Predicting Popularity of YouTube Videos - Project Report

Report/Analysis in Slides Form: Predicting Popularity of YouTube Videos -Presentation
Code: Predicting Popularity of YouTube Videos - Code
Data Set: Trending YouTube Video Statistics

Problem Statement

The purpose of this project is two folds:

1. Find the most viewed genre of YouTube videos in an economically challenged country like Mexico and an economically stable country like the USA? By focusing on this data YouTube can advertise to the audience of these countries with the videos that can potentially get them more followers/subscribers and in turn more ad revenue
2. Predict popularity class of YouTube videos and use the information to push popular content to new markets and engage more audiences.

Data Wrangling:

The data from Kaggle was mostly clean and ready to use. Two dataframes, one for each country was created and merged together after wrangling on each was performed individually. The data type of the date-time column was changed into datetime object. The category of the videos was represented by a numerical id therefore using a json file downloaded from YouTube_API, category names were mapped to category ids of all the videos. One of the category id in Mexico's data frame was not present in the json file. Therefore those 29 videos were deleted from the data frame. All the duplicates and null valued rows were dropped. Lastly a country column was added to the individual data frames and merged together to form a combined data frame.

Exploratory Analysis

The YouTube data contains video trends from 2017 to 2018. For exploratory analysis several measurements were collected and data analysis was performed between the two years. In order to assess how the viewership changed from year to year for the USA and Mexico several graphs were plotted. Some of the questions explored for each country are:

1. How many unique videos were in each category and how many average views each one received?
2. For each category how were the average views different from 2017 and 2018?
3. Which categories showed prominent changes in 2018 compared to 2017?
4. Were there any correlations between video's views, likes, dislikes and comment count?

Figures 1 - 4 show the bar plots between unique videos and average views per category for the USA and Mexico in 2017 and 2018. Figures 5 and 6 show the difference in average views from

2017 and 2018 and percentage change in the average views in each category for both countries.

Analyzing the Graphs:

The **unique videos versus average views** in each category for Mexico and USA in 2017 and 2018 are shown in Figure 1 through Figure 4. First thing that jumps out from these figures is that the **Mexican audience prefers new content (new unique videos) than American Audiences**. For example in 2017 for Mexico, there are over 2000 unique videos in Entertainment however for the USA there are only 455 unique videos(Figure 1 and Figure 3). Even though there are more unique videos **in Mexico, the average number of views is less than the USA.** For example in the Entertainment category the number of average views is only 250,000.For the USA in the same category there are 1.3M views.

Music videos incur the most average views every year in both countries. The most viewed Music category at 7M average views is in 2018 in the USA. Another observation from this data is that the Pets & Animal category seems to be second to the Music category in terms of average views in Mexico for 2017 and 2018. For the USA the second most viewed category is Film & Animation in both 2017 and 2018.

The videos in the **Shows category dominated over other categories in the USA and brought in 553% more average views in 2018**. In contrast, **views from Mexico for the Shows category dropped 88% in 2018**. The most average views gained in Mexico were in the Science & Technology category. **Mexico viewers watched more of the Science & Technology category videos that gained 240% more average views in 2018.** Other categories that lost average views in Mexico in 2018 were the Gaming and the Comedy categories. Interestingly none of the categories dropped below zero in the number of average views in the USA. However large the change, it was still positive.

