AMAZON CAPSTONE PROJECT

TABLE OF CONTENT

	K mean to determine highly demanded products. Model 8 - Product Clustering of main category "Office Products" using K mean to determine highly demanded products. Model 9 - Time series analysis to forecast demand of top selling product in Digital Music Model 10 - Timer series analysis to forecast demand of top selling product in Office Products.
8. Reference	References.

INTRODUCTION

Amazon Product Review Analysis The year was 1994 when Jeff Bezos launched Amazon from his garage. In 1995, the first product launched by Amazon was a book in 50 states and in 45 countries within 30 days. (Oberlo 2021) Within 26 years, Amazon became the world's largest online retailer and a household name. The Amazon name has become synonymous with online shopping and continues to grow by developing new products, acquisitions, and numerous service offerings to enlarge the customer base. Nowadays, almost 150.6 million people turn to the Amazon app for most everything. Several types of research have proven that customers trust Amazon. (Statista 2019) On average, the small and medium-sized businesses located in the USA sell more than 4,000 items per minute (Amazon 2019), which leads to millions of product reviews on Amazon.

PROVIDED DATA SET

Main categories – CDs & Vinyl and Office Products

In which we have 6 files:

1. CDs & Vinyl Meta data

2. Office Products Meta data

3. CDs & Vinyl Review data

4. Office Products Review data

5. CDs & Vinyl Rating data

6. Office Products Rating data

OBJECTIVES

Case Study: Amazon Product Review Analysis

- ❑ To develop an automated system to analyse and monitor an enormous number of reviews.
- ❑ By monitoring the entire review history, analyse tone, language, keywords, and trends over time to provide valuable insights that increase the success rate of existing and new products and marketing campaigns.

SCENARIOS

Scenario 1:

Inventory Optimization and Demand Forecasting Optimize inventory management by identifying the product categories (clustering as an outcome of text processing) on the customer review data. Predict what kind of products could be in demand (Time Series Analysis).

Scenario 2:

Customer Retention and Sentiment Forecasting Customer retention strategy through feedback analysis (customer classification and clustering as an outcome of analyzing the review text). Trend and seasonality analysis to predict how frequently a particular category of customer would shop in the future. (Time Series Analysis) Time Series component: Trend, Seasonality Analysis to predict how frequently this the customer would buy new products.

PYTHON LIBRARIES & PACKAGES USED IN THIS PROJECT

LIBRARIES/PACKAGES	USE
Json	to work with json file
Pandas	to create and manipulate with dataframes
numpy	to work with numpy arrays
gzip	to extract work file from zip file
nltk	working with nlp algorithms, Vadersentiment
sklearn	to working with machine learning algorithms
sklearn.linear	Logistic Regression (Classification algorithm)
sklearn.feature extraction.text TfidfVectorizer	To convert text to numerical based on TF-IDF score
nltk.corpus	to detect stopwords
Re	Regular expression operations
sklearn.metrics: classification report accuracy score f1 score accuracy score precision score mean squared error confusion matrix	Classification report Evaluation metric
sklearn.model Traintestsplit	Train test split
Time	To check the processing time
sklearn.preprocessing: LabelEncoder	To convert categorical to numerical And scaling the data for modelling

MinMaxScaler StandardScaler	
Warnings warnings.filterwarnings('ignore')	To ignore the warnings
sklearn.cluster: KMeans	For Cluster Formation
statsmodels.tsa.seasonal: seasonaldecomposition	Time series components for seasonal decomposition
statsmodels.tsa.stattools: adfuller	To find the stationarity of the data
statsmodels.graphics.tsaplots: plot acf plot pacf	To plot ACF and PACF plots
statsmodels.tsa.arima.model: ARIMA	To build the ARIMA model
statsmodels.tsa.statespace.sarimax: SARIMA	To build the SARIMA model
matplotlib.pyplot seaborn	Visualization tool

DATA DICTIONARY

META DATA DICTIONARY

‘title’: the name of the product on amazon
‘brand’: displays the Brand of the products.
‘rank’: product’s performance compared to product of same category
‘main_cat’: Main Category of the product
‘price’: The cost of the product in Dollars ($)
‘asin’: ASIN stands for Amazon Standard Identification Number
‘category’: organized all their products under categories & sub-categories, structuring them like a tree and branching them into specificity.

REVIEW DATA DICTIONARY

‘reviewerID’: Unique Identification of reviewer/ customer

‘reviewTime’: Date on which review is published
‘verified’: Whether review is published by the actual buyer or not
(bool – True/False)
‘reviewText’: The review published by the customer
‘overall’: rating entered along with review
‘summary’: summarized version of reviewText.
‘asin’: Amazon Standard Identification Number of the product for which the customer has provided the review

RATING DATA DICTIONARY
‘rating’: the rating provided by the customer (out of 5)
‘reviewerID’: Unique Identification of reviewer/ customer
‘date’: date on which rating was published
‘asin’: Amazon Standard Identification Number of the product for which the customer has provided rating

DATA CLEANING

For Meta Data:

1. Deleting the unnecessary columns like [‘tech1', 'fit', 'tech2', 'imageURL', 'imageURLHighRes', 'description', 'similar item', ''details'', 'date', 'vote', 'image', 'style'] from the meta dataset.

2. Primary category, sub category and product type are extracted from the category column with the help of for loop and slicing technique. The null values present in primary category or sub-category were treated by using conditional imputation. For all the values present in primary category was matched with the values of main_cat column and mode were used to impute the null values. Similarly using conditional and relational logic between primary and sub-category column, the null values of sub category were imputed using mode corresponding to the mode of sub category of the corresponding primary category.

3. Likewise, rank is extracted from rank column.

4. Preformed data cleaning on the columns as there were empty list ([]) in the data which were not read has null values. The data was cleaned and pre-processed by using for loop. In the loops ‘&’ and ‘&amp’ is replaced with ‘and’. Finally, replaced [] and ‘ ’ with np.nan, function of ‘numpy’ library used to mark empty cells as null and then fill that null values relevant values.

5. Price column had various noise values. The noise values were initially replaced with np.nan. Special Characters like ‘$’, ‘-’, ‘,’ were removed from the price column. Finally the data type of ‘price’ column was changed from string to float. Likewise, rank column was also pre-processed and converted to float data type.

6. In Brand column there was also empty list [] therefore replaced them with nan and eventually replace them with Unknown.

7. Finally, there existed null values in only 3 columns rank, price and product type which were imputed later on.

8. Converted into .csv format and extracted the Dataframe.

For Review data:

1. Deleting unnecessary olumns like vote, image, style from the review data.

2. Checked for the null values where columns like – ‘reviewerName’, ‘reviewText’ and ‘summary’ contained of null values.

3. The null values in ‘reviewerName’ were imputed with ‘Amazon Customer’.

4. For ‘reviewText’ and ‘summary’ we have used user defined function called categorize. For summary column, if overall = 1 it will be imputed with ‘one star’, if overall=2 then it will be imputed with ‘two star’, if overall=3 then it will be imputed with ‘three star’, if overall=4 then it will be imputed with ‘four star’, if overall=5 then it will be imputed with ‘five star’.
   Similarly for reviewText, , if overall = 1 it will be imputed with ‘poor’, if overall=2 then it will be imputed with ‘bad, if overall=3 then it will be imputed with ‘average, if overall=4 then it will be imputed with ‘good’, if overall=5 then it will be imputed with ‘great’.

5. Then for Unixreviewtime column was convert into date time format and the reviewdate was extracted.

6. Converted into .csv format and extracted the Dataframe.

For Rating data:

1. First Rename all the columns name by comparing and analysing the data with other datasets.

2. Then the Unixreviewtime is converted into datetime format and date is extracted.

3. Duplicate rows were observed which were dropped from the dataset.

4. Converted into .csv format and extracted the Dataframe.

Merging of Datasets:

1. Office product review file is first merged with office product meta data. Null values are checked and imputed using mode as the number of null values was extremely low. Duplicate rows were observed which were dropped from the dataset.

2. Same process was applied to the CDs & Vinyl data as well.

3. Finally, both merged files were concatenated and a final Data frame of review data was achieved.

4. Using iterative imputer price and rank null values are imputed in the data

5. Converted into .csv format and extracted the Data frame.

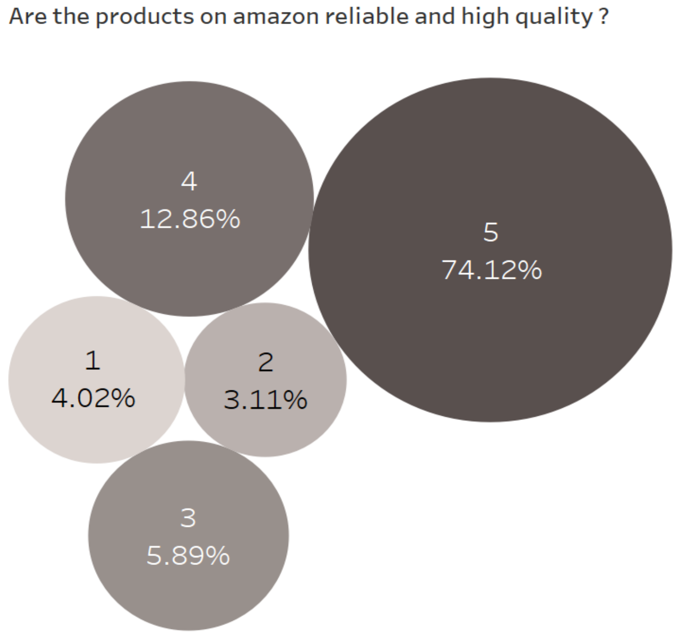
6. Similarly, Office product rating file is first merged with office product meta data. Null values are checked and rows containing mostly null values are dropped using dropna function. Same process was applied to the CDs & Vinyl data as well.

7. Finally, both merged files were concatenated and a final Data frame of review data was achieved.

8. Using iterative imputer price and rank null values are imputed in the data

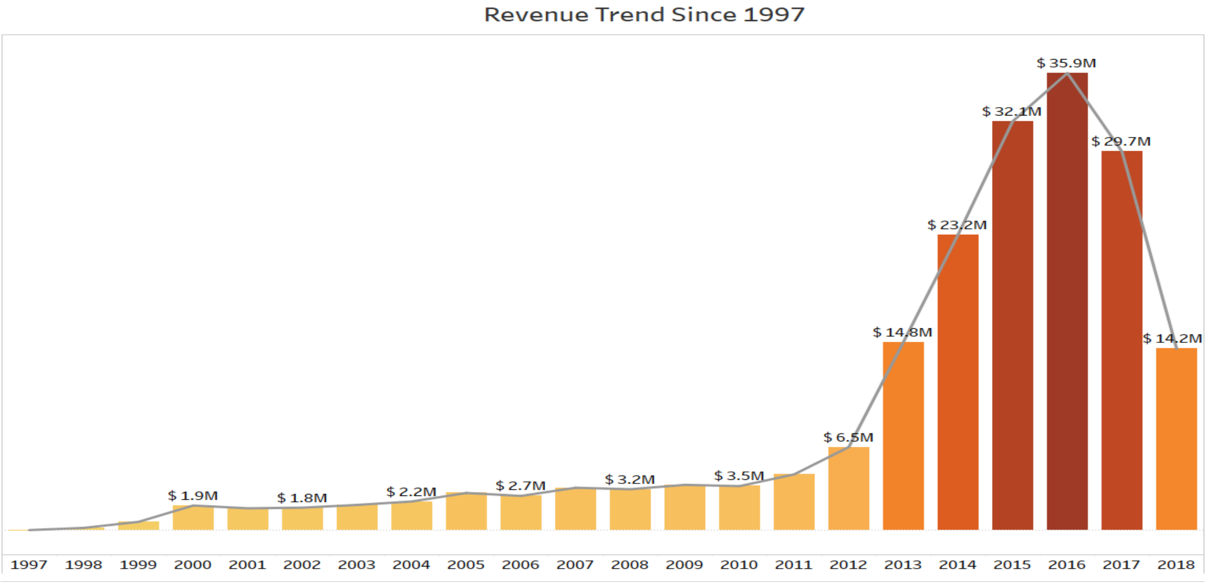
9. Converted into .csv format and extracted the Data frame.

PRELIMINARY ANALYSIS



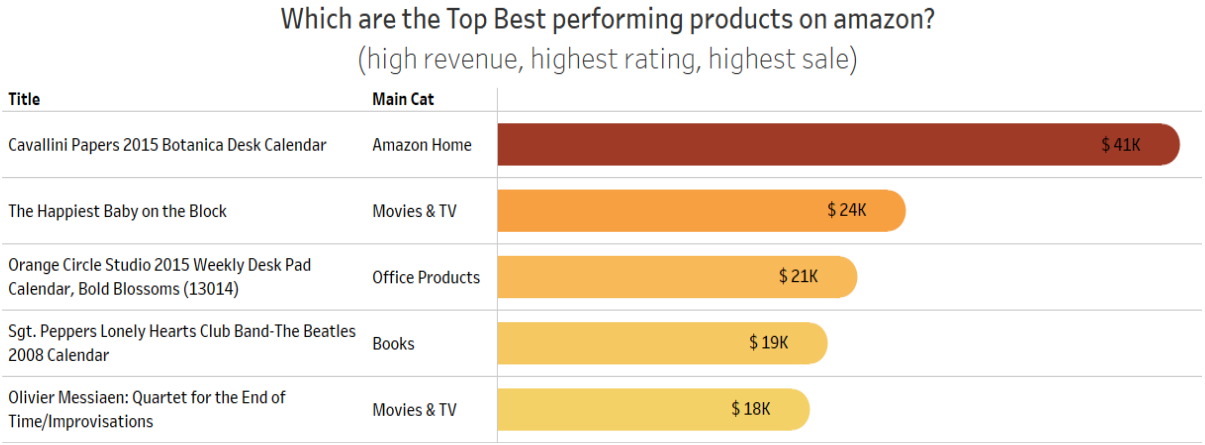
Yes, with the help of this graph we can see that most of the product have high rating which means customer are happy with the products. Most of the products have high rating so we can say that products on Amazon are reliable and high quality.

What is the Revenue Trend since 1997?



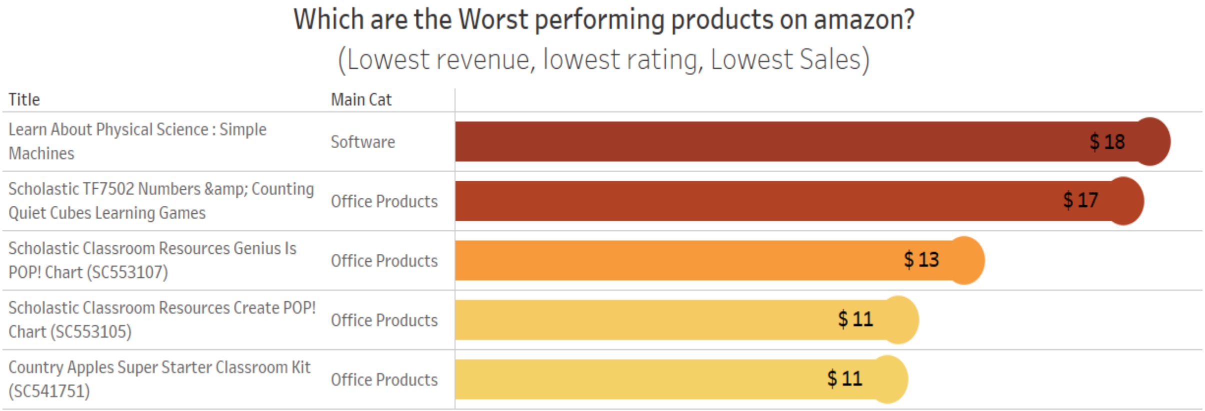
At the beginning of year 1997 total revenue increased slowly till 2012, post 2012 total revenue started increasing with high growth rate. It has the highest sales in the year 2016 after which it started declining rapidly.

Which are the Top 5 best performing products on Amazon?



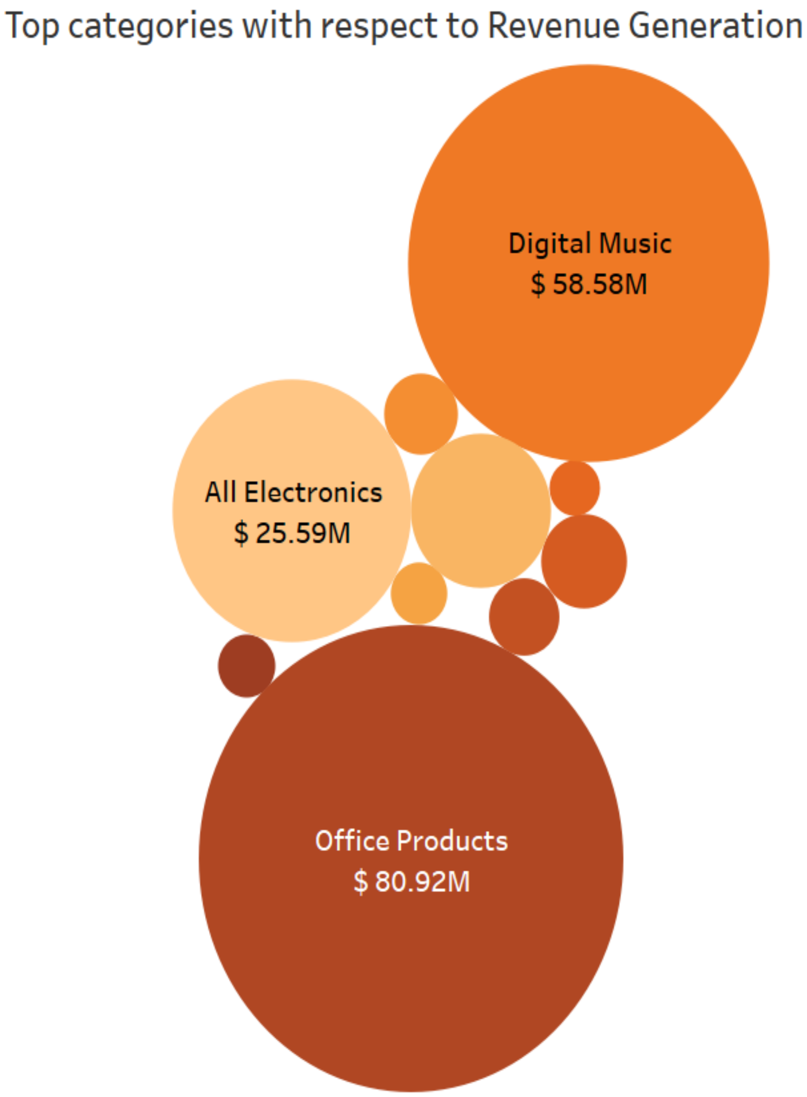
So this are the top 5 best performing products on Amazon on the basis of high revenue, high rating and highest sale. In which cavallini papers 2015 botanica desk calender from Amazon home is the highest as compare to the others**.**

Which are the worst performing products on Amazon?



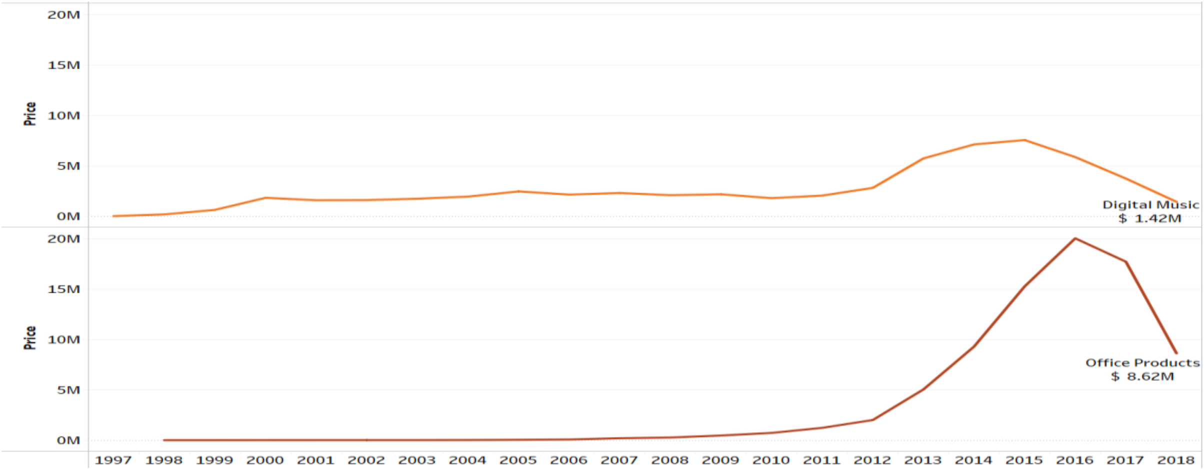
So this are the top 5 worst performing products on Amazon on the basis of lowest revenue, lowest rating and lowest sales. In which country apples super starter classroom kit from office products is the worst as compare to the others**.**

What are the top categories with there Revenue?



