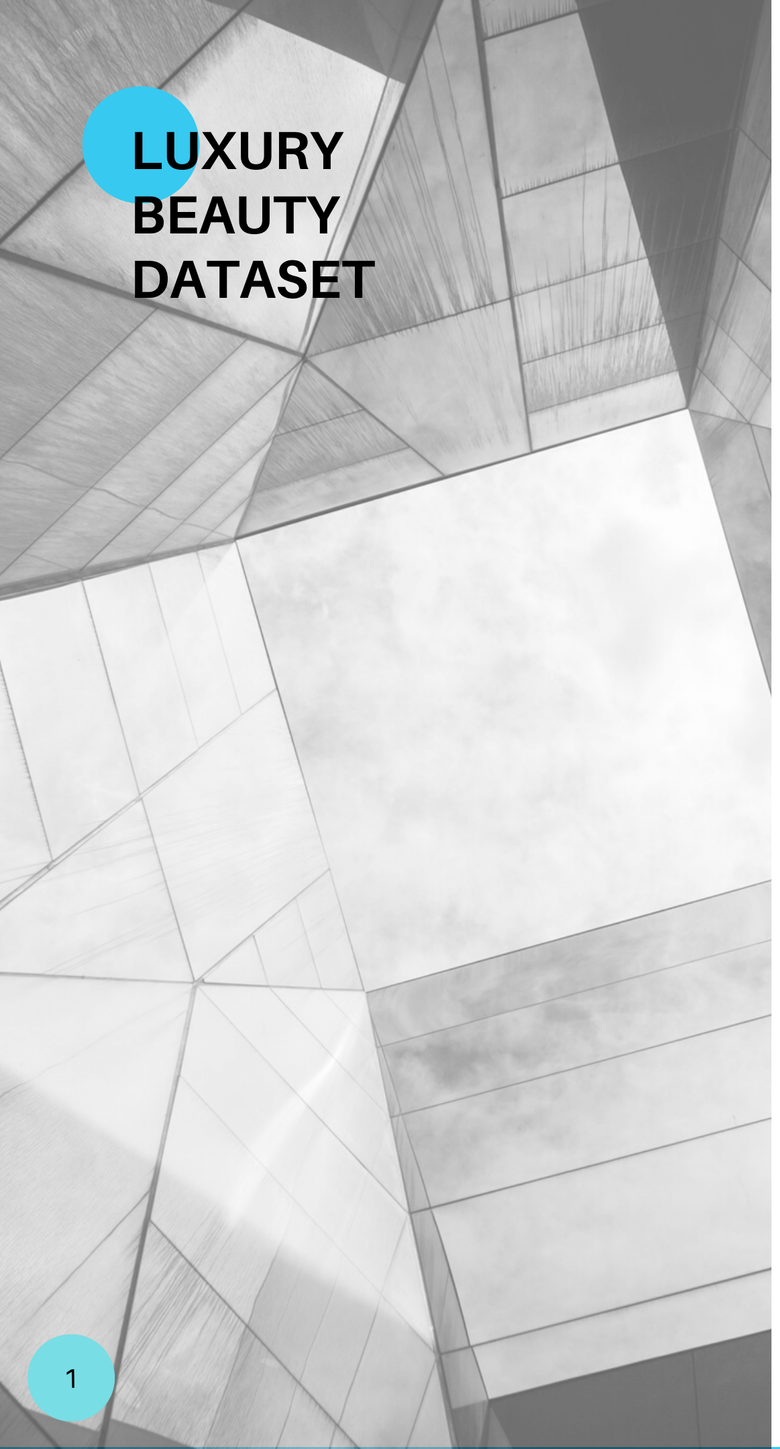


LUXURY
BEAUTY
DATASET

AMAZON PRODUCT REVIEW ANALYSIS

GUIDED BY : DR. AMIT KUMAR

PRESENTED BY :

UDIT SINGH CHANDEL
SHIKHAR NIGAM
VIPUL SHARMA
BHARATH KUMAR R

AGENDA

- I — INTRODUCTION
- O — OBJECTIVES
- D — DESCRIPTIVE ANALYSIS
- D — DATA DICTIONARY
- **P** — PLUG AND PLAY DATASET
- E — EXPLORATORY DATA ANALYSIS
- S — SENTIMENT ANALYSIS
- P — PRODUCT SEGMENTATION
- C — CUSTOMER SEGMENTATION
- T — TIME SERIES SENTIMENT FORECASTING

INTRODUCTION

AMAZON

- Amazon is an American multinational technology company focusing on e-commerce, cloud computing, online advertising, digital streaming, and artificial intelligence.
- Amazon's global headquarters are in more than 40 owned and leased buildings spread across Seattle's adjacent South Lake Unions.
- Jeff Bezos, the CFO of amazon has different roles in the stock market worldwide and has been an inspiration to many.
- Initially, it was an online marketplace for books, later the amazon buyer Sales approach warmed up so well, the company surpassed all it's competitors in a very short time.

OBJECTIVES

1. TO CLASSIFY THE PRODUCT REVIEW BASED ON REVIEW SENTIMENTS.
2. PRODUCT SEGMENTATION FOR RECOMMENDING PRODUCTS TO THE CUSTOMERS BASED ON THEIR REVIEWS.
3. CUSTOMER SEGMENTATION TO IDENTIFY AND PREVENT CHURNING OUT CUSTOMERS BASED ON REVIEWS GIVEN.
4. FORECAST OF CUSTOMER SENTIMENTS TO ESTIMATE PRODUCT DEMANDS IN THE FUTURE.
5. TO IMPORT KEY INSIGHTS FROM THE AVAILABLE DATASETS TO MAKE INFORMED BUSINESS DECISIONS.

DESCRIPTIVE ANALYSIS

Datasets used

- **5 Core of Luxury Beauty**

This particular dataset carries information mainly about Product reviews and ratings.

- **Metadata of Luxury Beauty**

To support our Analysis and further, columns from Metadata are essential aspects when it comes to analysis of different forms.

- **Ratings of Luxury Beauty**

The dataset beholds information about customer ratings and customer reviewer ID which can support our statements on Customer Sentiments.

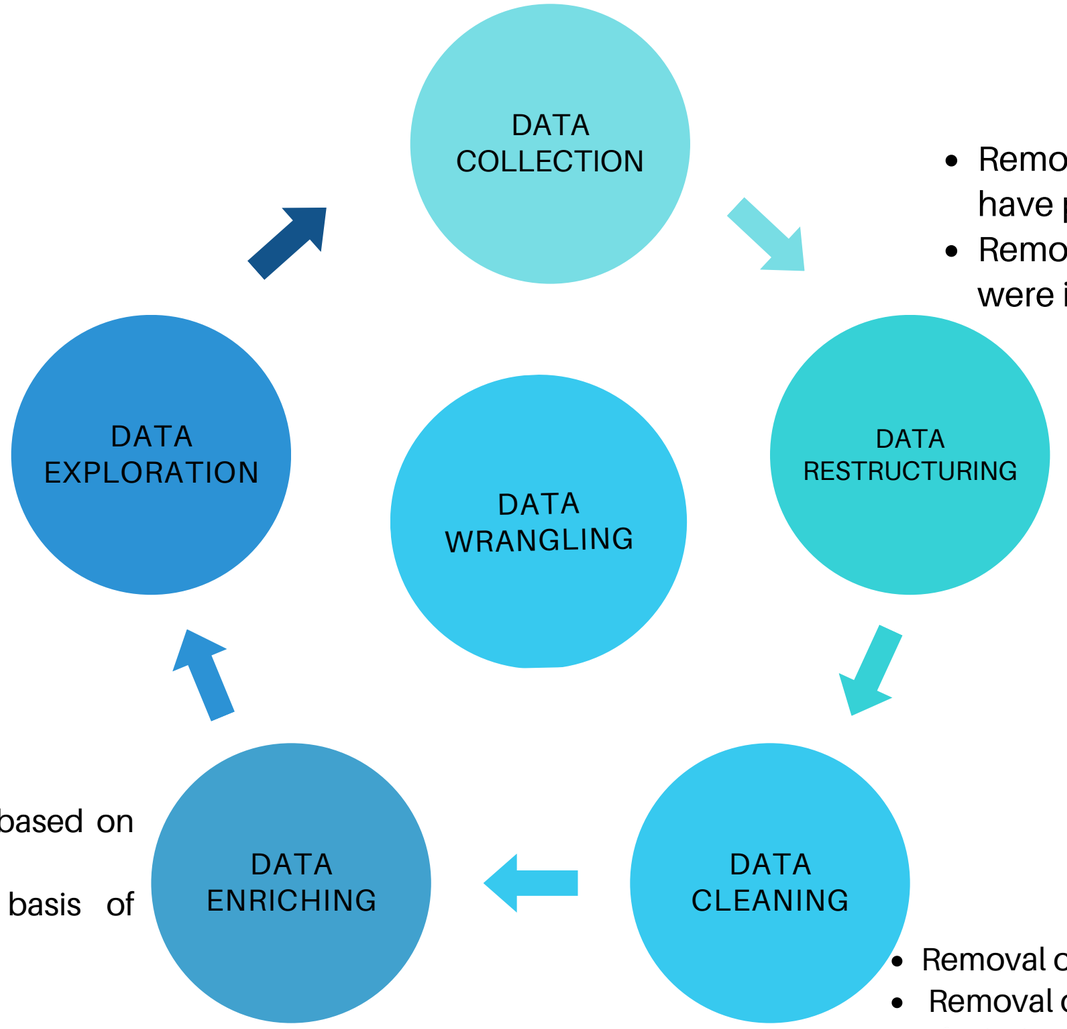
```
Shape of our Luxury_MetaData is: (12299, 19)
Shape of our Luxury_RatingData is: (574628, 4)
Shape of our Luxury_CoreDatais: (34278, 12)
```

DATA DICTIONARY

RATING	Contains custoer ratings for a particular product
VERIFIED	Depicts if the Reviewer is verified customer or not
REVIEW TIME	It contains the information of dateon which Review is published
PRODUCT ID	Is also known as ASIN which stands for Unique amazon standard ID number
REVIEWER NAME	Depicts the name of the Reviewer
REVIEW TEXT	It contains Reviews posted by the Customers
SUMMARY	Summarized version of Review Text
RANK NO	It contains the information of how the products are ranked
MAIN CATEGORY	It contains the information of Main category
PRICE	It contains the price information of the products

DATA WRANGLING

- Gathering data of the amazon products along with their brand information, customer information, category information etc. from the provided data files.



