Capstone Project Report

Amazon Product Review Analysis

Prime Pantry Data

Under Guidance Of
Dr. Amit Kumar

Presented By :
Sania Inamdar ● Subhranil Mondal ● Tejaswini Dhangar ● Yash Singh

Problem Statement

Customer Retention and Sentiment Forecasting

Customer retention strategy through feedback analysis (Customer classification & clustering as an outcome of analyzing the review text).Trend & seasonality analysis to predict how frequently a particular category of customer would shop in the future.

Data Source

- Amazon Prime Pantry data consist of information of Prime members and non-Prime members in selected areas to buy non perishable food items and household supplies in everyday package sizes.
- The data we have to download was **5-core** and **metadata** from the Base dataset (https://jmcauley.ucsd.edu/data/amazon/). These two file contains reviews and rating related columns.
- After that we are merging the two dataset to get our final dataset to work on.

Description of the treatment on the data

1 — 2 — 3 — 4

Dropping Non-Required Column

Checking Null Values

Data Cleaning

Filling Null Values

- Price column is most important column, but we can see that in price column null values are present & it is in string format. So we cleaned the price column,filling the null values & convert the datatype into float.

- Correcting the date column & change the date format from 06 13, 2015 to 2015-06-13.

- We are merging the review & summary column as review_text because review and summary contents are similar.

Preprocessing Analysis

Description of the Dataset

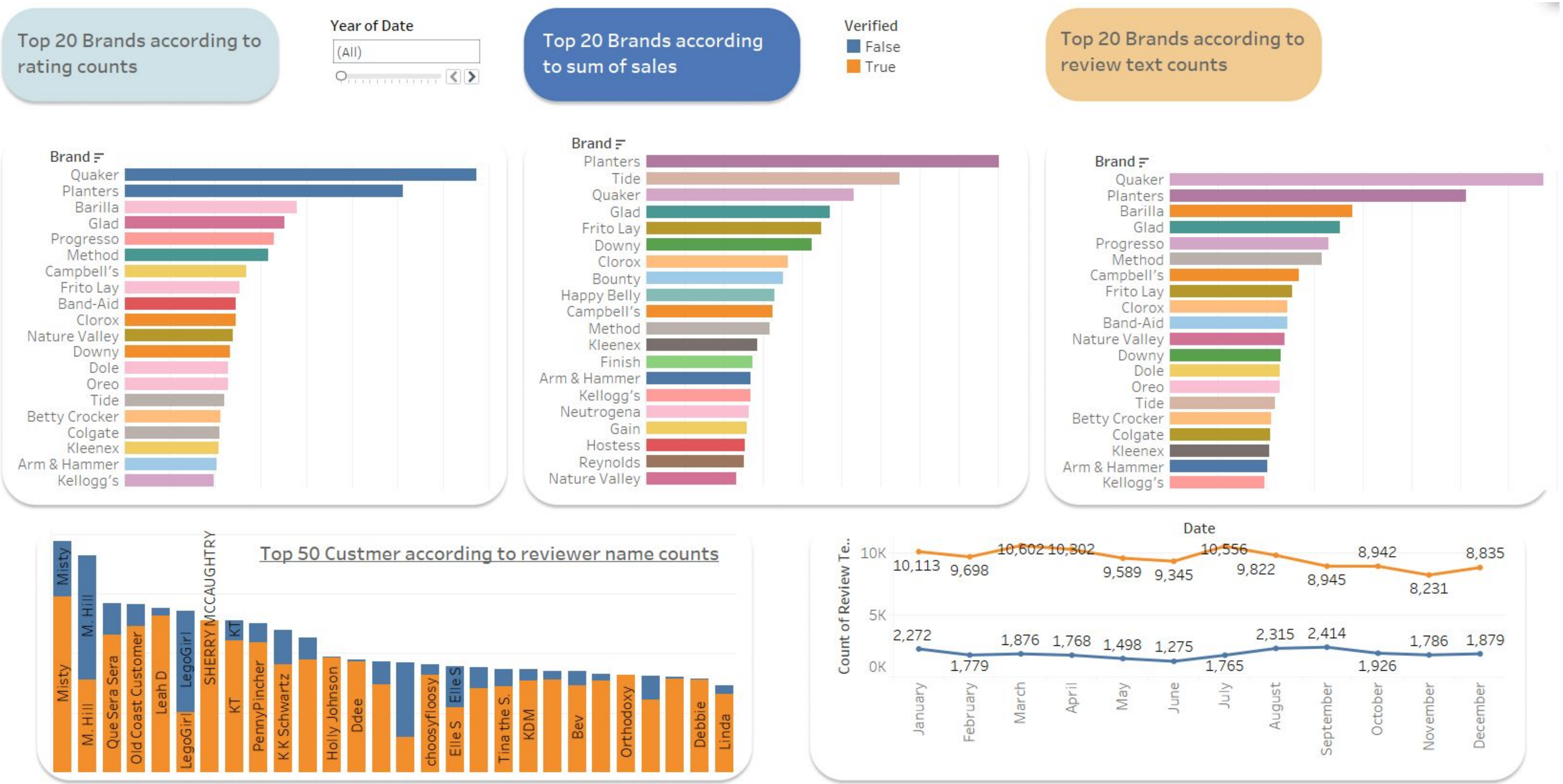
After merging the 5 Core & Metadata on '**asin**'

COLUMNS	DESCRIPTION
OVERALL	Rating given to product
VERIFIED	User is verified or not
REVIEWTIME/DATE	Date of the review
REVIEWERID	Id of the user
ASIN	Product ID
REVIEWER NAME	Name of the reviewer
REVIEWER TEXT	Review text of reviewer
SUMMARY TEXT	Summary of review text
UNIXREVIEWTIME	Time of the review
VOTE	Helpful votes of the review
IMAGE	Images that users post after they have received the product
STYLE	A dictionary of the product metadata, e.g., "format" is "hardcover"
CATEGORY	List of categories the product belongs to
TECH1	The first technical detail table of the product
DESCRIPTION	Description of the product
FIT	
TITLE	Name of the product
ALSO BUY	Also buy product
TECH2	The second technical detail table of the product
BRAND	Brand name
FEATURE	Bullet-point format features of the product
RANK	Sales rank information
ALSO VIEW	Also view products
DETAILS	Product Details
MAIN CATEGORY	URL
SIMILAR ITEMS	Similar product table
DATE	Date of review
PRICE	Price in US dollars (at time of crawl)
IMAGEURL	URL of the product image
IMAGEURLHIGHRES	URL of the high resolution product image

After Preprocessing size of dataset (137769, 17)

COLUMNS	DESCRIPTION
OVERALL	Rating given to product
VERIFIED	User is verified or not
REVIEWTIME/DATE	Date of the review
REVIEWERID	Id of the user
ASIN	Product ID
REVIEWER NAME	Name of the reviewer
REVIEWER TEXT	Review text of reviewer
SUMMARY TEXT	Summary of review text
UNIXREVIEWTIME	Time of the review
DESCRIPTION	Description of the product
TITLE	Name of the product
ALSO BUY	Also buy product
BRAND	Brand name
FEATURE	Bullet-point format features of the product
RANK	Sales rank information
ALSO VIEW	Also view products
PRICE	Price in US dollars (at time of crawl)

Exploratory Data Analysis



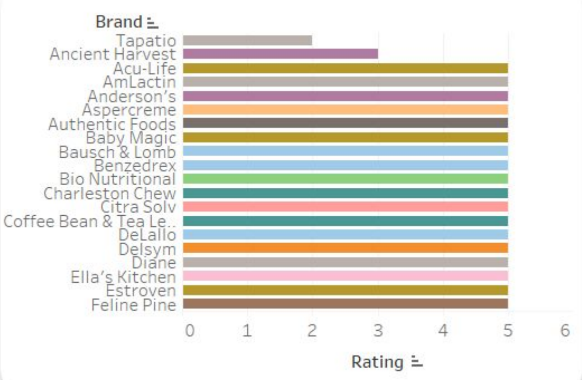
Exploratory Data Analysis (Contd.)

