YouTube video views in the US and UK: Exploratory data analysis, statistical inferences, and classification

R. S. Elias

January 2020

Abstract

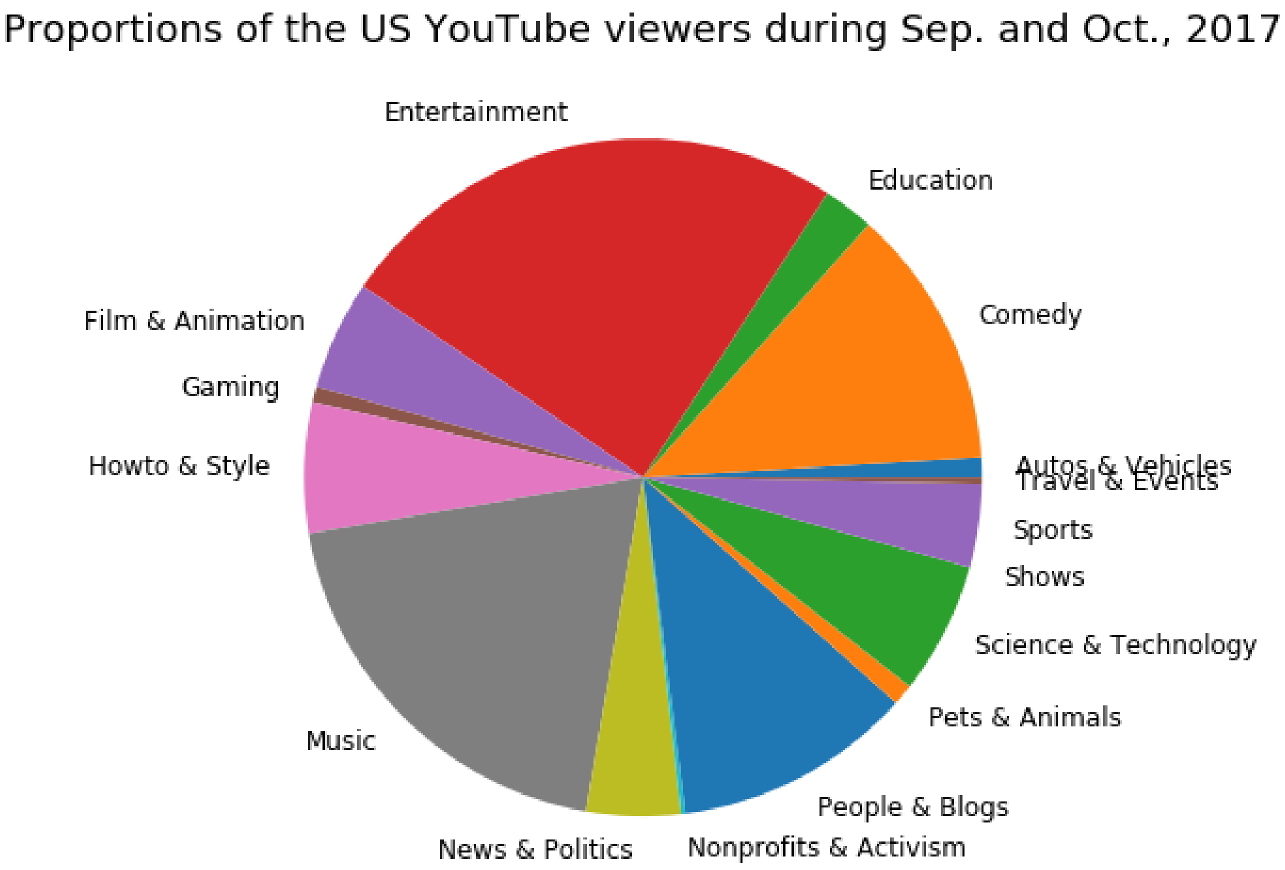
- This project analyzes YouTube video dataset from the US and the UK. There two parts in the project: exploratory data analysis and classification.
- Exploratory data analysis includes data loading, cleaning, joining of dataframes, and visual investigations.
- Classification of YouTube US viewers includes feature engineering, label encoding, scaling, and applying and comparing two machine learning models -- logistic regression and random forest classifier.

Dataset description

- 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
- 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]. There are 16 unique category_id.

