



# PROJECT DOCUMENTATION

AMAZON PRODUCT REVIEW ANALYSIS

- Ravali K
- Bhargavi M
- Sruthi Varsha G
- Umesh Chandra M

# Introduction:

Amazon is an American multinational technology company focusing on e-commerce, cloud computing, online advertising, digital streaming, and artificial intelligence. It has been referred to as "one of the most influential economic and cultural forces in the world", and is one of the world's most valuable brands. It is one of the Big Five American information technology companies, alongside Alphabet (Google), Apple, Meta (Facebook), and Microsoft. Because of its popularity and ubiquity, Amazon is really the place where people actually spend time and write detailed reviews, unlike other platforms where consumers have to be nudged. Amazon review data analysis can tell companies a lot about their product, even elements that they might not have thought of, as the example we will be analyzing later will demonstrate.

It is observed that the maximum number of customers look at product reviews before they make a purchase. Survey results show that positive product reviews are a key factor for purchasing by 57% of Amazon buyers (Statista, 2019). As product reviews are often the deciding factor for many customers, it's very important to have a well-automated system for monitoring them. The traditional manual process of Amazon product reviews is time-consuming and inefficient when millions of reviews are being posted all the time. It doesn't show any trend or patterns over time. Moreover, it is tough to understand customers' sentiment towards any product or its delivery.

## Objective of this project:

- ❖ To monitor and analyze customer reviews to gain valuable insights on customer preferences and behavior.
- ❖ To use the insights gained from customer reviews to improve the success rate of existing and new products

# **Scenarios to be solved:**

## **Scenario 1: 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.

## **Methods used to reach objective:**

- Data Collection
- Exploratory Data Analysis
- Sentiment Analysis
- Classification on customer retention
- Time series analysis on demand product and sentiments
- Product Recommendation
- Clustering on customer segmentation
- Conclusion

# **Data Collection:**

## **Datasets used:**

The dataset contains product reviews and meta data from Amazon, including 142.8 million reviews spanning from May 1996 to October 2018.

In Amazon website, they are so many categories, among those I will use 4 datasets i.e.,

- ❖ Office Products
- ❖ Office Products meta data
- ❖ Cd's and Vinyl
- ❖ Cd's and Vinyl meta data

**Refer to the below link for the datasets.**

<https://nijianmo.github.io/amazon/index.html>

# Reading the Datasets:

We will be using the 5-core dataset and also the meta data of the categories of office products and cd vinyl. As the data's are in the json format we are using the following the function to read them as pandas dataframe.

## Pandas data frame

This code reads the data into a pandas data frame:

```
import pandas as pd
import gzip

def parse(path):
    g = gzip.open(path, 'rb')
    for l in g:
        yield json.loads(l)

def getDF(path):
    i = 0
    df = {}
    for d in parse(path):
        df[i] = d
        i += 1
    return pd.DataFrame.from_dict(df, orient='index')

df = getDF('reviews_Video_Games.json.gz')
```

# Understanding the Data:

## The columns of 5-core data:

- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewerName - name of the reviewer
- vote - helpful votes of the review
- style - a dictionary of the product metadata, e.g., "Format" is "Hardcover"
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)
- image - images that users post after they have received the product.

## The columns of Metadata:

- asin - ID of the product, e.g., 0000031852
- title - name of the product
- feature - bullet-point format features of the product
- description - description of the product
- price - price in US dollars (at time of crawl)
- imageUrl - url of the product image
- imageURL - url of the high resolution product image
- related - related products (also bought, also viewed, bought together, buy after viewing)
- salesRank - sales rank information
- brand - brand name
- categories - list of categories the product belongs to
- tech1 - the first technical detail table of the product
- tech2 - the second technical detail table of the product
- similar - similar product table

## The data shape of office products:

```
Total Shape of Office supplies data: (800357, 12)
Total Shape of Office meta data: (315458, 19)
```

## The data shape of Cd vinyl:

```
Total Shape of CD Vinyl data: (1443755, 12)
Total Shape of CD Vinyl meta data: (516914, 19)
```

## Extracting the useful information from metadata:

We have extracted columns like brand, category, main cat from meta data and merged with the 5-core data on the basis of asin for further analysis.

After merging the meta and 5-core data the shape of office products is:

```
Shape after merging: (915962, 15)
```

After merging the meta and 5-core data the shape of Cd vinyl is:

```
Shape after merging: (1722761, 15)
```

## Data Preprocessing and cleaning:

Data preprocessing involves the transformation of the raw dataset into an understandable format. Preprocessing data is a fundamental stage in data mining to improve data efficiency. The data preprocessing methods directly affect the outcomes of any analytic algorithm.

Preprocessing has been done in two separate notebooks for office products and cd vinyl data.

## Missing values check:

The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values

The presence of missing values reduces the data available to be analyzed, compromising the statistical power of the study, and eventually the reliability of its results. In addition, it causes a significant bias in the results and degrades the efficiency of the data

## Percentage of Missing values in Office Products Data:

```
overall      0.000000
verified     0.000000
reviewTime   0.000000
reviewerID   0.000000
asin         0.000000
style        35.197530
reviewerName 0.017577
reviewText   0.026639
summary      0.015940
unixReviewTime 0.000000
vote         89.286128
image        98.659224
category     0.157539
brand        0.157539
main_cat     0.157539
dtype: float64
```

## Percentage of Missing values in CD Vinyl Data:

```
reviewerID   0.000000
asin         0.000000
reviewerName 0.003947
verified     0.000000
reviewText   0.017820
overall      0.000000
reviewTime   0.000000
summary      0.015556
unixReviewTime 0.000000
style        3.973738
vote         65.290310
image        99.753767
category     8.525268
brand        8.525268
main_cat     8.525268
dtype: float64
```

→ **Missing values Imputation:** Checked for null values and imputed it with mode and the null values less than 1% or nearly equal to zero are dropped and dropped the columns that had more than 50% null values.