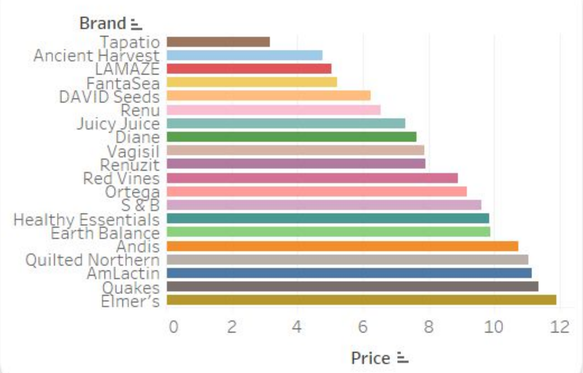
Bottom 20 Brands according to rating counts

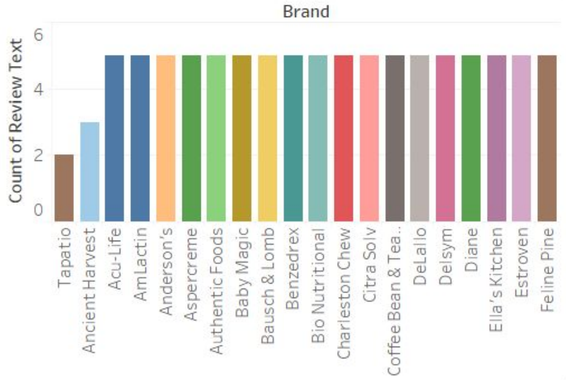
Bottom 20 Brands according to sum of sales

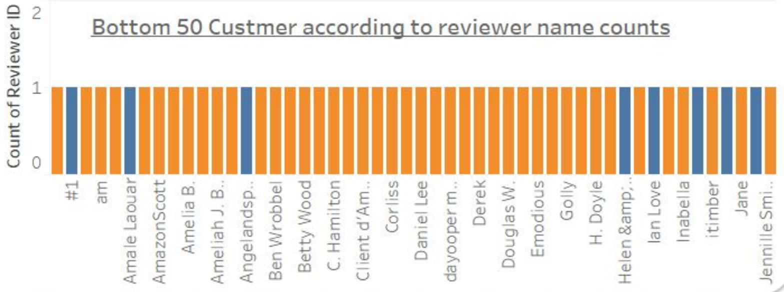
Verified
False
True

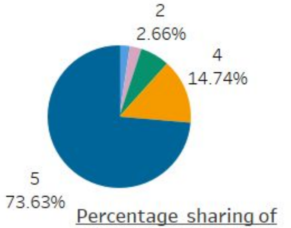
Bottom 20 Brands according to review text counts



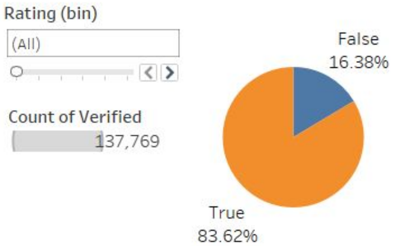








Percentage sharing of rating by each category



Percentage of verified and non-verified customers

Text Preprocessing & Sentiment Analysis

Text Cleaning

We are removing non-word characters, white spaces, digits, punctuations, stop words from the review_text column.

Text Processing

Then we are going for the lemmatization to reduce the variability in the words. And we are converting the lists into strings and storing the preprocessed data into clean_text column.

Sentiment

We are using **TextBlob** to calculate polarity of each reviews. And based on polarity we are calculating the sentiments of the reviews. And saved the data in csv format for further use.

- Text Blob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.
- Polarity score lesser than or equal to zero is termed ‘Negative’, Score greater than or equal to 0.5 is termed ‘Positive’, and score ranging between 0 to 0.5 is termed ‘Neutral’.

EDA on Sentiment Analysis

AMAZON Prime Pantry Customer Sentiment Analysis

amazon PrimePantry

Total No Of Customers
137,769

Most Satisfied Customer
Sherry Maccaughtry
124 Positive Review

Most Disatisfied Customer
Lisa M.Rainer
67 Negative Review

Total Customer Sentiment

Revie..	Revie..	Analysis egative	Neutral	Positive	Grand..
AMMNGUJ..	Misty	17	117	60	194
A35Q0RBM..	M. Hill	10	170	2	182
AXK37UZY8..	Que Sera Se..	44	34	64	142
AKPG8VQB..	Old Coast C..	15	76	50	141
A2H9H3BV..	Leah D	36	41	61	138
A13J2PGKN..	LegoGirl	17	117	2	136
ANDVNCX6..	SHERRY MC..	4		124	128
A25DP3DW..	KT	20	95	13	128
A26K3T6L5..	PennyPinch..	9	113	3	125
A3F9UAX22..	K K Schwartz	15	39	66	120
A1UVH2I7..	AmyJean B..	28	30	55	113
A2737NVZ9..	Holly Johns..	36	23	38	97
A276RHM6..	Ddee	16	16	63	95
A3EF7PUYT..	Gary R. Jor..	17	2	74	93
A2YKWYC3..	ShannonOn..	11	78	3	92
A3JXARXF2..	choosyfloosy		52	39	91
A1JN63QB..	Elle S	17	65	7	89
AM0LVUTS..	Patricia Geib	28	54	6	88
A6V5N3VG..	KDM	4	35	48	87
A3MPHEJ0..	Tina the S.	9	42	36	87
AA3ORMFS..	Bev	54	16	15	85
A2SQX1JEI..	oldtimeroc..	26	12	47	85

Satisfied Customers

Revie..	Reviewer Name	Analysis Positive
ANDVNCX6..	SHERRY MCCAUGHTRY	124
A3EF7PUYT..	Gary R. Jordan	74
A1BT9J2I6..	Debbie	72
A3F9UAX22..	K K Schwartz	66
AXK37UZY8..	Que Sera Sera	64
A276RHM6..	Ddee	63
A3G3NLQP..	Debby P	61
A2H9H3BV..	Leah D	61
AMMNGUJ..	Misty	60
A3DFMKBG..	Angelica Kate	55
A1UVH2I7..	AmyJean Barclay	55
A1K1KL78Y..	Bernard DiStefano	55
A3TJYI860..	Jimbo's Account	53
A27UQ3GD..	Jodi Walbaum-Vaniman	53
A1ZB822AY..	Amazon Customer	53
A1PO4QCL..	kittin katt	53
A1Z91URY7..	Patricia D.	52
A1XQ0F01C..	Mrs. J.	52
ALRNOXWP..	Alana H Mars	50
AKPG8VQB..	Old Coast Customer	50
AUDKWPXR..	C. C. Taylor Lechem	49
A3PM301Y..	johnny horton	49

