Predicting Popular Video Genres

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. Can the most popular future YouTube video be predicted by the given data?

Data Mining and Data Wrangling

Data Mining: Kaggle has an enormous amount of data on YouTube's Video Statistics. It can be found at Trending YouTube Video Statistics
The data includes 16 columns (fields) in its raw form for both the USA and Mexico videos. There are a total of 40,949 records found in the USA csv file and 40,451 records found in Mexico csv files. The data for the most part is clean.

Data Wrangling:

1. The data was loaded from csv files available on Kaggle into data frames for the USA and Mexico. Jupyter notebook was used to build the code in PythonThere were several columns that were not relevant to do the analysis. These were dropped from the data frames using ".drop()" method in pandas.
2. The data in its raw form doesn't contain category(genre) by name. It only has an id associated with it. To solve this issue, json files for the categories for the USA and Mexico were downloaded from www.googleapi.com/youtube/v3/. After downloading the response using "requests.get()", the response was converted to a json nested dictionary using ".json()" method. The key "items" in the nested dictionary has the translation of interest i.e cateogory_id: category_name. A simple "for" loop function is written to extract that dictionary out. The id field is however a string and therefore it needed to be changed into integer in order to use .map() method on the data frame to add an additional column of "category_name" to the data frame.
3. After addition of the column "category_name", there were 252 rows with NaN in the new column for Mexico. The category id of 29 had no translation in Mexico's json category file therefore NaN was written in the newly added column for 29 rows. After investigating what those videos might be, it was determined they were all different categories and since the data frame for Mexico had 40+ rows, dropping those 29 rows would not make a huge difference
4. The Mexico data frame was appended to the USA data frame after adding a new column "country" to distinguish the rows. A reset_index was needed to make index set again with the new data added

5. Lastly, dropped all the duplicates from the combined data frame. Publish_time and trending_date fields were changed to DateTime data type using pd.to_datetime. Changing Trending_Date was not straightforward. A function was written to change format YYDDMM into ISO80010

Exploratory Analysis

The code for building data frames and plotting several graphs for exploratory analysis can be found here

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?

These questions were answered by plotting bar plots and correlation matrices in Python. 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. Figure 7 shows correlation matrix between video's views, likes, dislikes, and comment_count.

Analyzing the Graphs:

The **unique videos versus average views** in each category for Mexico and USA in 2017 and 2018 are shown in Fig1 through Fig4. First thing that jumps out from these figures is that the **Mexican audience watches unique videos more 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(Fig1 and Fig3). 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 category. Interestingly none of the categories dropped in the number of average views in the USA. However minimal the change, it was still positive.

A review of correlation between average views, likes, dislikes and comment count between the two countries show that **there is a slight correlation between likes/ views, dislikes/views and comment count/views for videos watched in Mexico.**

