Project Milestone 2

Data Cleaning

In this milestone, you are required to finish the data cleaning, including but not limited to

1. **Data cleaning**

1.1. Handling missing data

Missing data was found only in the video description column, this column is completely eliminated because the description of the video predominantly does not affect the view count.

1.2. Random sampling (if you get a large dataset, because the limitation of XLMiner is 10,000 records.)

As for random sampling, the original data set had around 40950 rows, from this a random dataset of 10000 rows was chosen using the RAND() function in excel. This is done to make sure there is a proper mix of high and low view count videos.

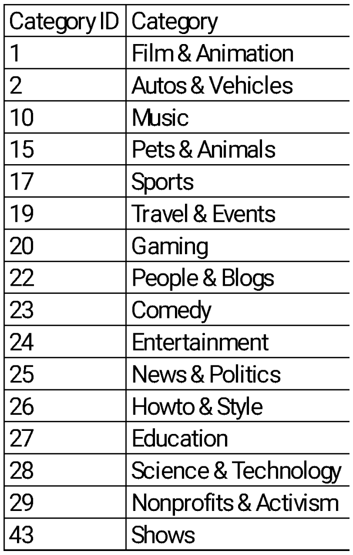
(Upload your original dataset and cleaned dataset after the first two steps)

1.3. Identify numerical variables vs. categorical variables

Before identifying the variables, the raw data needed to be transformed for variables to be usable. Variables like video title, channel name, and video tags were in the form of text in the original dataset. Using excel formulas, these variables were transformed into numerical variables like Title words (number of words in the title), Channel characters (number of characters in the channel name), and Tag words (number of words in the video tags). Other than that, in the raw data, it can be seen that the published time variable was depicted in a single column with published date (yy-mm-dd) and time bundled together. This column was transformed into two categories: Published month (January – December) and published time (AM/PM). After all the transformations, the following are the numerical and categorical variables in the dataset:

- Numerical:
    - Views – Number of views on a video
    - Title_wrods – Number of words in the video title
    - Channel_char – Number of characters in the channel name
    - Tag_words – Number of words in the video tags
    - Likes – Number of likes on the video
    - Dislikes – Number of dislikes on the video
    - Comment_count – Number of comments on the video
- Categorical:
    - Publish_month – Month of the year the video is published in
    - Publish_time – Time of the data the video is published (AM/PM)
    - Comments_disabled – If the comments are disabled on the video (true/false)
    - Ratings_disabled – If the ratings are disabled on the video (true/false)
    - Category_id – This number pertains to different categories that the video falls in

- The following table shows the different categories that pertain to their respective category IDs



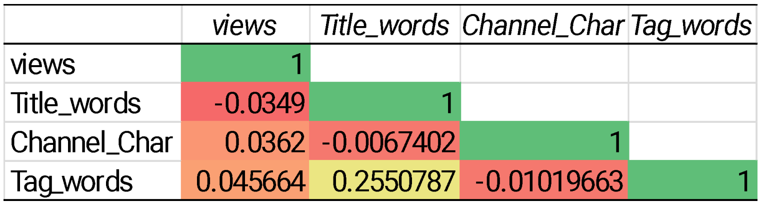
Category ID	Category
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism
43	Shows

1.4. Based on your questions, which variable is the dependent variable? Which variables could be used as independent variables? (**If any variables in your original dataset is not used in your analysis, please list your reasons**)

- Based on the questions chosen , the dependent variable for initial analysis would be the **Views** or the number of views on the video. Other dependents (likes, dislikes, comments) will be analyzed later as a cumulative popularity measure of the videos.
- The independent variables in both Numerical and categorical would be as follows:
- Numerical:
    - Title_wrods – Number of words in the video title
    - Channel_char – Number of characters in the channel name
    - Tag_words – Number of words in the video tags
- Categorical:
    - Publish_month – Month of the year the video is published in
    - Publish_time – Time of the data the video is published (AM/PM)
    - Comments_disabled – If the comments are disabled on the video (true/false)
    - Ratings_disabled – If the ratings are disabled on the video (true/false)
    - Category_id – This number pertains to different categories that the video falls in
- From the original data, the variables like thumbnail link, trending date, and published year have been removed from the analysis given their reduced relevance to our questions and DV.

