Springboard Data Science Career Track

Capstone Project II

Title of the project:

YouTube video views in the US and UK: Exploratory data analysis, inferential statistics, and classification

R.S. Elias

January 2020

Table of contents

1. Introduction:

YouTube is a video-sharing platform based in California, USA. “Three former employees of PayPal, Chad Hurley, Steve Chen, and Jawed Karim, devised this platform in February 2005. Google bought the site in November 2006. YouTube allows users to upload, view, rate, share, add to playlists, report, comment on videos, and subscribe to other users. It offers a wide variety of user-generated and corporate media videos” (source: Wikipedia).

1.1. Brief description of the project:

The project covers exploratory data analysis, inferential statistics, and application of machine learning (ML) classification algorithms to a dataset of YouTube video views in the US and UK. In the exploratory data analysis, behavioral patterns of YouTube viewers, for example, number of YouTube video views, likes, and dislikes in the US and UK are analyzed. In the inferential statistical analysis, distributions of number of views in the US and UK are examined based on the null hypothesis that there is no significant difference between the distribution of number of views in the US and UK. In the application of classification algorithms, two ML models -- logistic regression and random forest -- are applied to classify the US YouTube videos according to a computed label, and a comparison among the models built are presented.

1.2 Short summary of findings:

It is observed that distributions of both US and UK views are right-skewed. Few YouTube videos are viral with extremely high number of views. The mean number of views in the US is 939,102; while the mean number of views in the UK is 1,110,467. As there are some extreme number of views, standard deviation of both distributions are very high. The standard deviation of the US views is 2,147,691, which 2.3 times larger than its mean.

In the UK, there are more number of views in the categories of Music, Entertainment, and Sports. However, in the US, there are more number of views in the categories of Comedy, News & Politics, Education, Science & Technology. It is found that the correlation coefficient between number of likes and views is 0.83.

In terms of the applications of ML classification models, it was found that Random Forest classifiers outperform Logistic Regression classifiers in terms of precision and recall.

2. Approach:

2.1. Data Acquisition and wrangling:

The data source is https://www.kaggle.com/datasnaek/YouTube. The dataset contains 6 files -

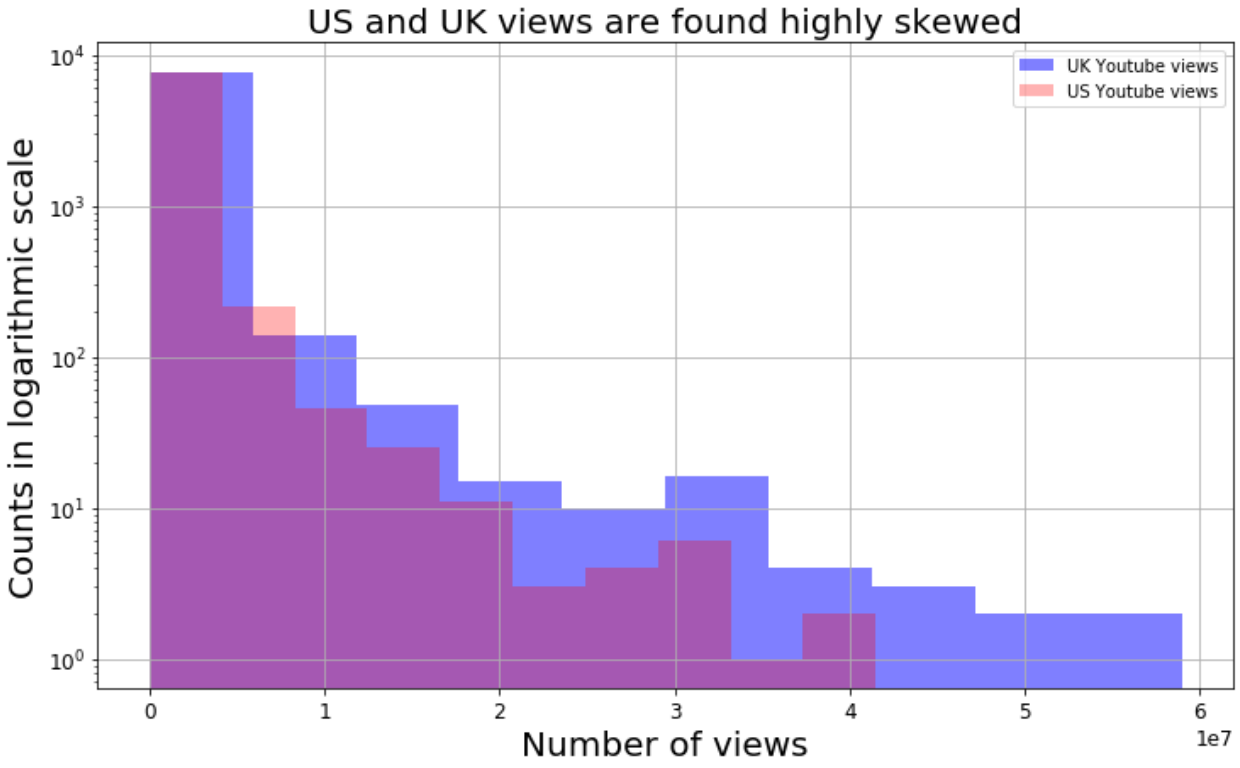
- 3 files on US viewers: UScomments.csv, USvideos.csv, US_category_id.json and
- 3 files on UK data: UKcomments.csv, UKvideos.csv, UK_category_id.json.

