Final Report- Digital Media Intelligence

Shrivathsa Gopala Krishna Kumar, Rahul Niranjan Srinivas, Truc Huynh

Contents

Dataset description:

The dataset includes several months of data on daily trending YouTube videos. The data is captured in the USA and is stored in a csv file. The Column data includes:

Table 1: Dataset Description

Columns	Data Type
1. video_id	String
2. trending_date	Date
3. title	String
4. channel_title	String
5. category_id	Integer
6. publish_time	Date
7. tags	String
8. views	Integer
9. likes	Integer
10. dislikes	Integer
11. comment_count	Integer
12. thumbnail_link	String
13. comments_disabled	Boolean
14. ratings_disable	Boolean
15. video_error_or_removed	Boolean
16. description	String

Purpose of the project:

Vloggers on YouTube are provided with visual analytics on the content they upload by default, but they do not get the overall visualization of their competitor. Analyzing trending videos may provide publishers with the ability to predict the trend. Thus, they can develop their future content based on the analysis. Our analysis provides market forecast and intelligence which reveals information on the viewership, likes and dislikes. This prediction is based on many parameters and knowledge that our models have acquired by analyzing most trended videos that have been uploaded on YouTube. So, this can be extremely useful and profitable for YouTube channels who depend on their channels as a source of income.

Intended audience:

We assume the audience of this report to know the basics of:

1. R programming language
2. Algebra
3. Statistics.

Importing and tidying the data:

The data is being imported using the read.csv() function. Although this is slower compared to the read_csv(), it is important that the integrity and the proper structure of data be maintained. A single function has been constructed which handles the data import and tidying. It can be seen below:

The code above shows the function used to import and clean data. After the data is imported, the data is tidied. The process followed for tidying the data is as follows:

1. The date is processed so that it is in a readable format. And, only the required component of the date which is the month and days of week are selected.i.e. The variable trending_date and publish_time are cleaned to obtain the respective month.
2. The same process is repeated for both, trending_date and publish_time.
3. These variables are changed to a factor type to help in further analysis.
4. Two new variables are created which are basically containing the values of publish_time and trending_date. But, now, they are renamed to publish_month and trending_month respectively.
5. Then, the videos unnecessary to our project, i.e., the ones with no tags or no description are being omitted for the sake of convenience. This will not have a big impact on the analysis because most of them are outliers and losing them does not affect the analysis greatly.

Then, a new csv file is generated by using the write.csv() function. The new file is renamed as the same name as the input file, but with the a 'new' appended at the beginning of the filename. So, a new csv file containing the tidied data is generated once the function is called. The input to a function is a file. Eg: tidy_dataset(CAvideos.csv). So, once this function is called, the CAvideos.csv is tidied and another file called newCAvideos.csv is generated. This holds the tidied data. After the data is tidied, the structure of the data frame is as follows:

Table 2: Tidy Data Description

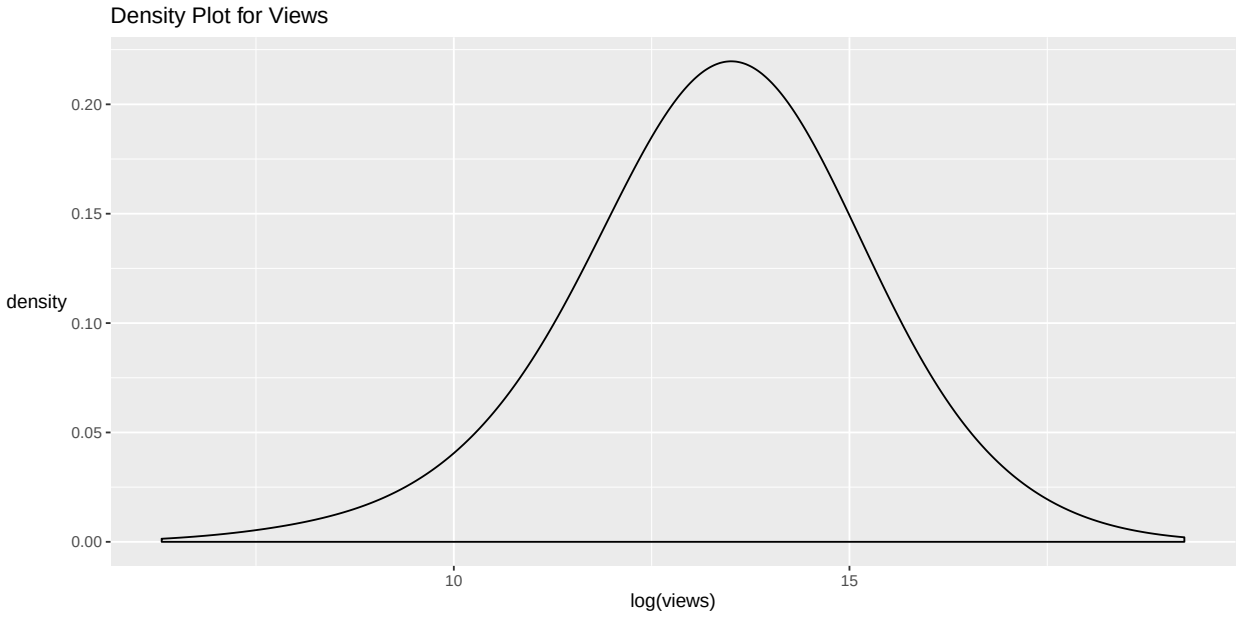
Columns	Data Type
1. trending_month	Factor
2. title	String
3. channel_title	String
4. category_id	Integer
5. publish_month	Factor
6. tags	String
7. views	Integer
8. likes	Integer
9. dislikes	Integer
10. comment_count	Integer
11. comments_disabled	String
12. ratings_disable	String
13. video_error_or_removed	String
14. description	String
15. day_of_week	Factor