→ Checked the data and replaced the links in 'main cat' column as related category and also the space was replaced with mode.

→ Dropped the unwanted columns and duplicate records in data.

→ Changed the date column to datetime type and overall column as int16.

Converted the pre-processed data to csv files. Again in new notebook we had merged the two csv files and then again repeated the steps to check whether the data is in correct form or not. Now the data is ready for EDA and Modelling.



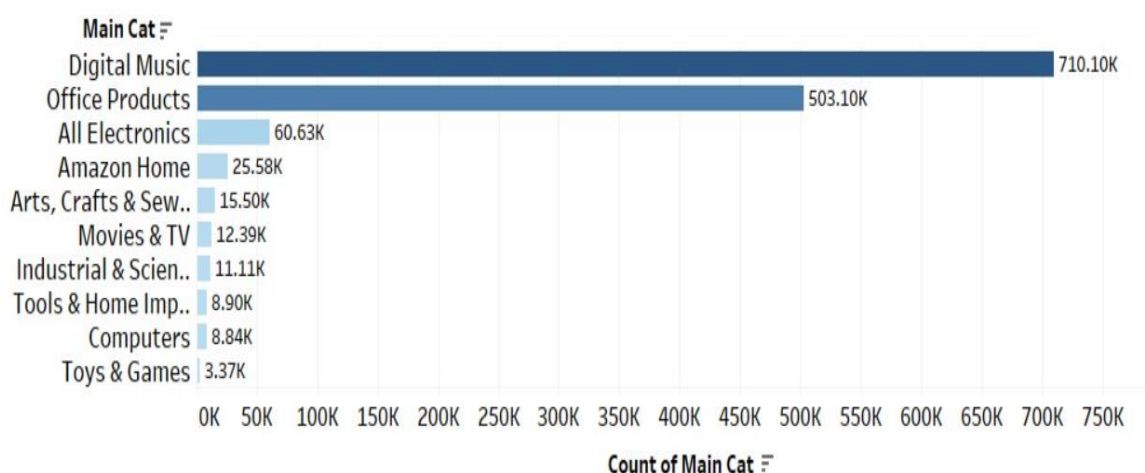
## Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques.

It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

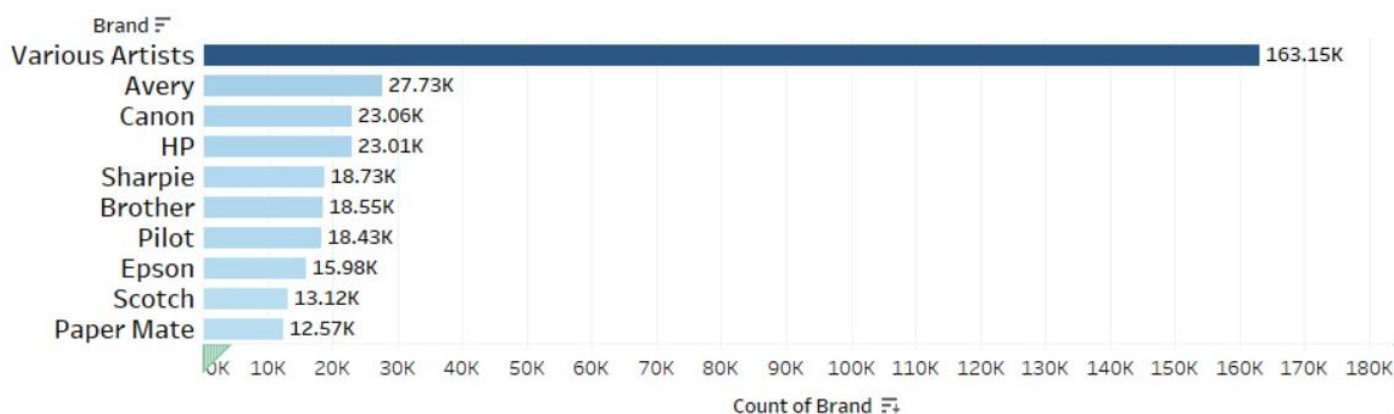
## UniVariate Analysis:

### Top 10 Main Categories



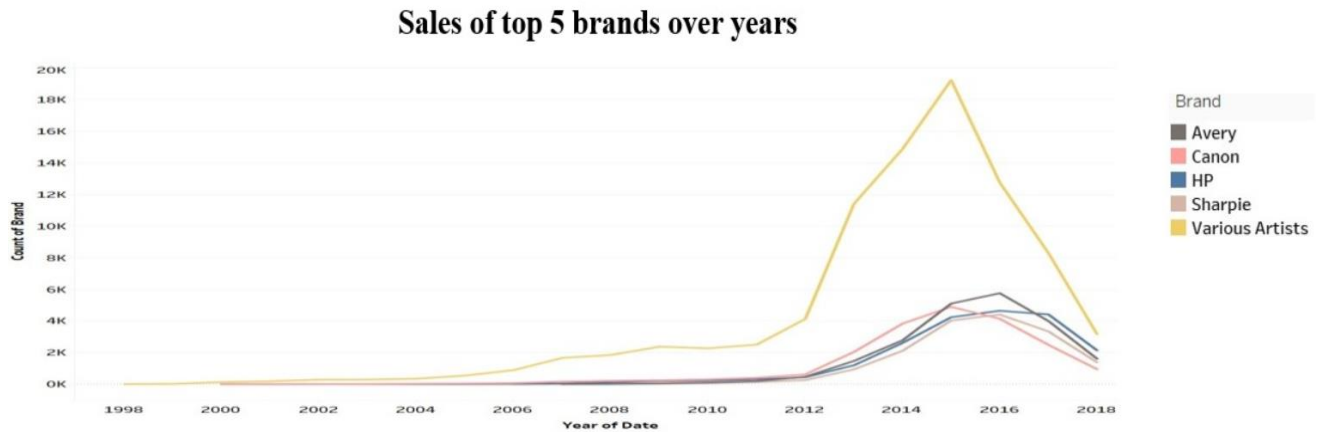
**Observation:** Digital Music, Office Products and All Electronics are the top 3 Main Categories preferred by the customers.