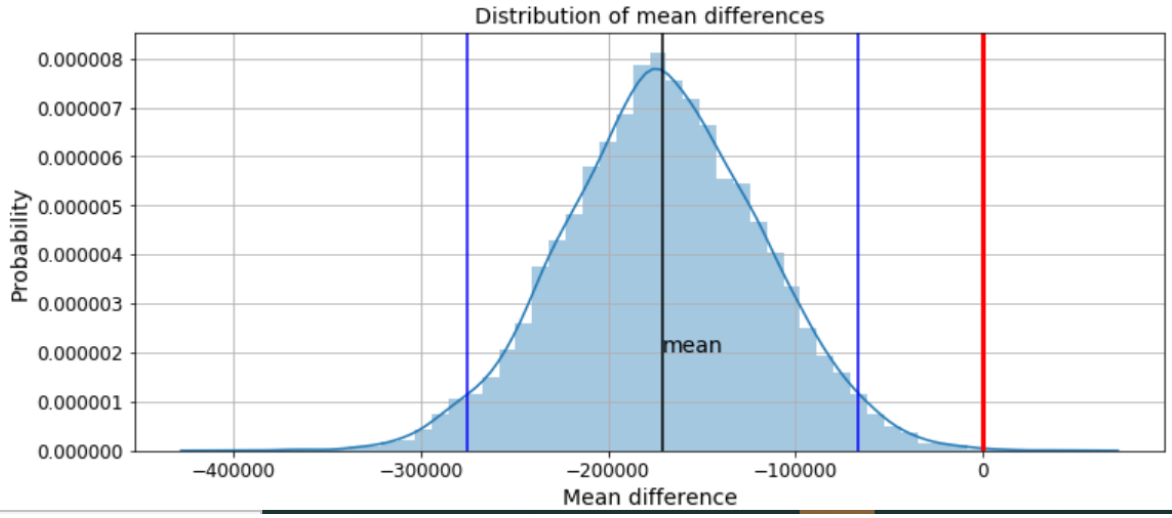
- USvidoes.csv has following 11 columns: ['video_id', 'title', 'channel_title', 'category_id', 'tags', 'views', 'likes', 'dislikes', 'comment_total', thumbnail_link', 'date'] where unique 'video_id's are 2364.
- UScomments.csv has 4 columns: ['video_id', 'comment_text', 'likes', 'replies'], where unique 'video_id's are 2266.
- US_category_id.json has [category_id, kind, etag, item_snippet], where unique category_id are 16.

Data cleaning:
Data types, number of null and non-null values are checked by using df.info(). Some unusual dates, for example, "26.0903jeumSTSzc" are changed to “26.03”. Date format was as “string” type. It has been changed to “datatime64”.