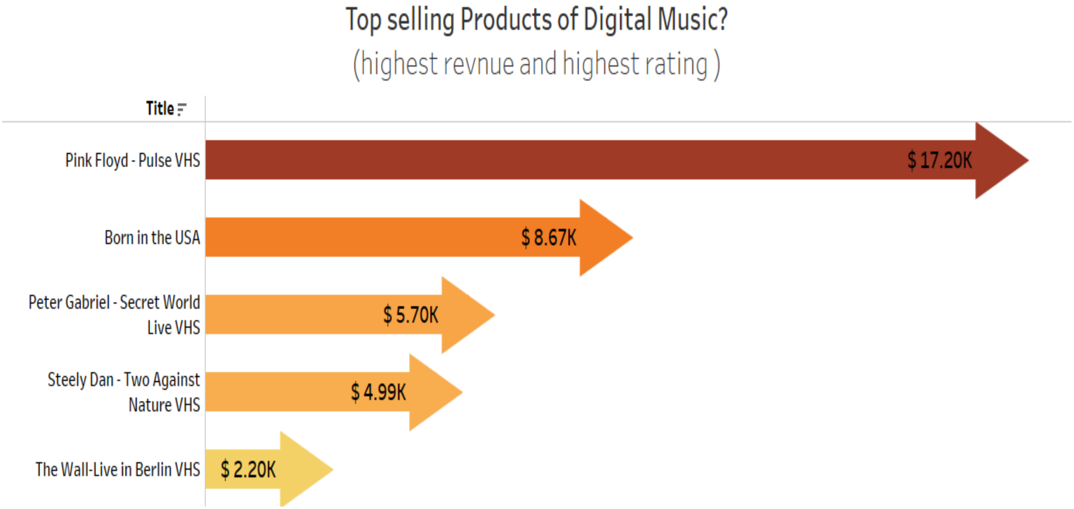
So this are the top categories with respect to Revenue Generation. In which office products has the highest revenue after that digital music and all electronics.

What is the Revenue Trend of the top selling categories?



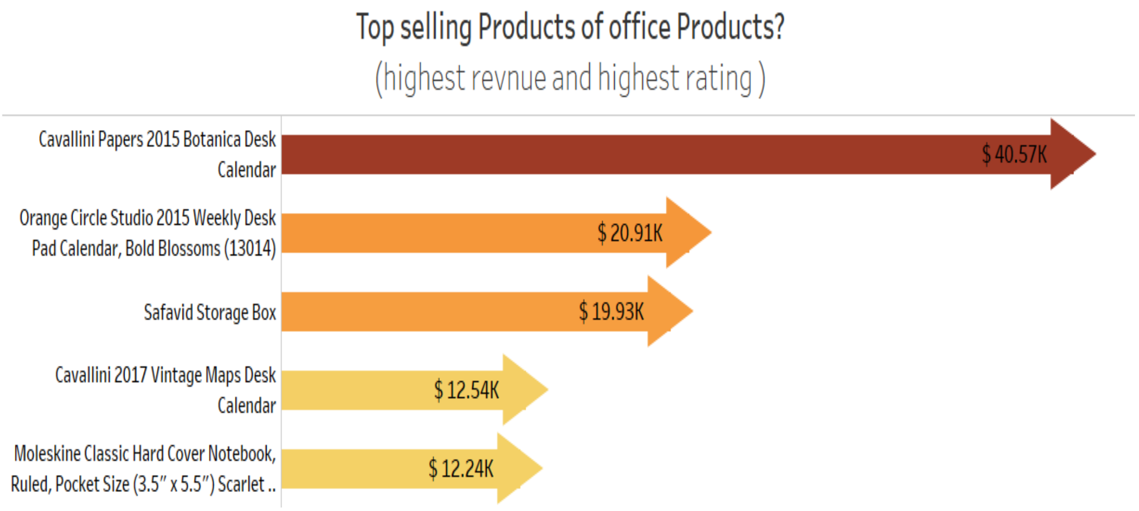
So with help of this above graph we can see that for digital music it is always increasing from the beginning and it is highest at the year of 2015 and after that its decreasing but for office products at the beginning we can see a steady line till 2012 which means there is no variation in the sales but after 2012 it starts increasing and it is highest in the year 2016 and starts falling down.

Top Selling Products of Digital Music?



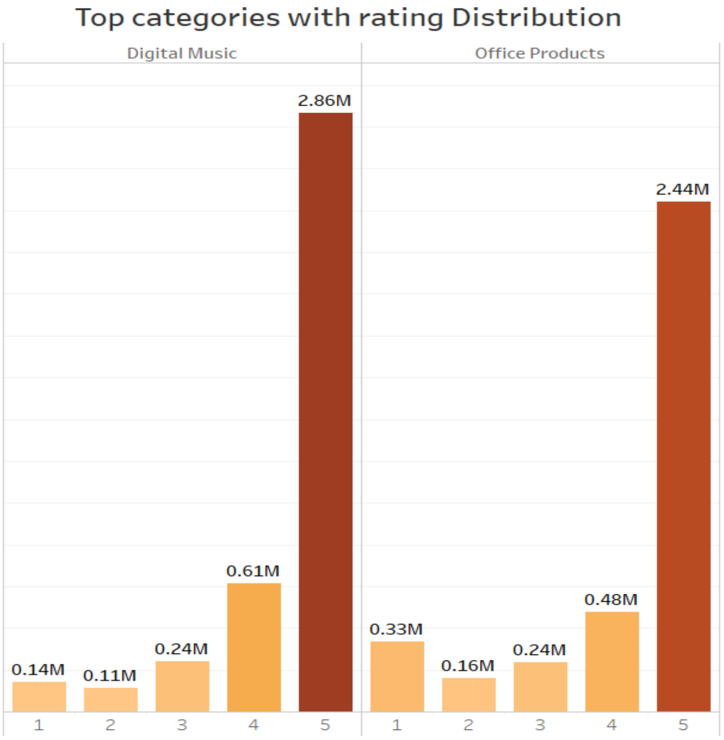
So this are the top selling Products of digital music on the basis of highest revenue and highest rating. In which pink Floyd-pulse VHS has the highest revenue with highest rating as compare to others.

Top Selling Products of Office Products?



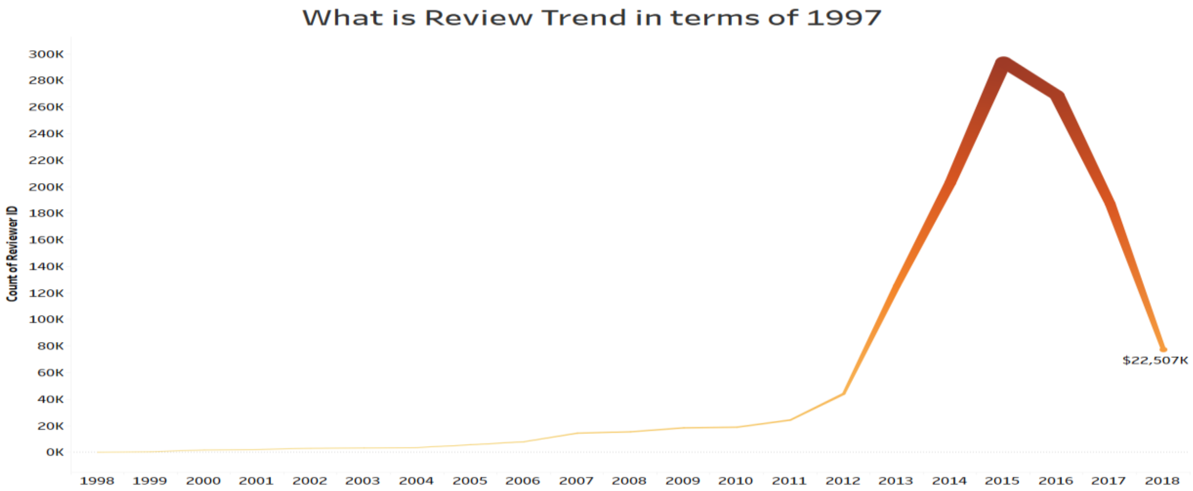
So this are the top selling Products of Office Products on the basis of highest revenue and highest rating. In which cavallini papers 2015 botanica desk calender has the highest revenue with highest rating as compare to others.

Top categories with there rating distribution?



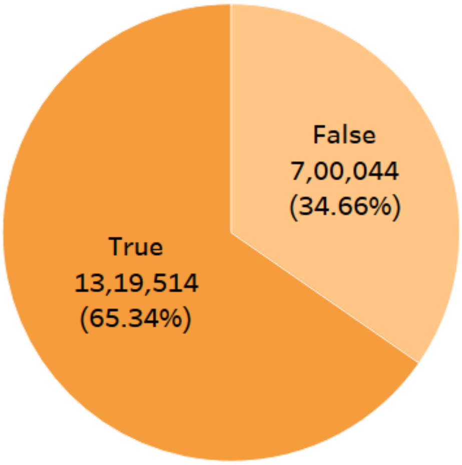
So with the help of bar charts we can see that rating 5 has the highest distribution in both the categories.

What is review Trend in terms of 1997?



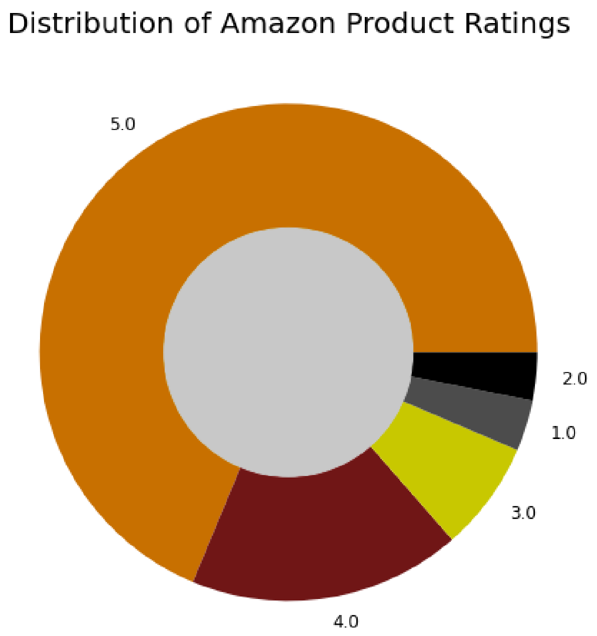
At the beginning of the year 1998 it starts increasing slowly till 2012 and after 2012 we can see a difference in the increase in the review till 2016 and in the year 2016 it has highest review and after that it starts falling down.

Should we consider all reviews ?



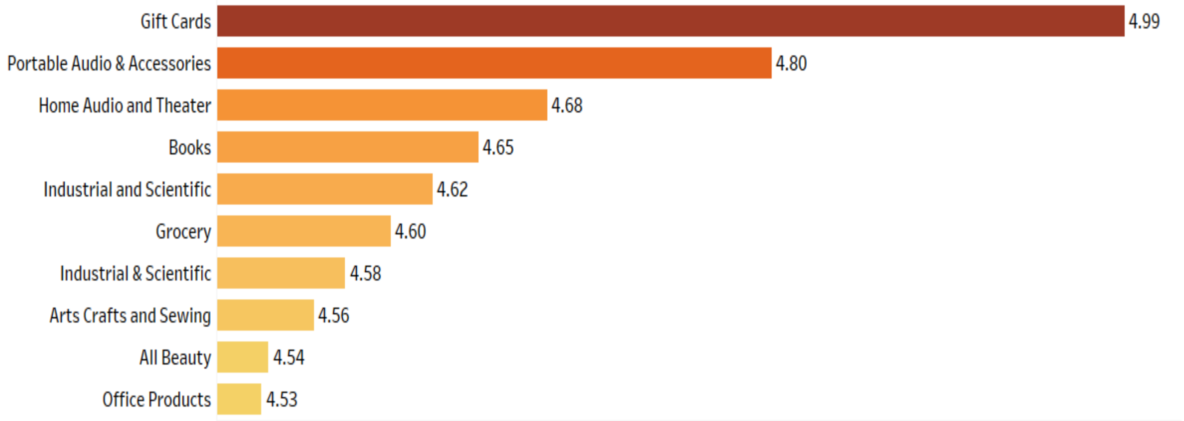
Since nearly 35% reviews are not by verified customers, any analysis should be made on only verified reviews

Distribution of Amazon product Ratings?



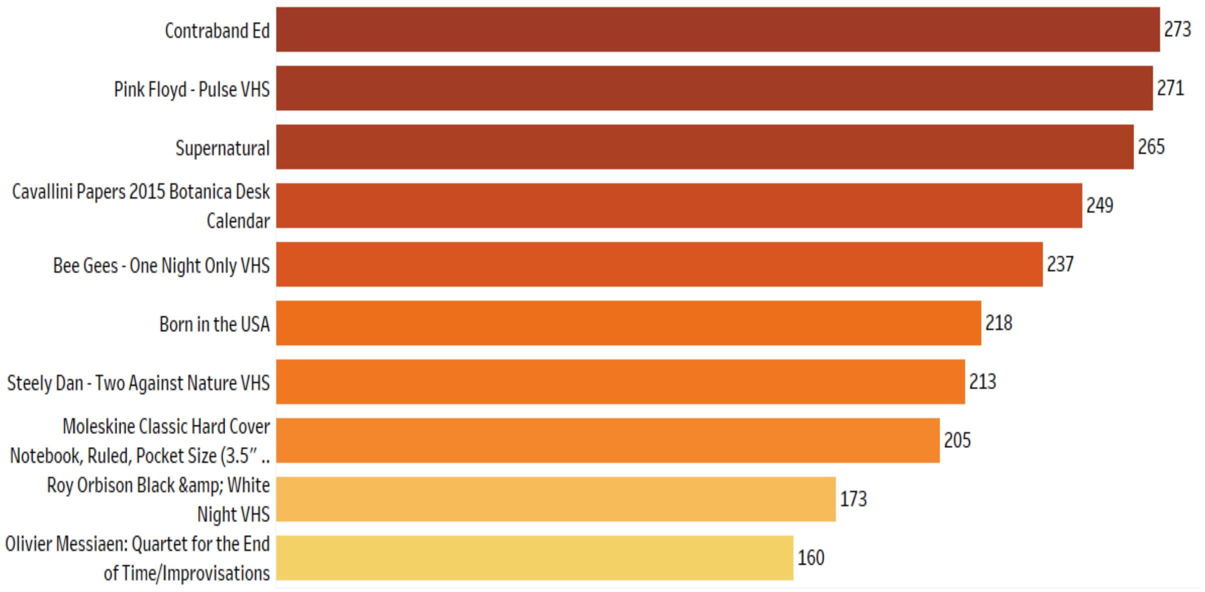
We can see with the help of the donut chart the rating 5 has the highest distribution of Amazon products as compares to others rating.

Top 10 Categories with highest average overall score



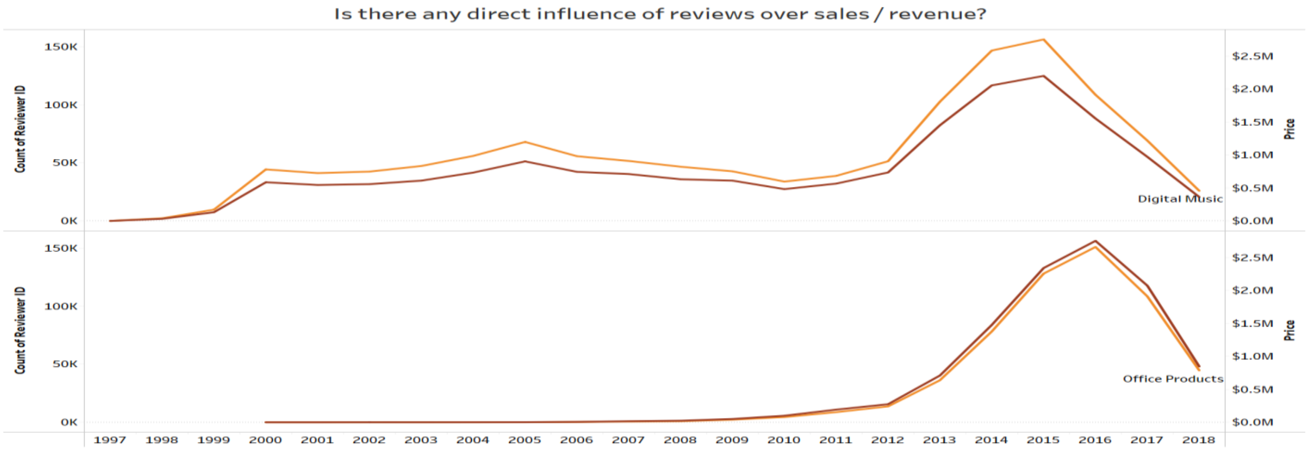
So this are the top-10 categories with highest average overall score in which Gift Cards has the highest average overall score as compare to others.

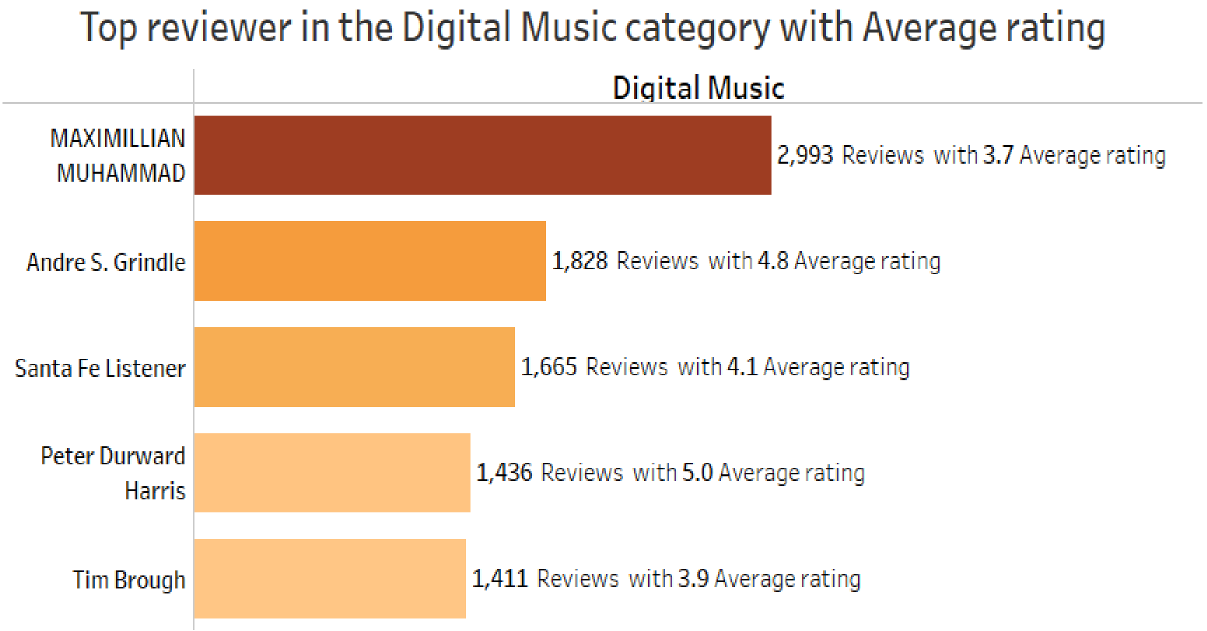
Title of Top 10 products have received highest number of reviews



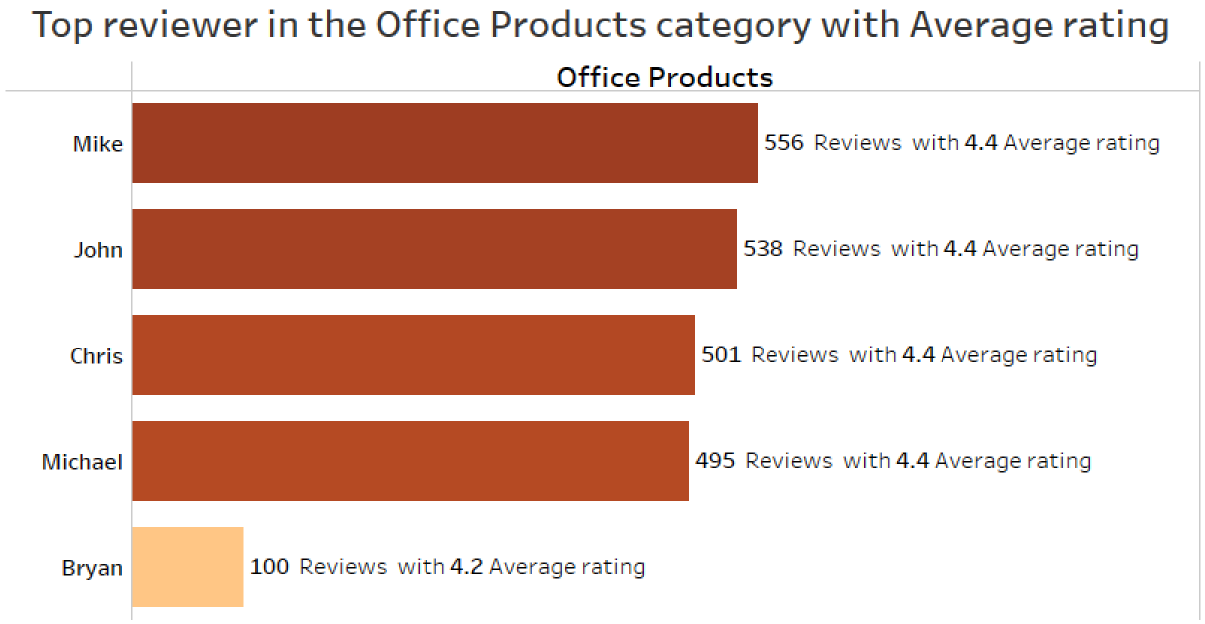
These are the title of top 10 products which have received highest number of reviews. In which contraband Ed has the highest number of reviews.