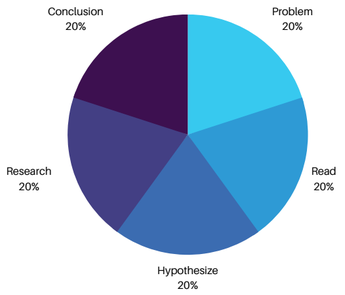
- Removal of those fields which have plenty of missing values.
- Removal of those fields which were insignificant to our analysis

- Removal of noise in our data.
- Removal of those records which had plenty of missing values for essential fields.

- Missing value imputation based on meaningful categorization.
- Merging our tables on basis of matching fields.

- Save our final data frame as csv file for future use

EXPLORATORY DATA ANALYSIS



LINK TO TABLEAU DASHBOARD

https://public.tableau.com/app/profile/shikhar.nigam/viz/finalprojecteda/Dashboard3

```
Shape of our Final dataset: (23349, 16)
Unique Products in our Final dataset: 1212
Unique Reviewers in our Final dataset: 3803
```

- As a data analyst to achieve the dataset where we could perform our analysis, the raw dataset has been well observed and preprocessed.
- Initially, 5 CORE, Metadata, and Reviews have been considered and feature extraction has been based on the information each column had to offer.
- Then to impute the missing values, careful steps are taken by grouping data based on product ids.
- Eventually, we end up with almost 23,350 rows of data and 16 valuable columns.
- We have 1212 unique products in our final dataset of Luxury beauty.

Sentiment Analysis

Sentiment Analysis is an approach that identifies the emotional tone behind any particular text or reviews

1

TEXT-PREPROCESSING

- PROBLEMREMOVE ALL NON-LETTERS AND NON-SPACES EXCEPT FOR HYPHENS AND DIGITS
- REMOVE ALL NUMBERS EXCEPT THOSE ATTACHED TO A WORD
- REMOVE ALL HYPHENS EXCEPT BETWEEN TWO WORDS
- REMOVE ALL URLS
- REMOVE MULTIPLE SPACES
- REMOVE STOP WORDS
- LEMMITIZATION

2

TEXT VECTORIZATION

- **TF-IDF**
- **COUNTVECTORIZER USING N-GRAMS(1,2)**

3

CREATING THE MODELS

- MULTINOMIAL NAIVE BAYES
- LOGISTIC REGRESSION WITH ONE VS REST CLASSIFIER
- DECISION TREE
- RANDOM FOREST

4

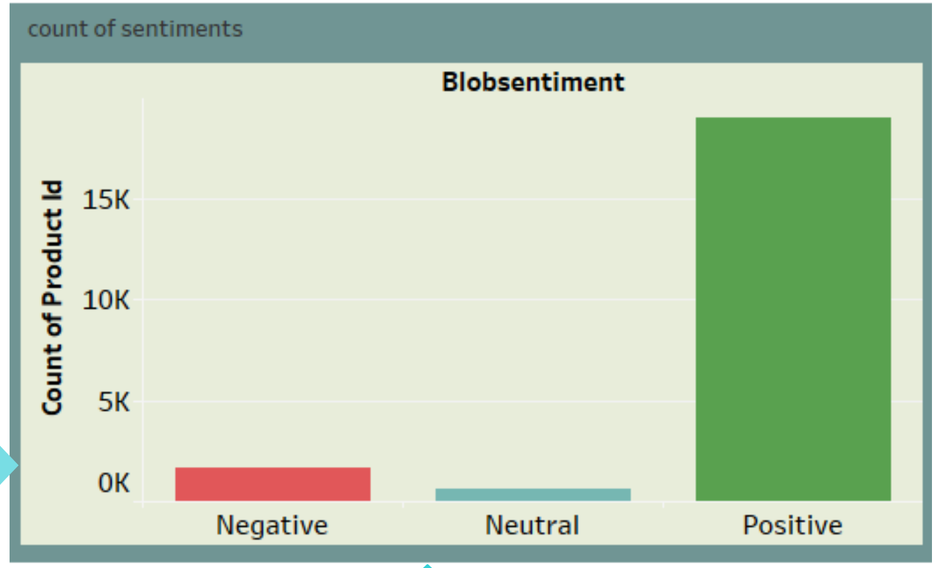
CONCLUSION

Interpret the results.

TEXT PREPROCESSING

We have used TextBlob library to find the sentiments of the reviews. TextBlob is a python library for Natural Language Processing (NLP).

final_review	polarity	blobsentiment
handcream beautiful fragrance not stay protect...	0.483333	Positive
no evidence product improvement similarly pric...	-0.083333	Negative
wonderful hand lotion seriously dry skin stay ...	0.346250	Positive
good hand		



Since or data is imbalanced so we need to do text augmentation

We have used the Wordnet library to create some artificial reviews of the negative and neutral reviews, which are under-represented by using synonyms of those reviews.

After removing the bias in the data, this the proportion of different sentiments which we will be using to create the classifier model

Negative	0.548735
Positive	0.262076
Neutral	0.189189

COMPARISON OF DIFFERENT MODELS

MODEL WITH TF-IDF VECTORIZER	ACCURACY	FI-SCORE
MULTINOMIAL NAIVE BAYES	0.9038	0.8707
LOGISTIC REGRESSION (ONE V/S REST CLASSIFIER)	**0.9299**	**0.933**
DECISION TREE CLASSIFIER	0.7668	0.8122
RANDOM FOREST CLASSIFIER	0.8857	0.853

MODEL WITH COUNT VECTORIZER N-GRAMS(1,2)	ACCURACY	FI-SCORE
MULTINOMIAL NAIVE BAYES	0.915	0.9005
LOGISTIC REGRESSION (ONE V/S REST CLASSIFIER)	0.927	0.912
DECISION TREE CLASSIFIER	0.761	0.808
RANDOM FOREST CLASSIFIER	0.893	0.847

AS WE CAN SEE LOGISTIC REGRESSION WITH ONE VS REST CLASSIFIER OFFERS BEST ACCURACY AND F-1 SCORES. SO WE WILL BE USING THIS MODEL FOR CLASIFYING THE REVIEWS

Prediction on unseen data

	product_id	title	final_review	blobsentiment	prediction_sentiment
0	B0000530HU	Aqua Velva After Shave, Classic Ice Blue, 7 Ounce	hey aqua velva man absolutely love stuff year ...	Positive	Positive
1	B00006L9LC	Citre Shine Moisture Burst Shampoo - 16 fl oz	buy hope help rid dandruff begin winter reason...	Negative	Negative
2	B00021DJ32	NARS Blush, Taj Mahal	truly high quality blush go smoothly need touc...	Positive	Positive
3	B0002JHI1I	Avalon Organics Wrinkle Therapy CoQ10 Cleansin...	coq10 essential healthy youthful skin naturall...	Positive	Positive
4	B0006O10P4	ZUM Zum Bar Anise Lavender, 3 Ounce	decide splurge oz bar buy wonderfully strong s...	Positive	Positive

- Above table shows the output of Logistic Regression model on unknown review text
- Thus we are able to achieve the very first objective of making a well automated review system.

The Stages of Product Segmentation

1 **OBJECTIVE**

PRODUCT SEGMENTATION

2 **REASON**

TO CLUSTER THE PRODUCTS FOR RECOMMENDING BEST PRODUCTS TO OUR CUSTOMERS AND IMPROVING THE BAD PRODUCTS FOR TIME AHEAD.

