Predicting Popularity of YouTube Videos

Fariha Baloch
Springboard - Data Science Track

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 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. 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

- Columns that were not relevant to this analysis were dropped, e.g.
  - 'Video_id ,thumbnail_link, description, tags, channel_title

- The data doesn't contain category(genre) by name. A json file, available here, was downloaded and used to map category names to category id

- Dropped the rows with NaN and date columns were formatted to DateTime object

- Mexico data frame was appended to the USA data frame. Reset_index was performed. All the rows that were duplicates were dropped from the data.
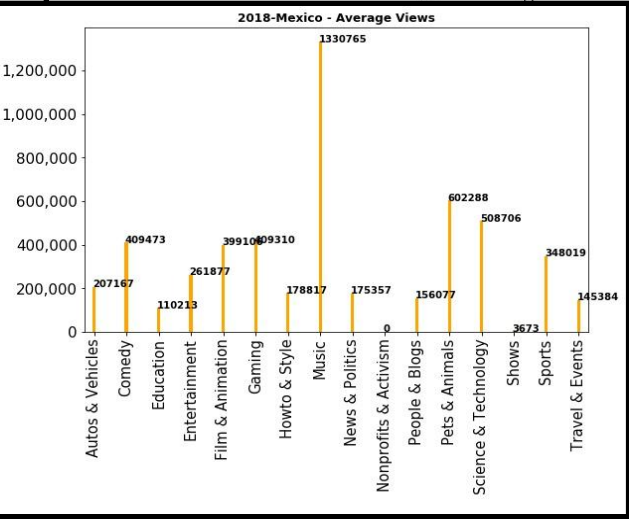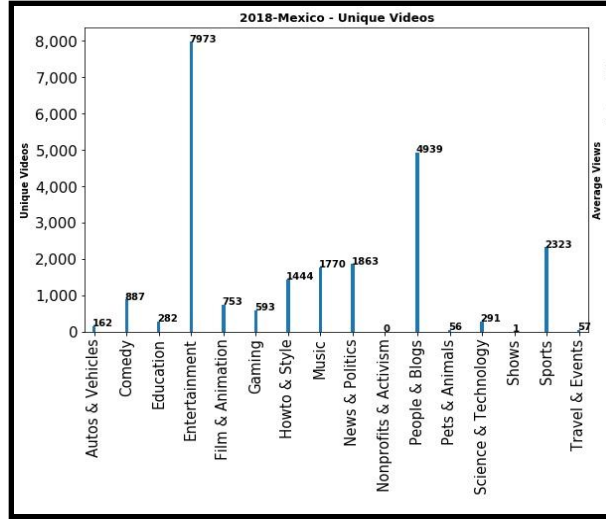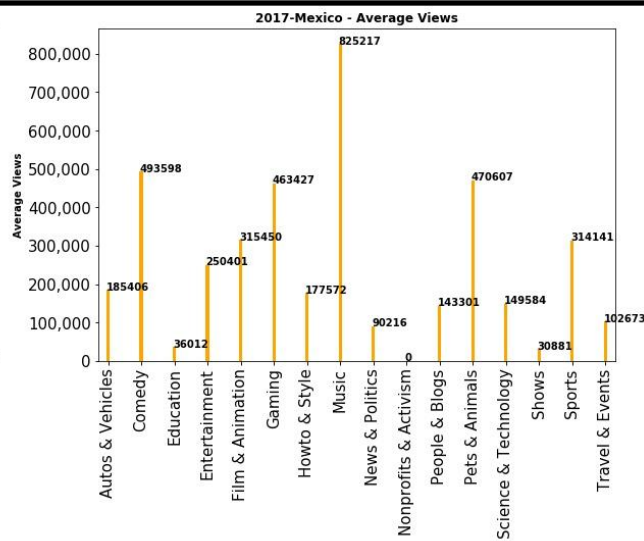
Exploratory Analysis

- To see the video viewing trend in both countries, several graphical analysis were performed

- Some of the questions asked were

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

Mexico - Unique Videos vs Average Views, 2017 - 2018

2017-Mexico - Unique Videos

Category	Value
Autos & Vehicles	44
Comedy	282
Education	123
Entertainment	2074
Film & Animation	205
Gaming	140
Howto & Style	477
Music	461
News & Politics	360
Nonprofits & Activism	0
People & Blogs	1274
Pets & Animals	13
Science & Technology	103
Shows	2
Sports	627
Travel & Events	21

2017-Mexico - Average Views

Category	Value
Autos & Vehicles	185406
Comedy	493598
Education	36012
Entertainment	250401
Film & Animation	315450
Gaming	463427
Howto & Style	177572
Music	825217
News & Politics	90216
Nonprofits & Activism	0
People & Blogs	143301
Pets & Animals	470607
Science & Technology	149584
Shows	30881
Sports	314141
Travel & Events	102673

2018-Mexico - Unique Videos

Category	Value
Autos & Vehicles	162
Comedy	887
Education	282
Entertainment	7973
Film & Animation	753
Gaming	593
Howto & Style	1444
Music	1770
News & Politics	1863
Nonprofits & Activism	0
People & Blogs	4939
Pets & Animals	56
Science & Technology	291
Shows	1
Sports	2323
Travel & Events	57