1.3. Data visualization

1.3.1 Correlation table of numerical variables, **comment** on the correlations

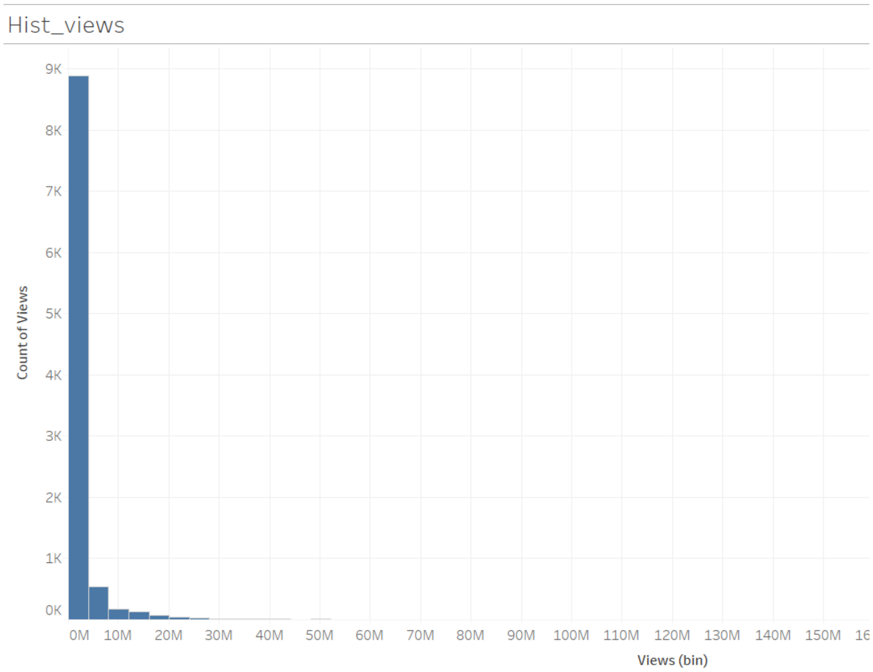


	views	*Title_words*	*Channel_Char*	*Tag_words*
views	1			
Title_words	-0.0349	1		
Channel_Char	0.0362	-0.0067402	1	
Tag_words	0.045664	0.2550787	-0.01019663	1

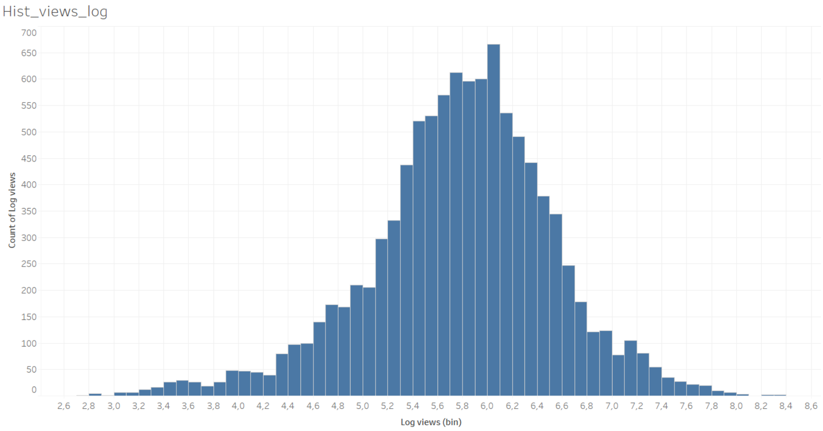
From the correlation table, it can be seen that the numerical variables have small influence on the dependent variable. Within that influence, the Title_wrods has a negative correlation, which means that a greater number of words in the title lower the view count. And the other variables have a positive correlation which means higher channel name characters and words in the video tag would relate to higher view count. It needs to be noted that this is not very significant as seen in the table.