3 **APPROACH**

Clustering has been performed on the basis of Product ratings and Sentiments

4 **MODELS USED**

The clustering approaches we have performed here are of 2 types:

1. K-means Algorithm
2. Agglomerative Clustering

5 **RESULT**

We have categorized products into :

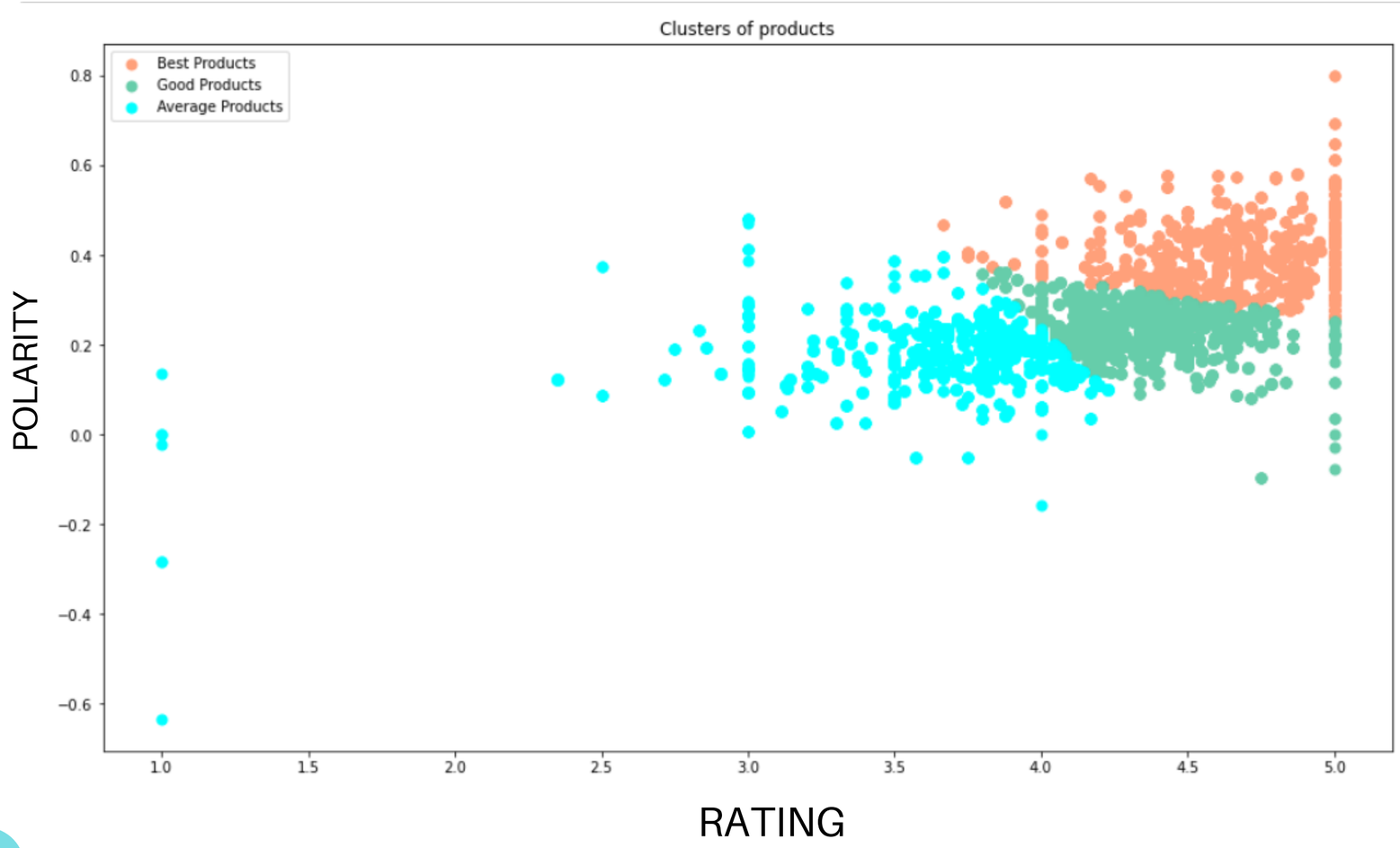
1. Best products
2. Good products
3. Average products

COMPARISON OF OUR MODELS

We have used the Silhouette score as the performance metric for comparing the model's effectiveness.

	model	silhoutte_score
0	K-means	0.401922
1	Agglomerative clustering	0.377144

As silhoutte score of K-Means algorithm is better we will use this model for our product segmentation

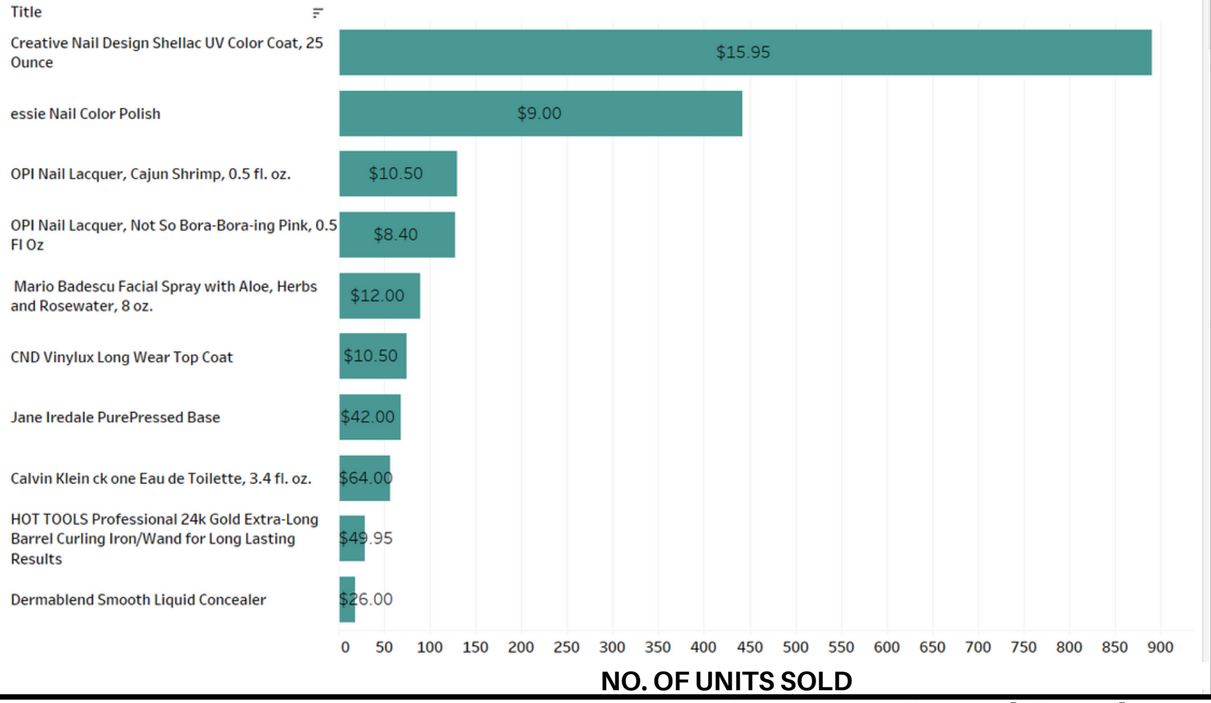


This graph is showing the categorization of products into :

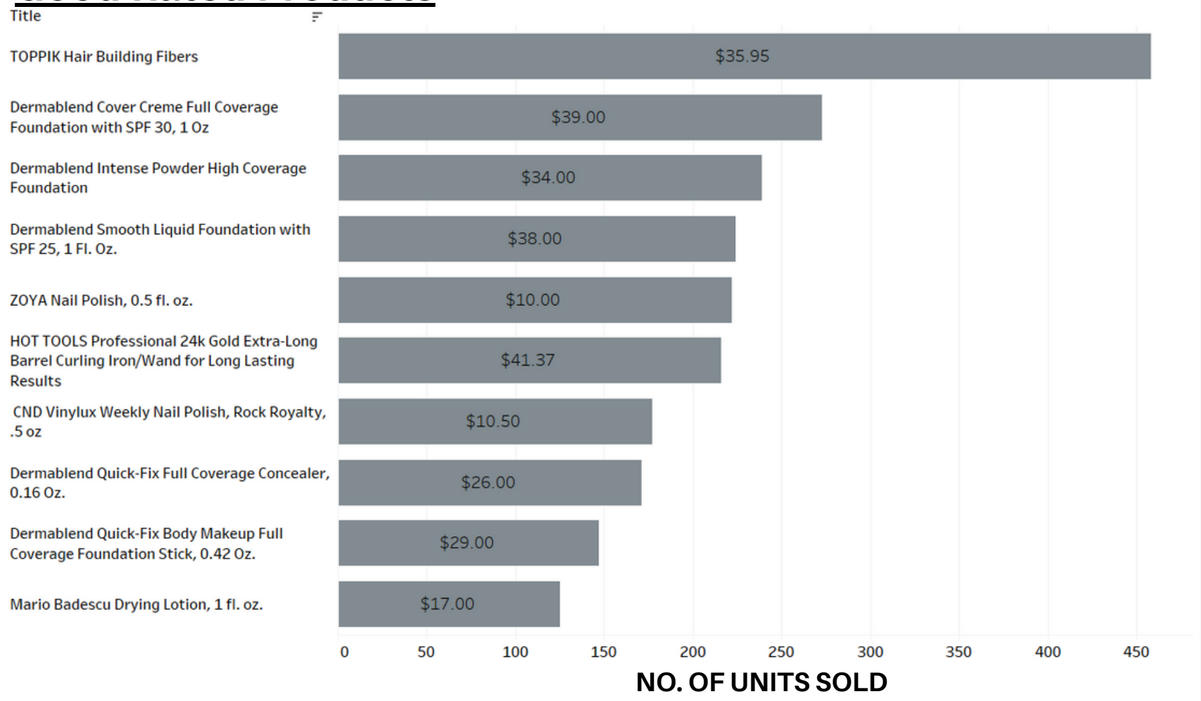
1. Best products - Having high polarity and high rating
2. Good products - Having high ratings but a comparatively lower polarity
3. Average products - Having low polarity and low rating

