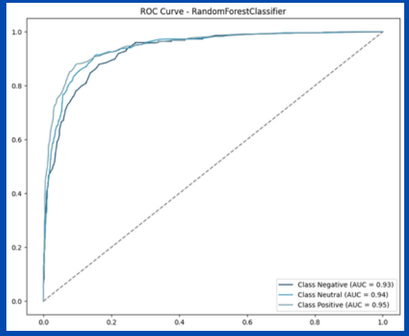

ROC Curve - LogisticRegression
Class Negative (AUC = 0.94)
Class Neutral (AUC = 0.95)
Class Positive (AUC = 0.97)

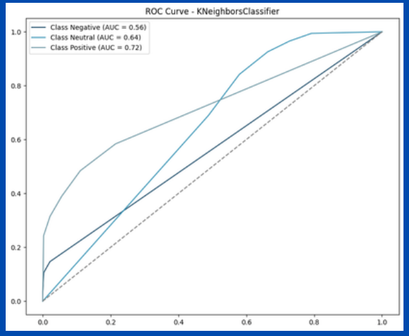

ROC Curve - XGBClassifier
Class Negative (AUC = 0.94)
Class Neutral (AUC = 0.96)
Class Positive (AUC = 0.97)

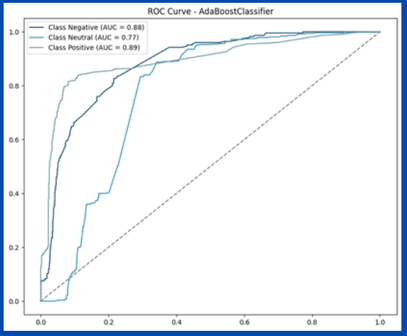

ROC Curve - DecisionTreeClassifier
Class Negative (AUC = 0.72)
Class Neutral (AUC = 0.87)
Class Positive (AUC = 0.88)