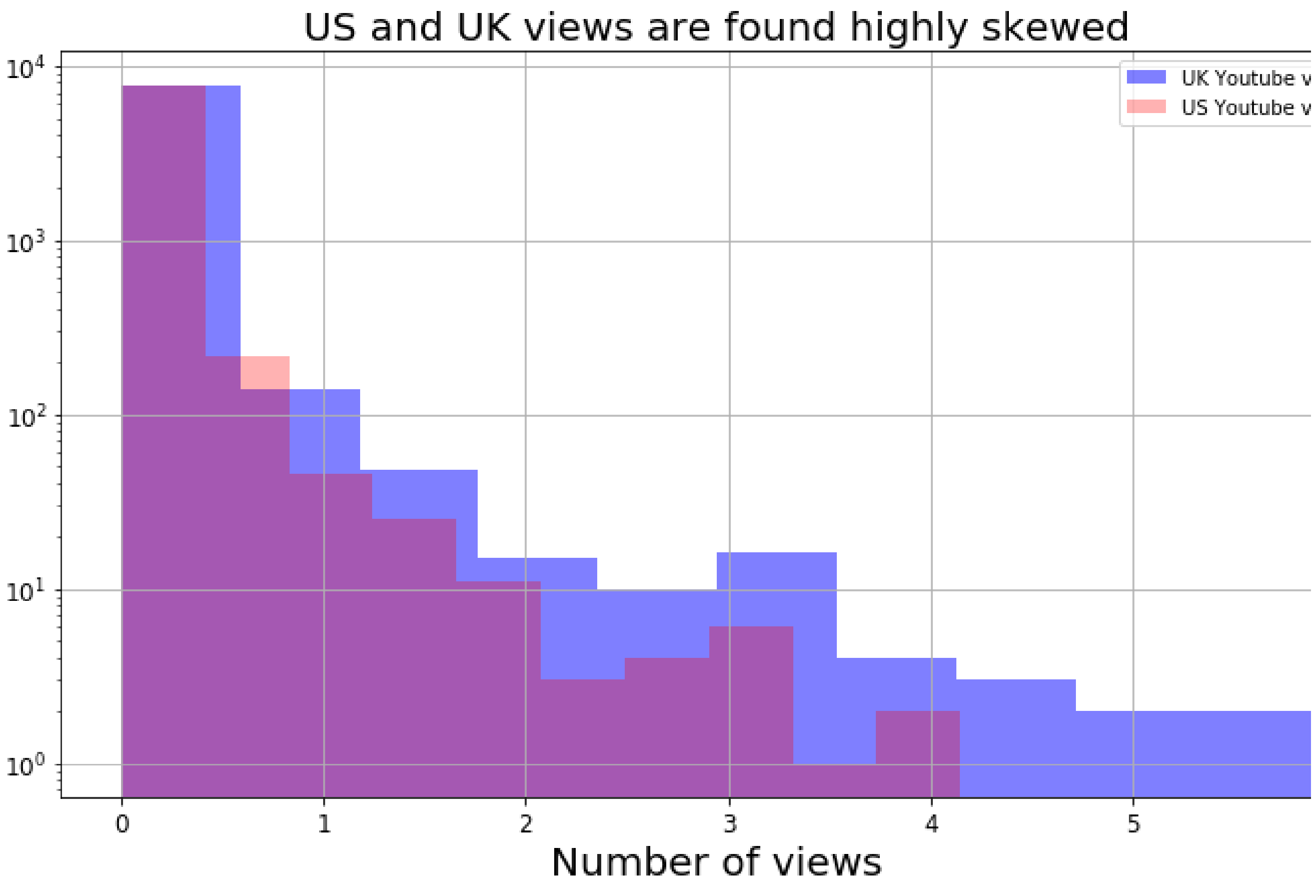
Exploratory data analysis

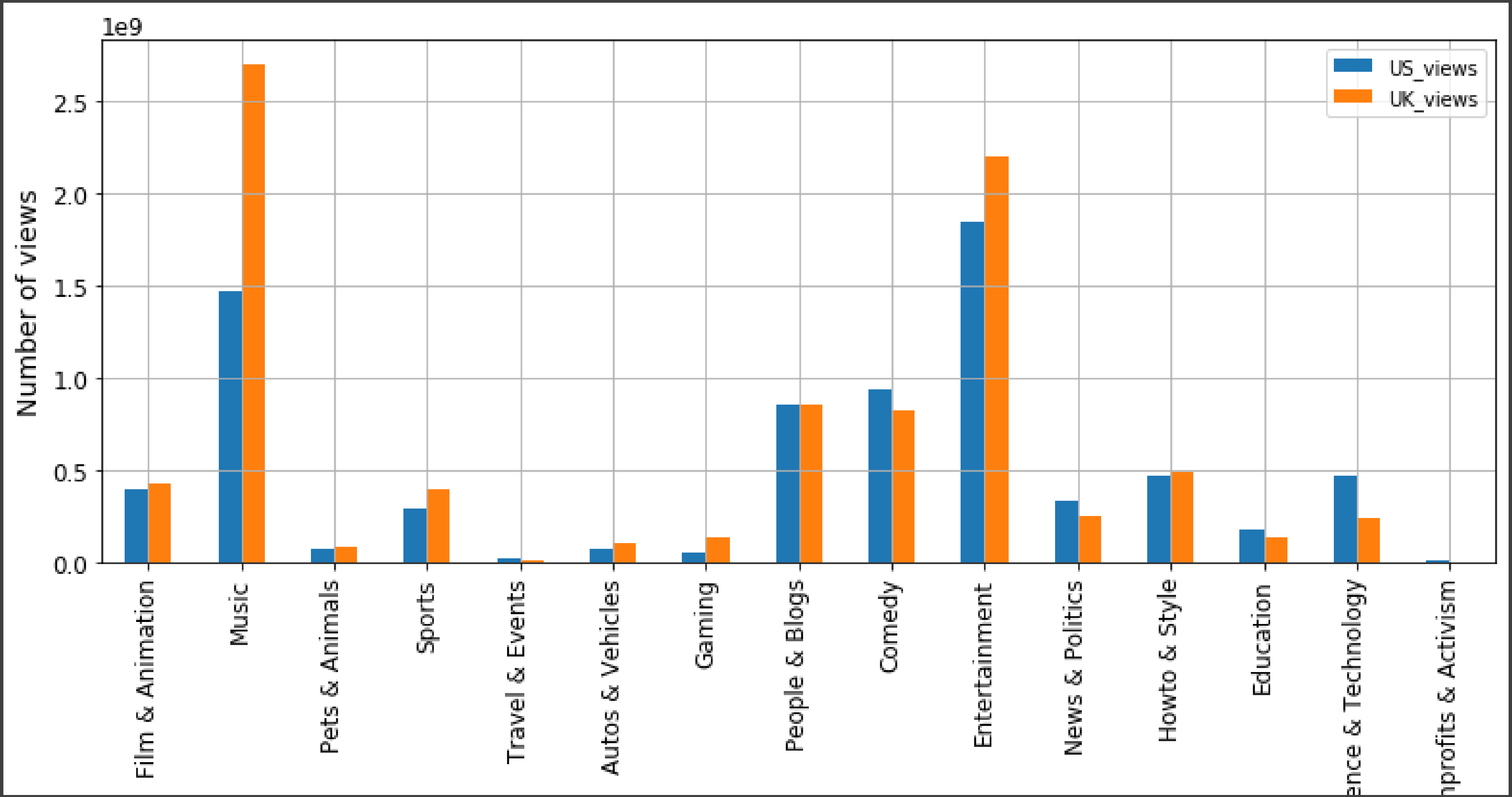
- In the UK, there are a greater number of views in the categories of Music, Entertainment, and Sports.
- However, in the US, there are a greater number of views in the categories of Comedy, News & Politics, Education, Science & Technology.