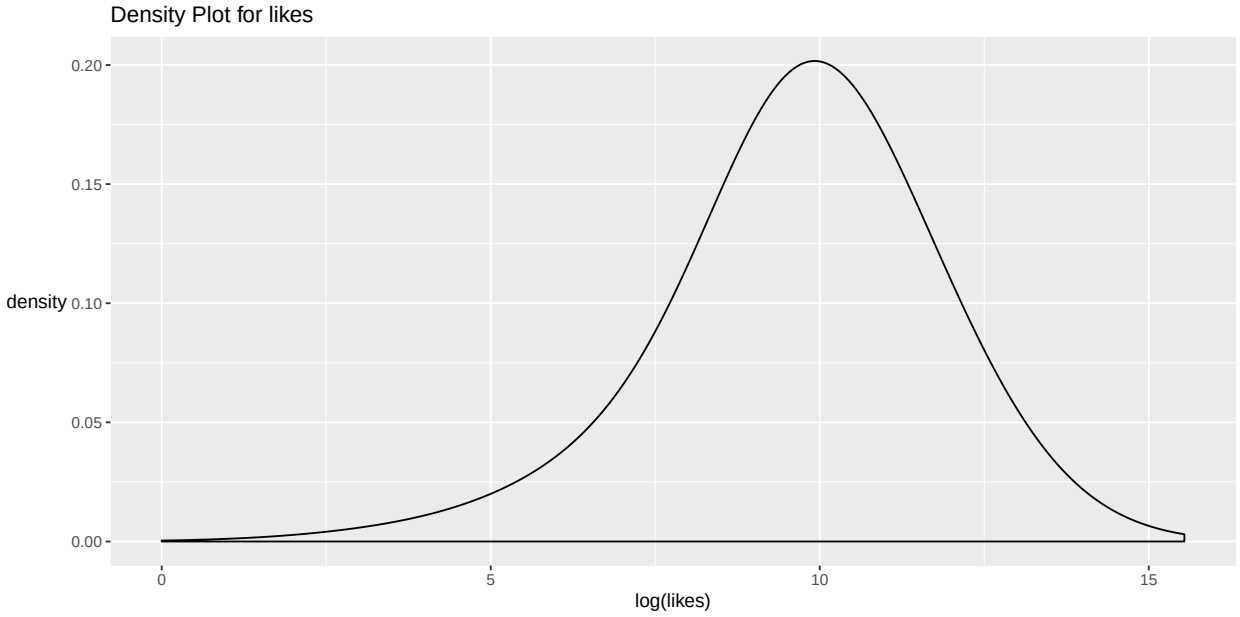
Exploratory data analysis:

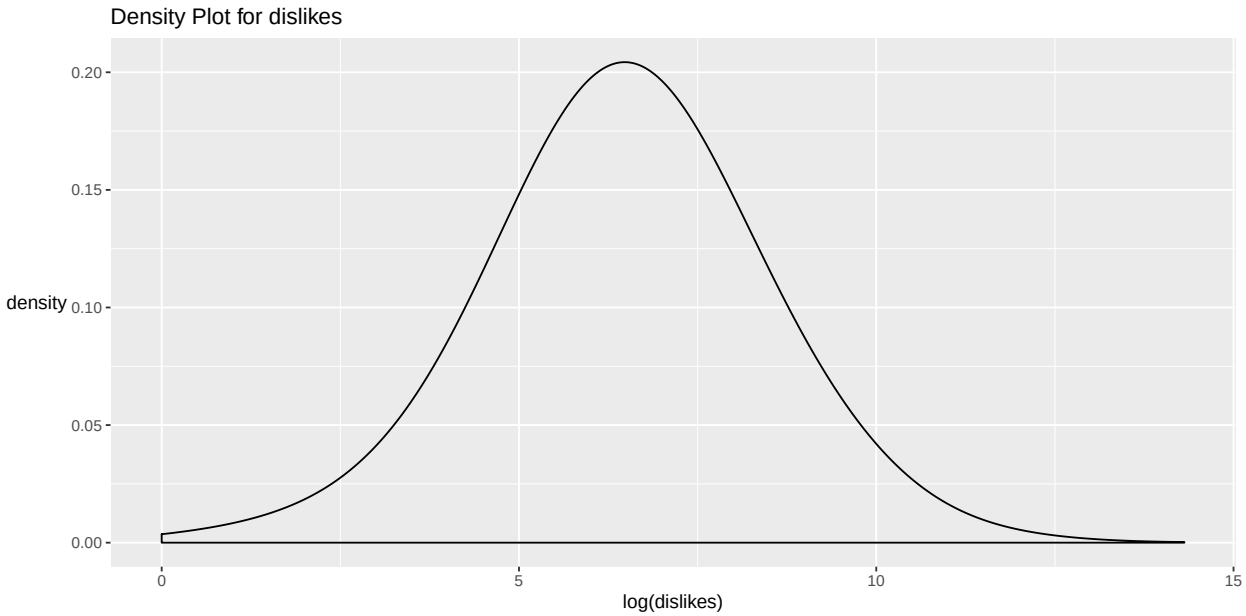
Exploratory Data Analysis is the critical process of performing initial investigations on the obtained data in order to build hypotheses, expose anomalies present and check assumptions.

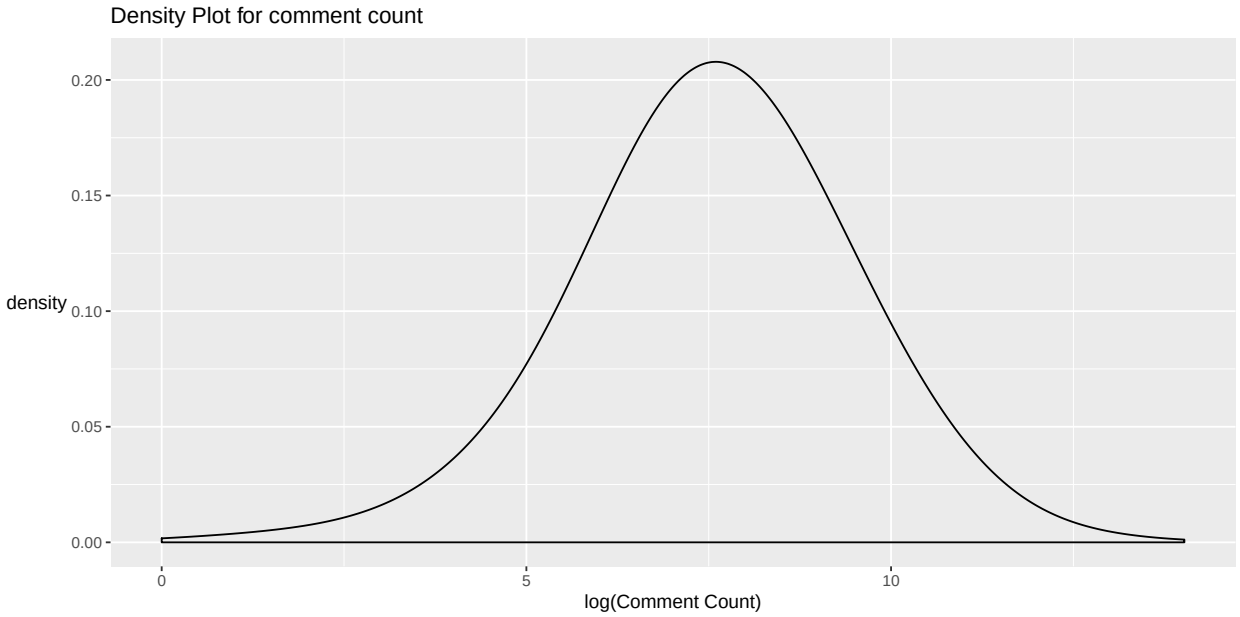
Study of distribution of variables in the dataset:

Density plot:



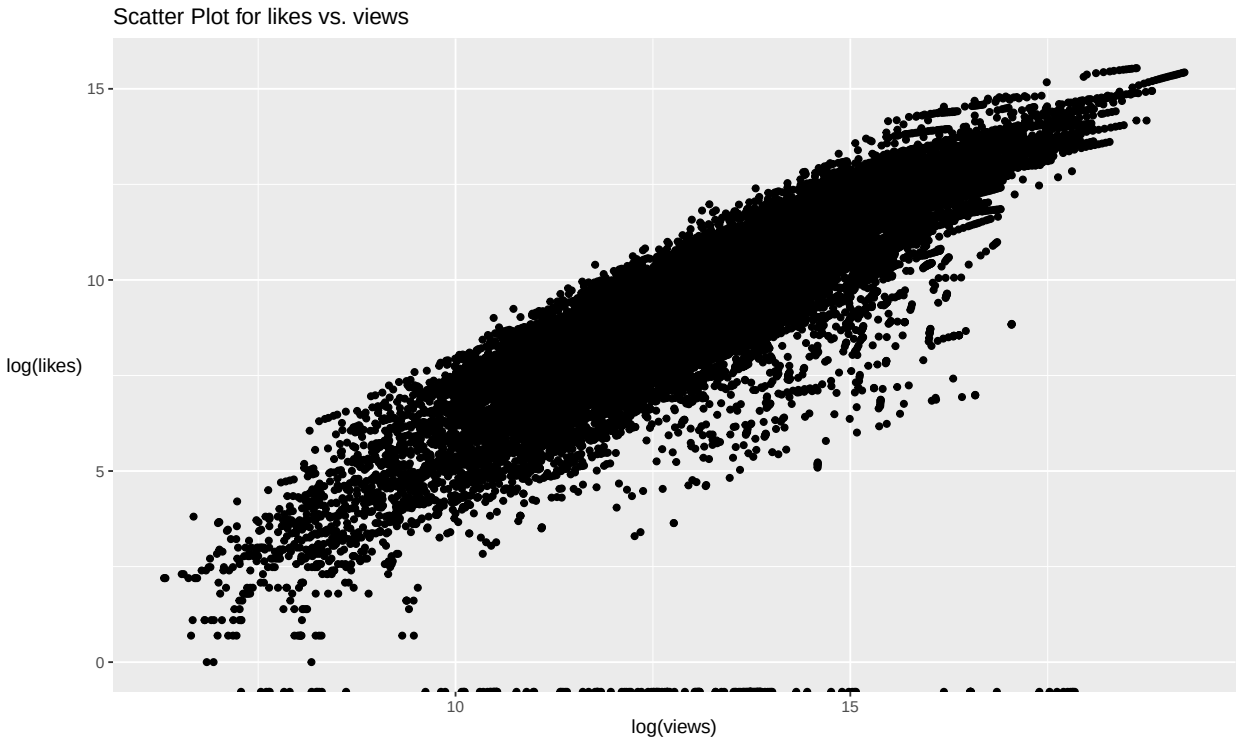


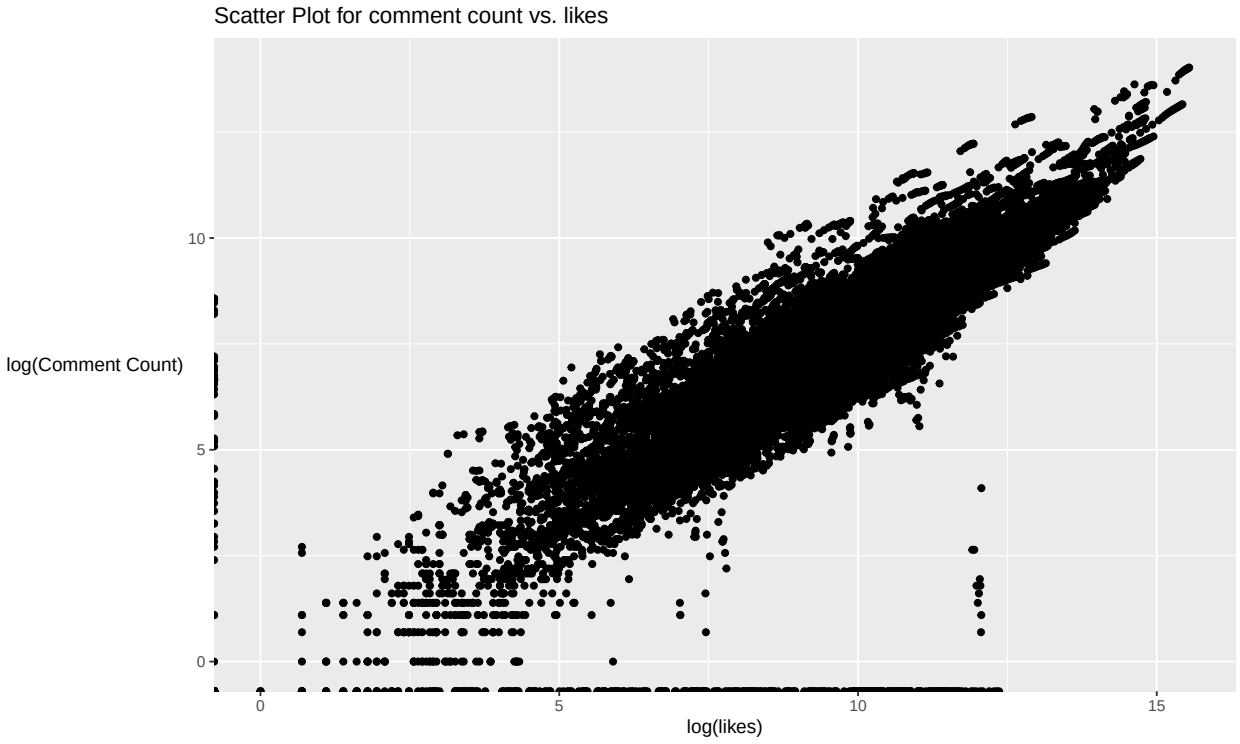




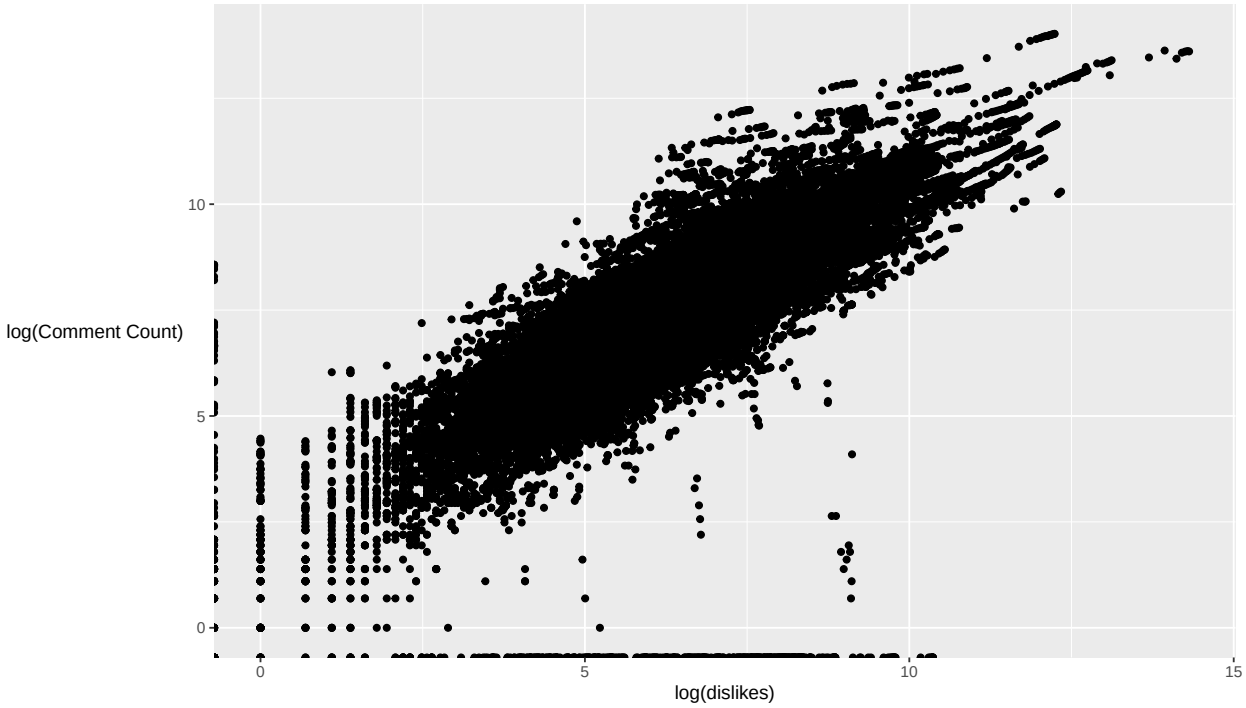
From the Density plots, we conclude that all the continuous variables follow the normal distribution. This brings us to the next step for verifying correlation whether it is valid or spurious. We term a correlation is Spurious when the correlation matrix shows a strong correlation coefficient, but the scatterplot does not show any possible linear curve fit. We perform this step, in order to eliminate spurious correlations. In this step we draw scatter plots to the strong and moderately correlated variable pairs.

Scatter plot:

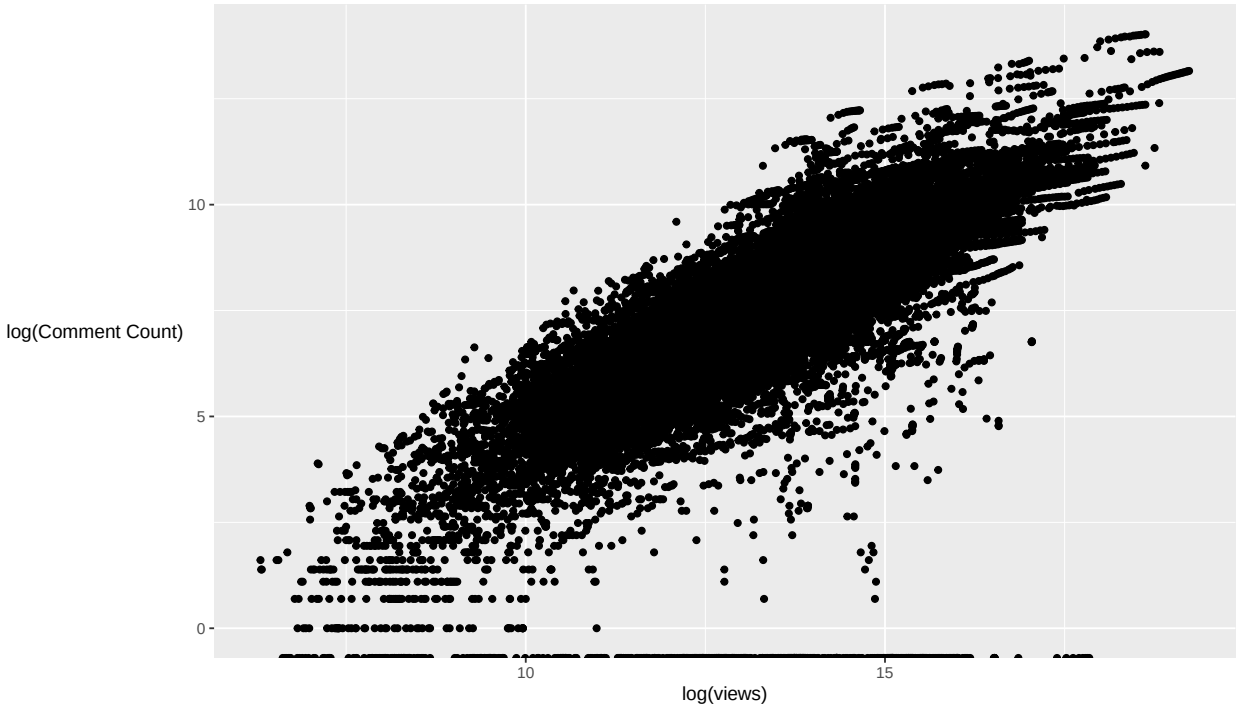


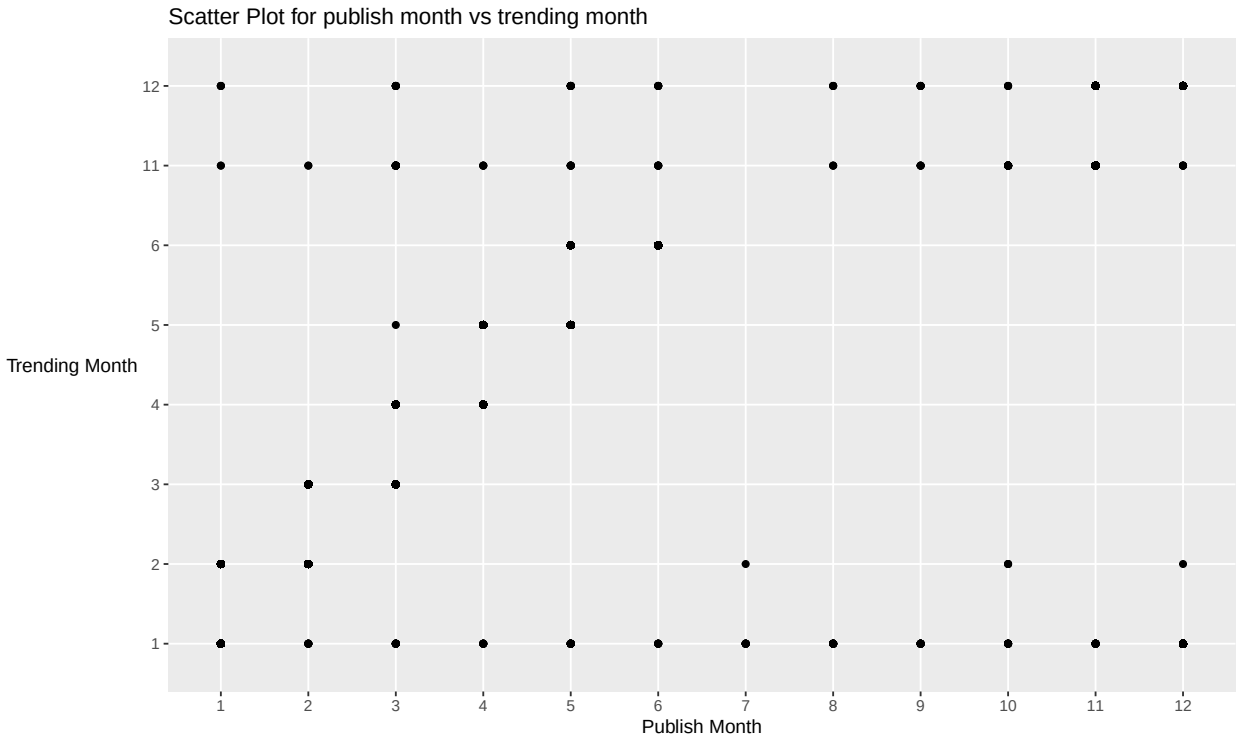


Scatter Plot for comment count vs. dislikes



Scatter Plot for views vs. comment count





From the Scatter plots, we conclude that all the selected variable pairs are not spurious. Now that we have proven that the correlation is not spurious, we go ahead to assume that the relationship is valid.

As a result of Exploratory Data Analysis, we will know the following:

1. Correlation map for the variables present in the dataset
2. Initial graphical visualizations

Correlation study:

Correlation definition:

Correlation is the bivariate (two variable) analysis that measures the strength of association between two variables. It is also used to determine the direction of the association. The strength of the association is expressed by the correlation coefficient, whose value varies between -1 and +1. A value of +1 indicates a strong positive correlation, a value of -1 indicates a strong negative correlation and a value of 0 indicates there is no correlation. Positive correlation meaning the value of the response variable tends to change in the same direction for every change in the explanatory variable. Negative correlation meaning value of the response variable tends to change in the opposite direction for every change in the explanatory variable. The closer the coefficient value is to zero the weaker the strength of correlation.