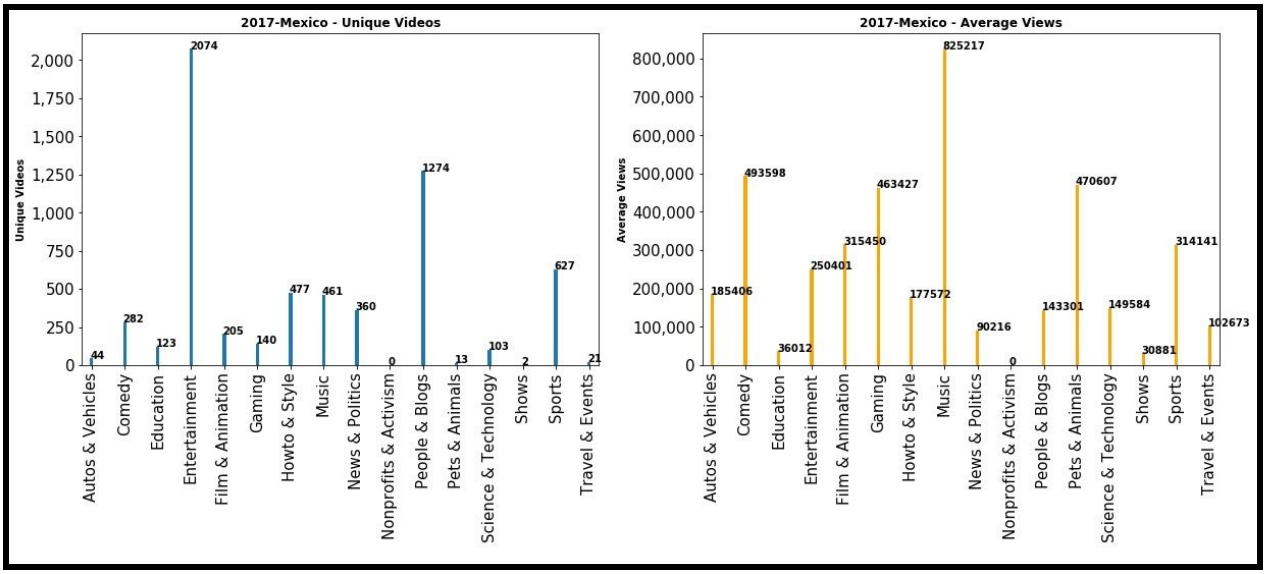

Figure 1: Mexico - Unique Videos vs Average Views, 2017

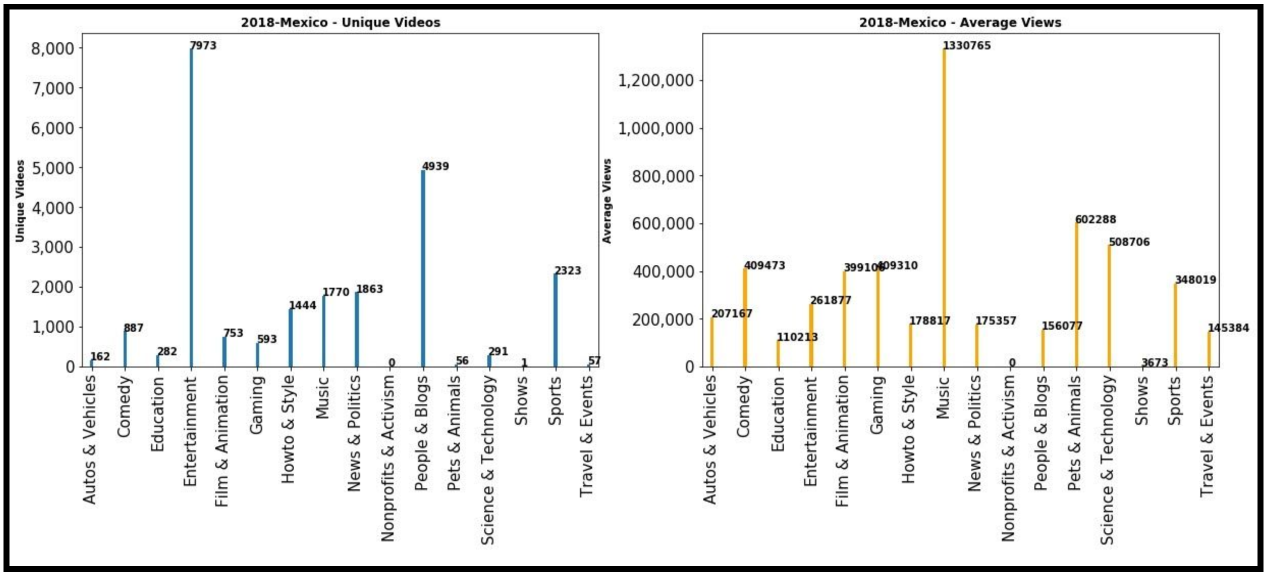

Figure 2: Mexico - Unique Videos vs. Average Views, 2018

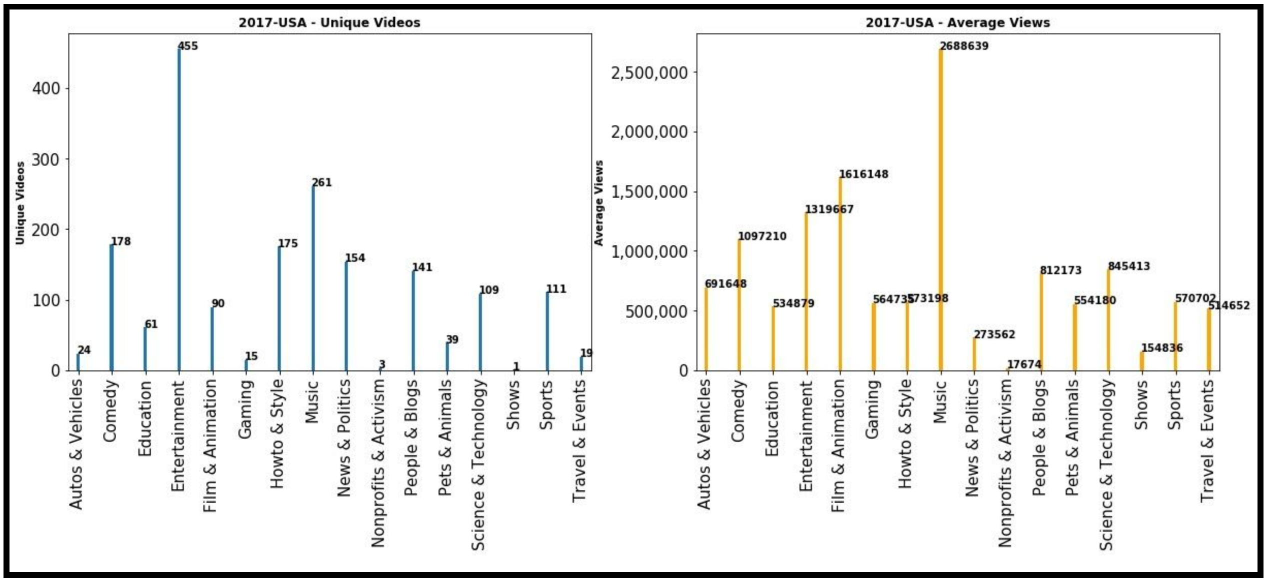

Figure 3: USA - Unique Videos vs. Average Views, 2017

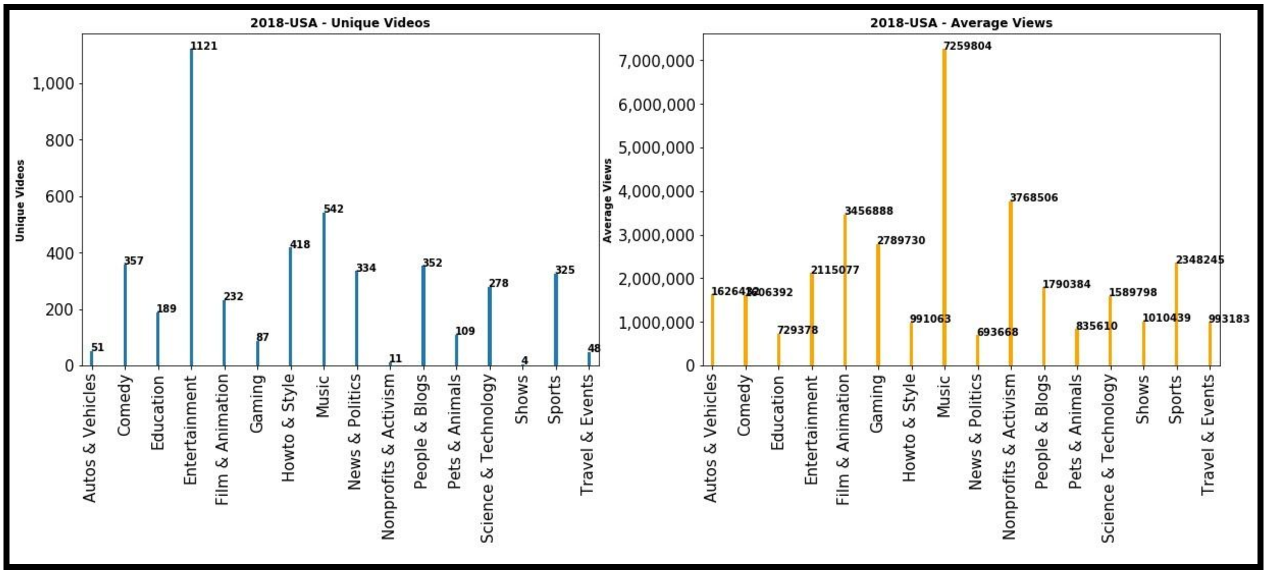

Figure 4: USA - Unique Videos vs. Average Views, 2018

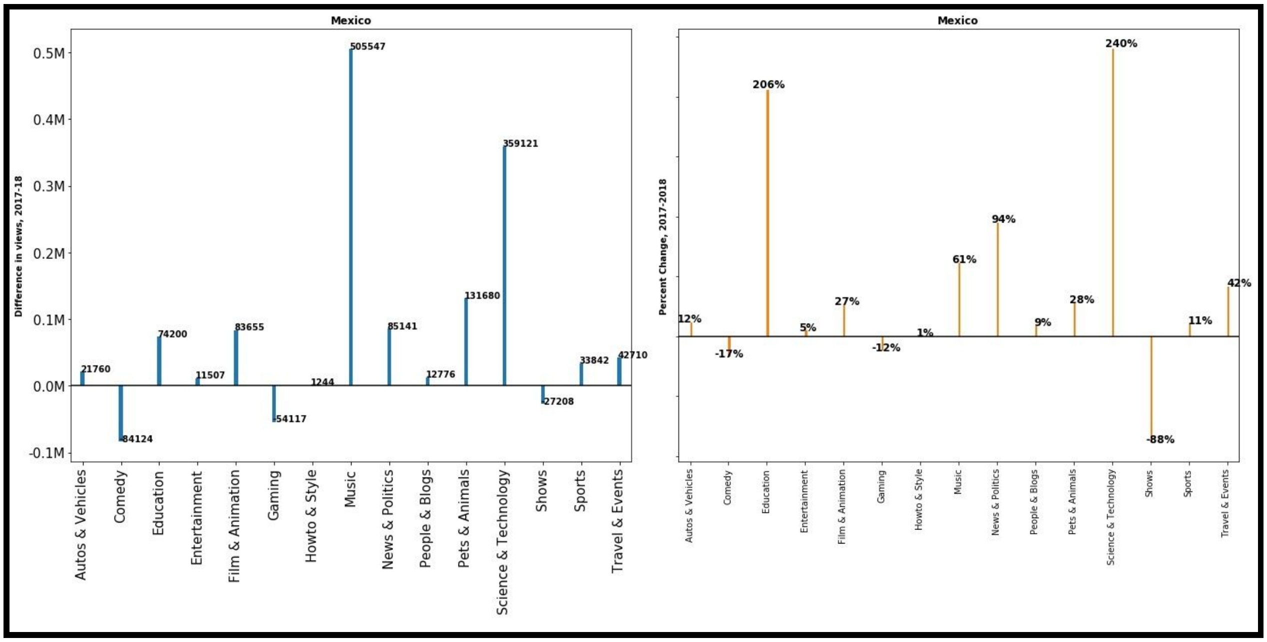

Figure 5: Mexico - Difference in views from 2017 to 2018

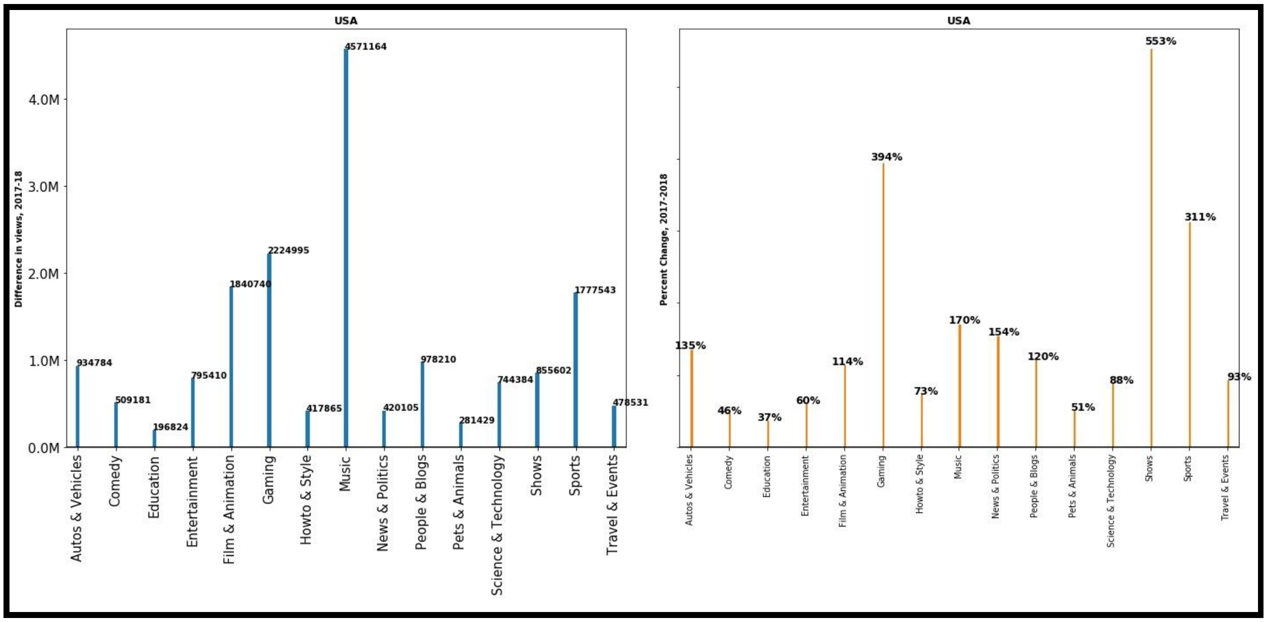

Figure 6: USA - Difference in views from 2017 to 2018

Statistical Analysis

For statistical analysis three hypothetical scenarios were considered.

Scenario 1: For YouTube to consider investing properly in people, infrastructure etc, they need to know the audience engagement with the videos in both countries. For this they need to know the views distribution and a 95% confidence interval around the population mean.