Comparison of distributions of number of views in the US and UK

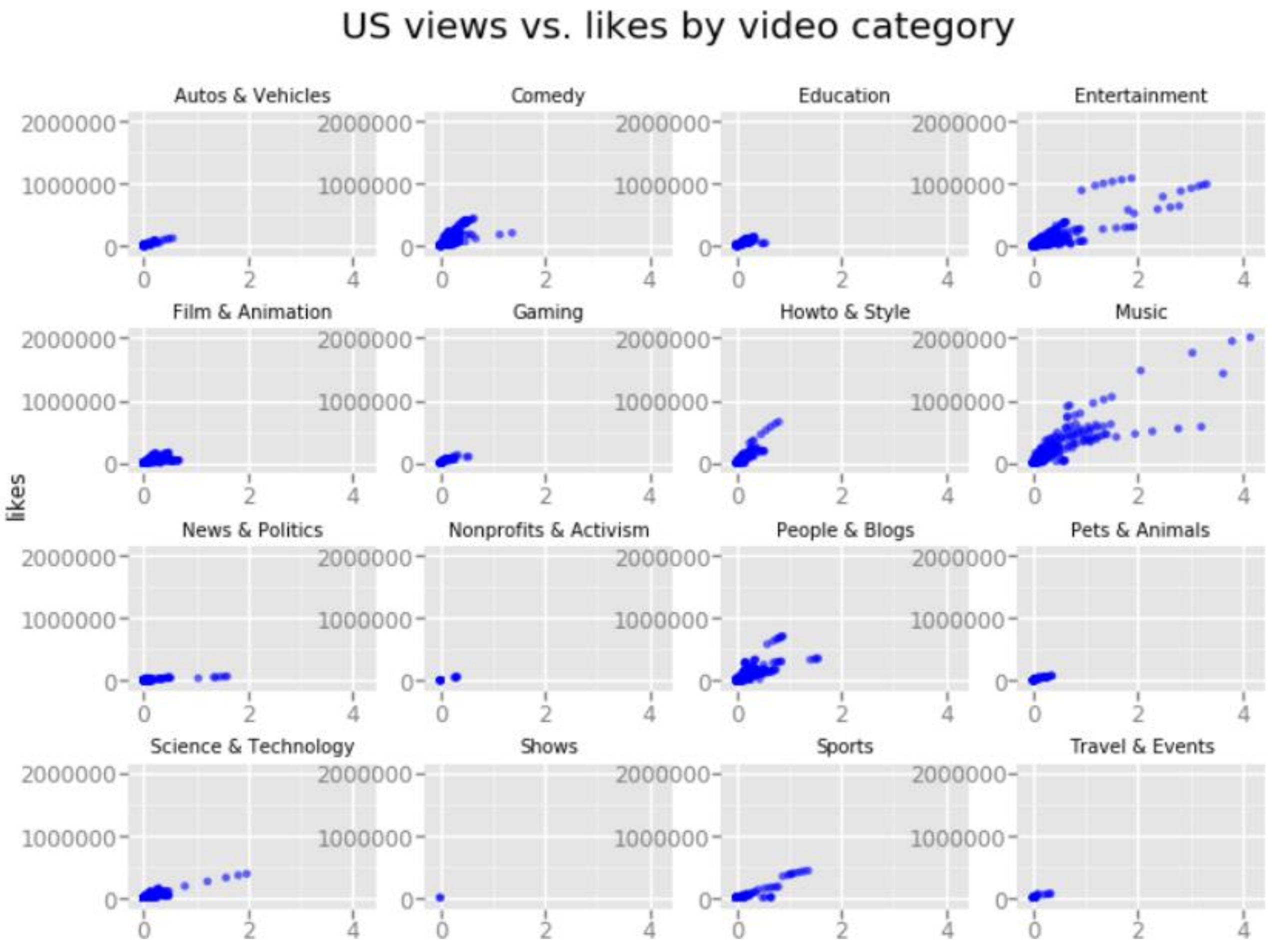
- Distributions of both US and UK views are right-skewed.
- Because, few videos are viral with extremely high number of views.




1e9
Number of views
0.0
0.5
1.0
1.5
2.0
2.5
US_views
UK_views
Film & Animation
Music
Pets & Animals
Sports
Travel & Events
Autos & Vehicles
Gaming
People & Blogs
Comedy
Entertainment
News & Politics
Howto & Style
Education
ence & Technology
nprofits & Activism

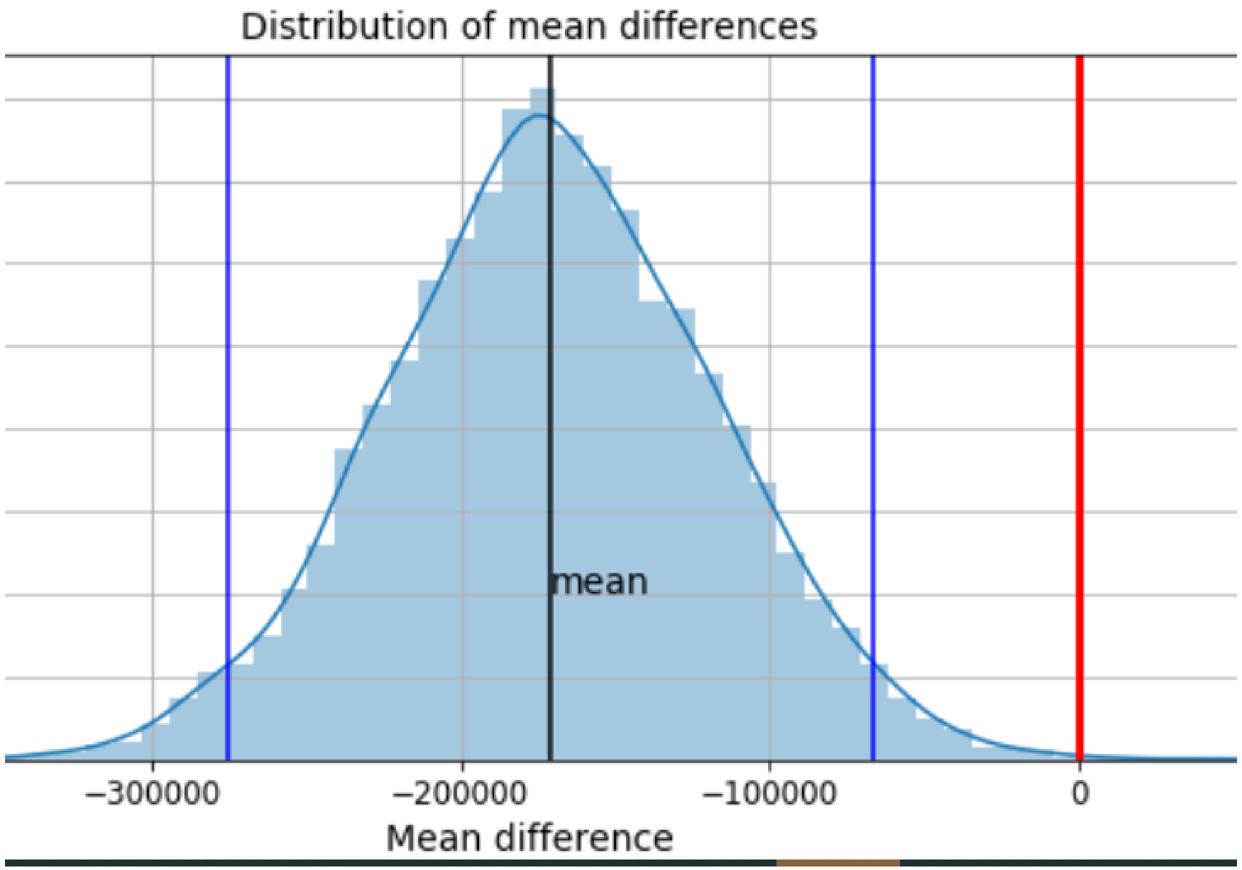
Category-wise scatterplots between number of views and number of likes in the US

