SC1015 Mini Project: YouTube Insights

Done by:
Aaron Ang (U2122736A)
Deuel Teo (U2222284J)
Russell Lim (U2221750L)

Table of Contents

Problem Formulation

Motivation of our Project
Intended Outcome of Project

Core Analysis & Machine Learning

Further analysis and Machine Learning techniques utilised

Exploratory Data Analysis

Data extraction, preparation, cleaning and analysis

Summary & Conclusion

Conclusive findings and insights

Introduction: Problem Formulation

About YouTube

- Most popular and widely used video hosting website.
- More than 2.6 billion users as of 2023 with nearly 122 million users daily.
- Over 51 million channels as of 2023 for users to view

Viral Videos

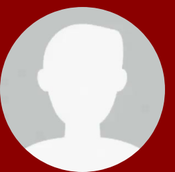
- Videos that spread rapidly through the internet
- Gains popularity and attention with users
- Videos that gain widespread attention may bring attention to channels in turn

What we wish to discover

What can YouTube Channels do to ensure that videos go viral?

Datasets Chosen

- Dataset containing metadata for Channels and Videos on YouTube, over 400K data samples
- Contains metrics such as likes, comments, views and subscribers and genre, etc.
- Additional CSV for video category included.



Exploratory Data Analysis

Preparing the Data

- Handling missing values (category id got no title) by merging two different datasets in the same github repository
- Creating ‘date difference’ time series data to represent difference between original published date and current date
- Made sure that the dataset does not have null values and that it is all unique values
- We used pre-processing min-max scalar to fit the response variable (videoViewCount) and the predictor variable (dateDifference)

What we wish to find out

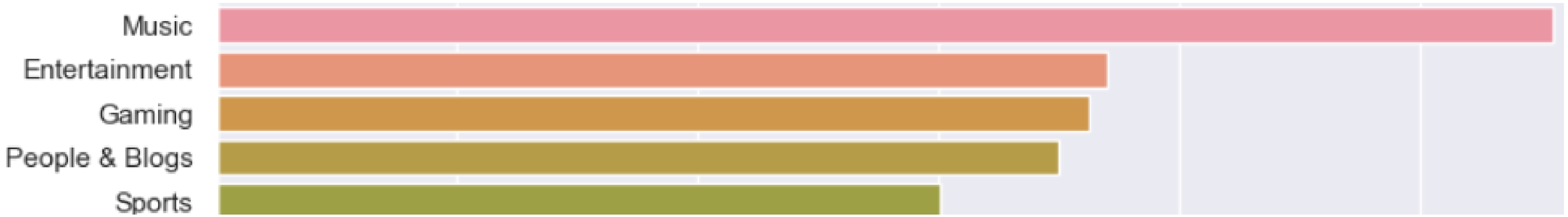
- We wanted to explore the relationship between genres and videos, and how the element of time affects the view counts of videos or channels.

We will do this by:

1. Exploring how many videos are there in each Genre
2. Exploring the new dataset that was created including the 'dateDifference' variable
3. Exploring dataset based on time series

1.1 Highest number of videos according to category

Top 5 Categories:



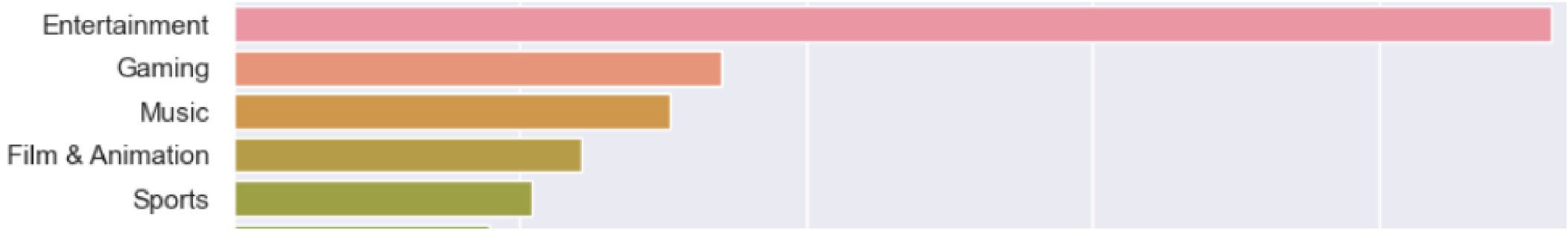
Analysis:

- It seems that youtube videos are generally produced in music, entertainment and gaming category.
- Thus, it appears that youtube videos are mainly used for people to enjoy music, entertainment and watch games.

However, the total number of videos is not an indication of how many views a video will have.

