**Preliminary Work and Plan: Youtube Data Analysis**

Bowen Feng Howard Hu Jing Lin Yuchen Zhao

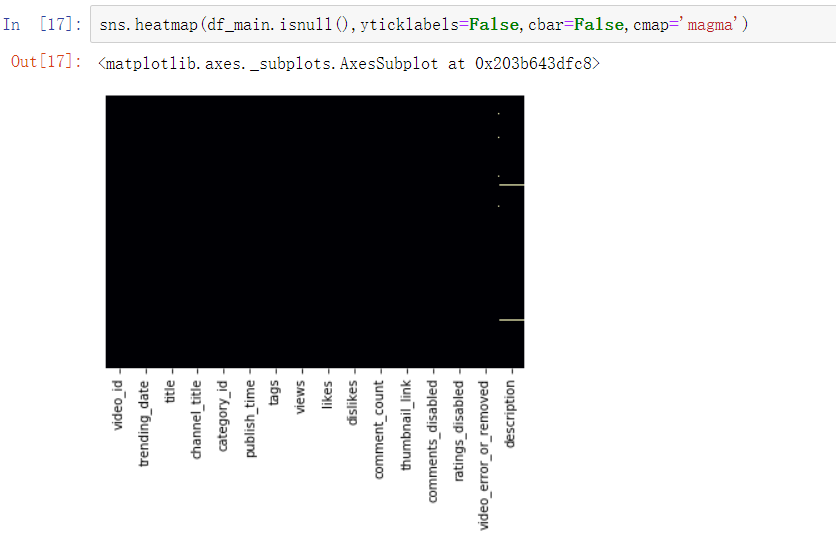
# **1. Preliminary Work**

Provide a list of things that you accomplished and learned about your data. Next to each, including the leader of the effort and anyone else who helped.

1. We downloaded Dataset(US Trending Video) from Kaggle, and loaded the dataset with pandas, we have 40949 records and 16 features.



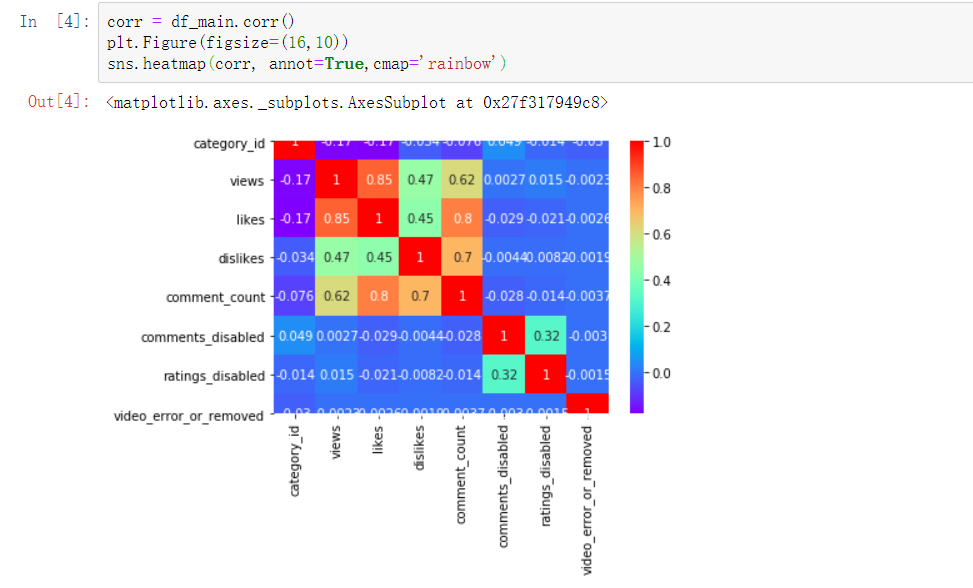
1. We create a dictionary that explains category\_id from the JSON file. Then we merge a new column called category into the original data frame for more efficient reading and analysis.
2. We plot the heatmap to check the missing data in our dataset, the white-ish color indicates the missing data in the corresponding feature:



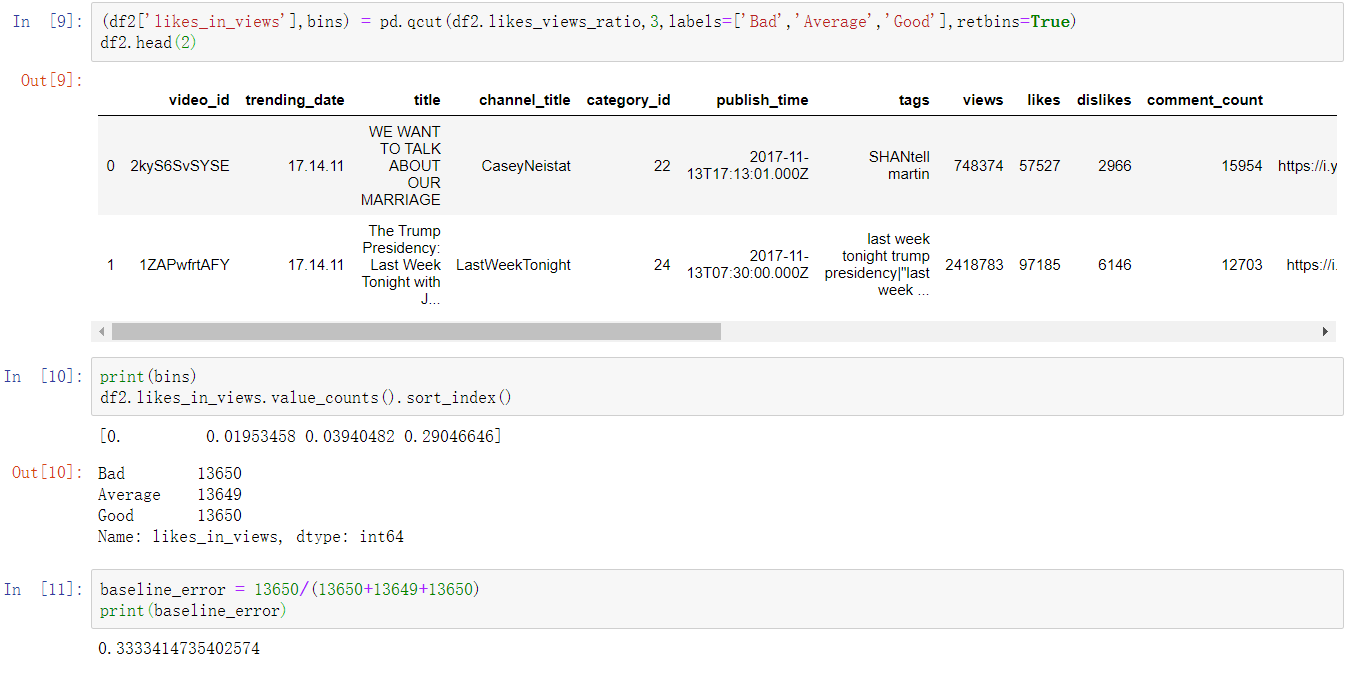
1. Using data visualization packages like panda and matplotlib to analyze the relationship between each different feature of trending video. Then we are trying to find common between top trending videos such as :
   1. How many numbers of likes and comments do trending videos have?
   2. What are the most popular tags and categories of trending videos?

We may raise more questions or change our questions.

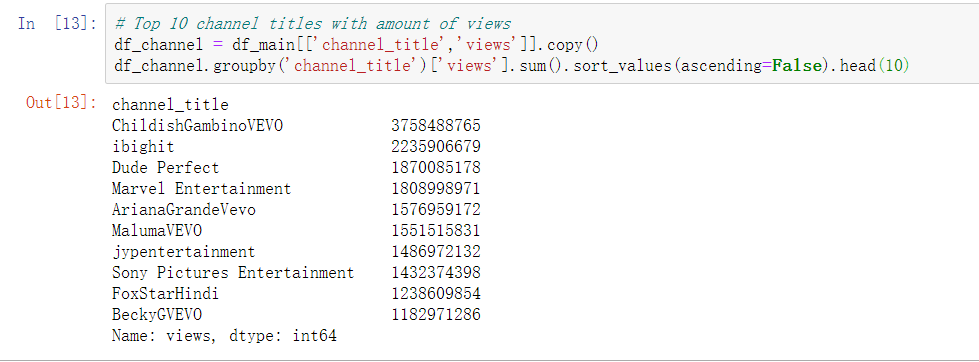
1. We use the pandas’ corr() function to find out the correlations among these features.



1. We visualize that:
   1. **Views** and **Likes** are correlated with 85%. We can conjecture that more views of a video brings more likes.
   2. **Likes** and **the number of Comments** are correlated with 80%. We would say the number of likes of a video has a positive relationship with its number of comments.
   3. **Dislikes** and **the number of Comments** are correlated with 70%. Although the number of dislikes has a positive relationship with its number of comments, its relationship is weaker than the relationship between likes and comments. We conclude that whether people like or dislike a video, they are willing to give a comment to express their idea.
2. The baseline error we found based on the likes-views ratio is about 0.333:



1. Based on our dataset, the top 10 channels with most total views are the following:



1. As top trending videos are likely to have a considerable number of likes, dislikes, and comments, we could find out from the correlation graph that all interactions with the video are correlated with each other.

# **2. Goals and Schedule**

|  |  |  |
| --- | --- | --- |
| **Goal** | **Description** | **Due date** |
| Minimum | 1. Analyzing the relationship between different features of the video:    1. Likes and dislikes vs. view    2. Comments vs. likes and dislikes    3. Top trending tags 2. Analyzing what factors affect how popular a YouTube video will be    1. Identifying the relationship between the popularity and factors the video has | End of 8th week |
| Reasonable | 3.Regression problem   1. Use feature engineering to extract useful features from raw data. We delete irrelevant features such as video description. 2. Separate data to train and predict dataset 3. Use the regression algorithm and calculate the RMSE | End of 9th week |
| Stretch | 1. Predictions more detailed information about the trending video like the views and likes ratio with    1. Tags    2. Titles | Final project due date |

**Detailed work schedule:**

|  |  |
| --- | --- |
| Task | Due date |
| Analyzing the relationship between different features of the video | 5.6 |
| Analyzing what factors affect how popular a YouTube video will be | 5.10 |
| Binary classification among different features | 5.17 |
| Predictions on more details of the trending video | 5.24 |
|  |  |