- 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.

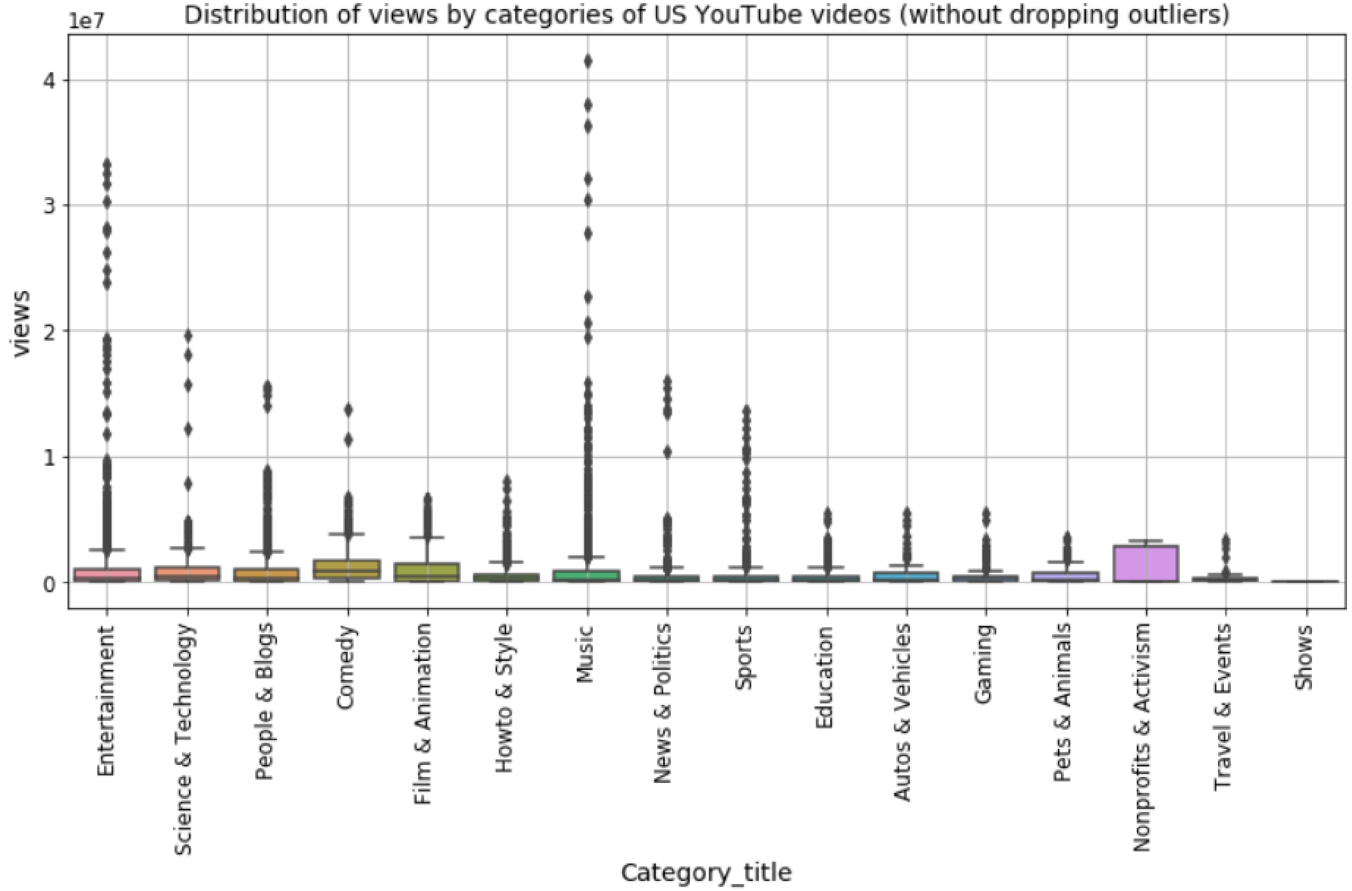


Statistical inference

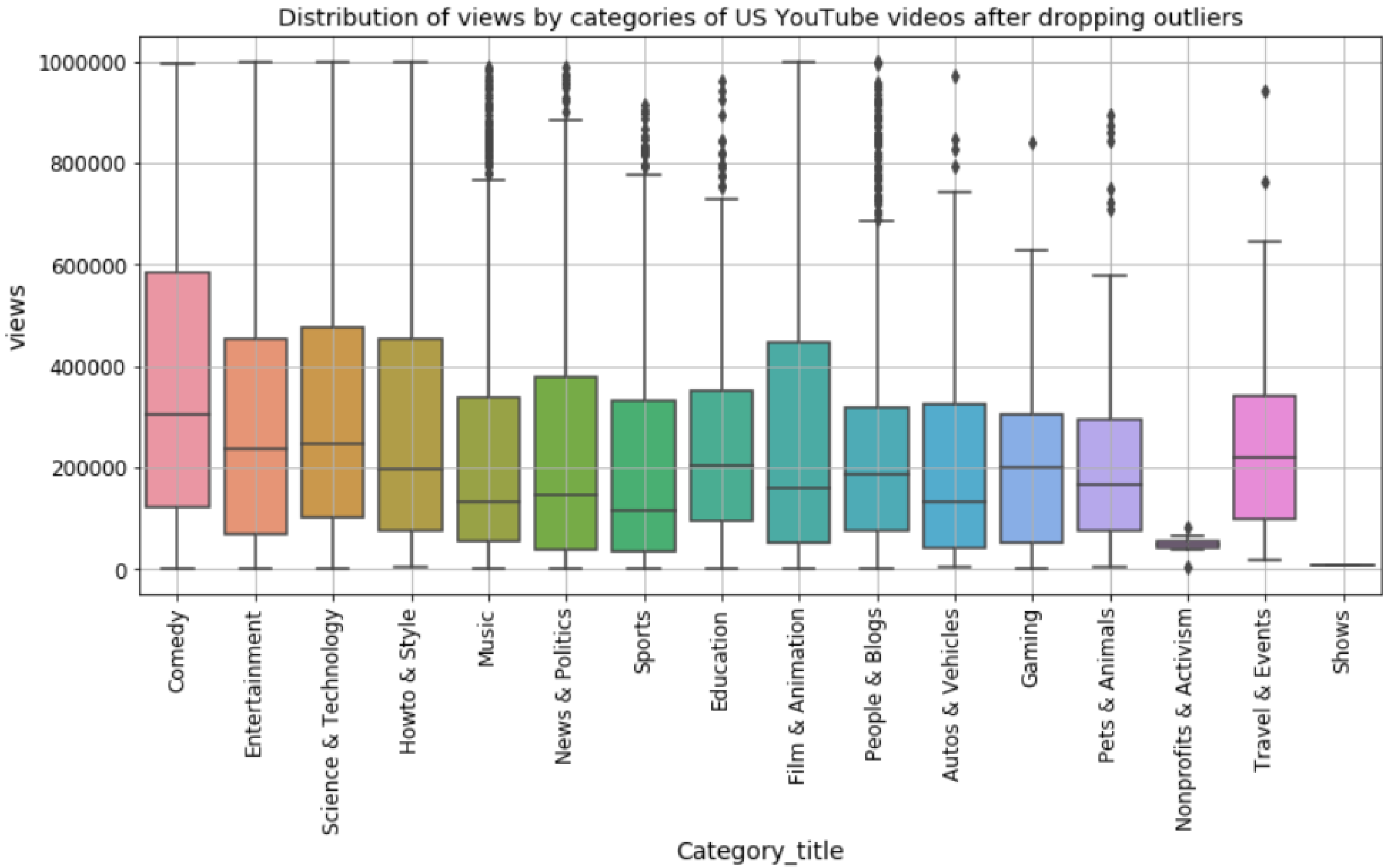
- From the figure, we observe that mean difference of 0 is outside the 95% confidence interval and falls into the zone of rejection of the null hypothesis.