1.2 Highest views according to category

Top 5 Categories:



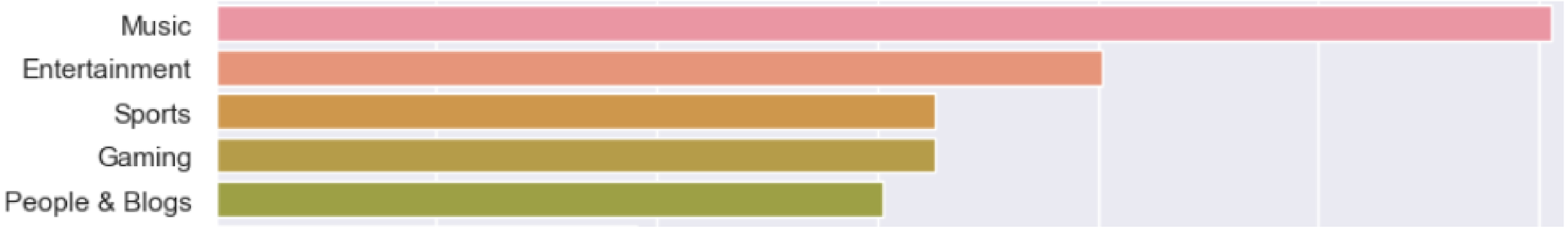
Analysis:

- Thus, it appears that youtube videos that generate the most views entertainment, gaming and music videos.

However, are there any other factors that might allude to people watching videos?

1.3 Highest elapsed time according to category

Top 5 Categories:



Analysis:

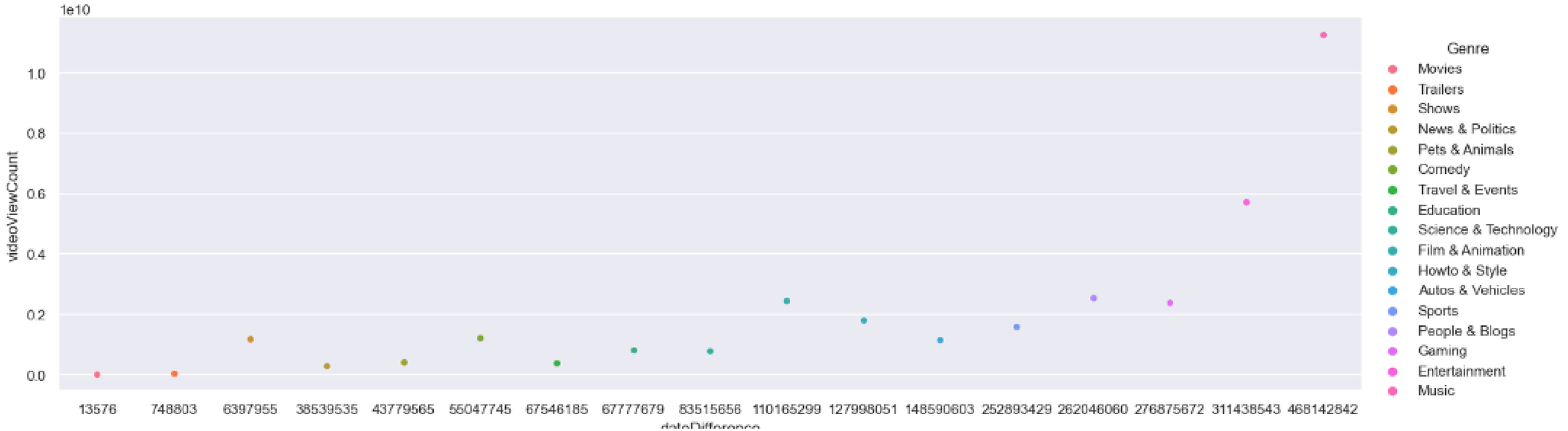
- Thus, it appears that youtube videos that generate the highest elapsed time are Music, Entertainment and Sports.

Conclusion:

We can see that in all the variables explored, Music, Entertainment and Gaming are consistently at the top of the categories, so content creators that want a video to go viral would have to create videos in the following genres:

- Music
- Entertainment
- Gaming

2.1 Highest difference in date published according to category



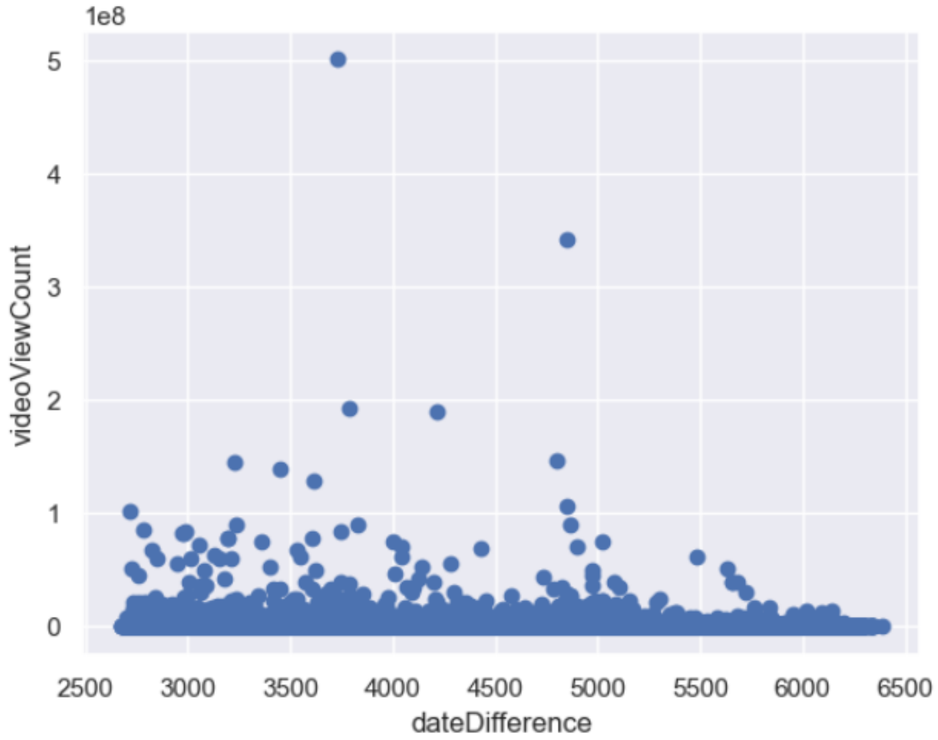
Analysis

This shows us that there might be a correlation between the longer a video is posted to whether a video will go viral as the top genres here:

- Music
- Entertainment
- Gaming

This relates to the top videos in number and view count.

2.2 Plot data based on videoViewCount and dateDifference



We can see that the clustering between the videoviewcount and dateDifference is pretty much the same. This tells us that there is actually no much correlation between datedifference and videoviewcount

Analysis:

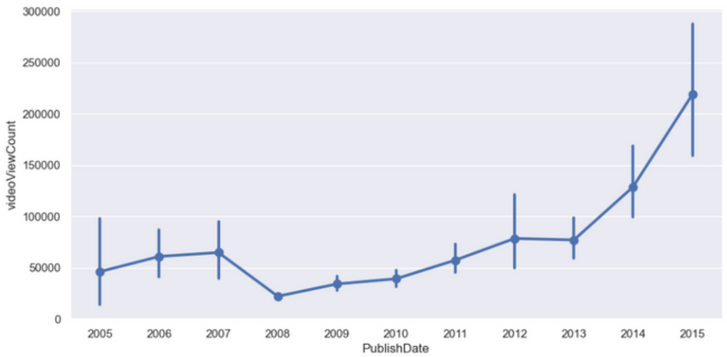
- We have see that the longer a video is posted for does not actually have any correlation with whether it will go viral.

Conclusion:

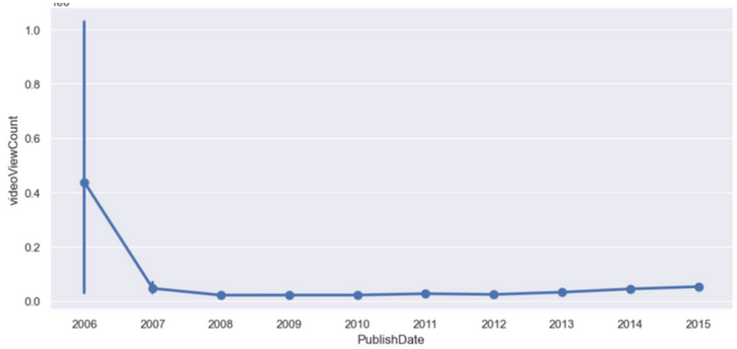
- Thus, we will have to find other factors of how a video might go viral.

3.1 Exploring the trend of the view of videos according Genre

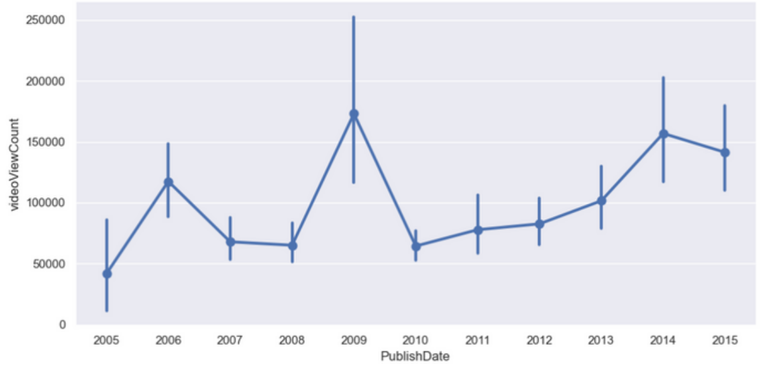
Entertainment



Gaming



Music



We can see that the trend for the top genres:

- Music
- Entertainment
- Gaming

Other than gaming which has stayed mostly consistent, the other top genres will most likely continue to rise in the years to come.

Analysis:
We have see that the genres that are top in videoviewcount are not only top in view count but also have a rising trend in the years.