1.3.2 Histogram of numerical variables (choose at least three variables you are most interested), **comment** on the distributions

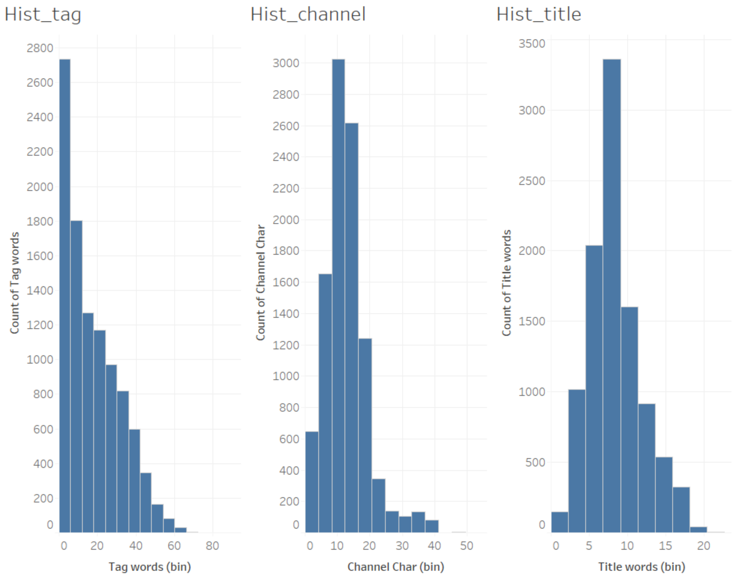
The histogram of the Views variable can be seen as follows:



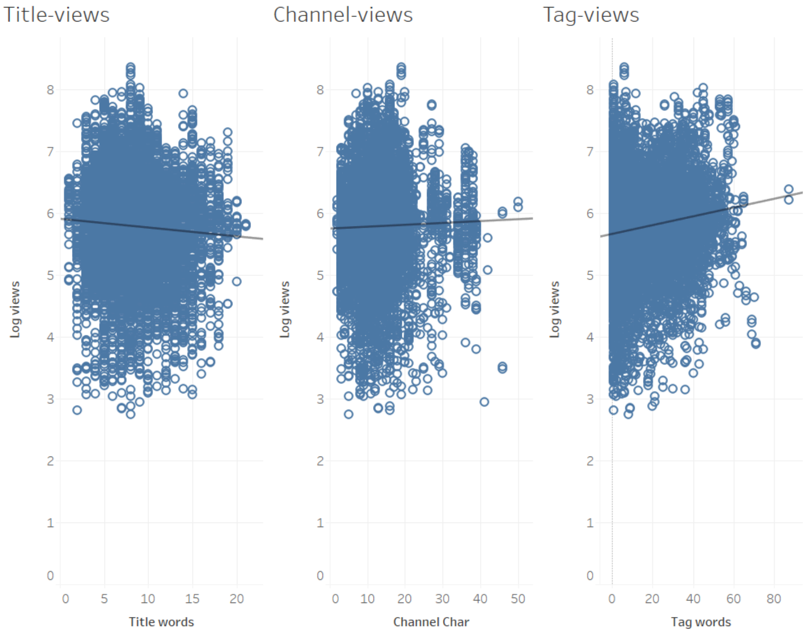
As it can be clearly seen, the view count column needs to be normalized in order to have a better representation because the numerical differences between the view counts of different videos is very high which in turn is resulting in the skewed representation. The normalized view count with logarithmic transformation to the base 10 would give a histogram as follows:



This shows a normal distribution with a better representation of the data. The other histograms of the numerical variables can be seen in the following picture. Each of the variables has few outliers which are essentially not errors but just extreme values which are important for analysis.



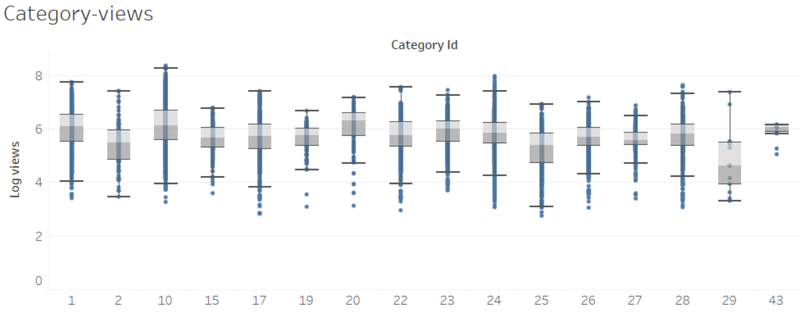
1.3.3 Scatterplot of the Dependent variable (if your DV is a numerical variable) vs independent variables (choose at least three variables you are most interested), **comment** on the relationships