- We conclude that the mean number of views in the US and UK are statistically significant at 5% confidence level.
- Negative difference shows that number of Us views are less than the number of UK views.



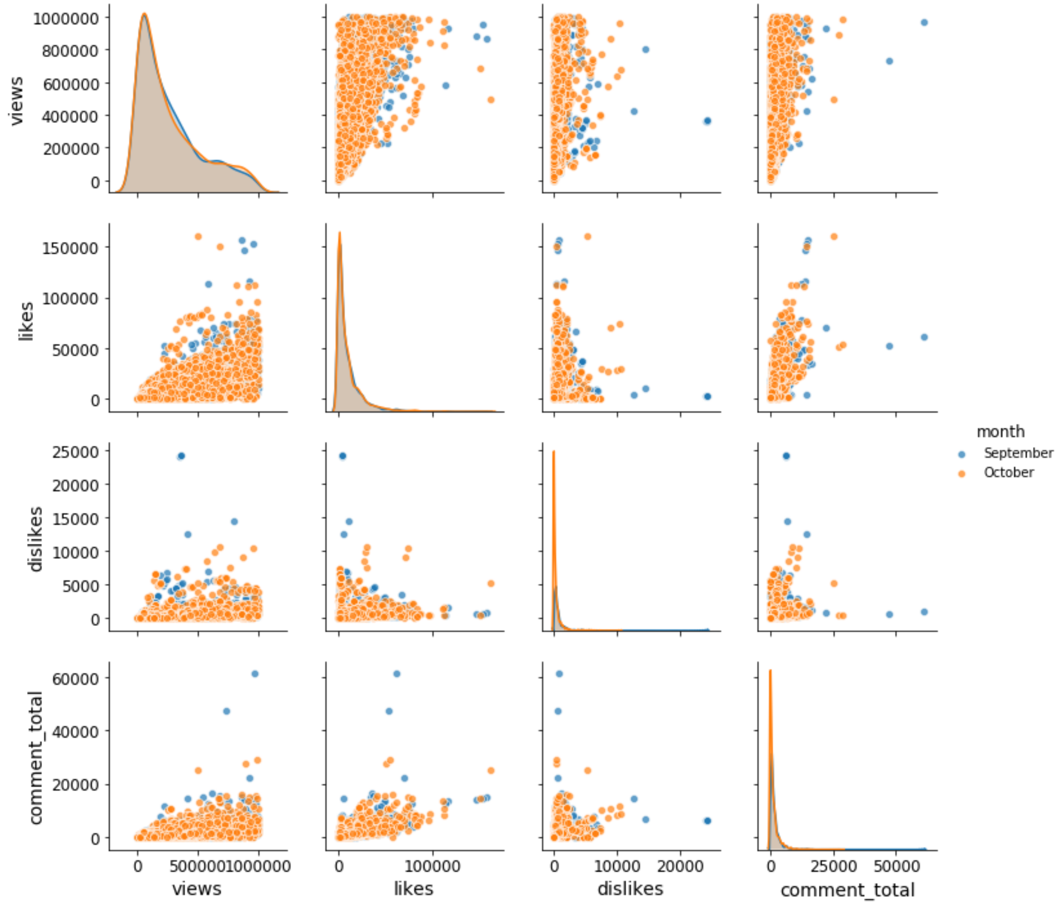
US YouTube views distribution by video category (without dropping outliers)