Scenario 2: Does the number of videos in each category depend upon the country it is viewed in? or country has no effect on the video categories and the number of videos in each category?

Scenario 3: How does the expected value of the video views change as time progresses?YouTube could be interested in this data to see how and when to place advertisements and attract traffic to click on a new ad. More interestingly how does the expected value of the difference in the views progress?

Scenario 1:

Using the clean data that was produced during data wrangling, populations' statistics were inferred. With 95% confidence Level, population's mean was determined using bootstrap inference method.

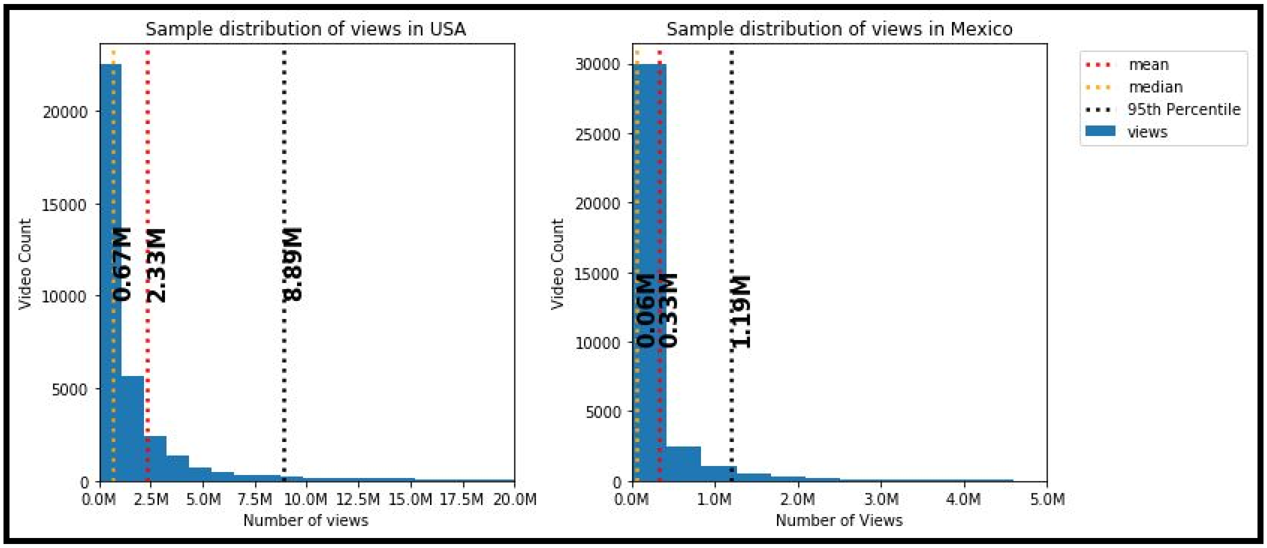

Figure 7: Sample Statistics - USA/Mexico

SUMMARY OF STATISTICS:

USA SAMPLE STATISTICS - VIEWS

- 95% of the views are under: 8.89M
- Other statistics of the views in the USA:
    - Sample Mean= 2.33M
    - Sample Standard Error = 7.26M
    - Sample Median= 0.67M

The PEAK of the distribution is closer to it's median 0.67M, rather than mean 2.33M
This is expected behavior from a left skewed distribution

Mexico SAMPLE STATISTICS - VIEWS

- 95% of the views are under: 1.19M
- Other statistics of the views in the Mexico:
    - Sample Mean= 0.33M
    - Sample Standard Error = 1.45M
    - Sample Median= 0.06M

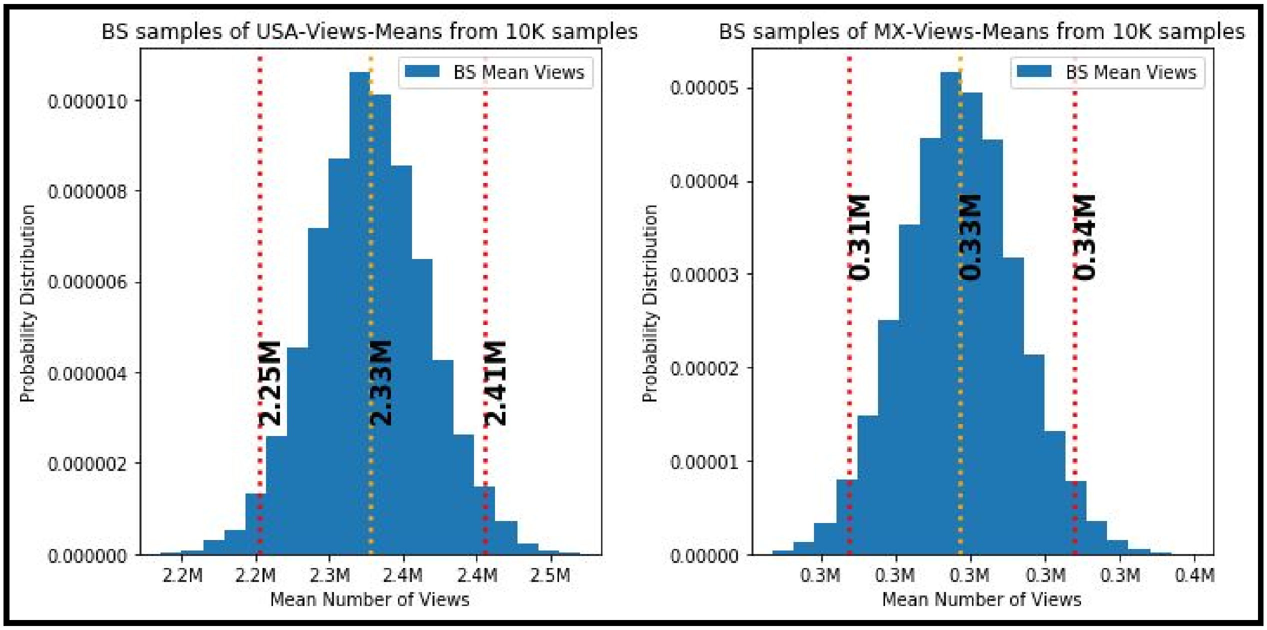

Figure 8: Inference of population mean with 95% CI

USA:Confidence interval (95%) for population mean is [2.25M ,2.41M]
Mexico:Confidence interval (95%) for population mean is [0.31M ,0.34M]

Scenario 2:

To determine if the number of videos in each category depend upon the country it is viewed in, Pearson chi-squared statistical hypothesis test was used. The analysis was performed using chi2_contingency and chi2 scipy packages. The null hypothesis set to be true was that the country and category variables are independent. Alternative was that the two variables are dependent on each other.

Results:
Probability=0.95, Critical=25.00, Stat=7532.27
Stat is greater than critical value therefore variables are Dependent (reject H0)
Significance=0.05, p=0.00
P-value is less than alpha (significance level) therefore variables are Dependent (reject H0)

Analysis of Scenario2

Since results show that stat returned from the chi2-test is greater than the critical value, the two variables are dependent. Also the p-value returned from the chi2- test is less than alpha showing that the two variables are dependent.

Scenario 3:

Expected Number of Views: To find the expected number of views of a video, changing over different numbers of periods, Bayseian inference method was used.

The video chosen for this analysis is titled “Childish Gambino - This is America”. It is the most viewed video in the USA’s YouTube Data. It was introduced on May 8th 2018 and collected views in abundance till it reached 225M on June 2nd 2018.

A bar plot is drawn in Figure 9 to visualize the increment in number of views as the days progress. The data for this specific video was collected for 25 days. It can be observed that the mean of the views for this video is 150M. The rate of change increases as the days increase

The Bayesian model showed a slightly different picture of the same data. Figure 10 shows that the change in views happened between day 5 and 6. Expected value of views went from 75M to 180M views