2018-Mexico - Average Views

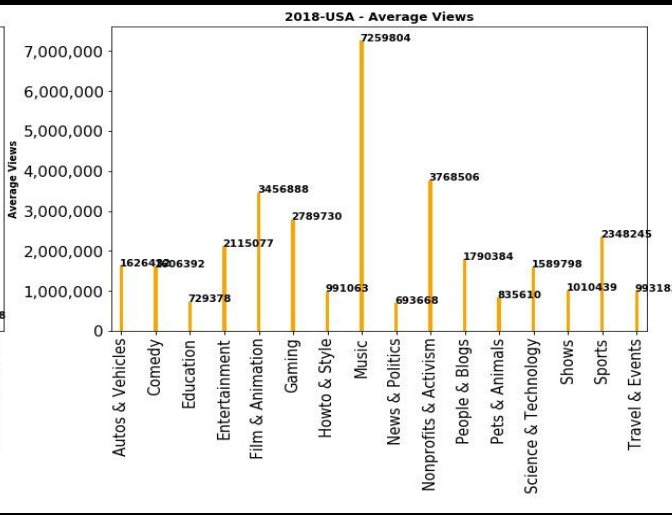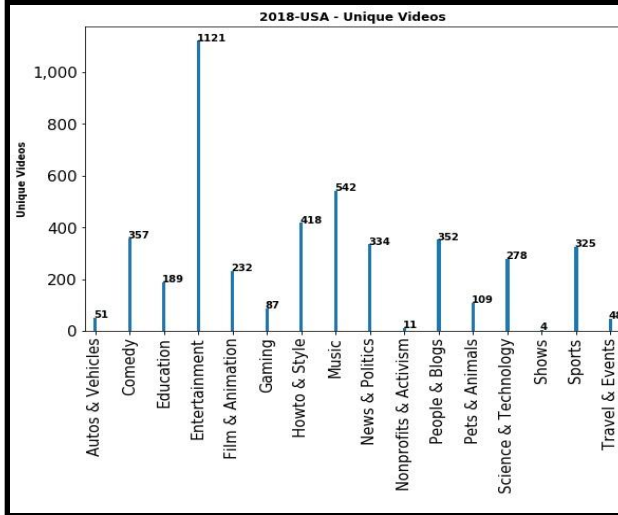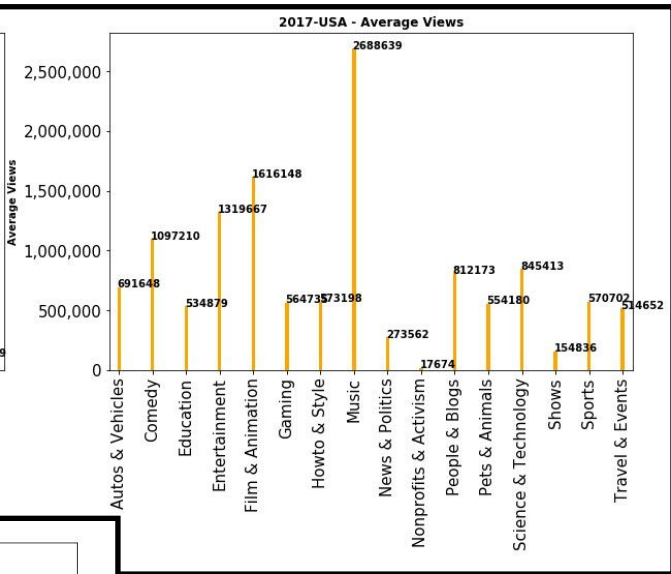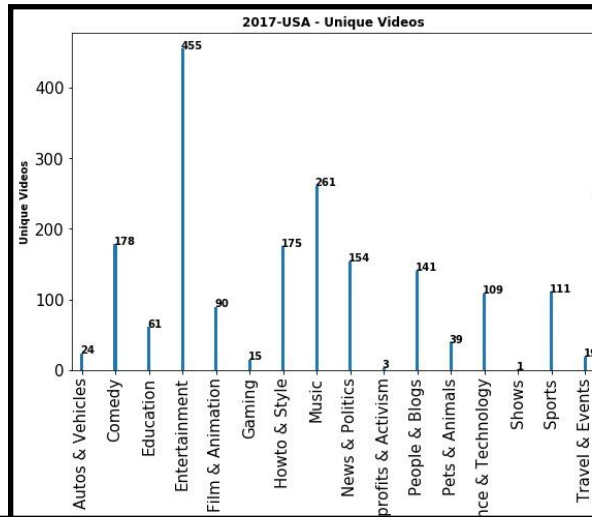| Category	Value
Autos & Vehicles	207167
Comedy	409473
Education	110213
Entertainment	261877
Film & Animation	399106
Gaming	409310
Howto & Style	178817
Music	1330765
News & Politics	175357
Nonprofits & Activism	0
People & Blogs	156077
Pets & Animals	602288
Science & Technology	508706
Shows	3673
Sports	348019
Travel & Events	145384