PRODUCT SEGMENTS

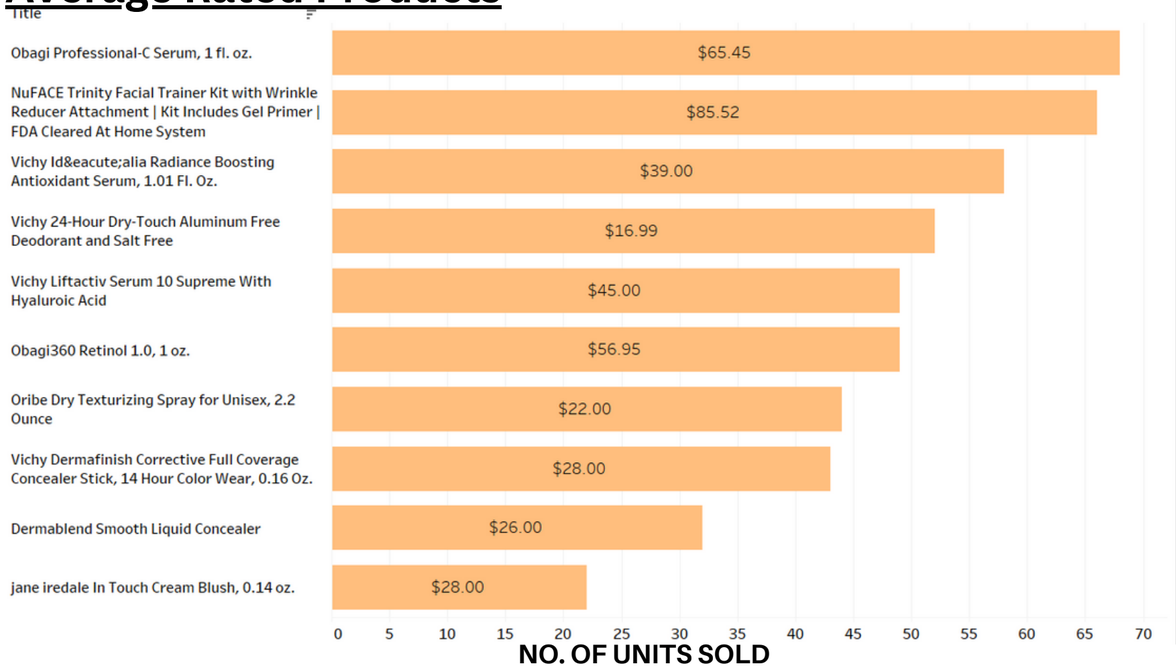
Best Rated Products



Good Rated Products



Average Rated Products



Customer Segmentation

OBJECTIVE	REASON	APPROACH	MODELS USED	RESULT
CUSTOMER SEGMENTATION	1. To segment the customers into different groups so as to target our promotions according to their behavior. 2. To adopt measures for Dissatisfied customers like feedback, promotion campaigns, and discounts so that they do not churn out	Clustering has been performed on the basis of Product ratings and Sentiments of reviews given.	1. K-means Algorithm 2. Agglomerative Clustering	Identified 3 clusters of customers: 1. Highly Satisfied Customers 2. Satisfied Customers 3. Dis-Satisfied Customers

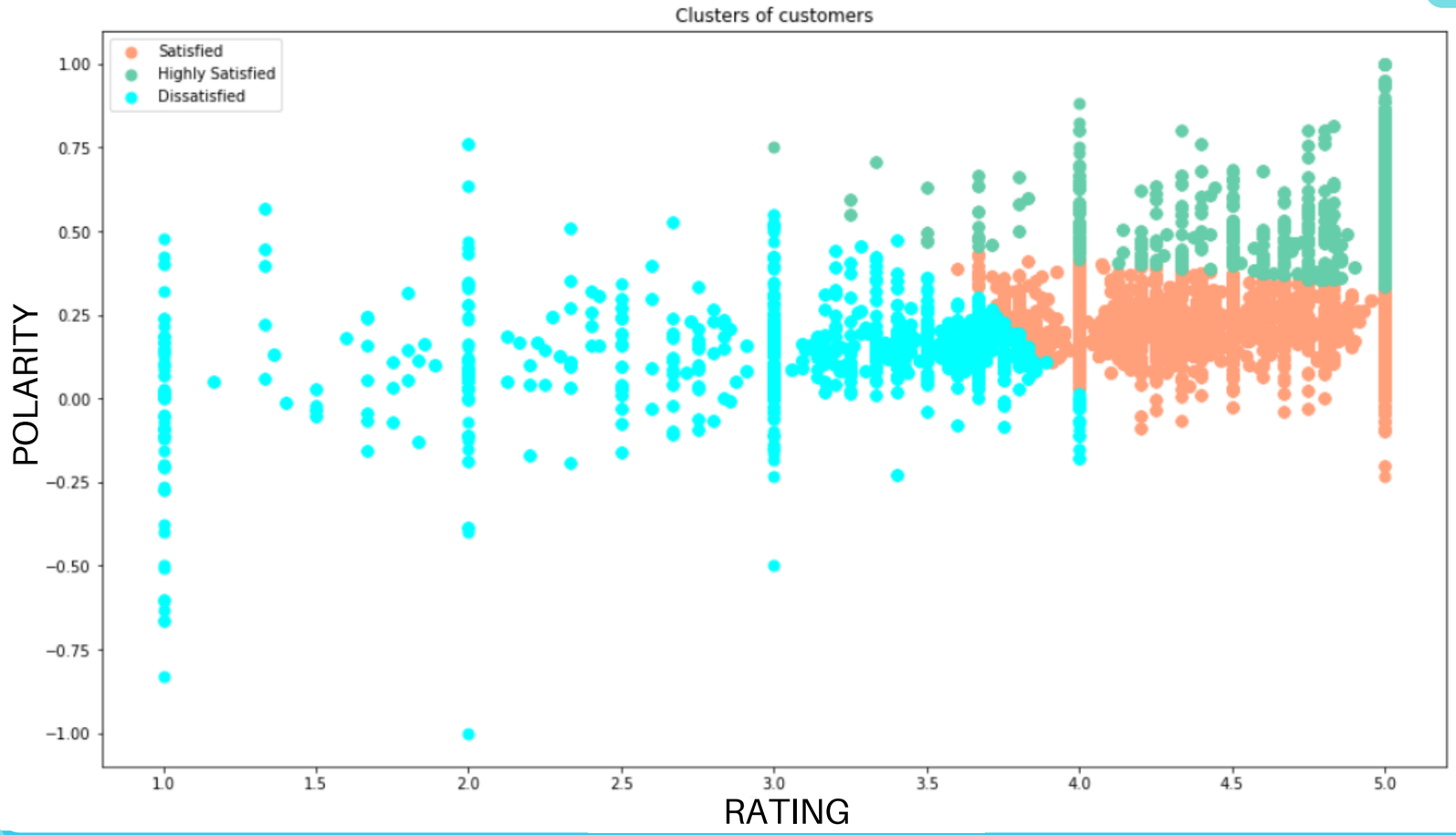
COMPARISON OF MODELS

:

	model	silhoutte_score
0	k-means	0.426565
1	agglomerative	0.338547

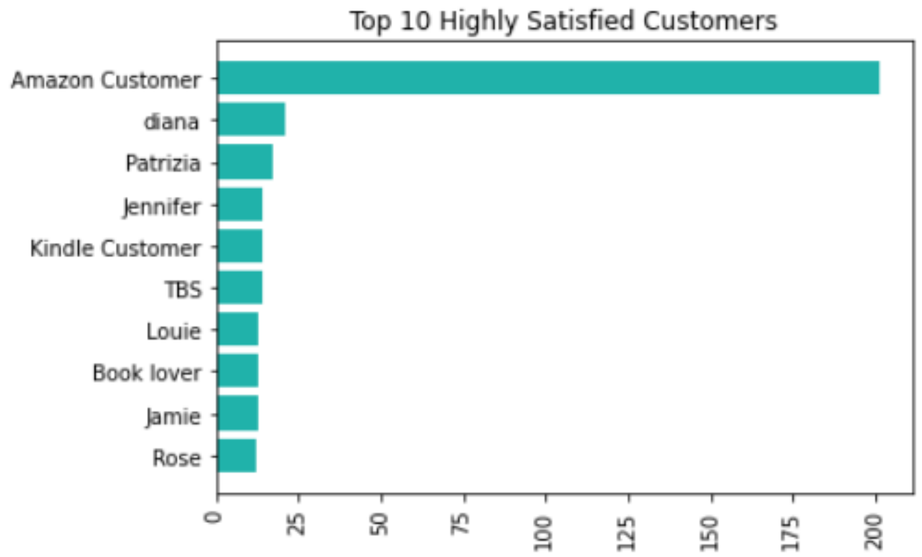
We have used Silhoutte score as the performance metric for comparing the model's effectiveness.