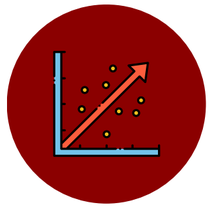
Conclusion:
Thus, the videos posted in those genres with a rising trend will most likely be viral in the years to come.

Top video categories

The top video categories correlated to the subscriber count, channels views and trend are **Gaming**, **Music** & **Entertainment**

Core Analysis & Machine Learning

Regression Techniques Used



Linear Regression

Standard Prediction of variable value with another variable

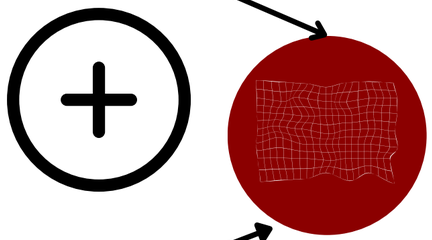
Regularisation

Lasso Regression

LR Equation = LeastSquare Objectives + α * (sum of the absolute value of coefficients)

Ridge Regression

L2 Penalty



Elastic Net Regression

Combining Lasso Regression (L1 Penalty) and Ridge Regression (L2 Penalty) to regularise the Linear Regression

Analysis Optimal Regression Method

Regression Methods	Linear	Lasso	Ridge	Elastic-Net
R^2 (Train)	0.76740	0.76050	0.77115	0.76605
R^2 (Test)	0.75594	0.77099	0.74523	0.75826
Difference	0.01146	-0.01049	0.02592	0.00779
Difference (%)	1.49	-1.38	3.36	1.02

From above table, we can observe that out of the 4 regression methods used, Elastic-Net Regression produce a result that that is most minimal difference (of 1.02%) between accuracy score for the trained model and the test model.

This is inline with our theoretical findings that Elastic-Net being the combination of Lasso and Ridge regression is the optimal regression method.

Clustering/Classifications Techniques Used

K-Means

Clustering method that aims to divide data points into K clusters in a way that the sum of the squared distance between the objects and assigned cluster is minimised.

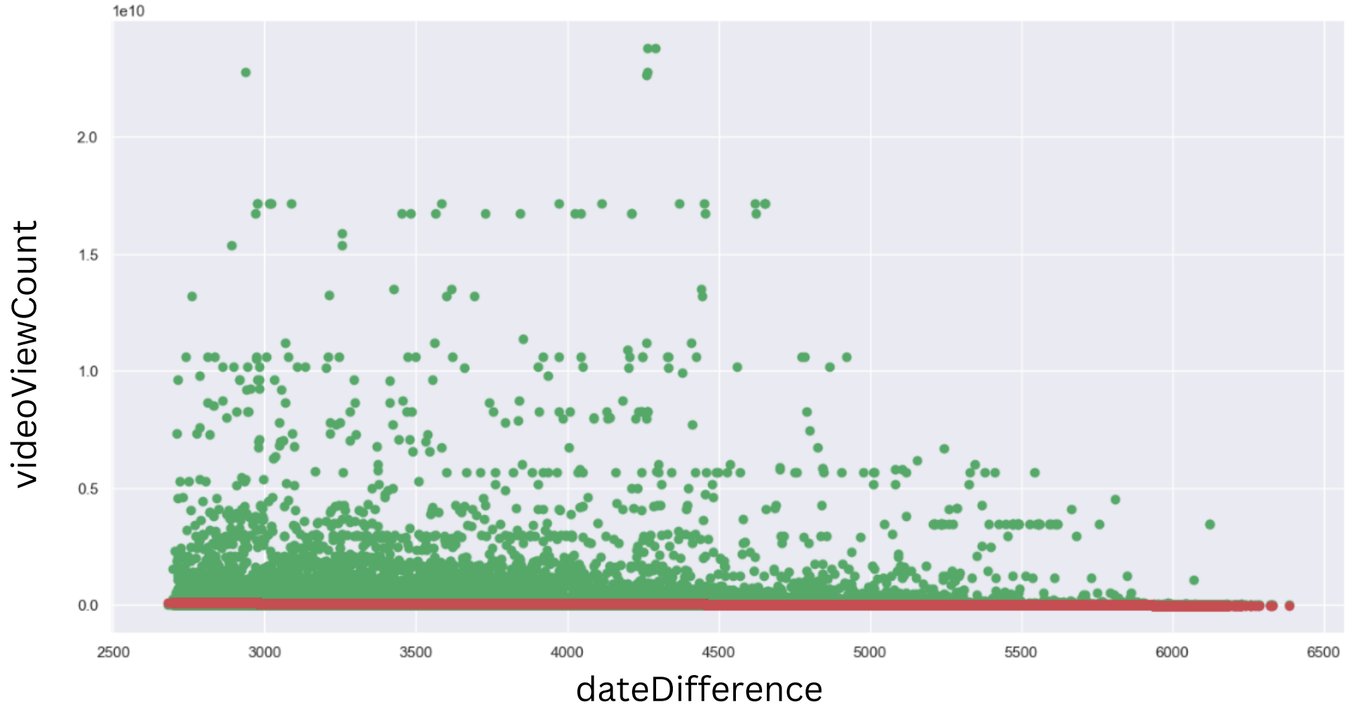
Helper Functions/Charts
MinMax Scaler (transforming data)
Elbow Plot (finding the optimal K value)