Ratios:
2017:
Entertainment: 121 av views/video
Music: 1790 av views/video

2018:
Entertainment: 32 av views/video
Music: 751 av views/video

USA -
Unique Videos vs Average Views, 2017 - 2018

2017-USA - Unique Videos

Category	Unique Videos
Autos & Vehicles	24
Comedy	178
Education	61
Entertainment	455
Film & Animation	90
Gaming	15
Howto & Style	175
Music	261
News & Politics	154
Nonprofits & Activism	3
People & Blogs	141
Pets & Animals	39
Science & Technology	109
Shows	1
Sports	111
Travel & Events	19

2017-USA - Average Views

Category	Average Views
Autos & Vehicles	691648
Comedy	1097210
Education	534879
Entertainment	1319667
Film & Animation	1616148
Gaming	564735
Howto & Style	573198
Music	2688639
News & Politics	273562
Nonprofits & Activism	17674
People & Blogs	812173
Pets & Animals	554180
Science & Technology	845413
Shows	154836
Sports	570702
Travel & Events	514652

2018-USA - Unique Videos

Category	Unique Videos
Autos & Vehicles	51
Comedy	357
Education	189
Entertainment	1121
Film & Animation	232
Gaming	87
Howto & Style	418
Music	542
News & Politics	334
Nonprofits & Activism	11
People & Blogs	352
Pets & Animals	109
Science & Technology	278
Shows	4
Sports	325
Travel & Events	48

2018-USA - Average Views

Category	Average Views
Autos & Vehicles	1626412
Comedy	2006392
Education	729378
Entertainment	2115077
Film & Animation	3456888
Gaming	2789730
Howto & Style	991063
Music	7259804
News & Politics	693668
Nonprofits & Activism	3768506
People & Blogs	1790384
Pets & Animals	835610
Science & Technology	1589798
Shows	1010439
Sports	2348245
Travel & Events	993183

Ratios:
2017:
Entertainment: 2900 av views/video
Music: 10,301 av views/video

2018:
Entertainment: 1886 av views/video
Music: 13,394 av views/video

Unique Videos vs Average Views - Analysis

- Mexican audience prefers to watch new content (more unique videos per category) than American audience
- American audiences watch same videos again and again (more average views per category)
- Even though there are more unique videos in Mexico dataframe, the average number of views (overall) is less than the USA
- Music category has the most average views every year in both countries
- Pets & Animal category seems to be second to the Music in Mexico (average views)
- Film & Animation is second in terms of average views in the USA

Year to Year Difference in Average Views
2017 - 2018
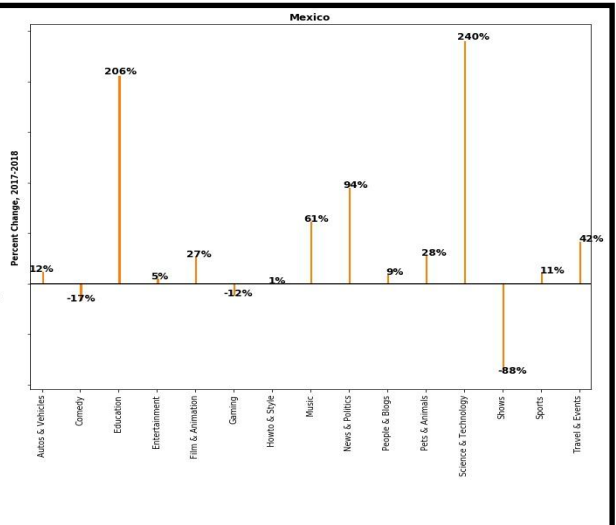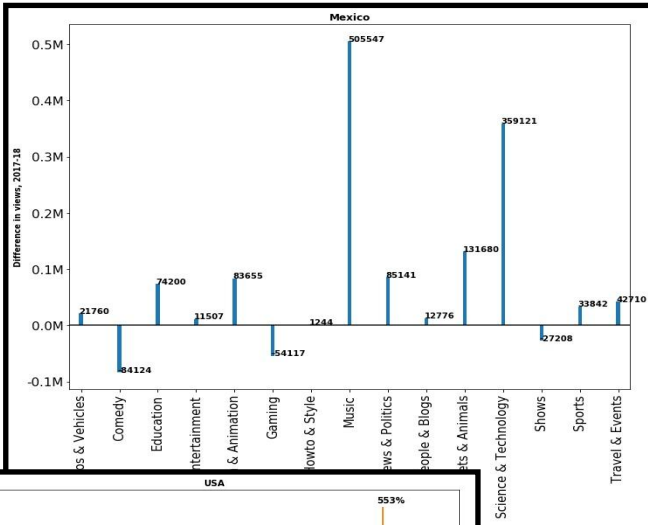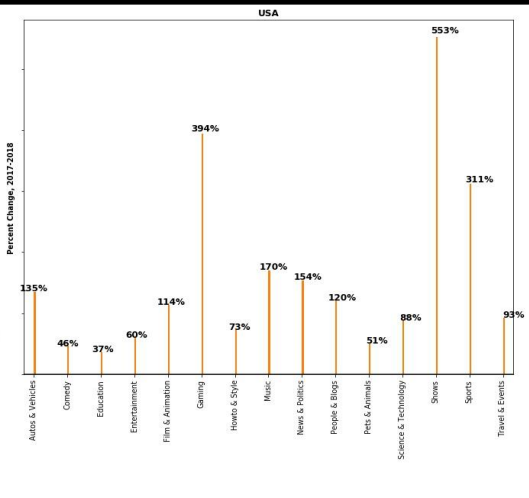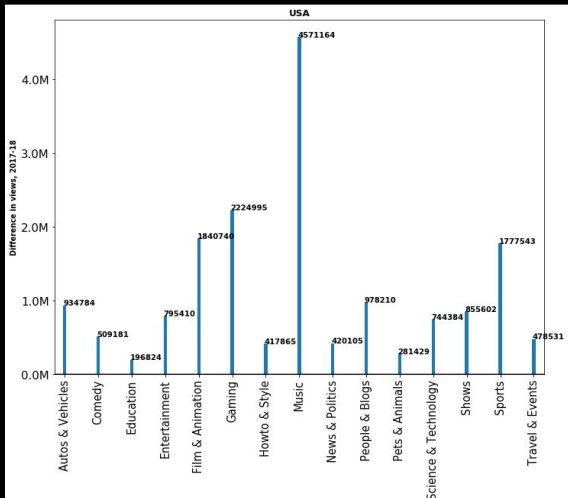
Year to Year Difference in Average Views - Analysis