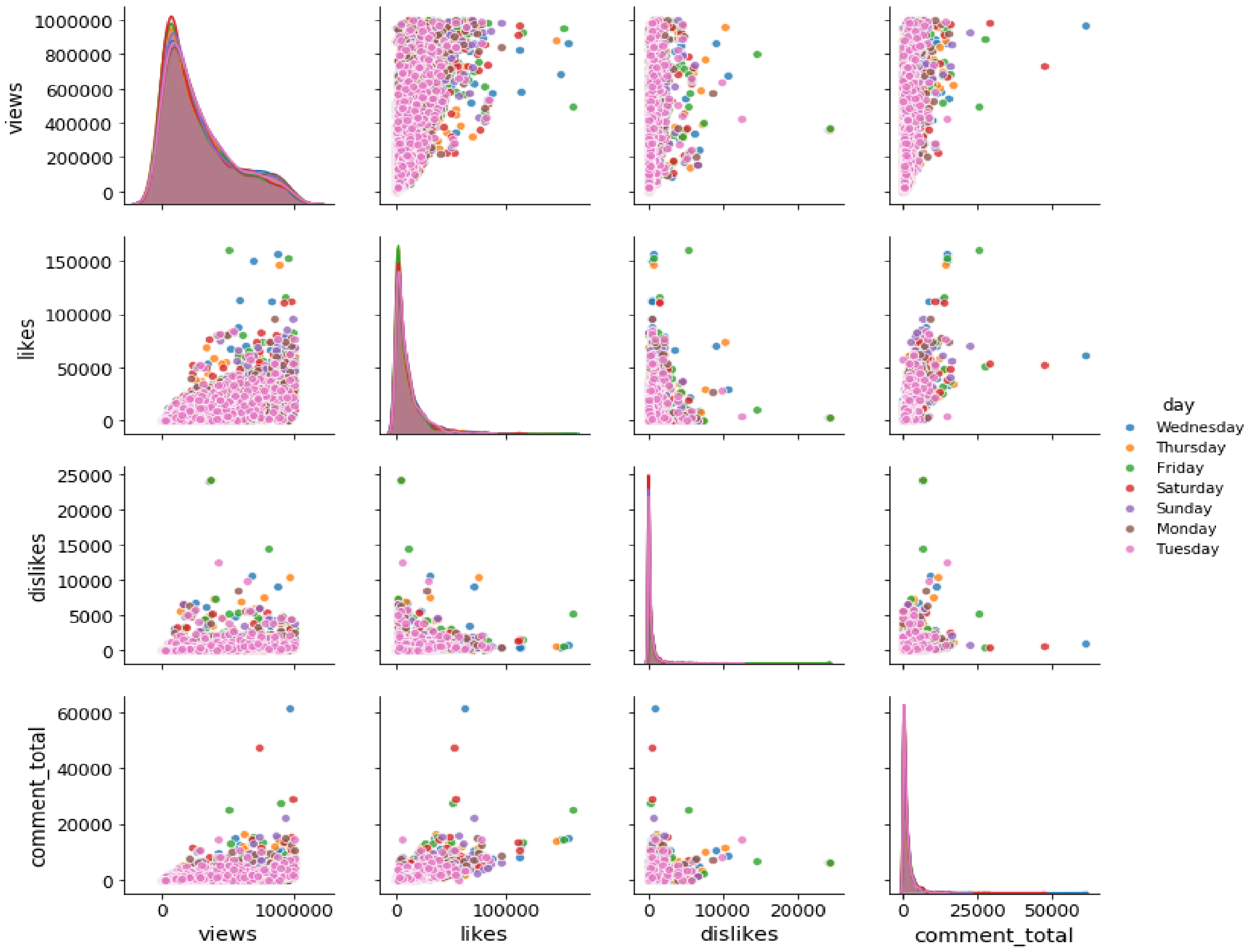
YouTube views distribution by category (after dropping outliers)



Feature engineering: Effect of month (September or October) on number of views, likes, dislikes, and comments