K-Nearest-Neighbours

Uses proximity to make predictions of what class an individual data point belong to

Helper Functions/Charts
Classification Report (F1-score, Precision & Recall)

Regression between videoViewCount and dateDifference




```
Explained Variance (R^2) of model       : 0.0026407392357435944
Mean Squared Error (MSE) of model       : 2.1161028468199376e+17
Root Mean Squared Error (RMSE) of model : 460011178.86633337

Explained Variance (R^2) of test        : 0.0021692447455570196
Mean Squared Error (MSE) of test        : 2.261252504236057e+17
Root Mean Squared Error (RMSE) of test  : 475526287.8365461
```

Initial Expectations

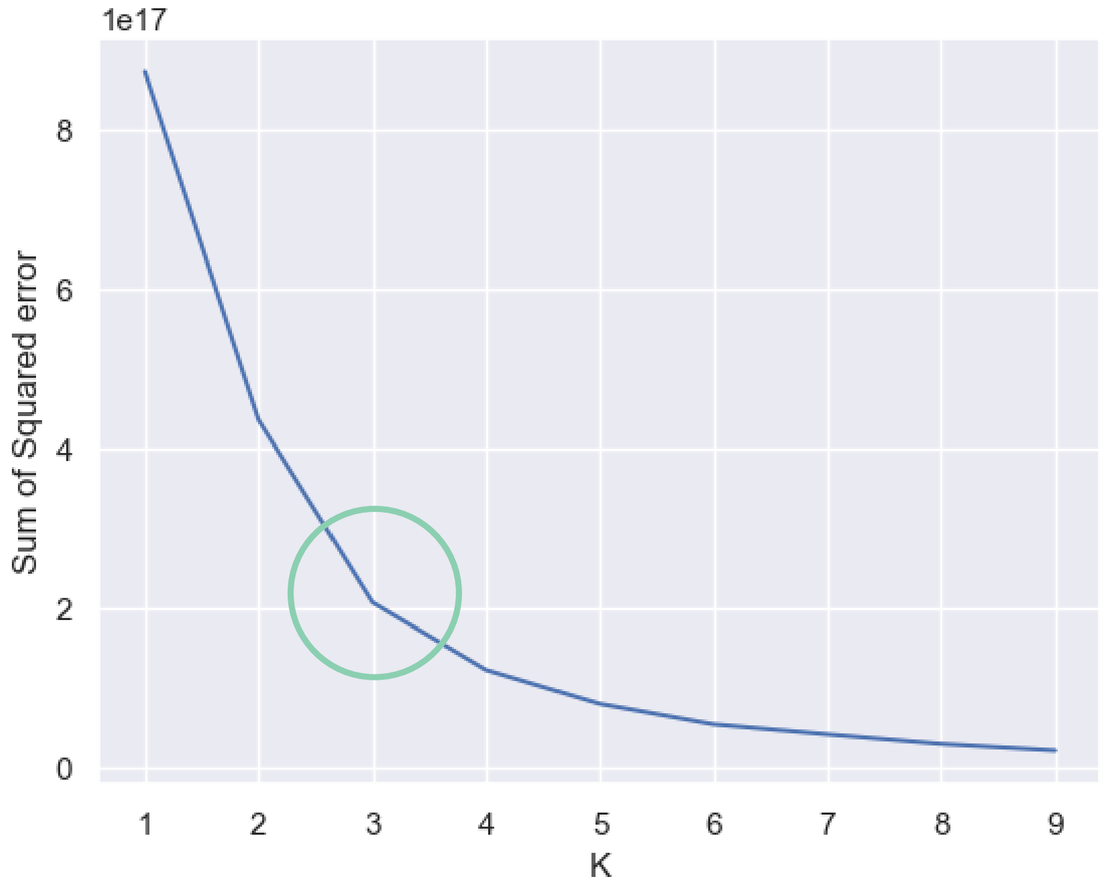
Before performing the linear regression between videoViewCount and dateDifference, we anticipated that the longer video is on YouTube, the more people would have watched the video.

Results

However, from the regression done between the predictor and response variables, we have realised that interestingly the element of time does not play a strong factor in popularity of videos, in particular YouTube.

(K-Means) Elbow Plot for Time Series vs videoViewCount

```
SSE = []
kRange = range(1,10)
for k in kRange:
    km = KMeans(n_clusters=k)
    km.fit(df[['videoViewCount','dateDifference']])
    SSE.append(km.inertia_)
```



Description of Elbow Plot

The elbow plot as what its name represents (the minimum point of the curved region) is the optimal K-value.