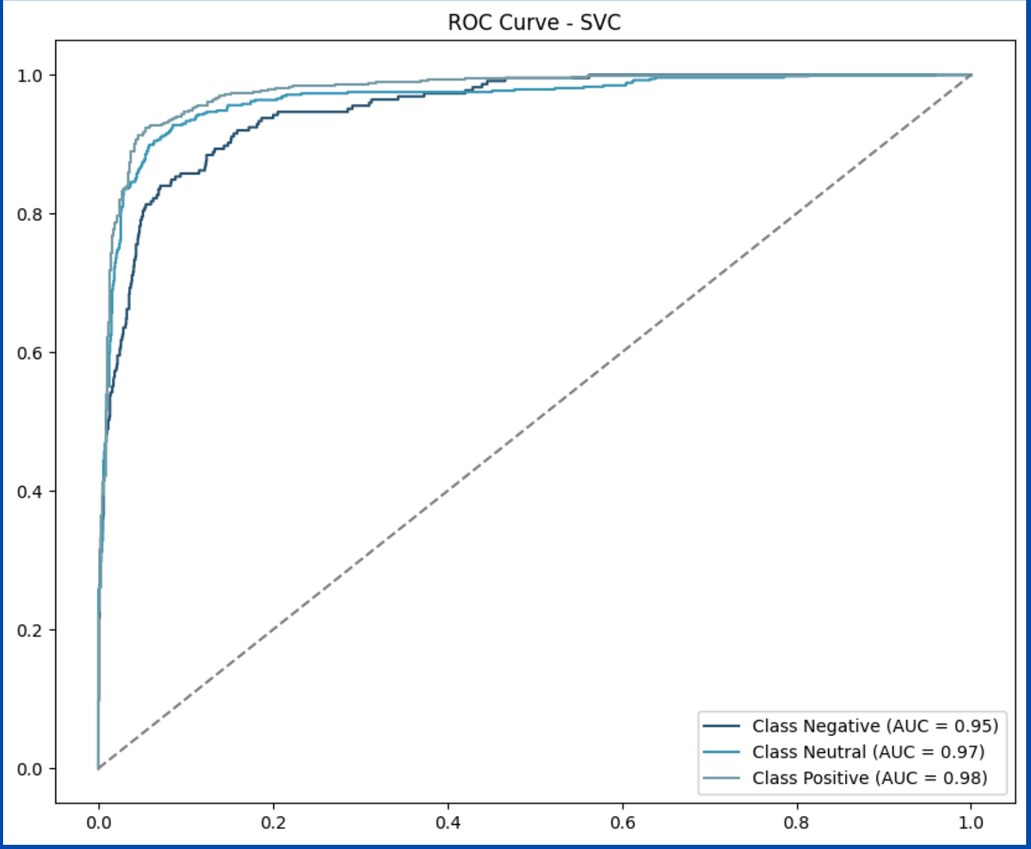

ROC Curve - LGBMClassifier
Class Negative (AUC = 0.93)
Class Neutral (AUC = 0.95)
Class Positive (AUC = 0.96)

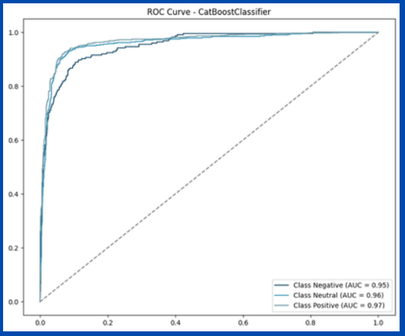

ROC Curve - GradientBoostingClassifier
Class Negative (AUC = 0.92)
Class Neutral (AUC = 0.94)
Class Positive (AUC = 0.94)


ROC Curve - RandomForestClassifier
Class Negative (AUC = 0.93)
Class Neutral (AUC = 0.94)
Class Positive (AUC = 0.95)

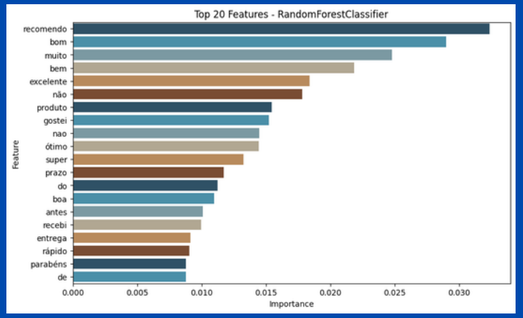

ROC Curve - KNeighborsClassifier
Class Negative (AUC = 0.56)
Class Neutral (AUC = 0.64)
Class Positive (AUC = 0.72)

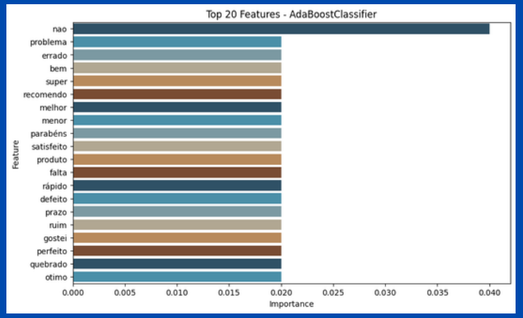

ROC Curve - AdaBoostClassifier
Class Negative (AUC = 0.88)
Class Neutral (AUC = 0.77)
Class Positive (AUC = 0.89)