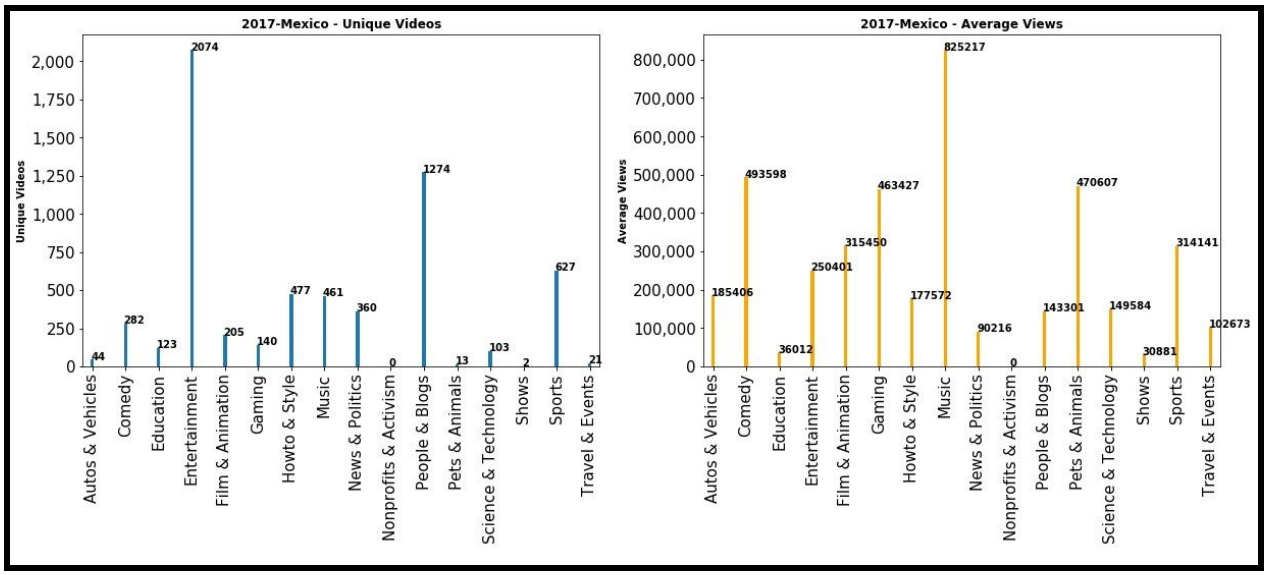

Figure 1: Mexico - Unique Videos vs Average Views, 2017

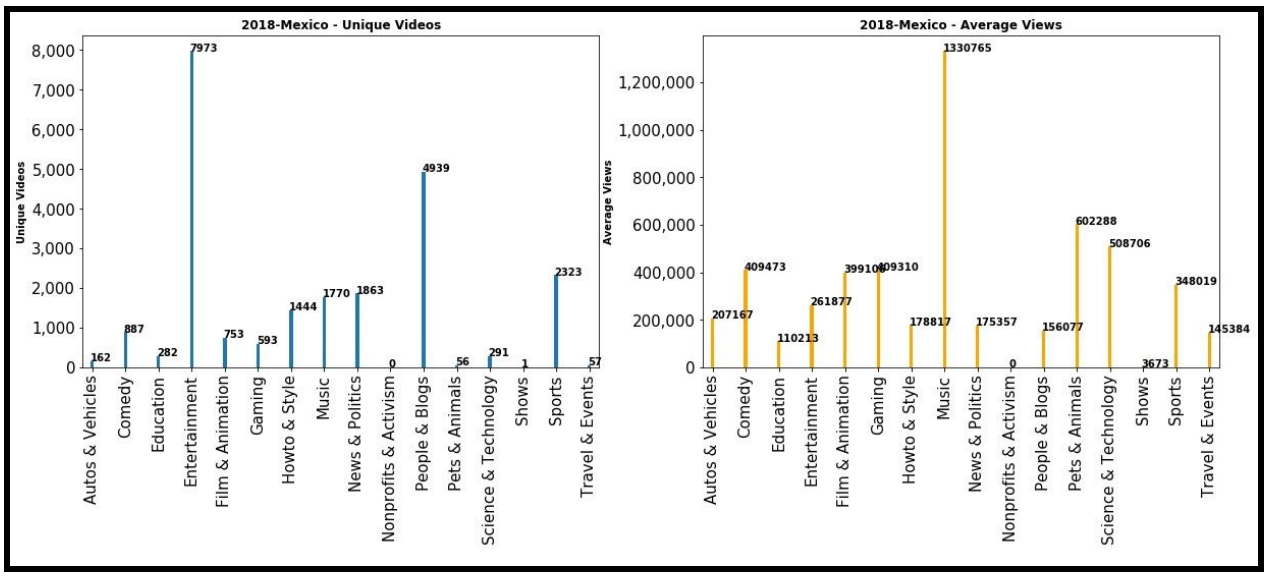

Figure 2: Mexico - Unique Videos vs. Average Views, 2018

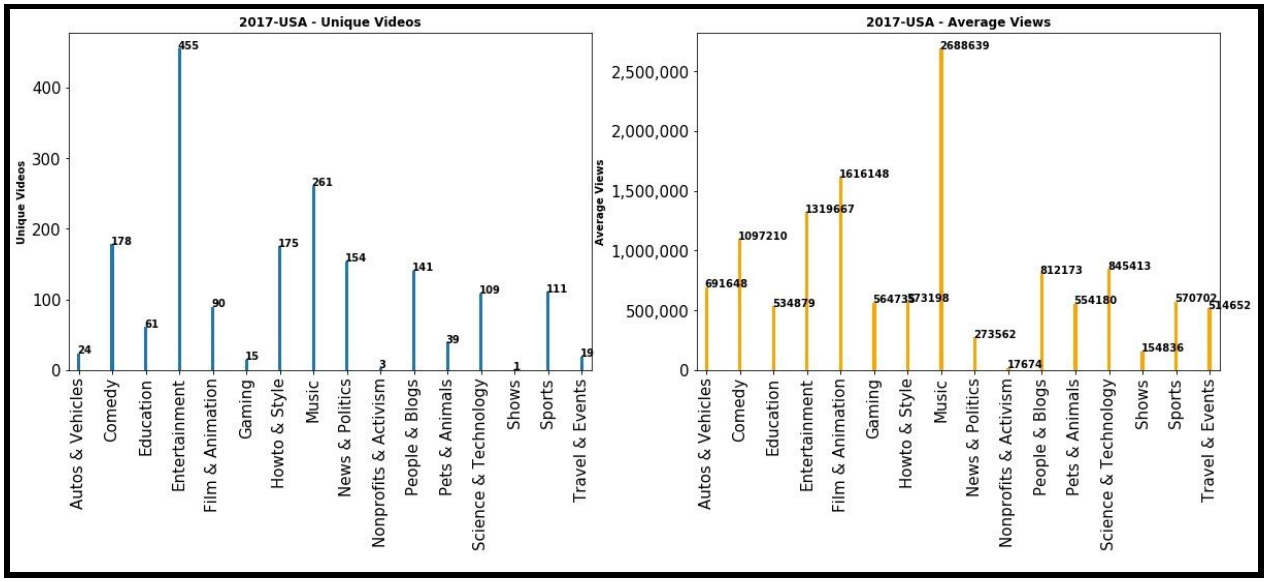

Figure 3: USA - Unique Videos vs. Average Views, 2017

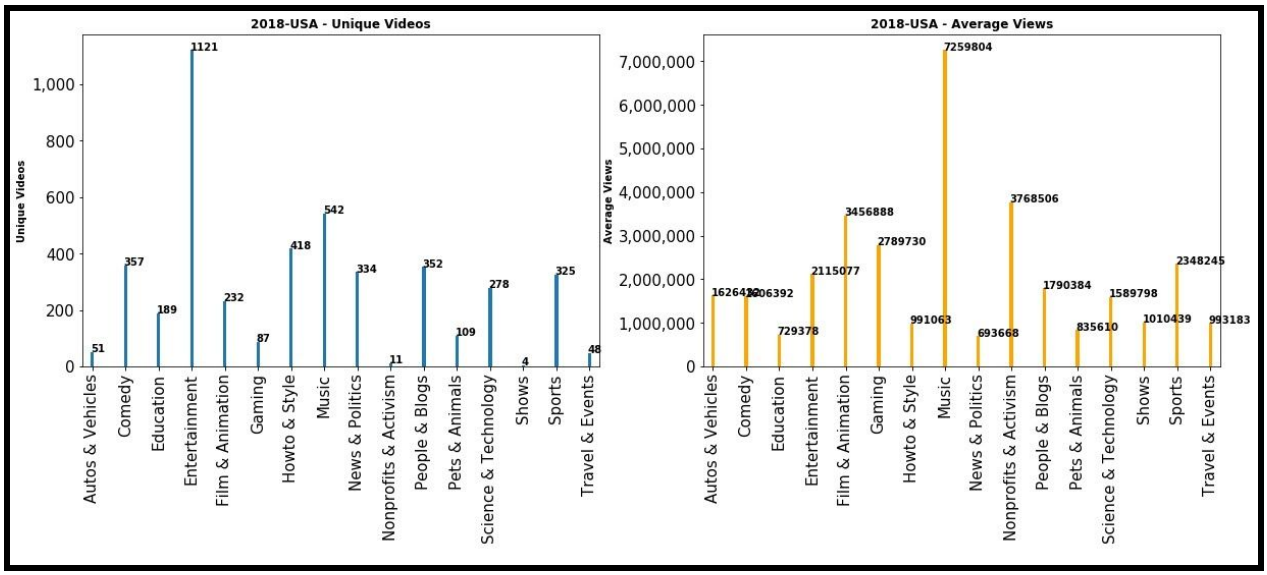