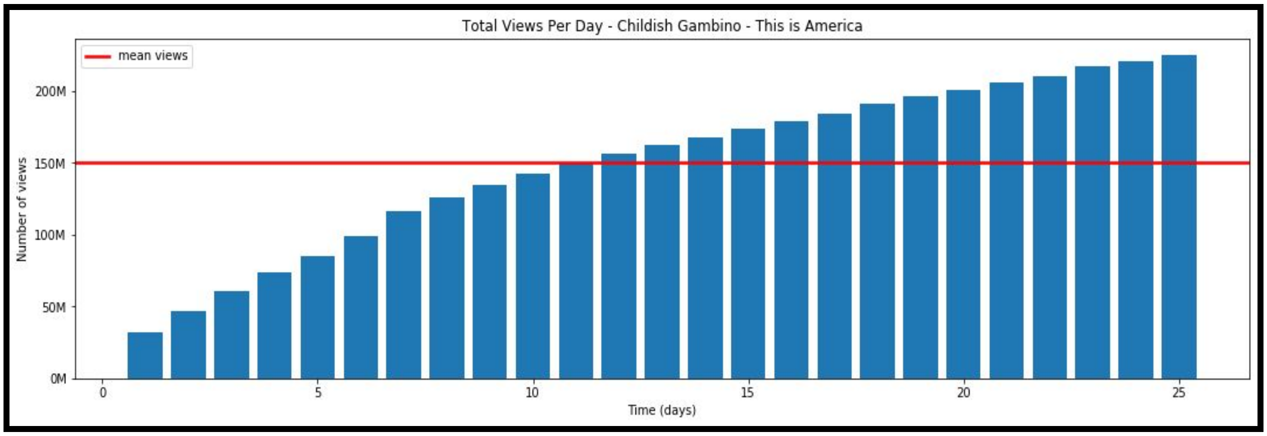

Figure 9: Views per Day - Childish Gambino - This is America

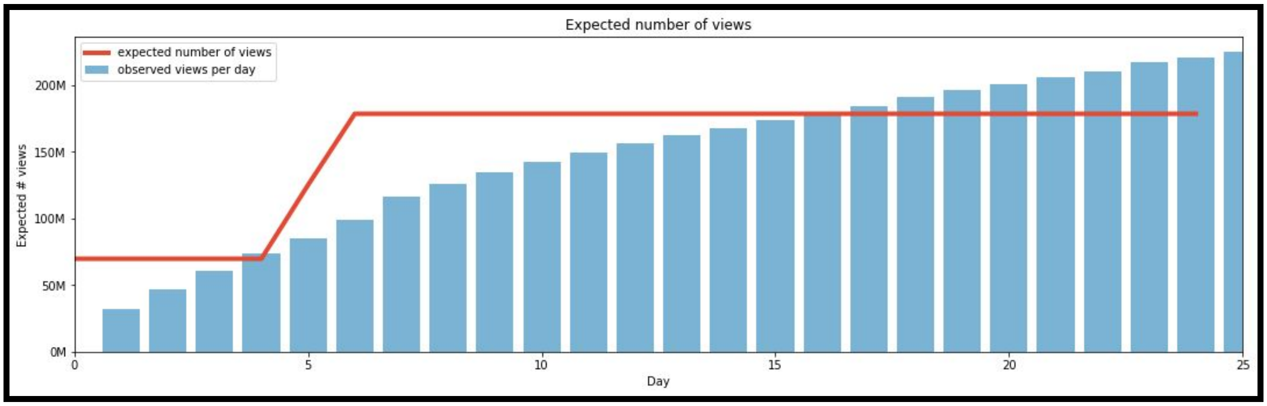

Figure 10: Expected number of views per day using Bayesian Inference - Childish Gambino

<u>Expected Number of the Difference in the Views</u>: For the same video, a bar plot of the difference in views between the consecutive days is drawn in Figure 11. Using a similar Bayesian inference model, the expected number of differences in number of views in three periods was calculated and plotted in Figure 12.

It can be observed from Figure 11, that the most difference in views occurred on day 5 which was predicted by our model previously i.e it chose day 5/6 accurately to predict the period where change happened. It can be observed that there are more periods where the change happened. The simplicity of the previous model made it predict only one such period. The mean of the difference in views is 8.2M

The model predicted that there are three dips in the expected values of the differences in views. Around day5, day 10 and day 21. The model representation of the difference in expected values of the differences in views is shown in Figure 12. As any popular video eventually starts declining in viewership as time passes, this model gives a glimpse of when the most views can be achieved and when the decline starts.

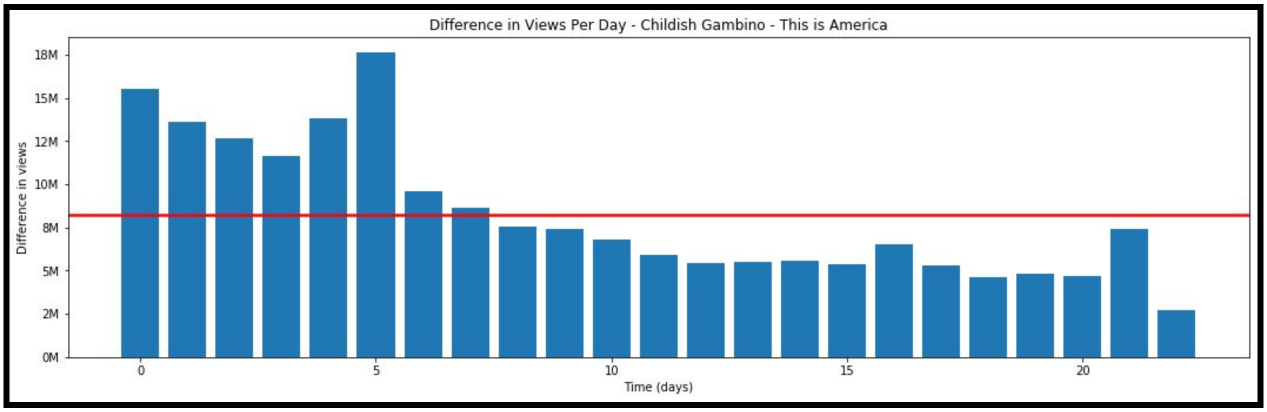

Figure 11: Difference in Views per Day - Childish Gambino - This is America

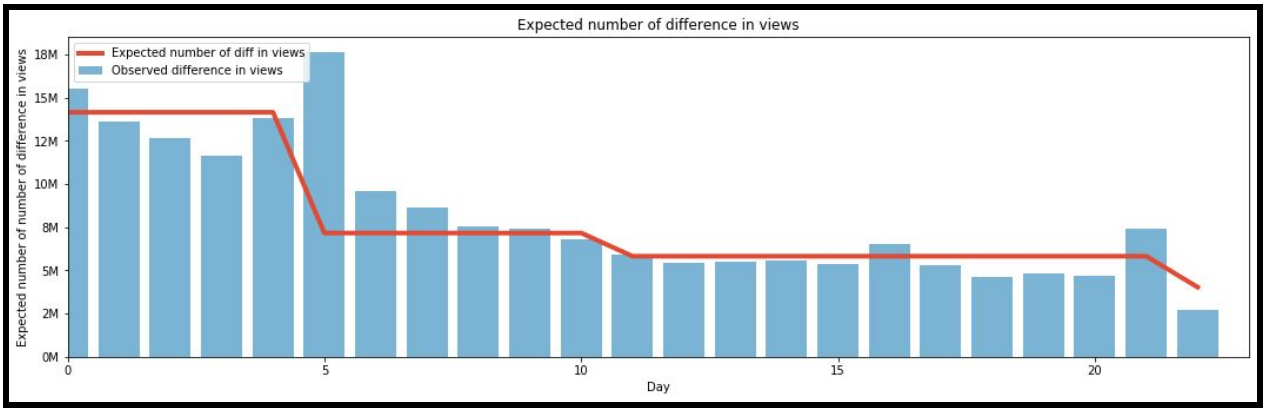

Figure 12: Expected number of difference in views per day using Bayesian Inference - Childish Gambino

Predicting Popularity of the Videos using Machine Learning Methods:

This portion of the project deals with the second problem statement given at the beginning of the report. The scenario with more details can be viewed as follows

Scenario:

YouTube (the Client) would like to identify which group of videos is popular i.e. which group gets the most views and engagement from the viewers. This could help with pushing those videos to new markets,identifying traits that make them popular and incentify the viewers to make videos with such traits. YouTube is looking to minimize Type II errors of the prediction model i.e. labeling videos that are popular as not popular should be kept at a minimum. If there is an increased number of incidences where popular class video is labeled as not popular class video, YouTube will not advertise those videos to the new markets and lose possible new audiences.