- 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
- Mexico viewers watched more of the Science & Technology category videos that gained 240% more average views in 2018
- Gaming and the Comedy categories also dropped average views from 2017 to 2018 in Mexico
- None of the categories dropped below zero in the number of average views in the USA. However large the change, it was still positive.

Statistical Analysis

- ○ **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(a):

Mean, Std, and median for the sample data are shown below

<u>SUMMARY OF STATISTICS:</u>

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

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
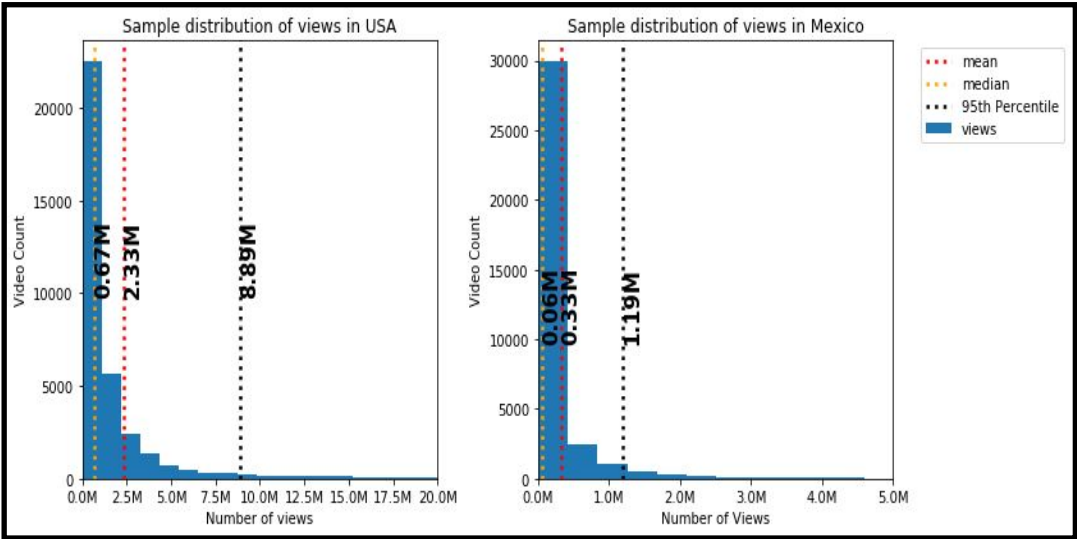
Scenario 1(b):

Inferring population mean using bootstrap method with 10,000 bootstrap samples

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

Mexico:Confidence interval (95%) for population mean is **[0.31M ,0.34M]**
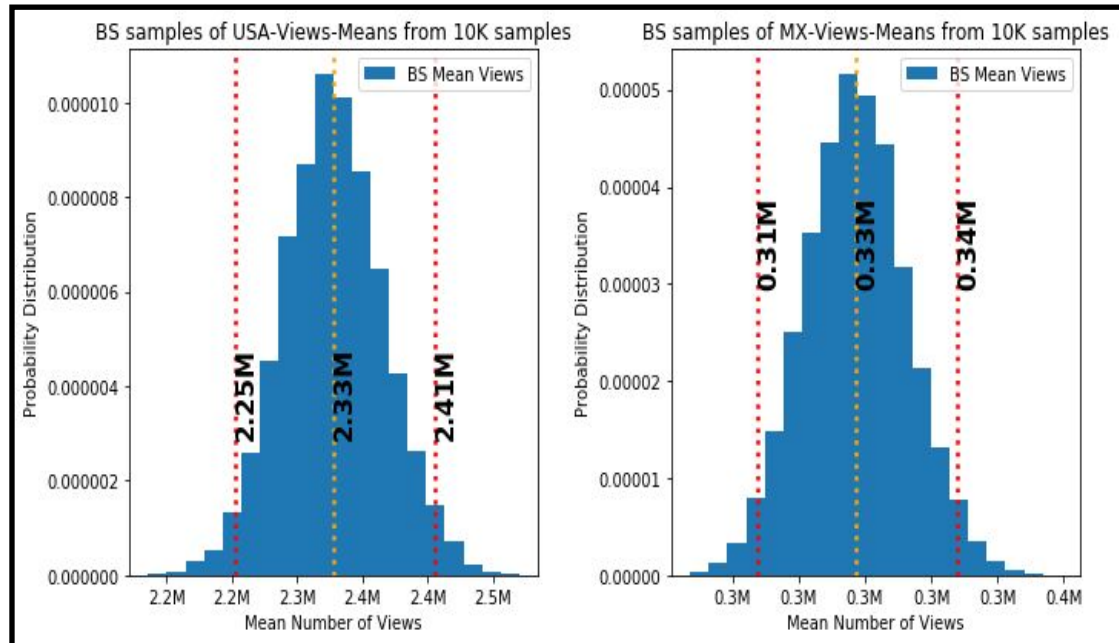
Scenario 2:

To determine if the number of videos in each category depend upon the country it is views in, Pearson chi-squared statistical hypothesis test was performed.

H0: The country and category variables are independent

Ha: 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 Scenario 2:

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(a)

Bayesian inference was used to find expected value of song "Childish Gambino - This is America" (most viewed song in the USA.
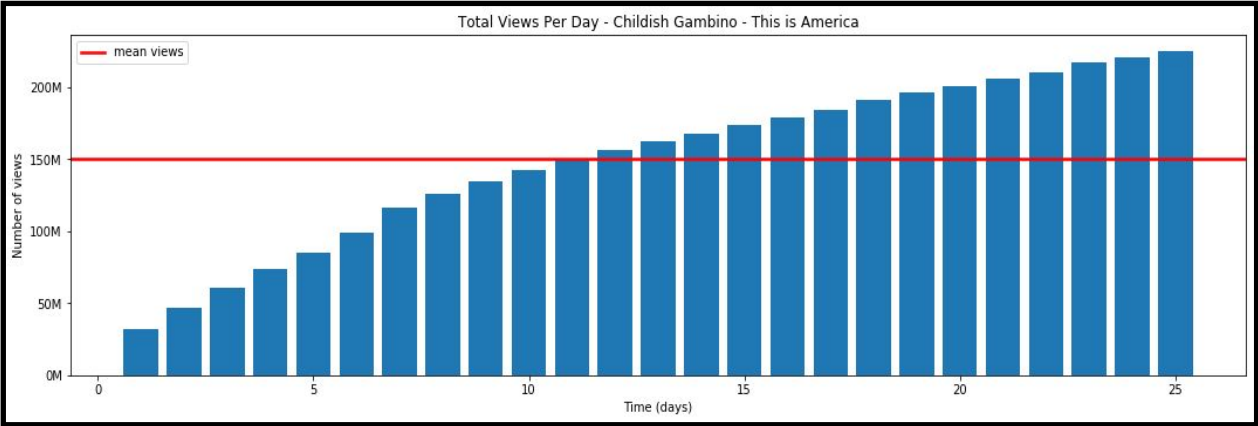
Views per Day
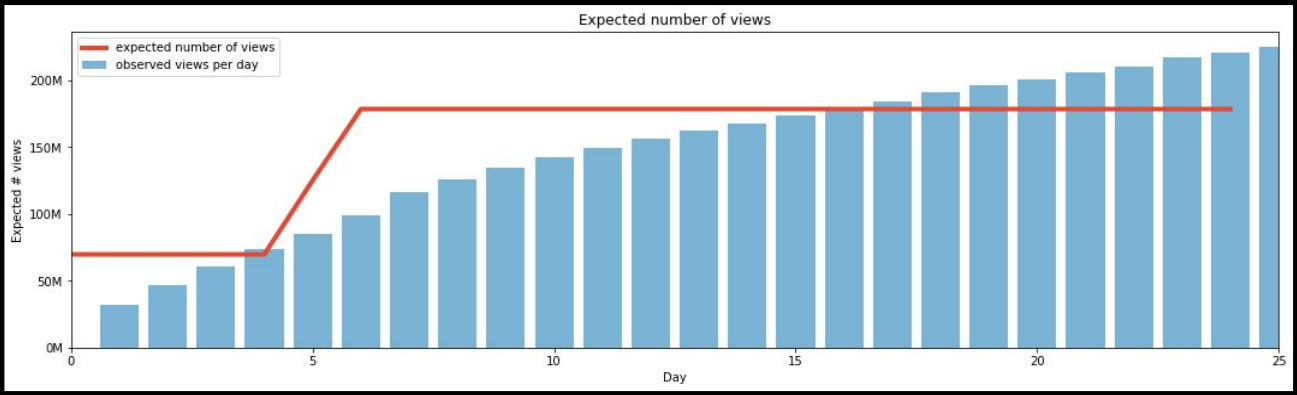
Expected number of views per day using Bayesian Inference
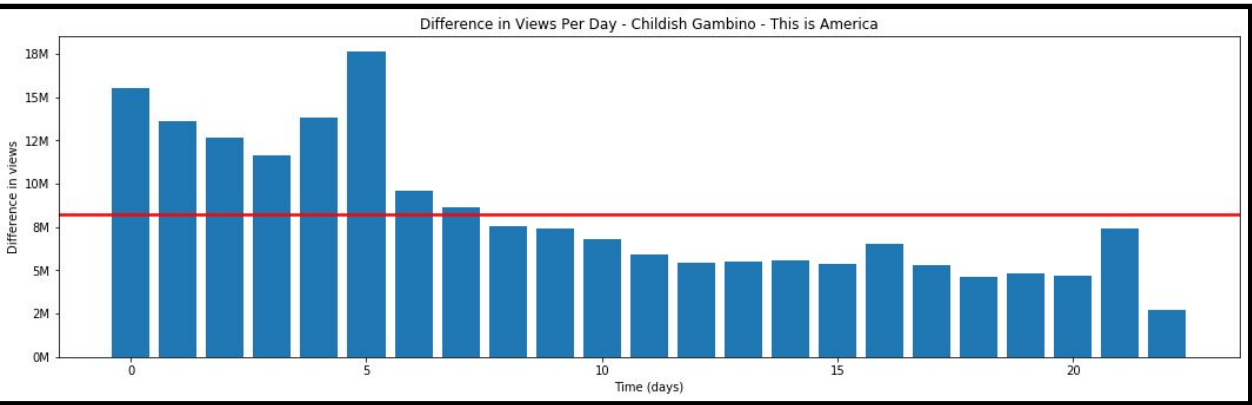
Analysis of Scenario 3(a)