2.2. Storytelling and inferential statistics:

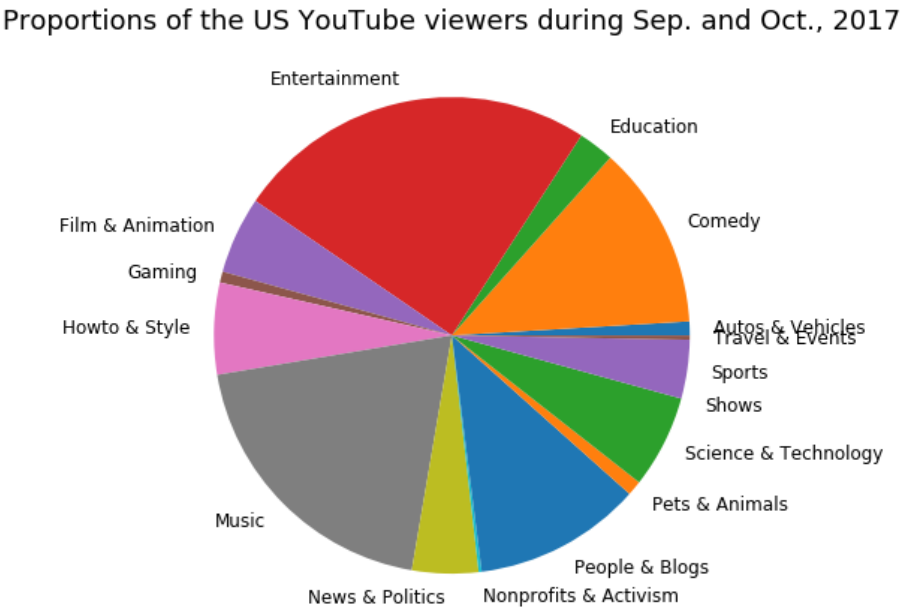
2.2.1. Distribution of the total number of views in the US and UK:
It is observed that few videos are viral with extremely high number of views and hence distributions of both US and UK views are right-skewed.





$\mu 1$ = mean number of US YouTube views
$\mu 2$= mean number of UK YouTube views
Null hypothesis, H0: $\mu 1 - \mu 2 = 0$
Alternative hypothesis, H1: $\mu 1 - \mu 2 \neq 0$
Confidence level, α = 5%
From the above figure, we observe that mean difference of 0 is outside the 95% confidence interval and falls into the zone of rejection of the null hypothesis. We conclude that the mean number of views in the US and UK are statistically significant at 5% confidence level. Negative difference shows that number of Us views are less than the number of UK views.

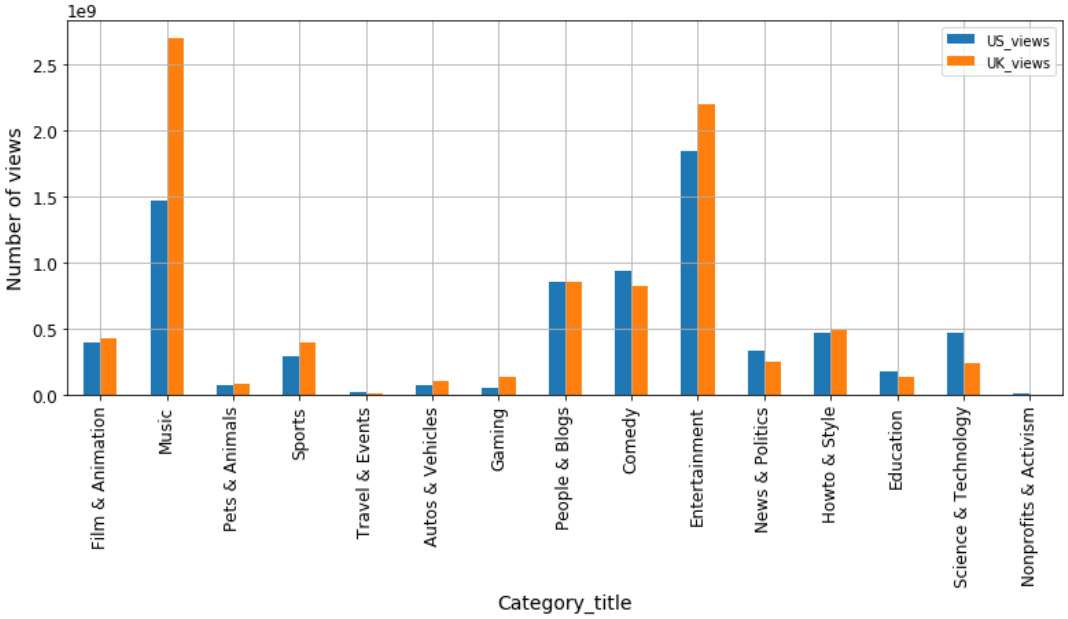
2.2.2. Proportion of YouTube views by video categories in the US:



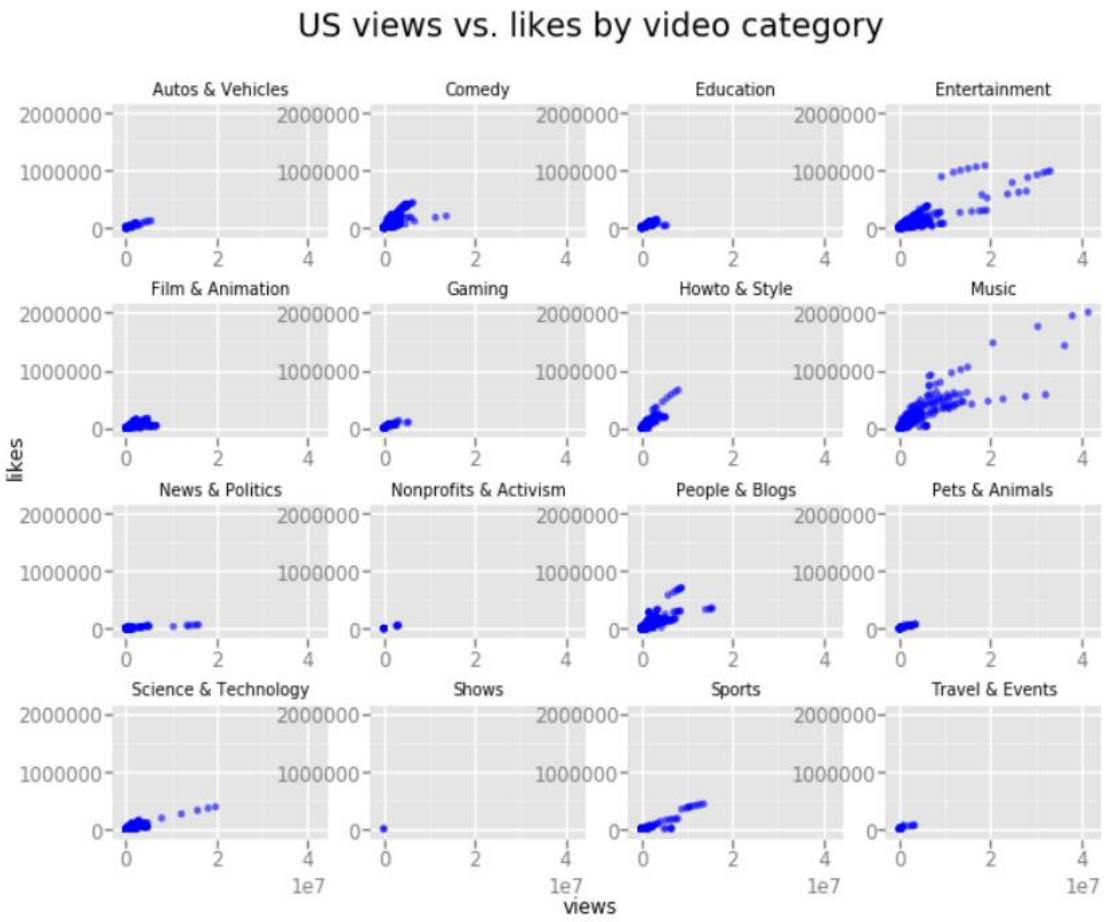
In the UK, there are a greater number of views in the categories of Music, Entertainment, and Sports. However, in the US, there are a greater number of views in the categories of Comedy, News & Politics, Education, Science & Technology.

2.2.3. Comparison of YouTube views in the US and UK:

Proportion of YouTube views by video categories in the US and UK:

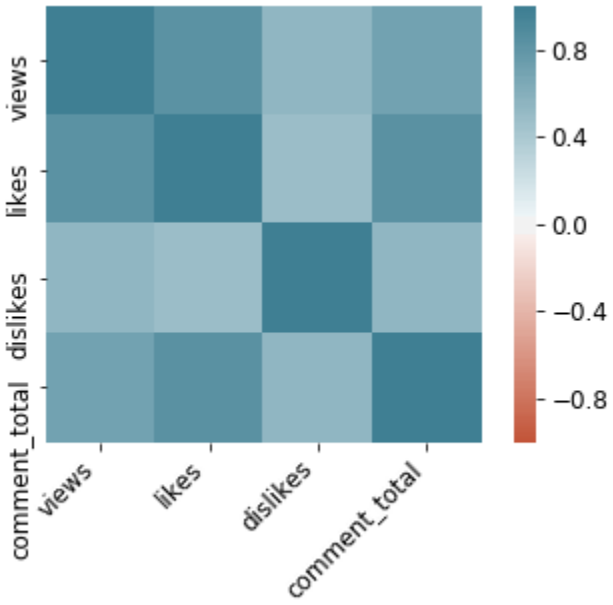


2.2.4. Category-wise scatterplots between number of views and number of likes in the US:



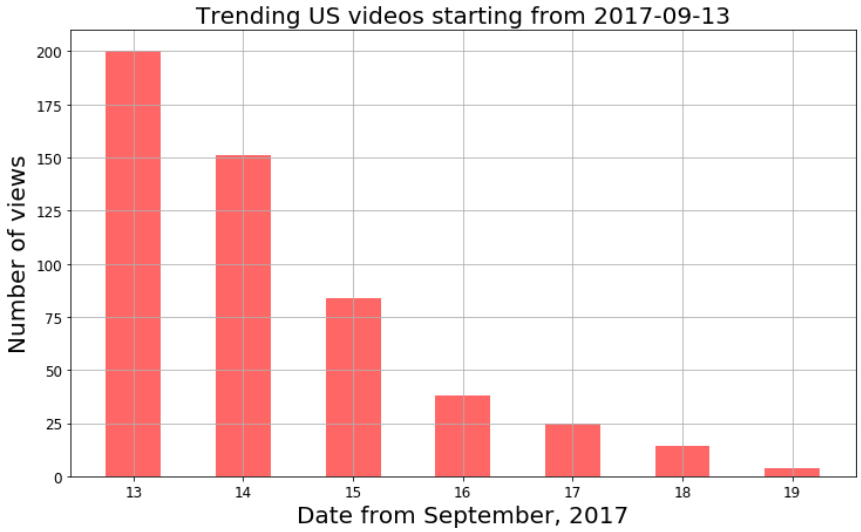
Category-wise scatterplots between number of views and number of likes show positive trends between number of views and number of likes. This means that the videos with higher number of views have higher number of likes.

2.2.5. Correlation coefficients among number of views, likes, dislikes, and comments:



It is found that the correlation coefficient between number of views and likes is 0.83, between number of dislikes and number of views is 0.54, and between number of comments and views is 0.72.

2.2.6. **Trending YouTube video views in the US**:



The above figure shows the number of views of particular videos that were high on September 13, 2017, declines rapidly by September 19, 2017.This shows that the number of views declines quickly over a week.

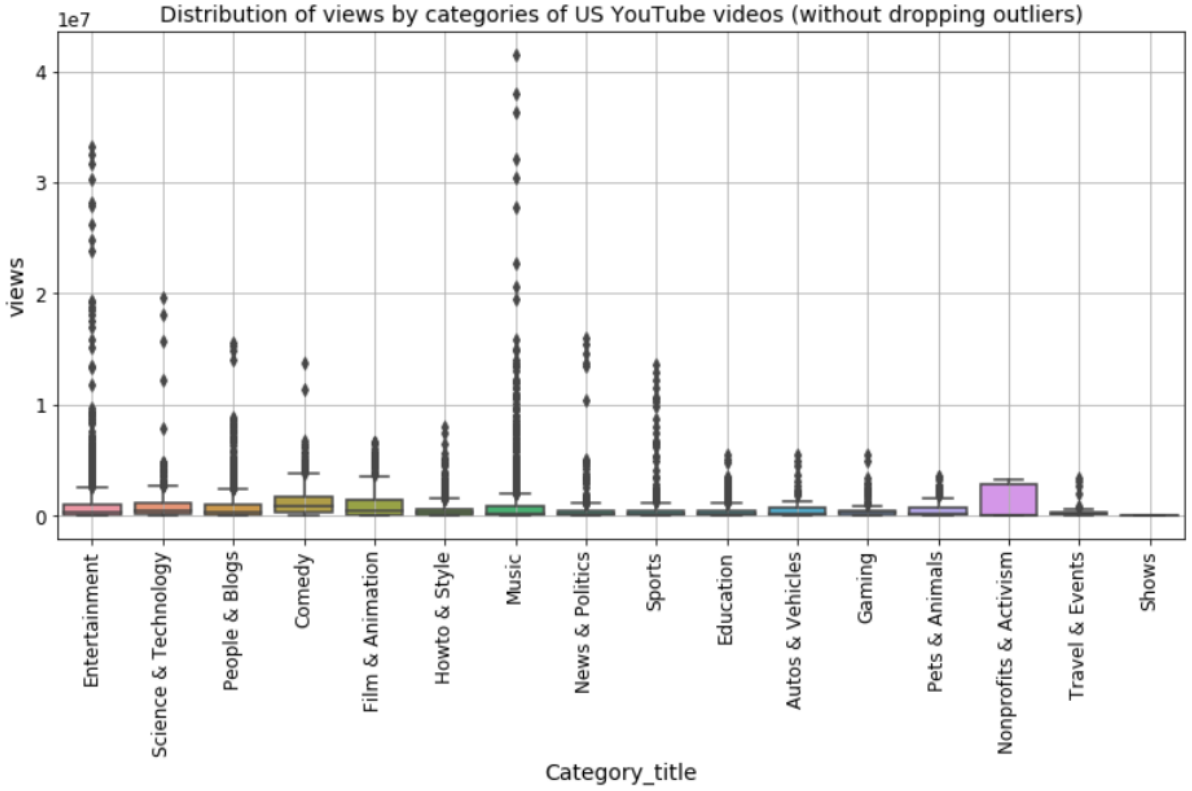
2.3. Baseline models:

2.3.1. Data processing for classification models:

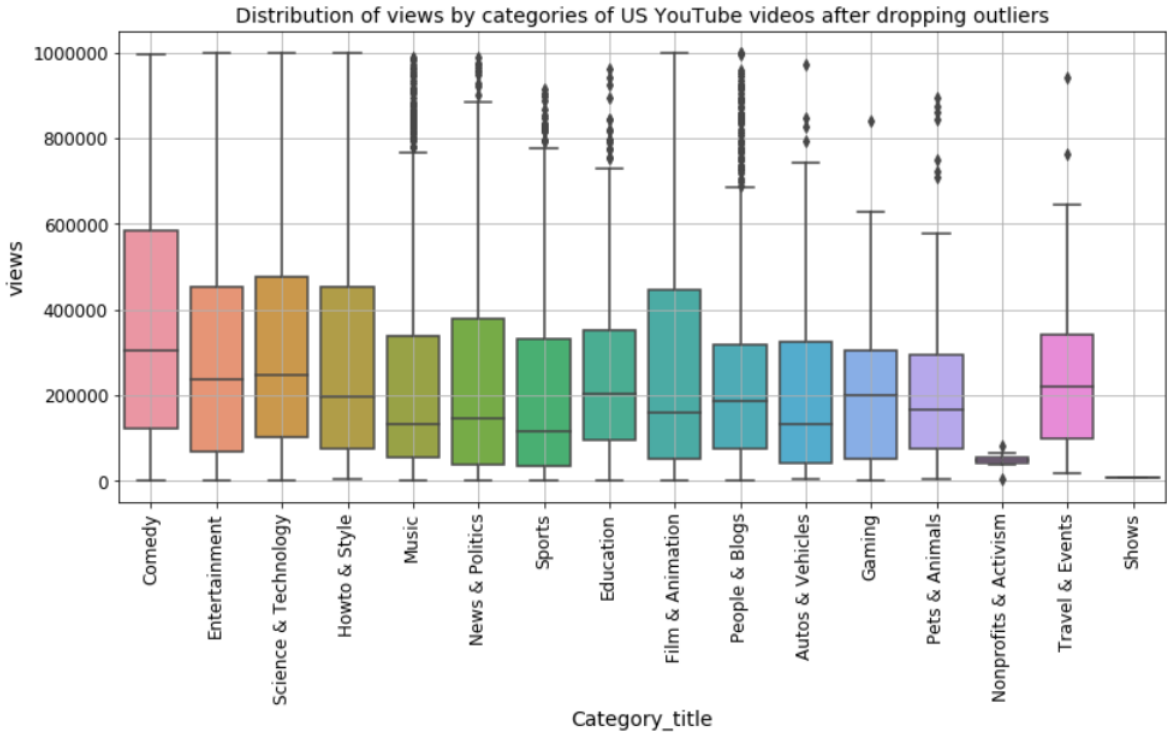
Following table shows input features (*channel_title*, *Category_title*, *views*, *likes*, *dislikes*, *comment_total*, *day*, *month*, *Weekday/weekend*) and output (*viewer_class*).

	channel_title	Category_title	views	likes	dislikes	comment_total	day	month	Weekday/weekend	viewer_class
6057	Life Noggin	Education	440393	14362	390	1575	Sunday	October	weekend	extreme
6058	Business Insider	News & Politics	55762	1265	760	1873	Sunday	October	weekend	low
6059	LP	Music	142908	7088	68	437	Sunday	October	weekend	medium
6060	YouTube FanFest	Entertainment	24532	2148	77	0	Sunday	October	weekend	low
6061	National Science Foundation	Science & Technology	144039	1574	59	0	Sunday	October	weekend	medium