### Top 10 Brands

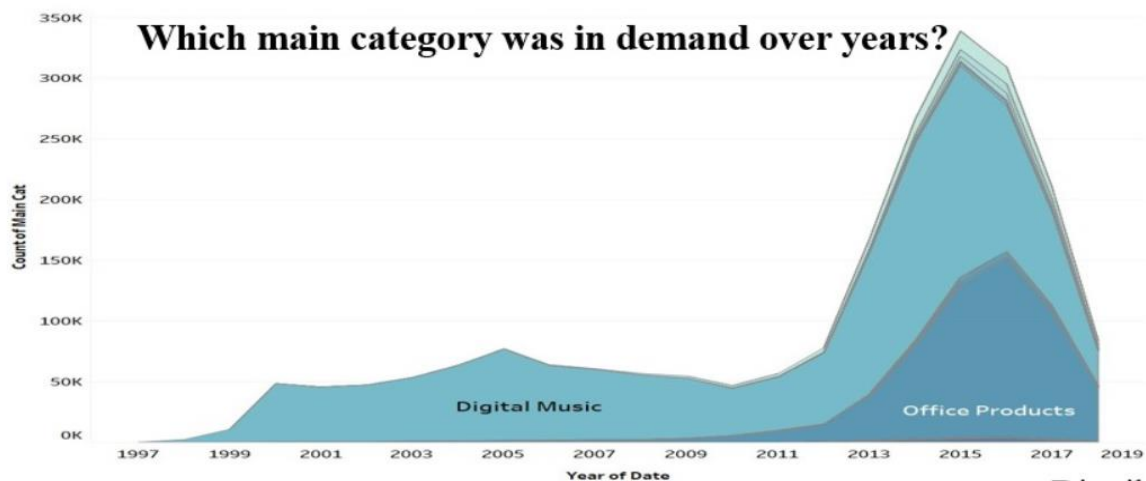


**Observation:** Various Artists, Avery and Canon are the top 3 most preferred brands by the customers.

## Bivariate Analysis:



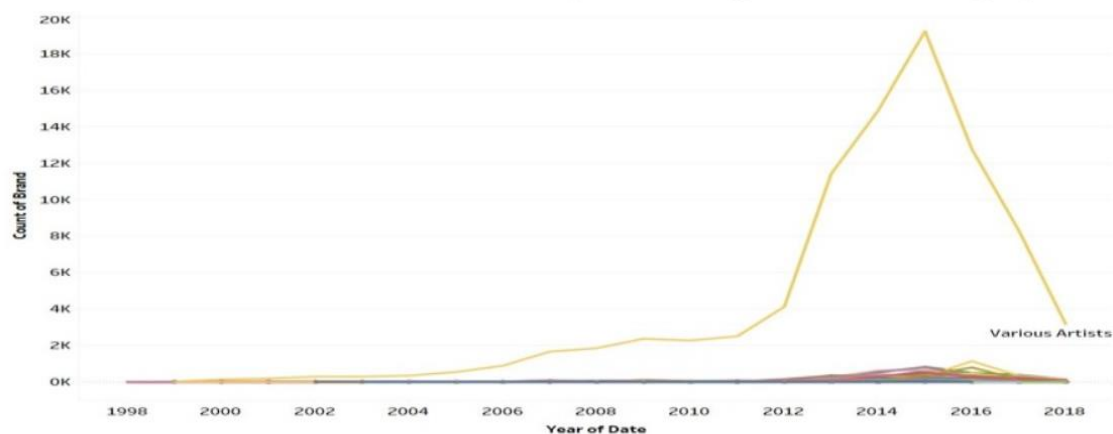
**Observation:** Avery tops the highest sales over the years compared to other brands.



**Observation:** Digital Music and Office Products are in demand over the years.

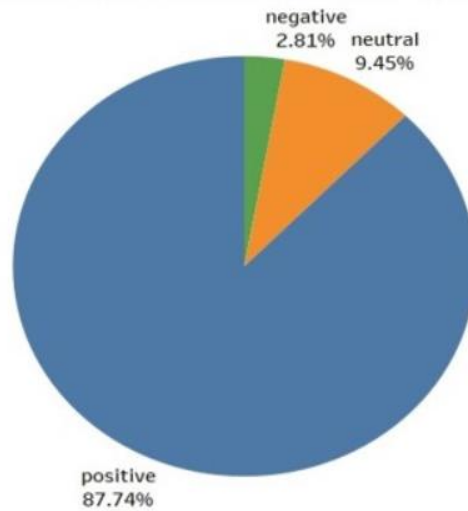
## Multivariate Analysis:

**Which brand was in demand over the years of Digital Music Category?**



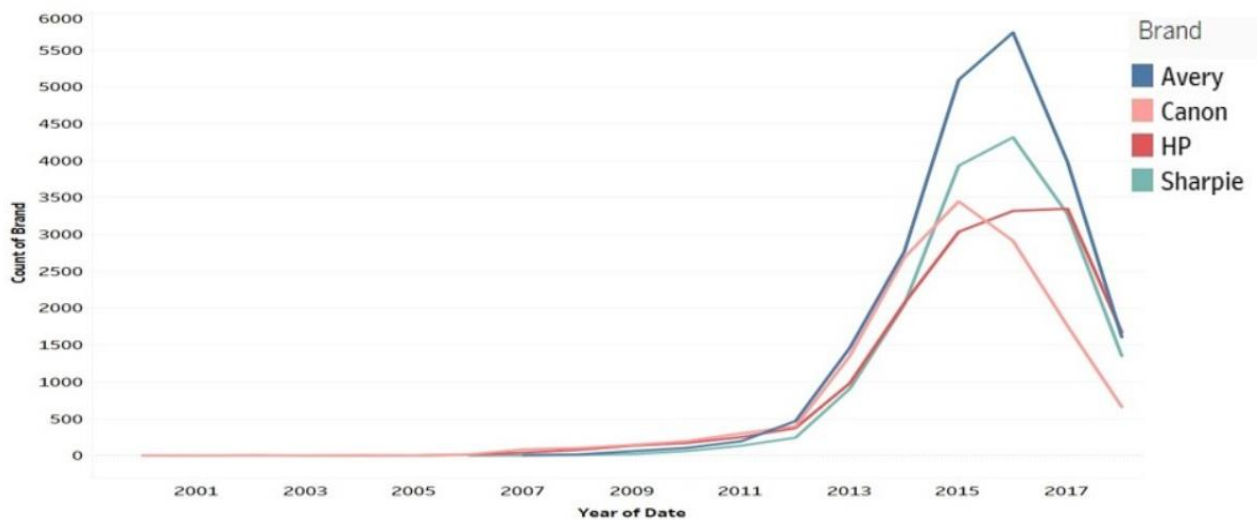
**Observation:** The brand Various Artists was in demand over the years of Digital Music Category

### Distribution of sentiment of **Digital Music** with Various Artists' brand



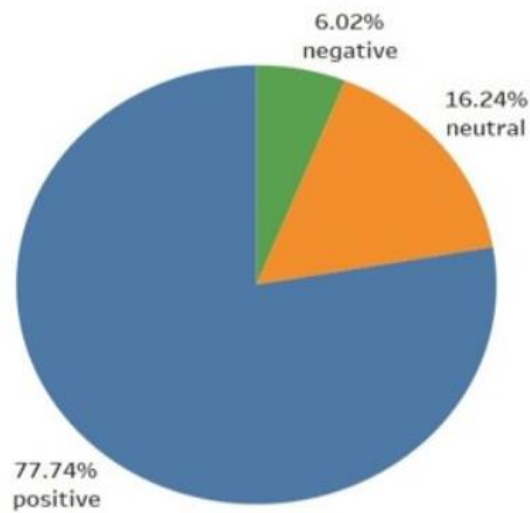
**Observation:** The Brand Various Artists has about 87.74% of positive Sentiments, 9.45% of Neutral Sentiments and 2.81% of Negative sentiments.

### Which brand was in demand over the years of the Office Products Category?

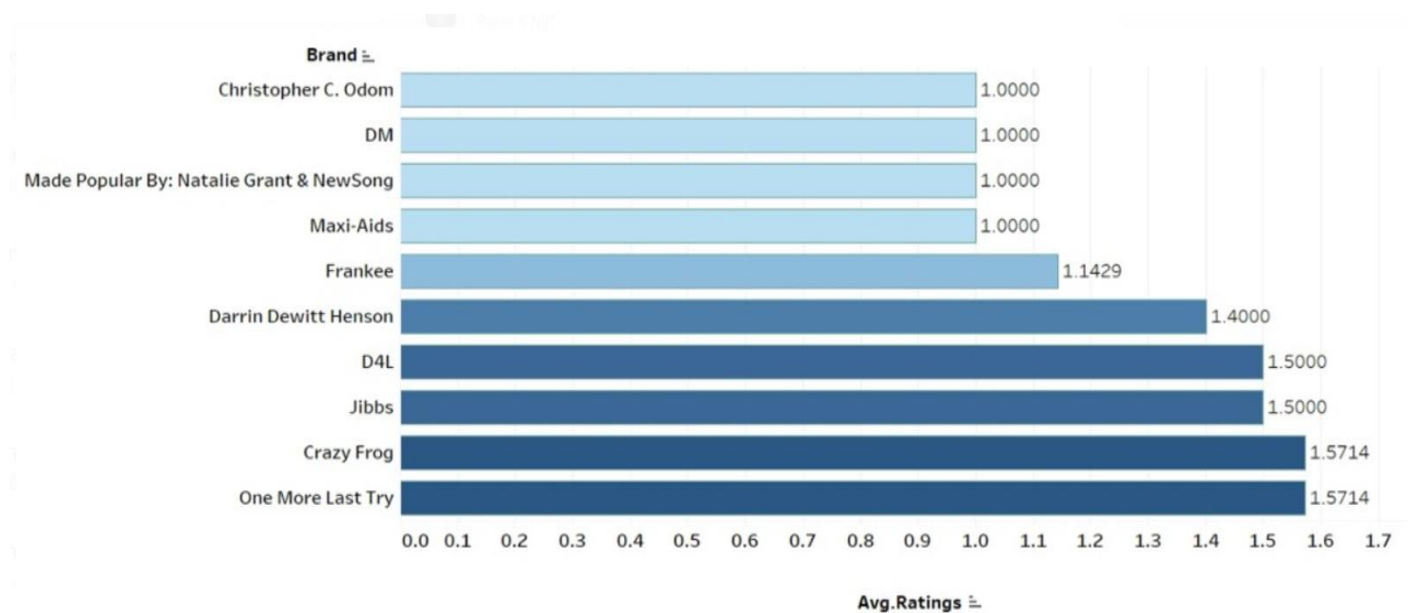


**Observation:** The brand Avery was in demand over the years of Office Products Category.

## Distribution of sentiment of Office Products with Avery brand



**Observation:** The Brand Avery has about 77.74% of positive Sentiments, 16.24% of Neutral Sentiments and 6.02% of Negative sentiments.



