VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF APPLIED SCIENCE

MT2013 - PROBABILITY AND STATISTICS

Assignment Report - Group 30

Analyze and evaluate contributing factors to the view-count of BBC videos on YouTube platform with the assistance of various tools

---

Under the guidance of: Dr. Nguyen Tien Dung

HO CHI MINH CITY, MAY 2023

Member list & Workload

No.	Full name	Student ID	Contribution
1	Nguyen Van Thanh Dat	2152055	20%
2	Nguyen Khanh Nam	2153599	20%
3	Le Van Phuc	2152241	20%
4	Nguyen Quang Thien	2152994	20%
5	Tran Minh Tuan	2152336	20%

Contents

1 Introduction

In this given assignment, our team members have opted for using Kaggle, which is not only a reliable and valuable website for precise up-to-date information ranging from various fields of life but also offers strong tools and resources to assist us in reaching desired data science objectives. We hope to find a source of data that is abundantly clear enough and also has plentiful contributing factors, which can pave the way for our team to carry out a comprehensive analysis and easily conclude in the most sensible way later on.

Figure 1: Kaggle

Throughout several times of discussion, we end up selecting sort of data in a conducted survey of a channel named BBC on YouTube platform. This is considered to be supportive as it contains over 10,000 figures from different videos published with numerous key factors *(explained in detail in the section 3.2.)* included inside, which are:

- Category
- Duration
- Licensed Content
- Like-count
- Dislike-count
- Comment-count
- View-count

Based on the figures and elements collected, our team has made up our mind for the main goal of the assignment to analyze and check **which kind of factors mentioned above mostly exert a decisive effect on the view-count of BBC videos** with the help of several tools, namely *Multi-factor ANOVA Test*, *Multiple Linear Regression Models using R*.

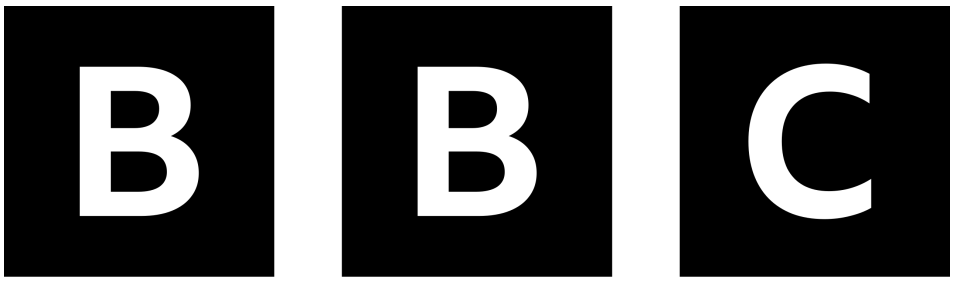

Figure 2: BBC Channel

2 Theoretical Basis

2.1 Multiple Linear Regression

2.1.1 Basic concept

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. Multiple linear regression can be use when we want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable.

2. The value of the dependent variable at a certain value of the independent variables.

Assumptions of multiple linear regression

Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor. The Residuals vs Fitted and Normal Q-Q graph are used to ensure.

Normality: The data follows a normal distribution. This assumption is confirmed by the usage of Normal Q-Q graphs as well as the Residual Histogram.

Independence of observations (Multicollinearity): In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to make sure these before developing the regression model. If two independent variables are too highly correlated ($r^2 > 0.6$), then only one of them should be used in the regression model.

Homogeneity of variance (Homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable. Or simply, standard deviation are equal for all points. Scale-Location and Residuals vs Fitted Graph are useful when supporting the validation of this assumption.

Multiple linear regression formula

$$y = \alpha + \beta_1 X_1 + ... + \beta_n X_n + \epsilon, \text{ where:}$$

y: the predicted value of the dependent variable.
α: the y-intercept.
$\beta_i, i = 1, 2, ...n$: the regression coefficient of the i^{th} variable X_i.
ϵ: model error.

2.1.2 Interpreting Linear Regression Model Output in R

Consider the following example:

```
> LRmodel2<-lm(view_count ~ duration_sec + like_count + comment_count, data = youtube)
> summary(LRmodel2)

Call:
lm(formula = view_count ~ duration_sec + like_count + comment_count,
    data = youtube)

Residuals:
     Min       1Q   Median       3Q      Max
-1158958   -57370   -32768    26626   608549

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  81640.3095  1400.7131  58.285   <2e-16 ***
duration_sec   -32.8822     3.6404  -9.033   <2e-16 ***
like_count      65.8127     0.7488  87.895   <2e-16 ***
comment_count   89.6119     4.5851  19.544   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109600 on 10472 degrees of freedom
Multiple R-squared:  0.5159,    Adjusted R-squared:  0.5157
F-statistic:  3720 on 3 and 10472 DF,  p-value: < 2.2e-16
```

Figure 3: Linear model example.

Now, let's briefly discuss each component of the output.

Formula Call: The first item shown in the output is the formula R used to fit the data.

Residuals: The Residuals section of the model output breaks it down into 5 summary points. When assessing how well the model fit the data, we should look for a symmetrical distribution across these points on the mean value zero (0). In our example, we can see that the distribution of the residuals do not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points.

Coefficients:

Estimate: The Estimate column is the estimated effect, also called the regression coefficient or r^2 value.

Standard Error: The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable.

t-value: The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between speed and distance exist.

Pr(>t): The Pr(>t) acronym found in the model output relates to the probability of observing any value equal or larger than t. A small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between speed and distance. Typically, a p-value of 5% or less is a good cut-off point. The 'Signif. Codes' associated to each estimate. Three stars (or asterisks) represent a highly significant p-value.

Residual Standard Error: Residual Standard Error is measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term ϵ.

Multiple R-squared, Adjusted R-squared: The R-squared (R^2) statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. R^2 is a measure of the linear relationship between our predictor variable (speed) and our response / target variable (dist). It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). In multiple regression settings, the R^2 will always increase as more variables are included in the model. That's why the adjusted R^2 is the preferred measure as it adjusts for the number of variables considered.

F-Statistc: F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be depends on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H_0 : There is no relationship between speed and distance). The reverse is true as if the number of data points is small, a large F-statistic is required to be able to ascertain that there may be a relationship between predictor and response variables.

2.1.3 Residuals versus Leverage

A residuals vs leverage plot is a type of diagnostic plot that allows us to identify influential observations in a regression model.

Each observation from the dataset is shown as a single point within the plot. The x-axis shows the leverage of each point and the y-axis shows the standardized residual of each point.

Leverage refers to the extent to which the coefficients in the regression model would change if a particular observation was removed from the data set. Observations with high leverage have a strong influence on the coefficients in the regression model. If we remove these observations, the coefficients of the model would change noticeably.

Standardized residuals refers to the standardized difference between a predicted value for an observation and the actual value of the observation.

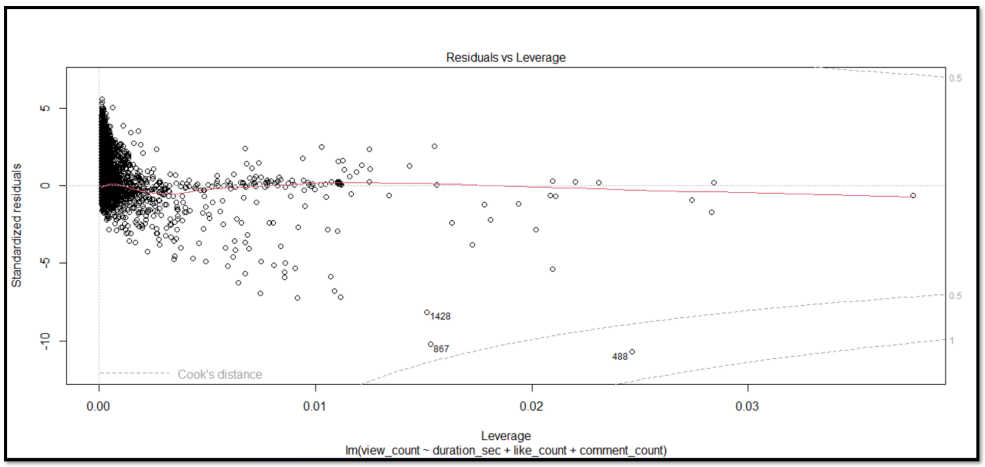

Figure 4: An example of the Residuals vs Leverage plot.

If any point in this plot falls outside of Cook's distance (the red dashed lines) then it is considered to be an influential observation. In this example, there is a point falls outside of the dashed line(observation number 3381). This means that this regression model has 1 influential points.

2.1.4 Scale-Location Graph

A scale-location plot is a type of plot that displays the fitted values of a regression model along the x-axis and the the square root of the standardized residuals along the y-axis.

When looking at this plot, we check two things:

1. Verify that the red line is roughly horizontal across the plot. If it is, then the assumption of homoscedasticity is likely satisfied for a given regression model. That is, the spread of the residuals is roughly equal at all fitted values.

2. Verify that there is no clear pattern among the residuals. In other words, the residuals should be randomly scattered around the red line with roughly equal variability at all fitted values.

Let's consider the following example. We can observe two points from the Scale-Location plot for this regression model.

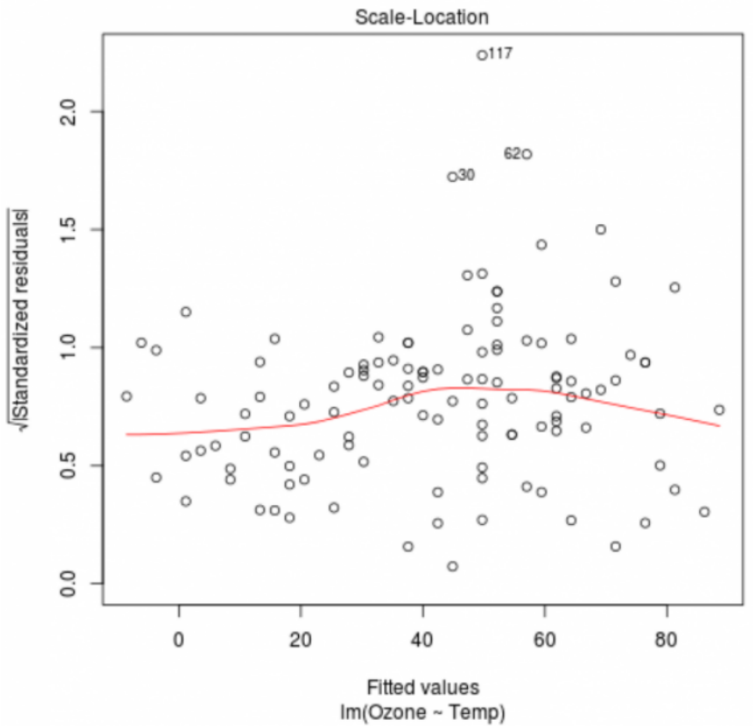

Figure 5: An example of the Scale-Location plot.

Firstly, the red line is roughly horizontal across the plot. This proves that the assumption of homoscedasticity is sastified for a given regression model.

Secondly, the residuals should be randomly scattered around the red line with roughly equal variability at all fitted values.

2.1.5 Residuals vs Fitted Graph

The Residual vs Fitted plot allows us to detect several types of violations in the linear regression assumptions.

In the plot, the fitted values $\widehat{y}$ is sketched on the x-axis and the residuals y - $(\widehat{y})$ are represented on the y-axis. The Residuals and Fitted plot is mainly useful for investigating:

1. Whether Linearity holds. This is indicated by the mean residual value for every fitted value region being close to 0. In R this is indicated by the red line being close to the dashed line.

2. Whether Homoscedasticity holds. The spread of residuals should be approximately the same across the x-axis.

3. Whether there are outliners. This is indicated by some 'extreme' residuals that are far from the rest.

Let's take a look of the following plot for the car data set.