Disatisfied Customers

Reviewer ID	Reviewer Name	Analysis Negative
92ZKEZI137M1	Lisa M. Rainer	67
1K8CSVXQUE6HG	Doyle s.	60
A3ORMFSVHSH0	Bev	54
1800579078GCU	Nancy E. Overstreet	50
XK37UZY8UPYP	Que Sera Sera	44
2H9H3BVFNS3Y0	Leah D	36
2737NVZ9F2P6I	Holly Johnson	36
8I6H0K9Y5EEC	Michael C. Lounsbury	35
24BYYMKE7518V	Ginger Pollock	35
2511IWU7F25ZG	Bo Sheriff	33
2EBHVSUECXI3I	Matein	31
ZVIQ5SU7XPD5	Michael Chaffin	30
3T2LT2KQQNNQ1	Norman R.	30
1MYK03VRIKETH	Sabrina	30
NTWFVZZKSBSF	Red Wine Kline	29
FGH4CSZ4M5CM	JanieP	29
3BXXHSF0S3A0L	Janet Bliss	29
1U39SOPOEDBLL	John B.	29
1IS09GJPJV629	Carey Probst	29
1F1YCHCIY8752	2tall4u	29
M0LVUTSAU0J5	Patricia Geib	28
IQLP3JDOU91X	Amazon Customer	28

Total Sentiment
137,769

Positive
63,940 (46.49%)

Neutral
46,270 (33.64%)

Negative
27,323 (19.87%)

EDA on Sentiment Analysis (Contd.)

AMAZON Prime Pantry Brand Recommendation Analysis

Total No of Brands

137,769

Most Recommended Brand

Product B000WHTWD2 from Brand Mott's

355 Positive Review

Most Non-Recommended Brand

Product B0014CZUZ8 from Brand Planters

138 Negative Review

Overall Brand Sentiment

Asin	Brand	Analysis Nega..	Neutral	Positive	Grand To
B0014CZUZ8	Planters	138	228	295	66
B00HZ6X8..	Tide	112	174	287	57
B0014CV2IW	Band-Aid	105	115	200	42
B0014D5N..	Glad	104	158	255	51
B000Q5NG..	Progresso	77	104	143	32
B00UJIXP0A	Planters	67	90	126	28
B0014CUC..	Quaker	64	99	175	33
B00SH4ORI2	Campbell's	63	53	64	18
B0111K4XHS	Planters	54	65	149	26
B01506PS..	Frito Lay	52	66	129	24
B000RYBU..	Dole	50	28	95	17
B00R2O5JD0	Oreo	48	41	105	19
B000OV4FY2	Dole	47	57	138	24
B00JGZBTCI	Downy	46	62	155	26
B00R1J3N84	Frito Lay	45	56	119	22
B00IDODTA..	Frito Lay	45	49	114	20
B000Q5LW..	Progresso	44	35	70	14
B00R1H0A..	Nature Vall..	43	42	103	18
B000YK0S58	Band-Aid	43	67	70	18
B00IB1D0IO	Barilla	42	30	106	17
B000Q5K1..	Dole	42	50	109	20
B00IDODUY..	Frito Lay	41	46	94	18

Recommended Brands

Asin	Brand	Analysis Positi..
B000WHTWD2	Mott's	355
B0014CZUZ8	Planters	295
B00HZ6X8QU	Tide	287
B0014D5NVS	Glad	255
B0014E84TK	Nutella	252
B00PC5CI8K	Rice Krispies Treats	245
B0013FRRMK	Method	212
B0010XS2CE	Softsoap	209
B00PTX8N52	Dawn	206
B0014CV2IW	Band-Aid	200
B00I9JK8GU	Ziploc	182
B00NP79AI8	Nestle Pure Life	180
B0014CUCU6	Quaker	175
B0014E2SOC	Scotch-Brite	174
B00IAE6PWA	Pepperidge Farm	167
B015TYDZTE	Bounty	166
B00QTX0D18	Kellogg's	165
B00RGVOUIO	Ritz	164
B00099XPD6	Carnation	161
B00NAZ532W	Viva	160
B00JGZBTCI	Downy	155
B0014CSG5Y	Huy Fong	154
B0111K4XHS	Planters	149

Non- Recommended Brands

Asin	Brand	Analysis Nega..
B0014CZUZ8	Planters	138
B00NAZ532W	Viva	122
B000WHTWD2	Mott's	118
B00HZ6X8QU	Tide	112
B01B4N9Y6C	Cottonelle	106
B00PC5CI8K	Rice Krispies Treats	106
B0014CV2IW	Band-Aid	105
B0014D5NVS	Glad	104
B0014E2SOC	Scotch-Brite	98
B0014E84TK	Nutella	90
B0014D2CIU	Reynolds	84
B015TYDZTE	Bounty	82
B00NP79AI8	Nestle Pure Life	81
B00IAE6PWA	Pepperidge Farm	77
B000Q5NG78	Progresso	77
B000WCZ57Y	Terra	68
B011DX5G92	Charmin	67
B00UJIXP0A	Planters	67
B00099XPD6	Carnation	67
B00NAZ5JQ2	Cottonelle	66
B0014CUCU6	Quaker	64
B013OV42NW	Slim Jim	63
B00SH4ORI2	Campbell's	63

Total Sentiment

137,769

Positive

63,940 (46.49%)

Neutral

46,270 (33.64%)

Negative

27,323 (19.87%)

Classification Analysis

Objective: To build best supervised ML model to predict the sentiments of unseen reviews.

1. **Gathering Data**
   Defining target column & Feature column.
2. **Splitting Data**
   Splitting the data into train test split at 70:30 ratio
3. **Vectorization**
   Converting text data into numeric by vectorization methods.
4. **Model Building & Evaluation**
   Build 4 model to obtain the best evaluated model for the data
5. **Hyper Parameter Tuning**
   Hyper Parameter Tuning of best model to improve the model performance.
6. **Prediction**
   Predicting the unseen data by best model to observe the performance.

Classification Analysis

Classification Models

Models used for Classification

1	**Logistic Regression**	• Logistic regression is a statistical method for predicting binary classes. • It computes the probability of an event occurrence.
2	**Multinomial Naive Bayes**	• Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in NLP. • The algorithm is based on the Bayes theorem and predicts the tag of a text
3	**Decision Tree Classifier**	• A decision tree is a flowchart-like tree structure where an internal node represents feature, the branch represents a decision rule, and each leaf node represents the outcome.
4	**KNN Classifier**	• KNN is a non-parametric learning algorithm. Non-parametric means there is no assumption for underlying data distribution.