Reasons for performing correlation analysis:

We perform Correlation analysis in order to determine the strength of association between variables in the dataset. By doing so we will be able to know the potential attributes about each variable.

Correlation is strictly used to test association between variables, it does not make any assumptions whether one variable dependent on the other.

Different types of correlations:

In [1] it is given that, the three types of correlation are:

1. Pearson Product Moment Correlation
2. Spearman Rank Correlation
3. Kendall Rank Correlation

Pearson correlation:

It is a technique used to measure two quantitative and continuous variables[2].

Spearman rank correlation:

It is a technique used when we need to measure correlation between two ranked variables or when we want to measure the correlation between a quantitative variable and a ranked variable[3].

Kendall rank correlation:

It is a technique used when we need to measure the correlation between pairs of bivariate points; the co-ordinates are measured individually to declare each point in the graph as concordant or discordant with respect to the other points on the graph.

Example:

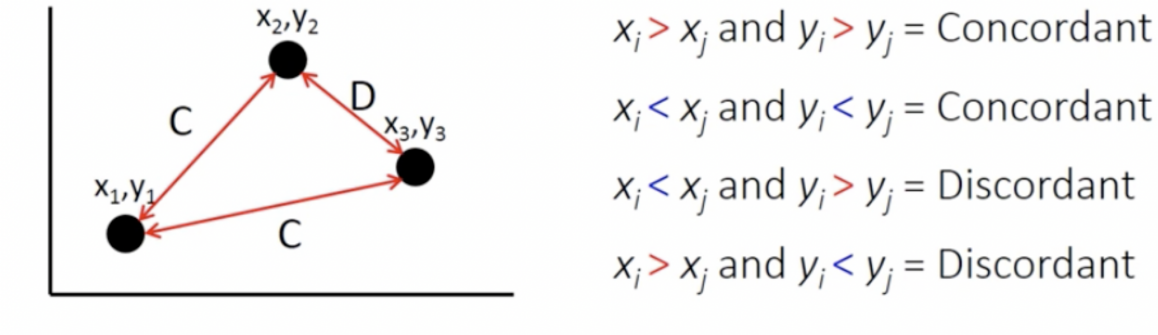
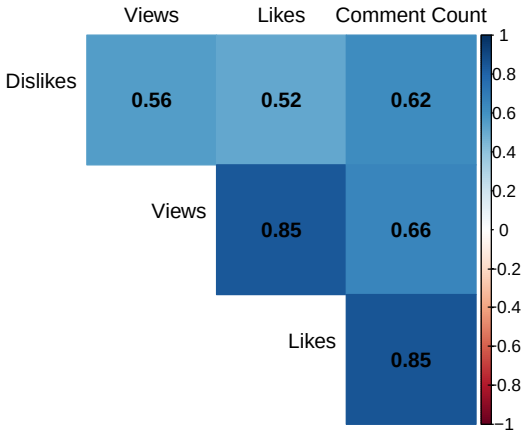


