**Name: Megha Bista**

**Project Name: Natural Language Processing (NLP) Analysis of Taylor Swift’s Song Lyrics**

**Introduction:**

This project will mainly analyze how the topics of Taylor Swift's songs have changed over time and whether they affect Album Sales using NLP. We will also use models such as Logistic Regression, Naive Bayes, Decision Tree, and Random Forest to analyze the results.

**Dataset:**

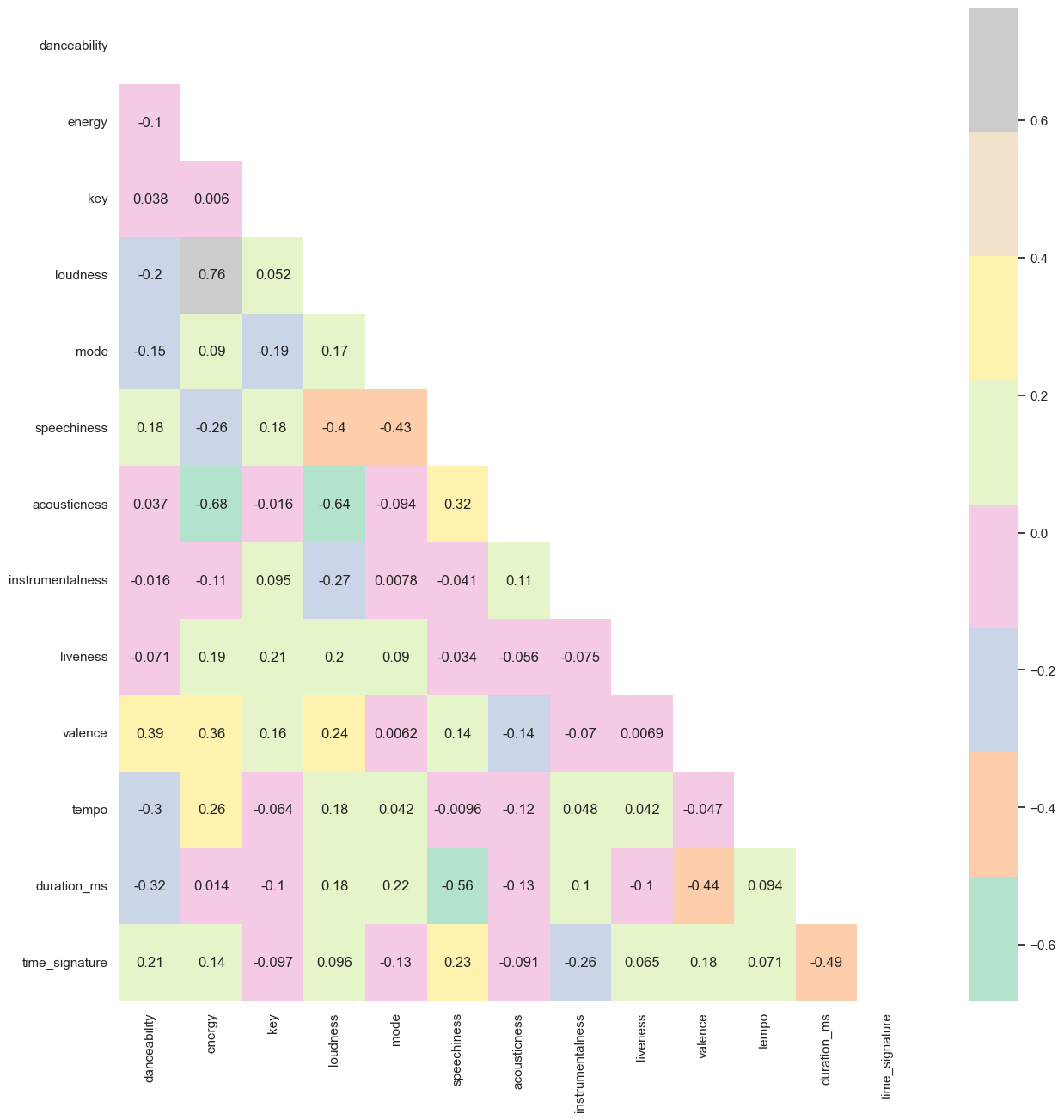