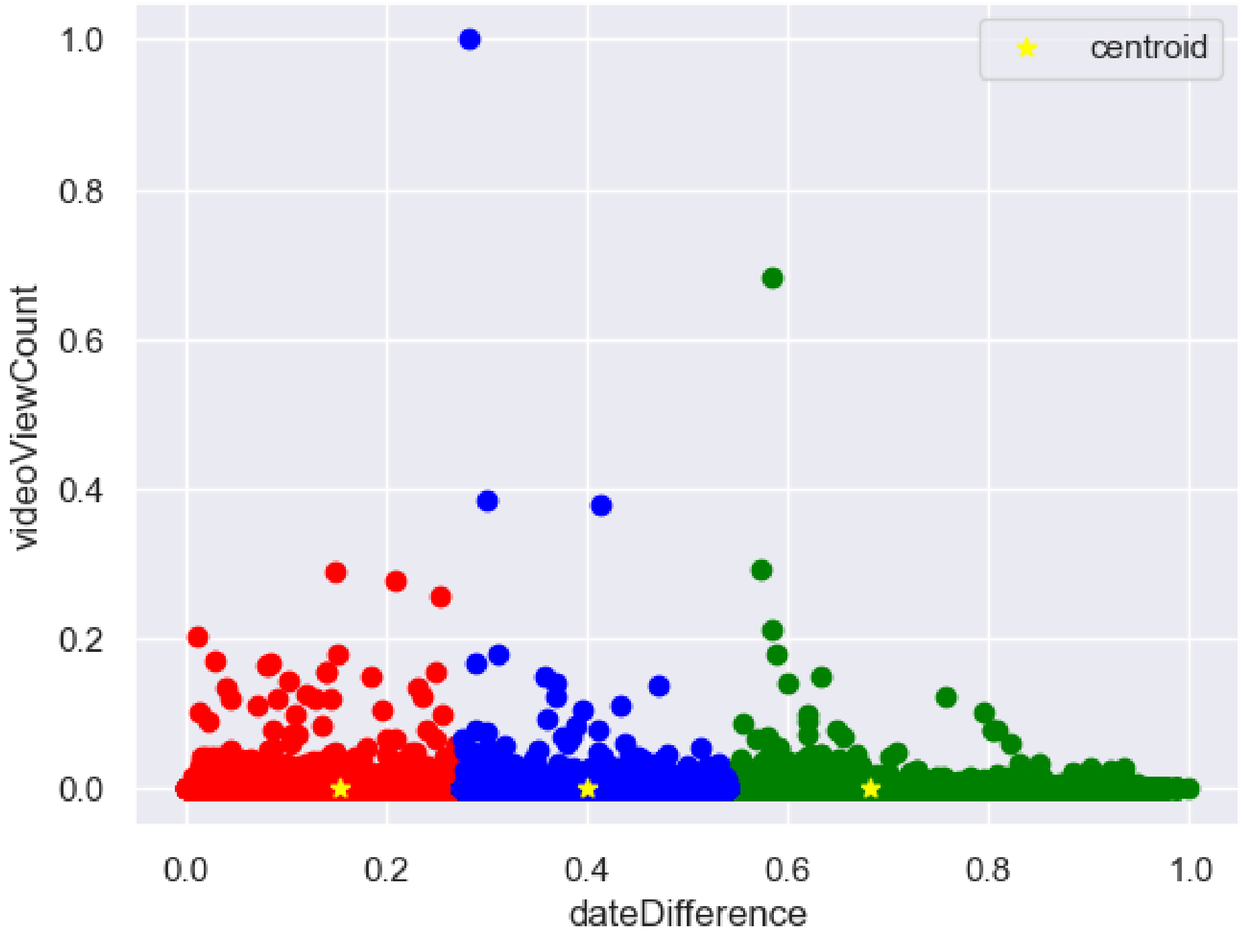
This optimal K-value in turn represents the optimal number of clusters the data points can be segregated into.

Results

From the plot, we can observe that the optimal K value is 3.

This means that the data points can be divided into 3 unique clusters.