Is there any direct influence of reviews over sales/revenue?





So this are the top reviewer in Digital music category with average rating. In which Maximillian Muhammad has the highest number of reviews with average Rating of 3.7 .



So this are the top reviewer in Office Products category with average rating. In which Mike has the highest number of reviews with average Rating of 4.4 .

MODELLING

Model 1 – Sentiment analysis using NLP

Analyzing The Product Reviews Sentiment.

Procedure

1. Importing the libraries and final review dataset.

2. The Dataset is huge therefore, its recommended to sample the data for easy & quick conclusion.

3. Sampling the data with respect to the overall column, so that equal sample be extracted.

4. Removing all the stops words and do the all the text cleaning process.

5. For sentiment analysis we have used VaderSentiment (Valence Aware Dictionary and sEntiment Reasoner)

6. Generating sentiment category for the sample data for future analysis.

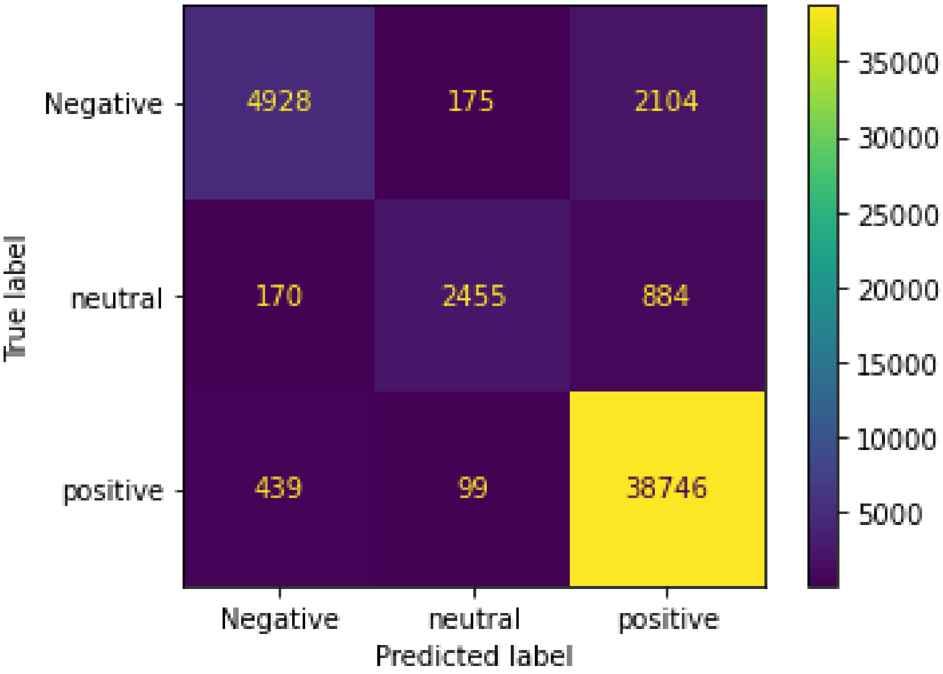
7. After Generating sentiments, TF-IDF Vectorizer Model is created and trained with the train data on behalf of 80:20 ratio.

8. Checking different classifier model for best positive sentiment F1-score.

9. For Modelling We Are Using Logistic Regression along with One-vs-Rest values for test data is predicted and Metrics are calculated.

10. Basically confusion metrics is plot to know that how many words belongs to which sentiment and find out which sentiment is truly predicted.

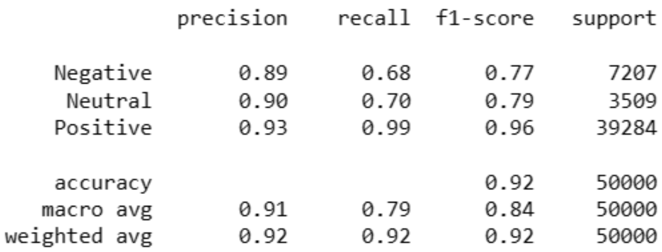
11. Confusion Metrics is plotted to check how many predicted values are correctly predicted.



From the above confusion metrics we can see that most of the words belongs to positive sentiments and which are truly predicted.

12. Classification replot is shown to know that which sentiments has the highest F1-score and which also say the accuracy of the model.

13. After Computing the classification report the F1 score of predicting Positive sentiment is the highest with 96% F1



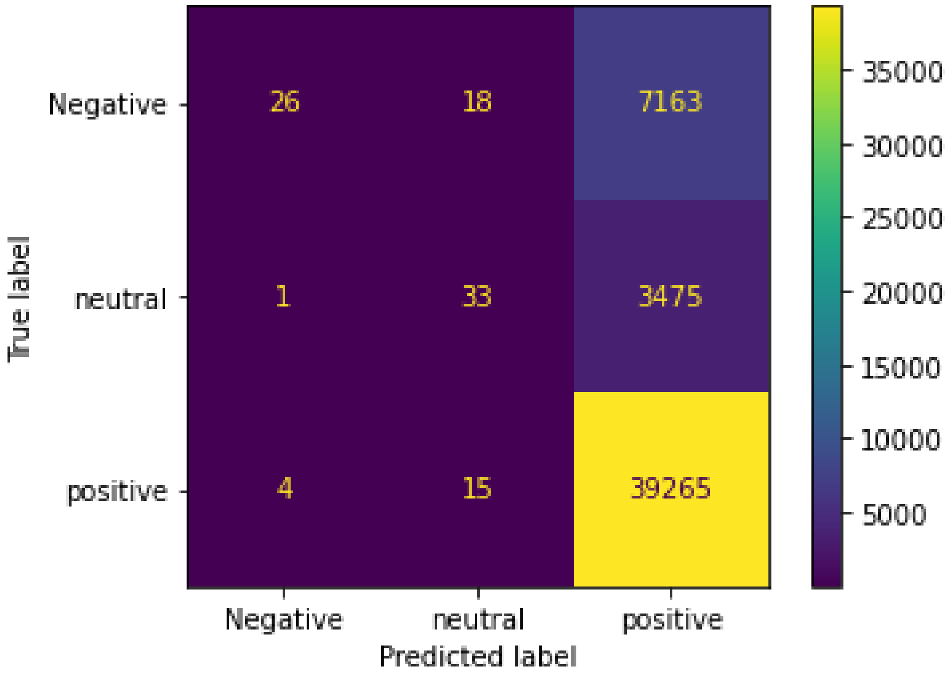
```
              precision    recall  f1-score   support

    Negative       0.89      0.68      0.77      7207
     Neutral       0.90      0.70      0.79      3509
    Positive       0.93      0.99      0.96     39284

    accuracy                           0.92     50000
   macro avg       0.91      0.79      0.84     50000
weighted avg       0.92      0.92      0.92     50000
```

14. From the classification report we can see that 77% of F1-score of negative sentiments 79% of neutral and 96% of positive sentiments.

15. Which says the individual F1-score of every sentiment and positive sentiments has the highest F1-score compares to others.

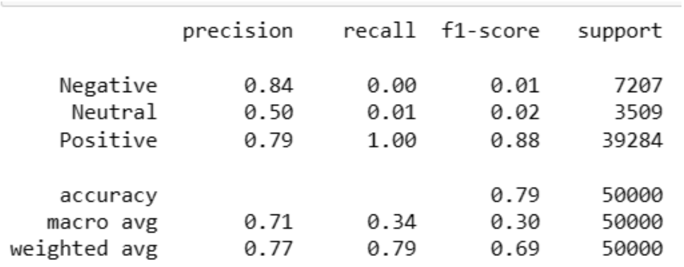
16. Again Doing Modelling Using MultinominalNB Classifier for test data is
predicted and Metrics are calculated.

17. Confusion Metrics is plotted to check how many predicted values are
correctly predicted.



From the above confusion metrics we can see that most of the words are positive sentiments and which are truely predicted.

18. After Computing the classification report the F1 score of predicting Positive sentiment is the highest with 88% F1



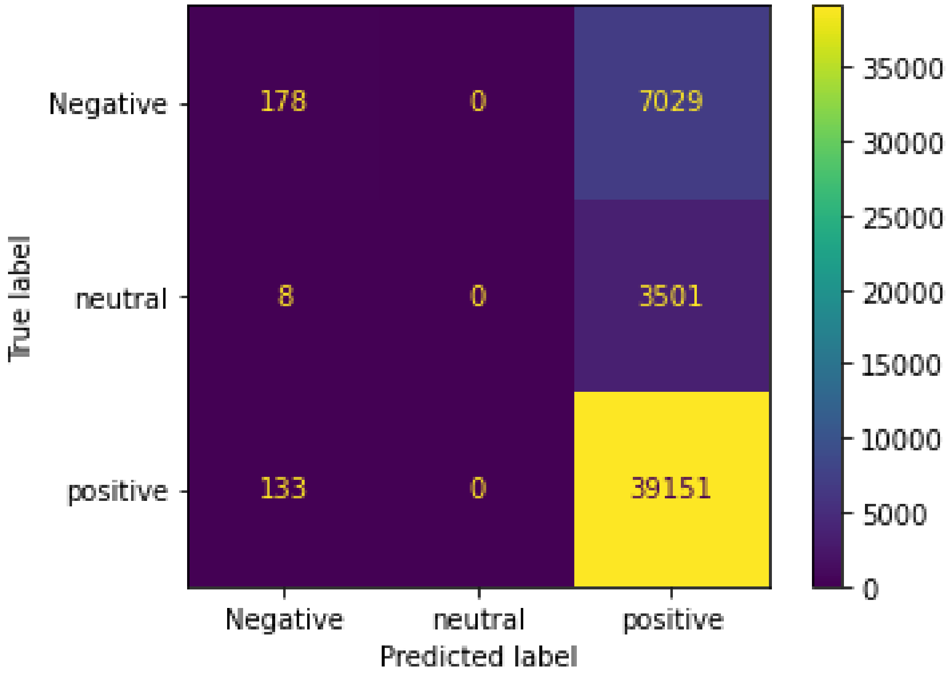
```
              precision    recall  f1-score   support

    Negative       0.84      0.00      0.01      7207
     Neutral       0.50      0.01      0.02      3509
    Positive       0.79      1.00      0.88     39284

    accuracy                           0.79     50000
   macro avg       0.71      0.34      0.30     50000
weighted avg       0.77      0.79      0.69     50000
```

From the classification report we can see that 10% of F1-score of negative sentiments 20% of neutral and 88% of positive sentiments.

Which says the individual F1-score of every sentiment and positive sentiments has the highest F1-score compares to others sentiments.

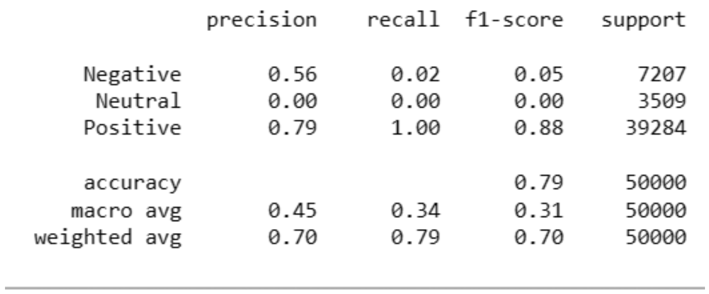
19. Again Doing Modelling Using DecisonTreeClassifier for test data is predicted and Metrics are calculated.

20. Confusion Metrics is plotted to check how many predicted values are correctly predicted.



From the above confusion metrics we can see that most of the words are positive sentiments and which are truely predicted.

21.After Computing the classification report the F1 score of predicting Positive sentiment is the highest with 88% F1.



```
              precision    recall  f1-score   support

    Negative       0.56      0.02      0.05      7207
     Neutral       0.00      0.00      0.00      3509
    Positive       0.79      1.00      0.88     39284

    accuracy                           0.79     50000
   macro avg       0.45      0.34      0.31     50000
weighted avg       0.70      0.79      0.70     50000
```

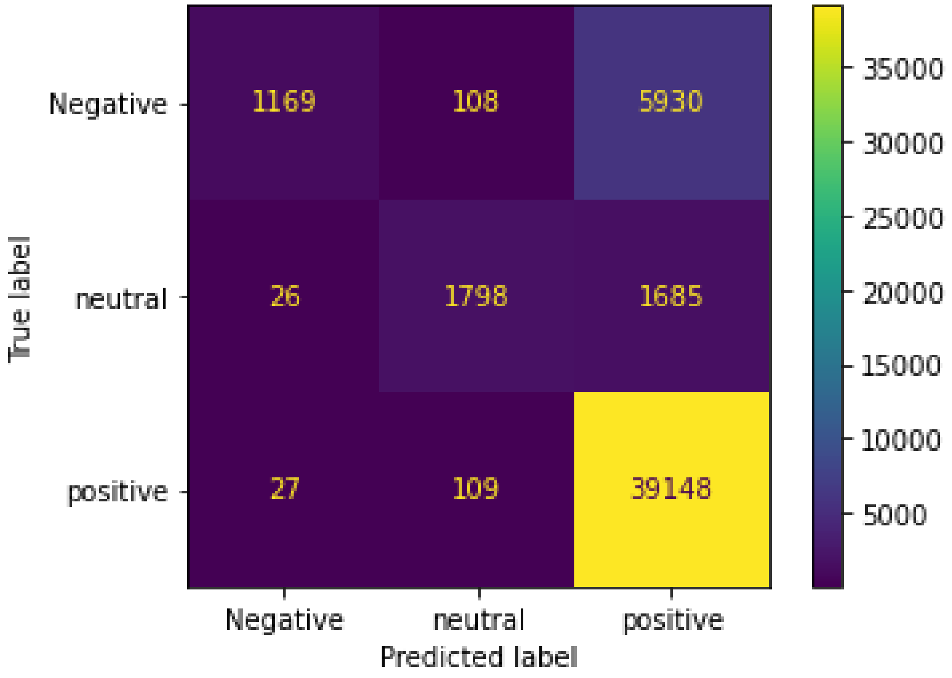
From the classification report we can see that 0.5% of F1-score of negative sentiments 0% of neutral and 88% of positive sentiments.

Which says the individual F1-score of every sentiment and positive sentiments has the highest F1-score compares to others sentiments.

22.Again Doing Modelling Using RandomTreeClassifier for test data is predicted and Metrics are calculated.

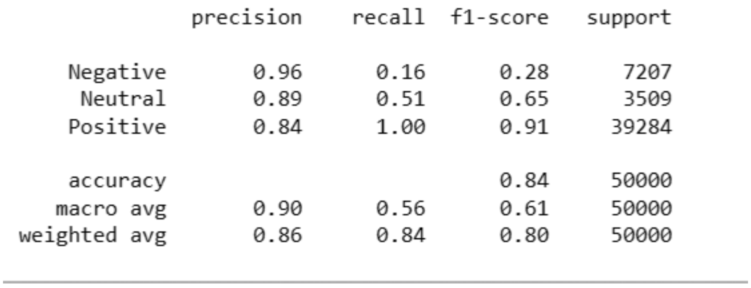
23. Confusion Metrics is plotted to check how many predicted values are

correctly predicted.



From the above confusion metrics we can see that most of the words are positive sentiments and which are truely predicted.

24. After Computing the classification report the F1 score of predicting Positive sentiment is the highest with 91% F1.



```
              precision    recall  f1-score   support

    Negative       0.96      0.16      0.28      7207
     Neutral       0.89      0.51      0.65      3509
    Positive       0.84      1.00      0.91     39284

    accuracy                           0.84     50000
   macro avg       0.90      0.56      0.61     50000
weighted avg       0.86      0.84      0.80     50000
```

From the classification report we can see that 28% of F1-score of negative sentiments 65% of neutral and 91% of positive sentiments.

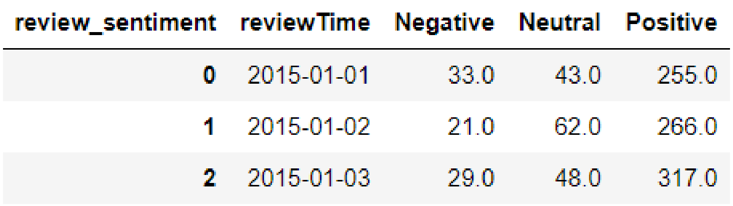
Which says the individual F1-score of every sentiment and positive sentiments has the highest F1-score compares to others sentiments.

25. So by comparing all model F1-score of of predicting Positive sentiment is highest of 96% F1 of model Logistic Regression with OVR Classifer has the highest F1-score of positive sentiment compare to others models.

26. So we will consider model of logistic Regression for sentiments analysis.

MODEL-2 - Time series analysis for sentiment forecast

1. Importing important libraries.
2. Importing data – Data with sentiment generated through vadersentiment library.
3. Only the ‘true’ verified reviews are considered for analysis and reviews posted after 01.01.2015.
4. Count of different sentiments is taken using pivot table function of ‘pandas’ library.

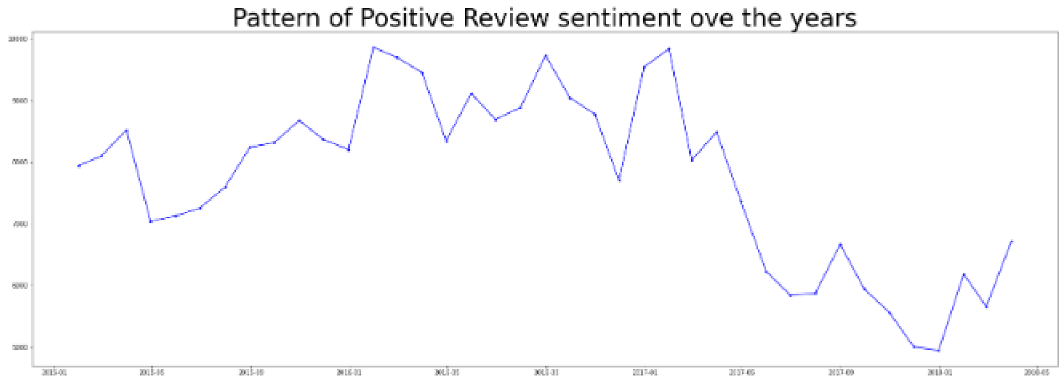


review_sentiment	reviewTime	Negative	Neutral	Positive
0	2015-01-01	33.0	43.0	255.0
1	2015-01-02	21.0	62.0	266.0
2	2015-01-03	29.0	48.0	317.0

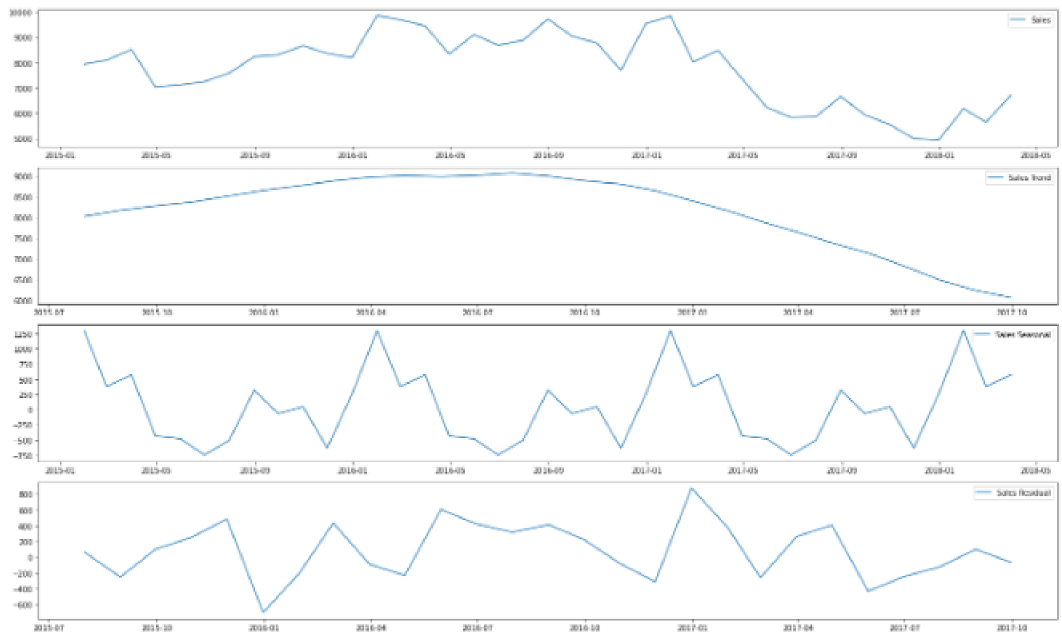
5. The data type of date column in the table set to date time format using to_datetime function of pandas
6. The dates are grouped by monthly format taking sum of each sentiment for each month. The data type of date is again set as date time format and the data is sorted in ascending order using 'sort_values' function of pandas. Finally, the index is set as date.
7. Three data frames are created for positive, neutral and negative sentiment separately where count of each sentiment is present on monthly basis namely pos_df, neu_df and neg_df sentiment.