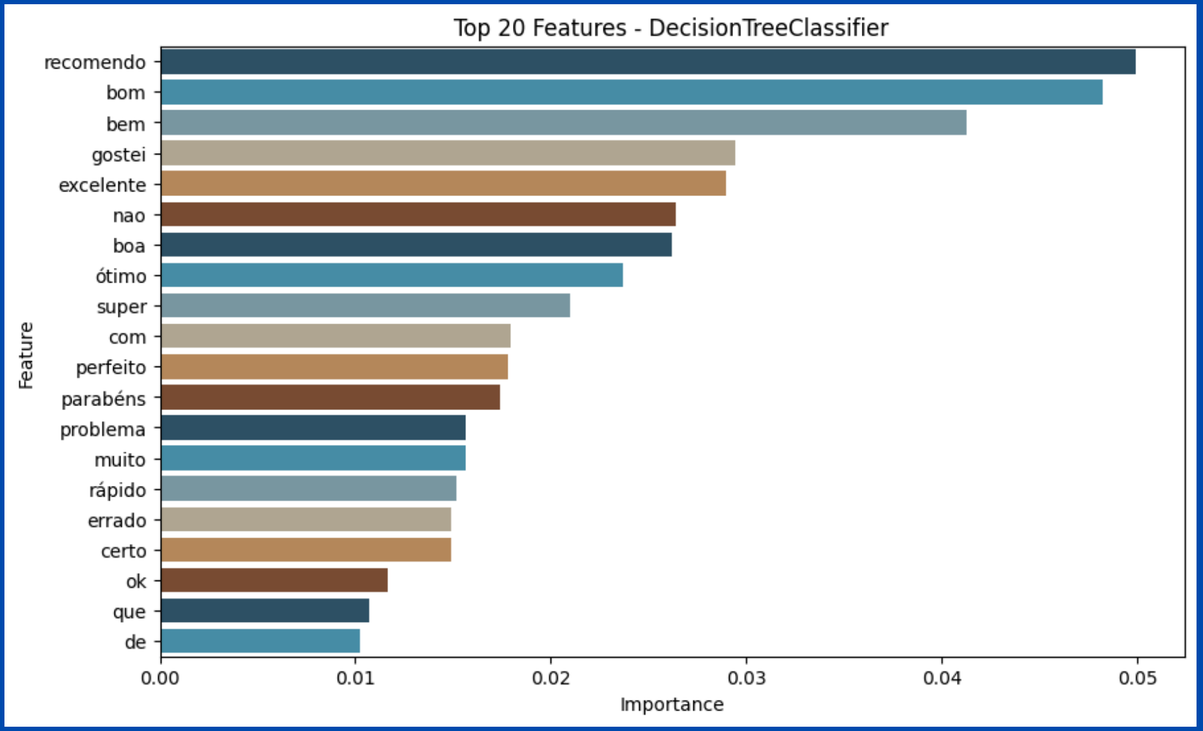

ROC Curve - SVC
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Class Negative (AUC = 0.95)
Class Neutral (AUC = 0.97)
Class Positive (AUC = 0.98)

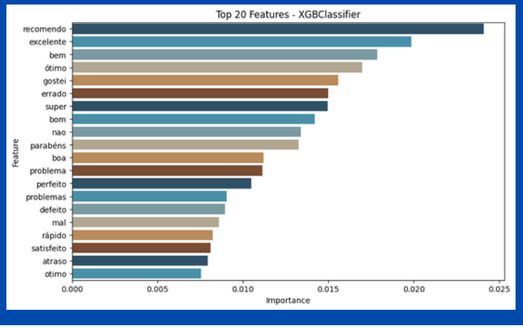

ROC Curve - CatBoostClassifier
Class Negative (AUC = 0.95)
Class Neutral (AUC = 0.96)
Class Positive (AUC = 0.97)