Model Summary

Vectorization Techniques Used:

1. **Bag Of Words:** Bag of words counts the occurrence of each words but doesn't preserve the order of the sentence.
2. **Count-Vectoriser:** It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. But it will consider the order of the words.
3. **TF-IDF:** Term Frequency Inverse Document Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus (data-set).

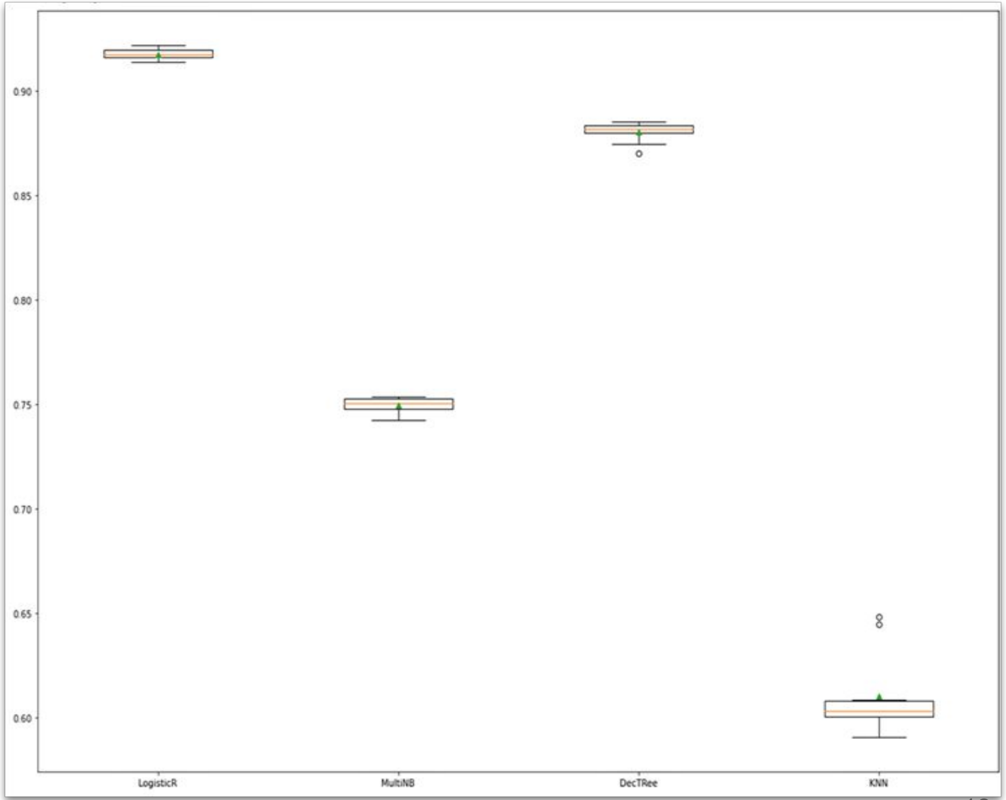
- We can observe that in Logistic Regression by Bag Of Word method we are getting highest Accuracy & F1 score.

Model	Technique	Accuracy	F1-Score
Logistic Regression	**BOW**	**0.93873**	**0.938424**
	CV	0.929226	0.928783
	TFIDF	0.921322	0.921278
Multinomial NB	BOW	0.751382	0.736585
	CV	**0.794637**	**0.789675**
	TFIDF	0.765251	0.746678
Decision Tree Classifier	BOW	0.86912	0.866888
	CV	**0.873097**	**0.87125**
	TFIDF	0.865677	0.863494
KNN Classifier	**BOW**	**0.720274**	**0.678415**
	CV	0.677844	0.618092
	TFIDF	0.661914	0.641593

Tf-Idf Vectorization method with Cross Validation

- **Cross-validation** is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

- From the visuals we can say that the **Logistic Regression** model gets the higher **F1 score (91.8%)** among all the models.



Hyperparameter Tuning

According to model evaluation **Logistic Regression** gives us better **accuracy & F1 score** compared to others.

- Considering **L1,L2** penalties and logarithmically spaced **C values**.
- We are applying Grid search cv on this.

```
Train Accuracy :  0.9823974396275822
Test Accuracy  :  0.9512656386383473
f1-Score Test  :  0.9512871008844906
Classification Report :
              precision    recall  f1-score   support

    Negative       0.96      0.95      0.95      8067
     Neutral       0.93      0.93      0.93     13737
    Positive       0.97      0.97      0.97     19440

    accuracy                           0.95     41244
   macro avg       0.95      0.95      0.95     41244
weighted avg       0.95      0.95      0.95     41244

------------------------------------------------------
Confusion Matrix :
 [[ 7636   420    11]
 [  310 12790   637]
 [   22   610 18808]]
```

Model	Accuracy	F1 Score
Logistic Regression With TFIDF(Before HP Tuning)	0.921322	0.921278
Logistic Regression With TFIDF(After HP Tuning)	0.951266	0.951287

Prediction on Random Data

After hyper parameter tuning as our model getting 95% accuracy. So here we are testing our model on random data.

Predicting the random data on best model:

```
## Prediction on random data
#######################################
data=["Taste not to be believed. Buy a box for my office every week",
      """These are delicious and healthy snacks!  I with they were more affordable because they're really tasty and convenient.
      I purchased these because they're lower in sugar than many other brands and really enjoy them.""",
      "I like most of the flavors but this one is my favorite so far!!","""Excellent. Only complaint is they stick to the wrapper and are hard to remove.
      But taste and nutrition are great."""]
data=vector.transform(data)
lr_tf_model.predict(data)
```

```
array(['Negative', 'Positive', 'Neutral', 'Neutral'], dtype='<U8')
```

Clustering Algorithms

1	K-Means	It classifies the dataset by dividing the samples into different clusters of equal variances.
2	Agglomerative Hierarchical	The Agglomerative hierarchical algorithm performs the bottom-up hierarchical clustering.
3	DBSCAN	The areas of high density are separated by the areas of low density.

Clustering Analysis

To construct groups or clusters while ensuring that the observations are as similar as possible with sentiments.

1. **Feature Extraction**

   Selecting the columns for cluster formation

2. **Cluster Selection**

   Selecting the number of clusters by Elbow curve, K distance graph, Dendrogram

3. **Model Building**

   Building 3 Models to obtain the best clusters

4. **Model Evaluation**

   Evaluating best model by cluster formation on Scatter plot

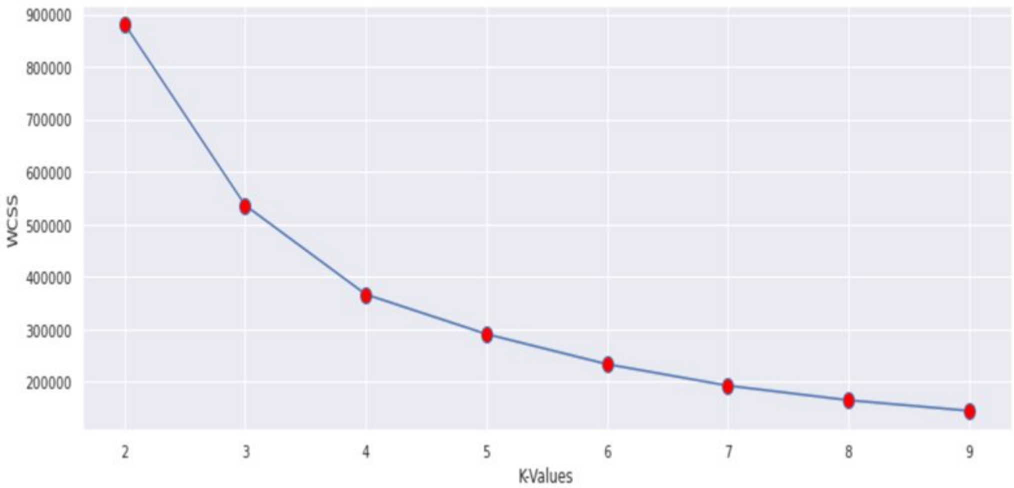
[Clustering Analysis]

