Project Report on

Amazon Product Reviews for Musical Instruments and Digital Music

Submitted By

Sr.no	**Name of students**
1	Shreyas Kulkarni
2	Chilamakuri Bhavesh
3	Yukti Goyal

Under the Guidance of

Ms. Komilla Bhatia

Table of Contents

ABSTRACT

We developed this project to understand the types of reviews present in Amazon. In raw data all columns are not correctly organized and its very hard work to maintain the information for the same. If you want to find review which contains specific words, manually it will take a lot of time. It only makes the process more difficult and hard. This aim of the project is to automate the work performed like cleaning the review texts, making data readable for machine, dropping unnecessary columns, fetching words out of the data and creating desired columns, predict future ratings and sentiment analysis. Our project aims to provide a complete solution to the needs and Based on this information you can take decision regarding your business development. This project presents an approach for sentiment analysis to check sentiments of user reviews by classifying them as per their corresponding polarity along with a graphical visualization of the analysis of data. The implementation makes use of Python and its libraries for Natural Language Processing and Data Visualization. The project also contains a simple predictor tool for user reviews given as input at run-time. The usage of Python package TextBlob for sentiment analysis presents the necessary outcomes. The approach along with its implementation, evaluation, visualization and conclusion are the aspects of the project.

INTRODUCTION

Customer reviews and ratings on ecommerce platforms like Amazon have an influence on the product reputation and they are important for prospective buyers before they decide to make purchases. Text mining methods like Sentiment Analysis can be used to uncover customer opinions and understand the general customer sentiment about a product. Amazon dataset consists of ratings and reviews as feedback for its products. The ratings are numeric, ranging from 1 to 5, while the feedback is in the form of text. While ratings can be useful, looking at the overall performance of a product, it has been observed that the actual sentiment of the customers is better reflected in their comments. Another important factor that buyers might consider while buying products on Amazon would be price. Capturing the actual sentiment through comments can provide the company with an understanding of how particular categories are in demand, which products need to be kept or removed, etc. Using manual way to go through the feedback comments is certainly a tedious task, as the amazon musical instruments data itself contains five lakh plus reviews. A sentiment analyzer can be used in such a case, that can check the sentiment of the comments given by a customer. This project elaborates the implementation of text processing ,sentiment analysis of user reviews and future predictions of ratings.

Data Description

META DATA

Columns	Description
Description	description of the product
Title	name of the product
Tech1	the first technical detail table of the product
Tech2	the second technical detail table of the product
Brand	Brand Name
Feature	bullet-point format features of the product
Rank	Rank Information
Main_Cat	Category of Products
Date	Date of Order
Price	Price of Products
Asin	ID of the Product
ImageUrl	Url of Image uploaded

Table 1.1: List of columns and their description contained in metadata

Reviews Data

Columns	Description
Overall	rating of the product
Verified	Reviews True or False
ReviewTime	Time of review (raw)
ReviewerID	ID of the reviewer
Asin	ID of the product
Style	A dictionary of the product metadata
ReviewerName	Name of the reviewer
ReviewText	Text of the review
Summary	Summary of the review
UnixReviewTime	Time of the review (unix time)

Table 1.2: List of columns and their description contained in reviews data

Libraries

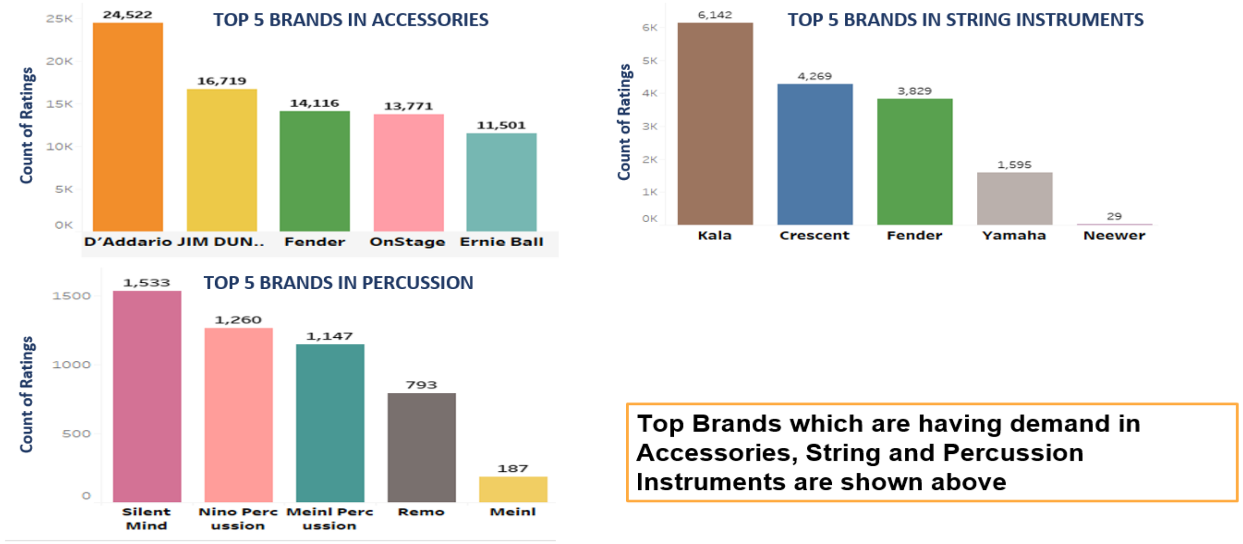
1. import pandas as pd
2. import numpy as np
3. import pandas as pd
4. import json
5. import re
6. import nltk
7. from textblob import TextBlob
8. from collections import Counter
9. from nltk.tokenize import sent_tokenize, word_tokenize
10. from sklearn.feature_extraction.text import TfidfVectorizer
11. from sklearn.preprocessing import LabelEncoder
12. from sklearn.preprocessing import MinMaxScaler
13. from sklearn.cluster import KMeans
14. import matplotlib.pyplot as plt
15. import seaborn as sns
16. from statsmodels.tsa.api import SimpleExpSmoothing
17. from wordcloud import WordCloud, STOPWORDS
18. from PIL import Image
19. from wordcloud import ImageColorGenerator
20. import warnings
    warnings.filterwarnings("ignore")