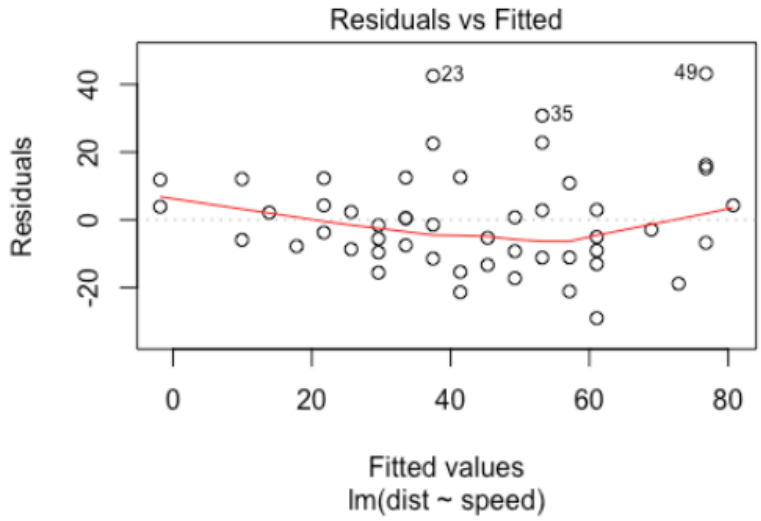

Figure 6: An example of the Residuals vs Fitted plot.

Here we see that linearity seems to hold reasonably well, as the red line is close to the dashed line. We can also note the heteroscedasticity: as we move to the right on the x-axis, the spread of the residuals seems to be increasing. Finally, the observations 23, 35, 49 may be outliners.

Considering another example:

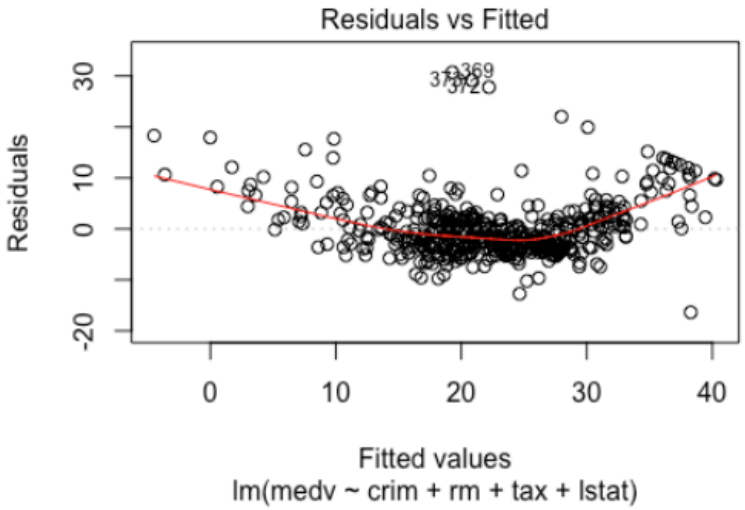

Figure 7: An example of the Residuals vs Fitted plot.

In this plot, the linearity is violated, there seems to be a quadratic relationship. Whether there is homoscedasticity or not is less obvious, we will need to investigate the others plot.

2.1.6 Normal Q-Q plot

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption.

However, It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot.

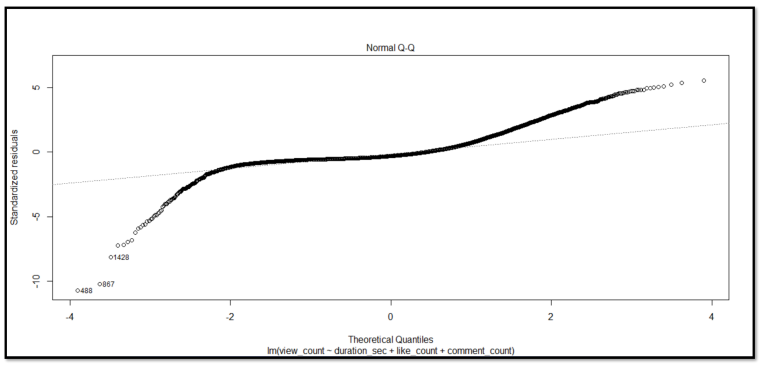

Figure 8: An example of Q-Q plot.

2.2 ANOVA (One-way and Welch's)

An ANOVA (Analysis of Variance) test is a statistical test used to determine whether there is a significant difference between two or more categorical groups by testing the difference in means using the variance.

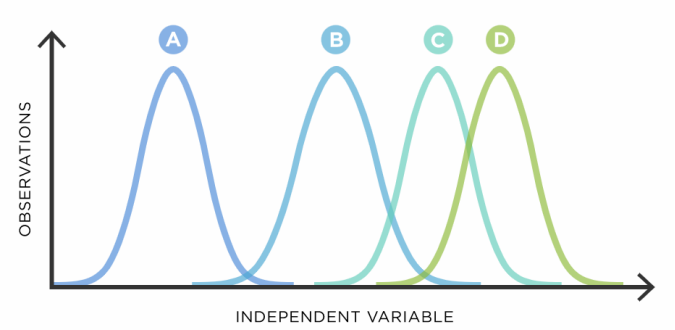

Figure 9: ANOVA diagram

One-way ANOVA has a categorical independent variable (a factor) and a normally distributed continuous dependent variable (such as the range or ratio level). Independent variables divide items into two or more mutually exclusive classes, categories, or groups.

For instance, One-way ANOVA is used to examine the purchasing patterns of several customers for the same product depending on factors like gender, age, income, geography, and so on.

Let I will be the number of treatments and μ_1, μ_2, ..., μ_I are the expected values of population 1, 2, ..., I. Wanting to test the following:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_I & \text{(Null hypothesis)} \\ H_1 : \exists\ \mu_i, \mu_j, (i \neq j) & \text{(Alternative hypothesis)} \end{cases}$$

Having the following assumptions:

+ The data are random and independent.
+ Also following the Normal distribution.
+ Homogeneity of variance between groups of independent variables.
  (The variance is approximately the same for each group of samples)

Sum of squares:

$$SST = \sum_i \sum_j (x_{ij} - \overline{x})^2$$

$$df(SST) = N - 1$$

$$SSTr = \sum_i \sum_j (\overline{x_i} - \overline{x})^2$$

$$df(SSTr) = I - 1$$

$$SSE = \sum_i \sum_j (x_{ij} - \overline{x_j})^2$$

$$df(SSE) = N - I$$

$$SST = SSE + SSTr$$

Where: SST is the Total Sum of Squares, SSTr is the Sum of Square of treatment, SSE is the sum of squares of error, $\overline{x}$ is the mean of all treatments, $\overline{x_i}$ is the mean of treatment i, N is the number of all observations.

Mean sum of errors:

$$MSE = \frac{SSE}{df(SSE)}$$

Mean sum of treatments:

$$MSTr = \frac{SSTr}{df(SSTr)}$$

Having the statistic $F = \frac{MSTr}{MSE}$. Rejecting H_0 if $F > F_{\alpha,I-1,N-1}$

The traditional ANOVA or **One-way ANOVA** still performs well when the data are normal, equal variances, and either balanced or unbalanced. To put it another way, use **Welch's ANOVA** if your data contains unequal variances, but utilize **One-way ANOVA** if the only problem is with the varying sample sizes.

The choice of method to use depends on the specific characteristics of the data and research question. In general, if the sample sizes of the groups being compared are roughly equal and the variances are only slightly different, you may still be able to use basic ANOVA, as it is fairly robust to violations of the equal variance assumption.

On the other hand, if the variances are substantially different or the sample sizes are unequal, it may be more appropriate to use the **Welch's ANOVA** or the Brown-Forsythe test. The **Welch's ANOVA** is a good option when the sample sizes are disparate or the variances are significantly different, whereas the Brown-Forsythe test is better suitable when the sample sizes are about similar but the variances are different.

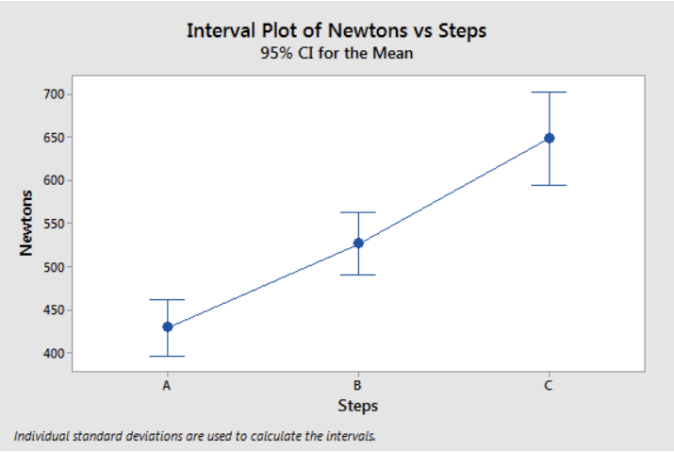

Figure 10: For these sample data, Welch's analysis of variance is a smart choice

The ground reaction forces produced by jumping off stairs of various heights are an example's data. Graph the data to understand the scenario. The chart above is an interval plot that displays the group means and 95% confidence intervals. The ranges, which are based on the unique standard deviations for each group, have various visual appearances.

2.3 Kruskal-Wallis Test

A non-parametric (distribution-free) alternative to the one-way ANOVA is the Kruskal-Wallis test, commonly referred to as the Kruskal-Wallis H test or the Kruskal-Wallis ANOVA. When the ANOVA assumptions are not satisfied or there is a considerable departure from the ANOVA assumptions, the Kruskal-Wallis test is useful. (It is preferable to utilize ANOVA since it is a little bit more potent than non-parametric tests if the data conforms to the ANOVA assumptions.)

When contrasting differences between two or more groups, the Kruskal-Wallis test is employed. For the purpose of computing test statistics and p values, the Kruskal-Wallis test makes no assumptions about any particular distribution (such as the normal distribution of samples).

The sample mean ranks or medians are compared in the Kruskal-Wallis test, which distinguishes it from the ANOVA, which compares sample means.

Kruskal-Wallis' Assumptions

+ The independent variable should have two or more independent groups.

+ The observations from the independent groups should be randomly selected from the target populations.

+ Each subject should only provide one response, and observations are collected separately from one another with no correlation between groups or within groups.

+ The dependent variable should be continuous or discrete.

Kruskal-Wallis test Hypothesis

- If each group distribution is not the same, H_0: All group means are equal versus H_1: At least, one group means different from other groups.

- In terms of medians (when each group distribution is the same), H_0: Population medians are equal versus H_1: At least, one population means different from other populations.

Kruskal-Wallis test statistic

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \right)$$

Where:

+ N is the total observation in all groups (total sample size).

+ k is the number of groups.

+ n_j is sample size for the i^{th} group.

+ R_j is the sum of ranks of j^{th} group.

H is approximately chi-squared distributed with $df = k - 1$. The p-value is calculated based on the comparison between the critical value and the H value. If $H \geqslant$ critical value, rejecting the null hypothesis and vice versa.

Dunn Test for Multiple Comparisons

Dunn Test should be used to determine the exact groups that differ when the Kruskal-Wallis test findings are statistically significant.

A non-parametric pairwise multiple comparison methods called the Dunn test is used to compare the medians of three or more groups. It is often referred to as the Dunn-Bonferroni test or the Bonferroni-Dunn test. It is employed when the assumption of equal variances fails or when non-normal data is present. Additionally, the Tukey HSD (using the results of the One-way ANOVA test to determine which group is significantly different) and the Dunn test have a similar basis in concept.

The Dunn test is typically used after rejecting the null hypothesis in a Kruskal-Wallis test, which tests whether there are any significant differences between the groups. If the Kruskal-Wallis test is significant, the Dunn test is then used to determine which groups differ significantly from each other.

The conditions to use the Dunn test

+ The data are non-normally distributed or violate the assumption of equal variances.

+ There are three or more groups to compare.