K Means Clustering

Feature Selection:

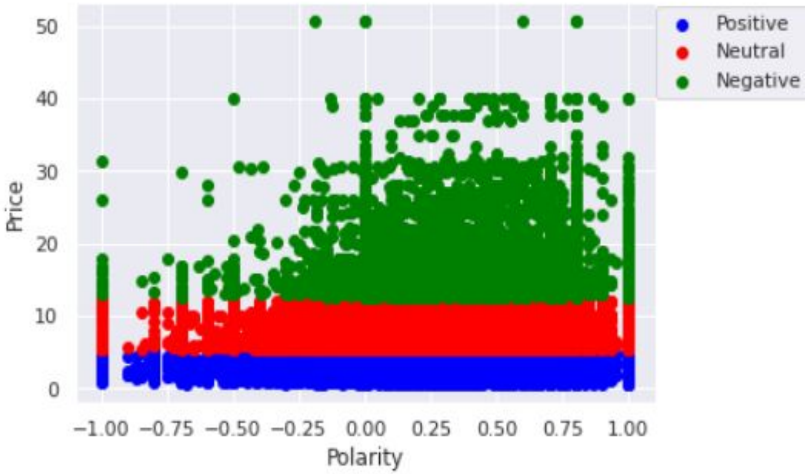
- Rating
- Price
- Polarity

Elbow method: To find the optimal value for k



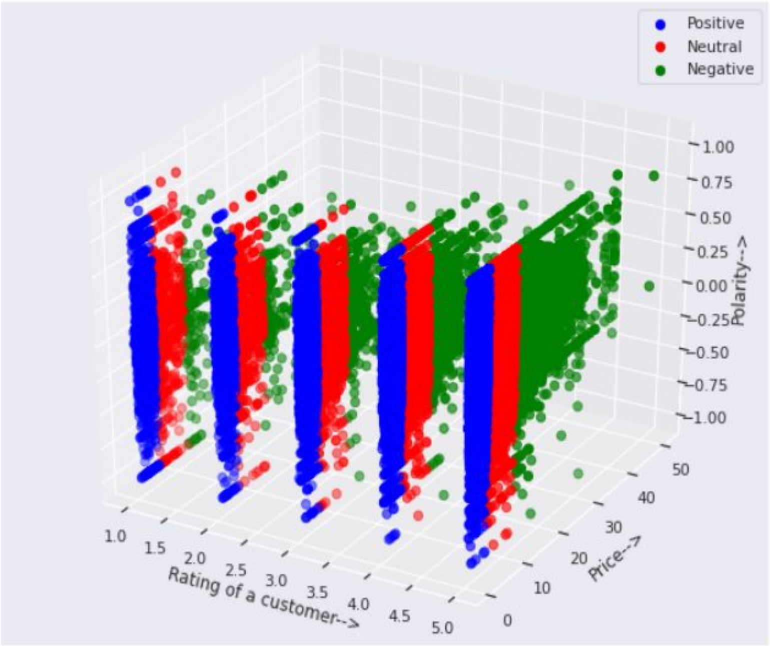
Clustering Evaluation

- Getting Optimum Clusters By **K-Means**

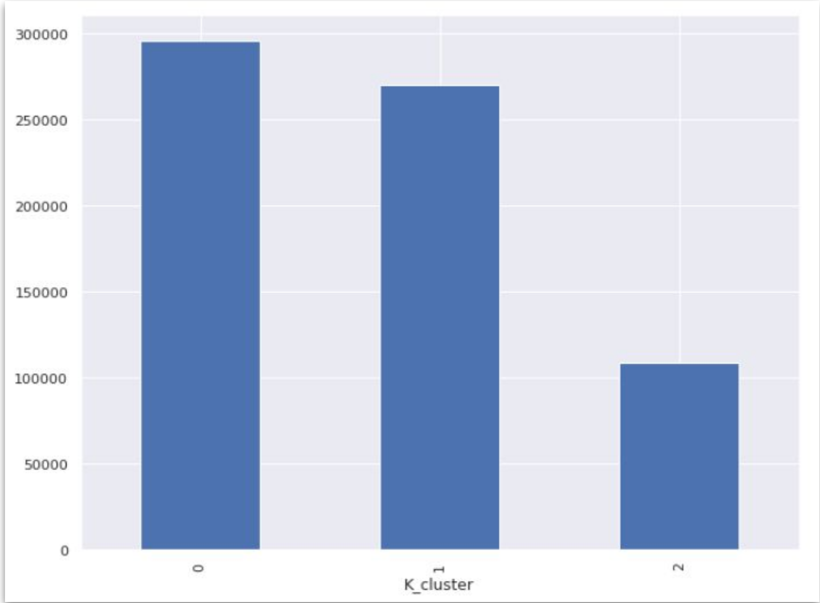


From 2D plotting we can't interpret properly.

By 3D plotting now we can Interpret the clusters properly.



Result of Clustering



Cluster ‘0’=’Positive’
Cluster ‘1’=’Neutral’
Cluster ‘2’=’Negative’

- The clusters 0 represents products with positive review and having maximum sales.
- The cluster 1 represents products with neutral reviews.
- Cluster 2 represents products with negative reviews having lesser sales compare to other two clusters.
- So, we need to focus on products belonging to cluster 2 such as Tapatio, Ancient Harvest. So we can increase the quality of those products.

Time Series Analysis

Objectives: Through Time Series Analysis we are going to predict/forecast the future sentiment of the customers shopping on Amazon Prime Pantry.