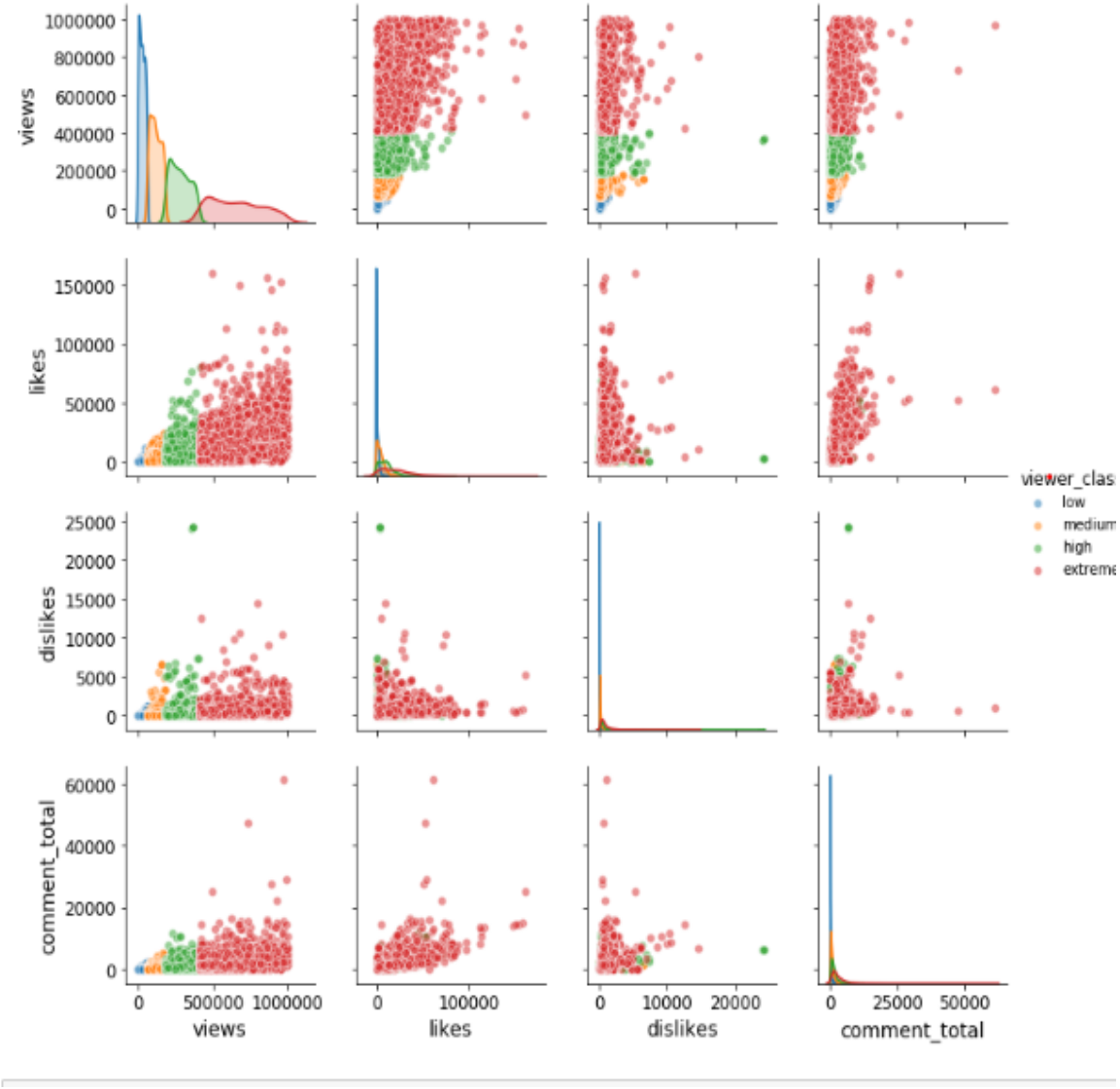
Following figure shows the distribution of number of views in the US before outliers are dropped. It is found that there are extreme views of music and entertainment categories.



After dropping outliers, we have following distribution of number of views in the respective categories.



The next figure shows a pairplot that shows four classes (low, medium, high, and extreme) with relationship to number of views, likes, dislikes, and comment_totals.