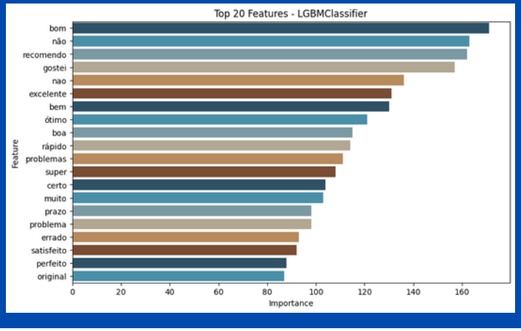
FEATURE IMPORTANCES

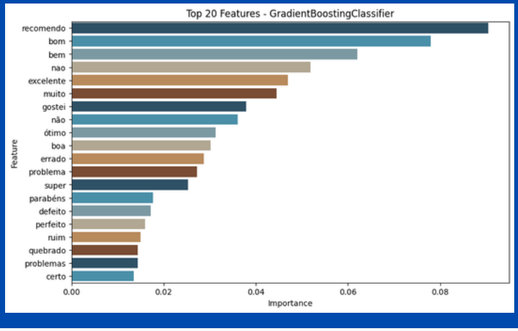


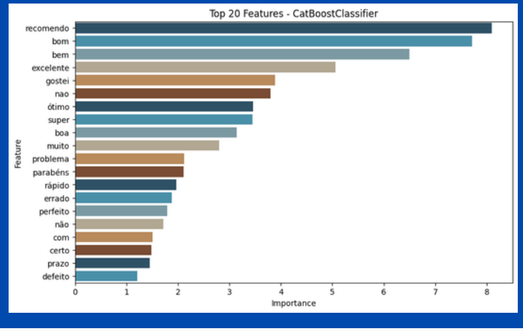












CLASSIFICATION REPORT

Classification Report for GaussianNB

	precision	recall	f1-score	support
Negative	0.16	0.56	0.24	225
Neutral	0.39	0.36	0.37	525
Positive	0.71	0.31	0.43	983
accuracy			0.36	1733
macro avg	0.42	0.41	0.35	1733
weighted avg	0.54	0.36	0.39	1733

Classification Report for DecisionTreeClassifier

	precision	recall	f1-score	support
Negative	0.65	0.47	0.55	225
Neutral	0.77	0.86	0.81	525
Positive	0.90	0.90	0.90	983
accuracy			0.83	1733
macro avg	0.77	0.74	0.75	1733
weighted avg	0.83	0.83	0.83	1733

Classification Report for RandomForestClassifier

	precision	recall	f1-score	support
Negative	0.83	0.41	0.55	225
Neutral	0.78	0.85	0.81	525
Positive	0.86	0.92	0.89	983
accuracy			0.83	1733
macro avg	0.83	0.73	0.75	1733
weighted avg	0.83	0.83	0.82	1733

Classification Report for LogisticRegression

	precision	recall	f1-score	support
Negative	0.80	0.51	0.62	225
Neutral	0.81	0.88	0.84	525
Positive	0.91	0.95	0.93	983
accuracy			0.87	1733
macro avg	0.84	0.78	0.80	1733
weighted avg	0.87	0.87	0.86	1733

Classification Report for AdaBoostClassifier

	precision	recall	f1-score	support
Negative	0.74	0.48	0.59	225
Neutral	0.68	0.91	0.78	525
Positive	0.93	0.83	0.88	983
accuracy			0.81	1733
macro avg	0.78	0.74	0.75	1733
weighted avg	0.83	0.81	0.81	1733

Classification Report for XGBClassifier

	precision	recall	f1-score	support
Negative	0.83	0.57	0.67	225
Neutral	0.79	0.95	0.86	525
Positive	0.95	0.92	0.93	983
accuracy			0.88	1733
macro avg	0.85	0.81	0.82	1733
weighted avg	0.88	0.88	0.88	1733

Classification Report for LGBMClassifier

	precision	recall	f1-score	support
Negative	0.80	0.55	0.65	225
Neutral	0.80	0.93	0.86	525
Positive	0.94	0.92	0.93	983
accuracy			0.88	1733
macro avg	0.85	0.80	0.81	1733
weighted avg	0.88	0.88	0.87	1733

Classification Report for KNeighborsClassifier

	precision	recall	f1-score	support
Negative	0.89	0.07	0.13	225
Neutral	0.38	0.93	0.54	525
Positive	0.90	0.39	0.54	983
accuracy			0.51	1733
macro avg	0.72	0.46	0.40	1733
weighted avg	0.74	0.51	0.49	1733

Classification Report for GradientBoostingClassifier

	precision	recall	f1-score	support
Negative	0.85	0.48	0.61	225
Neutral	0.74	0.92	0.82	525
Positive	0.92	0.89	0.91	983
accuracy			0.85	1733
macro avg	0.84	0.77	0.78	1733
weighted avg	0.86	0.85	0.84	1733

Classification Report for SVC

	precision	recall	f1-score	support
Negative	0.81	0.60	0.69	225
Neutral	0.83	0.92	0.87	525
Positive	0.93	0.94	0.94	983
accuracy			0.89	1733
macro avg	0.86	0.82	0.83	1733
weighted avg	0.89	0.89	0.88	1733