+ The Kruskal-Wallis test is significant.

It is commonly recommended to use the Dunn test only when there is a strong reason to believe that the data violate the assumptions of parametric tests like the ANOVA and Tukey HSD test.

2.4 Z-test

Z-test is Parametric Test, where the Null Hypothesis is less than, greater than, or equal to some value. A z-test is used if the population variance is known, or if the sample size is larger than 30, for unknown population variance.

We will use z-test for two samples to determine whether the means of two independent samples are significantly different from each other. The test involves calculating the difference between the sample means and comparing it to the expected difference.

$$\textbf{z score} = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where:

+ $\overline{x_1} - \overline{x_2}$: Different between sample mean.

+ $\mu_1 - \mu_2$: Different between population mean.

+ σ_1, σ_2 : Population standard deviation.

+ n_1, n_2 : Sample size.

2.5 Other methods

2.5.1 Shapiro-Wilk test

The Shapiro–Wilk test can be used to decide whether or not a sample fits a normal distribution, and it is commonly used for small samples.

How Does the Shapiro-Wilk Test Work?

The Shapiro-Wilk test first quantifies the similarity between the observed and normal distributions as a single number: it superimposes a normal curve over the observed distribution as shown below. It then computes which percentage of our sample overlaps with it: a similarity percentage.

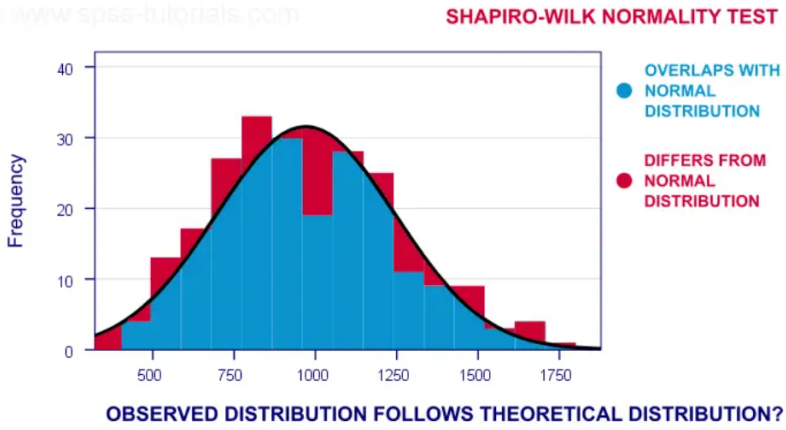

Figure 11: How Shapiro-Wilk Test Work.

The test gives you a W value; small values indicate your sample is not normally distributed (you can reject the null hypothesis that your population is normally distributed if your values are under a certain threshold). The formula for the W value is:

$$\mathbf{W} = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

We wish to test the following hypothesis:

$$\begin{cases} H_0 : \text{The sample is normally distributed.} \\ H_1 : \text{The sample is not normally distributed.} \end{cases}$$

If the p-value is greater than α, we fail to reject H_0 and conclude it follows normal distribution; otherwise, the sample is not normally distributed.

2.5.2 Kolmogorov-Smirnov Test

The One-Sample **Kolmogorov–Smirnov test** procedure compares the observed cumulative distribution function for a variable with a specified theoretical distribution, which may be normal, uniform, Poisson, or exponential. The **Kolmogorov–Smirnov test** is computed from the largest difference (in absolute value) between the observed and theoretical cumulative distribution functions. This goodness-of-fit test tests whether the observations could reasonably have come from the specified distribution.

```
ks.test(sample,"pnorm")
```

2.5.3 Levene Test

Levene's test is an inferential statistic used in statistics to assess if the variances of a variable obtained for two or more groups are equal. According to several common statistical techniques, the variances of the populations from which different samples are chosen are equal. This assumption is tested using Levene's test.

The homogeneity of variance or homoscedasticity null hypothesis, which states that the population variances are identical, is examined. When k is greater than two samples, it compares the variances of those samples.

Given a variable Y with sample size of N is divided into k subgroups, where N_i is the sample size of the i^{th} subgroup, the Levene Test statistic is defined as:

$$W = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} N_i(\overline{Z_i} - \overline{Z})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(\overline{Z_{ij}} - \overline{Z_i})^2}$$

Where Z_{ij} can have one of the following three definitions:

+ $Z_{ij} = |Y_{ij} - \overline{Y_i}|$ where $\overline{Y_i}$ is the **mean** of the i^{th} subgroup.
+ $Z_{ij} = |Y_{ij} - \hat{Y}_i|$ where $\hat{Y}_i$ is the **median** of the i^{th} subgroup.
+ $Z_{ij} = |Y_{ij} - \overline{Y_i'}|$ where $\overline{Y_i'}$ is the 10% **trimmed mean** of the i^{th} subgroup.

$\overline{Z_i}$ are the group means of the Z_{ij} and $\overline{Z}$ is the overall mean of the Z_{ij}.

The three choices for defining Z_{ij} determine the robustness and power of the Levene's test. When we talk about robustness, we imply the test's capacity to avoid misidentifying uneven variances when the underlying data are not normally distributed and the variables are actually equal. The test's power is its capacity to identify unequal variances when they actually exist.

3 Data analysis

3.1 Data description

We are permitted to find information pertaining to our primary area of specialization, which is Computer Science and Engineering, in this activity. Therefore, we choose the *BBC YouTube Videos Metadata*.

Figure 12: BBC Channel

- Using YouTube Data Tools one can access the metadata for the BBC channel and utilize this amazing data set to analyze the impact of categories, videos, comments, likes, dislikes, and so on. The data set contains 20 attributes.
- Collecting video metadata over a course of 13 years starting from 2007.

3.2 Importing Data

Firstly, we need to include some necessary libraries and read the data. In order to read the data set and return the normal data set in R, we utilize the command **read.csv()** and then print the data frame to observe by **head(youtube,5)**.

```
1 library(dplyr)
2 youtube<- read.csv("D:/Phúc/IT university/General/probability and
  statistics/assignment/bbc.csv")
3 head(youtube,5)
```

Here the result of the running code above:

```
> youtube<- read.csv("D:/Phúc/IT university/General/probability and statistics/assignment/bbc.csv")
> head(youtube,5)
  position               channel_id channel_title    video_id
1        1 UCCj956IF62FbT7Gouszaj9w           BBC 8qHOpGdjB_U
2        2 UCCj956IF62FbT7Gouszaj9w           BBC 1qeS-rOoBSw
3        3 UCCj956IF62FbT7Gouszaj9w           BBC JMfkBavl1ks
4        4 UCCj956IF62FbT7Gouszaj9w           BBC T_6RRmkLOSs
5        5 UCCj956IF62FbT7Gouszaj9w           BBC 3-mayD_9Yg8
          published_at
1 2020-08-13T15:00:02Z
2 2020-08-13T14:30:04Z
3 2020-08-13T05:50:21Z
4 2020-08-12T13:00:13Z
5 2020-08-12T11:00:02Z
                                                                 video_title
1                 Colin Robinson's Origins of the Species - What We Do In The Shadows | BBC
2                      Maisie Smith and Zack Morris on EastEnders' latest teen wedding - BBC
3 A-level results to arrive in year with no exams - Covid-19: Top stories this morning - BBC
4                                    8 signs you're in survival mode and how to start living - BBC
5                            The secret Heathrow lounge that costs □□2700 just to get in! | QI -

1
2
3 Subscribe and □□□□ to OFFICIAL BBC YouTube □□□□ https://bit.ly/2IXqEIn Stream original BBC programmes FIRST on BBC iPlayer □□□□ https://bbc.in/2J18jYJ □□□□ Subscribe and □□□□ @BBC News
4                                                                                 Subscribe and □□□□ to OFFICIAL BBC YouTube □□□□ https://
5
  video_category_id video_category_label duration duration_sec dimension
1                24        Entertainment  PT5M23S          323        2d
2                24        Entertainment  PT3M15S          195        2d
3                27            Education PT14M48S          888        2d
4                27            Education  PT3M50S          230        2d
5                24        Entertainment  PT1M52S          112        2d
  definition licensed_content view_count like_count dislike_count
1         hd                1        738         76             7
2         hd                1        512         55             9
3         hd               NA      19888        326            50
4         hd                1      14515        324           532
5         hd                1      15644        331            14
  favorite_count comment_count
1              0             4
2              0            13
3              0           128
4              0           282
5              0            22
```

Figure 13: The data frame originally

As can be shown from the figure, we have several terms that can be explained and understood:

- **position:** stands for numerical order.
- **channel_id:** determines the address of the channel (BBC).
- **channel_title:** gives the name of the channel conducted (BBC).
- **video_id:** provides the address of a published video.
- **published_at:** determines the specific date introducing a video to the public.
- **video_title:** stands for the name of a video created by a publisher.
- **video_description:** provides information that a publisher wants to convey to viewers.
- **video_category_id:** stands for the code of a video type.
- **video_category_label:** determines a specific type of video.
- **duration:** gives us the length of a video
- **duration_sec:** converts the duration into seconds.
- **dimension:** defines whether a video is 2D or 3D.
- **definition:** defines its quality, HD (high definition) or SD (standard definition).
- **caption licensed_content:** defines whether a video has copyright.
- **view_count:** determines the number of views gained from the public.
- **like_count:** provides the total figure for "like" gained from the public.
- **dislike_count:** defines how many people pressed the "dislike" button.
- **favorite_count:** determines the number of people marking a video as his/her favorite one.
- **comment_count:** stands for the total comments from viewers.

3.3 Data cleaning

Based on the input file, we consider several columns that are unuseful, so we decide to remove those with the following reasons:

- **video_id**, **video_title**, **video_description:** They provide other information that has nothing to do with the number of total views as well as not enough evidence to analyze.
- **published_at:** Since BBC is an internationally popular channel, so in case we opt this for analysis, we cannot conclude standard time for publishing to seek the highest viewers corresponding to all countries around the world.
- **channel_id** and **channel_title:** They are all the same with all rows among the data as it is conducted from BBC channel on YouTube platform.
- **duration:** This column is basically similar to **duration_sec**, which already converted into seconds, so we can base on **duration_sec** instead for analysis.
- **position:** This column just tracks the order of the data collected in the given file.
- **video_category_id:** Due to the fact that this column *video_ category_ id* and *video_ category_ label* determines the same object as the former is the code number standing for the latter, e.g. *video_ category_ id* 24 is corresponding to *video_ category_ label* Entertainment. Therefore, our team decides to analyze directly on the column *video_ category_ label* instead.
- **favorite_count:** This column shows all data values defined as Zero, so it is not evident to use.
- **dimension:** With this column, the data just contains 2 different values named *2D* and *3D*, but the latter one only appears 1 time over 10,000 figures from the data by using R shown in below:

```
table(youtube$dimension)
```

```
> table(youtube$dimension)

    2d     3d
12455      1
```

Figure 14: The number of 2D and 3D datas

Then we use the following segment code to remove these columns.

```
> youtube <-(youtube [, c(9,11,13,14,15,16,17,19)])
```

Because the videos which are not licensed content having the NULL value, so we have to replace it to 0 value.

```
youtube["licensed_content"][is.na(youtube["licensed_content"])]<- 0
```

Using the command **is.na(youtube)** will return a new data frame which has null value. Therefore, the *sum* command can be used to calculate the total number of rows having null value.