(K-Means) Clustering of Data Points



Description of Scatter Plot

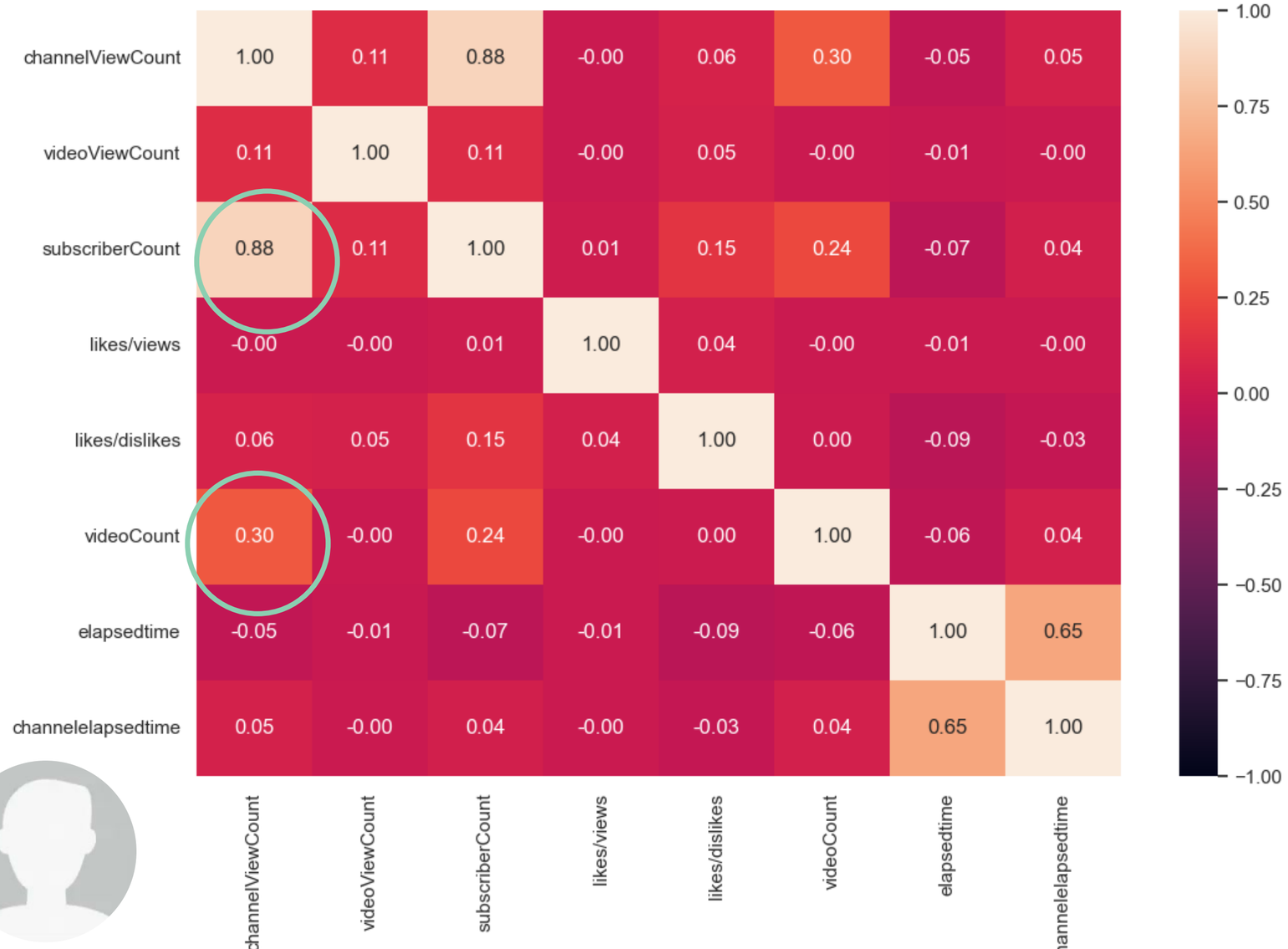
We can observe that three clusters (Red, Blue and Green) are largely determined by the dateDifference variable.

Results & Analysis

Even though the dateDifference in green cluster is quite high compared to that of in the red cluster, we can tell that the corresponding videoViewCount values are around the same and not very high.

This reinforces our previous finding that videoViewCount is not very dependent on dateDifference.

HeatMap between channelViewCount and other variables



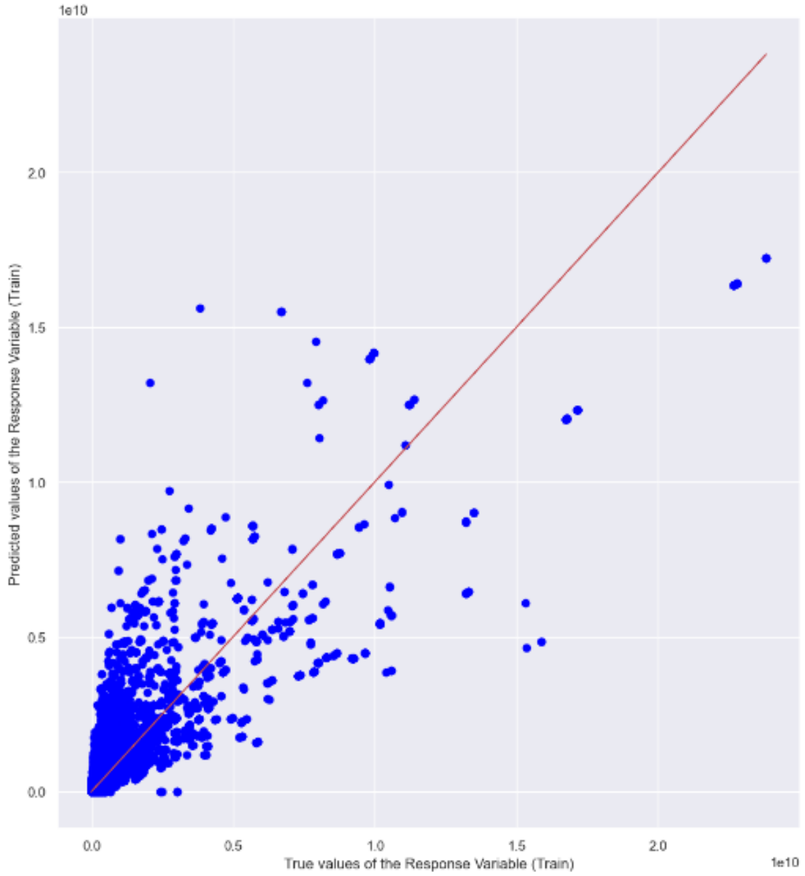
<u>Why did we look at the correlation matrix?</u>