For Positive sentiment forecast

8. The pos_df is resampled on 'Month'.
9. In order to make the data stationary last 7 rows are sliced for the data.
10. The pos_df is plotted to check the trend on the positive sentiment from 2015 till 2018

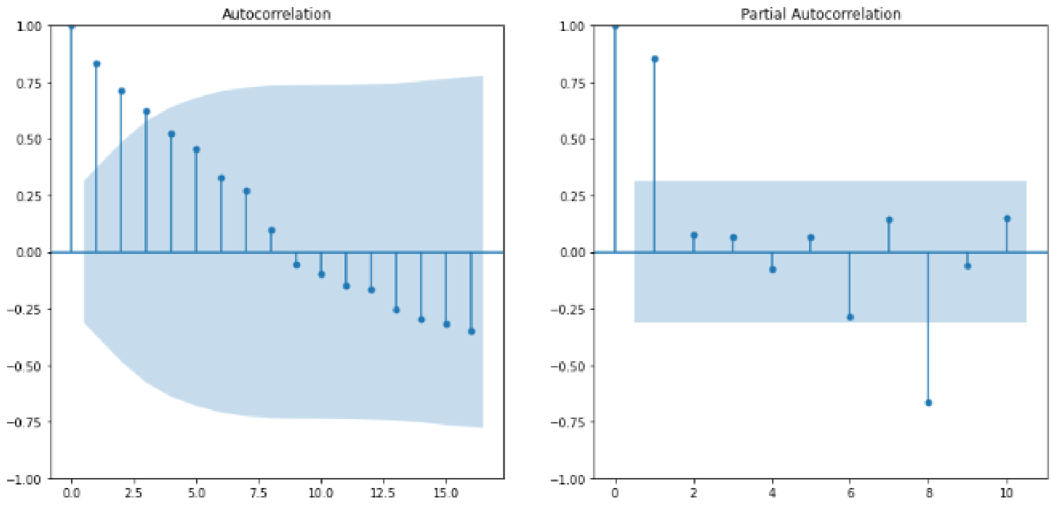


11. Seasonal decomposition of the data is performed to check the trend and seasonality in the data.



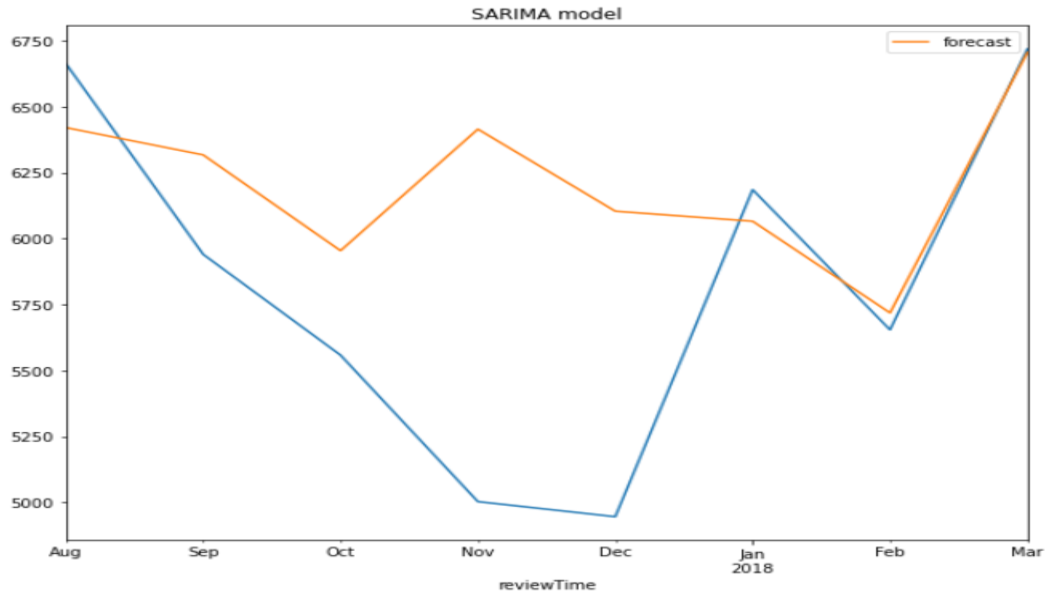
Since there is no trend but there is seasonality in the data, SARIMA model shall be made.

12. The stationarity of the data is checked, using Augmented dicky fuller test using a user defined function "checkStationarity" and setting the condition for p value to be less than 0.05 for stationary data.
13. Using the user defined function (checkStationarity) the stationary is checked. It was found out that the data is not stationary and differentiating the data once will make the data stationary. Hence value of d = 1 for modelling.
14. Pos_df is split into train1 and test1 data on 80:20 ratio.
15. Using ACF and PACF plot the value of q and p are found out respectively.

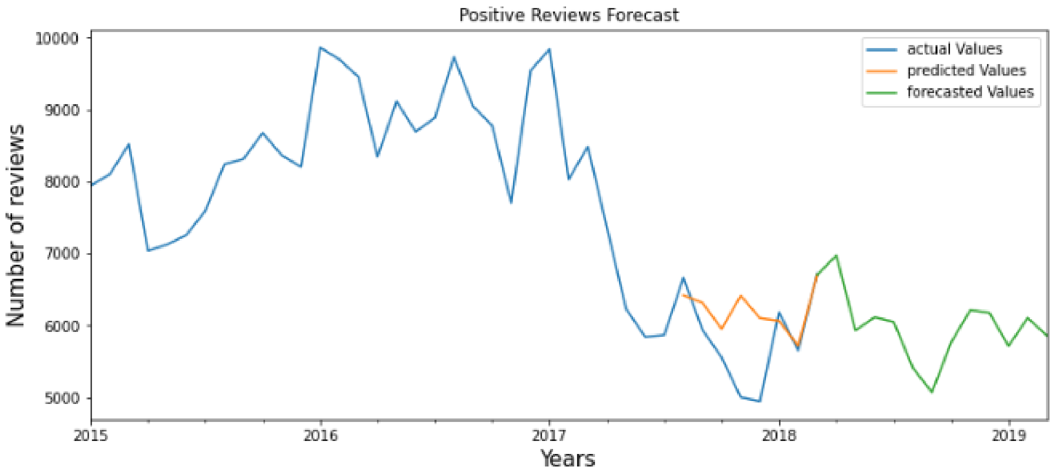


p = 2 and q = 3

16. For d = 1, p = 2 and q=3, different combinations of model are found out. If the p-value of any model is greater than 0.05 the model is declared as “Not a good model”.
17. For each of the above combinations of values, several metrics are calculated such as root mean squared error (RMSE), mean squared error(MSE), AIC and BIC. Out of these models, the model with lowest RMSE score is used for modelling.
18. Here for (p, d, q) = (1, 1, 0) there is least RMSE score of 632.46. Therefore, a SARIMA model, m1, with order as (1, 1, 0) and seasonal order as (1,1,1,15) is created and trained on train1 data.
19. Using m1 model, values on test1 data is predicted. Using the actual values and predicted values RMSE and MSE scores are calculated. Here the RMSE score is 680.75.



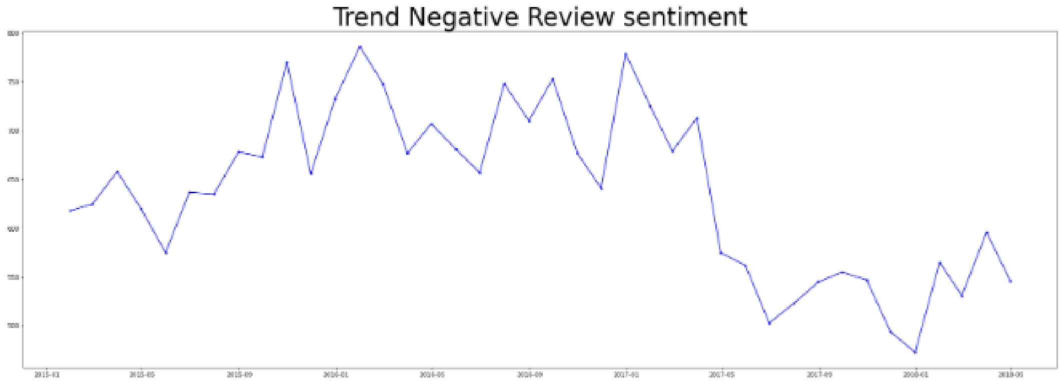
20. Using the m1 model next 12 months value of the positive sentiments is forecasted. The values are then plotted on the graph with actual and predicted values.



Conclusion: The positive sentiment is set to decline in future. There is possibility of small rise during last couple of months in 2018.

For Negative sentiment forecast

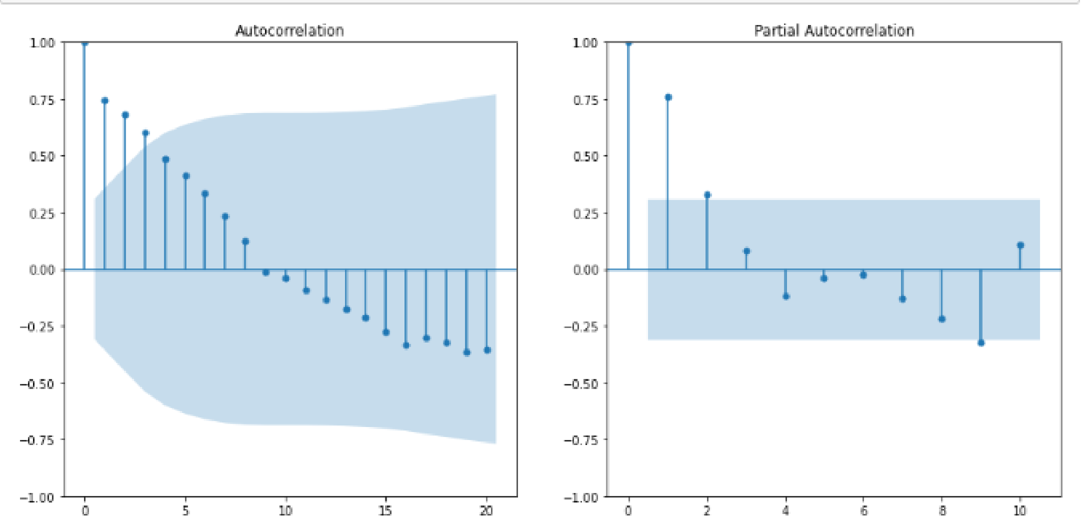
1. The neg_df is resampled on ‘Month’.
2. In order to make the data stationary last 6 rows are sliced for the data.
3. The neg_df is plotted to check the trend on the Negative sentiment from 2015 till 2018



4. Seasonal decomposition of the data is performed to check the trend and seasonality in the data.

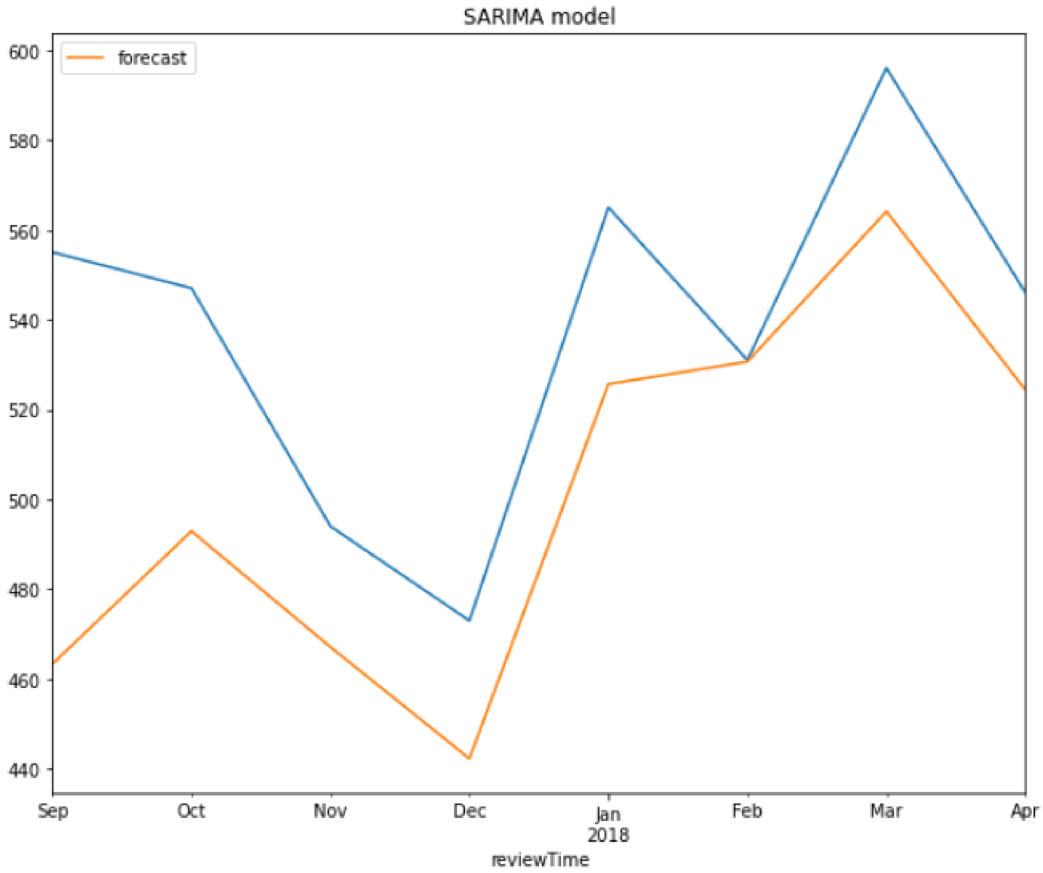
   Since there is no trend but there is seasonality in the data, SARIMA model shall be made.
5. The stationarity of the data is checked, using Augmented dicky fuller test using a user defined function “checkStationarity” and setting the condition for p value to be less than 0.05 for stationary data.

6. Using the user defined function (checkStationarity) the stationary is checked. It was found out that the data is not stationary and differentiating the data once will make the data stationary. Hence value of d = 1 for modelling.
7. The neg_df is split into train2 and test2 data on 80:20 ratio.
8. Using ACF and PACF plot the value of q and p are found out respectively.

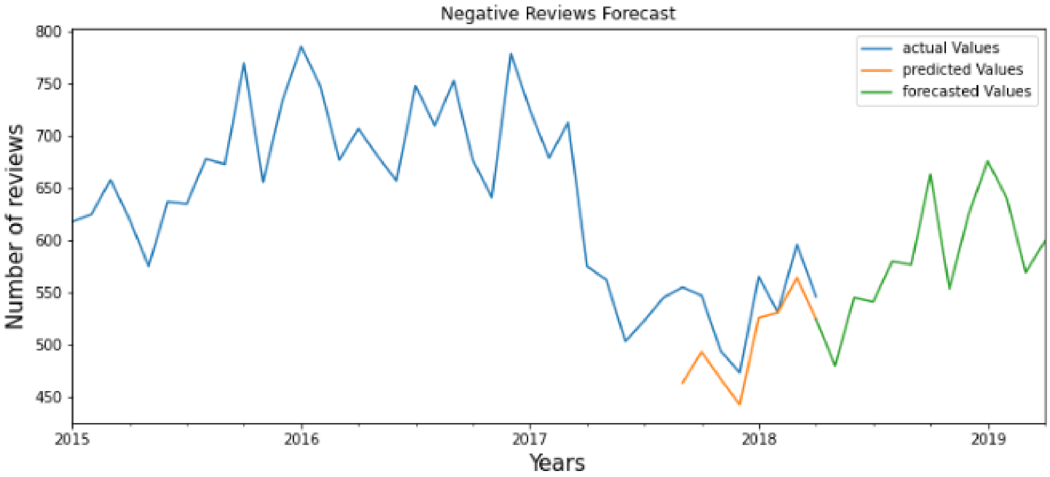


p = 3 and q = 3

9. For d = 1, p = 3 and q=3, different combinations of model are found out. If the p-value of any model is greater than 0.05 the model is declared as “Not a good model”.
10. For each of the above combinations of values, several metrics are calculated such as root mean squared error (RMSE), mean squared error(MSE), AIC and BIC. Out of these models, the model with lowest RMSE score is used for modelling.
11. Here for (p, d, q) = (0, 1, 1) there is least RMSE score of 36.63. Therefore, a SARIMA model, m2, with order as (0, 1, 1) and seasonal order as (1,1,1,18) is created and trained on train2 data.
12. Using m2 model, values on test2 data is predicted. Using the actual values and predicted values RMSE and MSE scores are calculated. Here the RMSE score is 44.74.



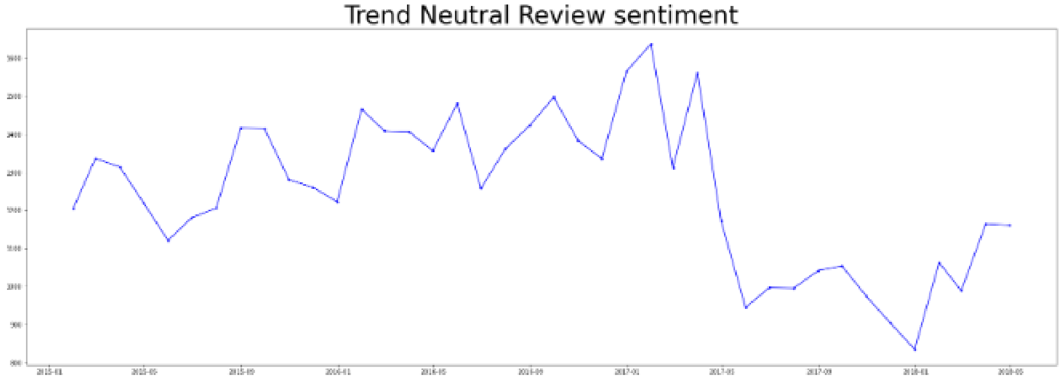
13. Using the m2 model next 12 months value of the Negative sentiments is forecasted. The values are then plotted on the graph with actual and predicted values.



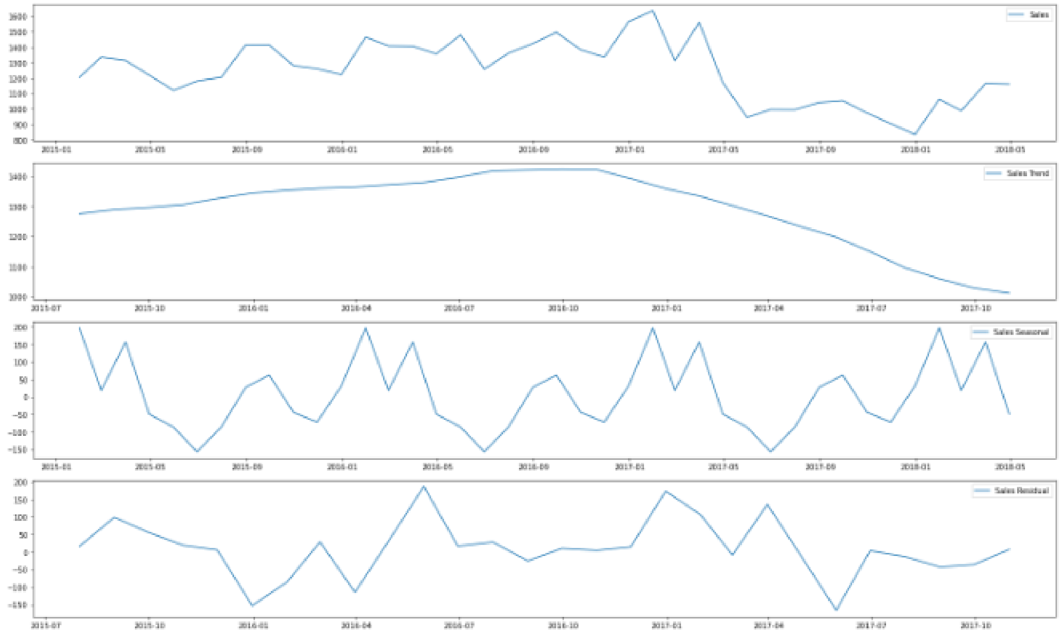
Conclusion: The negative sentiment is set to rise in future. There is possibility of small fall during last couple of months in 2018 and then rise again.