We can observe that there are only 99 cells of null value and this proportion is absolutely small. Consequently, we can delete those null value using the command **na.omit(youtube)** to create a new data frame which has no null value.

```
> sum(is.na(youtube))
[1] 99
> youtube<- na.omit(youtube)
> sum(is.na(youtube))
[1] 0
```

Then we decide to remove some categories in **video_category_label** which have a very small quantity of data, due to high precision of our conclusion. The larger the sample size, the more accurate the average values will be. Larger sample sizes also help researchers identify outliers in data and provide smaller margins of error. These following code help us show the frequency of each category and arrange it in descending order.

```
>freq_table <- table(youtube$video_category_label)
>freq_df <- as.data.frame(freq_table)
>names(freq_df) <- c("label", "count")
>freq_df <- freq_df[order(-freq_df$count),]
>head(freq_df, 15)
```

We have the following result:

```
> head(freq_df, 15)
                      label count
4             Entertainment  9128
2                    Comedy  1589
8                     Music   659
12          Pets & Animals    276
13  Science & Technology      206
9          News & Politics    201
15         Travel & Events     71
3                 Education     67
7           Howto & Style      44
14                  Sports     44
5         Film & Animation     29
11          People & Blogs     22
1         Autos & Vehicles     15
10 Nonprofits & Activism       7
6                   Gaming       1
```

Figure 15: The frequency of each category in descending order

And we just keep the category that have more than 200 data. These are Entertainment, Comedy, Music, Pets & Animals, Science & Technology and News & Politics.

```
1  >youtube <- youtube[youtube$video_category_label %in% c("Entertainment",
   "Comedy", "Music", "Pets & Animals", "Science & Technology", "News & Politics" ),
   ]
```

Because in this assignment our dependent variables is **view_count**, we need to remove outliers of this variable. First of all, we need to view the distribution of this variable.

```
1  > hist(youtube$view_count,xlab="ViewCount" ,probability=TRUE,main="Distribution of
   View count")
2  > lines(density(youtube$view_count), col="blue",lwd=2)
```

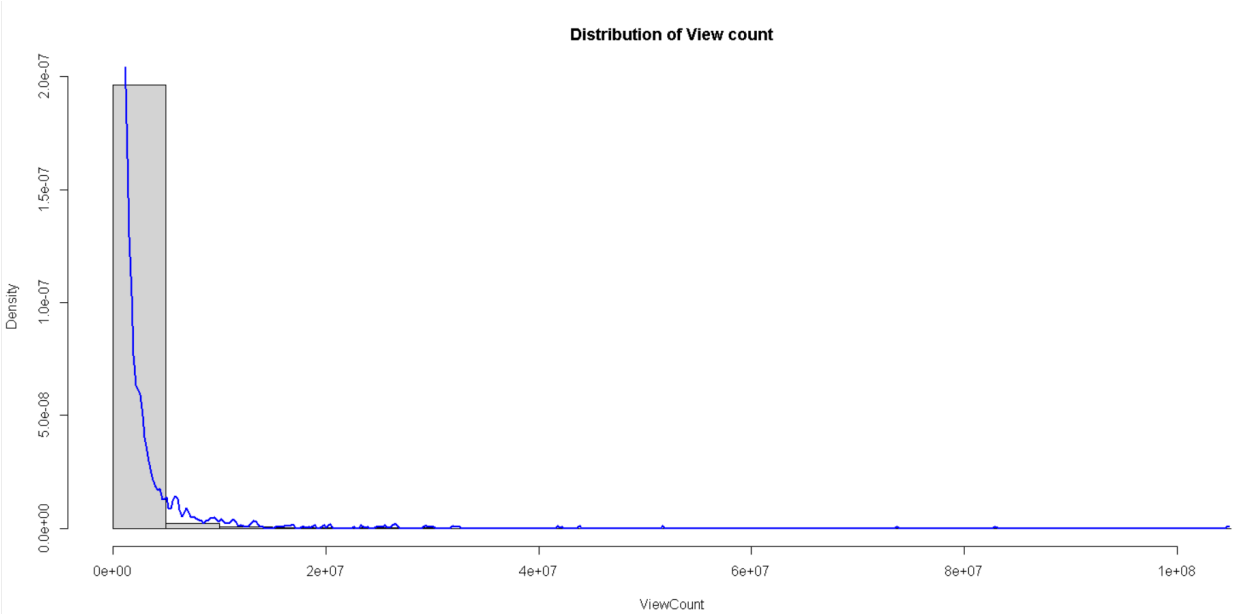

Figure 16: Histogram of **view_count** variable

As we can see, the view count does not distribute uniformly, it is mainly in the range from 0 to 1000000 and there are also some values are extremely higher. Therefore, we need to remove these outliers before analyzing them. In this case we use IQR method rule to find outliers.

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

Firstly, we get the IQR, first quartile and third quartile of **view_count** variable.

```
> summary(youtube$view_count)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
      512     45394    105162    551496    324184 104776527
> IQR(youtube$view_count)
[1] 278790.5
```

After calculating, we just need to remove all the data whose view count is higher than 742369.

```
>youtube<-youtube[youtube$view_count<=742369,]
```

And then we get the distribution of **view_count** variable after removing outliers.

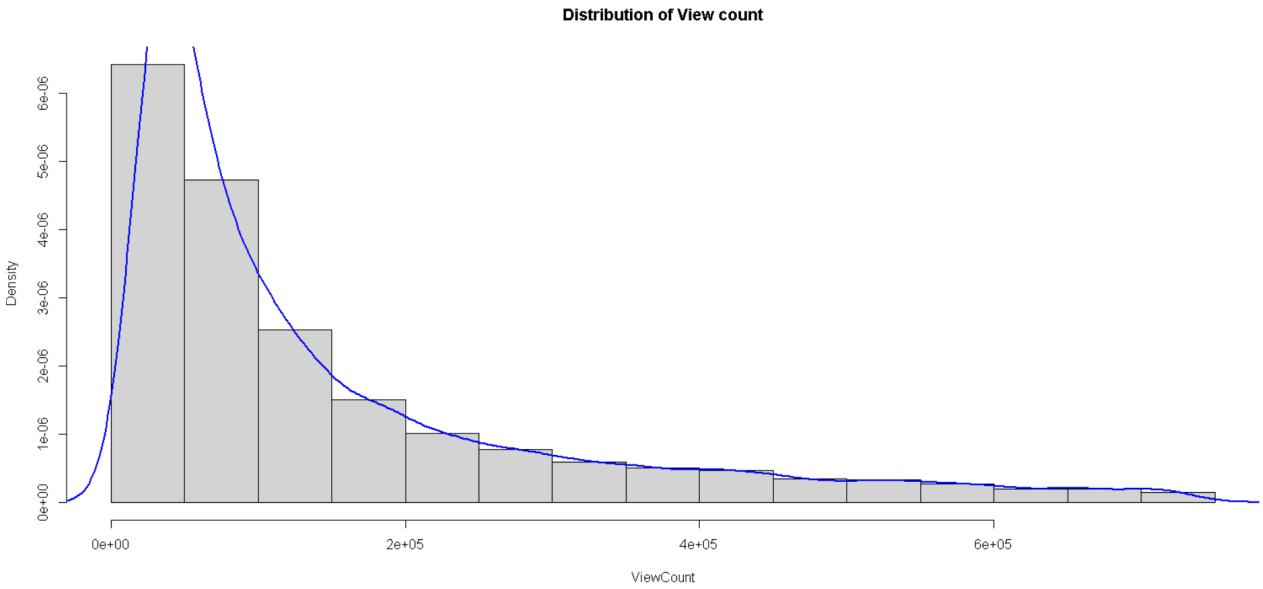

Figure 17: Histogram of **view_count** variable

After data processing, we can obtain the final filtered data frame below:

	video_category_label	duration_sec	definition	licensed_content	view_count	like_count	dislike_count	comment_count
1	Entertainment	323	hd	1	738	76	7	4
2	Entertainment	195	hd	1	512	55	9	13
5	Entertainment	112	hd	1	15644	331	14	22
6	News & Politics	965	hd	0	29095	462	47	159
7	Entertainment	824	hd	1	22185	322	36	66
8	News & Politics	651	hd	0	36984	476	64	201
9	Entertainment	288	hd	1	20301	364	26	29
10	News & Politics	731	hd	0	47528	533	233	549
11	Entertainment	231	hd	1	14594	188	22	34
12	News & Politics	845	hd	0	49028	474	282	799
13	Entertainment	1263	hd	1	34522	1168	32	155
14	News & Politics	1077	hd	0	52033	579	72	616
15	Science & Technology	405	hd	0	29402	505	73	144
16	Entertainment	73	hd	1	24329	399	134	145
17	Entertainment	211	hd	1	36823	1052	109	359
18	News & Politics	662	hd	0	62798	603	94	273
19	Comedy	309	hd	1	22033	252	36	33
21	Entertainment	108	hd	1	35203	336	46	89
22	News & Politics	886	hd	0	84551	911	53	227
23	Entertainment	254	hd	1	21845	463	34	59

Figure 18: The final source of data achieved after processcing

4 Data visualization

4.1 Analyzing the boxplot with the discrete variables

- Analyze **video_category_label**

```
> boxplot(view_count~video_category_label, data=youtube,
xlab="video_category_label", ylab="ViewCount", main="Distribution between view
count and category")
```

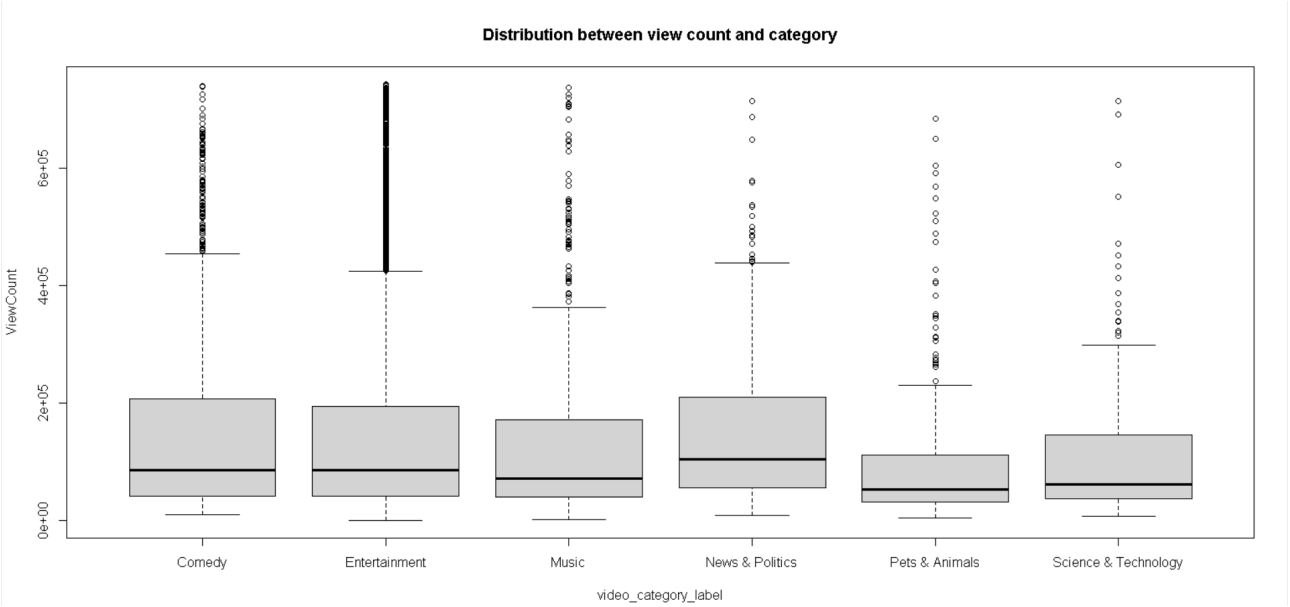

Figure 19: Boxplot of category

```
> table(youtube$video_category_label)
                Comedy          Entertainment                  Music
                  1396                   7901                    588
      News & Politics         Pets & Animals Science & Technology
                   192                    215                    184

```

The number of videos about entertainment accumulates the highest proportions. However, we can not conclude which category has the highest view by boxplot, so we need to perform a test later.