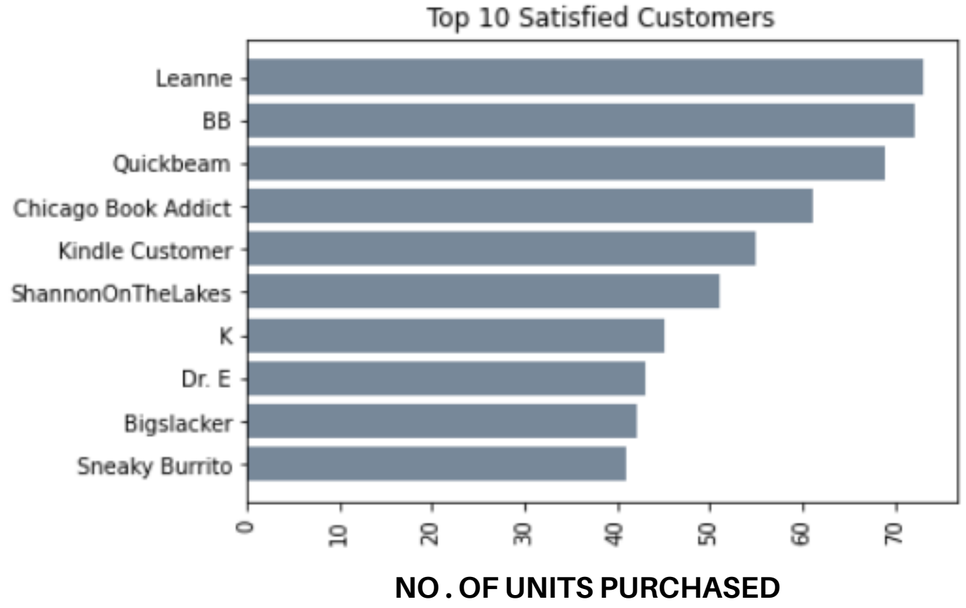
As silhoutte score of K-Means algorithm is better we will use this model for our CUSTOMER segmentation



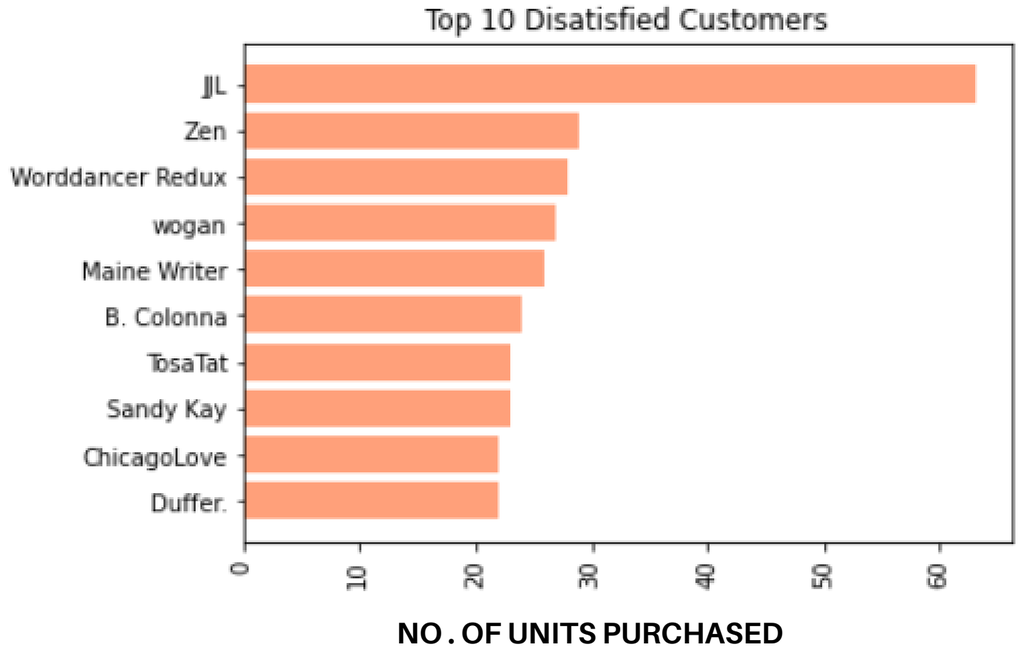
This graph is showing the categorization of customers into :

- Highly Satisfied - Having high polarity and high rating
- Satisfied - Having a high rating but the comparatively lower polarity
- Dissatisfied - Having low polarity and low rating





CUSTOMER SEGMENTS



Take feedback from Dis-satisfied customers so as to know their concerns and prevent their churn out

PRODUCT DEMAND FORECASTING BASED ON SENTIMENTS

The ability of forecasting a future event to take place is one of the key role of a business to make future stratigies, here we are performing the sentimental demand forecasting to find the demands of the products in future.

1

RESAMPLING OF DATA

- RESAMPLING IS FOR FREQUENCY CONVERSION WHICH ENSURES THAT THE DATA IS DISTRIBUTED WITH A CONSISTENT FREQUENCY.
- WE RESAMPLED OUR DATA ON MONTHLY WISE FOR THE COUNT OF PRODUCT ID'S AS PER THEIR SENTIMENTS.
- BY THIS WE CAN SIMPLY FIND THE COUNT OF SOLD PRODUCTS AS PER THEIR SENTIMENTS.

2

STATIONARITY TEST

- ADF TEST.
- USED LAG DIFFERENCING TECHNIQUE TO MAKE OUR DATA STATIONARY

3

STATISTICAL DECOMPOSITION

- CHECKING FOR THE TREND IN OUR DATA.
- CHECKING SEASONALITY IN OUR DATA.
- CHECKING RESIDUALS IN OUR DATA.