A bar plots in previous slide is to visualize the increment in number of views as the days progress

- The data for "Child Gambino - This is America" music video was collected for 25 days. It can be observed that the mean of the views for this video is 150M

-

- The Bayesian model showed a slightly different picture of the same data. Mean value or the expected value of the views derived by using Bayesian method shows that the change in number of views happened between day 5 and 6. Expected value of views went from 75M to 180M views
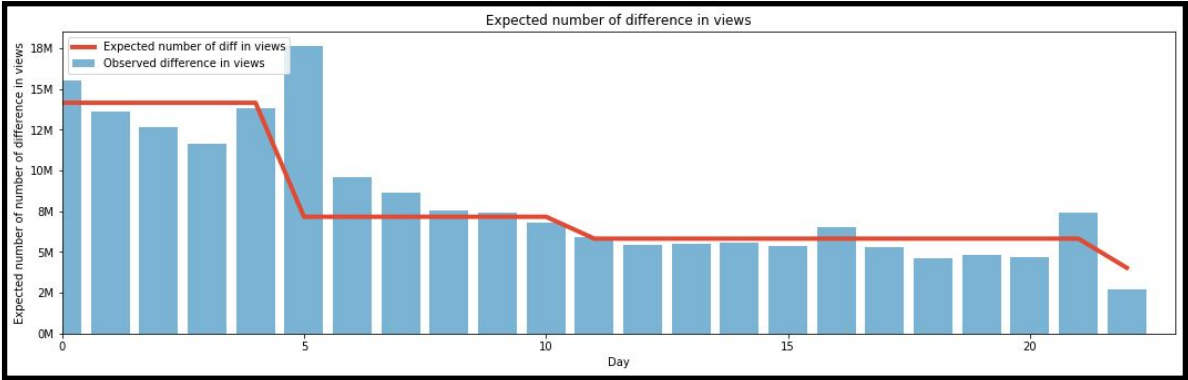
Scenario 3(b):
Bayesian inference was used to find expected value of difference in views of song "Childish Gambino - This is America" (most viewed song in the USA)

Difference in Views per Day

Expected number of difference in views per day using Bayesian Inference

Analysis of Scenario 3(b)

A bar plots in previous slide are to visualize the progress in number of differences in views as the days progress

- The difference in Views per day from the sample data clearly shows that between day 5 / 6 a spike in difference occurred which was predicted by model of expected number of views. The mean of the difference in views is 8.2M

- The Bayesian model for the expected value of differences in the views show that there are three dips in the difference of views at three different points. Day 5. 10, 21

- 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.

Machine Learning

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.
- 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

Feature Engineering

- **A new feature was created to weigh-in each video's engagement capabilities**

  Engagement Score (En)= Number of likes + Number of dislikes + 2 Number of comment_count*

- **Additional features added are, title_length, channel_title_length, 10 most common words in the title, publish day, year, date and hour**