- Analyze **definition**

```
> boxplot(view_count~definition, data=youtube, xlab="definition",
ylab="ViewCount", main="Distribution between view count and definition")
```

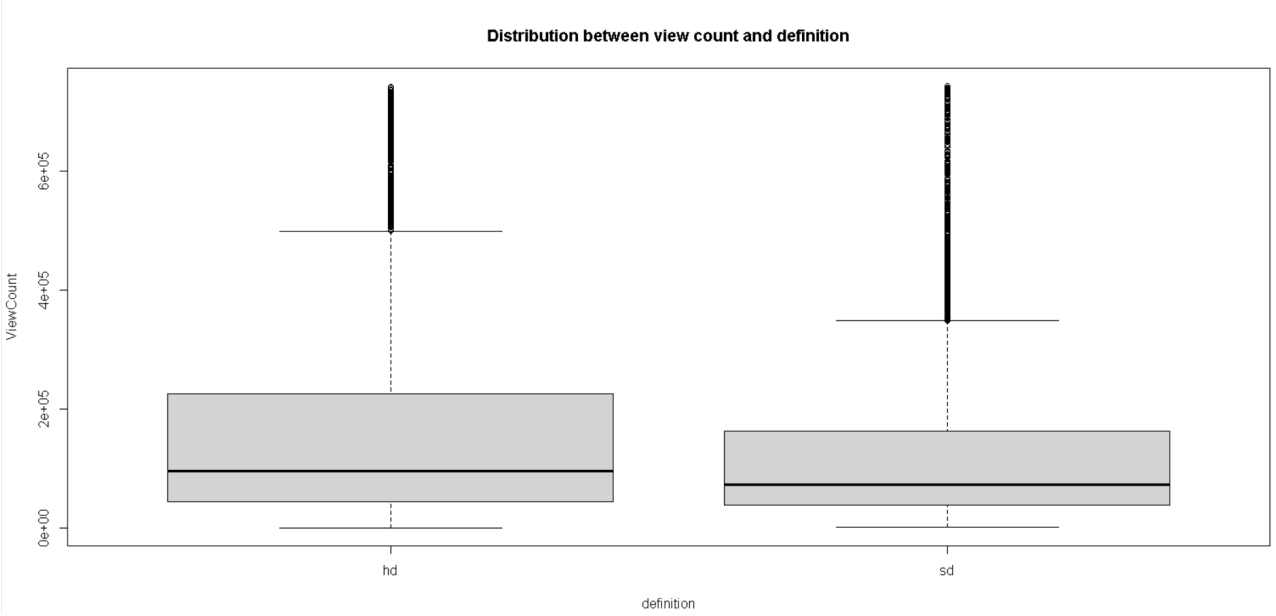

Figure 20: Boxplot of definition

```
> table(youtube$definition)
  hd   sd
5314 5162
```

The number of hd and sd videos are nearly the same. Although the boxplot indicates that hd videos have the higher view than sd videos, we can not sure anything until perform the test.

- Analyze **licensed_content**

```
> boxplot(view_count~licensed_content, data=youtube, xlab="licensed content",
ylab="ViewCount", main="Distribution between view count and licensed content")
```

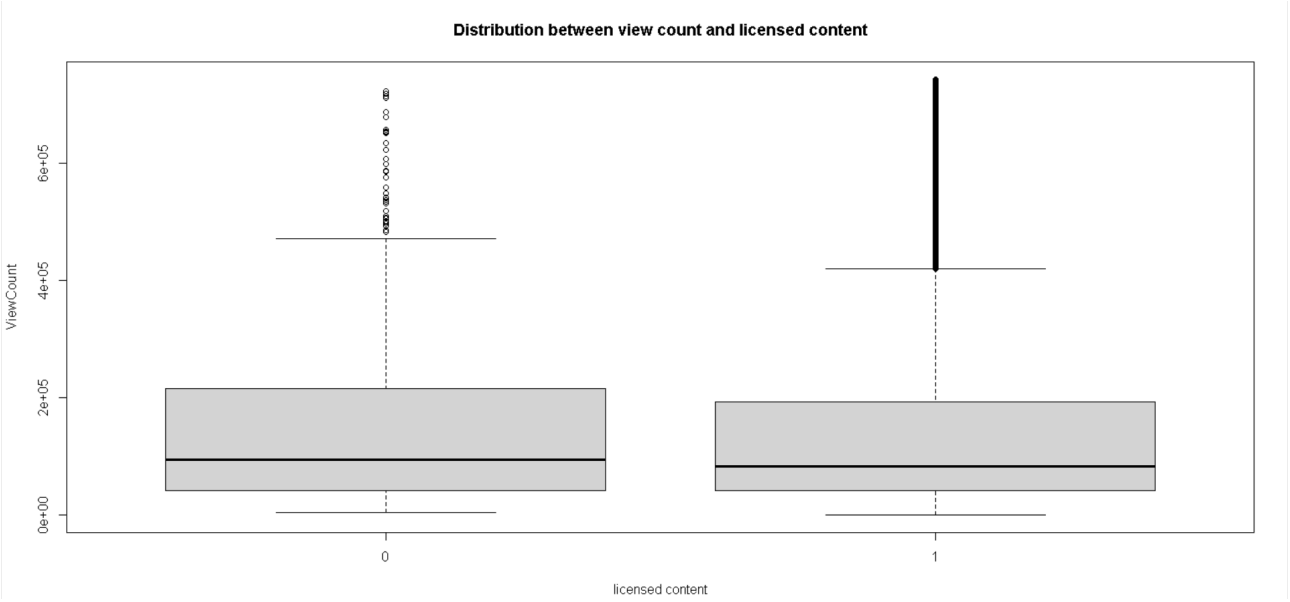

Figure 21: Boxplot of licensed content

```
> table(youtube$licensed_content)
    0     1
  498 9978
```

As we can see most of the videos are licensed. However, we also need to perform a test to clarify which type attracts more viewers and whether this variable really affect the view count of a video.

4.2 Analyzing the continuous variables

There are 4 continuous variables: **duration_sec**, **like_count**, **dislike_count**, **comment_count**. And we will use scatter plot to analyze these variables.

```
> plot(view_count ~ duration_sec, data=youtube)
> plot(view_count ~ like_count, data=youtube)
> plot(view_count ~ dislike_count, data=youtube)
> plot(view_count ~ comment_count, data=youtube)
```