**Observation:** The brands Christopher C.Odom and DM are the top 2 brands that are at high risk of customer retention.

**Interactive Dashboard Link:** [dashboard](#)

# Modelling

## **Sentimental Analysis:**

Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service, or idea.

### **Objective:**

To analyze the attitude or emotional state of the customer from their posted review text as positive, negative and neutral sentiments.

### **Main Libraries used:**

- Textblob
- CleanText & NeatText(for text cleaning)

### **Model used:**

- LogisticRegression (OVR)
- Fasttext
- MultinomialNB

**Columns used:** Reviewtext, Analysis

## **Data preparation for sentimental Analysis:**

- Created a user defined function to clean the data with predefined libraries like CleanText and NeatText (Removed the multiple spaces, numbers, special characters, stop words...etc.)
- Added a new column as polarity which gives the polarity score of the review text by using Textblob.

## Why Text Blob?

### Comparing Text Blob and Vader Sentiment Analyzer

```
For word ---> good ---> using Textblob (0.7 , Positive) ---> using vader (0.4404 , Positive)
For word ---> bad ---> using Textblob (-0.7 , Negative) ---> using vader (-0.5423 , Negative)
For word ---> perfect ---> using Textblob (1.0 , Positive) ---> using vader (0.5719 , Positive)
For word ---> tasteless ---> using Textblob (-0.6 , Negative) ---> using vader (0.0 , Neutral)
```

On Comparison of polarity scores obtained, Textblob gives a better polarity scoring than Vader.

- Calculated the sentiments (pos, neg, neu) based on the polarity score.
- Splitted the data into train and test data on the ratio of 80:20.
- Performed the modelling using fasttext, Logistic Regression and MultinomialNB.
- Evaluated the models using f1\_score.
- By comparing all the f1\_scores of models, Fasttext had given the best f1\_score.
- So, we have used the Fasttext model for our prediction.
- Now, our model is able to recognize the polarity of reviewtext.

## Prediction of the Model:

```
Enter a text      : I loved that particular book
```

```
-----
Actual Results
```

```
Polarity Score   : 0.4
```

```
This review is   : Positive
```

```
-----
Predicted Results
```

```
Label           : 2
```

```
Sentiment       : Positive
```

# Classification:

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

**Objective:** Prediction of Customer retention based on the categories.

**Goal:** And the ultimate goal of churn analysis is to reduce churn and increase profits. As more customers stay longer, revenue should increase, and profits should follow.

## Main Libraries used:

- sklearn
- flaml (AutoML)

## Model Used:

- AutoML (for finding the best algorithm)
- LGBMClassifier
- Logistic Regression.

**Columns Used:** Brand, main\_cat, polarity\_scores, target (created)

- The data is grouped on basis of main\_cat, brand with average polarity scores.
- The target is created by defining a function with a condition of getting 1 if the data is greater than equals 1 else 0 and passing the polarity scores which yields 1 and 0 based on that condition.

## Feature Engineering:

- Converting the object type columns into categorical type.
- Label Encoding: The columns brand and main\_cat is encoded using label encoder.
- Data Scaling: Independent features are scaled using MinMax Scaler.

## Setting Independent(X) and Dependent(Y) features:

X: brand\_encoded, main\_cat\_encoded

Y: target

- Data Splitting: Splitted the data into train and test data using train\_test\_split function in the ratio of 70:30.
- To select the best algorithm, we have used AutoML model.
- On fitting the train and test data, LGBMClassifier model is identified to be the best model by the AutoML model.
- Using LGBMClassifier model, we have performed the modelling.
- Accuracy of train and test data are obtained as 99%.
- Thus, the model is able to identify whether the brand and the category satisfied the customers or is at risk of customer retention.

## Prediction of the Model:

Enter category :All Beauty

Enter Brand :3M

Label : 1 (Not Left)

This brand have good ratings and most of the customers were enjoyed this brand products



# Time Series: Demand Forecasting

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series.

**Objective:** Predicting the Forecast of the demand products (Digital Music & Office Products).

## Main Libraries Used:

- statsmodel
- sklearn

## Model Used:

- TripleExponentialSmoothing
- SARIMA Model

**Columns Used:** date, main\_cat(count)

- We have identified and taken the top 2 products which are in demand as Digital music and Office products.
- By grouping the main\_cat and extracting its count over the period(date).

## Data Preparation for Time Series:

- The date column is set as the index and sorted.
- The date is resampled into months.
- The Stationarity check was done on the data and the data is made stationary for the model.
- The optimum alpha, beta and gamma values for the model is selected by checking its rmse values.

- Performed the modelling on the resampled stationary data using triple exponential smoothing (holt winter's method).
- Performed the SARIMA model on resampled data.
- RMSE scores are calculated for the two models.