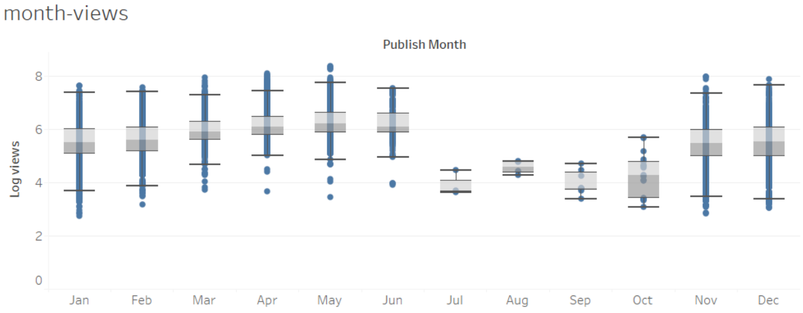


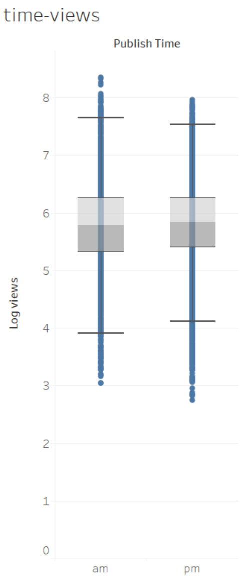
As seen in the correlation table, the scatter plots shown above, between the numerical variables and the dependent variable show a similar relationship. Where an increase in the title words slightly decreases the view count, and an increase in the Channel name characters and tag words slightly increases the view count.

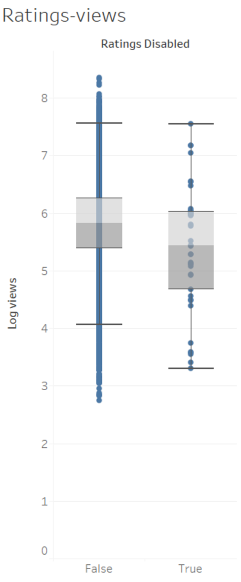
1.3.4 Boxplot of the Dependent variable (if your DV is a categorical variable) vs independent variables (choose at least three variables you are most interested), **comment** on the relationships

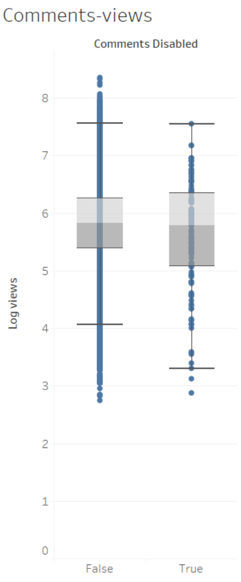
The box plots of the view count in relation to the category ID and the month published is shown below. It can be seen that category 10 which is the music category has the highest view count among the others, and the summer months of April, May and June have the higher view counts as compared to the other months.











The above box plots show the relationship between the view count and the other categories like time published, comments disabled and ratings disabled. It can be seen that the median view counts for each of these categories are pretty similar, with the videos published in the AM having a slightly higher view count. And as expected, the videos with comments and ratings not disabled have a higher median view count.

1.3.5 From the previous plots, detect outliers and find out whether they are errors or extreme values

As seen in the histograms and the box plots, there are a few outliers, but they are not errors, but just extreme values in each category that have an effect and are important for the analysis.

2. **In general, is there any potential issues with the dataset? Like if you could get data from the company directly, how would you extend current dataset?**

   Overall, it can be said that there are no potential issues with the dataset. All the transformations needed to make the data usable from the raw data have been meticulously made. And the behavior of the variables as seen in the visualization above is as expected in all of the cases. In general, the pattern observed was that the numerical variables have a lower effect on the view count as compared to the categorical variables which tend to impact the view count strongly, like the month published and the category of the video, which is natural and as expected.

   If the data from the company can be extracted directly, it would be nicer to have the transformations done in the original dataset with clear differentiation of published months, published time, and numerical variables, which would reduce the pre processing time of the data.

Please also attach a table of how you split the tasks among group members.

The final output of this milestone could be slides or word/pdf file. There is no page limit.

Please note that in practice data cleaning is important and time consuming. If you run analysis using a “bad” dataset, your model would not be reliable.

Please also note that I have high expectations of the final project and hope you could apply what you have learned in this class to your future work.

Last but not least, you are always welcome to discuss your project with me.