Data Preprocessing

1. Understanding the dataset and its relation is a key factor to start off with as without the same the work cant progress.
2. Merging the metadata and reviews data based on asin i.e product id is the start to the progress of the project.
3. Dropping unnecessary columns that don't add much value for the analysis is a must step else a chunk of unnecessary features will keep up showing.
4. ReviewTime was converted from object to a date time datatype for accurate results.
5. Data Visualization is the core base step for representing a dataset and understanding best parameters from the same which is simplified with the help of graphs.
6. Dealing with missing values is an important step which needs to be used correctly else biasness can generate in a dataset and lead to wrong information.
7. Analysis for multiple features tells the stories for top and the lowest numbers which depicts the best and the lows for a particular attribute.
8. Tableau was used for multiple visualizations and interactive dashboards were created to ease the analysis of the data and find multiple results with use of filters and actions over it.

Data Visualization

Musical Instruments Data



Top Brands which are having demand in Accessories, String and Percussion Instruments are shown above

Figure 2.1 Bar graph showing top brands based on reviews

Digital Music Data

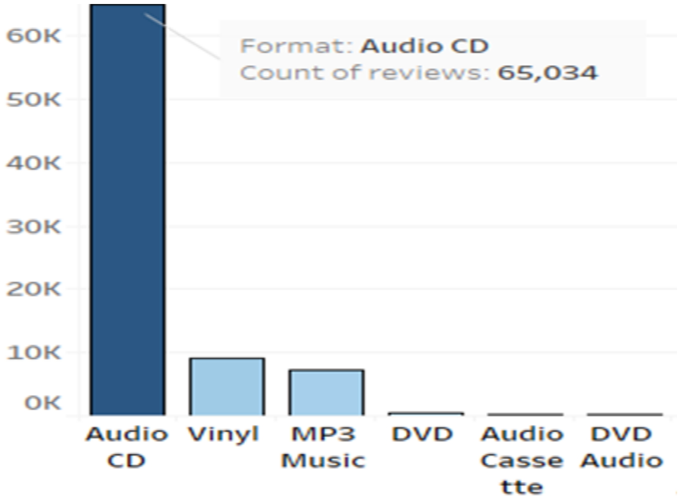
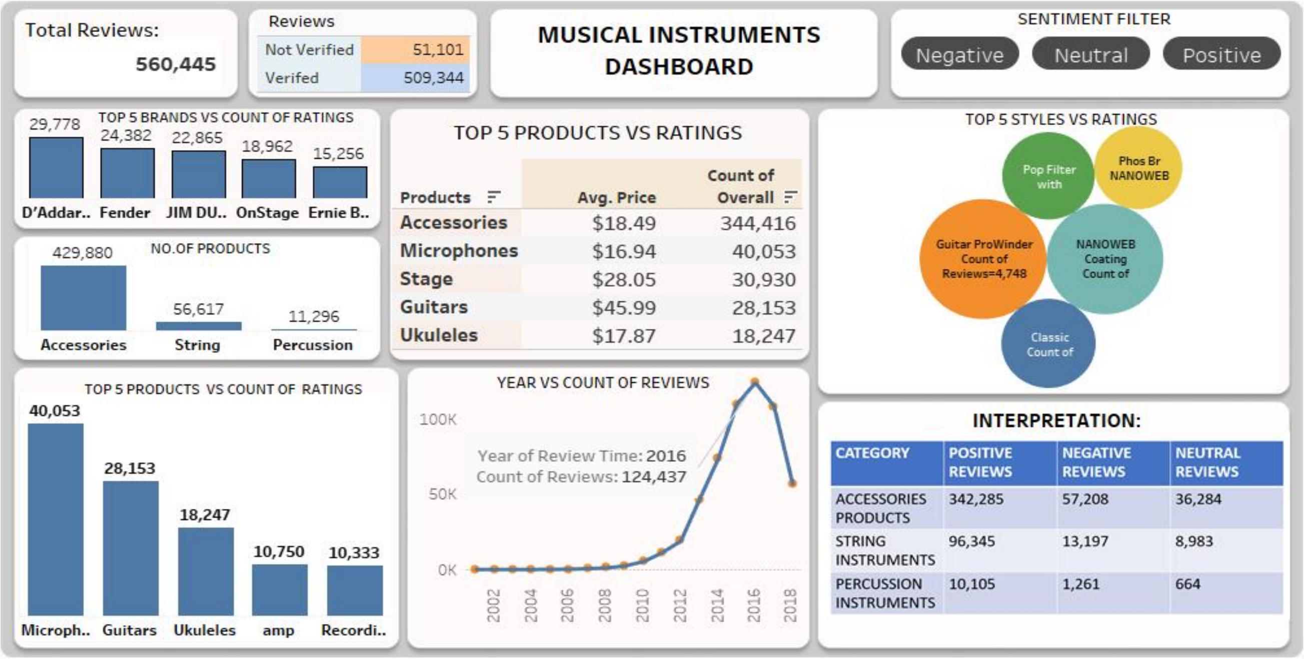