### **Digital Music**

Triple Exp. → RMSE = 813.11

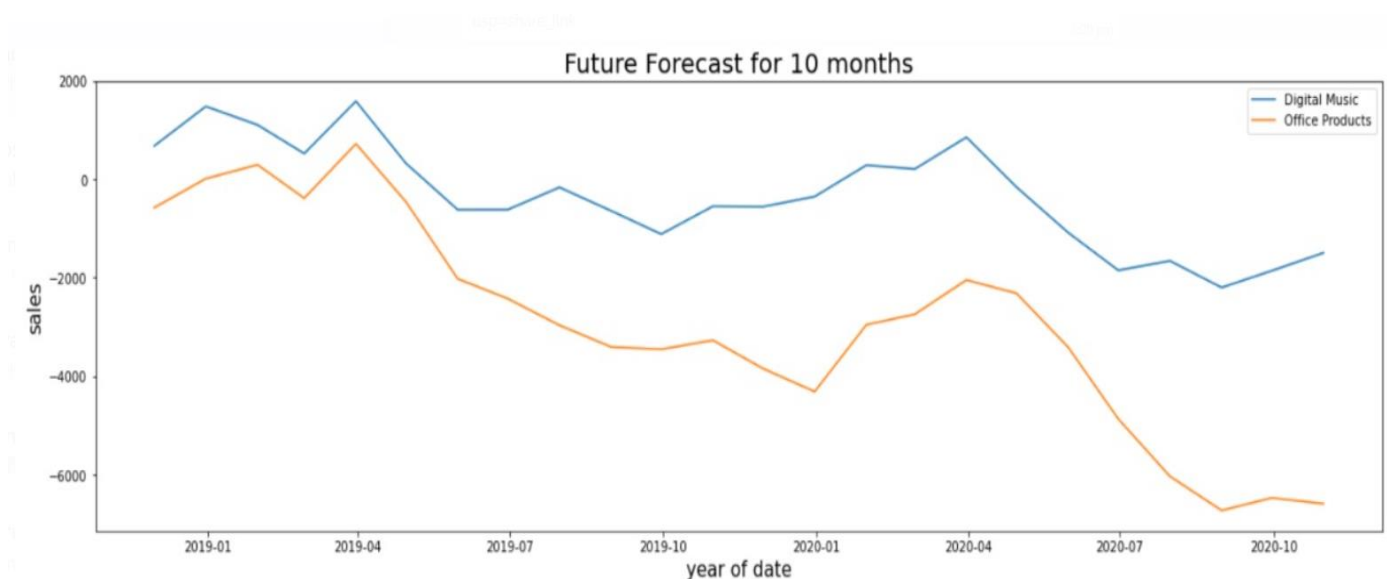
SARIMA → RMSE = 753.45

### **Office Products**

Triple Exp. → RMSE = 492.29

SARIMA → RMSE = 408.70

- By Comparing the RMSE of SARIMA and TripleExponentialSmoothing, SARIMA model has given the less error. So, we have predicted the future forecast using SARIMA Model



- Forecasted the future demand of the products for next 12 months.
- On forecasting the current demand products, it is observed that the future demand for Digital music and Office products are decreasing over the period but comparing the demand Digital Music is in demand than Office Products.

# Time Series: Sentiment Forecasting

**Objective:** Forecasting the sentiments of Digital Music and Office Products over the period of time.

## Main Libraries Used:

- statsmodel
- sklearn

## Model Used:

- TripleExponentialSmoothing
- SARIMA Model

**Columns Used:** date, polarity\_scores, sentiment

## Data Preparation for Time Series:

- The date column is set as the index and sorted.
- The date is resampled into Months.
- The Stationarity check was done on the data and the data is made stationary for the model.
- The optimum alpha, beta and gamma values for the model is selected by checking its rmse values.
- Performed the modelling on the resampled stationary data using triple exponential smoothing (holt winter's method) and the RMSE scores are calculated.
- Performed the modelling on resampled data with SARIMA model by using the best hyperparameters obtained by hyper tuning and the RMSE scores are calculated.