Clustering

- Log of En and Views was used as input to KMeans clustering method to identify classes in the data
- Three clusters produced the best boundaries for the classes
- The class distribution was imbalanced
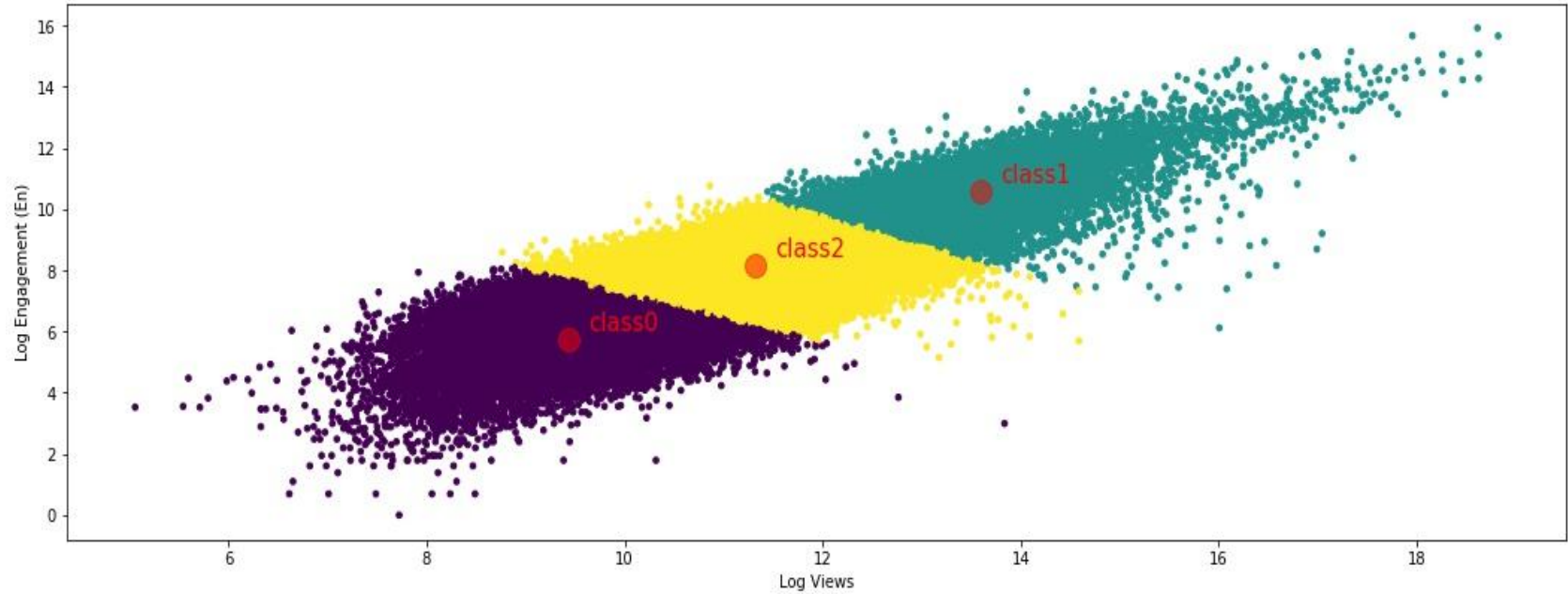
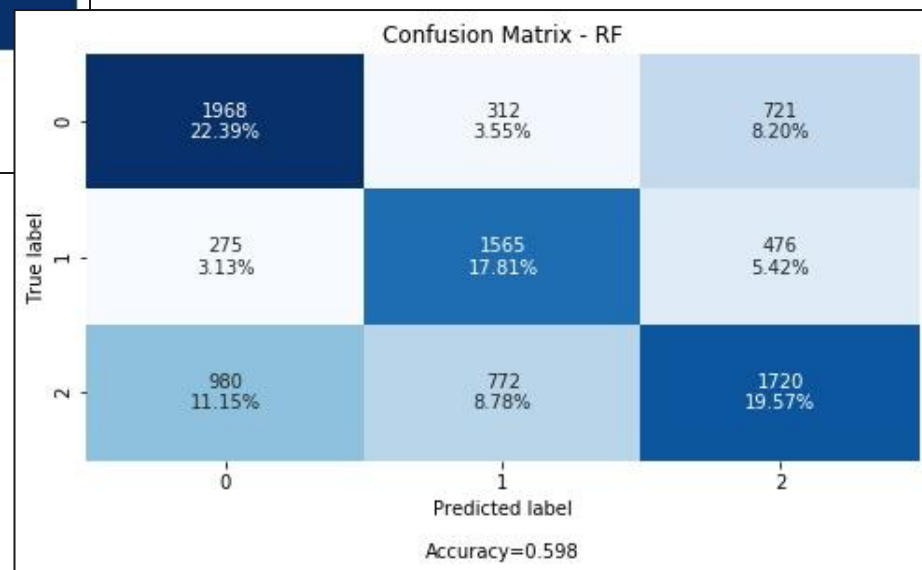|	Number of Views (e^x)	Engagement Score (e^y)	Video Count
Class 1	>=1.2 M	>= 60 K	9152
Class 2	0.2 M <= Views < 1.2 M	2900 <= En < 60K	13,851
Class 0	< 0.2 M	< 2900	12,150