Figure 2.2 Bar graph showing top formats based on reviews

Tableau Dashboards



Reviews	
Not Verified	51,101
Verifed	509,344

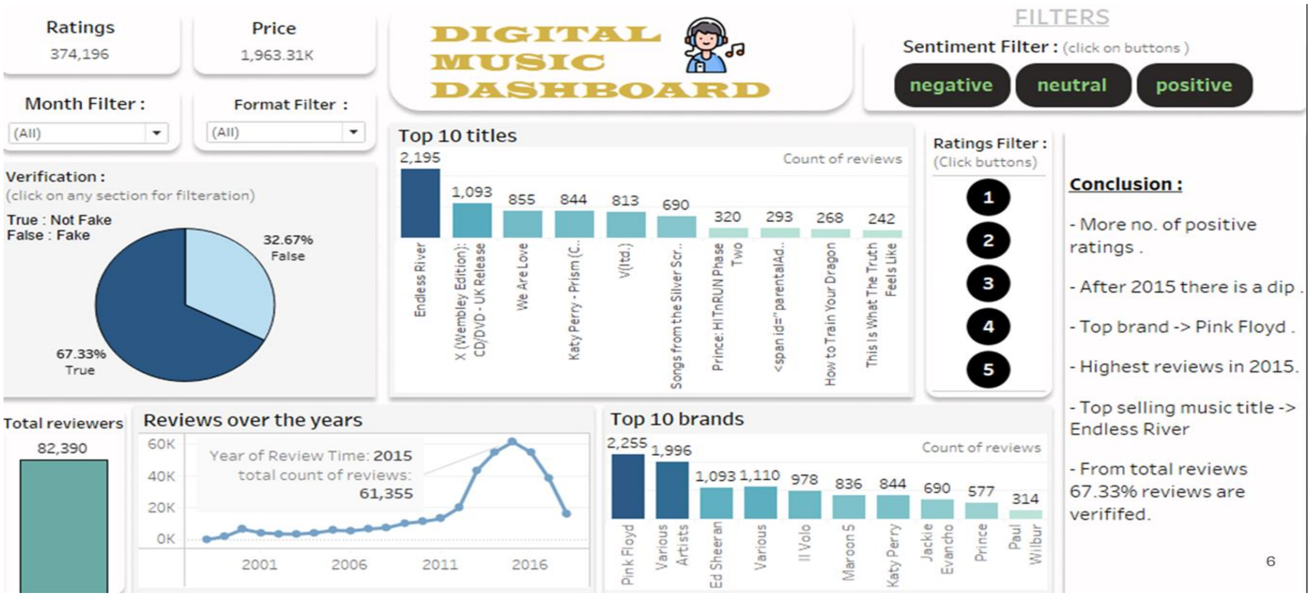
TOP 5 PRODUCTS VS RATINGS

Products	Avg. Price	Count of Overall
Accessories	$18.49	344,416
Microphones	$16.94	40,053
Stage	$28.05	30,930
Guitars	$45.99	28,153
Ukuleles	$17.87	18,247

INTERPRETATION:

CATEGORY	POSITIVE REVIEWS	NEGATIVE REVIEWS	NEUTRAL REVIEWS
ACCESSORIES PRODUCTS	342,285	57,208	36,284
STRING INSTRUMENTS	96,345	13,197	8,983
PERCUSSION INSTRUMENTS	10,105	1,261	664

Figure 3.1 Musical Instruments Dashboard

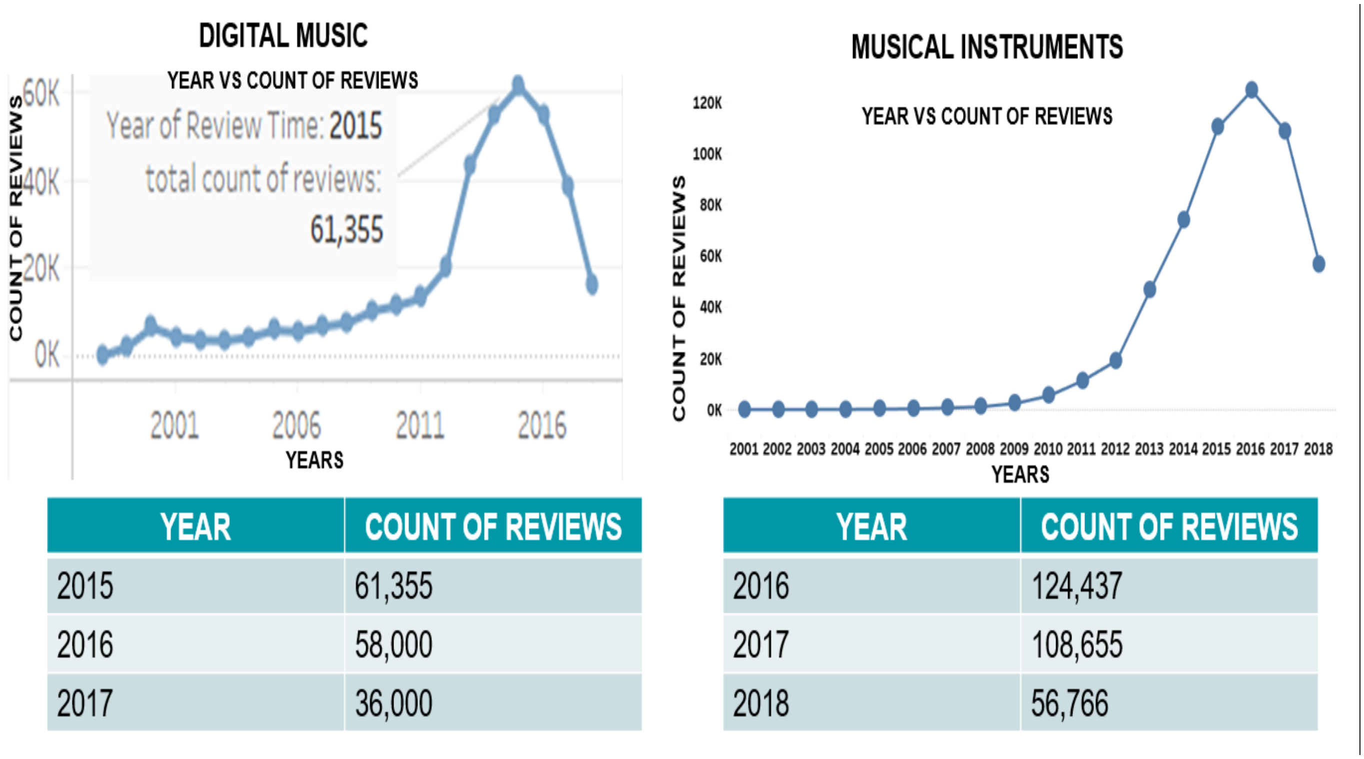


Conclusion :

- More no. of positive ratings .
- After 2015 there is a dip .
- Top brand -> Pink Floyd .
- Highest reviews in 2015.
- Top selling music title -> Endless River
- From total reviews 67.33% reviews are verififed.