Figure 1: Kendall Rank Correlation

Applications of correlation in our project:

Pearson correlation graphical map:



From the Pearson Correlation map, it is clear that,

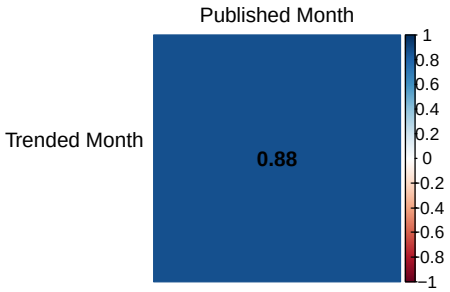
1. Views and likes
2. Comment count and likes

These are strongly correlated with a coefficient value of 0.85 and 0.85 respectively. Following which,

1. Comment count and dislikes
2. Comment count and views These pairs moderately correlated with a coefficient value of 0.62 and 0.66 respectively.

While we find the remaining pairs are also positively correlated, we ignore the correlation because the value of the remaining pairs relative to the above four pairs are weakly correlated.

Spearman correlation map:



From the Spearman Correlation map, it is clear that, Trending month and Published month are highly correlated, with a Coefficient value of 0.88.

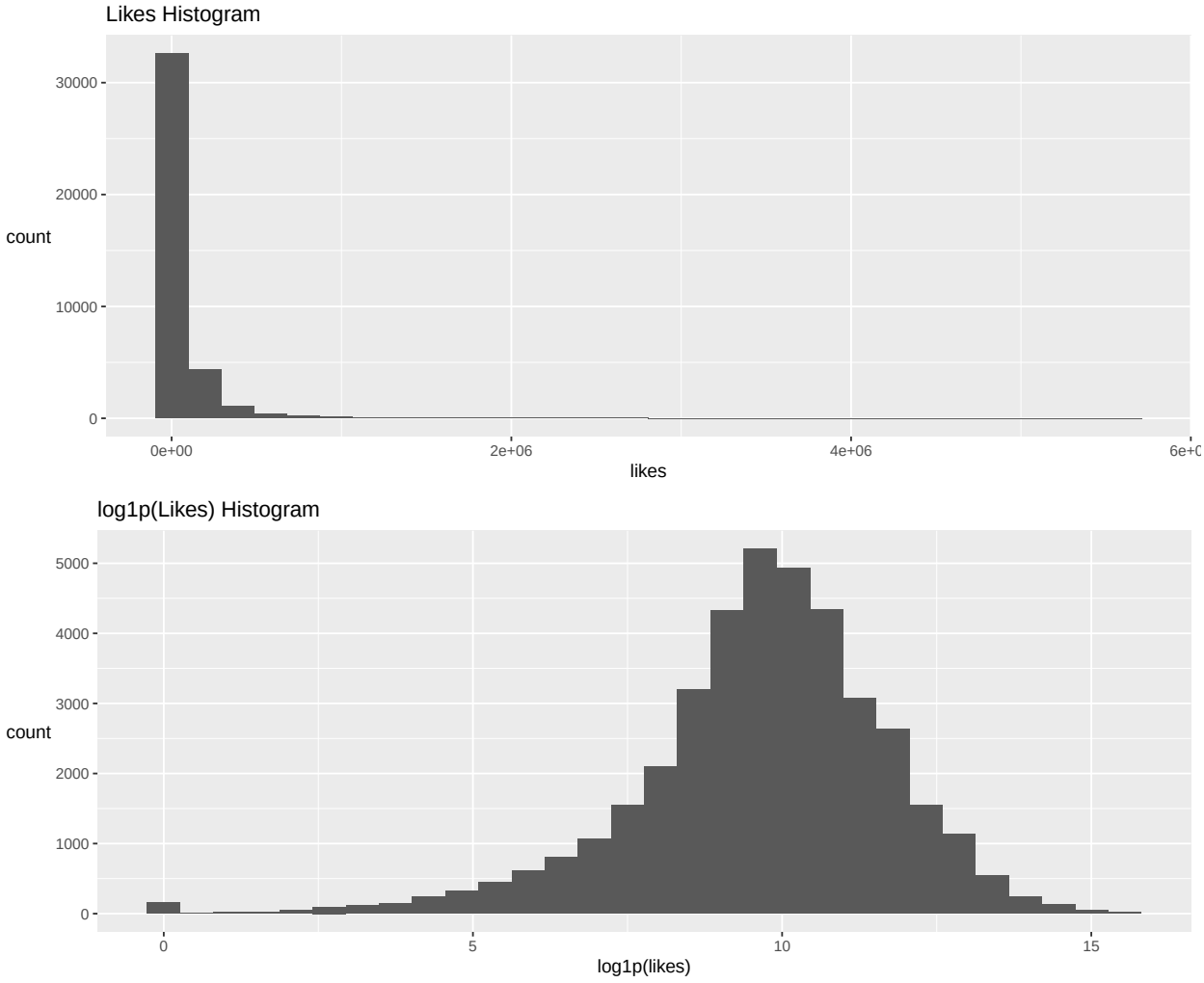
Hypothesis definition, analysis, testing and model selection:

Hypothesis 1:

Predict the number of likes for a trending video in the dataset.

Table 3: Hypothesis 1 Summary Statistics

mean likes	median likes	standard deviation of likes
75753.04	18560	232943.6



Taking natural log, Since log10 or log or log2 gives infinity for observations containing zero, we have taken log1p i.e log(1 + x), this keeps the value calculatable, as log is monotonic, the ordering is preserved. $|x| << 1$.