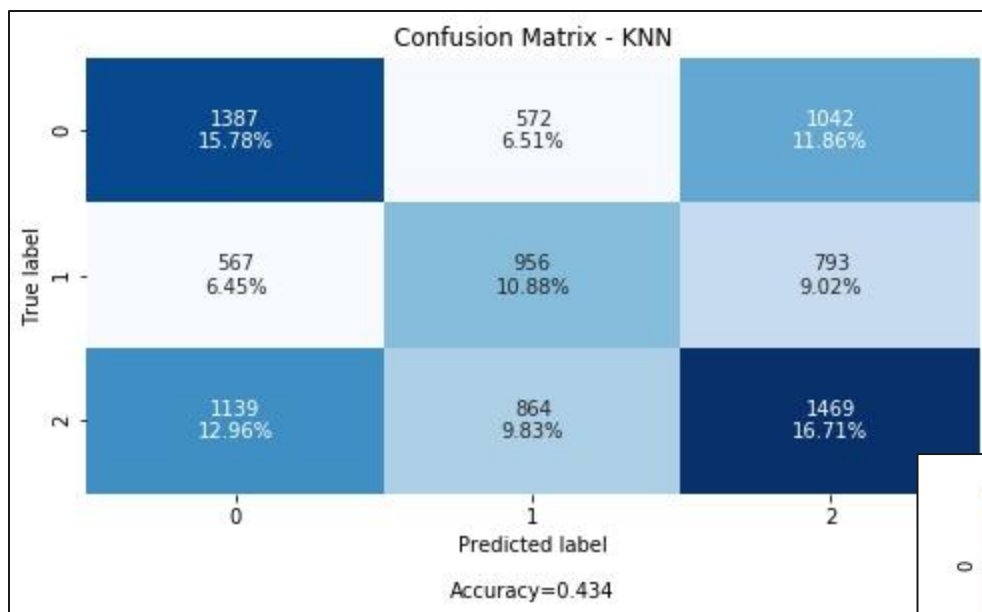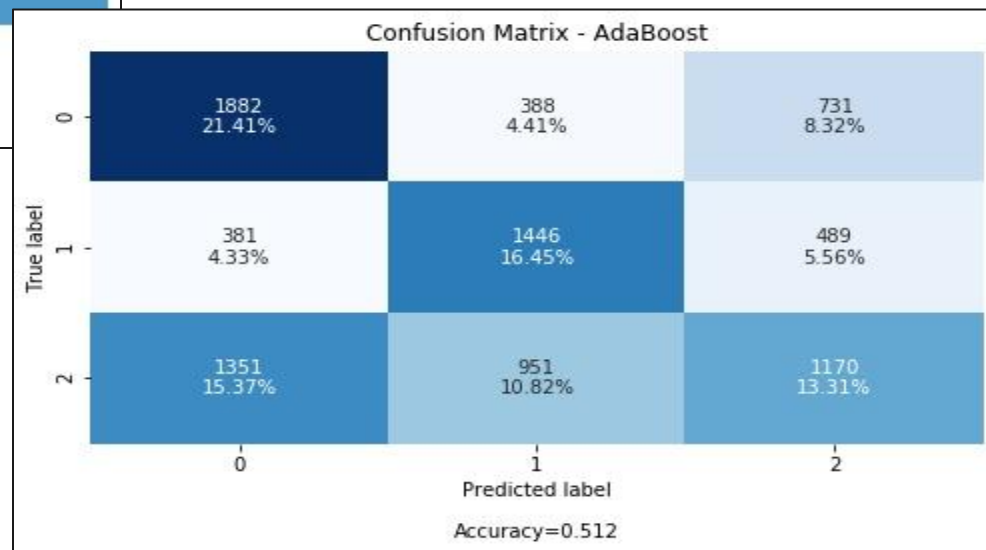Imbalanced Classes - Problematic for classification

Classification

- SMOTE (Synthetic Minority Oversampling Technique) was used to balance the imbalanced data
- GridSeachCV was used to find best parameters for an estimator
- Accuracy Score:
  - KNN: 58.34
  - Random Forest: 63.91
  - Gradient Boost: 54.84
  - AdaBoost: 52.65
- None of the test set accuracy scores are very high
- Random Forest is better than others

Confusion Matrix - KNN

	0	1	2
0	1387 / 15.78%	572 / 6.51%	1042 / 11.86%
1	567 / 6.45%	956 / 10.88%	793 / 9.02%
2	1139 / 12.96%	864 / 9.83%	1469 / 16.71%

True label / Predicted label

Accuracy=0.434

Confusion Matrix - RF

	0	1	2
0	1968 / 22.39%	312 / 3.55%	721 / 8.20%
1	275 / 3.13%	1565 / 17.81%	476 / 5.42%
2	980 / 11.15%	772 / 8.78%	1720 / 19.57%

True label / Predicted label

Accuracy=0.598

Confusion Matrix - GB

	Predicted 0	Predicted 1	Predicted 2
True 0	1939 / 22.06%	397 / 4.52%	665 / 7.57%
True 1	390 / 4.44%	1487 / 16.92%	439 / 4.99%
True 2	1253 / 14.26%	912 / 10.38%	1307 / 14.87%

Accuracy=0.539

Confusion Matrix - AdaBoost

	Predicted 0	Predicted 1	Predicted 2
True 0	1882 / 21.41%	388 / 4.41%	731 / 8.32%
True 1	381 / 4.33%	1446 / 16.45%	489 / 5.56%
True 2	1351 / 15.37%	951 / 10.82%	1170 / 13.31%

Accuracy=0.512

Classification

	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

- AdaBoost - Recall (0.52) and F-Score is best
- Type II errors will be kept at minimum

Class 1 is most important - more views and more engagement