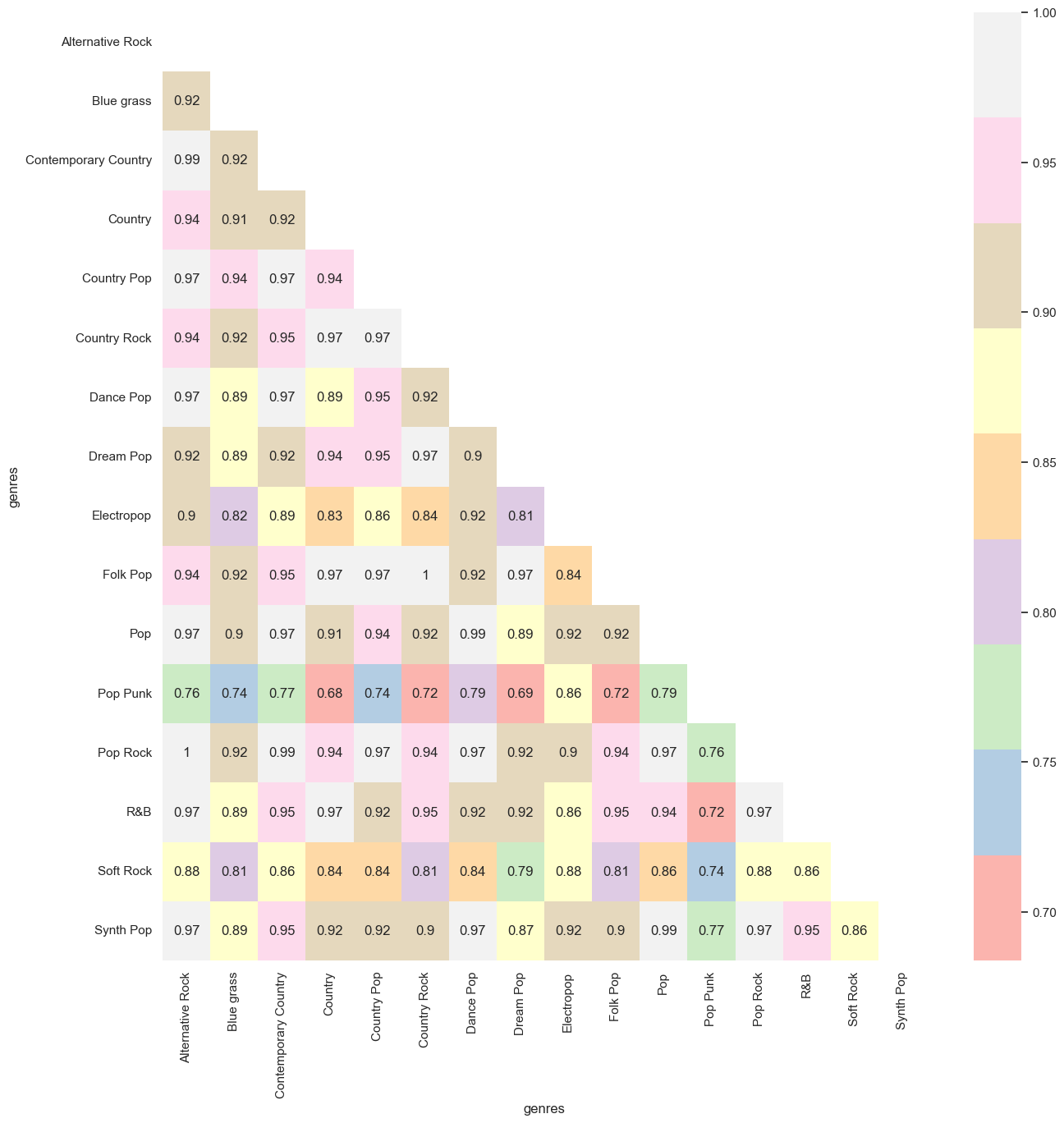
This dataset is available on Kaggle. The format of this data is a CSV file. It is about 218 KB in size. This dataset contains twenty columns which include track album, track artist, track title, and attributes such as danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valance, tempo, duration, time signature, track unique resource indicator (URI), track id, track lyrics, and genre.