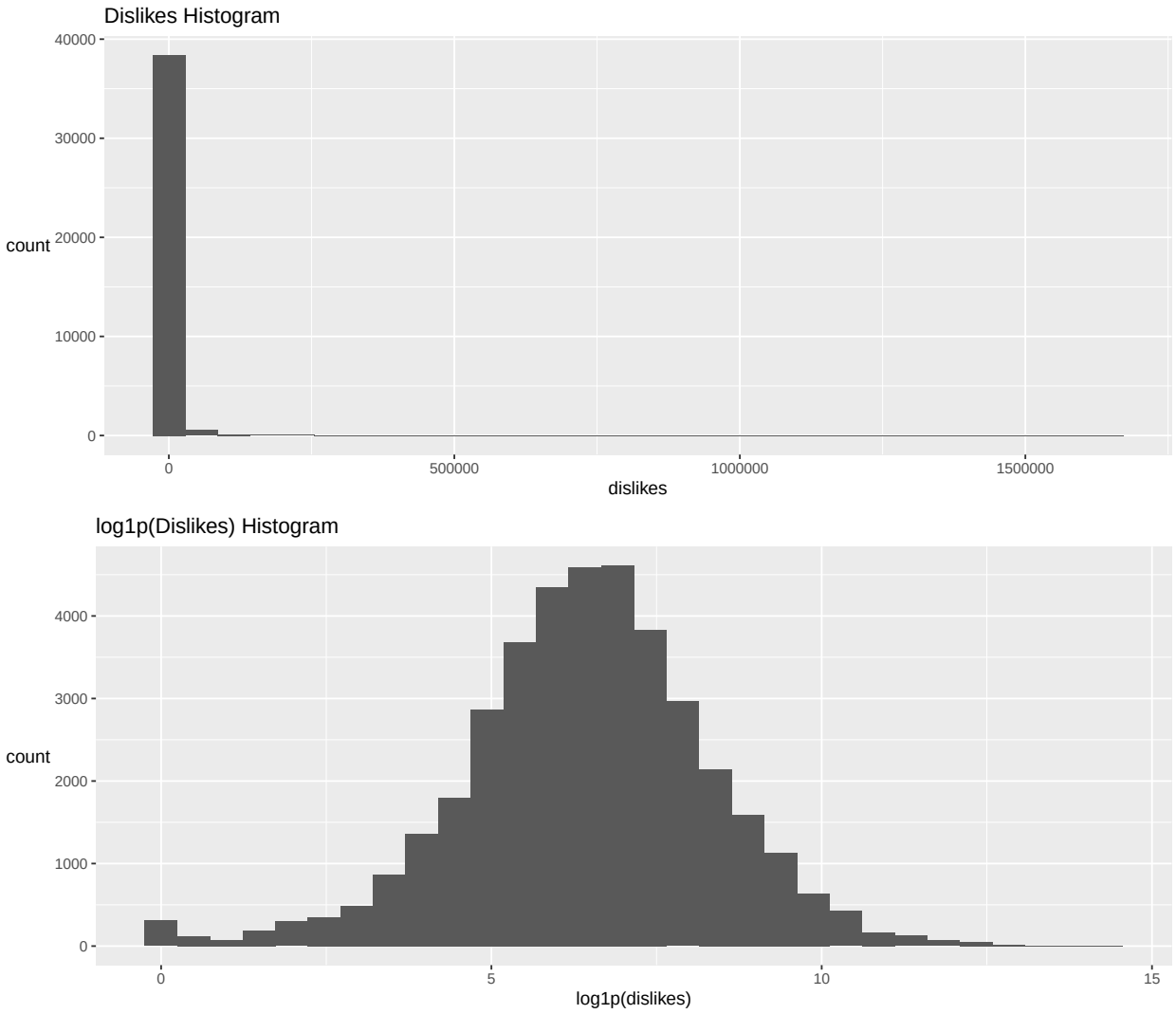
The high R squared value explains that the prediction is accurate up to 81.4% of the time. We arrive at this value because, the closer the R squared value is to 1.0, the closer it is to being perfectly accurate. There are many parameters to assess the model, we've chosen to go ahead with R squared method. To make sure that the model is unbiased, we have considered accuracy of the model when the testing dataset is used.

Hypothesis 2:

Predict the number of dislikes for a trending video in the dataset.

Table 4: Hypothesis 2 Summary Statistics

mean dislikes	median dislikes	standard deviation of dislikes
3563.972	641	23527.4



Taking natural log, Since log10 or log or log2 gives infinity for observations containing zero, we have taken log1p i.e log(1 + x), this keeps the value calculatable, as log is monotonic, the ordering is preserved. |x| << 1.

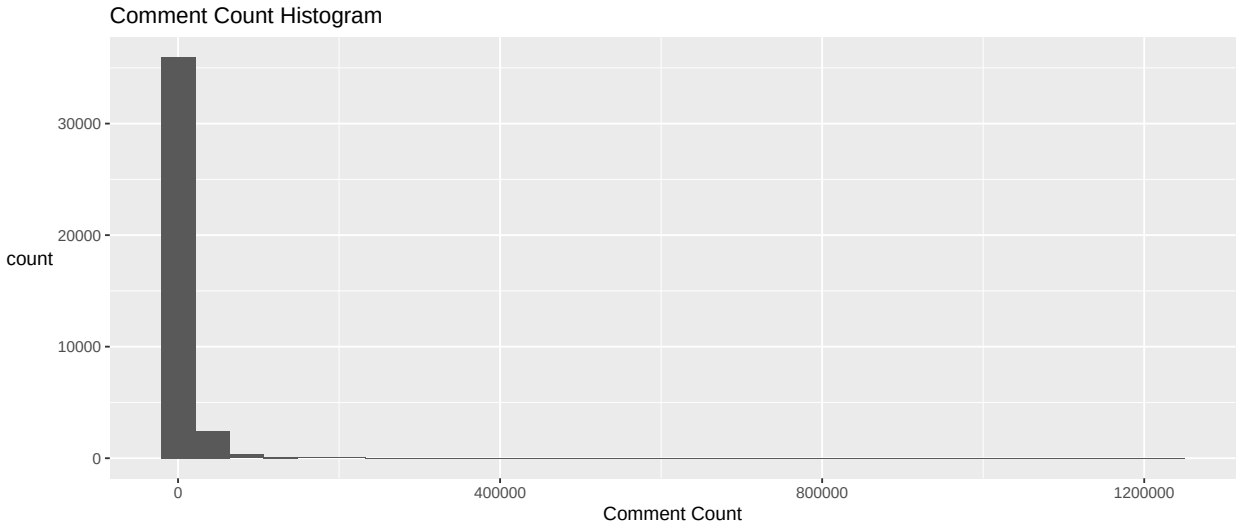
The high R squared value explains that the prediction is accurate up to 78.8% of the time. We arrive at this value because, the closer the R squared value is to 1.0, the closer it is to being perfectly accurate. There are many parameters to assess the model, we've chosen to go ahead with R squared method. To make sure that the model is unbiased, we have considered accuracy of the model when the testing dataset is used.