Approach:

Wrangling:

The combined (USA and Mexico) data frame contains over 76,000 rows. A lot of videos are present in the data frame multiple times with the statistics for each of the collection day. In order to avoid duplication of the statistics only the most recent data was kept for the analysis. This exercise brought down the total number of rows from over 76,000 to around 36,000.

Feature Engineering:

1. A new statistic was created using likes, dislikes and comment_count to show the engagement level of the users for the YouTube videos. The formula is given below. The comment_count feature is given twice the weight of likes and dislikes because if a user comments on the video he/she is technically putting in more effort compared to someone who hits likes/dislikes buttons

$$Engagement\ Score\ (En) = \ Number\ of\ likes\ +\ Number\ of\ dislikes\ +\ 2 * \ Number\ of\ comment\ count$$

2. Added additional features for title length (title_len)and channel title length (channel_title_len). These could be informative features for the classification models
3. The log of views (views_log) and log of En (En_log) were added as transformed features. This makes these features normally distributed and having these on the same scale helps with clustering of the data
4. Text Extraction - 10 common words from the “title’: The titles of the videos have a lot of different words and some are commonly used by the content creators repeatedly. Using sklearn’s CountVectorizer converted a collection of text documents (titles of the videos) to a sparse matrix with their respective word counts. The vocabulary formed after fit_transform method is extracted by using .get_feature_names() and saved in a list. The vocabulary is converted to an array and then to a data frame with feature names as columns. Stopwords in english and spanish and also some of the expressions used in the titles that are not ‘words’ are dropped from the vocabulary

Clustering:

Using KMeans in sklearn, clustering was performed using En_log and Views_log features. Plotted knee-elbow graph to find the optimal number of clusters. Following is the plot showing the optimal number of clusters is either 3, 4 or 5. The knee point is loosely defined as the point of maximum curvature in a system i.e. where the inertia is least changed from one point to the other. Inertia is an attribute to identify the sum of squared distances of the samples to the nearest cluster

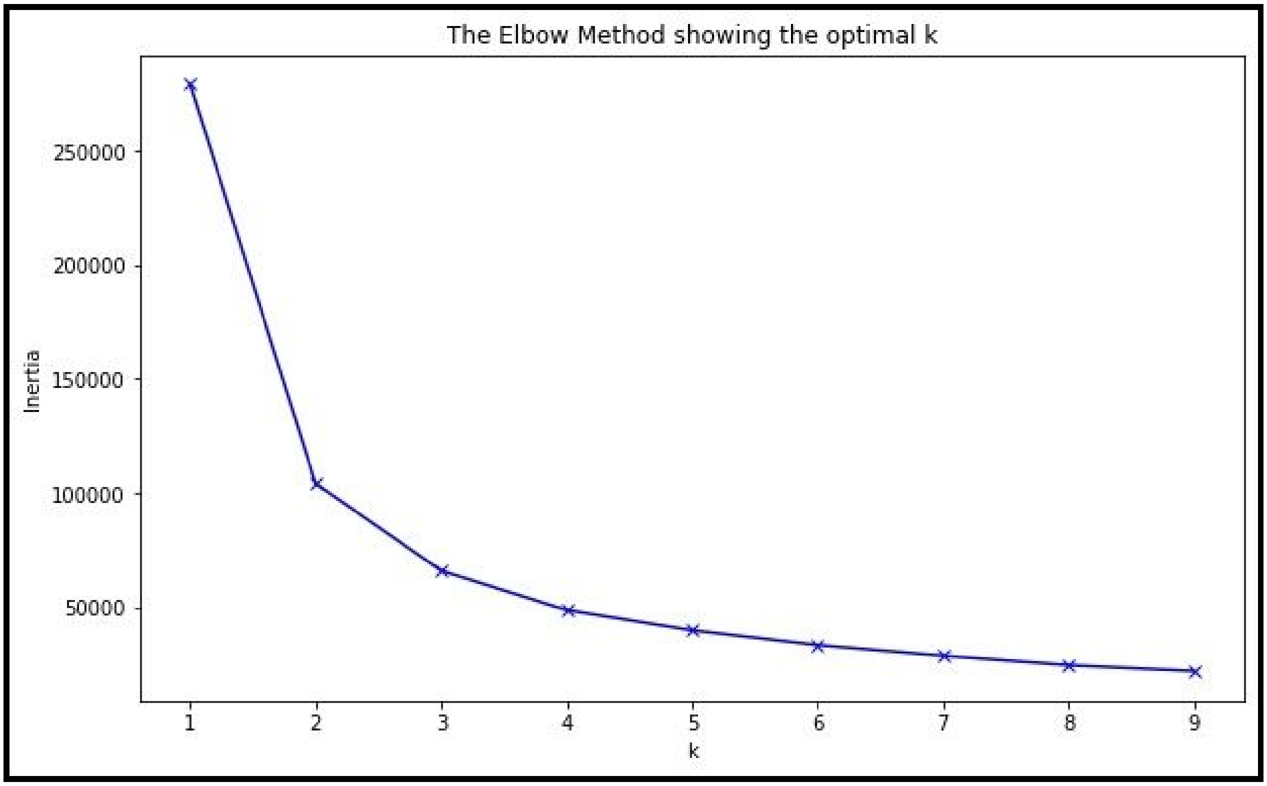

Figure 13: Knee-Elbow Plot

Classes were created using 3 clusters. Figure 14 shows the scatter plot of the distribution of classes over En_log and Views_log data.