Figure 4: USA - Unique Videos vs. Average Views, 2018

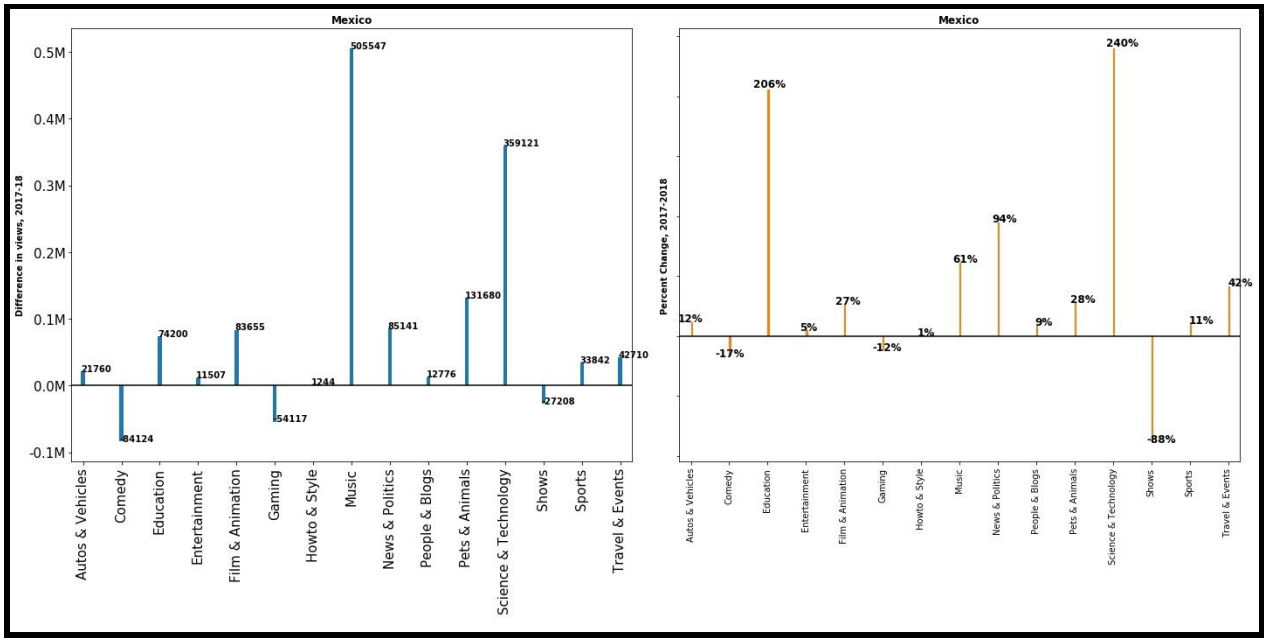

Figure 5: Mexico - Difference in views from 2017 to 2018

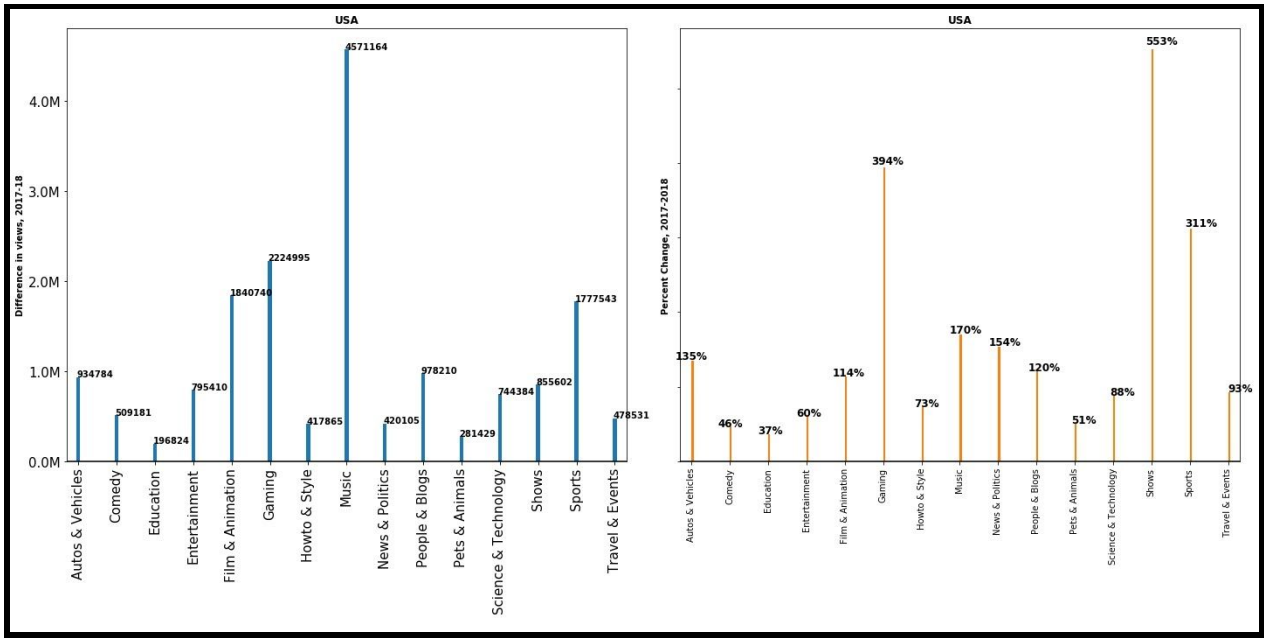

Figure 6: USA - Difference in views from 2017 to 2018

		views		likes		dislikes		comment_count	
	country	Mexico	USA	Mexico	USA	Mexico	USA	Mexico	USA
	country								
views	Mexico	1.00	0.72	0.92	0.34	0.93	-0.17	0.93	-0.11
	USA	0.72	1.00	0.78	0.79	0.69	0.37	0.70	0.43
likes	Mexico	0.92	0.78	1.00	0.42	0.97	-0.11	0.98	-0.04
	USA	0.34	0.79	0.42	1.00	0.37	0.83	0.34	0.87
dislikes	Mexico	0.93	0.69	0.97	0.37	1.00	-0.14	0.98	-0.07
	USA	-0.17	0.37	-0.11	0.83	-0.14	1.00	-0.17	1.00
comment_count	Mexico	0.93	0.70	0.98	0.34	0.98	-0.17	1.00	-0.10
	USA	-0.11	0.43	-0.04	0.87	-0.07	1.00	-0.10	1.00

Figure 7: Correlation Matrix

Statistical Analysis

The code for building data frames and plotting several graphs for exploratory analysis can be found here

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.