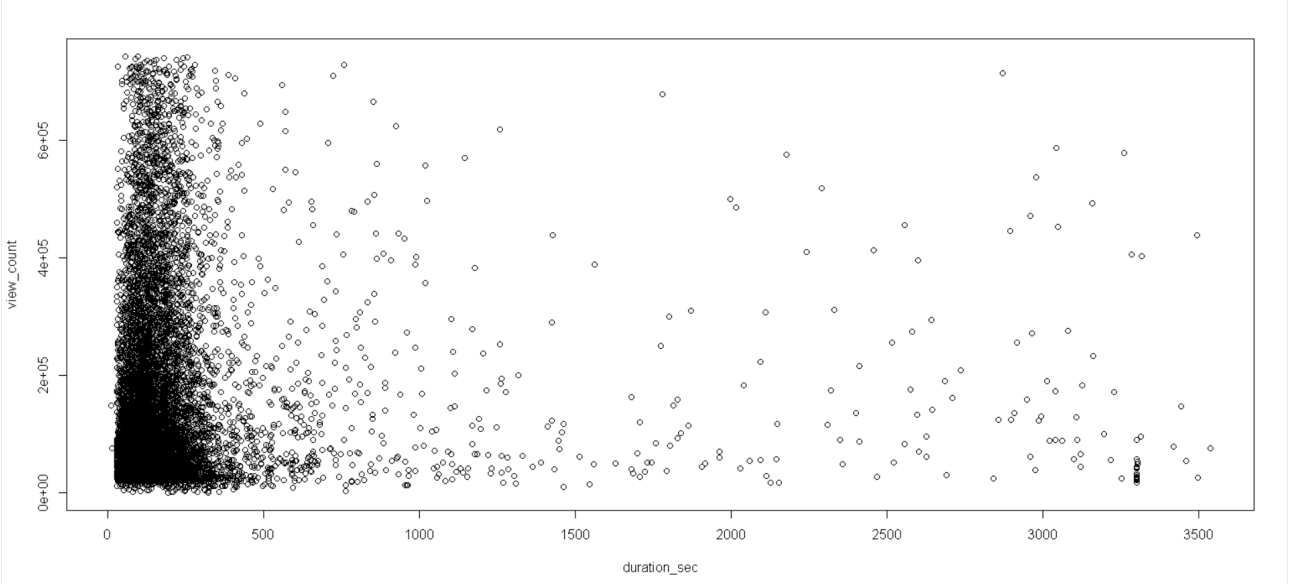

Figure 22: Scatter plot of duration

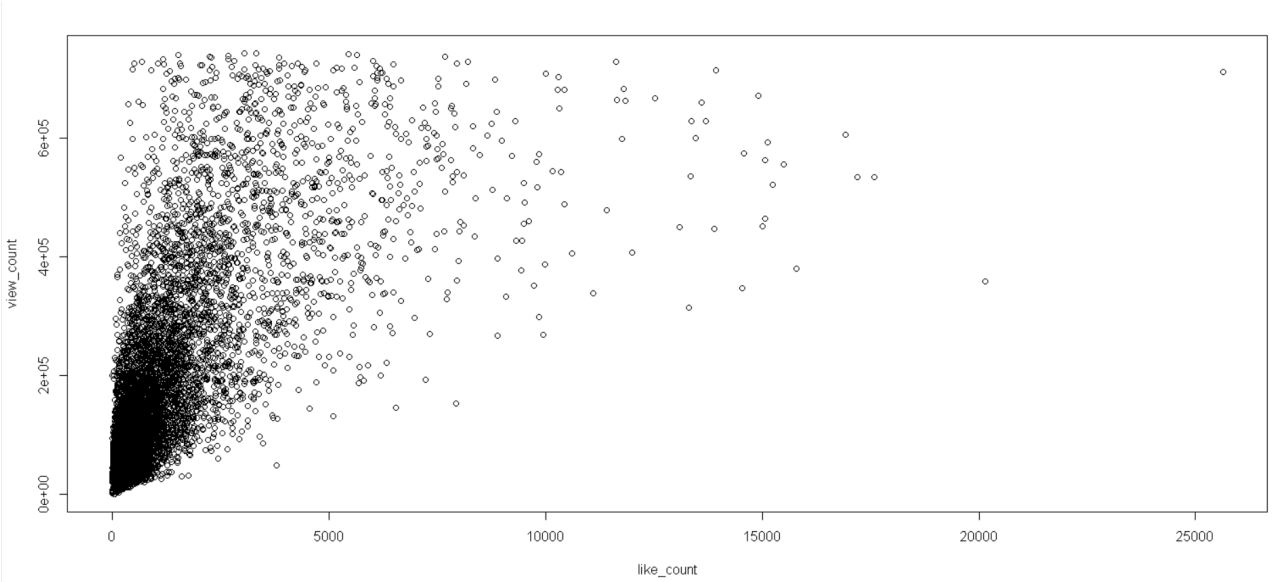

Figure 23: Scatter plot of like

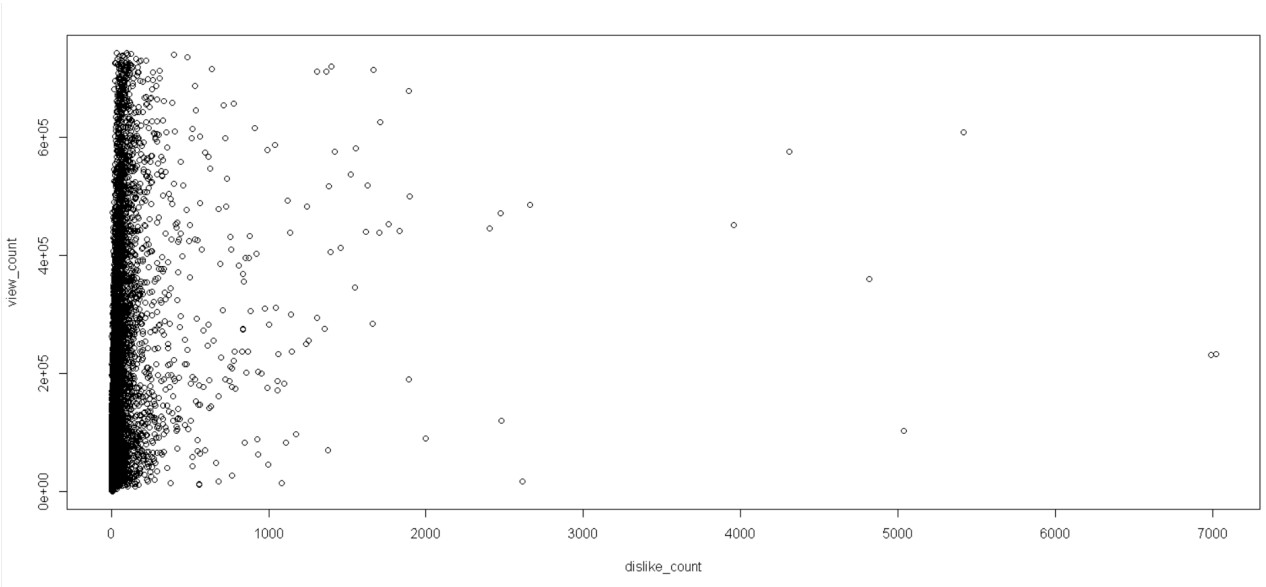

Figure 24: Scatter plot of dislike

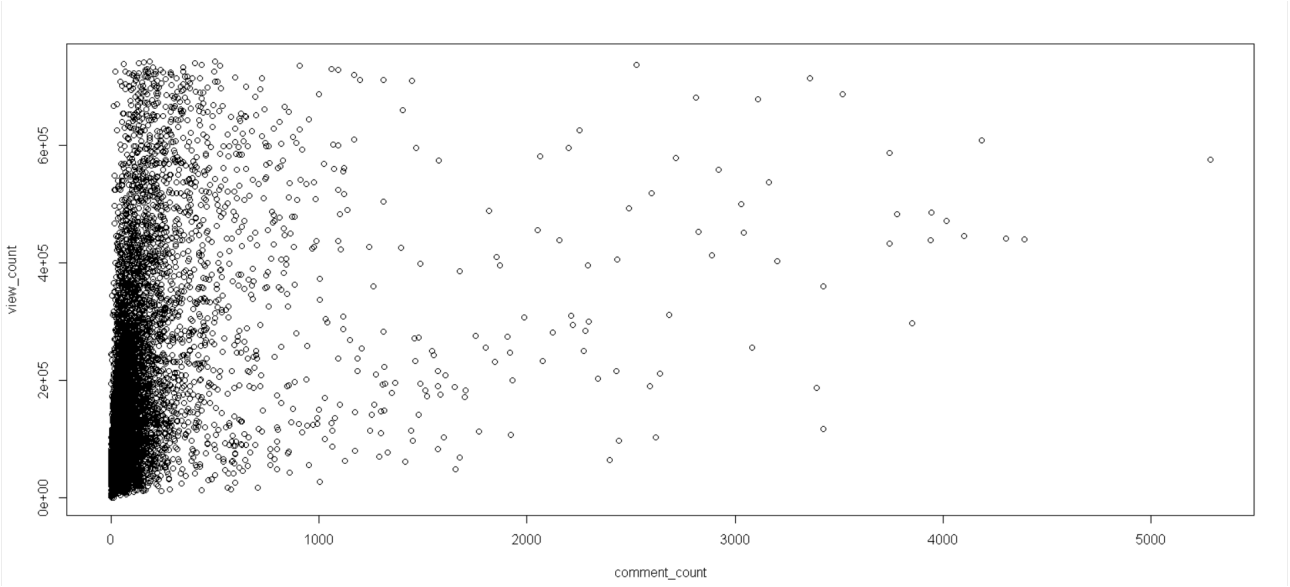

Figure 25: Scatter plot of comment

Four scatter plots show that **like_count** seem to be linearly related to view count. The higher likes are, the higher views are. And the rest of the variables seem not to correlate with the view count. However, we also need to perform a test to check the correlation between them.

4.3 Summary Data

```
> summary(youtube)
```

```
video_category_label  duration_sec      definition         licensed_content
Length:10476          Min.   :  11.0   Length:10476       Min.   :0.0000
Class :character      1st Qu.:  89.0   Class :character   1st Qu.:1.0000
Mode  :character      Median : 128.0   Mode  :character   Median :1.0000
                      Mean   : 198.2                      Mean   :0.9525
                      3rd Qu.: 199.0                      3rd Qu.:1.0000
                      Max.   :3536.0                      Max.   :1.0000
  view_count        like_count      dislike_count     comment_count
Min.   :   512   Min.   :    2   Min.   :   0.00   Min.   :   0.0
1st Qu.: 41036   1st Qu.:  163   1st Qu.:   7.00   1st Qu.:  21.0
Median : 83527   Median :  406   Median :  16.00   Median :  45.0
Mean   :149113   Mean   :  963   Mean   :  51.19   Mean   : 118.4
3rd Qu.:193616   3rd Qu.: 1062   3rd Qu.:  40.00   3rd Qu.: 107.0
Max.   :742108   Max.   :25629   Max.   :7022.00   Max.   :5284.0
```

Figure 26: Summary Data

The mean and the median of **view_count** variable are significantly different, so it may not be a normal distribution. However, performing test to test whether it is normal distribution or not is also needed to apply other methods.

5 ANOVA Test

5.1 Initialize variable to store the data

This section's major goal is to evaluate the correlation between each category, definition, and licensed content of video material, as well as the view counts for each one, and to determine how they all relate to one another. The **ANOVA** will be used to determine whether or not the sample groups' means are equal under the null hypothesis.

Applying **ANOVA** to analyze instead of analyzing in the usual way usually because we need to compare the average view counts of a data set based on multiple types of videos and to analyze multiple videos at the same time for instance. This method is more convenient than other methods because it is possible to quickly conclude the equality or difference between the mean of many variables equal to or more than 2. After testing, having **View Counts** as dependent and **Categories** as independent variables, as well as **View Counts** as dependent and **Definitions** as independent variables.

First of all, some code must be written to setting up the data before **ANOVA**:

```
# Changing 6 string names of the categories to an integer type to use ANOVA
youtube$label_id[youtube$video_category_label == "Entertainment"] <- 1
youtube$label_id[youtube$video_category_label == "Comedy"] <- 2
youtube$label_id[youtube$video_category_label == "Music"] <- 3
youtube$label_id[youtube$video_category_label == "Pets & Animals"] <- 4
youtube$label_id[youtube$video_category_label == "Science & Technology"] <- 5
youtube$label_id[youtube$video_category_label == "News & Politics"] <- 6
# Changing 2 string names of the definition to an integer type to use ANOVA
youtube$definition[youtube$video_category_label == "hd"] <- 0
youtube$definition[youtube$video_category_label == "sd"] <- 1
# Set up the data
cate = youtube[, c('label_id')]
def = youtube[, c('definition')]
licensed = youtube[, c('licensed_content')]
view = youtube[, c('view_count')]
```

Hypothesis testing

- Random and independent statistics.

- The total view counts of videos of different categories, definitions, and licensed contents which are based on normal distribution.

- Variance of total view counts of different categories, definitions, and licensed contents are equivalent.

5.2 Hypothesis related to normal distribution

5.2.1 Kolmogorov–Smirnov test:

This is the code for K-S test for R:

```
cate1 = subset(youtube, youtube$label_id == 1)
cate2 = subset(youtube, youtube$label_id == 2)
cate3 = subset(youtube, youtube$label_id == 3)
```

```r
cate4 = subset(youtube, youtube$label_id == 4)
cate5 = subset(youtube, youtube$label_id == 5)
cate6 = subset(youtube, youtube$label_id == 6)
cate1_view <- unique(cate1$view)
ks.test(cate1_view, "pnorm")
cate2_view <- unique(cate2$view)
ks.test(cate2_view, "pnorm")
cate3_view <- unique(cate3$view)
ks.test(cate3_view, "pnorm")
cate4_view <- unique(cate4$view)
ks.test(cate4_view, "pnorm")
cate5_view <- unique(cate5$view)
ks.test(cate5_view, "pnorm")
cate6_view <- unique(cate6$view)
ks.test(cate6_view, "pnorm")
defi_0 = subset(youtube, youtube$definition == 0)
defi_0 <- unique(defi_0$view_count)
ks.test(defi_0, "pnorm")
defi_1 = subset(youtube, youtube$definition == 1)
defi_1 <- unique(defi_1$view_count)
ks.test(defi_1, "pnorm")
license_0 = subset(youtube, youtube$licensed_content == 0)
license_0 <- unique(license_0$view_count)
ks.test(license_0, "pnorm")
license_1 = subset(youtube, youtube$licensed_content == 1)
license_1 <- unique(license_1$view_count)
ks.test(license_1, "pnorm")
```

After we used K-S test for each category. These are the results:

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  cate1_view
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 27: Normality test for entertainment data.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  cate2_view
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 28: Normality test for comedy data.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  cate3_view
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 29: Normality test for music data.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  cate4_view
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 30: Normality test for animal and pets data.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  cate5_view
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 31: Normality test for science and technology data.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  cate6_view
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 32: Normality test for news and politics.

According to the results, the p-values of 6 categories are approximately to 0, so they are not normal distribution.

Similarly, we do the same for each type of definition and each value of licensed video. We also got the same results (It's shown below).

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  defi_0
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 33: Normality test for hd videos.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  defi_1
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 34: Normality test for sd videos.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  license_0
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 35: Normality test for licensed videos.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  license_1
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 36: Normality test for non-licensed videos.

5.2.2 Levene's Test:

This is the code for Levene's test in R:

```
leveneTest(youtube$view_count~factor(youtube$label_id))
leveneTest(youtube$view_count~factor(youtube$definition))
leveneTest(youtube$view_count~factor(youtube$licensed_content))
```

After running above code, this is the result:

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    5   3.589 0.003033 **
      10470
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 37: Levene test for categories.

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    1  87.614 < 2.2e-16 ***
      10474
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 38: Levene test for definition.

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value  Pr(>F)
group    1  2.8258 0.09279 .
      10474
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 39: Levene test for licensed content.

We use Levene Hypothesis to test the Homogenity of Variance.
- H0: Variance of total posts of each categories are equivalent.
- H1: There are at least 2 categories whose Variance of total posts are not equivalent.
P - value <=0.05: reject H0, accept H1.
P - value >0.05: can not reject H0, continue to use ANOVA method.
According to the above results, we can conclude that: For categories and definition, we can reject H0, but we can't reject it for licensed test.

5.3 One-way ANOVA

Correlation between licensed content and view counts

When using One-way ANOVA, having collected data about one categorical independent variable and one quantitative dependent variable. The independent variable should consist of two or more categorical, independent groups. Typically, the One-way ANOVA is used when having three or more categorical, independent groups, it also can be used for just two groups (An independent-sample T-test is more commonly used for two groups commonly).

Set a new value for the variable **anova1** when using One-way ANOVA. The variance table is as follows.

```r
# Use oneway.test() to perform a One-way ANOVA
aonva1 <- oneway.test(view_count ~ licensed, data = youtube, var.equal = TRUE)
# Print the result
print(aonva1)
```

In this case, by setting var.equal = **TRUE**, requesting a traditional ANOVA assuming that the variances are equal across categories. (If the assumption of equal variances is not met, the results of the basic ANOVA may not be reliable. However, if the assumption is met, a standard ANOVA may be more powerful than a Welch's ANOVA.)

```
        One-way analysis of means

data:  view_count and licensed
F = 2.9281, num df = 1, denom df = 10474, p-value = 0.08708
```

Figure 40: Result of **anova1**

Because $p - value > 0.05$ so we cannot reject H_0 : Therefore, cannot conclude that the 2 licensed content values whose means of view counts are not equivalent.