1. **Splitting Sentiment**
   Split overall sentiments into 3 different column Positive,Neutral,Negative
2. **Resampling**
   Resampling the data Quarterly to get the continuity in the data
3. **Decomposition**
   To visualize time series components
4. **Check Stationarity**
   Checking the stationarity of the data by ADF test
5. **Time Series Model**
   ARIMA, SARIMA models on Positive, Neutral, Negative Sentiments
6. **Sentiment Forecasting**
   On the best model we will forecast the upcoming sentiments

Time Series Analysis

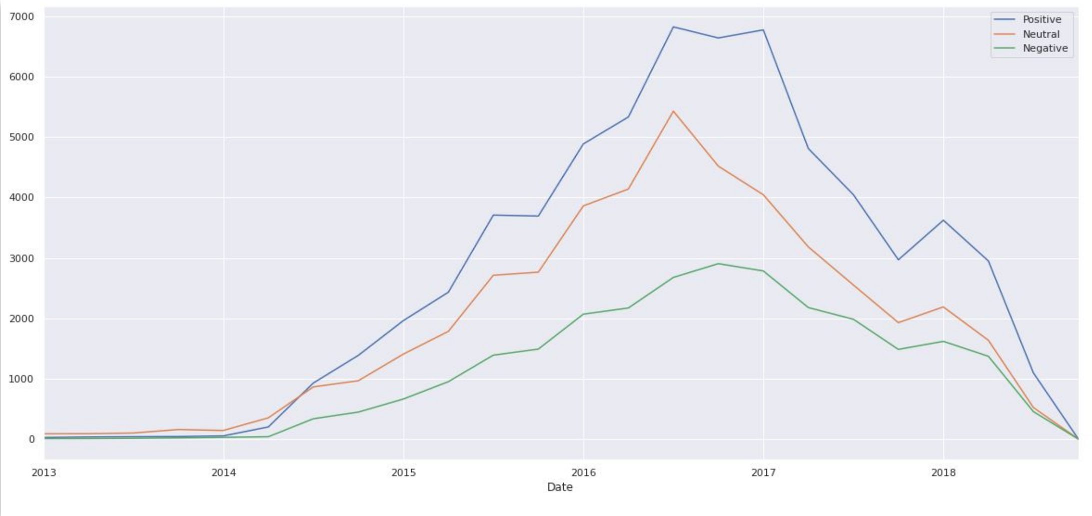
Time Series Analysis(Contd.)

Data Understanding :

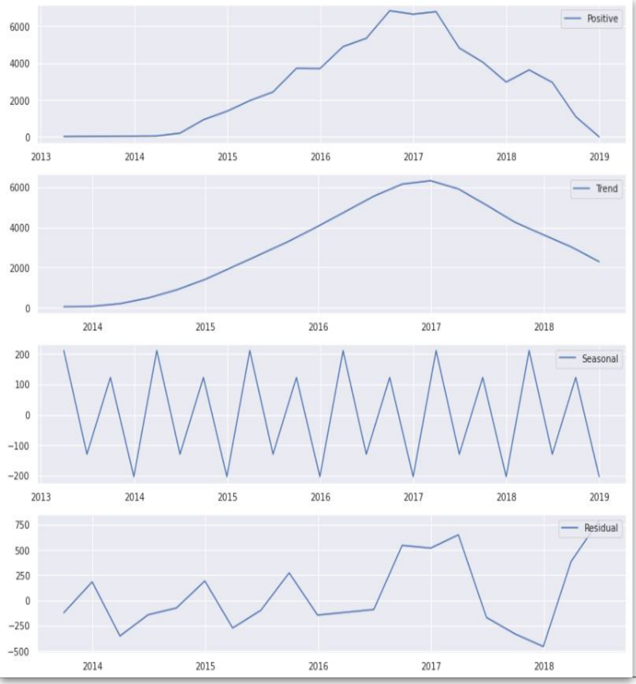
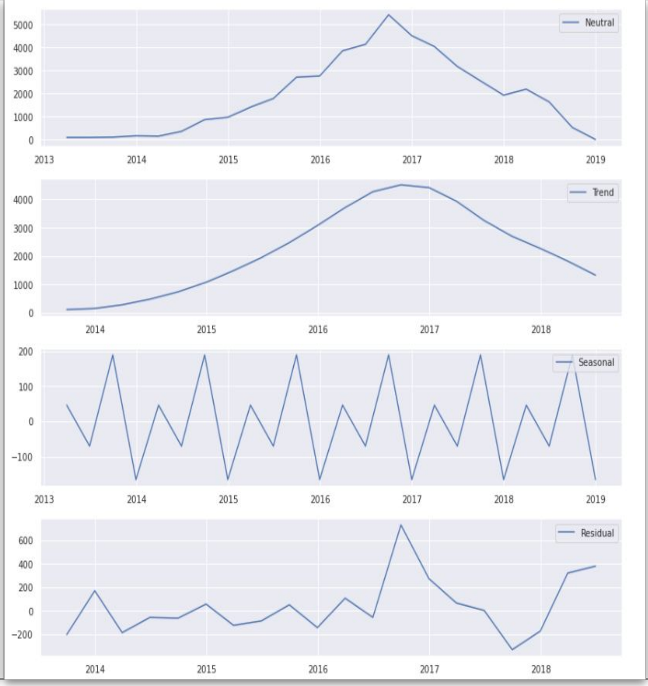
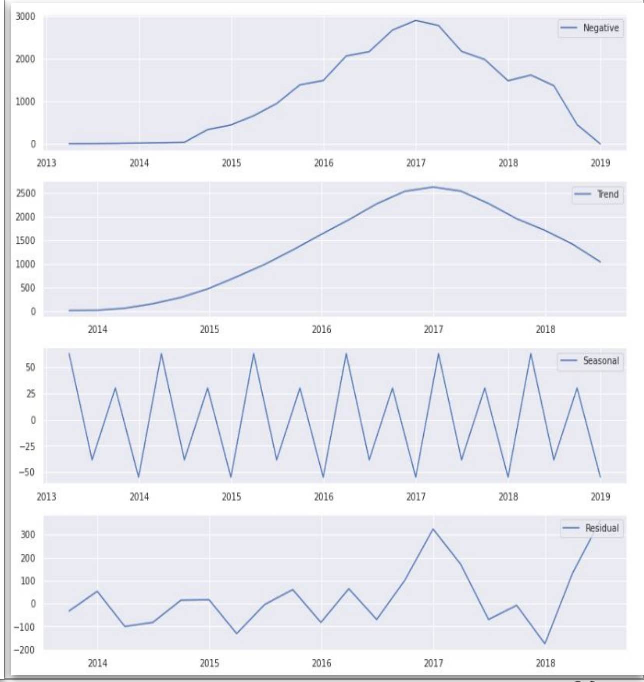
Split overall sentiments into 3 different column Positive,Neutral,Negative.

Analysis	Negative	Neutral	Positive	total
Date				
2006-06-30	0	3	0	3
2006-09-30	0	2	0	2
2006-12-31	0	3	0	3
2007-03-31	0	0	1	1
2007-06-30	0	0	0	0

Plot of Sentiments:



Decomposition Plots

Positive Sentiments	Neutral Sentiments	Negative Sentiments
		

Time Series Analysis(Contd.)

Stationarity Check

We have done stationarity check through Augmented Dickey Fuller test.

ACF

Autocorrelation function (ACF) is a measurement of how related the actual value is to the previous values including trend and seasonality.

PACF

Partial Autocorrelation function (PACF), unlike the ACF, finds the correlation between the residual values in the series,therefore it is only the partial function.