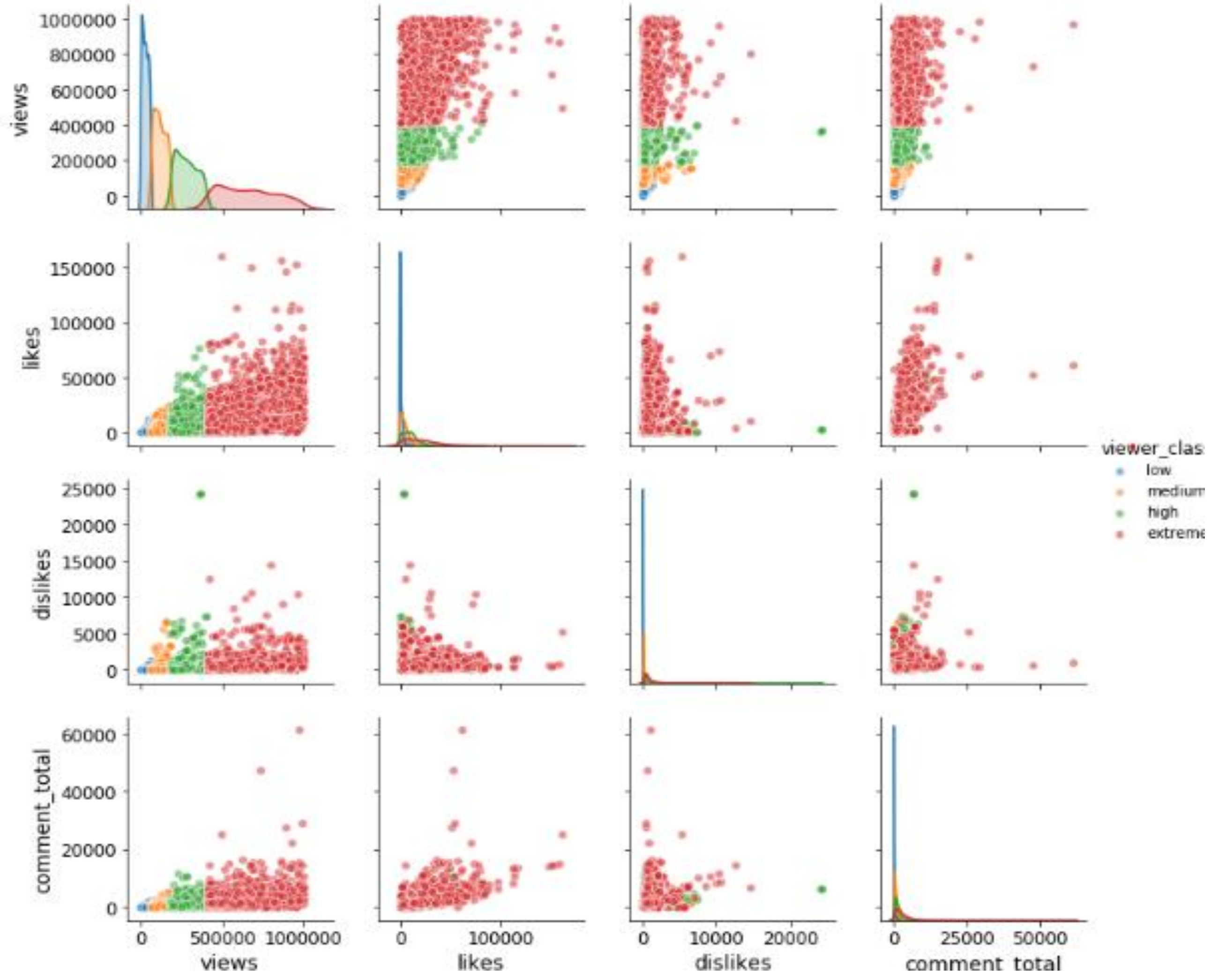


Effect of days (Monday, Tuesday, etc.) on views, likes, dislikes, and comments

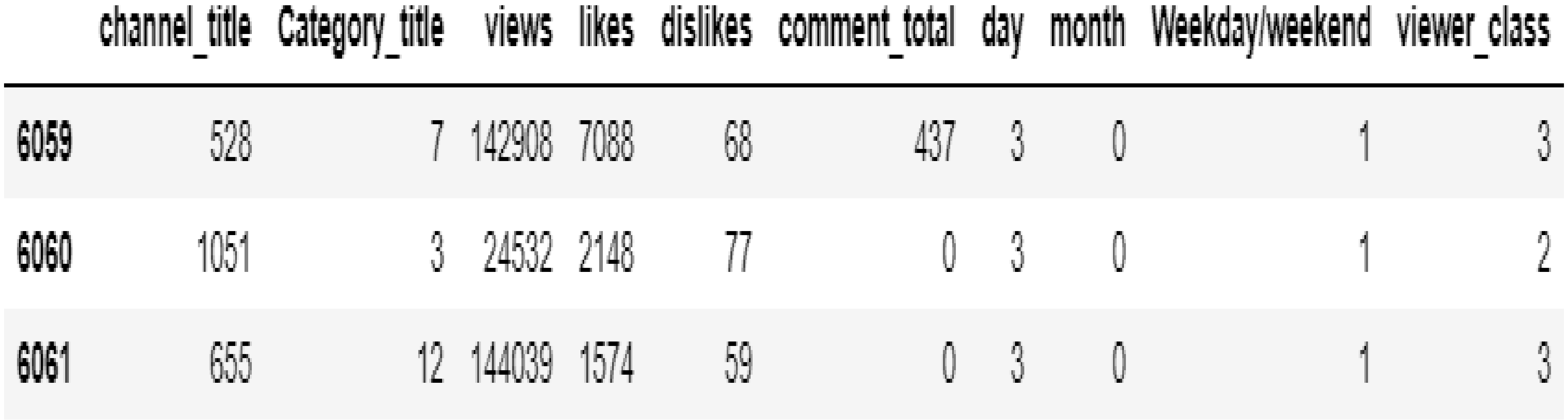


Four classes of video views

- Extreme (over 75th percentile)
- High (over 50th percentile)
- Medium (over 25th percent)
- Low (below 25th percentile)