4

MODELS USED

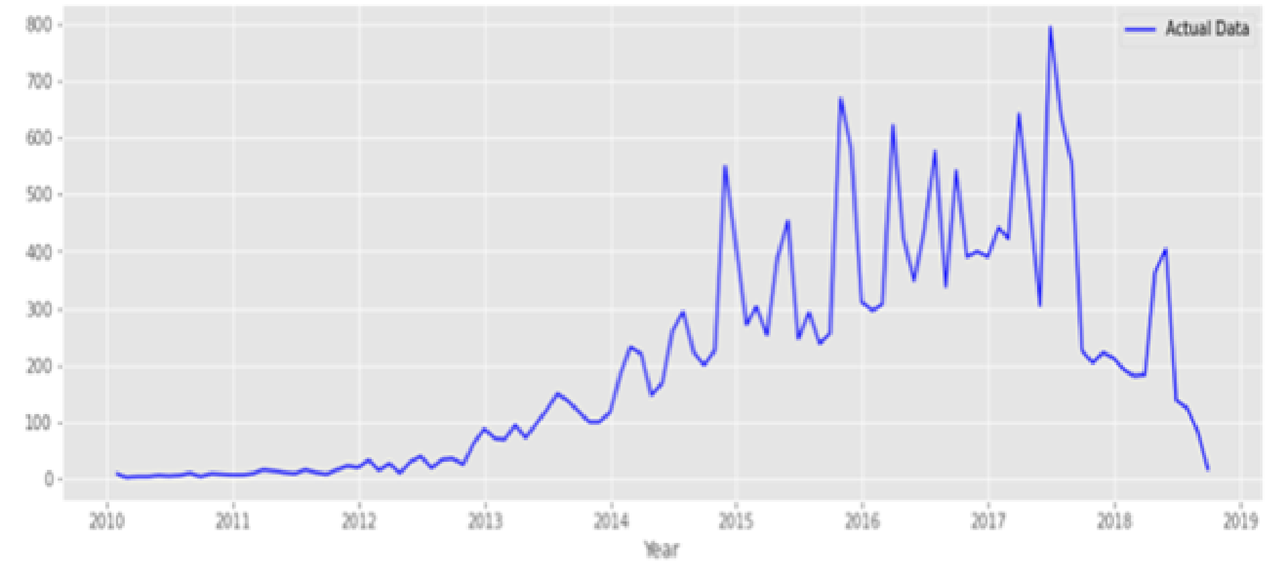
- ARIMA MODEL
- SARIMA MODEL

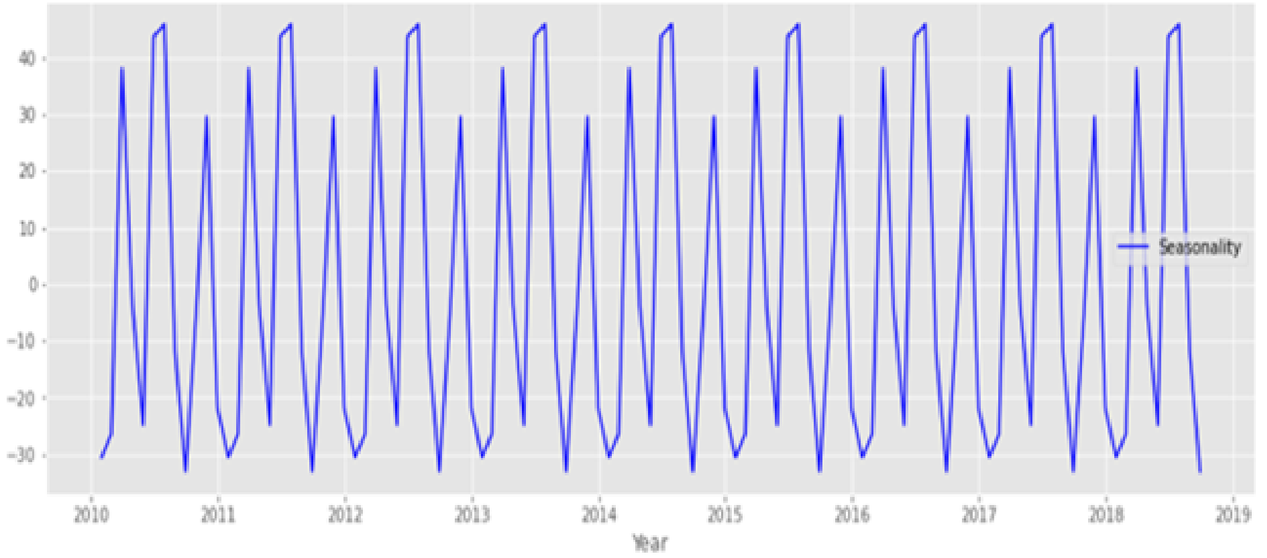
5

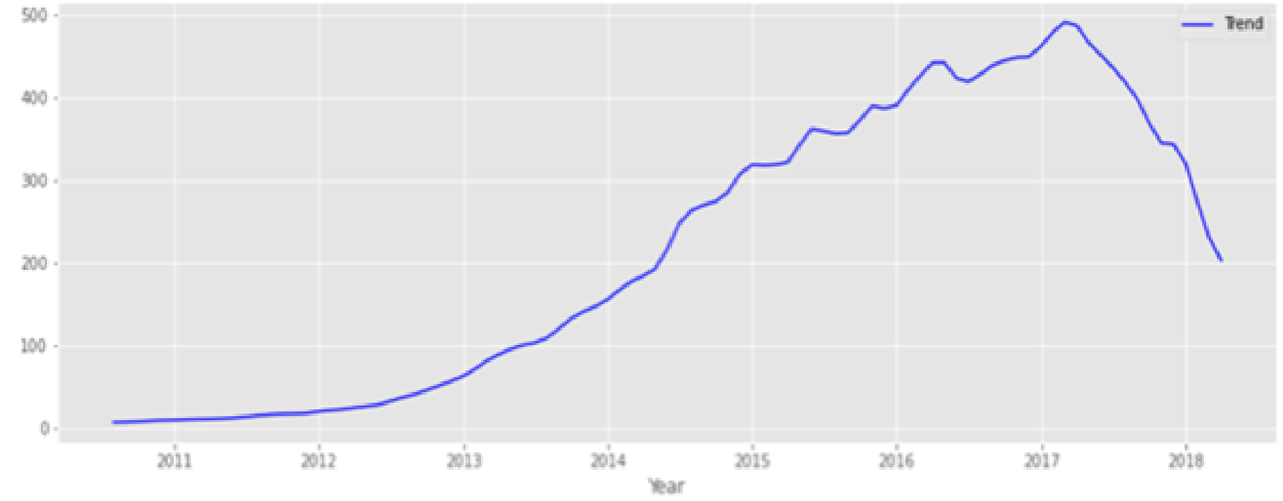
FORECASTING

Forecasted the demand of the products by choosing our best model.