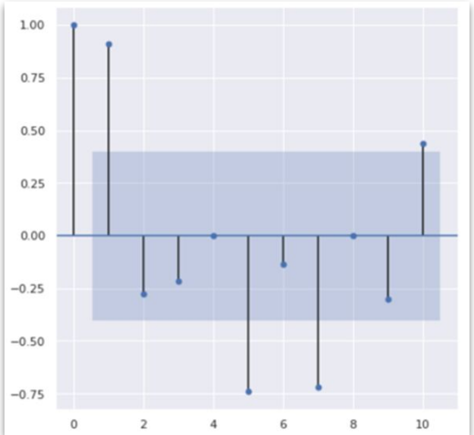
```
def checkStationarity(data):
    pvalue = adfuller(data)[1]
    if pvalue < 0.05:
        ret = "Data is Stationary. Proceed to model building"
    else:
        ret = "Data is not Stationary. Make it stationary"

    return(ret)
```

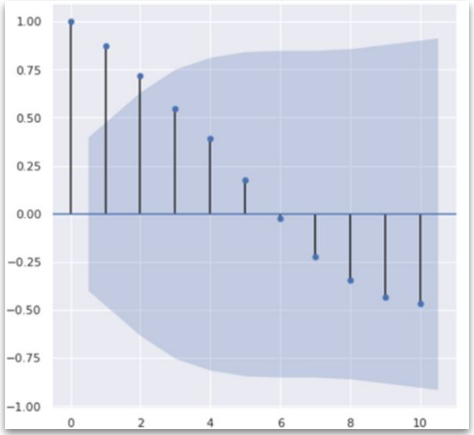
```
# checking for stationarity of the positive sentiment data
checkStationarity(Positive)
```

```
'Data is Stationary. Proceed to model building'
```

PACF PLOT (p)



ACF PLOT(q)



Time Series Models

ARIMA is a [p,d,q] model, p takes values upto previous p periods, q takes residuals upto q lags and d degree of difference done on data.

SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. All things remain same as ARIMA model but new seasonality elements are added.

Selecting p,q,P and Q values for positive sentiment based on low AIC Score.

	p	q	P	Q	AIC
32	1	2	0	2	221.144916
35	1	2	1	2	222.556143
14	0	2	0	2	227.318174
17	0	2	1	2	228.425169
26	1	1	0	2	239.514289

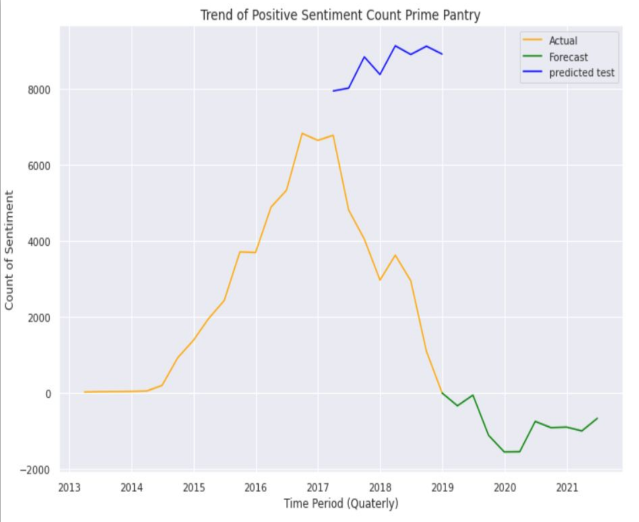
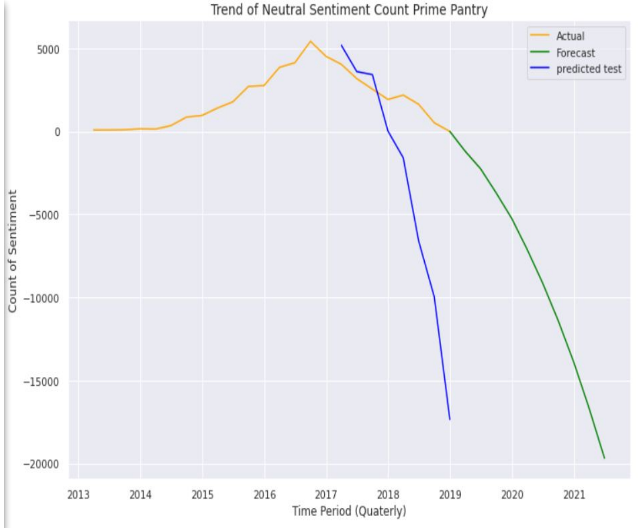
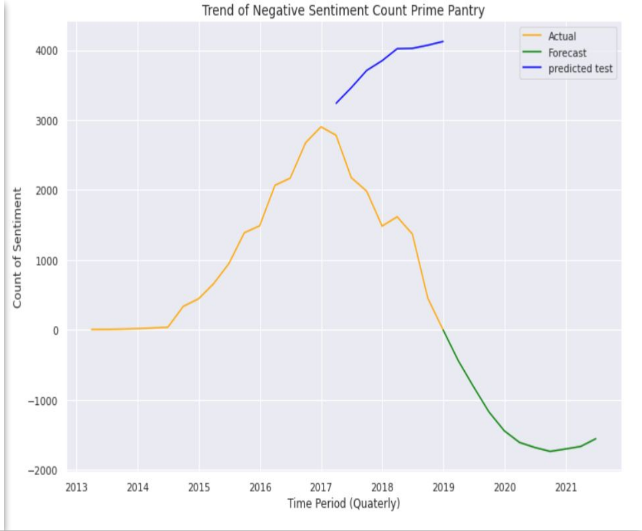
Selecting p,q,P and Q values for neutral sentiment based on low AIC Score.

	p	q	P	Q	AIC
14	0	2	0	2	158.042808
17	0	2	1	2	158.376535
32	1	2	0	2	159.790948
35	1	2	1	2	161.815118
26	1	1	0	2	176.791824

Selecting p,q,P and Q values for negative sentiment based on low AIC Score.

	p	q	P	Q	AIC
67	2	2	0	3	143.863235
43	1	2	0	3	143.938955
47	1	2	1	3	144.981551
71	2	2	1	3	145.694613
23	0	2	1	3	147.789784

Sentiments Forecasting

Positive Sentiment Forecasting	Neutral Sentiment Forecasting	Negative Sentiment Forecasting
		

1. From the plot we can say that the positive sentiments of customers will decrease till 2020 and after that it will increase, as per shopping habits of people.
2. From the plot we can say that the neutral sentiments of customers will gradually decrease ,as per shopping habits of people.
3. From the plot we can say that the negative sentiments of customers will decrease till 2020 and after that it will slightly increase, as per shopping habits of people.