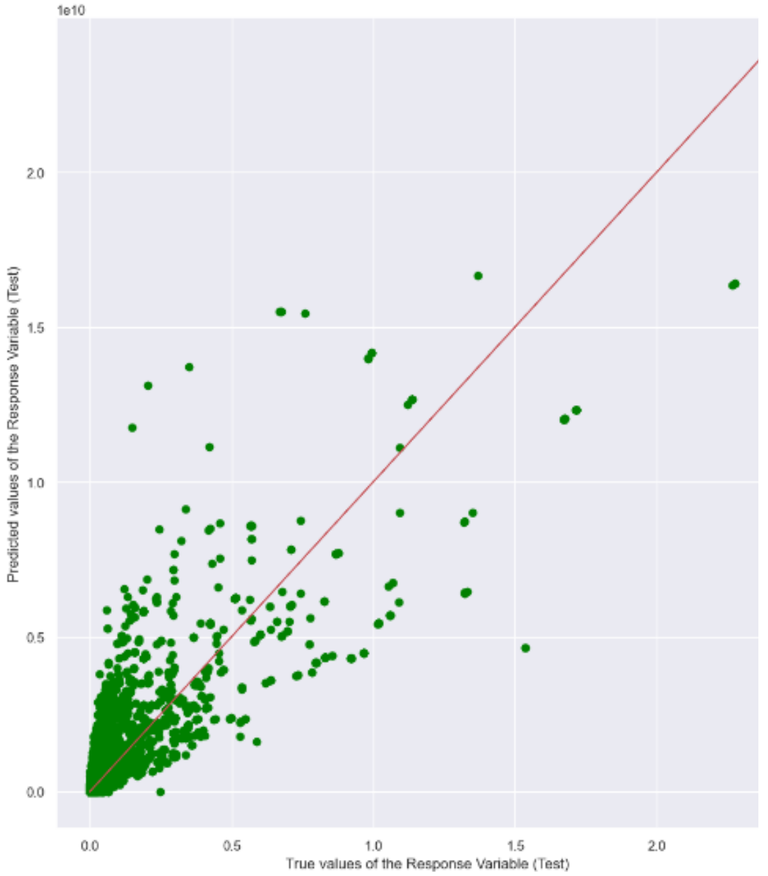
Since we have a mismatch between our expectations and reality, we decided that the correlation matrix help us understand the relationships between variables better.

<u>Observations</u>

Using the correlation matrix and the heat map, we discovered that subscriberCount has the highest correlation with channelViewCount with second being videoCount.

Regression between subscriberCount and channelViewCount






```
Elastic Net Regression
Intercept of Regression          : b =  [-6584745.71832883]
Coefficients of Regression       : a =  [682.27773035]

Goodness of Fit of Model         Train Dataset
Explained Variance (R^2)         : 0.7660464428341887
Mean Squared Error (MSE)         : 5.060281446743258e+16

Goodness of Fit of Model         Test Dataset
Explained Variance (R^2)         : 0.758258564879665
Mean Squared Error (MSE)         : 5.1875079007543784e+16
```

By comparing the regression plot between subscriberCount and channelViewCount and between videoViewCount and dateDifference, we can vividly observe that subscriberCount and channelViewCount has a better correlation than the latter.

Since we have previously determined that elastic net regression is the most optimal regression method, elastic net regression is employed here.

Summary Analysis

Key Factors and Recommendations

- Subscriber Count
  - ChannelViewCount
  - Likes per view
  - Likes per dislike
- Top Genres
  - Music
  - Entertainment
  - Gaming
- Video like count

Metrics with weak/no correlation

- Video Count
- Number of comments per video
- Elapsed Time

Interesting Insights

- Expectations
  - Longer Elapsed Time relates to higher viewer count
  - The higher number of videos posted will allude to more viral videos

- Actual Results
  - Difference in time between published date has little to no correlation to video views
  - Number of videos a channel has will not affect likelihood of increasing views for respective channel

Thank You