STATISTICAL DECOMPOSITION GRAPHS OF PRODUCTS DEMAND

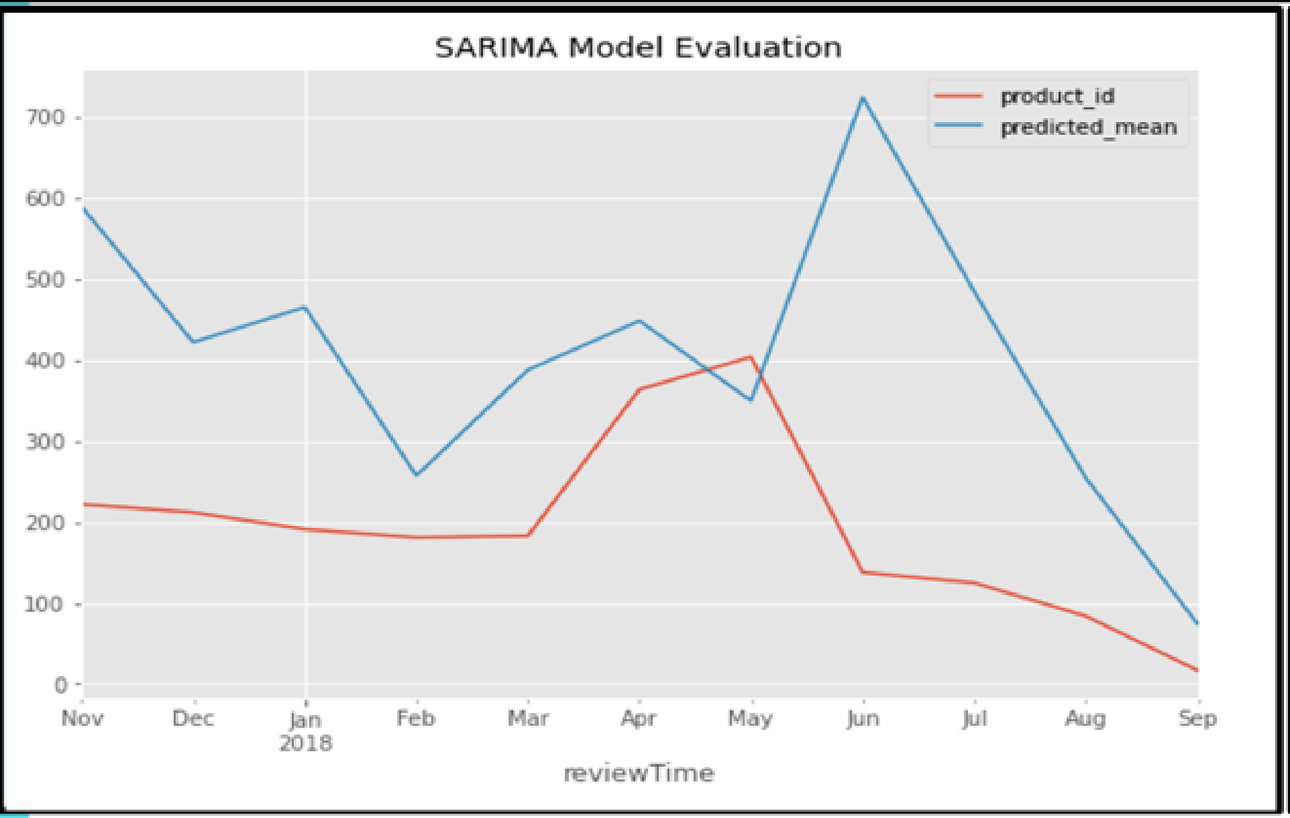


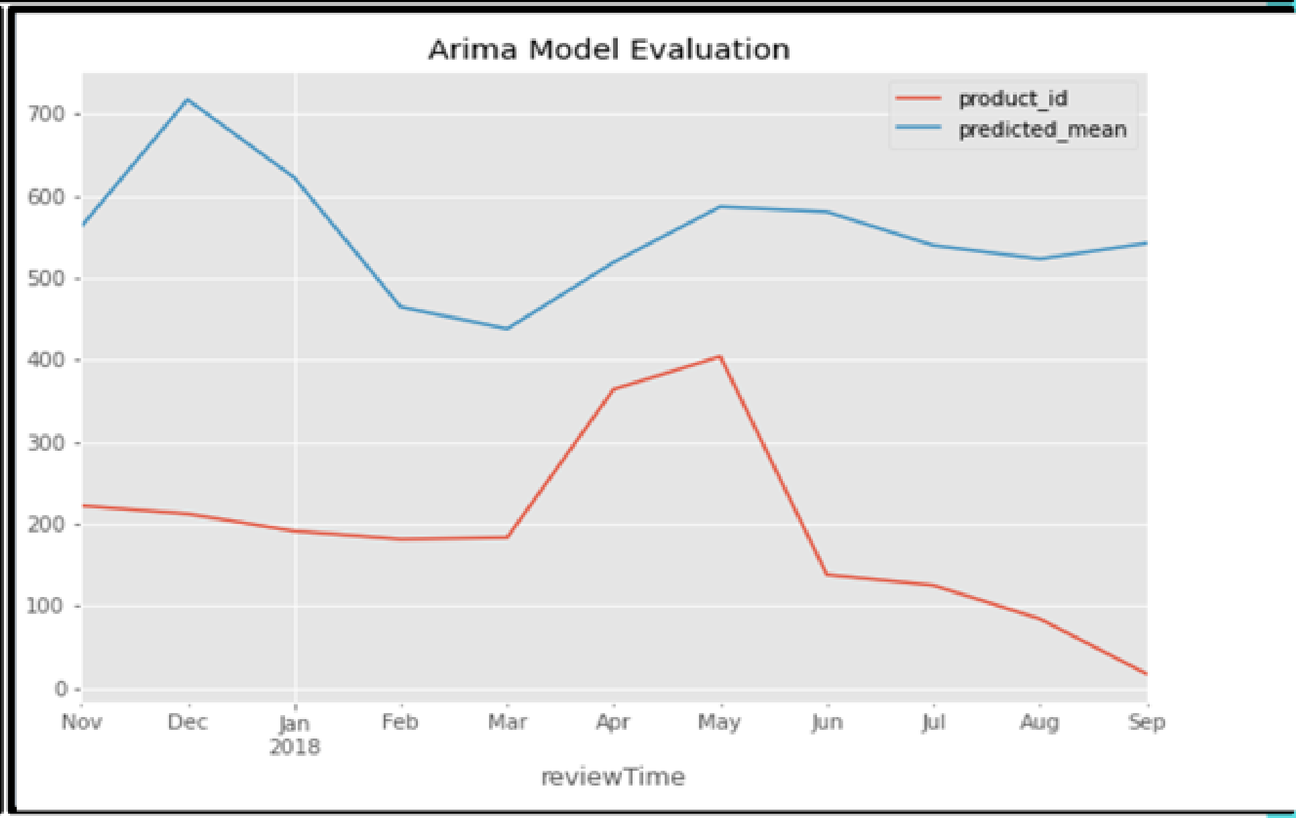


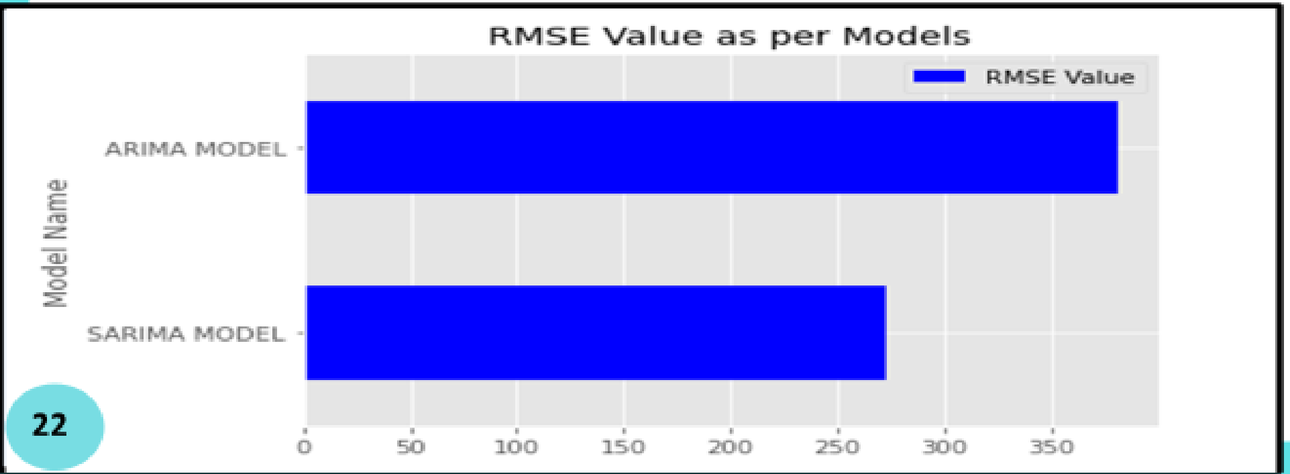


- **Trend: We observe that there is an increasing trend in the demand of the products till 2017 and after it starts decreasing.**
- **Seasonality: We observe a seasonality or similar pattern in demand of the products after a gap of 1 year.**
- **Residuals: We observe that the magnitude of the irregular data is increasing and decreasing propotional to the increament and decreament trend in the demand of products.**

COMPARISON OF MODELS (SARIMA,ARIMA) :





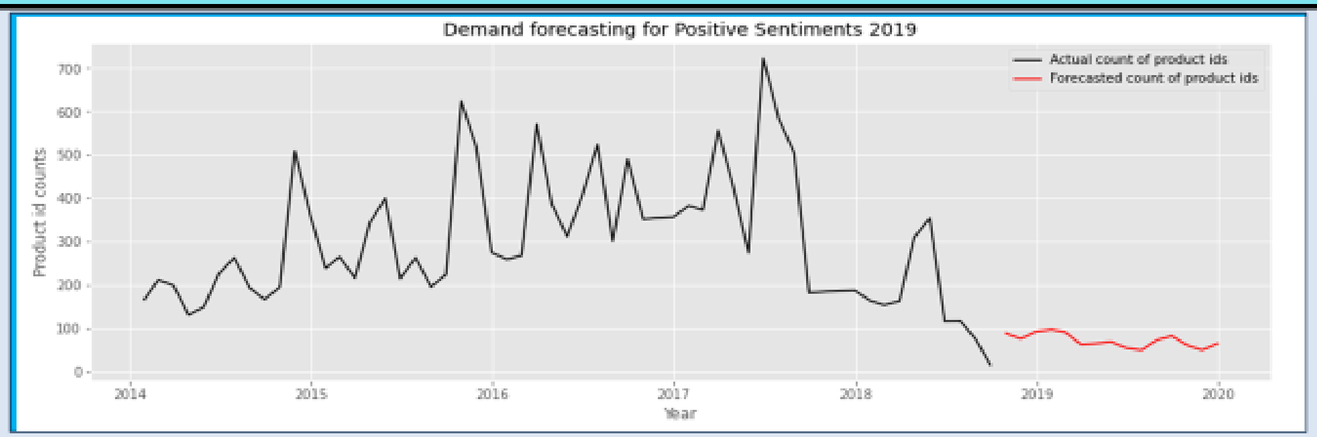


From these graphs we can clearly see that **SARIMA Model** giving the low error and the predicting values are more closer to the actual in comparison of **ARIMA Model**. And as we know that lower the RMSE value the better our model performs.

So, we chose **SARIMA Model** for the final forecasting of the demand of the products.

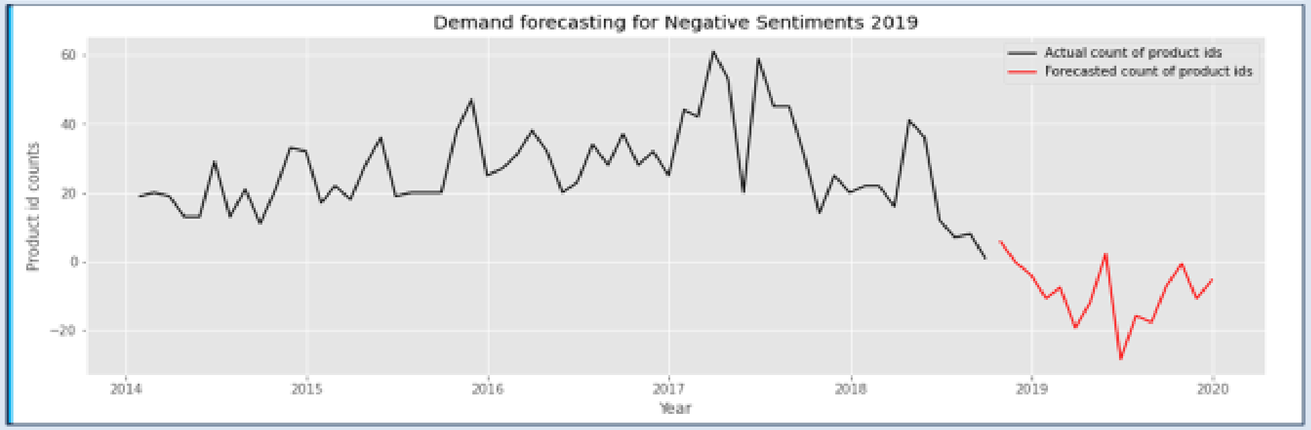
PRODUCT DEMAND FORECASTING BASED ON SENTIMENTS :

POSITIVE SENTIMENTS



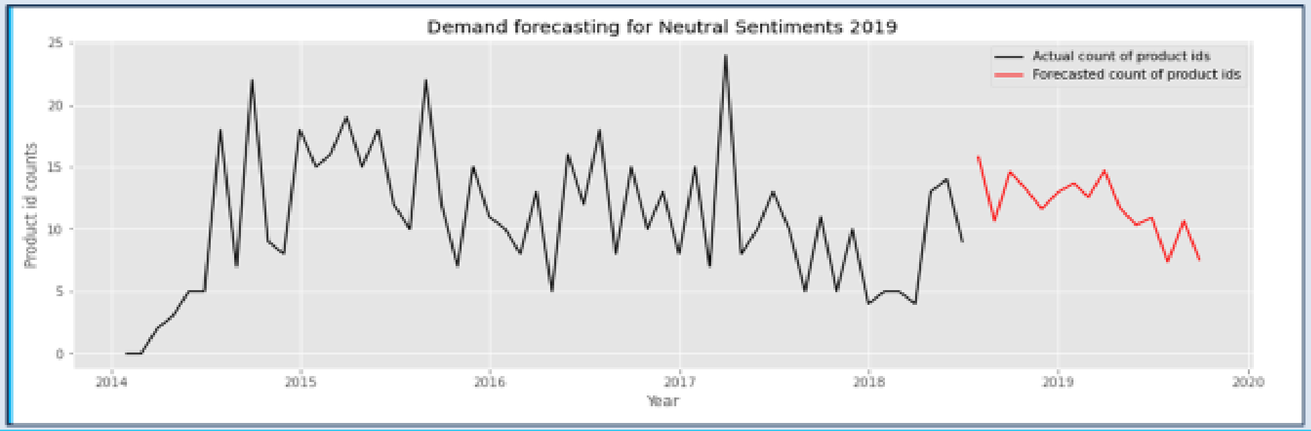
From this forecasted graph we can see that for the upcoming year 2019 the count of the products with positive sentiments will decreased,
So which means that the costumers will not be much satisfied with the service.