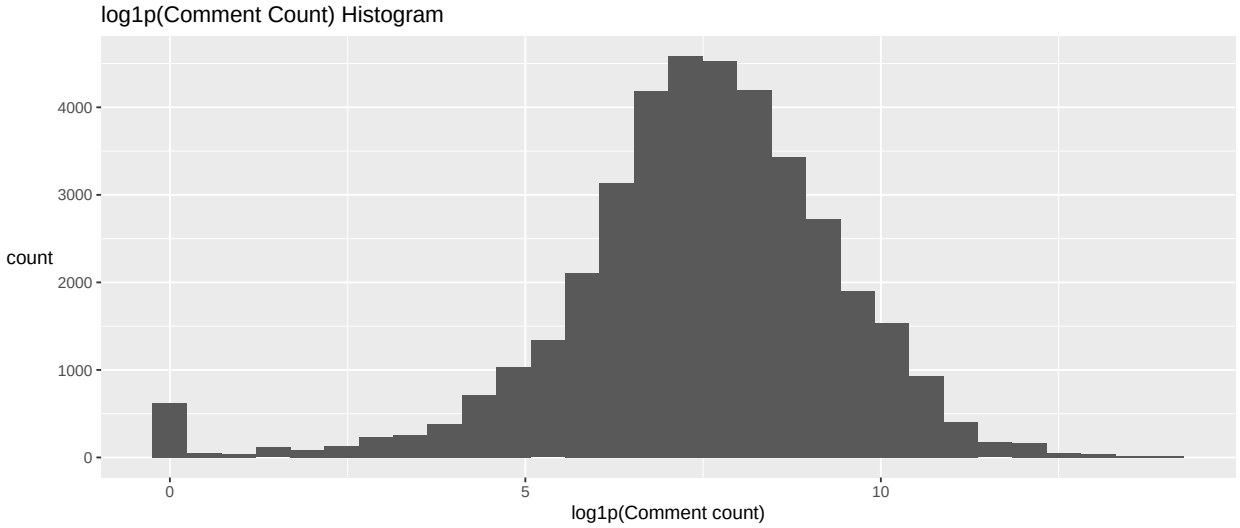
Hypothesis 3:

Predict the number of comments for a trending video in the dataset.

Table 5: Hypothesis 3 Summary Statistics

mean comment count	median comment count	standard deviation of comment count
8418.509	1897	34828.55





Taking natural log, Since log10 or log or log2 gives infinity for observations containing zero, we have taken log1p i.e log(1 + x), this keeps the value calculatable, as log is monotonic, the ordering is preserved. $|x| << 1$.

The high R squared value explains that the prediction is accurate up to 70.9% of the time. We arrive at this value because, the closer the R squared value is to 1.0, the closer it is to being perfectly accurate. There are many parameters to assess the model, we've chosen to go ahead with R squared method. To make sure that the model is unbiased, we have considered accuracy of the model when the testing dataset is used.

Reason for choosing the hypotheses:

As a result of Exploratory Data Analysis, we noticed that:

1. Views and likes
2. Comment count and likes
3. Comment count and dislikes
4. Comment count and views

Are strongly interdependent, however we make no assumption about the causal effect. The strong interdependence gives us an insight to further investigate to check the causality of this interdependence.

Hypotheses test:

We follow K-Fold cross validation technique, as it is best suited for small datasets[5]. (Please note that the following table has been modified to improve readability and to fit the table in a page)

Table 6: Hypothesis Test

Title	log1p(likes)	log1p(views)	log1p(comment count)
WE WANT TO TALK ABOUT OUR MARRIAGE	10.960027	13.52566	9.677528
The Trump Presidency: Last Week Tonight with John Oliver (HBO)	11.484382	14.69878	9.449672
Racist Superman | Rudy Mancuso, King Bach & Lele Pons	11.891595	14.97598	9.009692
Nickelback Lyrics: Real or Fake?	9.227492	12.74598	7.671827
I Dare You: GOING BALD!?	11.792343	14.55541	9.771041

The data of this table contains the column "training_cases" which conveys whether the data is training or testing data. The data has been used in a 2-Fold cross validation technique which has boolean values under the column "training_cases". "TRUE" indicates training data and "FALSE" indicates testing data. This has been done so that the model is unbiased and overfitting is prevented.

Reason and selection of model:

From the Pearson correlation map, we can conclude that there is strong interdependence on many continuous variables of the dataset. This tells us that there is a possibility of using simple linear regression because simple linear regression is useful for finding the relationship between two continuous variables in a dataset.

From the Spearman correlation map. It is evident that Trending date and Publish Date are strongly interdependent, however further investigation is required to help us predict this strong association.

Challenges:

Due to the COVID-19, and one of our team members stuck in Boston we were not be able to meet. We needed to strategize our communication to adapt to this situation.

We needed to come up with a model for predicting the causal effect for the strong correlation between the trending month and publish month. We still need to investigate further in order to come up with more suitable hypotheses for the same.

Conclusion:

We found this dataset on Kaggle.com and we came up with a novel approach of conveying the analysis by using shiny. The analysis predicts the number of likes, dislikes and comments. We learnt that we need to use linear regression depending on the scatter plot we visualize during the exploratory data analysis. We use a density plot to know the distribution of the response variables. We later introduced a correlation heat map. Although the idea of correlation coefficient matrix is not new, we found that most of the visualizations on correlation although informative, require a bit more effort to understand. We simplified it further to make it look more readable to the audience and attempted to increase the audience base for the analysis. To make it more interactive, we have used Shiny. This makes it easier for the user to use our model and interact in real time, by providing predictions in real time. Overall, we have made an attempt to make the information in our analysis more reproducible, redistributable and novel.

References:

1. Kerley, Christa. "What Are the Different Types of Correlations?" sciencing.com, https://sciencing.com/different-types-correlations-6979655.html. March 25 2020.
2. University of the West of England, Pearson's Correlation Coefficient. Retrieved March 25, 2020, from http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442.
3. McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.
4. Clapham, Matthew. E. Jan 30, 2016, 19: Non-parametric correlation, Retrieved March 25, 2020, from https://www.youtube.com/watch?v=bAstMHbytK0
5. AS (2017). JMP 13 Fitting Linear Models, Second Edition.