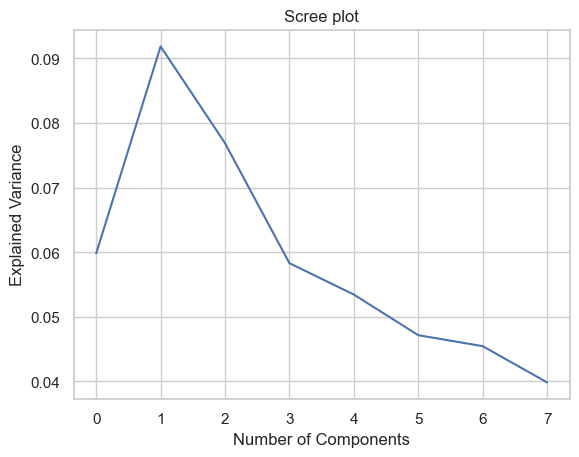
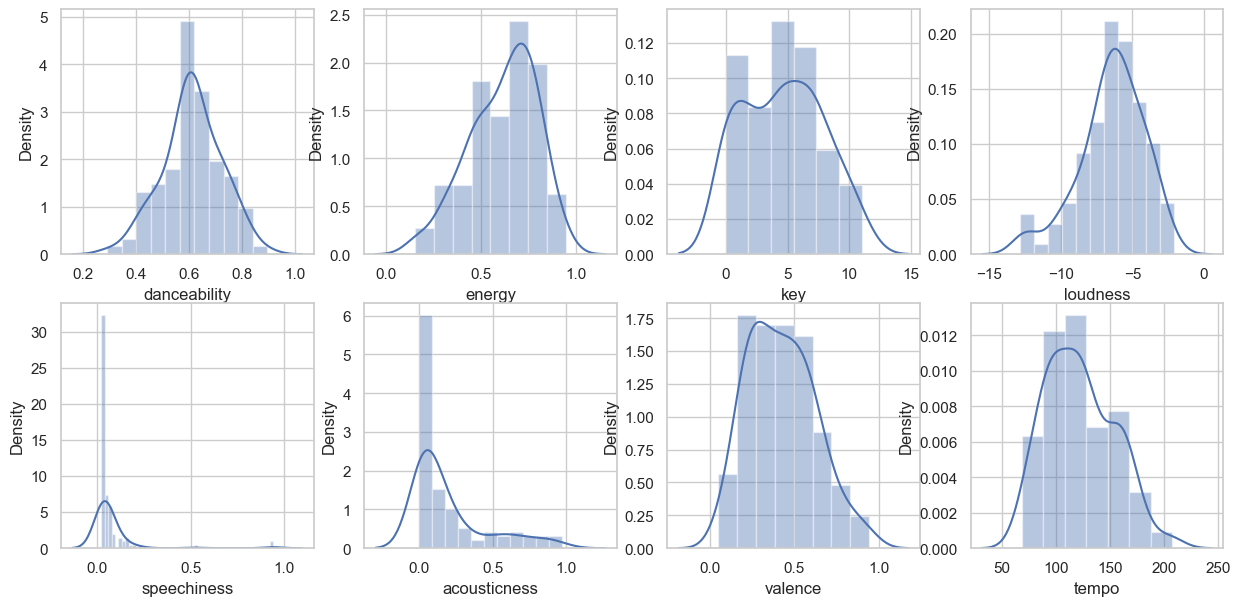
**Data Preparation:**