Figure 3.2 Digital Music Dashboard

Similarity Between Digital Music and Musical Instruments



YEAR	COUNT OF REVIEWS
2015	61,355
2016	58,000
2017	36,000

YEAR	COUNT OF REVIEWS
2016	124,437
2017	108,655
2018	56,766

Table 3.3: Similarity between Digital Music and Musical Instruments Trends

REASON (Change in Technology):-

- Cd’s, Dvd’s , VHS Tapes are replaced by Online Streaming Platforms.
- Some customers are having interest in Musical Instruments Applications in Mobile Phones and laptops.

Text Preprocessing

Regular Expressions (**Regex**) is an essential tool for text analytics. It is powerful in searching and manipulating text strings. Compared to the traditional approach for processing strings with a combination of loops and conditionals, one line of regex can replace many lines of code. The re module offers a set of functions that allows us to search a string for a match:

Function	**Description**
findall	Returns a list containing all matches
search	Returns a match object if there is a match anywhere in the string
split	Returns a list where the string has been split at each match
sub	Replaces one or many matches with a string

Table 4.1: List of functions and their description

Meta Characters

Character	**Description**
[]	Set of characters
\	Signals a special sequence (can also be used to escape special characters)
.	Any character (except newline character)
^	Starts With

$	Ends With
\w	Matches Word Character
\d	Matches Digits
*	Zero or more occurrences
+	One or More occurences
{}	Exact specified number of occurences

Table 4.2: List of characters and their description

Combination of these characters are used to clean the text data to get the desired output results. For example-

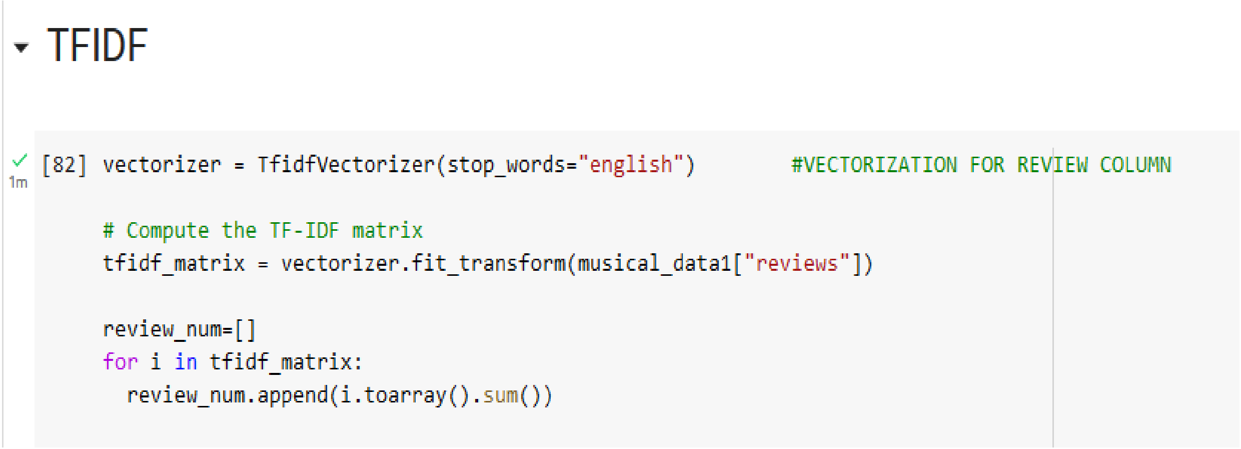
- Remove all non-letters and non-spaces except for hyphens and digits
    - text = re.sub("[^0-9A-Za-z\-]+", " ", text)
- Remove all numbers except those attached to a word
    - text = re.sub("(?<!\w)\d+", "", text)
- Remove all hyphens except between two words
    - text = re.sub("-(?!\w)|(?<!\w)-", "", text)
- Remove multiple spaces and lowercase everything
    - text = " ".join(text.split())
    - text.lower()

These expressions were used to clean raw text reviews to gain cleaned reviews without any symbols or numbers as an output and price column to remove $ symbols and text data present in it.

TF-IDF Vectorizer

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction. 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). A word that frequently appears in a document has more relevancy for that document, meaning that there is higher probability that the document is about or in relation to that specific word. A word that frequently appears in more documents may prevent us from finding the right document in a collection; the word is relevant either for all documents or for none. Either way, it will not help us filter out a single document or a small subset of documents from the whole set.

So then TF-IDF is a score which is applied to every word in every document in our dataset. And for every word, the TF-IDF value increases with every appearance of the word in a document, but is gradually decreased with every appearance in other documents.



```
vectorizer = TfidfVectorizer(stop_words="english")        #VECTORIZATION FOR REVIEW COLUMN

# Compute the TF-IDF matrix
tfidf_matrix = vectorizer.fit_transform(musical_data1["reviews"])

review_num=[]
for i in tfidf_matrix:
  review_num.append(i.toarray().sum())
```

Figure 4.3 TF-IDF Vectorization Code

Cluster Analysis

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. For the same first we convert categorical columns(sub category) into numerical using Label Encoder. Scaling is a method used to normalize the range of independent variables or features of data. Min-max scaling is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. Hence according to polarity (which represents sentiment), K-Means clustering is used to show the clusters based sentiments Positive, Neutral and Negative.