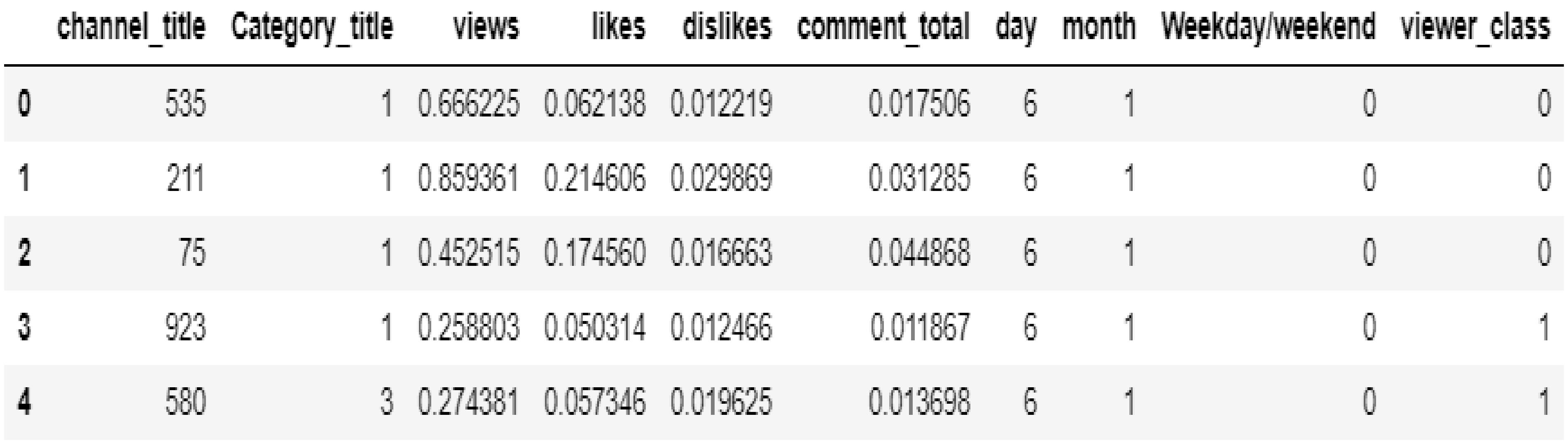


Label encoding for categorical features



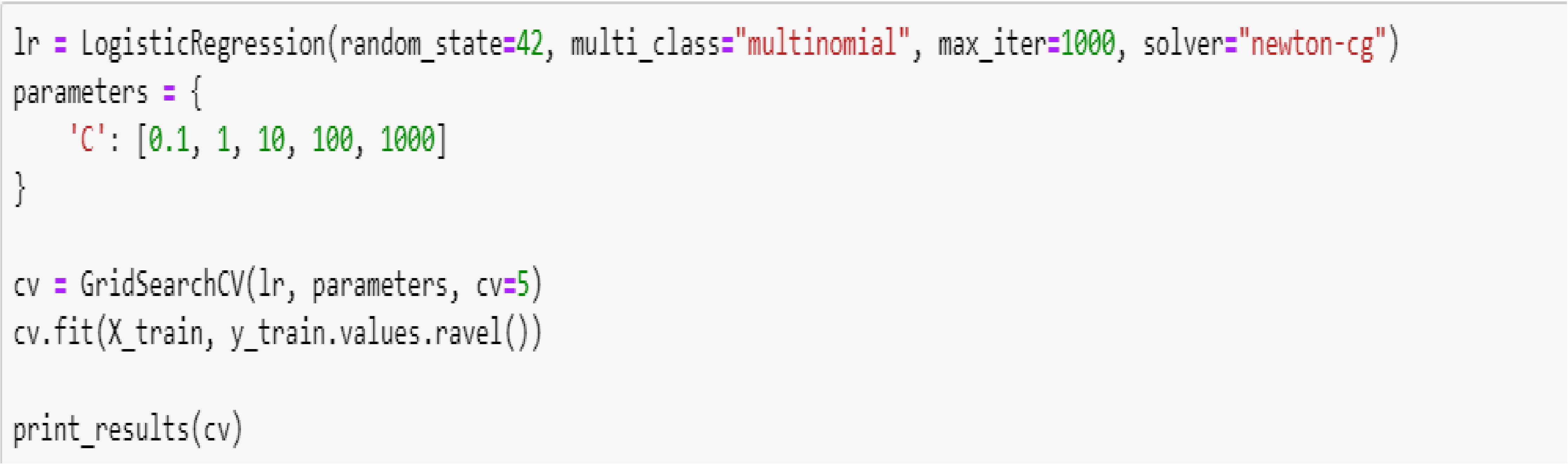
	channel_title	Category_title	views	likes	dislikes	comment_total	day	month	Weekday/weekend	viewer_class
6059	528	7	142908	7088	68	437	3	0	1	3
6060	1051	3	24532	2148	77	0	3	0	1	2
6061	655	12	144039	1574	59	0	3	0	1	3

Min-max encoding for numerical features



	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

Logistic regression with 5-fold cross-validation

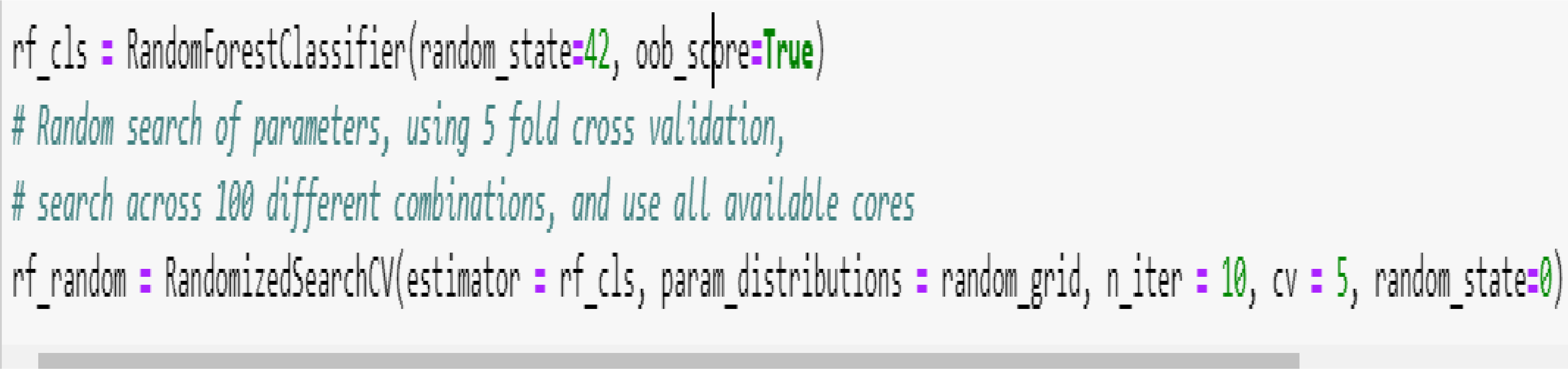


```
lr = LogisticRegression(random_state=42, multi_class="multinomial", max_iter=1000, solver="newton-cg")
parameters = {
    'C': [0.1, 1, 10, 100, 1000]
}

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

print_results(cv)
```

Random forest classifier with 5-fold cross-validation



```
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)
```

Comparison of 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

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.

Summary of findings

- 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.

Summary of findings

- 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.

Summary of findings

- 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.

Future works and recommendations for clients

- 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.
- By analyzing the YouTube dataset, it is recommended 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.