Classification Report for CatBoostClassifier

	precision	recall	f1-score	support
Negative	0.85	0.58	0.69	225
Neutral	0.82	0.94	0.87	525
Positive	0.93	0.93	0.93	983
accuracy			0.89	1733
macro avg	0.87	0.82	0.83	1733
weighted avg	0.89	0.89	0.88	1733

PERFORMANCE DATA COLLECTION AND HIGHLIGHTING

	Model	Accuracy	Precision	Recall	F1-Score
0	GaussianNB	0.356607	0.539848	0.356607	0.387535
1	DecisionTreeClassifier	0.830352	0.826630	0.830352	0.825665
2	RandomForestClassifier	0.834968	0.834989	0.834968	0.824912
3	LogisticRegression	0.868436	0.865989	0.868436	0.862371
4	AdaBoostClassifier	0.813041	0.831410	0.813041	0.811817
5	XGBClassifier	0.879977	0.884709	0.879977	0.876924
6	LGBMClassifier	0.877669	0.879503	0.877669	0.873799
7	KNeighborsClassifier	0.510675	0.740708	0.510675	0.487924
8	GradientBoostingClassifier	0.848817	0.857722	0.848817	0.843572
9	SVC	0.888055	0.887064	0.888055	0.884899
10	CatBoostClassifier	0.885747	0.887293	0.885747	0.882162

07

HAPPINESS ANALYSIS

HAPPINESS ANALYSIS

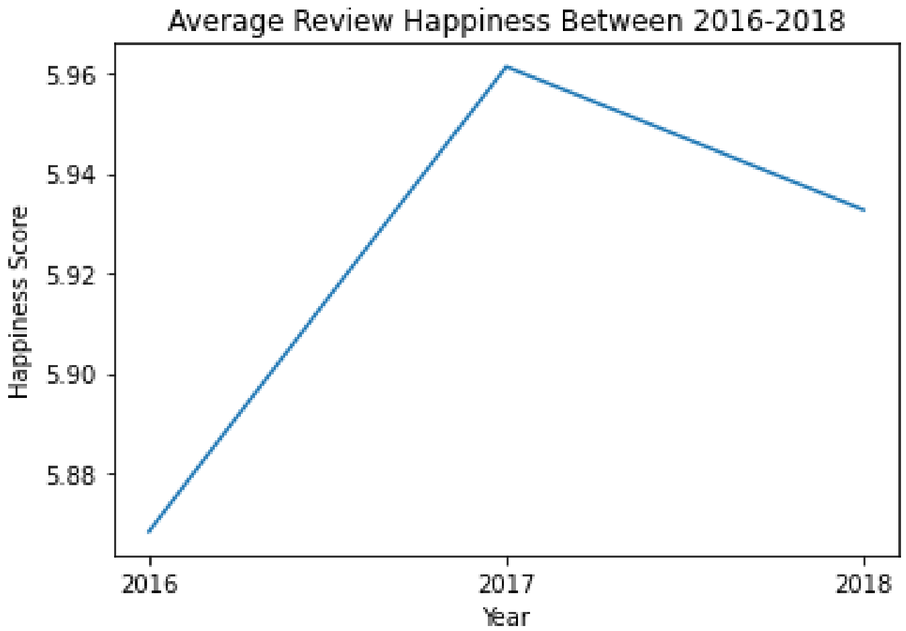
HAPPINESS ANALYSIS PREPROCESSING

- Translated reviews are further processed to obtain their valence score
- Happiness score are obtained based on the ANEW word list
- Further information such as “Valence” word count and proportion are also stored

1	Word	valence	arousal	dominance
2	aardvark	6.26	2.41	4.27
3	abalone	5.3	2.65	4.95
4	abandon	2.84	3.73	3.32
5	abandonment	2.63	4.95	2.64
6	abbey	5.85	2.2	5
7	abdomen	5.43	3.68	5.15
8	abdominal	4.48	3.5	5.32
9	abduct	2.42	5.9	2.75
10	abduction	2.05	5.33	3.02
11	abide	5.52	3.26	5.33
12	abiding	5.57	3.59	6.6
13	ability	7	4.85	6.55
14	abject	4	3.94	4.35
15	ablaze	5.15	6.75	4.58