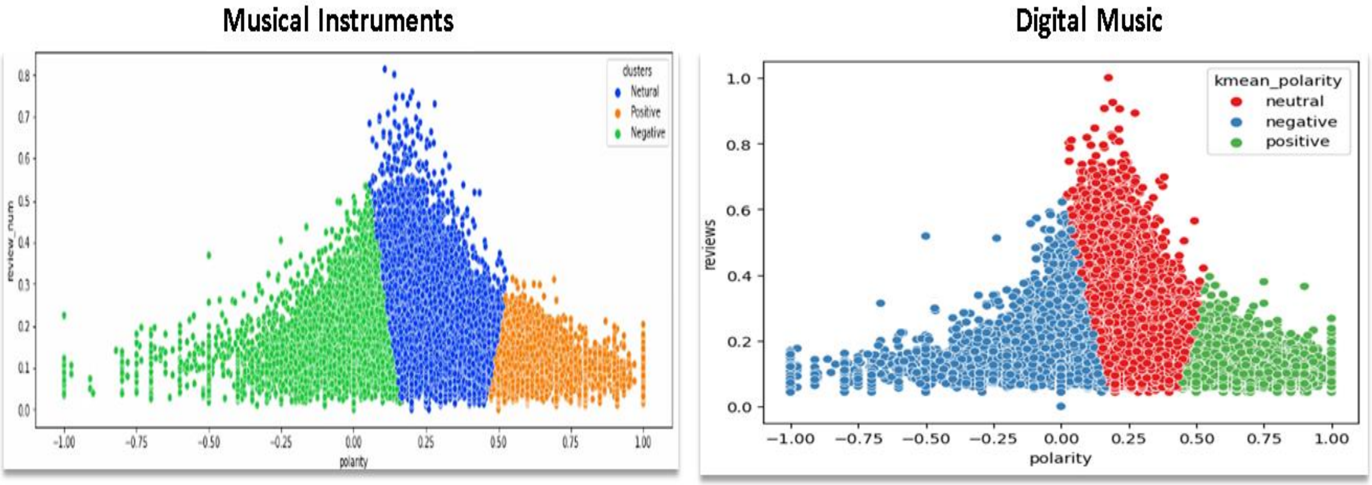


Figure 5 Clusters of reviews according to sentiments

Sentiment Analysis

Sentiment Analysis can help us decipher the mood and emotions of general public and gather insightful information regarding the context. It is a process of analyzing data and classifying the emotions of the text. TextBlob is a python library for Natural Language Processing (NLP) and actively used Natural Language ToolKit (NLTK) to achieve its tasks. It is a simple library which supports complex analysis and operations on textual data. It returns polarity and subjectivity of a sentence. Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment. **S**ubjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information**.** So, using this library , polarity scores were implemented for all the reviews. Determining the same, it was divided into 5 sentiment labels –

- Extreme Positive – polarity greater than 0.75
- Positive – polarity greater than 0 and less than 0.75
- Neutral – polarity equals to 0
- Negative – polarity lesser than 0 greater than -0.75
- Extreme Negative – polarity lesser than -0.75

After applying the same on both the data we found that most of the reviews are positive compared to negative reviews.

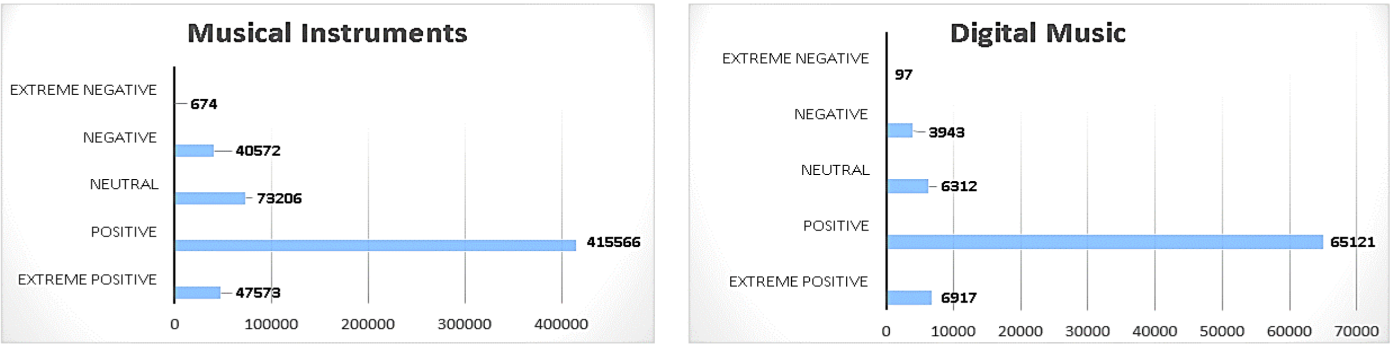


Figure.6: Bar graph showing reviews count according to sentiments

Word Cloud Analysis

Word Cloud was implemented as part of analysis on the reviews to get an idea about the words most frequently found in negative sentiment Reviews.

We have added some extra words to the stop words to remove the unnecessary words and to get most frequently used negative words.