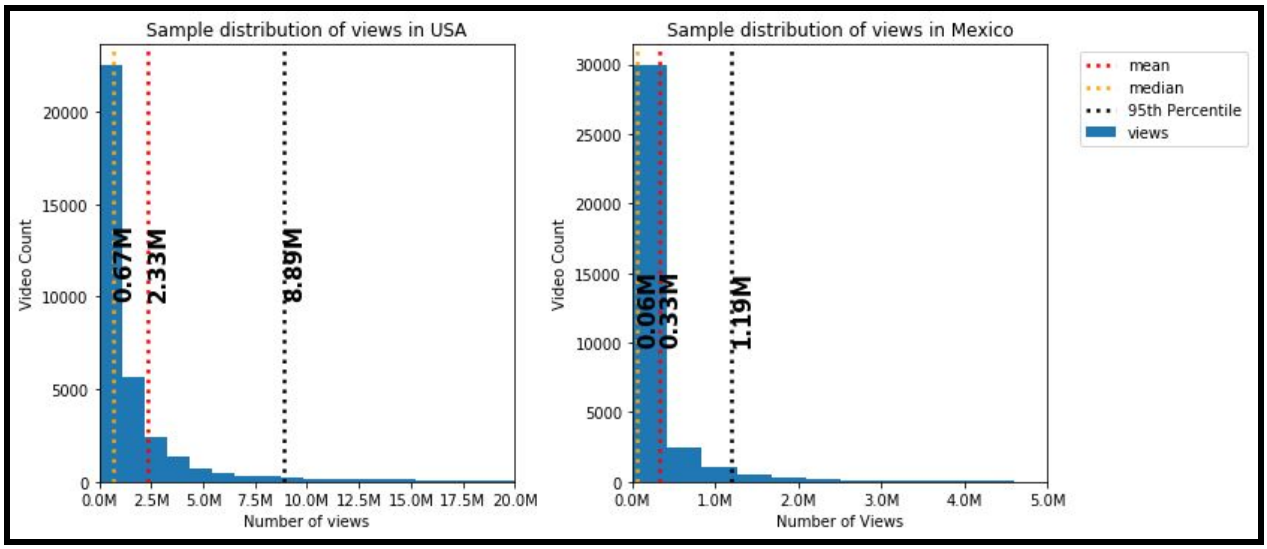

Figure 8: Sample Statistics - USA/Mexico

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

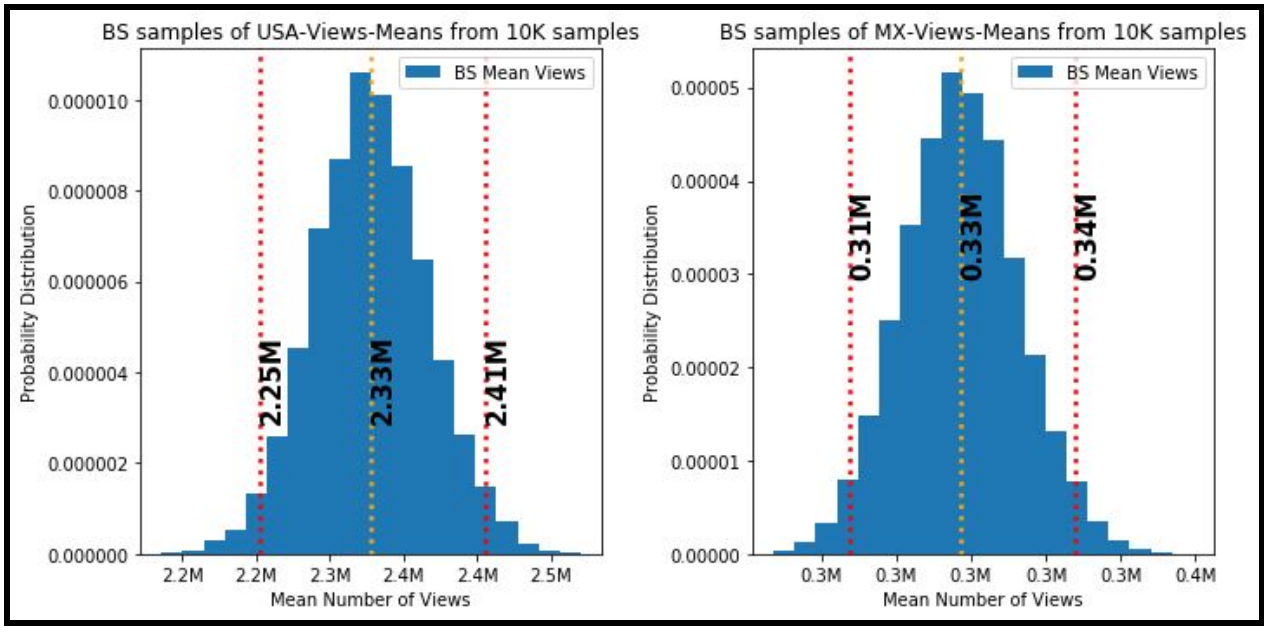

Figure 9: Inference of population mean with 95% CI

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 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, changed over different numbers of periods, Bayseian inference method was used. The details of how different a prior distributions were built and how the posterior distributions looked go here

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 10 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, however it can be seen from the bar chart in Figure 10 that the rate increases as the days go up.