5.4 Welch's ANOVA

Correlation between categories and view counts

Set a new value for the variable **anova2** when using Welch's ANOVA. The variance table is as follows.

```r
# Assuming your data is stored in a data frame called "youtube"
# Use oneway.test() to perform a Welch's ANOVA
aonva2 <- oneway.test(view_count ~ cate, data = youtube, var.equal = FALSE)
# Print the result
print(aonva2)
```

Having a reason to believe that the variances are unequal, it is generally better to use **Welch's ANOVA**, as it is more robust to violations of the equal variance assumption. Doing this by setting var.equal = **FALSE**.

```
        One-way analysis of means (not assuming equal variances)

data:  view_count and cate
F = 6.6076, num df = 5.00, denom df = 698.21, p-value = 5.066e-06
```

Figure 41: Result of **anova2**

Because $p - value < 0.05$ so we reject H_0, accept H_1 : There are at least 2 categories whose means of view counts are not analogous.

Correlation between definition and view counts

Set a new value for the variable **anova3** when using Welch's ANOVA. The variance table is as follows.

```r
# Assuming your data is stored in a data frame called "youtube"
# Use oneway.test() to perform a Welch's ANOVA
aonva3 <- oneway.test(view_count ~ def, data = youtube, var.equal = FALSE)
# Print the result
print(aonva3)
```

Welch's ANOVA is often preferable when there is cause to suspect that the variances are not equal since it is more resistant to violations of the equal variance assumption. Setting var.equal = **FALSE** will do this.

```
        One-way analysis of means (not assuming equal variances)

data:  view_count and def
F = 108.61, num df = 1, denom df = 10371, p-value < 2.2e-16
```

Figure 42: Result of **anova3**

Because $p - value < 0.05$ so we reject H_0, accept H_1 : There are 2 definition values whose means of view counts are not identical.

▷ **On the whole**, it is a good practice to check the assumption of equal variances using a test such as Levene's test, and then choose an appropriate method based on the results of the test. If the assumption is met, a normal ANOVA may be used. Otherwise, Welch's ANOVA or another non-parametric test may be used.

6 Kruskal-Wallis Test

Using the Kruskall-Wallis test when countering assumptions' failures in ANOVA. The test will be used to determine whether or not the sample groups' medians are equal under the null hypothesis. After testing, having **View Counts** as dependent and **Categories** as independent variables, as well as **View Counts** as dependent and **Definitions** as independent variables.

Correlation between categories and view counts

```
# Affected factor ~ influencing factor
kruskal.test(view ~ cate, data = youtube)
```

```
        Kruskal-Wallis rank sum test

data:  view by cate
Kruskal-Wallis chi-squared = 47.796, df = 5, p-value = 3.91e-09
```

Figure 43: Result of view counts by categories

From the Kruskal-Wallis test, the $p - value < 0.05$ indicates that there are significant differences in the view counts among the 6 categories.

Correlation between definition and view counts

```
# Affected factor ~ influencing factor
kruskal.test(view ~ def, data = youtube)
```

```
        Kruskal-Wallis rank sum test

data:  view by def
Kruskal-Wallis chi-squared = 109.54, df = 1, p-value < 2.2e-16
```

Figure 44: Result of view counts by definition types

From the Kruskal-Wallis test, the $p - value < 0.05$ indicates that there are remarkable variations in the view counts among the definition types.

Correlation between licensed content and view counts

```
# Affected factor ~ influencing factor
kruskal.test(view ~ licensed, data = youtube)
```

```
        Kruskal-Wallis rank sum test

data:  view by licensed
Kruskal-Wallis chi-squared = 1.8317, df = 1, p-value = 0.1759
```

Figure 45: Result of view counts by licensed content

From the Kruskal-Wallis test, the $p - value > 0.05$ indicates that there are no distinctions in the view counts among the licensed content.

▷ **To sum things up**, choosing an appropriate method based on the results of the test. When the assumption is not met (the data is non-normal distribution), using **the Kruskal-Wallis test**. After performing the test there are significant differences in the view counts among the definition types and 6 categories. Then let analyze which definition types or category has the highest view.

7 Which has the highest view among 6 categories

After we can conclude that at least one of the means is different among 6 categories, then we perform **Dunn-test** to determine which pairs of means are significantly different from each other.

```
 > dunnTest(view_count ~ video_category_label, data=youtube, method="bonferroni")
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Bonferroni method.

                                  Comparison         Z      P.unadj        P.adj
1                  Comedy - Entertainment  0.848723 3.960355e-01 1.000000e+00
2                           Comedy - Music  2.098153 3.589162e-02 5.383744e-01
3                    Entertainment - Music  1.836674 6.625804e-02 9.938706e-01
4                 Comedy - News & Politics -2.104363 3.534682e-02 5.302024e-01
5          Entertainment - News & Politics -2.554989 1.061911e-02 1.592867e-01
6                   Music - News & Politics -3.189689 1.424261e-03 2.136392e-02
7                   Comedy - Pets & Animals  5.470588 4.485451e-08 6.728176e-07
8            Entertainment - Pets & Animals  5.441926 5.270764e-08 7.906146e-07
9                     Music - Pets & Animals  3.734589 1.880219e-04 2.820328e-03
10         News & Politics - Pets & Animals  5.667651 1.447681e-08 2.171522e-07
11            Comedy - Science & Technology  2.807952 4.985765e-03 7.478648e-02
12     Entertainment - Science & Technology  2.622671 8.724344e-03 1.308652e-01
13              Music - Science & Technology  1.385947 1.657632e-01 1.000000e+00
14 News & Politics - Science & Technology  3.704744 2.116044e-04 3.174066e-03
15  Pets & Animals - Science & Technology -1.797968 7.218215e-02 1.000000e+00
```

From the table we can see that these following combinations are statistically significant difference from one another:
Comedy - Pets & Animals
Entertainment - Pets & Animals
News & Politics - Pets & Animals

Therefore, **Comedy**, **Entertainment**, **News & Politics** are 3 categories that are more popular than the others.

8 Which has the higher views among 2 definition types

In this section, we will perform z-test. It compares the means between the group to see which group has the higher mean.
H_0 - **hd** type view is less than or equal the **sd** type.
H_1 - **hd** type has the higher view than **sd**.

```
> hd<-subset(youtube, definition=="hd")$view_count
> sd<-subset(youtube, definition=="sd")$view_count
> sd(hd)
[1] 166248.9
> sd(sd)
[1] 146077.2
> z.test(x=hd,y=sd,alternative='greater',mu=0,
        sigma.x=166248.9,sigma.y=146077.2,conf.level=0.95)

        Two-sample z-Test

data:  hd and sd
z = 10.422, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 26816.2      NA
sample estimates:
mean of x mean of y
 164802.6  132960.9

```

Firstly, we create 2 **hd** and **sd** samples and then getting the standard deviation of each sample to pass to the z-test function. After performing z-test, p-value <0.05, then we reject H_0 and accept H_1. Therefore, we can conclude that **hd** type has the higher view than **sd**.

9 Multiple Linear Regression model

9.1 Overview

In this section, we are intending to build a linear regression model for researching the dependency of each variable on the others.

During the whole process, we will need to keep in mind the following criteria:

Linearity: Predictors in the model have a straight-line relationship with the dependent variable.

Normality: The residuals of the model should follow the normal distribution.

Homogeneity: Standard deviations are equal for all observations.

9.2 Multi linear regression model

According to the scatter plot above, we can see that the data we have maybe related to Linear regression.

We will then build the linear regression model which has *view_ count* as a dependent variable with the confidence level of 0.05.

We will firstly use the **lm()** command to build the linear regression model.

```
# Create a linear model
LRmodel1<-lm(view_count ~ duration_sec + like_count + dislike_count +
comment_count, data = youtube)

summary(LRmodel1)
```

After that, by using the **summary()** command, it will help us to demonstrate the statistic of the above model.

```
> LRmodel1<-lm(view_count ~ duration_sec + like_count + dislike_count + comment_count, data = youtube)
> summary(LRmodel1)

Call:
lm(formula = view_count ~ duration_sec + like_count + dislike_count +
    comment_count, data = youtube)

Residuals:
     Min       1Q   Median       3Q      Max
-1161981   -57373   -32819    26585   608508

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   81624.9209  1401.2164  58.253   <2e-16 ***
duration_sec    -32.7957     3.6460  -8.995   <2e-16 ***
like_count       65.8394     0.7513  87.632   <2e-16 ***
dislike_count     3.3147     7.6401   0.434    0.664
comment_count    87.9476     5.9783  14.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109600 on 10471 degrees of freedom
Multiple R-squared:  0.5159,    Adjusted R-squared:  0.5157
F-statistic:  2790 on 4 and 10471 DF,  p-value: < 2.2e-16
```

Figure 46: Model 1 statistics.

We can observe that almost of the variables have the confidence level that is less than 0.05, except for the **dislike_count** variable, therefore; this variable is not involved in the building process. We will then build a new model which name is model2 that have all the remaining variables.

```
LRmodel2<-lm(view_count ~ duration_sec + like_count + comment_count, data =
youtube)

summary(LRmodel2)
```

We got the following result.

```
> LRmodel2<-lm(view_count ~ duration_sec + like_count + comment_count, data = youtube)
> summary(LRmodel2)

Call:
lm(formula = view_count ~ duration_sec + like_count + comment_count,
    data = youtube)

Residuals:
     Min       1Q   Median       3Q      Max
-1158958   -57370   -32768    26626   608549

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  81640.3095  1400.7131  58.285   <2e-16 ***
duration_sec   -32.8822     3.6404  -9.033   <2e-16 ***
like_count      65.8127     0.7488  87.895   <2e-16 ***
comment_count   89.6119     4.5851  19.544   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109600 on 10472 degrees of freedom
Multiple R-squared:  0.5159,    Adjusted R-squared:  0.5157
F-statistic:  3720 on 3 and 10472 DF,  p-value: < 2.2e-16
```

Figure 47: Model 2 statistics.

From the table, the adjusted R-squared is 0.5157, which is above the average. Besides, all the predictors have a strong relationship with the dependent variable. Hence, we hope that there might be a chance of fitting a linear regression model in this case.

However, we still have to make sure our model meet the assumptions of linear regression models.

These following lines of code will help to specify our model.

```
# Plot 4 pictures
plot(LRmodel2)

# Plot residuals graph
h<-hist(LRmodel2[['residuals']] , breaks =5, density =40, col ="red", xlab
="Residuals", main ="Residuals histogram and normal distribution graph")
xfit<-seq(min(LRmodel2[['residuals']]), max(LRmodel2[['residuals']]), length =40)
yfit<-dnorm(xfit, mean = mean(LRmodel2[['residuals']]),
sd=sd(LRmodel2[['residuals']]))
yfit<-yfit*diff(h$mids[1:2])*length(LRmodel2[['residuals']])
lines(xfit, yfit, col= "black", lwd = 2)
```

The **plot()** command will give us the graphs of our model. There will be four graphs namely Residuals vs Leverage plot, Residuals vs Fitted plot, Scale-Location plot and Q-Q plot. Below code is to draw the Residual histogram to check if our model follows the normal distribution or not. The result is given as follow:

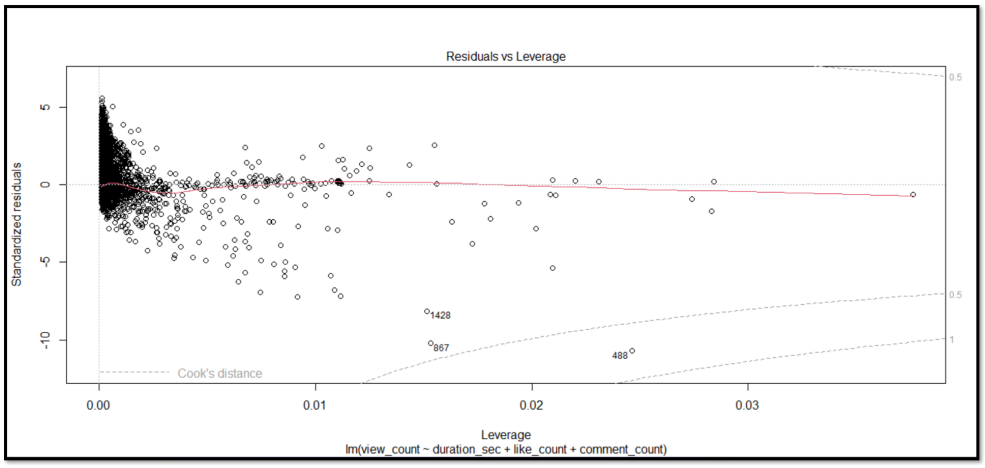

Figure 48: Residuals vs Leverage plot.