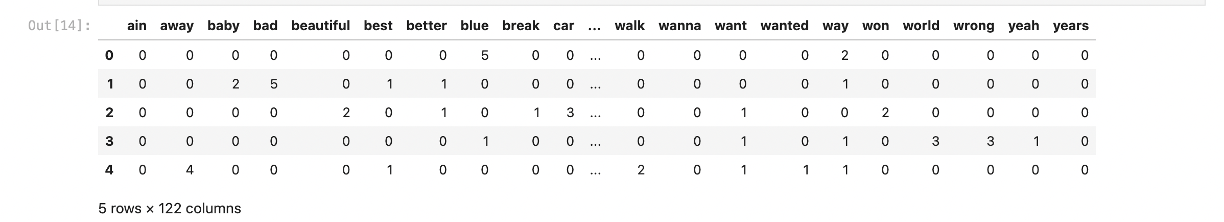
Data preparation was an important step to implement for this project since it is mainly based on Natural Language Processing (NLP). NLP is a common pre-processing technique that is used before text mining is performed. So for this project, I started the data preparation by adding two more columns to the data. Those two columns were “year\_released” and “world\_sales\_USD”. The values for those columns were acquired through Wikipedia. Adding these two columns was essential to answer our main question for this project. I also made sure to check if there were any null values. For the final data preparation process, I removed all punctuation and stop words from the track album lyrics column to prepare for NLP.

**Exploratory Data Analysis (EDA):**

Performing Exploratory Data Analysis (EDA) helped me to get a better understanding of this data since the data consists of twenty-two columns. Some of the techniques used in exploratory data analysis include visualizing the data with graphs and charts, correlation feature matrices, document-term matrices with unigram and bi-grams, and analyzing the data distribution. By exploring the data, I identified patterns, spotted outliers, and gained insights into the data such as Loudness and energy being the most correlated between features. Using the distribution plot for all audio features, I found that danceability had the highest distribution among all the features. The document-term matrix gave me an insight into the frequency of the occurrence of each term in the data.