The Bayesian model showed a slightly different picture of the same data. Figure 11 shows that the change in views happened between day 5 and 6. Expected value of views went from 75M to 180M views

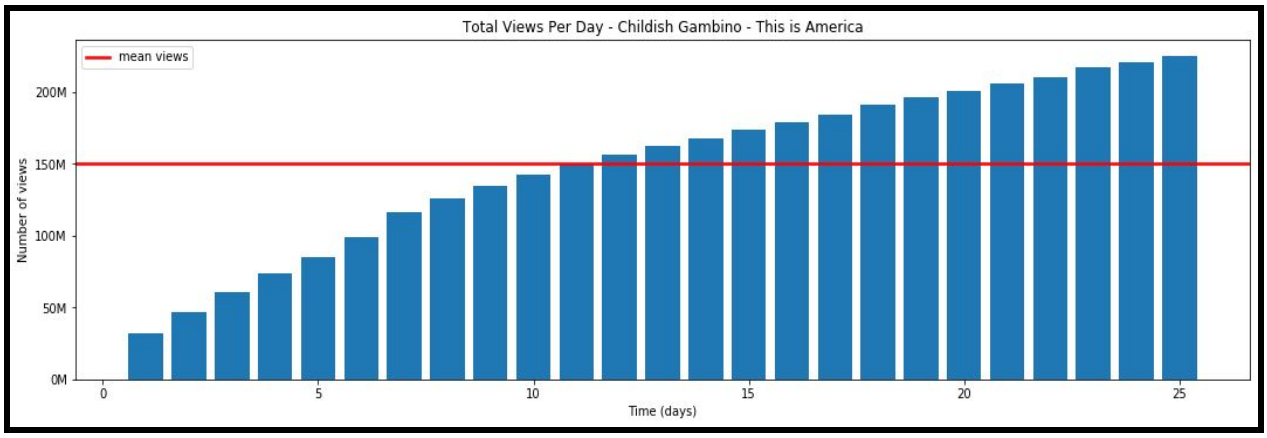

Figure 10: Views per Day - Childish Gambino - This is America

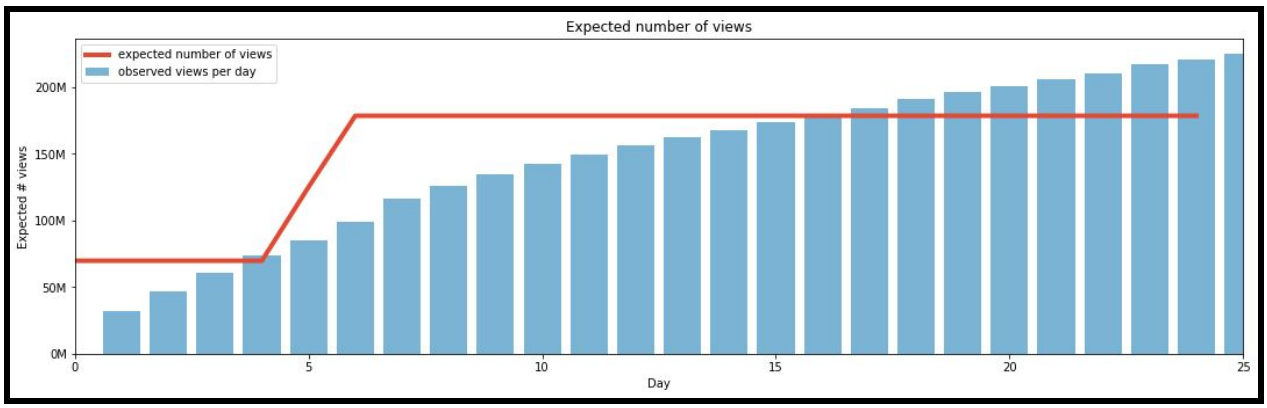

Figure 11: Expected number of views per day using Bayesian Inference - Childish Gambino

Expected Number of the Difference in the Views: For the same video, a bar plot is drawn in python, using the difference in views between the consecutive days. This is shown in Figure 12, Using a similar Bayesian inference model, the expected number of differences in number of views in three periods was calculated and plotted in Figure 13.