NEGATIVE SENTIMENTS



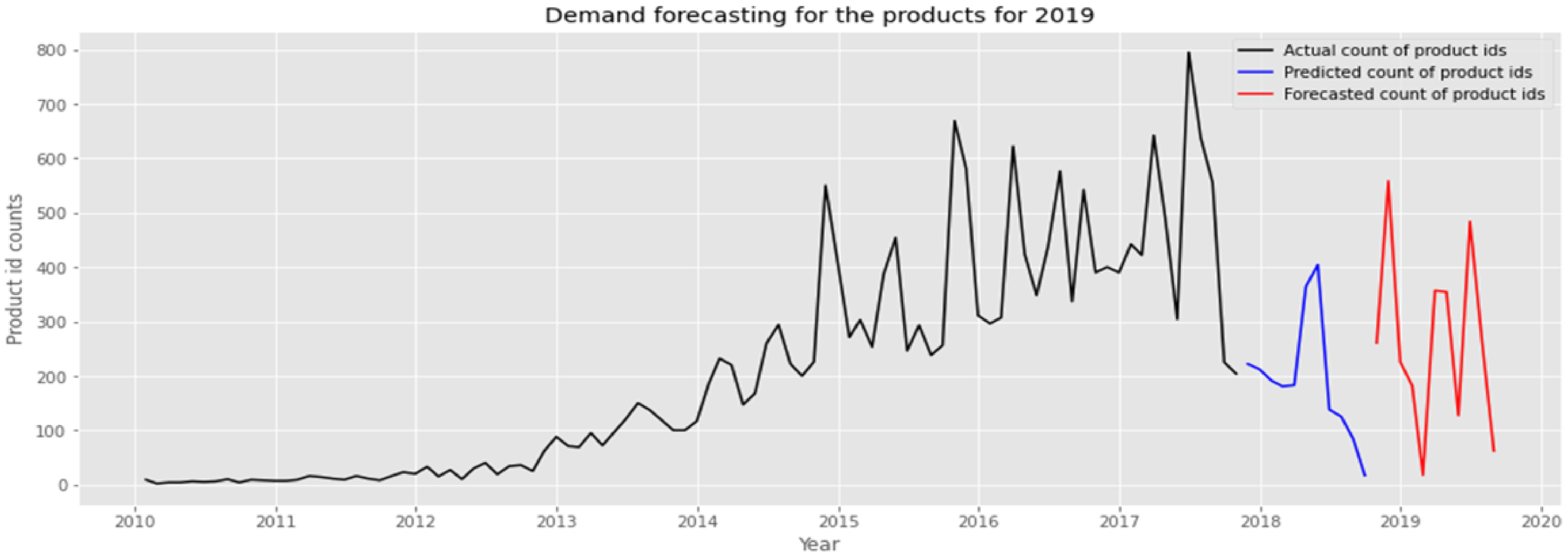
From this forecasted graph we can see that for the upcoming year 2019 the count of the products with negative sentiments will decreased in high rate, which means that the costumers will not be satisfied with the service, so company needs to work on the quality of the products.

NEUTRAL SENTIMENTS



From this forecasted graph we can see that for the upcoming year 2019 the count of the products with neutral sentiments will decreased,
So which means that the costumers who were giving neutral reviews there is a decline in shopping rate for them.

OVERALL PRODUCT DEMAND FORECASTING FOR 2019 :

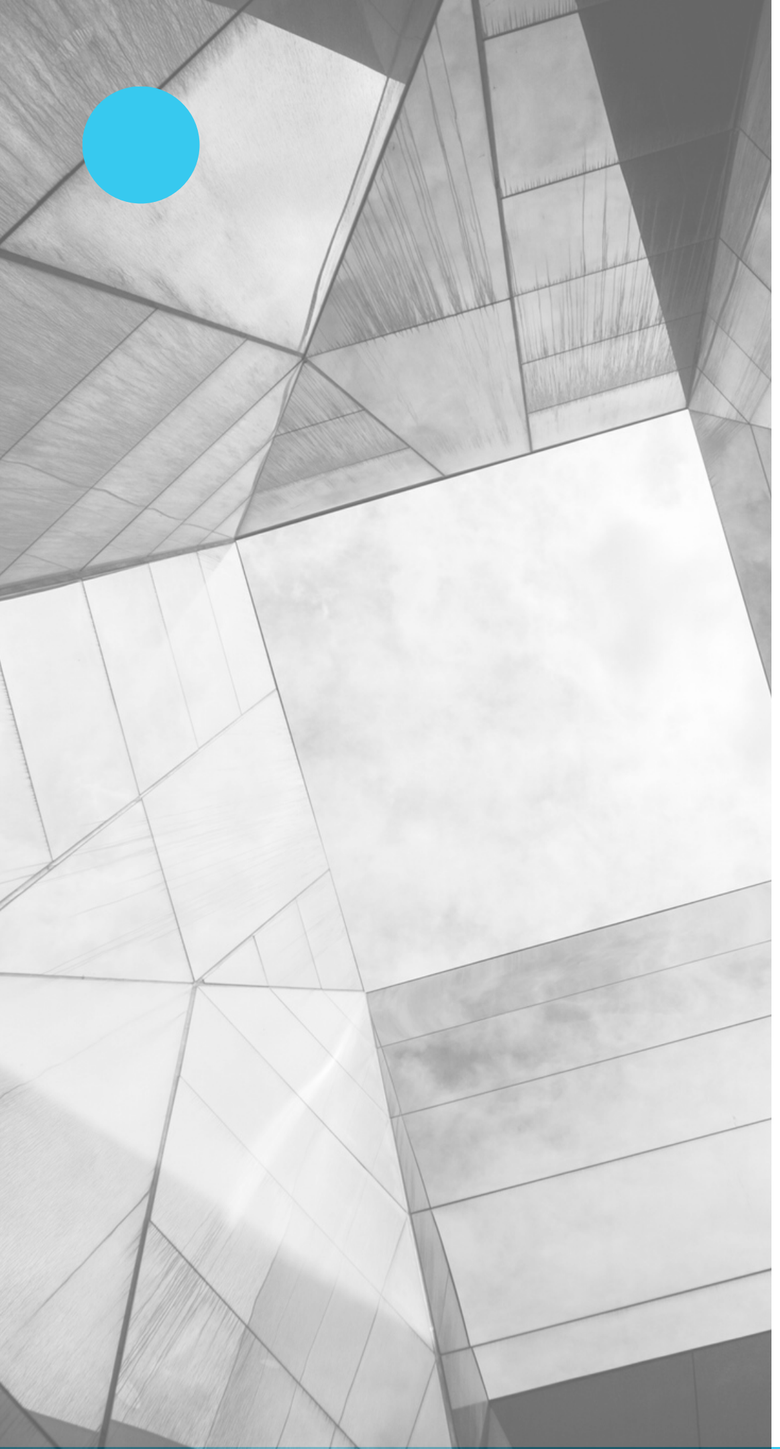


OBSERVATION:

- Here when we check for the overall products for all the sentiments ; we found that for the upcoming year 2019 the total count of the product id's will be decreasing which means that the sales and the demand of the products will decreased.
- Our hypothesis that as the positive ,neutral sentiments decreasing for the upcoming year the product demand will decreased got accepted.

CONCLUSIONS

- It has seen that among Luxury Beauty products the most popular products among customers are ELEMIS DAILY MOSITURE BOOST, ORIBE HAIR SHAMPOO, CREATIVE NAIL DESIGN SHELLAC and SKINMEDICA CREAM.
- From our EDA we can see there is a decreament in the purchase rate of our premium customers BB, Leanne and Quickbeam after 2017.
- Analysing our product clustering based on sentiments we observe that there is a significant difference in the sales of average rated product clusters as compared to best and good product clusters.
- Analysing our customer clustering based on sentiments we observe that there is a significant difference in the Dissatisfied and Satisfied customer clusters as compared to Highly satisfied customers clusters.
- From our time series forecasting we observe that the future sales and demand for products will fall as the customers will not be satisfied with the service.
- As the positive and neutral reviews are decreasing so company needs to relaunch the least rated products after working on their quality and decrease some price also and recommend our best products to the costumers.
- Company needs to provide some more festive, seasonal sales or discount offers to the costumers to attract them.



THANK YOU