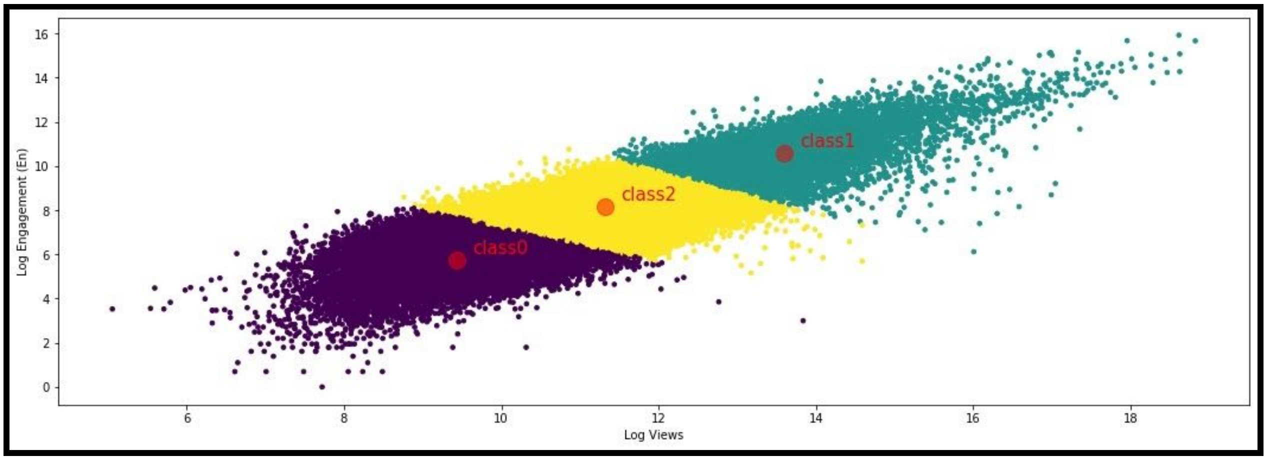

Figure 14: Class distribution with 3 clusters

Classification:

Classes Description:

	Number of Views (e^x)	Engagement Score (e^y)	Video Count
Class 1	>=1.2 M	>= 60 K	9152
Class 2	0.2 M <= Views < 1.2 M	2900 <= En < 60K	13,851
Class 0	< 0.2 M	< 2900	12,150

Table 1: Views and Engagement Scores

Majority of the videos are in class 2 as can be seen in Table 1. The classes are imbalanced and produce a challenge for classification methods. The imbalanced datasets when used with machine learning techniques have poor performance towards minority class and usually minority class is the most important class. SMOTE (Synthetic Minority Oversampling Technique) is a technique that oversamples the minority class. After splitting the data into train and test sets, the train set was oversampled for the minority class using SMOTE. The new distribution yielded 10,379 samples for each class.

Using KNN, Random Forest, Gradient Boost and AdaBoost prediction models, classification was performed. GridSearchCV was used to loop over multiple parameters of the models and find the best possible score for each model. The summary of the performances is shown in Table 2. Confusion matrices are shown in Figure 15 - 18

Accuracy Scores:

	Best Score	**Best Params**
KNN	0.583	{'n_neighbors': 1}
Random Forest	0.639	{'max_depth': 20, 'max_features': 6, 'n_estimators': 100}
Gradient Boost	0.548	{'n_estimators': 90}
AdaBoost	0.522	{'n_estimators': 140}

Table 2: Accuracy Scores and Best Parameters

Confusion Matrices:

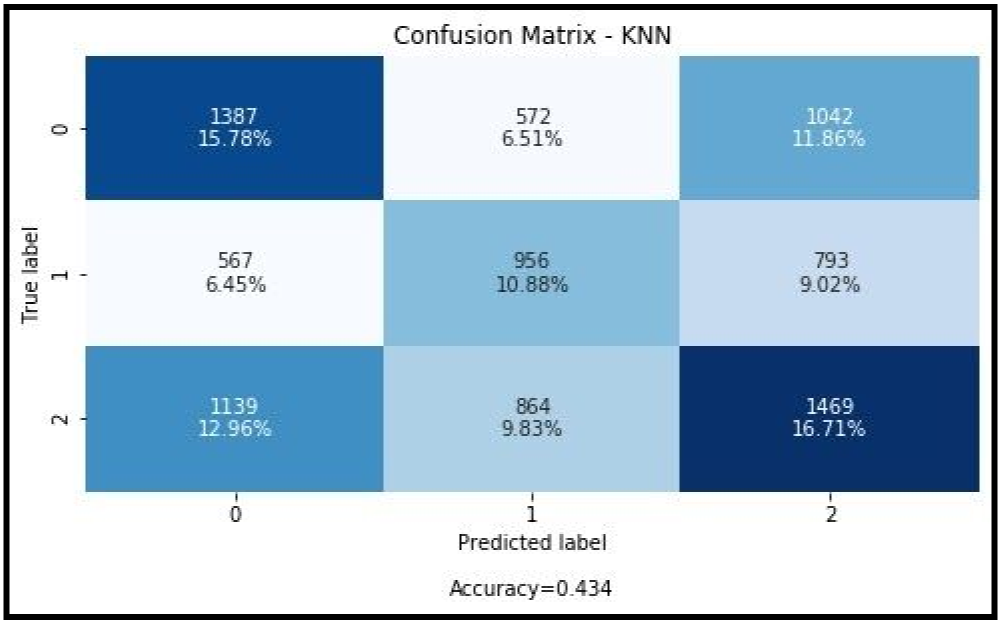

Figure 15: Confusion Matrix - KNN

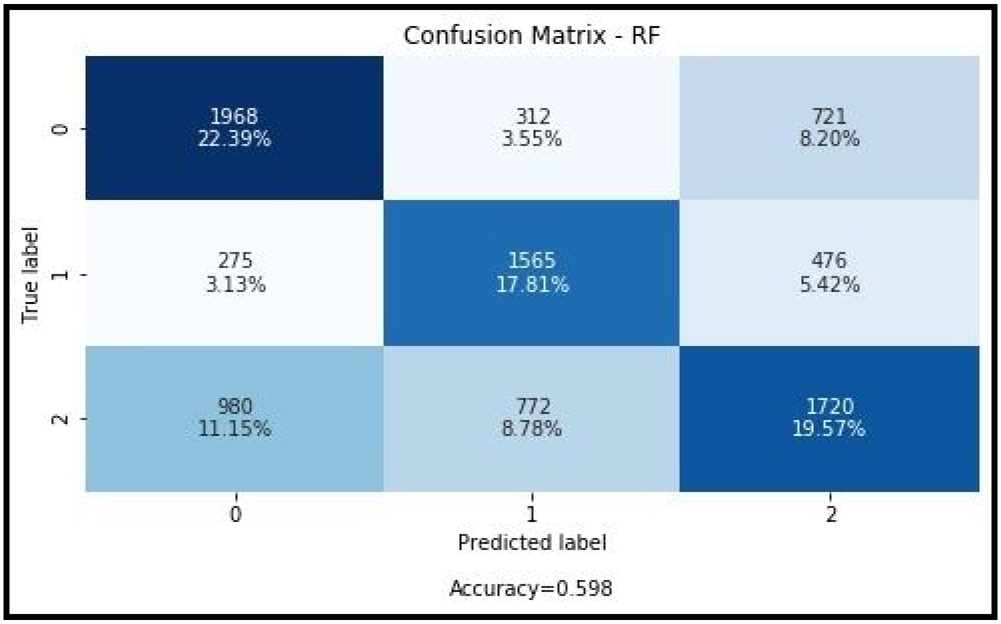

Figure 16: Confusion Matrix - Random Forest

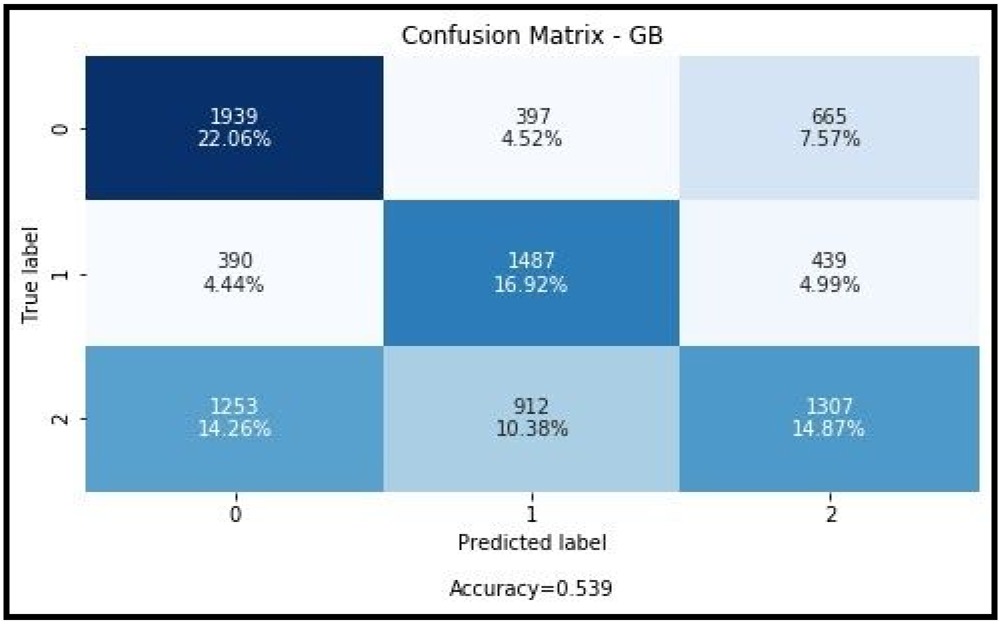

Figure 17: Confusion Matrix - Gradient Boost

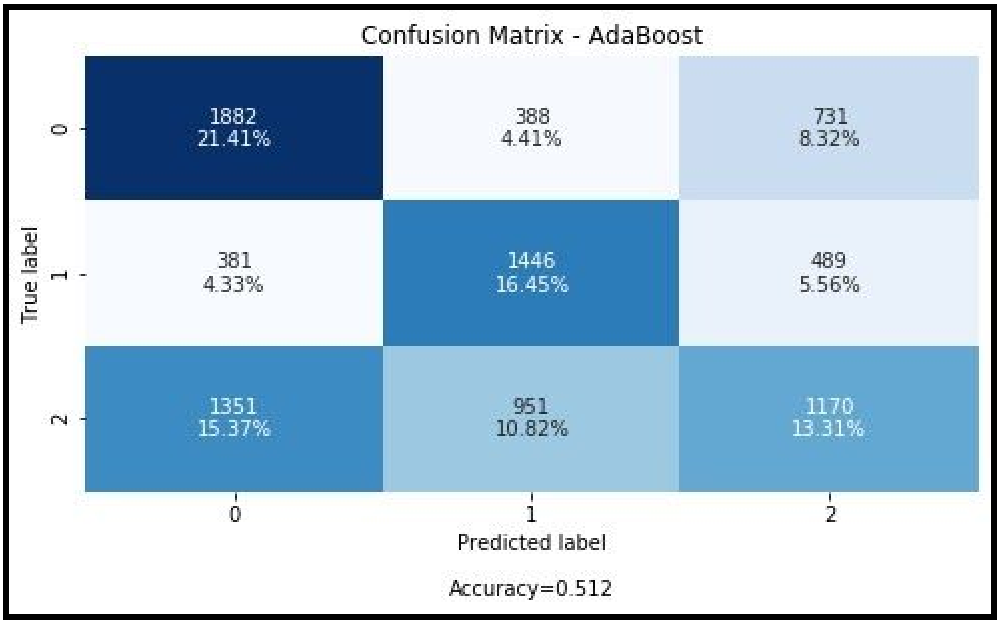

Figure 18: Confusion Matrix - AdaBoost

The accuracy of all the models (Table 2) is not very high. AdaBoost has the lowest accuracy and Random Forest has the highest accuracy. However, choosing a model for the prediction depends upon the goal at hand. The scenario to consider here is given in the second problem statement and the interpretation of the results and choosing an appropriate model will be done based on that scenario.

From Figure 14 the most relevant class to YouTube in class 1 since the videos in that class bring the highest views and engagement from the audiences. YouTube will advertise those specific videos to new markets to gain traction. If a class 1 is labeled as class 0 or 2, YouTube will not advertise that video and potentially not gain a new audience.Therefore YouTube is looking to minimize Type II errors of the prediction model. A high incidence of False Negatives will be the worst outcome compared to having high incidence of False Positives. Keeping this in mind the precision, recall and f-scores below for the four models are calculated for class 1 only (considered as positive class).

Precision/Recall/F-Scores - Class 1:

	Precision	**Recall**	**F-Score**
KNN	0.4	0.41	0.41
Random Forest	0.59	0.68	0.63
Gradient Boost	0.53	0.64	0.58
AdaBoost	0.62	0.52	0.57

Table 3: Precision, Recall and F-Score - class 1

From Table 3, since we are only concerned with having low Recall (goal is to have least number of False Negatives), Random Forest and Gradient Boost could be eliminated from consideration. Both Random Forest and Gradient Boost predicted more false negatives(high recall score) i.e. class 1 being labeled as class 0 or class 2 making YouTube drop some of the popular videos from being advertised and possibly losing new audience. AdaBoost and KNN seem to have a lower Recall and adequate Precision score. However since F-Score of AdaBoost is more than KNN's F-score it could be chosen as the model to achieve the required goal.

Conclusion:

To answer the first problem statement given at the beginning of the report, several data analysis methods were performed on the YouTube Data. After wrangling the data and visualizing it in different ways, it was concluded that in Mexico audiences prefer more variety in the content compared to the USA audience and the number of average views in the USA is more than average views in Mexico. In both countries the max views were brought in by the Music category. Using statistics specifically bootstrapping inference method on the views data, it was concluded with 95% confidence that the average views in the USA are between 2.25M to 2.41M and in Mexico the average views are 0.31M to 0.34M. Using this information YouTube can strategize decisions on where to add more support infrastructure and people.

The second part of the problem statement is to predict the popularity of the videos. Using KMeans clustering method, the videos were labeled with three classes. The KMeans algorithm was executed on log of the views and log of the new feature called Engagement score (En). This resulted in imbalanced class distribution therefore SMOTE was used to artificially balance the data. The ten most common words in the title of the videos were extracted using CountVectorizer and added to the data frame as additional features. The supervised learning algorithms KNN, Random Forest, Gradient Boost and AdaBoost were applied to the data.Since the goal was to keep Type II errors to a minimum, the recall score was given more importance than other matrices. The Recall score of Random Forest was the maximum among others and therefore was recommended as the choice of predicting the popularity of the videos given the data from YouTube.