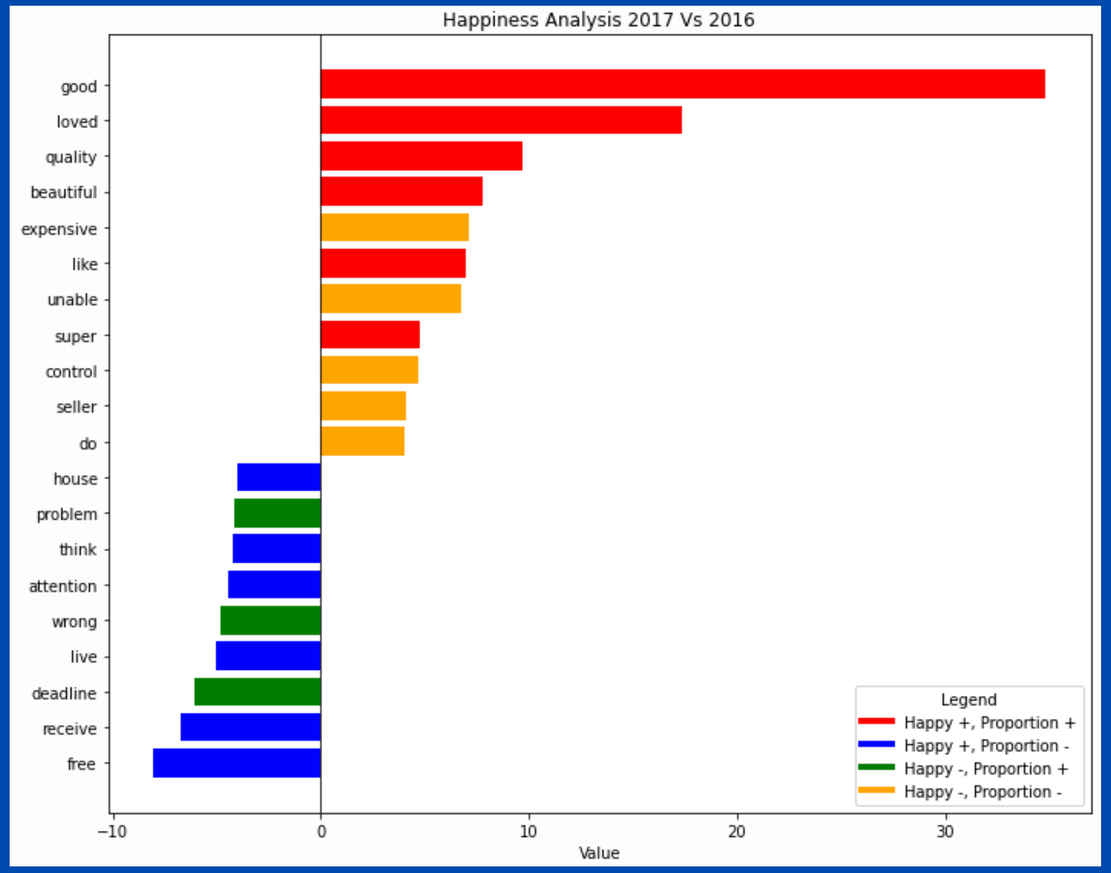
AVERAGE REVIEW HAPPINESS

- The average happiness score for each year is calculated from the Affective Norms for English Words (ANEW) valence scores.
- It is observed that there was an increase in happiness from 2016 to 2017
- This number dipped slightly at the end of the period. To investigate which words has the highest influence on the happiness levels, more analysis was done.