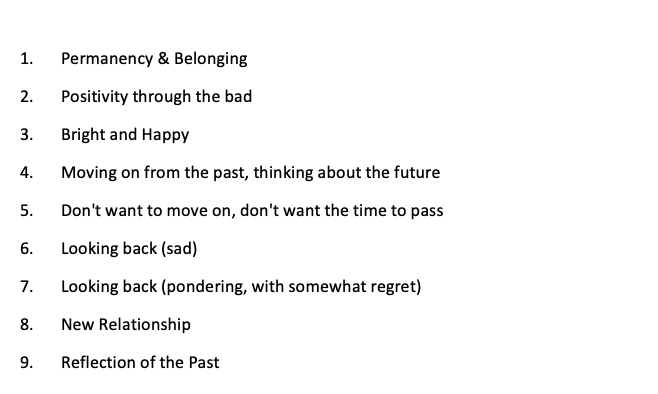
 

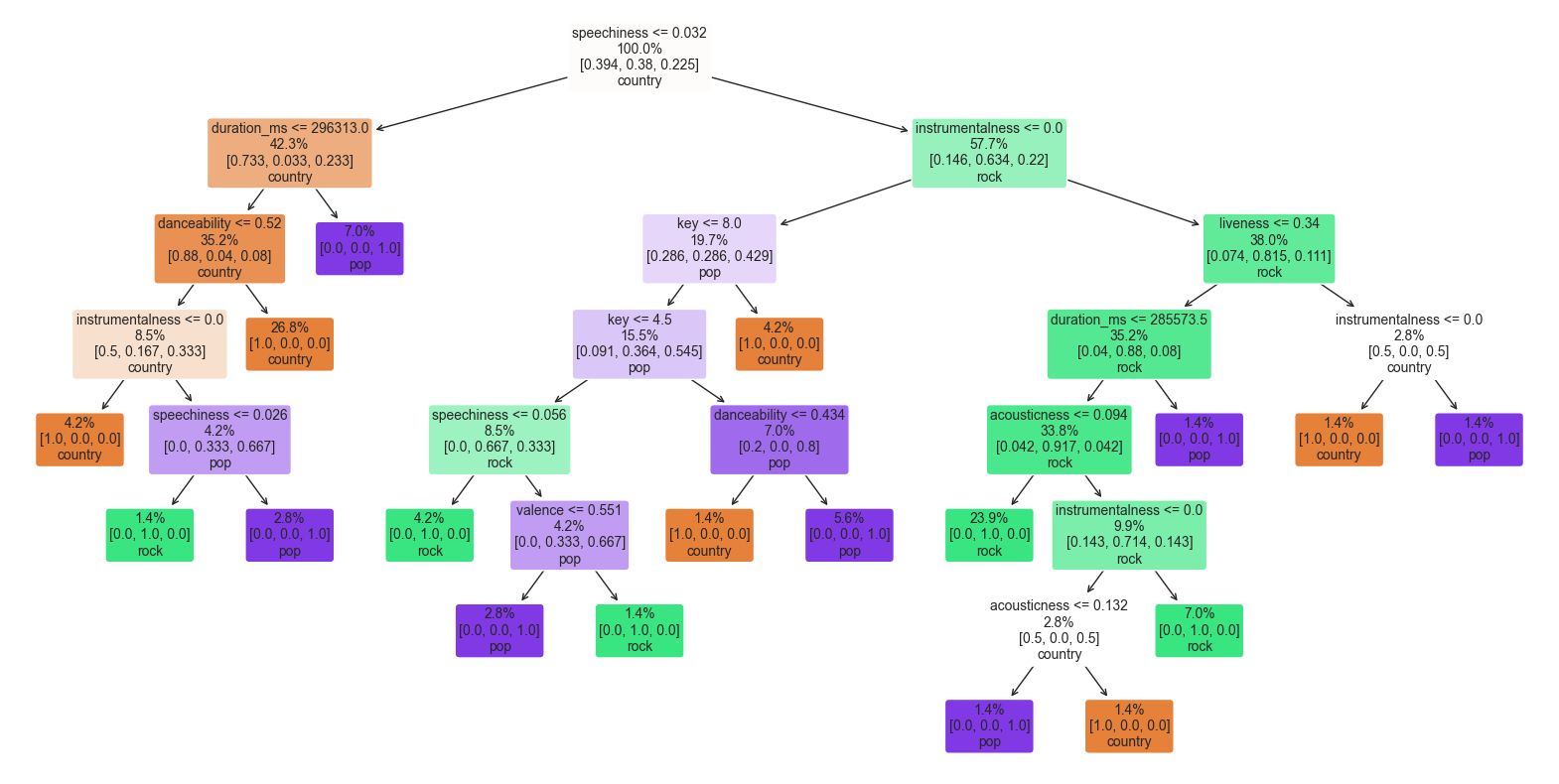
 

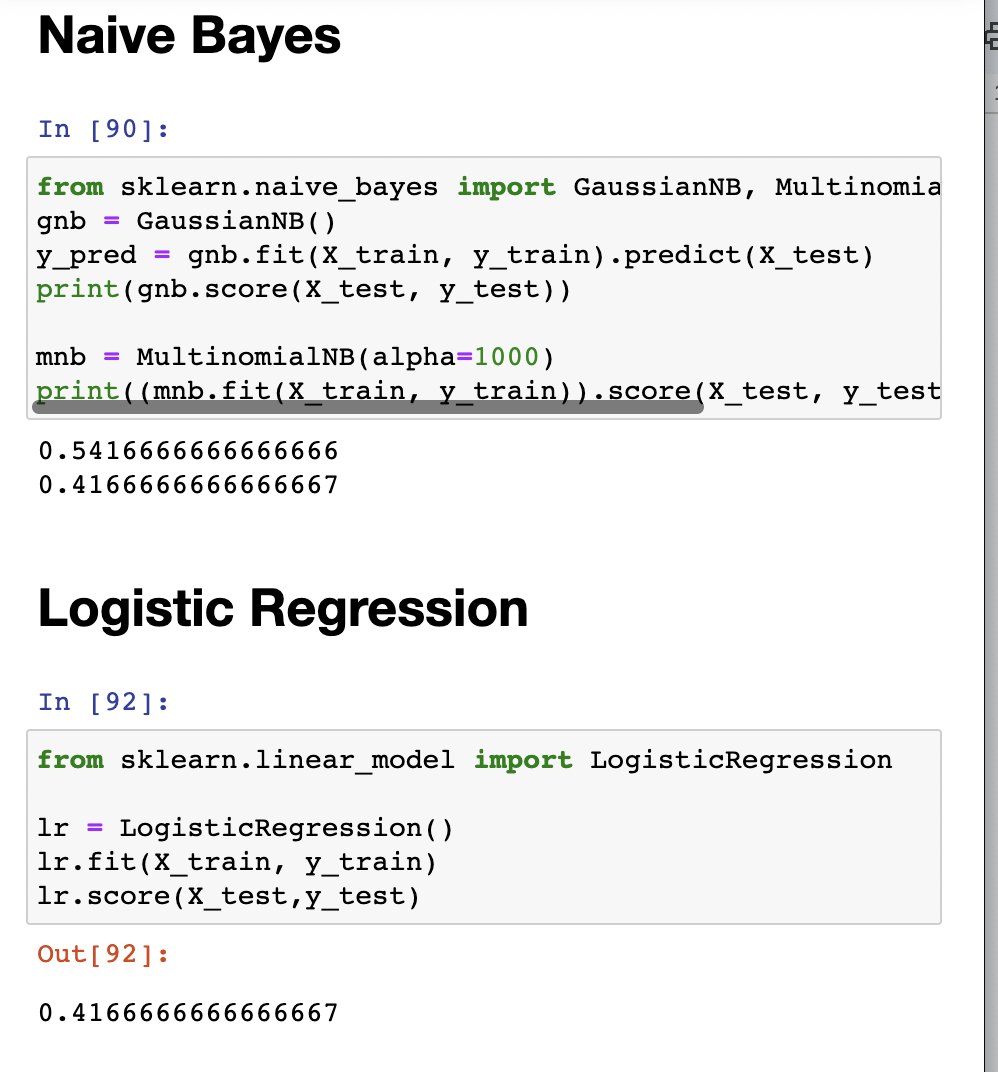
**Model Selection and Training:**

To select the most suitable model to solve this problem, I had to do several analyses beforehand and compare them. The first analysis I did was the Latent Semantic Analysis (LSA). LSA is a popular dimensionality-reduction technique that uses a matrix factorization technique called Singular Value Decomposition to identify the underlying relationships between words. The second analysis I did was the Non-Negative Matrix Factorization (NMF). NMF extracts words and creates a topic from the data by identifying themes and concepts that are common to a set of data. I also used the TF-IDF vectorizer. The TF-IDF vectorizer converts text into numerical vectors. It assigns a high value to terms that occur frequently in the data and assigns a low value to terms that occur frequently across the corpus. Then, I combined the LSA model with TF-IDF and NMF model with TF-IDF to compare which model gives the best output for K-means clustering. After several trials, I concluded that NWF with TF-IDF vectorizer and 9,10, and 11 topics produced the best topics. Then, I performed K-means Clustering on the NVM TF-IDF model and got 5 clusters. Those five clusters are about the Topics shown below. Based on the clusters, Cluster 0 was mostly about Topic 3, cluster 1 was mostly about Topic 5, cluster 2 was mostly about Topics 2 and 4, cluster 3 was mostly about Topics 6, 8, 7, and 0 and finally, cluster 4 was mostly about Topic 1, Some Topic 5 and 2.



I also performed a mean worldwide Sales per Cluster to answer the final question. Based on the values, I concluded that Cluster 3 has been a steady theme throughout Taylor Swift's career, while Clusters 0 and 1 have fluctuated; Clusters 2 and 4 have stayed relatively the same. Cluster 1 (looking back on the past, sad) was by far the most successful (in terms of Album Sales) and Cluster 0 (moving on from the past, thinking about the future) was the least successful. After answering the final question, I went ahead and divided the data into tests and train to evaluate my model. I performed decision trees, random forest, naive Bayes, and logistic regression. After evaluating all the models random forest gave me the highest value of 0.708 and the lowest value was obtained through logistic regression.



**Future Direction and Reflection:**

Based on the results acquired from the models above, I would like to use more complex models in the future. I think using complex models on this data will give better results and there won’t be any limitations. I would also like to continue working on this data by taking a different approach such as creating a recommendation system. I would also like to work with other datasets using the algorithms used on this dataset.