2.3.2. Label encoding for categorical features:

Categorical features, channel_title, Category_title, day, month, and weekday.weekend are label encoded. Following table shows that categorical features are label-encoded.

	channel_title	Category_title	views	likes	dislikes	comment_total	day	month	Weekday/weekend	viewer_class
6059	528	7	142908	7088	68	437	3	0	1	3
6060	1051	3	24532	2148	77	0	3	0	1	2
6061	655	12	144039	1574	59	0	3	0	1	3

2.3.3. Min-max scaling for numerical features:

MinMaxScaler, (data - min)/(max-min) scales 0 to 1. It preserves the same distribution and keeps the effect of outliers. Features, such as number of views, likes, and dislikes are scaled by min-max scaling from 0 to 1.

	channel_title	Category_title	views	likes	dislikes	comment_total	day	month	Weekday/weekend	viewer_class
0	535	1	0.666225	0.062138	0.012219	0.017506	6	1	0	0
1	211	1	0.859361	0.214606	0.029869	0.031285	6	1	0	0
2	75	1	0.452515	0.174560	0.016663	0.044868	6	1	0	0
3	923	1	0.258803	0.050314	0.012466	0.011867	6	1	0	1
4	580	3	0.274381	0.057346	0.019625	0.013698	6	1	0	1

2.3.4. Logistic regression:

Logistic regression with five-fold cross-validation and scikit-learn's GridSearchCV is utilized for classification of YouTube videos.

```
lr = LogisticRegression(random_state=42, multi_class="multinomial", max_iter=1000, solver="newton-cg")
parameters = {
    'C': [0.1, 1, 10, 100, 1000]
}

cv = GridSearchCV(lr, parameters, cv=5)
cv.fit(X_train, y_train.values.ravel())

print_results(cv)
```

2.4. Extended model:

2.4.1. Random Forest

Random forest is an ensemble machine learning model for classification and regression. It constructs multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

```
rf_cls = RandomForestClassifier(random_state=42, oob_score=True)
# Random search of parameters, using 5 fold cross validation,
# search across 100 different combinations, and use all available cores
rf_random = RandomizedSearchCV(estimator = rf_cls, param_distributions = random_grid, n_iter = 10, cv = 5, random_state=0)
```

3. **Summary of findings**:

   * YouTube videos in the US are classified based on their number of views.
   * Videos are labeled as "Extreme", of which the number of views is more than 75 percentile. Videos with number of views between the 50th and 75th percentile is labeled as "High". Similarly, "Medium" is in between the 25th and 50th percentile of views and "Low" is below the 25th percentile.
   * Label encoding refers to these classes as Extreme:0, High:1, Low:2, Medium:3.
   * Two machine learning models, Logistic Regression and Random Forest are utilized.
   * Dataset is randomly split into two segments: training dataset (80% of the dataset) and test dataset (20% of the dataset).
   * Model parameters are chosen by using training dataset with five-fold cross-validation.

* In the Logistic Regression model, the regularization parameter C is chosen by using grid search technique with cross-validation.
* In the Random Forest model, hyperparameters such as number of trees, maximum number of features, max. number of levels in each tree are chosen by using random search technique. Random search instead of grid search is used to reduce the running time to find these hyperparameters.
* Results demonstrate that random forest, with an accuracy of 74%, outperforms logistic regression, with an accuracy of 59%.
* In terms of precision, random forest has a precision of 83% for class 0 (extremely views video), while logistic regression has a precision of 75%. Precision is the ability to correctly diagnose positive features. In term of recall, for "extreme" videos, the recall is 66% by the linear regression and 86% by random forest. Recall is the ability of the model to retrieve relevant features.

Table: Comparison of performance between linear regression and random forest models

Performance measure	Classes videos	Linear Regression model	Random Forest model
Precision	Extreme	75%	83%
	High	48%	74%
	Low	64%	76%
	Medium	50%	65%
Recall	Extreme	66%	86%
	High	43%	61%
	Low	79%	82%
	Medium	49%	67%
F-1 score	Extreme	0.70	0.84
	High	0.46	0.67
	Low	0.71	0.79
	Medium	0.49	0.66

* Reasons that random forest performs better are:
  * Random forest is one of the most well-known ensemble models that combines a large number of independent decision trees and is trained over random and equally distributed subsets of a dataset.
  * Random forest models deal with overfitting by design.
  * Logistic regression works better where there exists a linear relationship. However, in the YouTube dataset the relationship between input features and output labels is non-linear, and hence the random forest performs better than the logistic regression.

4. Future works:

Neural networks can be used to build effective classification models. One way to extend the classification analysis is to apply neural network models and see whether it performs better than logistic regression and random forest classifiers.

5. Recommendation for clients:

By analyzing the YouTube dataset, it is found that music and entertainment videos are the most highly demanded video categories in both US and UK. The mean number of YouTube viewers in the UK is significantly higher than that of in the US. The videos that have higher number of views also have higher number of likes. In predicting classes of videos, "extreme" videos can be classified with more than 80% precision by the random forest model.