## Office Products: RMSE Scores of Models

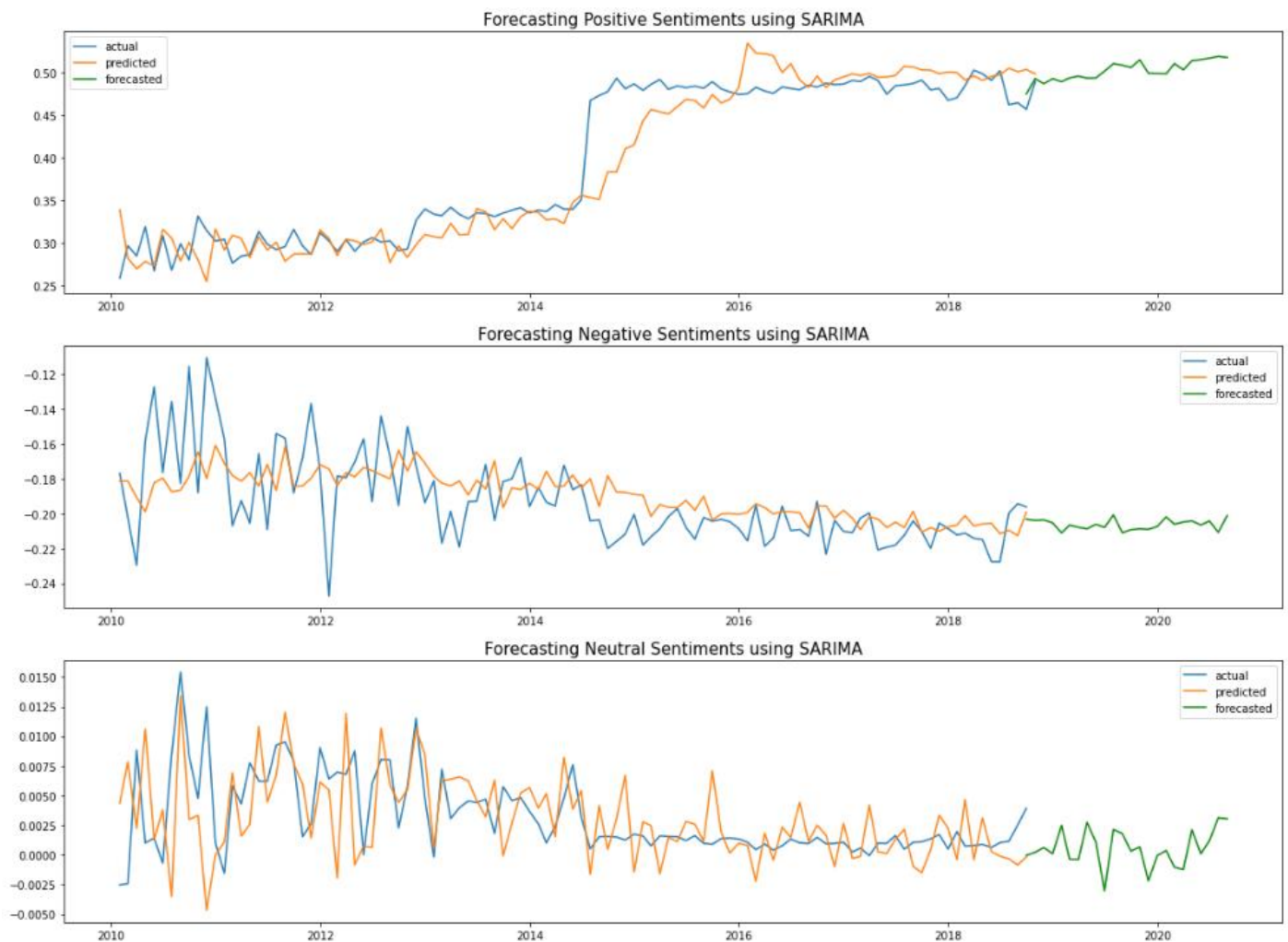
	Triple Exp Smoothing	SARIMA
Positive	0.142061	0.130982
Negative	0.058314	0.067175
Neutral	0.010037	0.009422

## Digital Music: RMSE Scores of Models

	Triple Exp Smoothing	SARIMA
Positive	0.027507	0.035215
Negative	0.043365	0.038591
Neutral	0.003547	0.004243

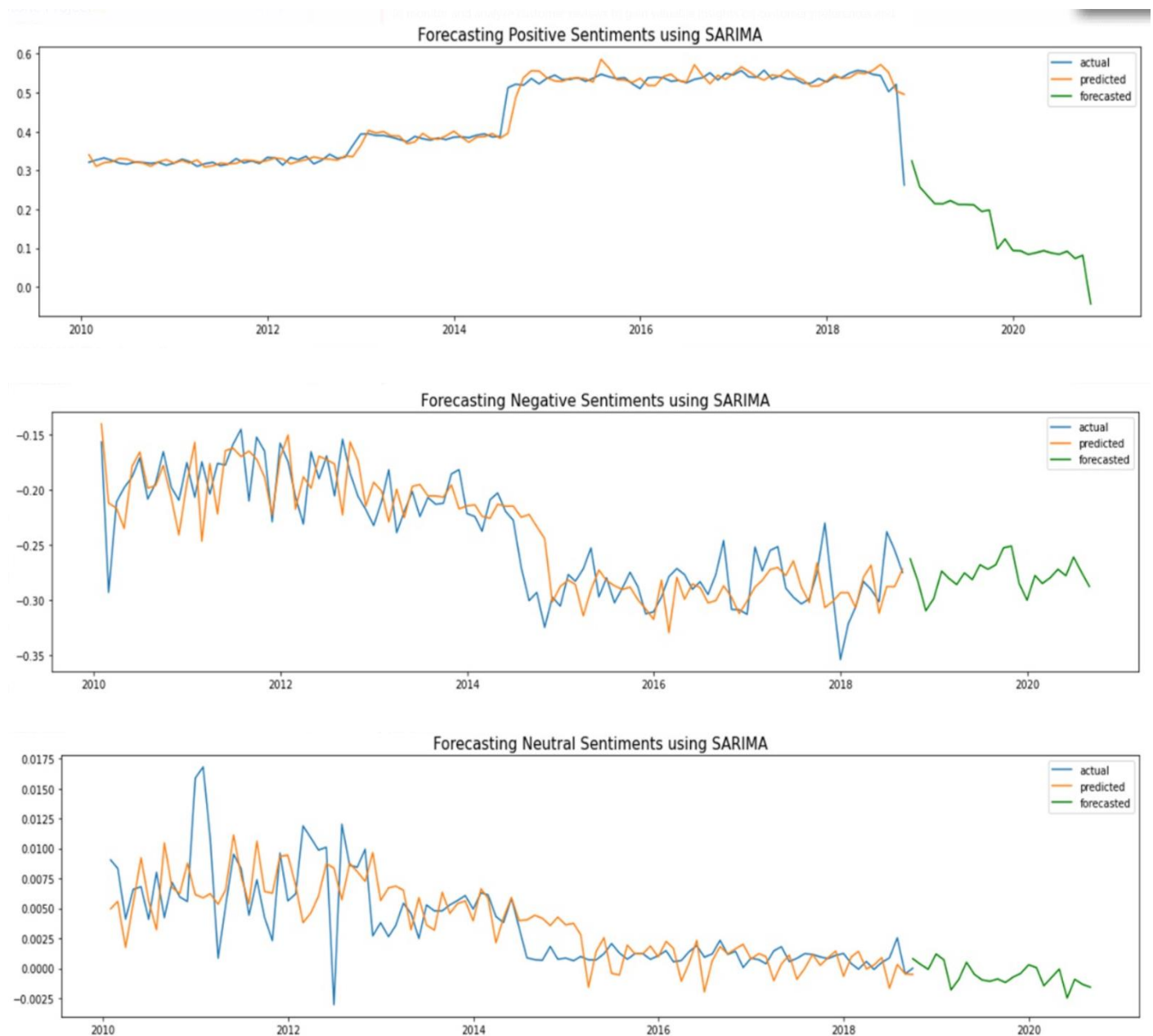
- The RMSE scores obtained by the models are almost equal and on considering the size of the data we have used SARIMA model for the forecasting of Sentiments.