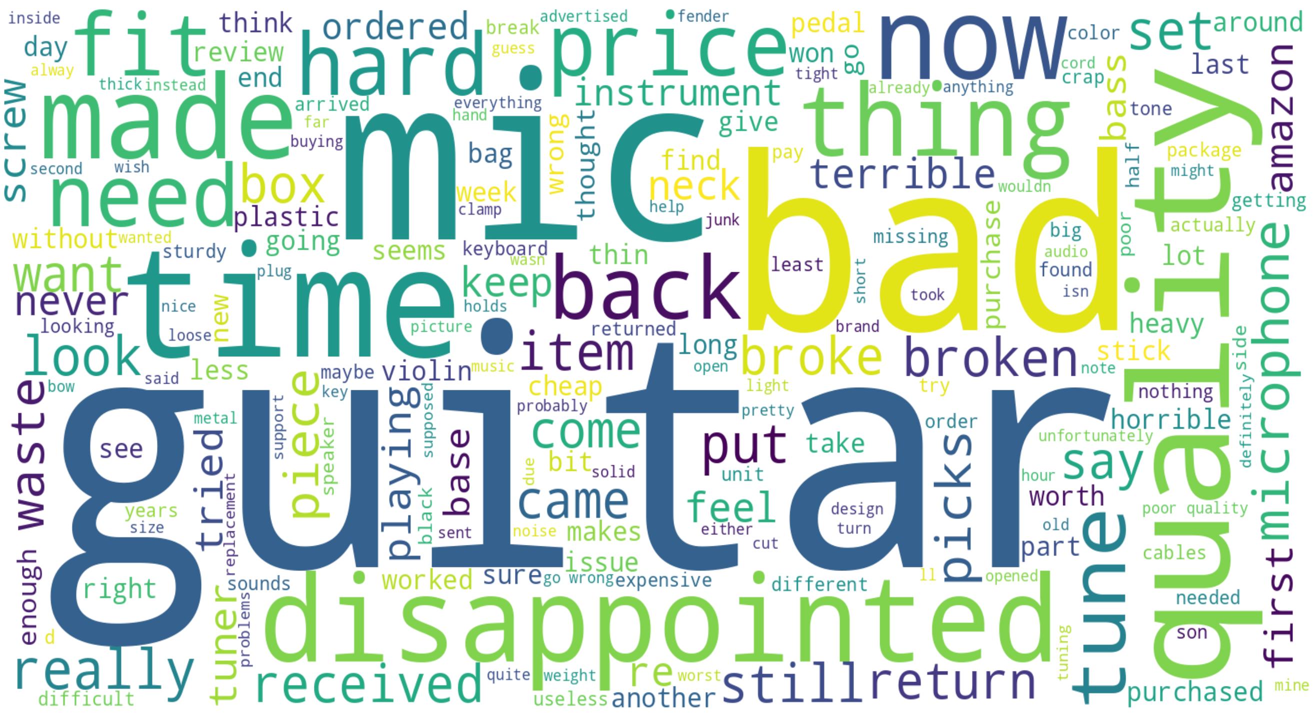


Figure 7: Word Cloud for Negative Sentiments

Time Series Analysis

Time series forecasting occurs when you make scientific predictions based on historical time stamped data. It involves building models through historical analysis and using them to make observations and drive future strategic decision-making. An important distinction in forecasting is that at the time of the work, the future outcome is completely unavailable and can only be estimated through careful analysis and forecasting models. Important factor to consider are –

Seasonality means that there are distinct periods of time when the data depicts a series of recurring patterns which shows that data has seasonality. **Trends** indicate whether a variable in the time series will increase or decrease in a given period. A trend can be upward or a downward trend. **Irregularities** can always occur, and we need to consider that when creating a prediction model. They present noise in historical data, and they are also not predictable. This is possible to check using Time Series Decomposition. A stationary time series has statistical properties or moments (e.g., mean and variance) that do not vary in time. Stationarity, then, is the status of a stationary time series. Conversely, non-stationarity is the status of a time series whose statistical properties are changing through time. The models used in the project are –

1) Arma (stationary data **)**
2) Arima (non-stationary data**)**

Using these both models on a resampled data (according to years) we could predict future ratings for our 3 subcategories String Instruments, Percussion Instruments and Accessories. The ratings varied from 4.2 to 4.7 for both data sets which proves the point that most of the reviews are positive when compared.