Conclusion

Customer we should focus on

	reviewerID	reviewerName	Analysis	Count_shopping	freq_time
18842	A35Q0RBM3YNQNF	M. Hill	Neutral	168	6 days 14:34:17.142857142
29907	ANDVNCX6JU4XW	SHERRY MCCAUGHTRY	Positive	124	0 days 08:19:21.290322580
765	A13J2PGKNMJG1K	LegoGirl	Neutral	118	11 days 11:35:35.593220338
29752	AMMNGUJK4HQJ5	Misty	Neutral	116	12 days 06:12:24.827586206
10362	A26K3T6L5NYO7L	PennyPincher	Neutral	111	10 days 00:25:56.756756756
10066	A25DP3DWUXSS48	KT	Neutral	93	11 days 22:27:05.806451612
17185	A2YKWYC3WQJX5J	ShannonOnTheLakes	Neutral	79	44 days 09:43:17.468354430
29271	AKPG8VQBS0MWR	Old Coast Customer	Neutral	78	10 days 15:23:04.615384615
20943	A3EF7PUYTF057Z	Gary R. Jordan	Positive	73	20 days 18:04:55.890410958
4764	A1JN63QBBNGB78	Elle S	Neutral	72	17 days 02:20:00
2756	A1BT9J2I6DC246	Debbie	Positive	72	14 days 19:40:00
26368	A92ZKEZI137M1	Lisa M. Rainer	Negative	67	4 days 05:22:23.283582089
21115	A3F9UAX22LLZWK	K K Schwartz	Positive	67	16 days 02:51:56.417910447
14668	A2O421DTA8J0RW	Dogs & Horses	Neutral	65	19 days 05:32:18.461538461
10501	A276RHM6BBPDTY	Ddee	Positive	64	20 days 06:45:00
2430	A1AB6D301MOTM0	Lynn G.	Neutral	63	7 days 10:17:08.571428571
32317	AXK37UZY8UPYP	Que Sera Sera	Positive	62	13 days 00:23:13.548387096
7833	A1W511P7B2QSQE	MariamG	Neutral	62	10 days 08:07:44.516129032
15417	A2R1HUYHXV7H18	Bugs	Neutral	61	41 days 21:14:45.245901639
12987	A2H9H3BVFNS3Y0	Leah D	Positive	61	10 days 22:01:58.032786885

Customer would likely to Churn out

	reviewerID	reviewerName	Analysis	Count_shopping	freq_time
0	A0526222H977CBZM4DK7	JAIME SCARPITTA	Negative	1	192 days
19987	A3AE8HSBCSLYX4	sly	Negative	1	35 days
6968	A1SU2TR45U1VB4	Chelsie Luchini	Neutral	1	74 days
19988	A3AE8HSBCSLYX4	sly	Neutral	1	35 days
6962	A1STPMGQSC12NS	valeri	Positive	1	226 days
19991	A3AEAMF75QS4WB	Amazon Customer	Neutral	1	199 days
19993	A3AECY1VS8T6V5	Y. Pope	Negative	1	814 days
20002	A3AG3ZHG78N4M5	Pat	Negative	1	683 days
20006	A3AGK4J9PHB6XE	Sbuxgirl36	Neutral	1	316 days
6952	A1SSOLFAUR915J	Crystal	Neutral	1	474 days

Conclusion(Contd.)

AMAZON Prime Pantry Brand Recommendation Analysis

Total No of Brands

137,769

Most Recommended Brand

Product B000WHTWD2 from Brand Mott's

355 Positive Review

Most Non-Recommended Brand

Product B0014CZUZ8 from Brand Planters

138 Negative Review

Overall Brand Sentiment

Asin	Brand	Analysis: Nega..	Neutral	Positive	Grand To
B0014CZUZ8	Planters	138	228	295	66
B00HZ6X8..	Tide	112	174	287	57
B0014CV2IW	Band-Aid	105	115	200	42
B0014D5N..	Glad	104	158	255	51
B000Q5NG..	Progresso	77	104	143	32
B00UJIXP0A	Planters	67	90	126	28
B0014CUC..	Quaker	64	99	175	33
B00SH4ORI2	Campbell's	63	53	64	18
B0111K4XHS	Planters	54	65	149	26
B015O6PS..	Frito Lay	52	66	129	24
B000RYBU..	Dole	50	28	95	17
B00R2O5JD0	Oreo	48	41	105	19
B000OV4FY2	Dole	47	57	138	24
B00JGZBTCI	Downy	46	62	155	26
B00R1J3N84	Frito Lay	45	56	119	22
B00ID0DTA..	Frito Lay	45	49	114	20
B000Q5LW..	Progresso	44	35	70	14
B00R1H0A..	Nature Vall..	43	42	103	18
B000YK0S58	Band-Aid	43	67	70	18
B00IB1D0IO	Barilla	42	30	106	17
B000Q5K1..	Dole	42	50	109	20
B00ID0DUY..	Frito Lay	41	46	94	18

Recommended Brands

Asin	Brand	Analysis: Positi..
B000WHTWD2	Mott's	355
B0014CZUZ8	Planters	295
B00HZ6X8QU	Tide	287
B0014D5NVS	Glad	255
B0014E84TK	Nutella	252
B00PC5CI8K	Rice Krispies Treats	245
B0013FRRMK	Method	212
B0010XS2CE	Softsoap	209
B00PTX8N52	Dawn	206
B0014CV2IW	Band-Aid	200
B00I9JK8GU	Ziploc	182
B00NP79AI8	Nestle Pure Life	180
B0014CUCU6	Quaker	175
B0014E2SOC	Scotch-Brite	174
B00IAE6PWA	Pepperidge Farm	167
B015TYDZTE	Bounty	166
B00QTX0D18	Kellogg's	165
B00RGVOUIO	Ritz	164
B00099XPD6	Carnation	161
B00NAZ532W	Viva	160
B00JGZBTCI	Downy	155
B0014CSG5Y	Huy Fong	154
B0111K4XHS	Planters	149

Non- Recommended Brands

Asin	Brand	Analysis: Nega..
B0014CZUZ8	Planters	138
B00NAZ532W	Viva	122
B000WHTWD2	Mott's	118
B00HZ6X8QU	Tide	112
B01B4N9Y6C	Cottonelle	106
B00PC5CI8K	Rice Krispies Treats	106
B0014CV2IW	Band-Aid	105
B0014D5NVS	Glad	104
B0014E2SOC	Scotch-Brite	98
B0014E84TK	Nutella	90
B0014D2CIU	Reynolds	84
B015TYDZTE	Bounty	82
B00NP79AI8	Nestle Pure Life	81
B00IAE6PWA	Pepperidge Farm	77
B000Q5NG78	Progresso	77
B000WCZ57Y	Terra	68
B011DX5G92	Charmin	67
B00UJIXP0A	Planters	67
B00099XPD6	Carnation	67
B00NAZ5JQ2	Cottonelle	66
B0014CUCU6	Quaker	64
B013OV42NW	Slim Jim	63
B00SH4ORI2	Campbell's	63

Total Sentiment

137,769

Positive

63,940 (46.49%)

Neutral

46,270 (33.64%)

Negative

27,323 (19.87%)

Business Insights

- We will recommend our classification model to predict the sentiments of the reviews related to prime pantry products.
- We will recommend our clustering model to identify in which cluster the new customers will fall based on their shopping habits.
- The positive sentiments of customers will decrease till 2020 and after that it will increase. So we can suggest the products from various brands that got moderate to low rating, try to improve the quality of the products so that the positive reviews remain constant or increase gradually.
- The neutral sentiments of customers will gradually decrease. So we may say that the neutral reviews will move towards either positive or negative reviews. So we can suggest the products from the brands that got the neutral reviews more, upgrade the quality or sales so that it will move towards positive reviews more.
- The negative sentiments of customers will decrease till 2020 and after that it will slightly increase. So we will recommend to increase the advertisements, simultaneously improve the quality of product and sales so that more people will buy those products & give more number of reviews.

Thank You