It can be observed from Figure 12, 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 13. 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.

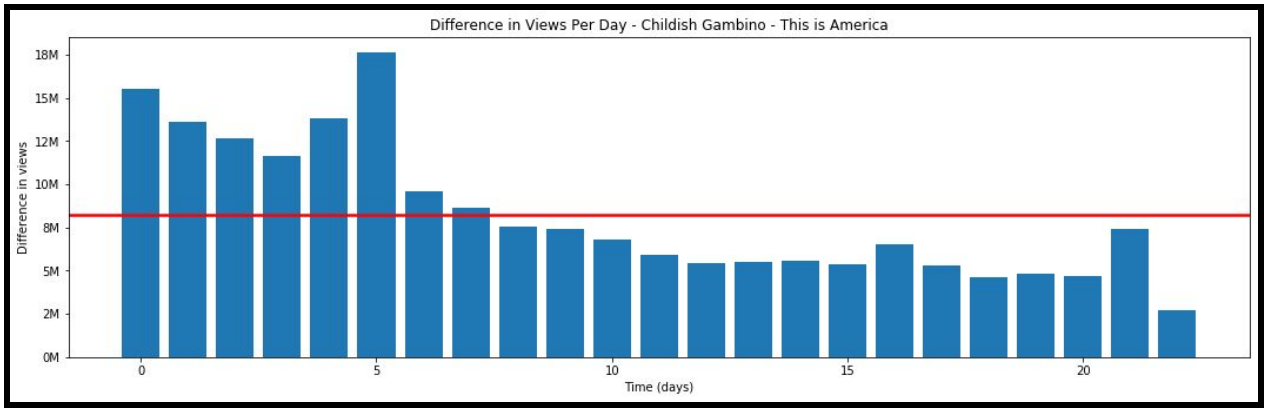

Figure 12: Difference in Views per Day - Childish Gambino - This is America

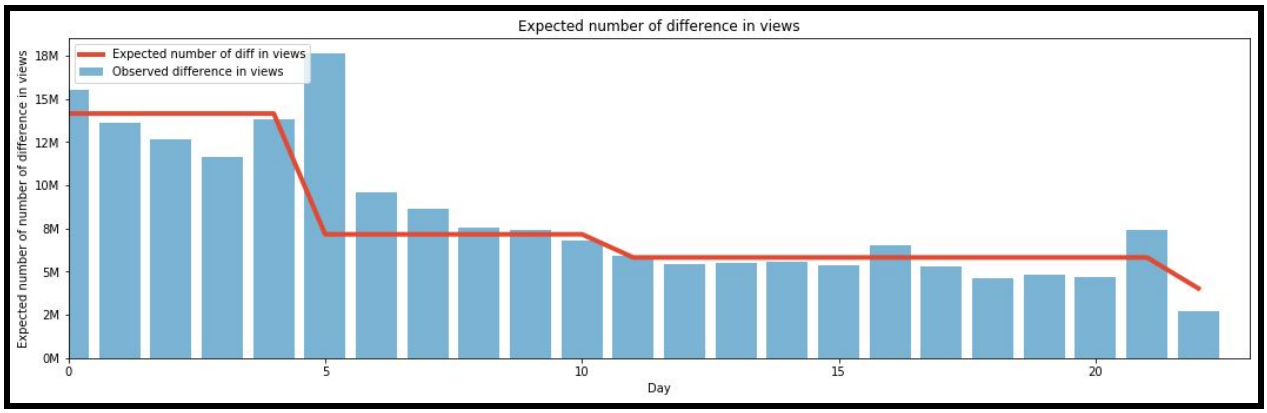

Figure 13: Expected number of difference in views per day using Bayesian Inference - Childish Gambino