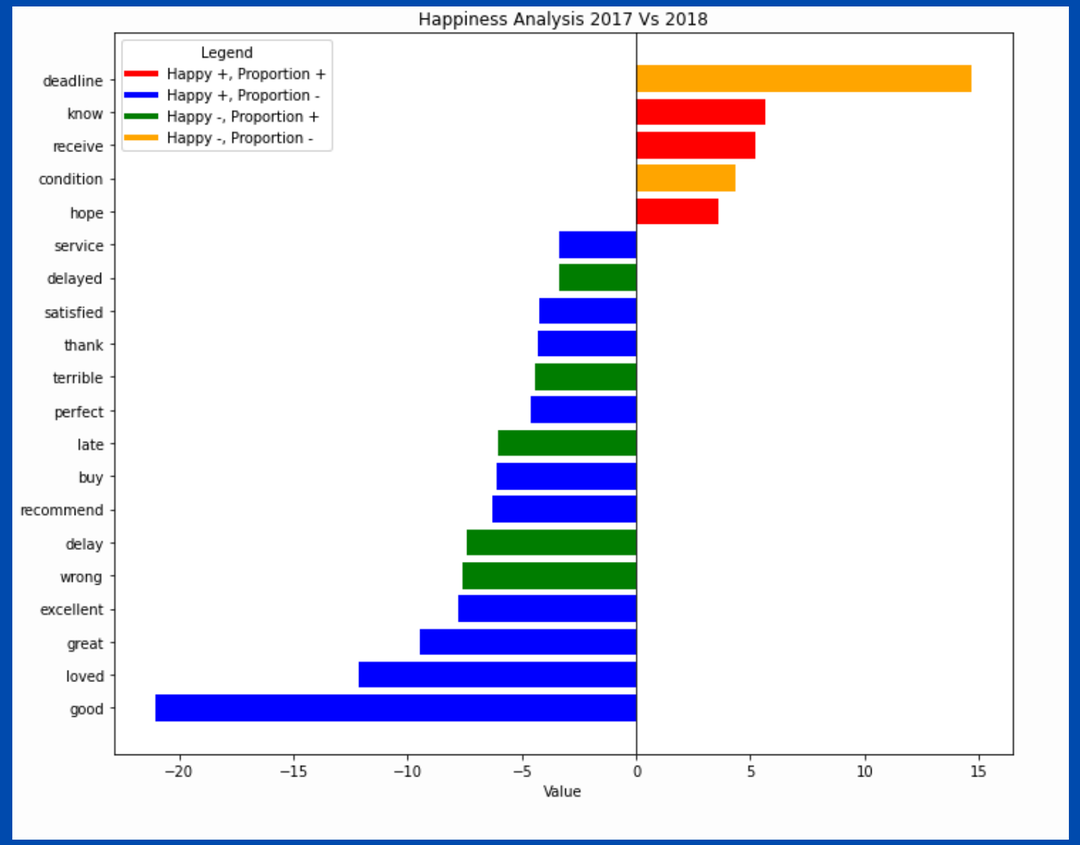
HAPPINESS ANALYSIS 2017 VS 2016

- The reference group is 2016 review and the comparison group is 2017
- From the top 20 words, 11 are positive and 9 are negative towards the happiness score
- Positive adjectives such as ‘good’, ‘loved’, ‘quality’, and ‘beautiful’ are used more often
- Interestingly, ‘expensive’ was being used less in 2017
- ‘Deadline’ which is used more often in 2017 may indicate slow deliveries or packaging time



HAPPINESS ANALYSIS 2018 VS 2017

- The word ‘deadline’ was used less often however words such as ‘late’ and ‘delay’, appears more often indicating shipment problem may still occur
- Words that affect happiness negatively are mostly positive words that are used less often, such as ‘good’, ‘great’, and ‘excellent’
- This may imply customers’ expectations had increased therefore, incentives might be introduced to increase overall satisfaction.



08 CONCLUSION

01

The analysis identified key trends in sales, product preferences, payment methods, and delivery times that significantly influence Brazilian e-commerce performance.

02

Changes in customer sentiment across years highlighted the impact of service quality, including delivery efficiency and product satisfaction, on overall happiness.

03

The analysis may lack contextual understanding of consumer behavior within Brazilian e-commerce. Further background information maybe needed to extract a complete insight

THANK YOU!

ANY QUESTION?