First of all, we will receive the Residuals versus Leverage plot which will help us to identify if our model has any influential point or not. From the residuals vs leverage graph, we can see that our model has one influential point which is observation number 488.

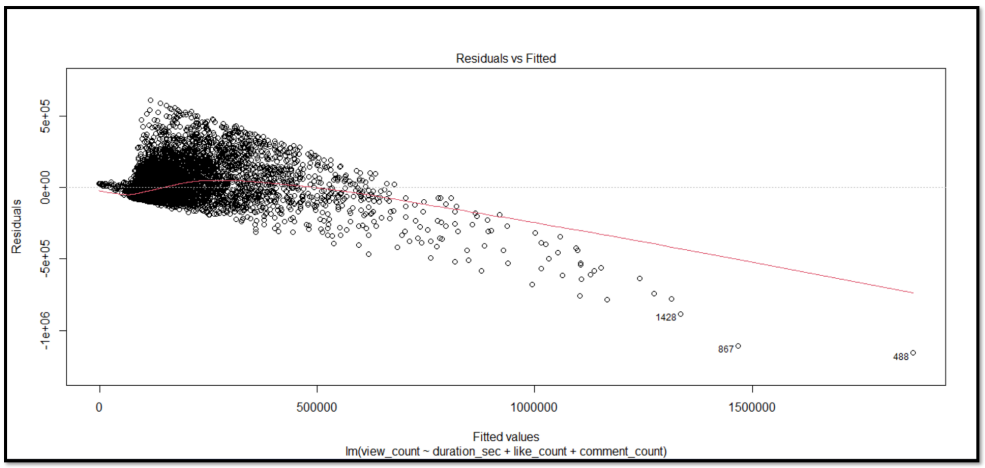

Figure 49: Residuals vs Fitted plot.

Next plot that we receive is the Residuals vs Fitted plot. It can be seen that the red line does not close to 0, in fact, it has a decreasing trend, which means that linearity is violated.

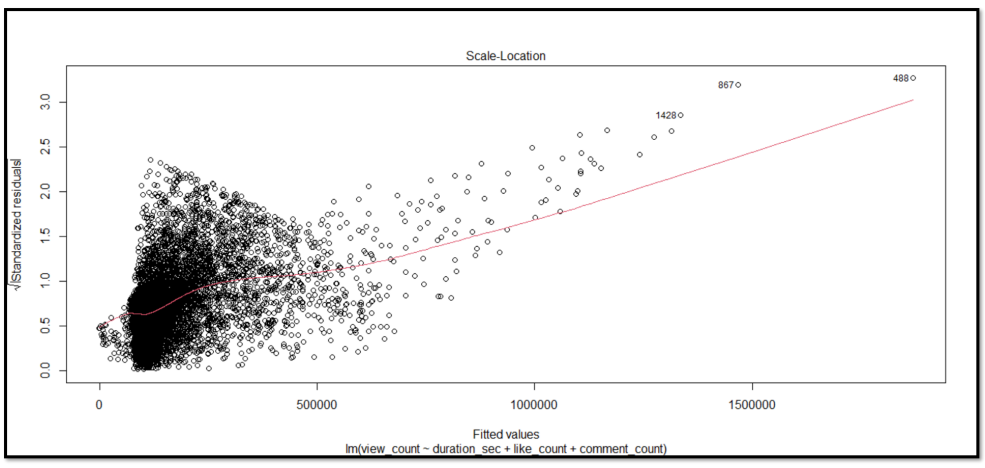

Figure 50: Scale-Location plot.

Since the spread of residuals in Residuals vs Fitted plot is not symmetric about the x-axis and the red line in Scale-Location graph is not approximately horizontal, the model does not meet homoscedasticity assumption.

Finally is the Q-Q plot and the residuals histogram.

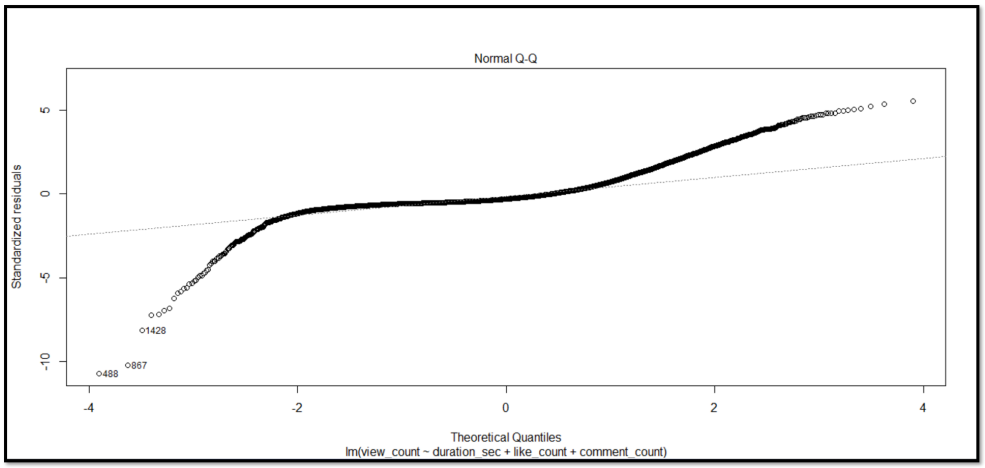

Figure 51: Q-Q plot.

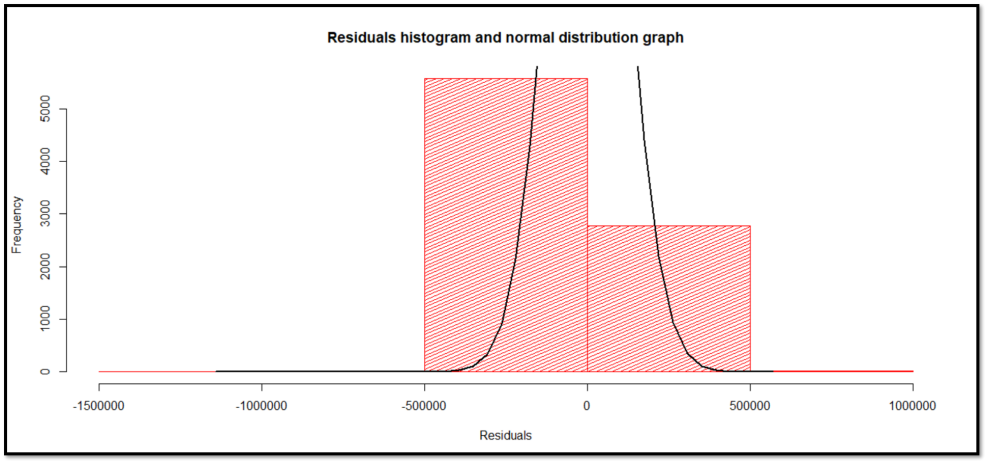

Figure 52: Residual histogram.

The Q-Q plot assesses if a set of data plausibly came from some theoretical normal distribution. In the Q-Q plot, it can be seen that those points are lying near the line, while only a few points are not lying near the line. We can conclude that this model is normally accurate, not 100% accurate.

In the residuals histogram, it seems that our model follows the normal distribution.

Although our model may follow the normal distribution, it failed in the linearity test and disobeys the homoscedasticity assumption. Therefore, our data does not fit the linear regression model.

10 Conclusion

All in all, based on the requirements of the given assignment, our team has channeled all-out efforts in order to research data sets and have numerous discussions to eventually end up with the data of BBC videos on YouTube and make up our mind to carry on this with a view to **Analyzing and evaluating contributing factors to the view-count of BBC videos on YouTube platform with the assistance of various kinds of tool** as mentioned initially. We apply the programming language R to process statistical data, including processing and analyzing unprocessed data to make them long-term essential data sources, *e.g. removing unused columns with plausible reasons*, or, even better, enabling generalization of the general situation and forecast of the data set.

During the research with some testing tools such as Kruskal-Wallis Test, Shapiro-Wilk Test, Kolmogorov-Smirnov Test, Levene Test, ANOVA Test, and Linear Regression Model, these lead us to a conclusion that the factors **category**, **definition**, **duration**, **like-count**, **comment** would be of fundamental importance to the affect of the number of views on BBC channel, and especially, we also found out some *categories* and detailed *definition* that can contribute to help BBC itself focus on those in order to get the most views from audiences.

What is more, by using analytical calculations and graphing throughout the assignment, it paves the way for us to better our skills in programming process, and also get used to the steps of how to arrange the correct sequence of implementation as well as several helpful tools based on the appropriate situations to support calculations and address sophisticated issues using computers.

Finally, our team fully appreciate Dr. Nguyen Tien Dung, who has dedicated to assist us in giving advice as well as the best solutions to figure out various problems we encountered during the time of fulfilling this assignment. As Sophomore students, actually, we cannot give a firm statement whether this is perfectly accurate, but at least, we have made the most of it by keeping researching on a daily basis and utilizing teamwork skills.

If you have any further questions, please contact us via this email: nam.nguyenolkmphy@hcmut.edu.vn

References

[1] Dataslice, Linear Regression Summary in R, https://www.youtube.com/watch?v=7WPfuHLCn_k

[2] Dataslice, Linear Regression Plots in R, https://www.youtube.com/watch?v=rfH7pCFvFT0

[3] GeeksforGeeks, Kolmogorov-Smirnov Test (KS Test), https://www.geeksforgeeks.org/kolmogorov-smirnov-test-ks-test/

[4] GeeksforGeeks, Kruskal-Wallis test in R Programming, https://www.geeksforgeeks.org/kruskal-wallis-test-in-r-programming/

[5] Jim Frost, Benefits of Welch's ANOVA Compared to the Classic One-Way ANOVA, https://statisticsbyjim.com/anova/welchs-anova-compared-to-classic-one-way-anova/

[6] Laerd Statistics, One-way ANOVA, https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php

[7] Stephanie Glen, Dunn's test: Definition, https://www.statisticshowto.com/dunns-test/

[8] Stephanie Glen, Kruskal Wallis H Test: Definition, Examples, Assumptions, SPSS, https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/kruskal-wallis/

[9] Stephanie Glen, Welch's ANOVA: Definition, Assumptions, https://www.statisticshowto.com/welchs-anova/

[10] Rebecca Bevans, Multiple Linear Regression | A Quick Guide (Examples), https://www.scribbr.com/statistics/multiple-linear-regression/

[11] Rebecca Bevans, One-way ANOVA | When and How to Use It (With Examples), https://www.scribbr.com/statistics/one-way-anova/

[12] Zach, How to Conduct a One-Way ANOVA in R, https://www.statology.org/one-way-anova-r/

[13] Zach, How to Perform Dunn's Test in R, https://www.statology.org/dunns-test-in-r/

[14] Zach, How to Perform Welch's ANOVA in R (Step-by-Step), https://www.statology.org/welchs-anova-in-r/

[15] Zach, Levene Test for Equality of Variances, https://www.statisticshowto.com/levene-test/