For Neutral sentiment forecast

1. The neu_df is resampled on ‘Month’.
2. In order to make the data stationary last 6 rows are sliced for the data.
3. The neu_df is plotted to check the trend on the Neutral sentiment from 2015 till 2018



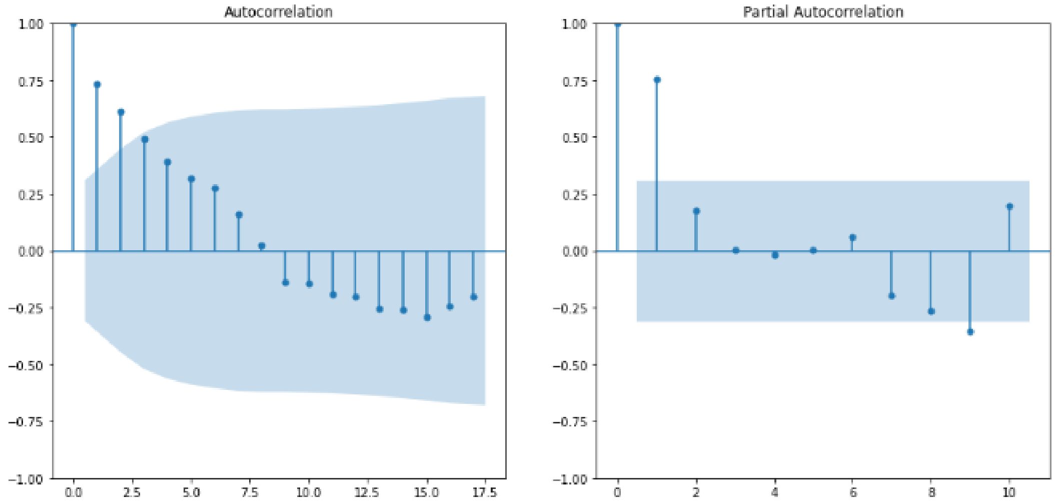
4. Seasonal decomposition of the data is performed to check the trend and seasonality in the data.



Since there is no trend but there is seasonality in the data, SARIMA model shall be made.

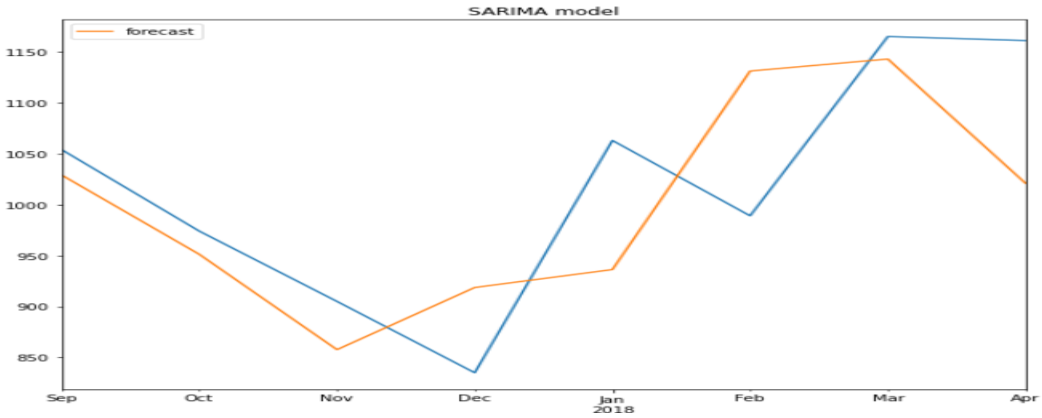
5. The stationarity of the data is checked, using Augmented dicky fuller test using a user defined function “checkStationarity” and setting the condition for p value to be less than 0.05 for stationary data.

6. Using the user defined function (checkStationarity) the stationary is checked. It was found out that the data is not stationary and differentiating the data once will make the data stationary. Hence value of d = 1 for modelling.
7. The neu_df is split into train3 and test3 data on 80:20 ratio.
8. Using ACF and PACF plot the value of q and p are found out respectively.

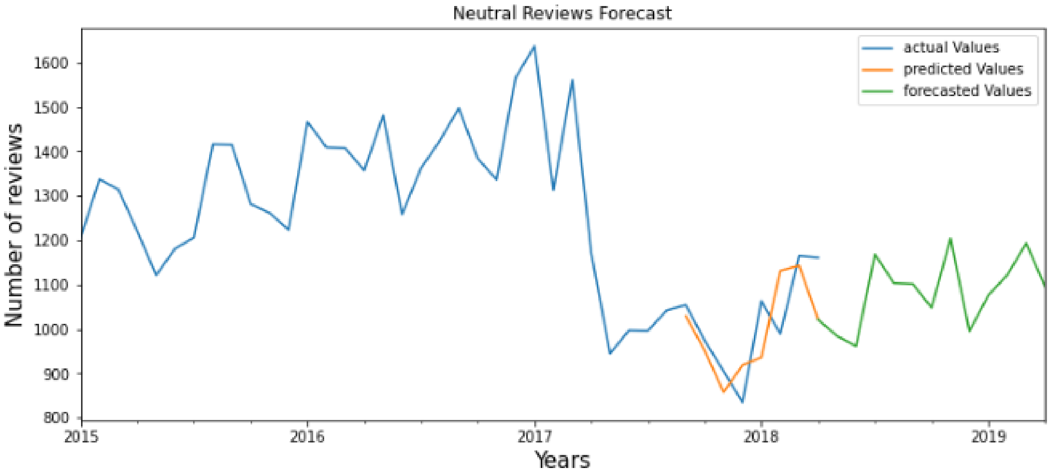


p = 2 and q = 2

9. For d = 1, p = 2 and q=2, different combinations of model are found out. If the p-value of any model is greater than 0.05 the model is declared as “Not a good model”.
10. For each of the above combinations of values, several metrics are calculated such as root mean squared error (RMSE), mean squared error(MSE), AIC and BIC. Out of these models, the model with lowest RMSE score is used for modelling.
11. Here for (p, d, q) = (2, 1, 0) there is least RMSE score of 109.04. Therefore, a SARIMA model, m3, with order as (2, 1, 0) and seasonal order as (1,1,1,15) is created and trained on train3 data.
12. Using m3 model, values on test3 data is predicted. Using the actual values and predicted values RMSE and MSE scores are calculated. Here the RMSE score is 91.34.



13. Using the m3 model next 12 months value of the Neutral sentiments is forecasted. The values are then plotted on the graph with actual and predicted values.



Conclusion: The Neutral sentiment is set to rise in future. There is possibility of small fall during last couple of months in 2018 and then rise again.

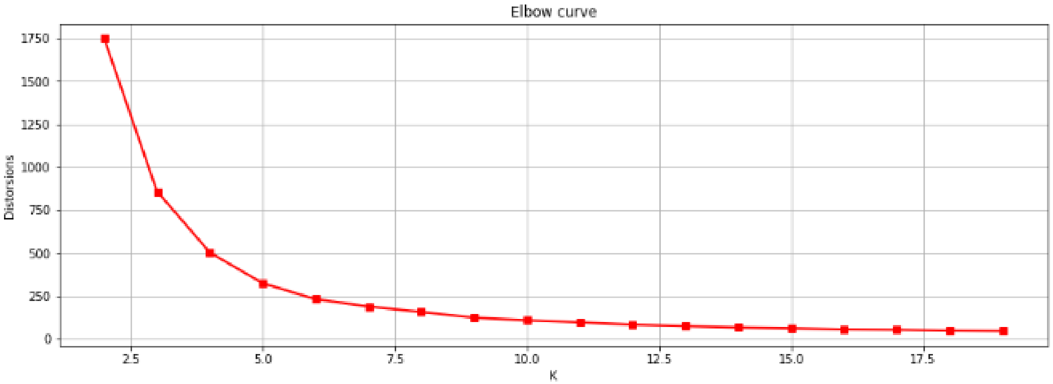
MODEL 3- Model – Customer clustering using K mean clustering model

1. Importing the necessary libraries.
2. Importing ‘customer_retention_office’ , ‘customer_retention_digital’, ‘final_reviews’ datasets (csv) as office, digital and review .
3. Using review data frame extracting price for all the products falling in the main_cat of ‘Digital Music’ for each unique reviewerID for all the verified reviews. Performed similar steps for the products

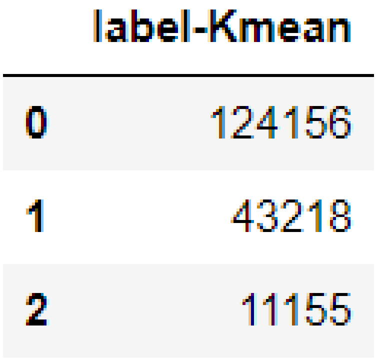
falling in the main_cat “Office Products”. The data frames are named as dig and off respectively.

4. In the office and digital data frames replacing the values in sentiment column with encoded data ie. 0 for Negative, 1 for Neutral and 2 for Positive.
5. Using the concat function of pandas, concatenated office and digital data frame and named the final data frame as data1.
6. Using the concat function of pandas, concatenated off and dig data frame and named the final data frame as data1.
7. Using groupby function on data2 where grouping is done on reviewerID and sum of price is taken for each reviewerID for total price spent by each customer. The data frame is named d1.
8. Using groupby function on data1 where grouping is done on reviewerID and mean of the sentiment is taken for each reviewerID as average sentiment. The data frame is named d2.
9. Using groupby function on data1 where grouping is done on reviewerID and count of reviewerID is taken for each reviewerID for number of orders placed by each reviewerID. The data frame is named d3.
10. Data Frame d1, d2 and d3 are merged on reviewerID. This data frame is named final_df.
11. Using data1 and sorting the values of reviewTime in ascending order and assigning a new variable xyz to the sorted data frame. Removing duplicated rows for multiple reviewerID entries and keeping only the first row. After removing/dropping the duplicates now the data frame contains each reviewerID with the date on which it made the first purchase. Creating a data frame start and only adding reviewerID and the reviewTime. Performing the same steps to get the last/recent date of purchase for each reviewerID and finally creating a data frame as end.
12. Merging start and end data frame on reviewerID. Only keeping the year of each date in the table by slicing it and renaming the date column from start table as year_first_order and end table as year_last_order. Subtracting the year_last_order from year_first_order to get the number of years a particular reviewerID has been the customer.
13. Created a new data frame as df2 which is the copy of final_df.
14. Considering column total_price _spent, avg_sentiment and total_order for modelling.

15. Scaling the data using MixMaxScaler and naming the scaled data frame as scaled_df.
16. Using Elbow curve for finding the value of K i.e. the number of clusters.

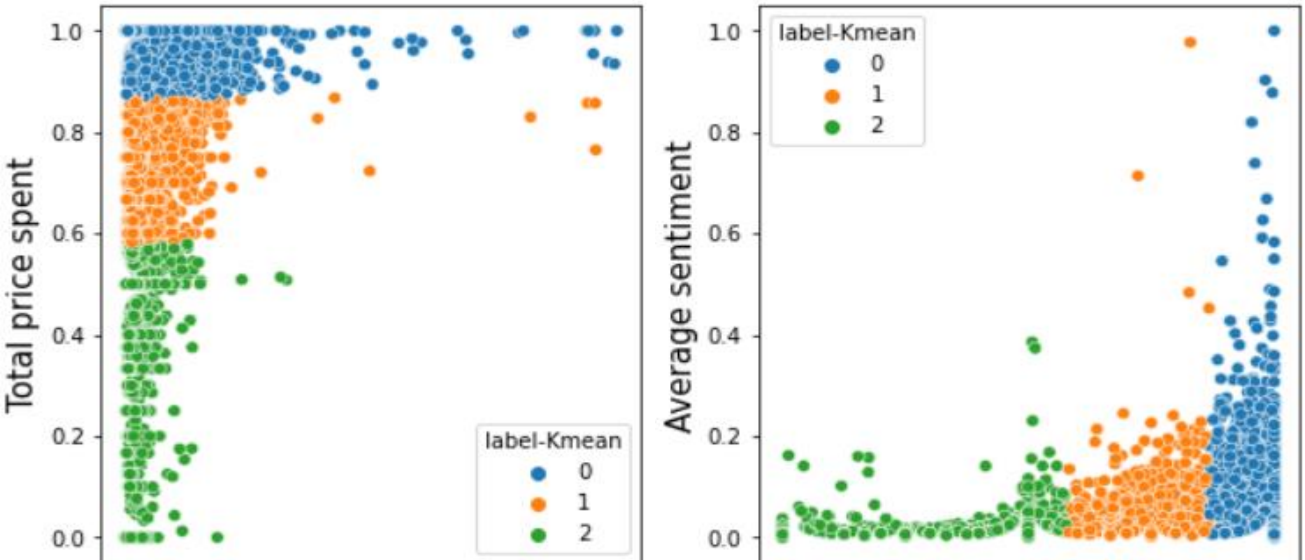


17. Also using a loop to calculate the silhouette score of each K value from the above elbow graph. Taking the values of K as 3
18. Creating a K means model using K = 3 ie. the number of clusters and training it on the df_r which is the copy of scaled_df
19. Using fit_predict function finding the designated label for each row/reviewerID.
20. The customer are divided among the following labels



	label-Kmean
0	124156
1	43218
2	11155

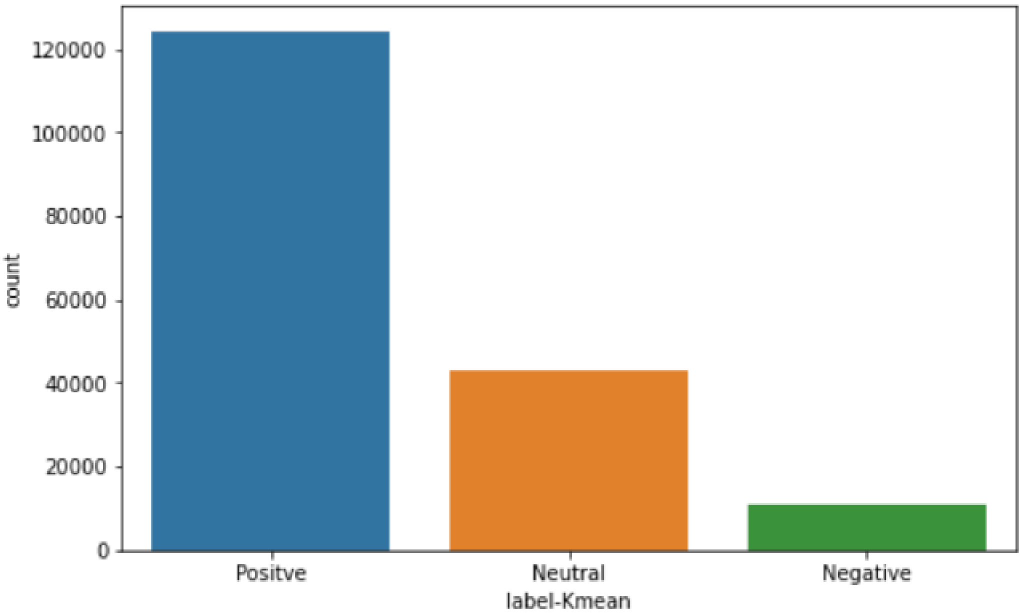
21. Calculating the silhouette score for the same which is 94.04. This silhouette score is considered a really good score for a model hence this is a good model for clustering
22. Using seaborn library for visualising the clusters in a scatterplot



Conclusion:

Label	Average Sentiment	Total price spent	Number of orde
0	Positive (1.7-2)	Low to high	Low to high
1	Neutral (1.2-1.7)	Low to Mid	Low to Mid
2	Negative (0-1.2)	Low	Low

23. Finding total count of each label : label 0 = positive label 1 = neutral and label 2 = negative

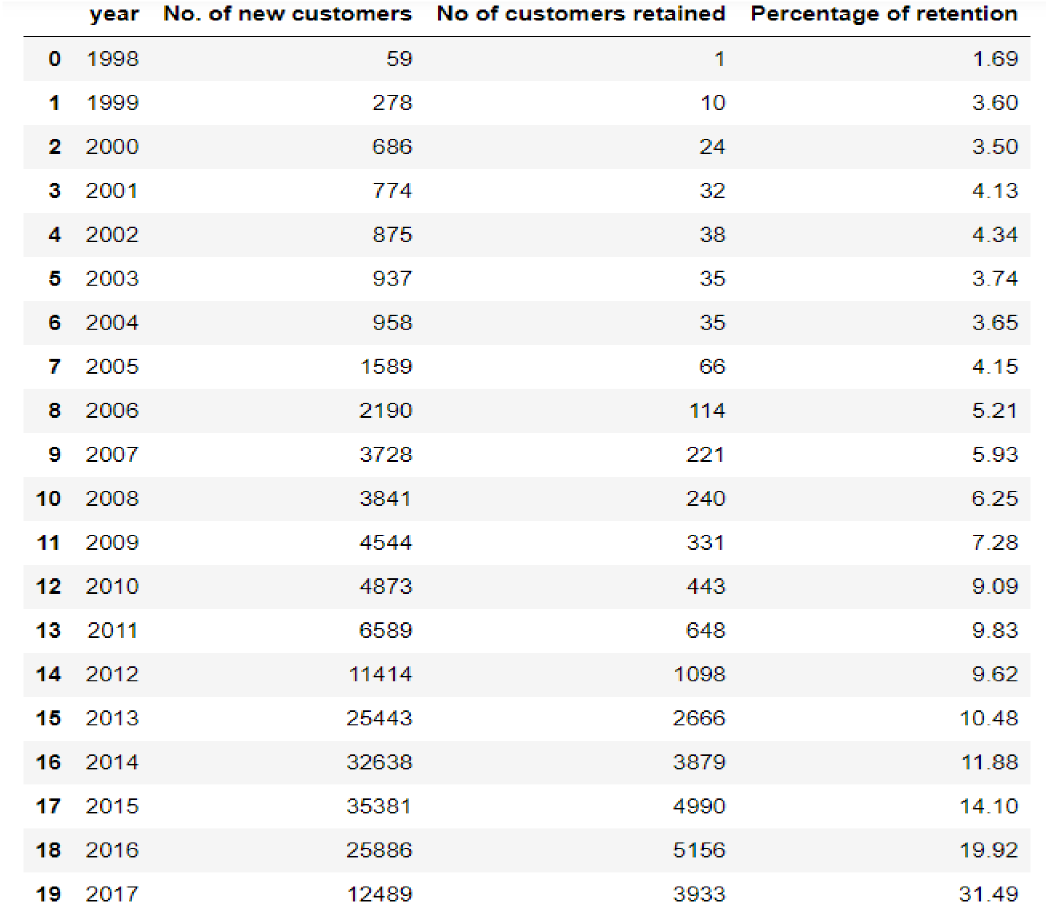


24. A final data frame is created with each label. The data frame includes reviewerID, total_price _spent, avg_sentiment and total_order, number of years they have been customers in Amazon, year of first

purchase and year of recent/last purchase. Hence extracting the three data frames with label 0 as “happy customers” label 1 as “average satisfied” and label 2 as “Not happy customer”.

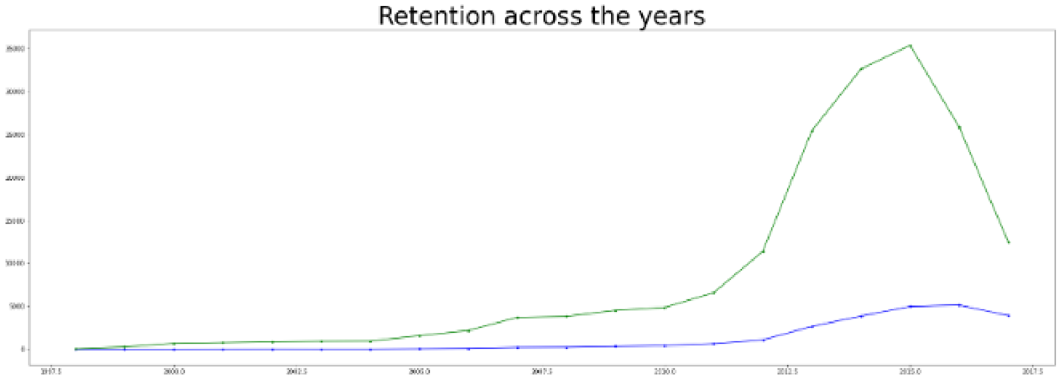
Model 4 – Timer series analysis for forecasting Customer retention for the next 3 years.

1. Importing necessary libraries
2. Importing rentention_year_customer data exported in ‘clustering of customer’ model.
3. This data frame contains unique reviewerID of all the customers along with the year of their first purchase, year of last purchase and total number of years they have been a customer on Amazon.
4. Using a for loop a data frame is created where number of new customer in each year is counted by analysis the year of first purchase of each customer. Further, it is checked whether these customers were retained will 2018 i.e. whether these customers shopped again 2018. Final data frame is achieved with relevant information.



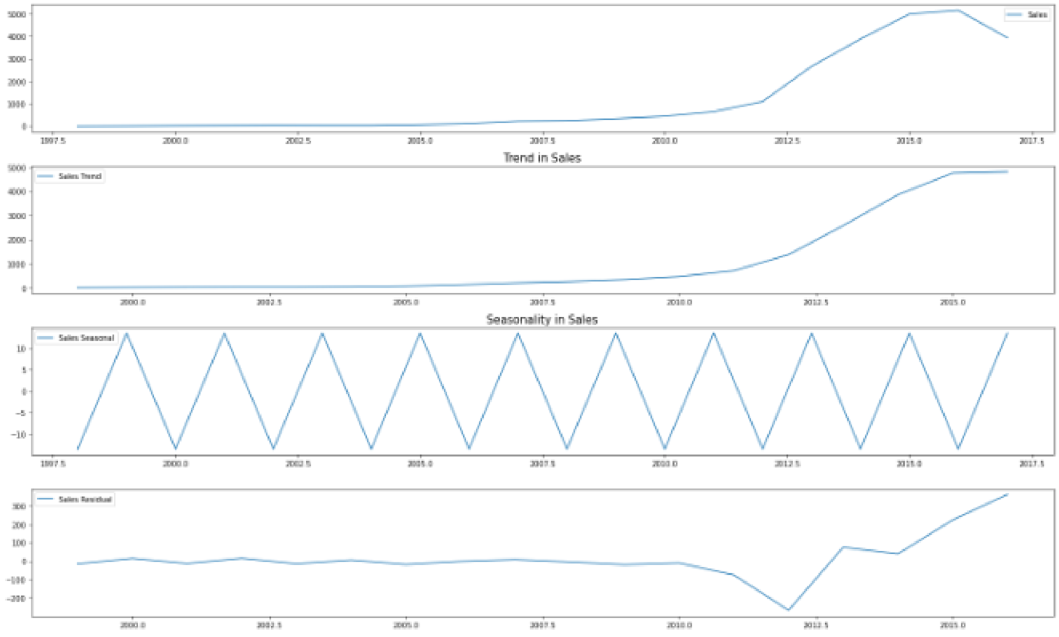
	year	No. of new customers	No of customers retained	Percentage of retention
0	1998	59	1	1.69
1	1999	278	10	3.60
2	2000	686	24	3.50
3	2001	774	32	4.13
4	2002	875	38	4.34
5	2003	937	35	3.74
6	2004	958	35	3.65
7	2005	1589	66	4.15
8	2006	2190	114	5.21
9	2007	3728	221	5.93
10	2008	3841	240	6.25
11	2009	4544	331	7.28
12	2010	4873	443	9.09
13	2011	6589	648	9.83
14	2012	11414	1098	9.62
15	2013	25443	2666	10.48
16	2014	32638	3879	11.88
17	2015	35381	4990	14.10
18	2016	25886	5156	19.92
19	2017	12489	3933	31.49

5. The data is plotted to check the spread of the data across the years i.e. starting from 1997 to 2018.



6. A data frame ‘reten’ is created containing the year (from 1997 to 2017) and the number of customer retained each year.

7. The data is checked for trend and seasonality using seasonal decomposition.

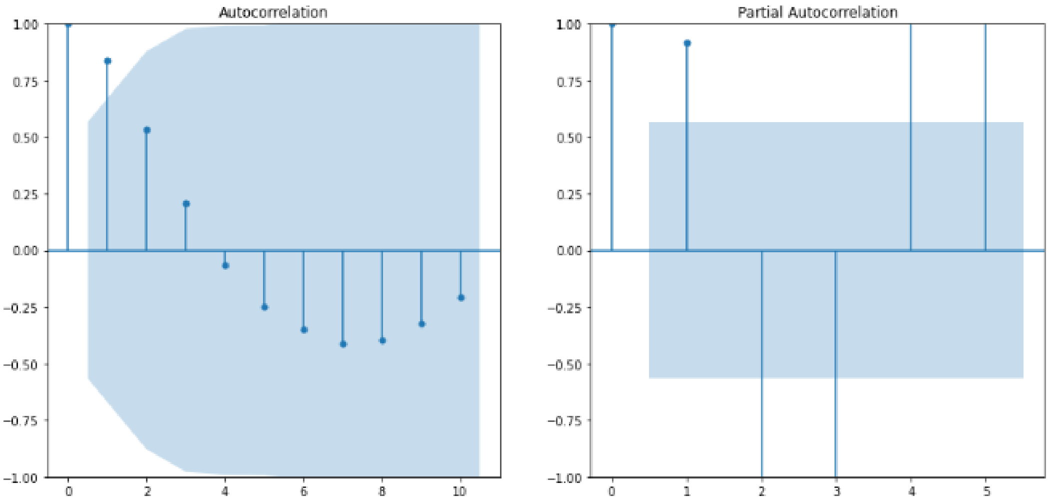


The data shows trend as well as seasonality.

8. The stationarity of the data is checked, using Augmented dicky fuller test using a user defined function “checkStationarity” and setting the condition for p value to be less than 0.05 for stationary data.
9. Using the user defined function (checkStationarity) the stationary is checked. It was found out that the data is not stationary and differentiating the data once will make the data stationary. After 8

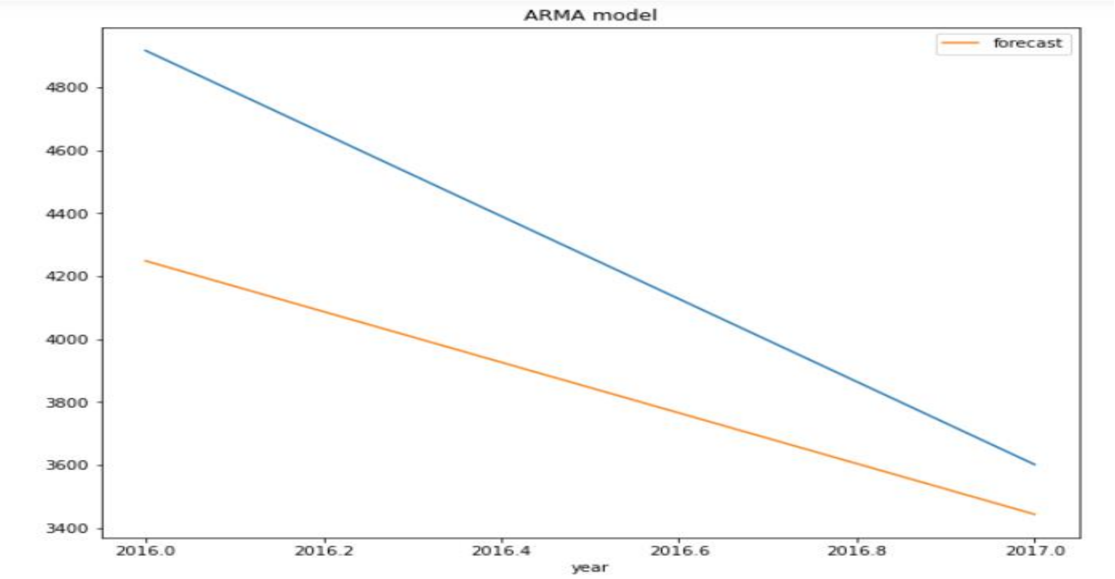
shifts the data became stationary, hence proceeding with the stationary data. (now d = 0)

10. The diff_reten, stationary data, is split into train1 and test1 data on 90:10 ratio.
11. Using ACF and PACF plot the value of q and p are found out respectively.

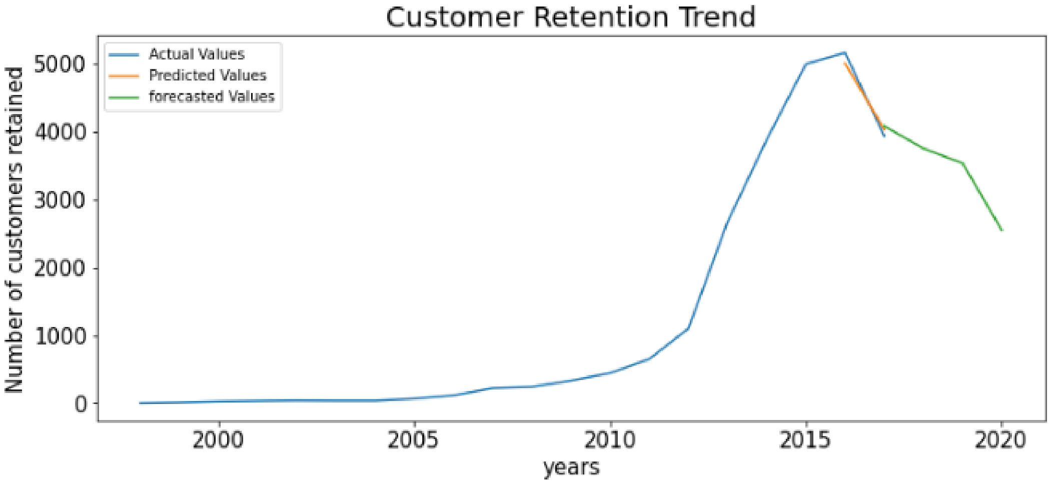


p = 5 and q = 1

12. For d = 0, p = 5 and q=1, different combinations of model are found out. If the p-value of any model is greater than 0.05 the model is declared as “Not a good model”.
13. For each of the above combinations of values, several metrics are calculated such as root mean squared error (RMSE), mean squared error(MSE), AIC and BIC. Out of these models, the model with lowest RMSE score is used for modelling.
14. Here for (p, d, q) = (5, 0, 1) there is least RMSE score of 307.79. Therefore, a ARMA model, m1, with order as (5, 0, 1) is created and trained on train1 data.
15. Using m1 model, values on test1 data is predicted. Using the actual values and predicted values RMSE and MSE scores are calculated. Here the RMSE score is 485.21



16. Using the m1 model next 3 years value of the Neutral sentiments is forecasted. The values are then plotted on the graph with actual and predicted values.

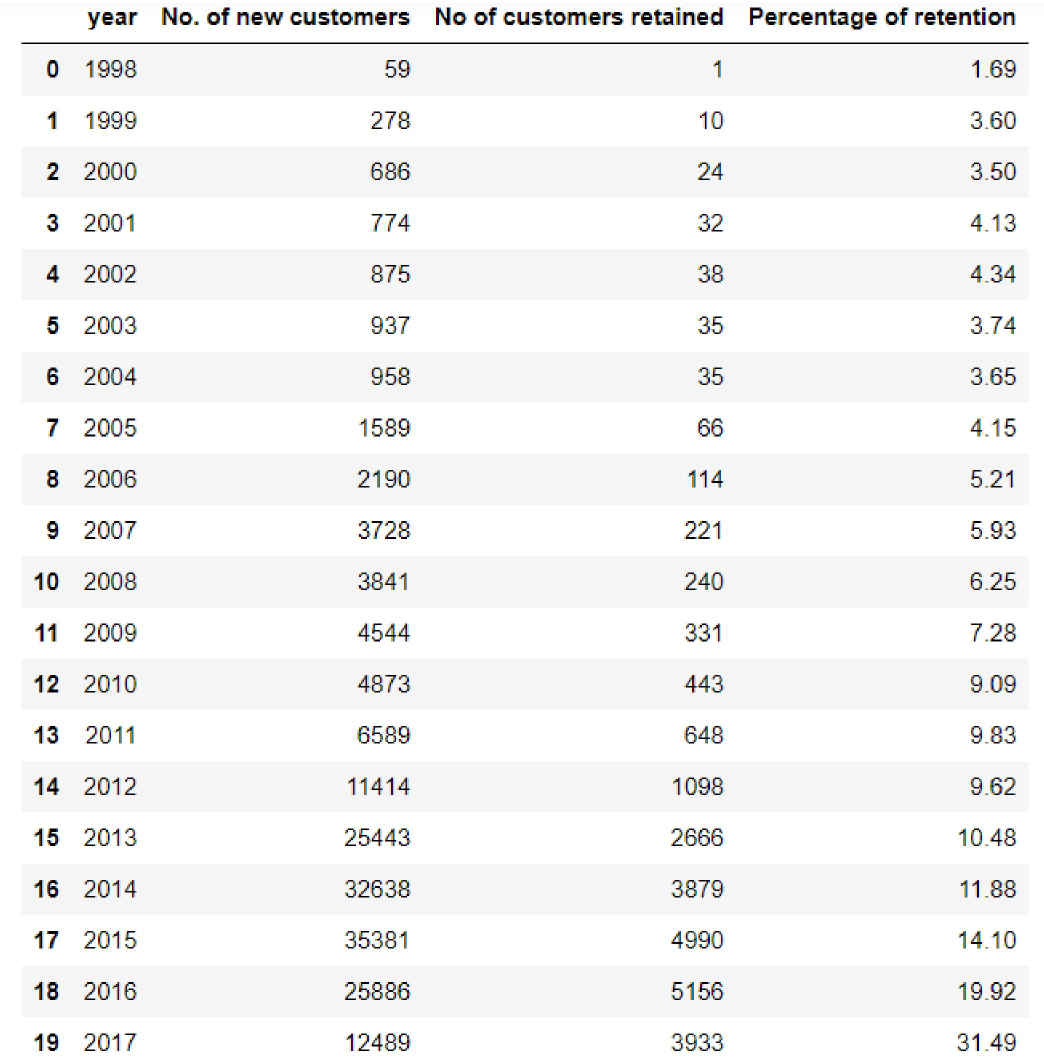


Conclusion: The number of customers retained for next year will remain fall each year.

Model 5 – Timer series analysis for forecasting new customers for the next 3 years.

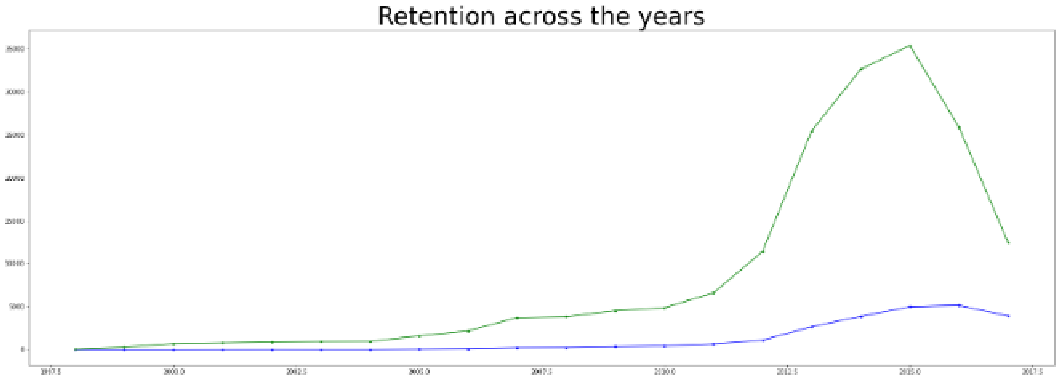
1. Importing necessary libraries
2. Importing rentention_year_customer data exported in ‘clustering of customer’ model.
3. This data frame contains unique reviewerID of all the customers along with the year of their first purchase, year of last purchase and total number of years they have been a customer on Amazon.

4. Using a for loop a data frame is created where number of new customer in each year is counted by analysis the year of first purchase of each customer. Further, it is checked whether these customers were retained will 2018 i.e. whether these customers shopped again 2018. Final data frame is achieved with relevant information.



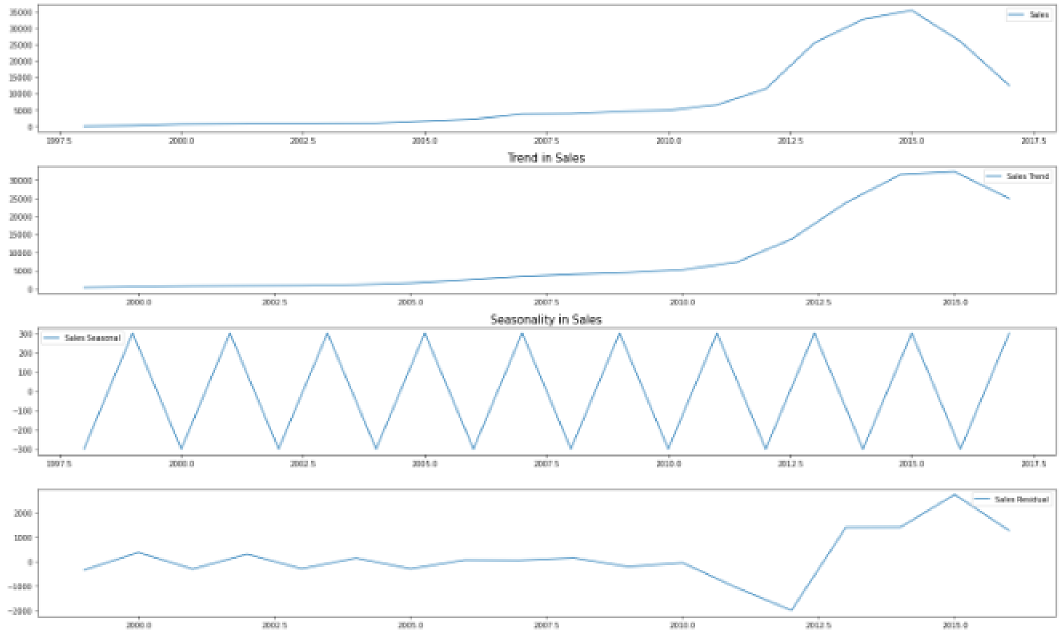
	year	No. of new customers	No of customers retained	Percentage of retention
0	1998	59	1	1.69
1	1999	278	10	3.60
2	2000	686	24	3.50
3	2001	774	32	4.13
4	2002	875	38	4.34
5	2003	937	35	3.74
6	2004	958	35	3.65
7	2005	1589	66	4.15
8	2006	2190	114	5.21
9	2007	3728	221	5.93
10	2008	3841	240	6.25
11	2009	4544	331	7.28
12	2010	4873	443	9.09
13	2011	6589	648	9.83
14	2012	11414	1098	9.62
15	2013	25443	2666	10.48
16	2014	32638	3879	11.88
17	2015	35381	4990	14.10
18	2016	25886	5156	19.92
19	2017	12489	3933	31.49

5. The data is plotted to check the spread of the data across the years i.e. starting from 1997 to 2018.



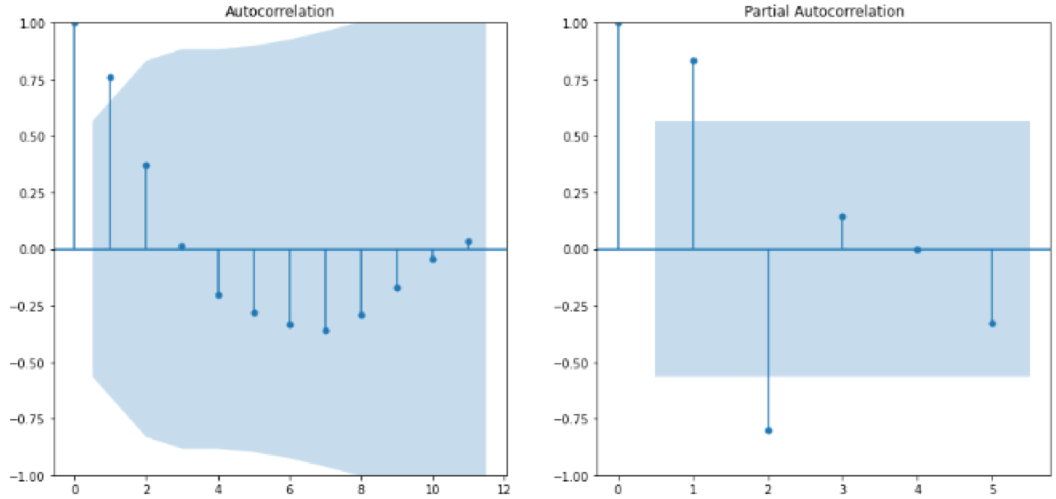
6. A data frame ‘reten’ is created containing the year (from 1997 to 2017) and the number of new customers each year.

7. The data is checked for trend and seasonality using seasonal decomposition.



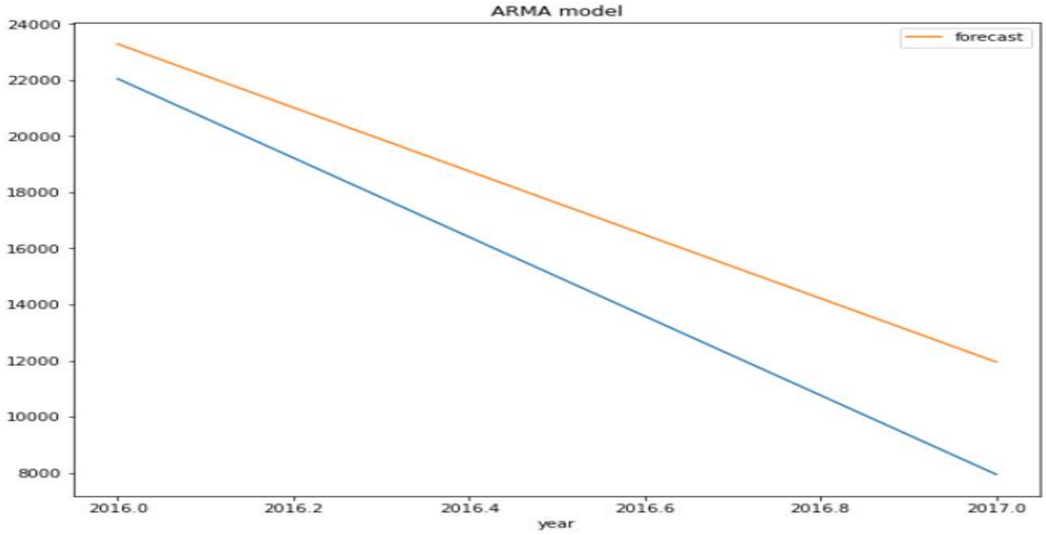
The data shows no trend however, there is seasonality in the data.

8. The stationarity of the data is checked, using Augmented dicky fuller test using a user defined function “checkStationarity” and setting the condition for p value to be less than 0.05 for stationary data.
9. Using the user defined function (checkStationarity) the stationary is checked. It was found out that the data is not stationary and differentiating the data once will make the data stationary. After 8 shifts the data became stationary, hence proceeding with the stationary data. (now d = 0)
10. The diff_reten, stationary data, is split into train1 and test1 data on 90:10 ratio.
11. Using ACF and PACF plot the value of q and p are found out respectively.

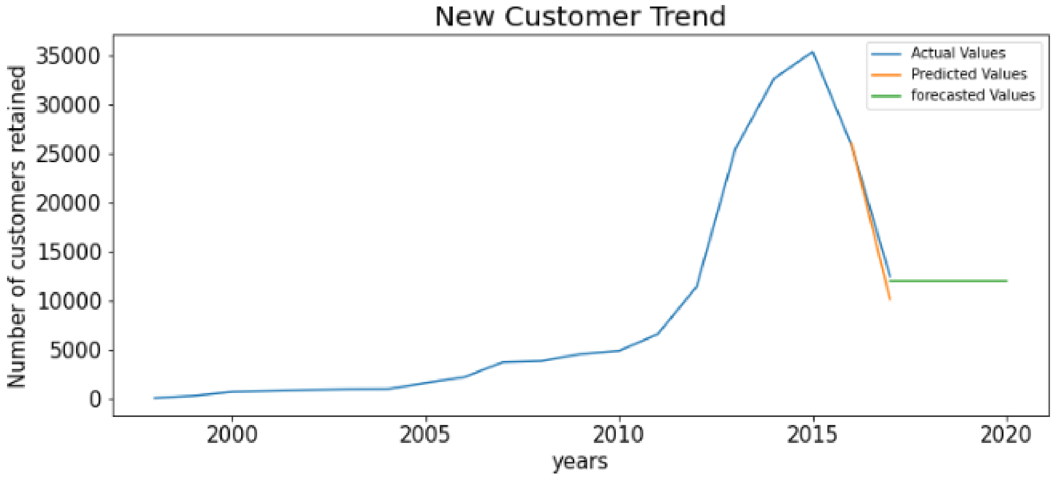


p = 2 and q = 1

12. For d = 0, p = 2 and q=1, different combinations of model are found out. If the p-value of any model is greater than 0.05 the model is declared as “Not a good model”.
13. For each of the above combinations of values, several metrics are calculated such as root mean squared error (RMSE), mean squared error(MSE), AIC and BIC. Out of these models, the model with lowest RMSE score is used for modelling.
14. Here for (p, d, q) = (0, 0, 1) there is least RMSE score of 2338.69. Therefore, a ARMA model, m1, with order as (0, 0, 1) is created and trained on train1 data.
15. Using m1 model, values on test1 data is predicted. Using the actual values and predicted values RMSE and MSE scores are calculated. Here the RMSE score is 2966.25



16. Using the m1 model next 3 years value of the Neutral sentiments is forecasted. The values are then plotted on the graph with actual and predicted values.



Conclusion: The number of new customers for next year will remain constant at 11952 each year

Model 6: Digital Music & Office Product Products Demand Analysis Using K-Means clustering

Procedure:

1. Importing the required libraries and importing the final reviews dataset.

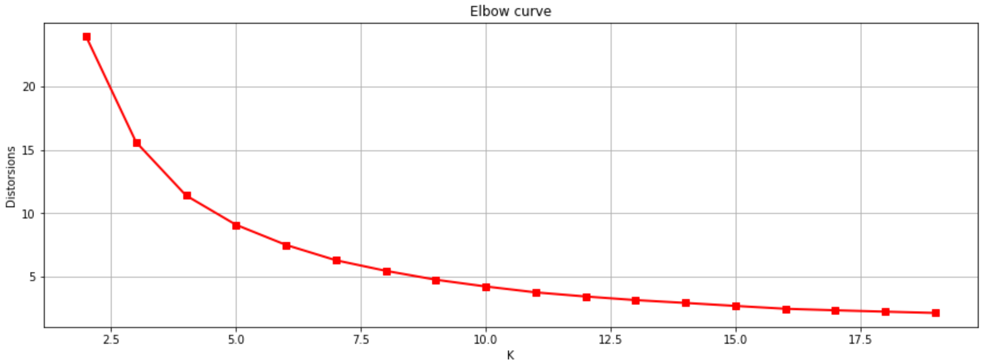
2. Digital Music & Office product category is extracted from the main category column.

3. Created a dataframe using groupby function of pandas, grouping it on ‘asin’ column to take count of particular asin.
4. Renaming the columns name from ‘asin’ to ‘Product ID’ & ‘count’ to ‘order_count’.
5. Created a dataframe using groupby function of pandas, grouping it on ‘asin’ and ‘price’ column to take total sales for particular ProductID.
6. Merging the dataframe of order count and total sales.

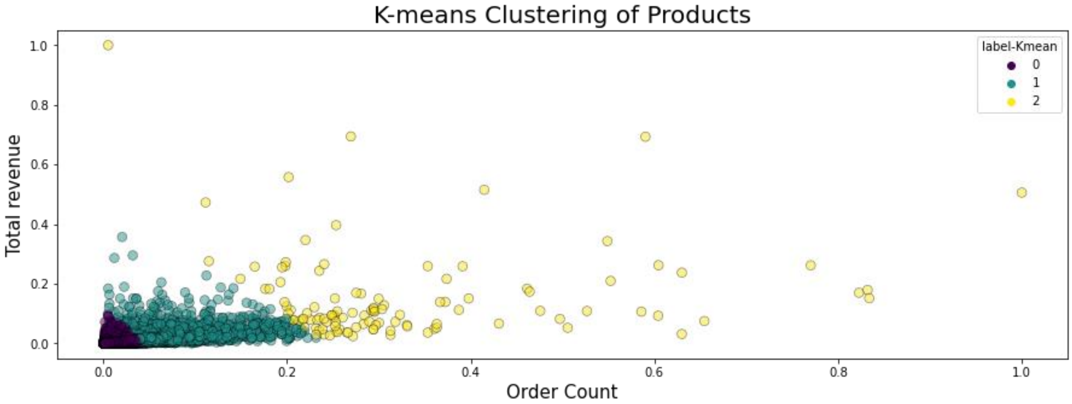
7. Sorting the merged dataframe by ordercount in ascending order.
8. Feature Engineering

Going further with data scaling using normalisation (MinMaxScaler) on column ‘order_count’ & ‘total_sales.

9. To find an optimal number of clusters using elbow method on the scaled dataframe.
   We are getting 5 number of clusters.



10. To check the goodness of cluster calculating silhouette scores where we’re getting a good score for 3 cluster.
11. Moving forward for modelling with K-Means clustering with 3 clusters.
12. Plotting a scatter plot for 3 clusters.



13. Label 2 denotes categorise with high price and high rating.
    Label 1 denotes categorise with low price and high rating.
    Label 0 denotes categorise with low price and low rating.

    By seeing the scatter plot, we can conclude that the Cluster 2 having ‘High’ number of order’s and generating high revenue.

14. Creating a dataframe name Df4 and merging productID, labels & class with order count & total sales.

15. Exploring our all 3 clusters we found that cluster number 2 is giving us high demanded products

16. Now, creating a new dataframe by taking some columns from the main data (Asin, tittle, main cat, primary category, sub cat) and merging It into the Df4. For extra information about the product.

17. Extracting the data where label = 1 and exported to a csv file to achieve a final list of highly demanding products for reference

Conclusion-

By applying K-Means clustering we can conclude that the Cluster labelled 1 is having the highest demanding products with respect to highest revenue.

Model 7: Digital Products Demand Analysis Using K-Means clustering

Procedure:

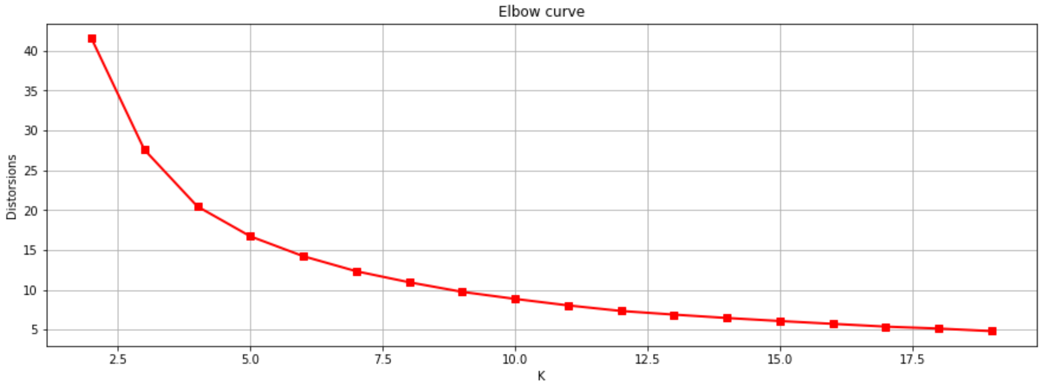
1. Importing the required libraries and importing the final reviews dataset.

2. Digital Music product category is extracted from the main category column.

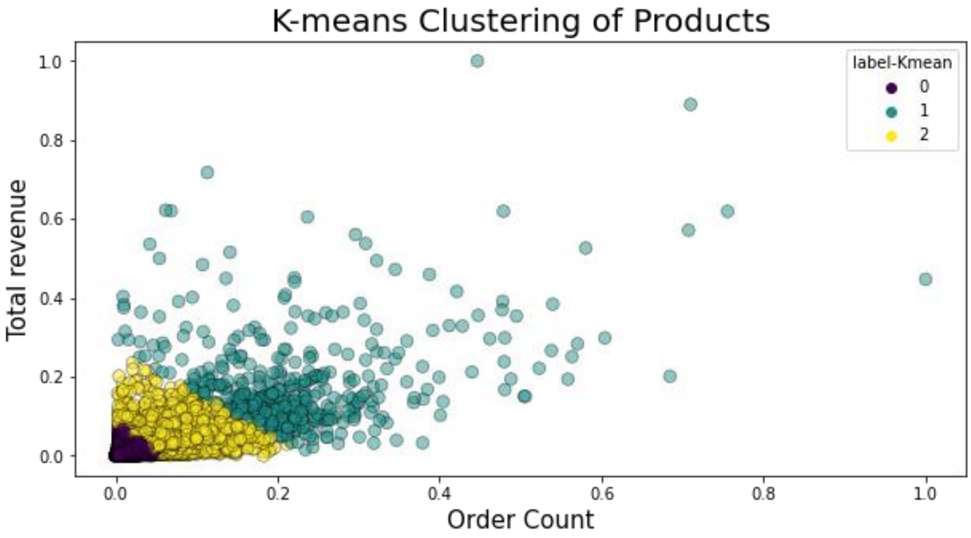
3. Created a dataframe using groupby function of pandas, grouping it on ‘asin’ column to take count of particular asin.
4. Renaming the columns name from ‘asin’ to ‘Product ID’ & ‘count’ to ‘order_count’.
5. Created a dataframe using groupby function of pandas, grouping it on ‘asin’ and ‘price’ column to take total sales for particular ProductID.
6. Merging the dataframe of order count and total sales.

7. Sorting the merged dataframe by ordercount in ascending order.
8. Feature Engineering
   Going further with data scaling using normalisation (MinMaxScaler) on column ‘order_count’ & ‘total_sales.

9. To find an optimal number of clusters using elbow method on the scaled dataframe.
   We are getting 5 number of clusters.



10. To check the goodness of cluster calculating silhouette scores where we’re getting a good score for 3 cluster.
11. Moving forward for modelling with K-Means clustering with 3 clusters.
12. Plotting a scatter plot for 3 clusters.



13. By seeing the scatter plot, we can conclude that the Cluster 1 having ‘High’ number of order’s and generating high revenue.

14. Creating a dataframe name Df4 and merging productID, labels & class with order count & total sales.
15. Exploring our all 3 clusters we found that cluster number 1 is giving us high demanded products

16. Now, creating a new dataframe by taking some columns from the main data (Asin, tittle, main cat, primary category, sub cat) and merging It into the Df4. For extra information about the product.
17. Extracting the data where label = 1 and exported to a csv file to achieve a final list of highly demanding products for reference

Conclusion-
By applying K-Means clustering we can conclude that the Cluster labelled 1 is having the highest demanding products with respect to highest revenue.

Model 8: Office Products Demand Analysis Using K-Means clustering

Procedure:

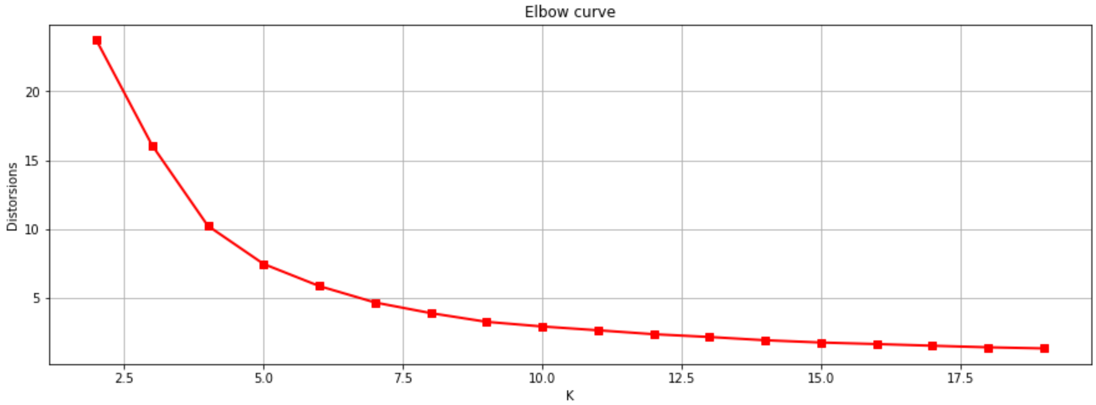
1. Importing the required libraries and importing the final reviews dataset.

2. Office Products category is extracted from the main category column.

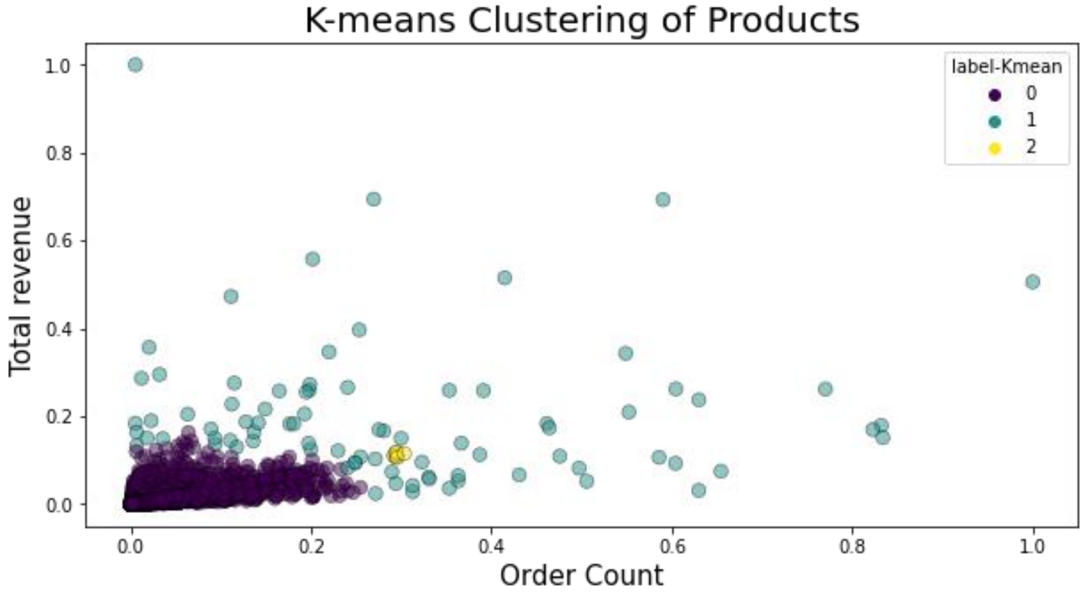
3. Created a dataframe using groupby function of pandas, grouping it on ‘asin’ column to take count of particular asin.
4. Renaming the columns name from ‘asin’ to ‘Product ID’ & ‘count’ to ‘order_count’.
5. Created a dataframe using groupby function of pandas, grouping it on ‘asin’ and ‘price’ column to take total sales for particular ProductID.
6. Merging the dataframe of order count and total sales.

7. Sorting the merged dataframe by ordercount in ascending order.
8. Feature Engineering
   Going further with data scaling using normalisation (MinMaxScaler) on column ‘order_count’ & ‘total_sales.

9. To find an optimal number of clusters using elbow method on the scaled dataframe.
   We are getting 5 number of clusters.



10. To check the goodness of cluster calculating silhouette scores where we’re getting a good score for 3 cluster.
11. Moving forward for modelling with K-Means clustering with 3 clusters.
12. Plotting a scatter plot for 3 clusters.



13. By seeing the scatter plot, we can conclude that the Cluster 1 having ‘High’ number of order’s and generating high revenue.

14. Creating a dataframe name Df4 and merging productID, labels & class with order count & total sales.
15. Exploring our all 3 clusters we found that cluster number 1 is giving us high demanded products
16. For further analysis by DB Scan, creating a dataframe name Df5 and merging productID, labels & class with order count & total sales.

17. Now, creating a new dataframe by taking some columns from the main data (Asin, tittle, main cat, primary category, sub cat) and merging It into the Df5 by dropping duplicates and ‘asin’ column, For extra information about the product
18. Extracting the data where label = 1 and exported to a csv file to achieve a final list of highly demanding products for reference

Conclusion-
By applying K-Means clustering & DB Scan we can conclude that the Cluster labelled 1, -1 is having the highest demanding products with respect to highest revenue.

Model 9 : Time Series Analysis for Demand Forecasting of Product (ASIN “B00SWBLS3C”) Digital Music

Procedure:

1. Importing the required libraries and importing the final reviews dataset.

2. Importing rating data as it will provide an actual insight of sales.

3. Converting the ‘date’ column into datetime data type.

4. Digital Music product category is extracting from the main category column.

5. Creating a dataframe named ’x’ using groupby function of pandas, grouping it on ‘asin’ & ‘price’ column to take total price for a particular asin.

6. Sorting the merged dataframe by ordercount in descending order.

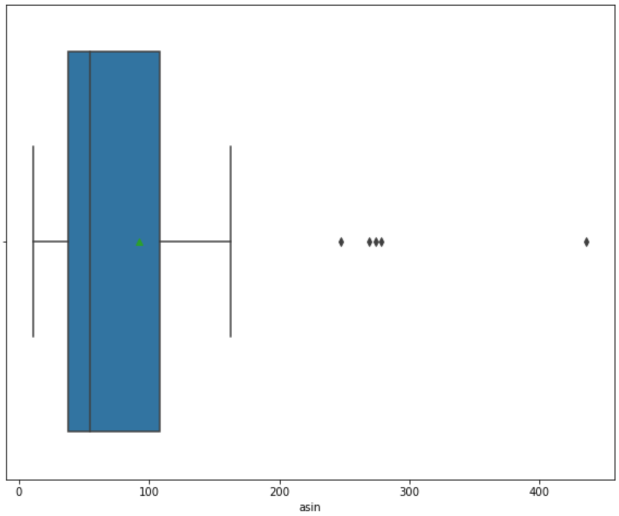
7. Creating a dataframe named ‘y’ using groupby function of pandas, grouping it on ‘asin’ & ‘price’ column to take count for a particular asin and renaming the column ‘price’ to ‘count’ & sorting the values by ‘count’ in descending order.

8. Merging the dataframe ‘x’ & ‘y’

9. Extracting the columns ‘title’,’asin’ & ’date’ for required ‘asin’.

10. Creating a dataframe named ’b’ using groupby function of pandas, grouping it on count of ‘asin’ column behalf of monthly to take number of products ordered in a particular month.

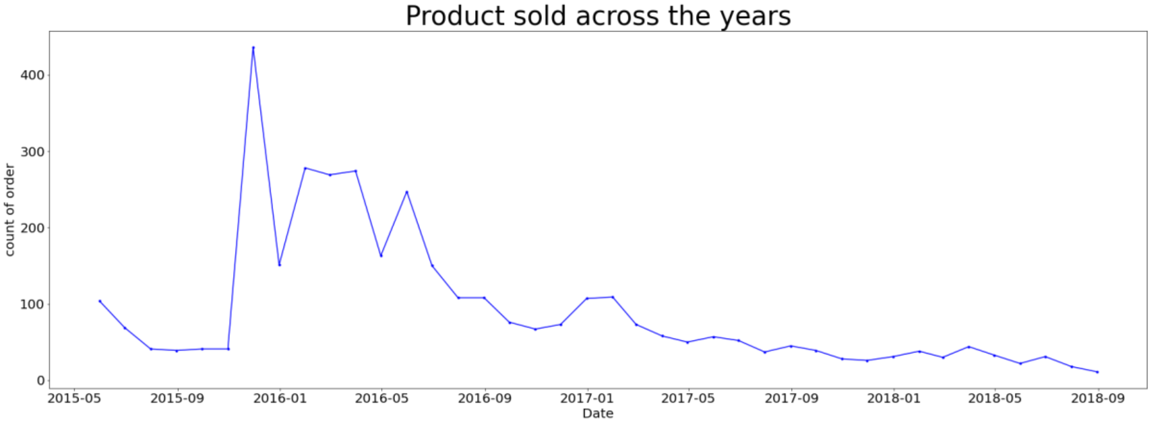
11. Resampling the data quarterly.



12. Modelling for inventory optimization.

13. Plotting the line plot to check the trend & seasonality of the data.

14. Plotting the box plot



15. From the above plot it is evident that there are few outliers in the data. Having outlier might impact the future modelling. However not all outliers are false values, some values can be the actual sales figures.

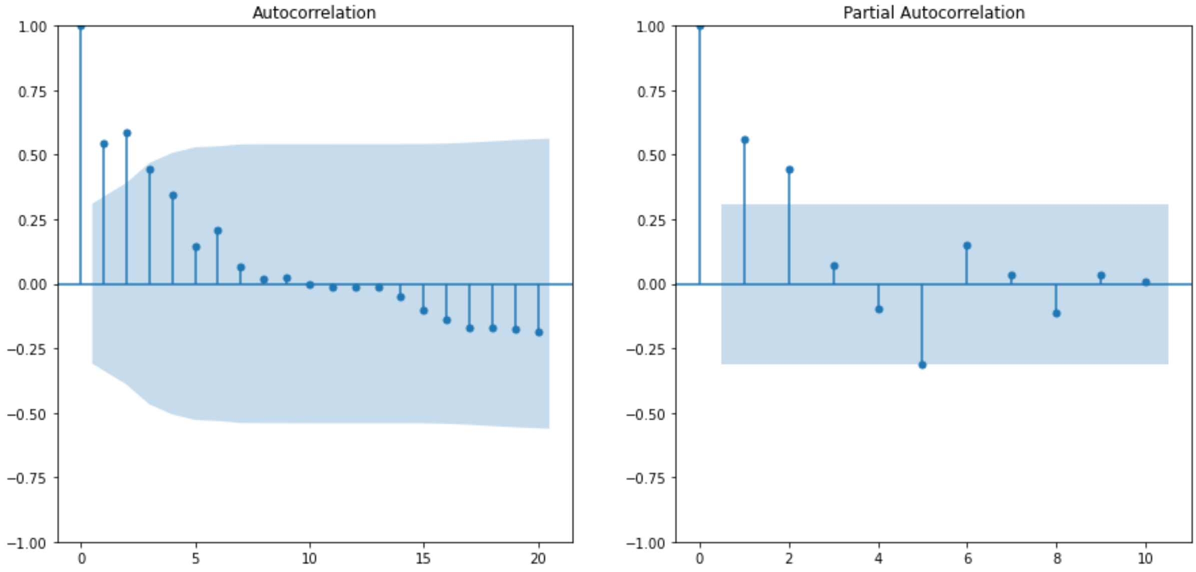
16. Decomposition to check the trend and seasonality in the data

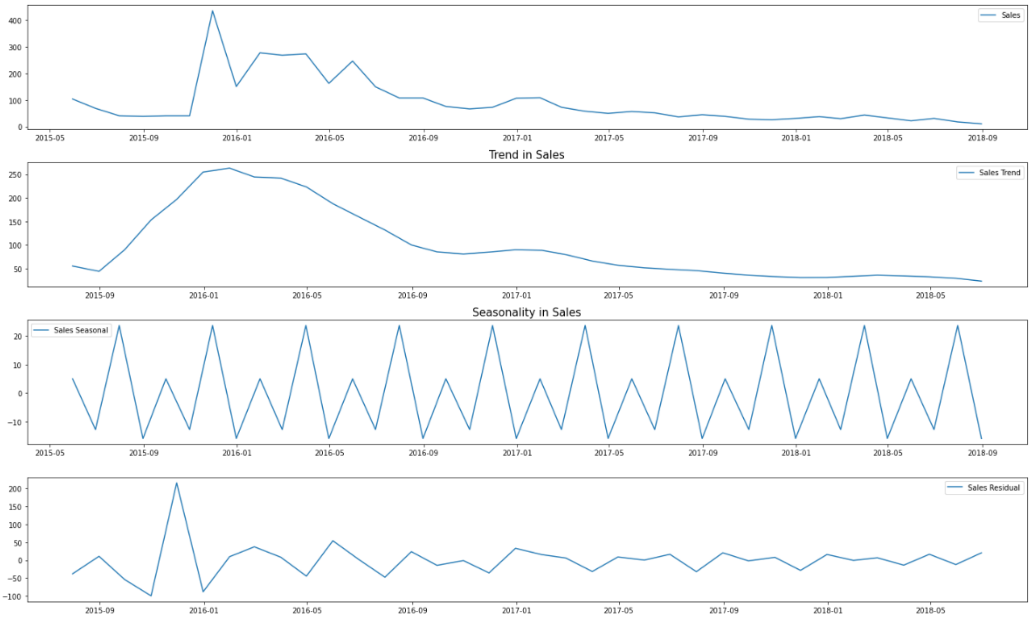
From the above plots we can observe that the data has visual evidence of seasonality as there is repetition in the pattern across a set period of time. However, the data lacks trend

17. Checking the stationarity of the data via adufler test

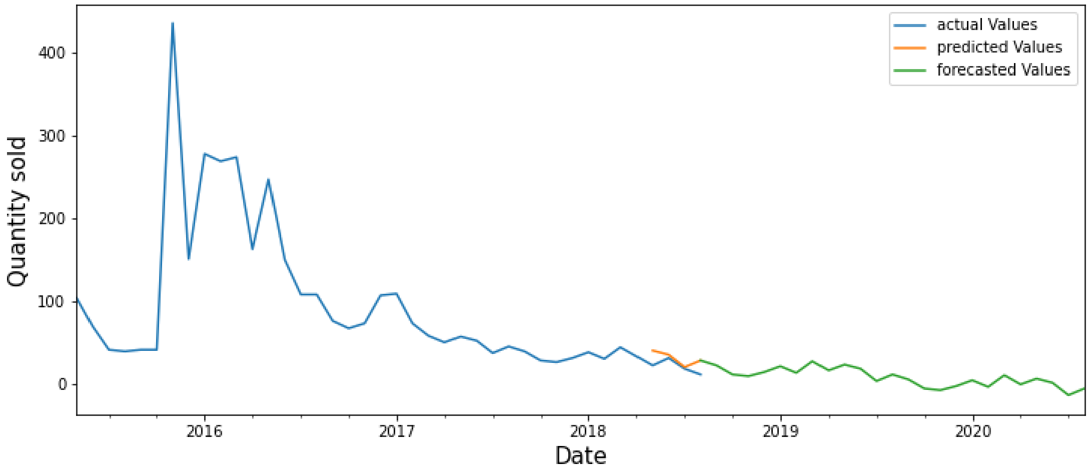
    By seeing the test value, we can conclude that our data is not stationary.

18. Now, we are differencing the data to make it stationary. After completion of differencing, we can see the data is stationary.
19. Before Modelling we have to plot the ACF and PACF for calculating the value of p and q as needed for modelling.





20. From the above plot the following area the max values: p = 3 , q = 2.
21. Splitting the data for training and testing by the ratio 90:10
22. Going further with Hyper-parameter tunning for modelling
23. By seeing the Summary Hyper-parameter tunning, we can conclude that our ‘Model 0 is the best model for forecasting.
24. Proceeding with SARIMA model, By generating p=0,q=0 &d=1 with seasonal order(P,d,Q,12)
25. After training the model we’re predicting our data and finding the error between the actual & predict values and also showing MSE & RMSE values.
26. Using the same model the values are forecasted for next 24 months.
27. Visualising the forecasted graph which shows our actual values & predicted values



Conclusion- Demand is likely to fall with seasonal pattern

Model 10:
Time Series Analysis for Demand Forecasting of Product (ASIN “B00AE9V3WQ”) Office Products

Procedure:

1. Importing the required libraries and importing the final reviews dataset.

2. Importing rating data as it will provide an actual insight of sales.

3. Converting the ‘date’ column into datetime data type.

4. Digital Music product category is extracting from the main category column.

5. Creating a dataframe named ’y’ using groupby function of pandas, grouping it on ‘asin’ & ‘price’ column to take total price for a particular asin.

6. Sorting the merged dataframe by ordercount in descending order.

7. Creating a dataframe named ‘merge’ using groupby function of pandas, grouping it on ‘asin’ & ‘price’ column to take count for a particular asin and renaming the column ‘price’ to ‘count’ & sorting the values by ‘count’ in descending order.

8. Merging the dataframe ‘x’ & ‘y’

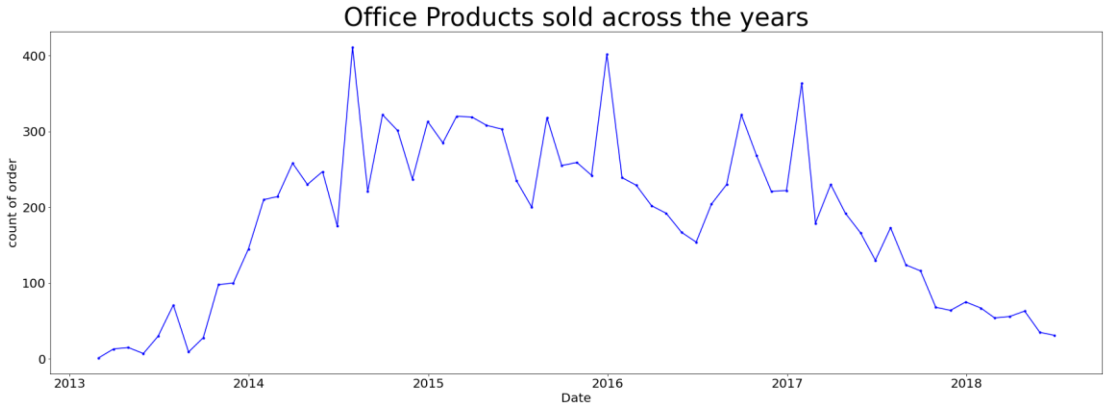
9. Extracting the columns ‘title’,’asin’ & ’date’ for required ‘asin’.(B00AE9V3WQ)

10. Creating a dataframe named ’b’ using groupby function of pandas, grouping it on count of ‘asin’ column behalf of monthly to take number of products ordered in a particular month.

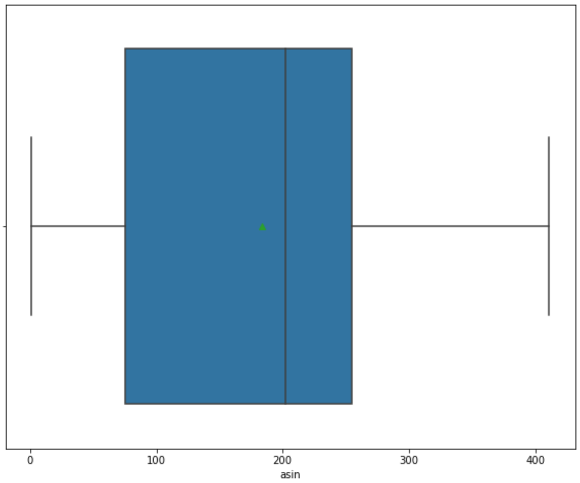
11. Resampling the data quarterly.

12. Modelling for inventory optimization.

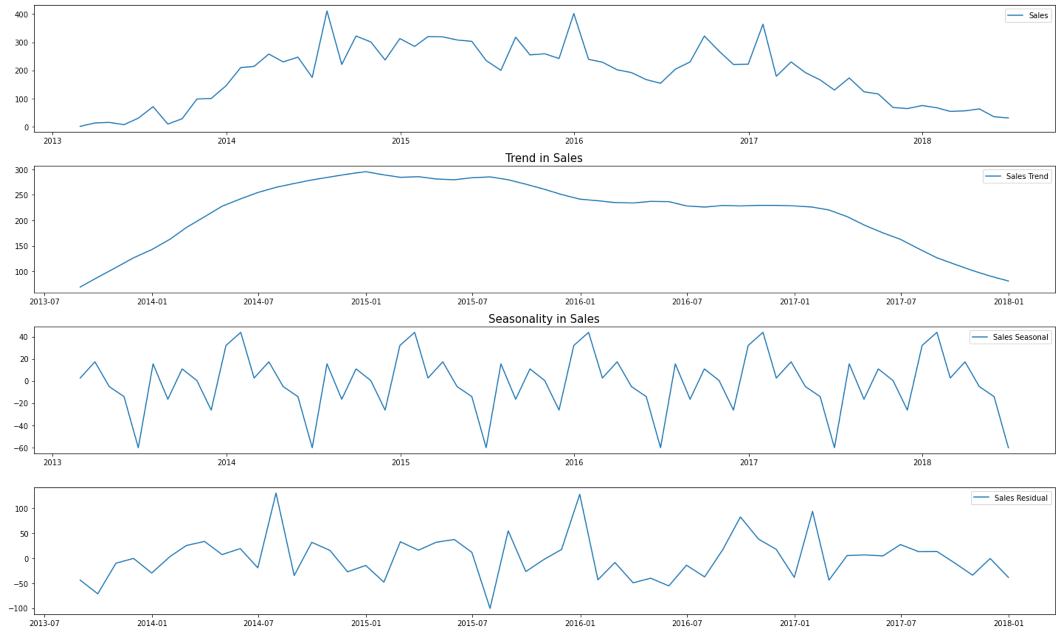
13. Plotting the line plot to check the trend & seasonality of the data.



14. Plotting the box plot



15. From the above plot it is evident that there are few outliers in the data. Having outlier might impact the future modelling. However not all outliers are false values, some values can be the actual sales figures.
16. Decomposition to check the trend and seasonality in the data



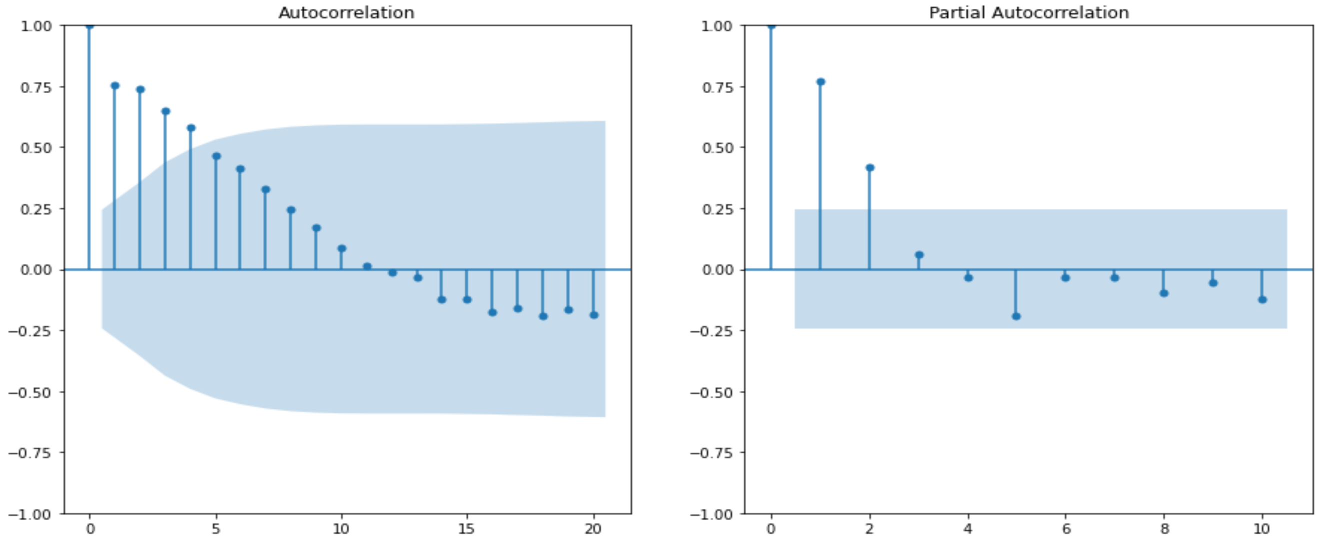
17. From the above plots we can observe that the data has visual evidence of seasonality as there is repetition in the pattern across a set period of time. However, the data lacks trend

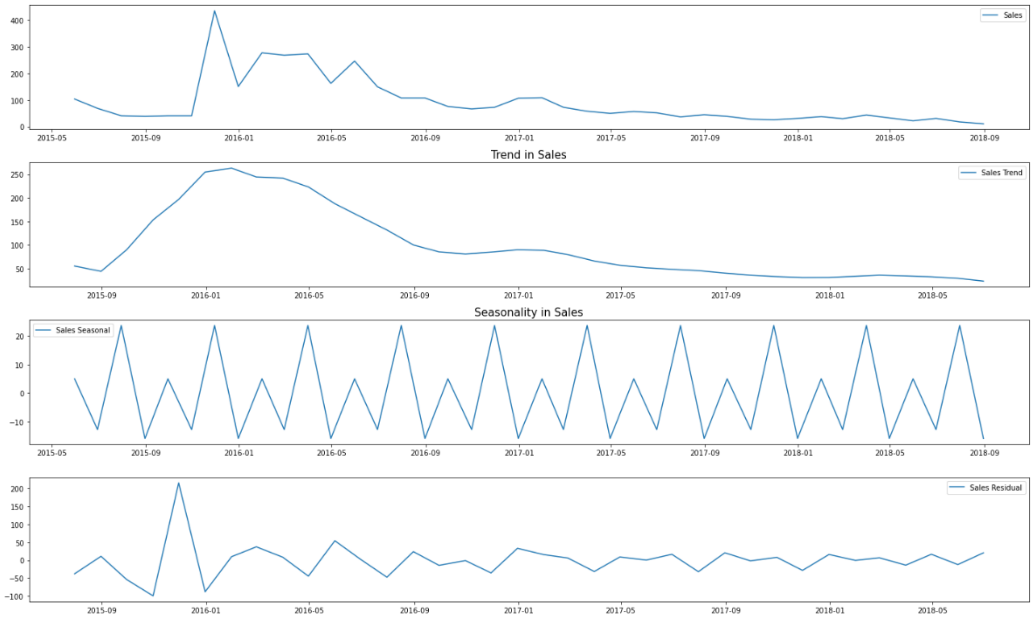
18. Checking the stationarity of the data via adufler test

    By seeing the test value, we can conclude that our data is not stationary.

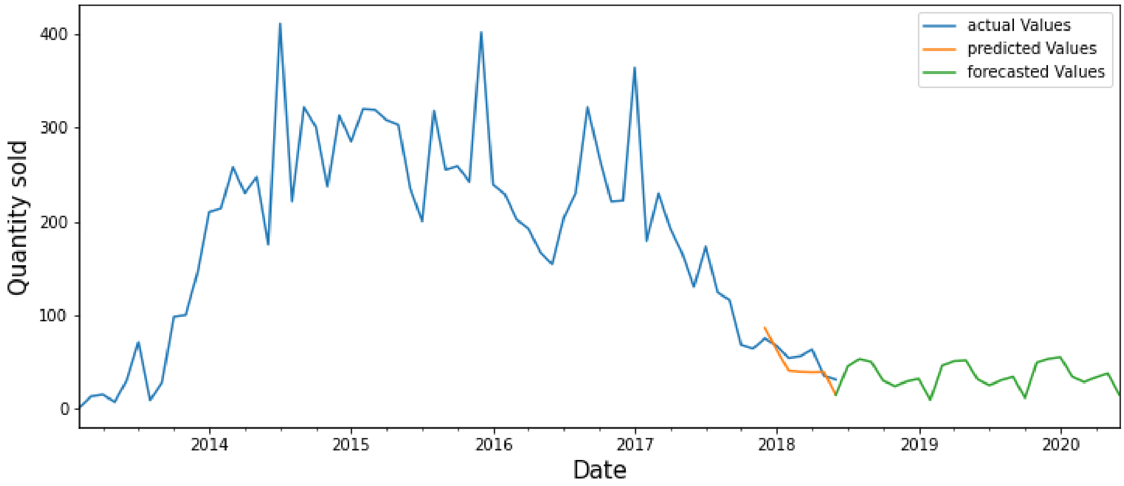
19. Now, we are differencing the data to make it stationary. After completion of 1 shift, we can see the data is stationary.

20. Before Modelling we have to plot the ACF and PACF for calculating the value of p and q as needed for modelling.





21. From the above plot the following area the max values: p = 2, q = 4.
22. Splitting the data for training and testing by the ratio 90:10
23. Going further with Hyper-parameter tunning for modelling
24. By seeing the Summary Hyper-parameter tunning, we can conclude that our ‘Model 9’ is the best model for forecasting.
25. Proceeding with SARIMA model, By generating p=0, q=0 &d=1 with seasonal order(P,d,Q,8)
26. After training the model we’re predicting our data and finding the error between the actual & predict values and also showing MSE & RMSE values.
27. Using the same model, the values are forecasted for next 24 months.
28. Visualising the forecasted graph which shows our actual values & predicted values



Conclusion- The demand for this product is likely to fall in future with seasonal patten

APPENDIX:

Link for Data Source →
https://jmcauley.ucsd.edu/data/amazon/

Working code book link→
https://drive.google.com/drive/folders/1e3tNUDN2PHsftue61Z3Ojbv13uk5qRmE?usp=share_link

REFERENCE
We have used these websites for reference:

Stack Overflow - Where Developers Learn, Share, & Build Careers
GeeksforGeeks | A computer science portal for geeks