Time Series Forecasting Graphs-

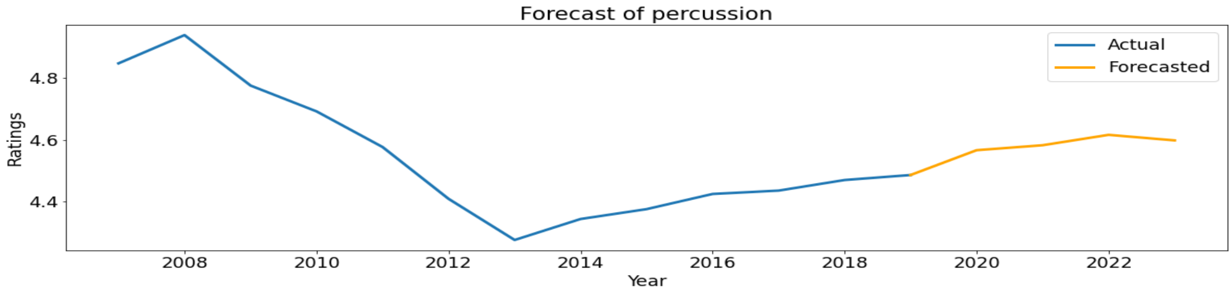


Figure 8.1 Time Series Forecasting for Percussion Instruments

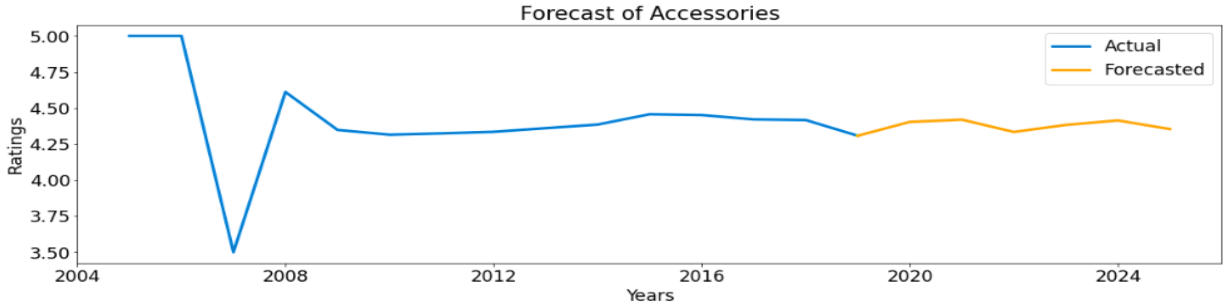


Figure 8.2 Time Series Forecasting for Accessories

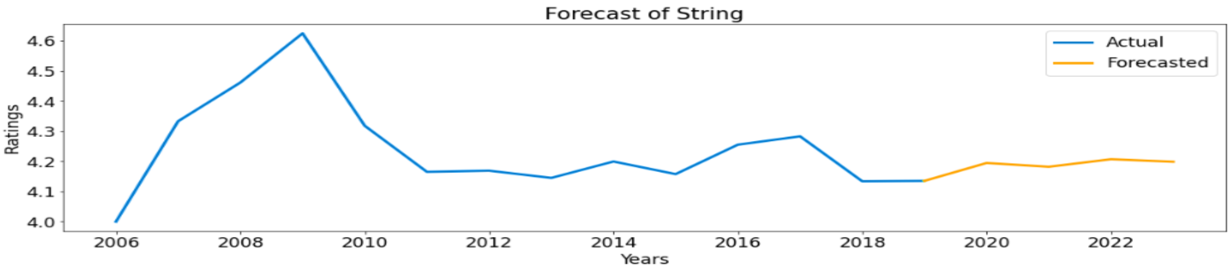


Figure 8.3 Time Series Forecasting for String Instruments

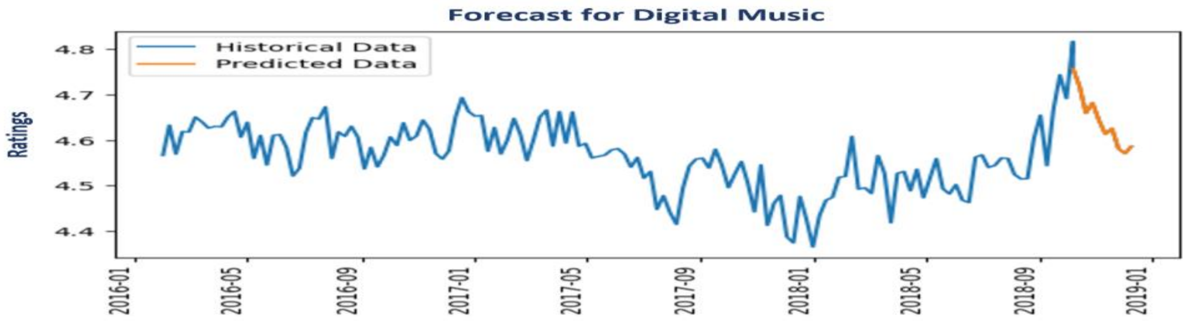
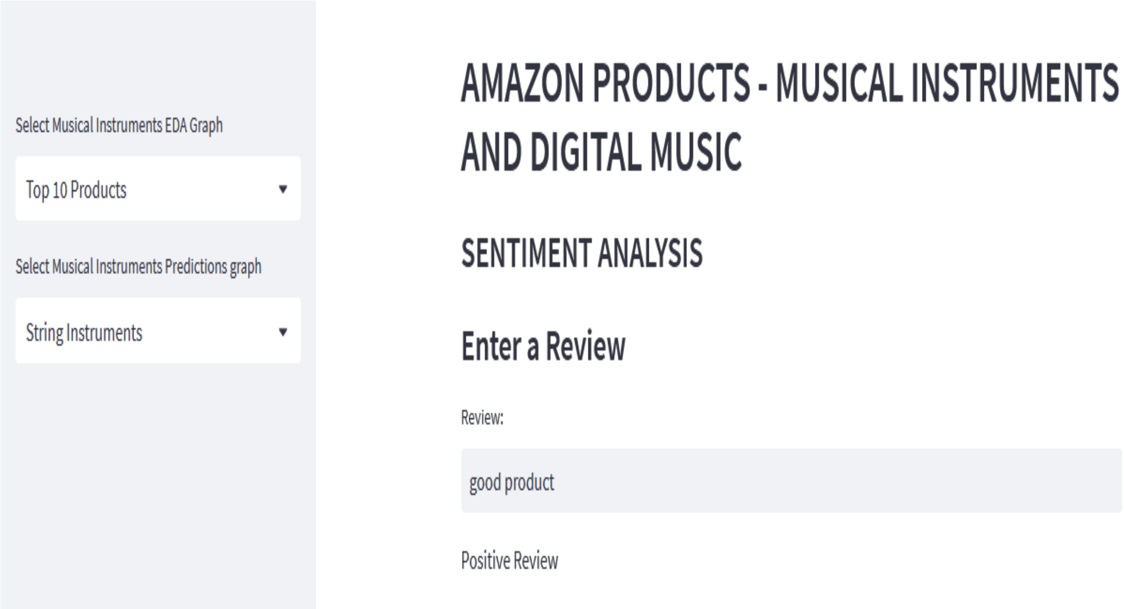


Figure 8.4 Time Series Forecasting for Digital Music

Streamlit

By using the streamlit library we have created frontend to our project which displays Data Analysis of top 10 products and brands for Musical Instruments and Digital Music and we have displayed Time Series Demand Prediction graphs for both Musical Instruments and Digital Music Category and we have Created a User Input Review Sentiment Analysis by using TextBlob.