## Forecast of Sentiments of Office Products:



From the above plots, the forecast of positive sentiments are increasing for next two years and the forecast of negative and neutral sentiments seems constant for next two years for Office Products Category.

## Forecast of Sentiments of Digital Music:



From the above graph, the forecasting of neutral and negative sentiment seems to be constant and the positive sentiment seems decreasing for next two years.

# Product Recommendation System:

A product recommendation system is a software tool designed to generate and provide suggestions for items or content a specific user would like to purchase or engage with.

**OBJECTIVE:** Building a recommendation system to recommend similar products to the products purchased by the customer.

**Main Libraries used:** sklearn

**Columns Used:** asin, brand, main\_cat

**Model Used:** kNearestNeighbors

- The duplicated data are removed.
- Label Encoding is done for main\_cat and brand column.
- Setting asin as index.
- Scaling is done to standardize the data for the model.
- The modelling is performed using kNearestNeighbor.
- The distances and indices are obtained from the model to identify the neighbors of the products.
- A function is defined to get the appropriate products that are similar to the purchased product using the indices obtained by the model.
- Thus, the similar products can be recommended to the customer based on their purchased product.

## Prediction of Model:

```
Purchased product is: 0439394058  
Category: Office Products  
Brand: Scholastic  
Recommended Products are:
```

	asin	main_cat	brand
0	0439784395	Office Products	Scholastic
1	0545114780	Office Products	Scholastic
2	0439654939	Office Products	Scholastic
3	0439731593	Office Products	Scholastic
4	0439731771	Office Products	Scholastic
5	0439506042	Office Products	Scholastic

# Clustering: Customer Segmentation

Clustering (also called unsupervised learning) is the process of dividing a dataset into groups such that the members of each group are as similar (close) as possible to one another, and different groups are as dissimilar (far) as possible from one another.

**OBJECTIVE:** To segment customers based on their sentiments

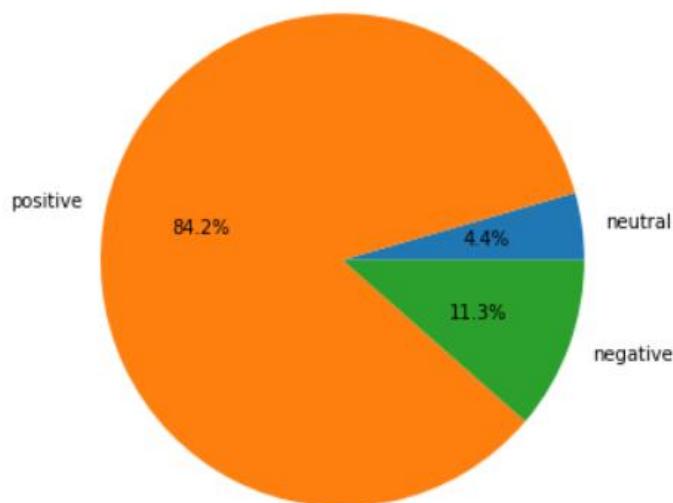
**Main Libraries Used:** sklearn

**Columns Used:** polarity(sentiments)

**Model Used:** Kmeans

- The sentiments in the polarity are vectorized using TfidfVectorizer.
- The modeling is performed and clusters are obtained based on the sentiments.
- From the clusters formed customers can be segmented based on their sentiments as satisfied and dissatisfied customers etc.
- Based on the reviewtext of the customers, the customers are segmented based on three clusters as positive, negative and neutral.

**Ratio of Clusters(sentiments):**





## Satisfied Customers:



## Dissatisfied Customers:













# FINAL CONCLUSION

- ❖ Most of the amazon customers were satisfied with the products quality and services provided.
- ❖ In Customer Retention some of the main categories and brand plays a main role in increase of retention. So, to reduce retention it is better to stop those brands.
- ❖ “Digital Music” is the categories that will be in high demand in the future.
- ❖ Based on the users interest the model automatically offers similar products.
- ❖ Customer segmentation is completely based on the review text, it clusters into 3 groups positive, negative, neutral.
- ❖ Finally, we can say that most of the amazon customers were satisfied with the services and products quality.

## Source codes:

-  **Office products Preprocessing** : [link](#)
-  **CD Vinyl Preprocessing** : [link](#)
-  **Final Preprocessing** : [link](#)
-  **Sentiment Analysis** : [link](#)
-  **Classification** : [link](#)
-  **Time Series Demand Forecast** : [link1](#)  
[link2](#)
-  **Sentiment Forecast** : [link1](#)  
[link2](#)
-  **Product Recommendation** : [link](#)

## References:

- [1] <https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b>
- [2] <https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524#:~:text=Subjectivity%20quantifies%20the%20amount%20of,opinion%20rather%20than%20factual%20information.>
- [3] <https://www.analyticsvidhya.com/blog/2018/05/24-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/>
- [4] <https://www.analyticsvidhya.com/blog/2021/09/guide-for-building-an-end-to-end-logistic-regression-model/>