Select Musical Instruments EDA Graph: Top 10 Products ▾

Select Musical Instruments Predictions graph: String Instruments ▾

AMAZON PRODUCTS - MUSICAL INSTRUMENTS AND DIGITAL MUSIC

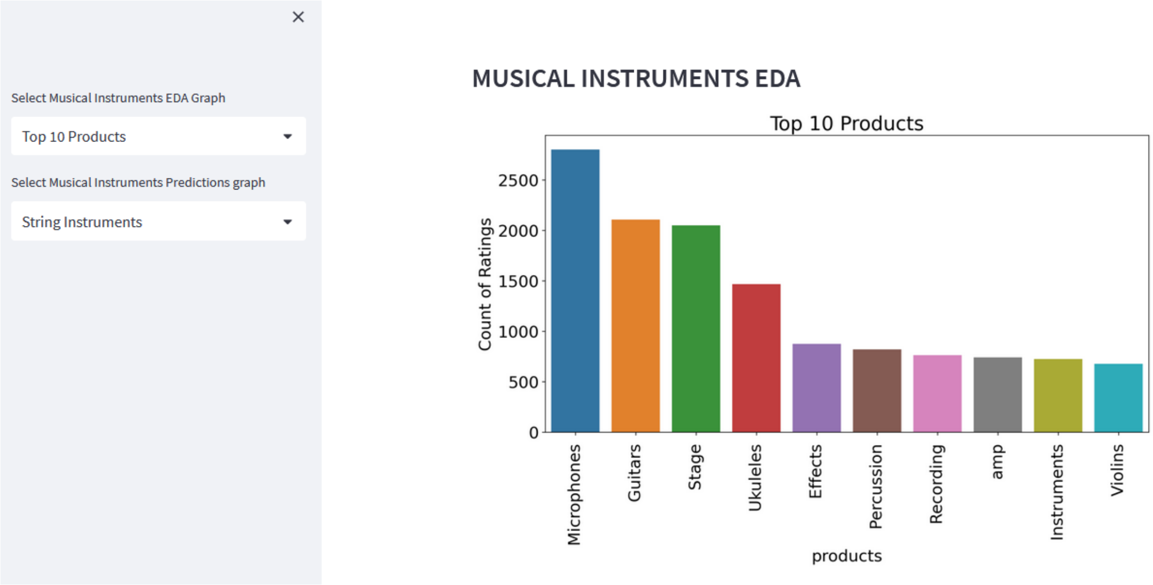
SENTIMENT ANALYSIS

Enter a Review

Review: good product

Positive Review

Figure 9.1 Streamlit – Sentiment Analysis



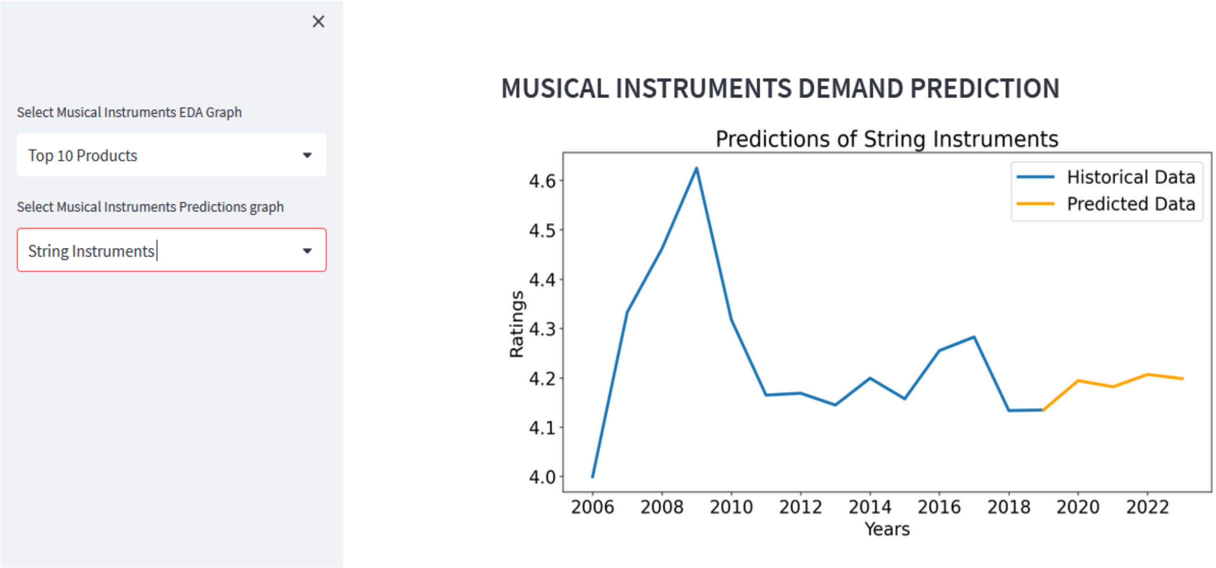
Select Musical Instruments EDA Graph

Top 10 Products

Select Musical Instruments Predictions graph

String Instruments

Figure 9.2 Streamlit – Exploratory Data Analysis(EDA)



Select Musical Instruments EDA Graph

Top 10 Products

Select Musical Instruments Predictions graph

String Instruments

Figure 9.3 Streamlit – Time Series Prediction

Business Conclusion

DIGITAL MUSIC - Inventory Optimization

FORMAT
➢ *Amazon Video* ➢ *Prime Video*

MUSICAL INSTRUMENTS – Inventory Optimization

CATEGORY	TOP BRANDS	TOP PRODUCTS
ACCESSORIES	➢ D'Addario ➢ Jim Dunlop ➢ Fender ➢ Onstage ➢ Gator, Yamaha	• Microphones • Recordings • Amplifiers • DJ Consoles • Speakers
PERCUSSION INSTRUMENTS	➢ Silent Mind ➢ Nino Percussion ➢ Menial Percussion ➢ Remo ➢ Gammon Percussion	• Drums
STRING INSTRUMENTS	➢ Kala ➢ Crescent ➢ Fender ➢ Yamaha ➢ Neweer	• Guitars • Ukleles • Violins • Violas • Cellos

- Keywords such as Terrible, Broken, Disappointed, Bad, etc… in reviews depicts that these are the reasons affecting the reviews.
- Most of the bad reviews are related to guitar. So more focus can be made on this product. As in string instruments category it is the most selling product.
- Audio Cd’s are comparatively expensive than online platform like Amazon video and Prime Video. Hence upcoming customers would prefer these platforms over others.
- In digital music, we can consider removing the formats like CD-R, MP3-CD, VHS-Tape having no reviews after a particular time period.

References

1. https://textblob.readthedocs.io/en/dev/
2. https://pypi.org/project/wordcloud/
3. https://streamlit.io/
4. https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews