Evaluating YouTube Data

[redacted]

Western Governors University

Table of Contents

A. Project Highlights

The project organizational need is for a Marketing company to decide on which YouTube Channels to target for their ads. The Research Question for this project that Data Analyst work will be performed to answer is: What YouTube Channels worldwide would be most beneficial to target for selling the companies advertisements and what factors influence the subscribers and views that the channel creates? This is an important organizational need for the marketing company as the channels with the most subscribers and views are going to generate the most attention to their advertisements, which in turn will attract the views to the product they are promoting.

The scope of the project is to gather and analyze a dataset with python in Jupyter Notebooks using Pandas, NumPy, Matplotlib, Seaborn and Pearson Libraries to complete the analysis. The dataset will first be imported to a Jupyter Notebook where it will then be cleaned and evaluated. The analysis will provide statistical evidence that will either validate or disapprove the hypothesis. The analyst will be analyzing comparisons between subscribers & views, subscribers & YouTuber, views & YouTuber, category & subscribers and category & views. Using these fields to compose statistical analysis by performing a correlation test as well practical analysis and creating appealing data visualizations by displaying scatter plots and bar charts that provide the marketing company with data insight. The project will end with a video presentation where I will show and explain the scripting used to complete the project and present the findings as a result. This can be used by key decision makers in the company to help the marketing company make an informed decision in the end.

The methodology used when creating this project was an Agile approach. I chose agile because of its iterative nature. This project was to provide analysis of a dataset for a marketing company to help them make an informed decision, so I chose the agile approach because of its adaptability and ability to adjust during each iteration. I wanted to involve the company during the process where they can give feedback and changes could be made to the scope of the project as needed as the deadline for this project was not of the highest priority. The progress of this project will be iterative working in sprints for each step of the way and then will be evaluated with feedback prior to moving on to the next step of the process until the project is completed. Listed below are the steps performed:

- Gather and Import Dataset
- Clean Dataset
- Evaluate and Analyze Dataset
- Create visualizations.
- Make Panopto video of Analysis Results

The tools used for the completion of this project and to produce the data analytical solution are python and Jupyter Notebooks which are provided from Anaconda. Python is the programming language that will be used throughout this project and Jupyter Notebooks is a platform that allows users to compile code in easy to work with code blocks that show your work clearly along the way. I have used the following libraries within Jupyter Notebook to complete this work; Pandas, NumPy, Matplotlib, Seaborn and Pearson libraries. Pandas Library was used to import the dataset used for this project into my Jupyter Notebook, which allowed me to do all the work needed to complete this project. The NumPy Library allows for performing mathematical functions and help dealing with null values. Matplotlib and

Seaborn Libraries were used to create plots and charts that paint a visual picture of the dataset and analysis. Finally, Pearson was used to perform descriptive statistics by performing a correlation test that resulted in statistical information about the relationships between fields by providing a correlation coefficient and P-Value of two variables.

B. Project Execution

The goal of this project is to provide a marketing company with analysis of YouTube dataset that will help the company make important business decisions based on the results of the data analysis. The project consists of first gathering the dataset and importing it to Jupyter Notebooks where the analysis will begin. The result of the project will provide clear analysis with statistical significance to the marketing company and create easy to understand visualizations that ultimately paint a picture that answers the business question by comparing multiple relevant fields. The final result of this project will be a video presentation of the data analyst's findings that clearly shows the benefit of investing in certain YouTubers. This ultimately will steer the business in the direction to make the best decision of which YouTuber to sell their advertisements to.

The dataset used is fully downloadable as a zip file, upon downloading the dataset I used Jupyter notebooks using python as the language to gather, clean, and analyze the dataset to gain vital insight that will ultimately be used by the marketing company to make final decisions on which YouTube channel to sell advertisements to. I used Pandas Library within my Jupyter Notebook to import the CSV file using read_csv. This allowed me to fulfill the next steps of the project where I performed cleaning and analyzing the dataset. I have used NumPy, Matplotlib and Seaborn Libraries to compose statistical analysis of the dataset as well as create visually pleasing data visualizations that both validate that analysis and create comparisons of fields such as: subscribers, views, category, and YouTuber that will provide further insight to the dataset. In performing this analysis, I was able to provide the necessary information the marketing company needed to make the most informed decision of which channels to target for selling advertisements to.

The methodology used when creating this project was an Agile approach. I chose agile because of its iterative nature. This project was to provide analysis of a dataset for a marketing company to help them make an informed decision, so I chose the agile approach because of its adaptability and ability to adjust during each iteration. I wanted to involve the company during the process where they can give feedback and changes could be made to the scope of the project as needed as the deadline for this project was not of the highest priority. The progress of this project will be iterative working in sprints for each step of the way and then will be evaluated with feedback prior to moving on to the next step of the process until the project is completed. Listed below are the steps:

- Gather and Import Dataset
- Clean Dataset
- Evaluate and Analyze Dataset
- Create visualizations.
- Make Panopto video presentation of Analysis Results

Project Timeline

Milestone	Projected Time	Start Date	End Date
Gather/Import Dataset	**2 Hrs.**	**9/20/2023**	**9/20/2023**
Clean Dataset	**4 Hrs.**	**9/20/2023**	**9/20/2023**
Evaluate and Analyze Dataset	**4 Hrs.**	**9/21/2023**	**9/21/2023**
Create Visualizations	**3 Hrs.**	**9/22/2023**	**9/22/2023**
Create Video presentation of Results	**5 Hrs.**	**9/25/2023**	**9/26/2023**

The plan and methodology had no variance from Task 2 as the planning, execution and methodology remained the same. The timeline did change a bit in comparison to what I had projected in Task 2. Some of the work took less time than initially anticipated so this allowed me to complete it a little earlier than anticipated.

C. Data Collection Process

The data selection and data collection plan did not differ at all from my plan in Task 2. The plan was to use a dataset YouTube Statistics 2023 to answer my business question, which was accessible to the public from Kaggle.com for Kaggle members. This is what I had outlined for the data selection and collection process in Task 2 and there was no deviation from that plan.

I did not run into any obstacles during the collection process as the dataset was a free accessible dataset provided by Kaggle.com and able to be saved as a CSV file allowing me to execute m plan put in place in Task 2.

I did not run into any unplanned data governance issues for this process at all as this was all planned out based on the work that had already been performed. The dataset provided by Kaggle is unrestricted and free to use and share with no licensing.

C.1 Advantages and Limitations of Data Set

The dataset used was great for my project but was not perfect, below are some examples of advantages and disadvantages:

Advantages

- The dataset used for this project was relatively clean which made it easy to work with. Due to this it did not take too much time to clean and evaluate.
- Another advantage is that the dataset was easily accessible. As stated, before this dataset was accessible for free by Kaggle that was as easy as clicking download.
- It had all the necessary fields required to perform the tasks outlined to complete this work.
- The dataset provided statistics from 2023 giving me information and insight that is relevant right now.

Disadvantages

- The dataset is limited to statistics from the year 2023. Having a dataset that had statistics for more years would have been beneficial to use to be able to evaluate these same trends overtime rather than having a more limited scope.
- Although the dataset was relatively clean and complete there was one column specifically where about 1/3 of its values were null values. This had to be taken dealt with appropriately as that many null values could really change the results of the analysis.

D. Data Extraction and Preparation

The data extraction portion of this project was straightforward and simple. I had to first become a member of Kaggle then I was able to download the CSV file from Kaggle.com. Once the dataset was downloaded to my computer, I imported the dataset into my Jupyter Notebook using python library Pandas read_csv function. Once the dataset was imported into a data-frame I was able to move to the preparation portion. In preparation I looked at the basic structure as well as areas that could potentially need to be changed to gain insight of the data frame. I specifically looked at the shape of the dataset, used the describe() function to learn more about the numerical data in the dataset, searched for any duplicates, and searched for null values using the isnull() function. In preparation to the analysis portion of this project I chose to deal with the null values by filling them with the average for some and dropping the others that did not contain too many to affect the data frame. I also chose to change the name of a column that was used a lot for simplicity and removed a few columns from the data frame that I would not be using in my analysis. Once these steps were performed, I was ready to start the analysis portion of the project.

E. Data Analysis Process

E.1 Data Analysis Methods

Exploratory and descriptive analysis were the primary methods used to complete this project. These methods were necessary to answer the business question, The hypothesis was that there are key relationships between different fields will play a major factor into which YouTuber was deemed most successful. I also hypothesized specifically that the YouTubers with the most views would have a positive correlation with those with the most subscribers. These key relationships will be used as a guide to help the marketing company make their final decision of which YouTuber to sell their ads to. I performed statistical measures when analyzing the relationship between subscribers and views to prove my hypothesis by completing a correlation test between the two variables and creating scatterplots of the variables to give a visual of their correlation. I found this to be most appropriate as these two variables are quantitative. I used exploratory analysis to explore the other relationships in the data frame by creating a series of bar charts. These charts were necessary and used to evaluate the relationships from a visual standpoint to help answer the business question,

E.2 Advantages and Limitations of Tools and Techniques

The tools that I use for the completion of this project and to produce the data analytical solution are python and Jupyter Notebooks which are provided from Anaconda. Python is the programming language that was used throughout this project and Jupyter Notebooks is a platform that allows users to compile code in easy to work with code blocks that show your work clearly along the way. I will be using the following python Libraries; Pandas, NumPy, Matplotlib, Seaborn and Pearson. Pandas Library was used import the dataset for this project into my Jupyter Notebook, which allowed me to do all the work needed to complete this project. The NumPy Library allowed for performing mathematical functions that were used such as the sum function and help dealing with null values. Matplotlib and Seaborn Libraries were used to create scatterplots and bar charts were used to explore relationships from the data frame and paint a visual picture of the dataset. Finally, Pearson Library was used to perform descriptive statistics by performing a Pearson correlation test that provided statistical information about the relationships between the subscriber and views fields.

Advantages

- The use of Jupyter Notebooks allowed me to import the dataset to work with and analyze, clean the data frame, compile code to perform exploratory analysis by creating visualizations to support my hypothesis and perform descriptive statistics to support my hypothesis all in one clean workplace.

Limitations

- It is hard to find any limitations to the tools that I used for this project, but one limitation from a data visualization standpoint is the ease of creating plots using Matplotlib vs a tool like Tableau. Tableau is a visualization tool that is perfect for exploratory analysis and is very simple and easy to use whereas using matplotlib with python is a much more technical way of exploring variables through data visualization.

E.3 Application of Analytical Methods

Many analytical methods were implemented to complete this project from start to finish. Below is the step-by-step process that was performed for completion:

1. Importing the dataset into the Jupyter Notebook using Pandas read_csv.
2. Exploring the data frame structure and searching for any completion issues in the data frame using the following functions:
    a. head() function used to generate a view of the first 5 rows and the fields in the data frame.
    b. describe() function was used to evaluate the numerical data in the data frame.
    c. info() function used to view the column types and see if there are any columns that are miss labeled.
    d. The shape function was used to understand the size of the data set in terms of the number of columns and rows present in the dataset.
    e. Isnull().sum function to view the number of null values from each column.

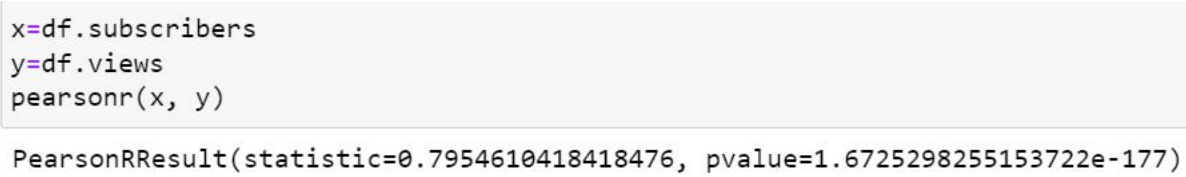
   f. Duplicated().sum to see if there are any duplicated values.

3. Cleaning the data frame
   a. Dealing with null values by filling with the average results for "subscribers_for_last_30_days"
   b. Using dropna() to drop the remainder null values.
   c. Rename "video views" to "views" for ease of work.
   d. Using drop() function to drop columns that were not needed to fulfill the scope of my project.
4. Performing exploratory analysis of the dataset and learning about relationships both programmatically and through data visualization
   a. Exploring the top 10 YouTubers that have the most subscribers and created a bar plot to convey the results visually using Matplotlib plot() function. This can be used to determine which YouTubers are the most successful.
   b. Exploring the top 10 YouTubers that have the most views and created a bar plot to convey the results visually using Matplotlib plot() function. This can be used to determine which YouTubers are the most successful.
   c. Exploring the top 10 YouTubers are the highest ranked. This can be used to determine which YouTubers are the most successful.
   d. Explore the relationship between subscribers and uploads by creating a scatterplot using plt.scatter()
   e. Exploring the top 10 most common YouTube categories by utilizing the value_counts function and sorting them in descending order and created a bar plot using Matplotlib plot() function. This gives a good representation of which categories are most popular among YouTubers.
   f. Exploring the relationship between categories & subscribers, categories & views and creating a bar plot using Matplotlib plot() function. This is important information as the results can be utilized in deciding about what YouTubers to choose to sell ads to. The company could look at YouTubers from the category with the most subscribers and views to try to find an edge and get their ads available to the largest audience.
5. Explore statistical significance of hypothesis (relationship between views and subscribers)
   a. Created a scatterplot to show the relationship of views and subscribers using Matplotlib plt.scatter() function.
   b. Performed Pearson correlation test from the between views and subscribers using the Pearson python library from scipy.stats. This results in a correlation coefficient and p- value for determining statistical significance between two quantitative variables.

F Data Analysis Results

F.1 Statistical Significance

To provide statistical significance of my solution I performed a Pearson correlation test between two quantitative variables: views and subscribers. The test results in a p-value and a correlation coefficient that will be used to gain insight on the project hypothesis. If the correlation coefficient is a positive value, we will know that it the variables have a positive relationship and a value that is closer to 1 indicates a strong relationship. Evaluation of the p-value we know that a value greater than 0.5 means that there is statistical significance while a value less than 0.5 indicates there is no statistical significance. The null hypothesis is that there is no correlation between views and subscribers, while the alternative hypothesis of my solution is that there is a positive correlation between views and subscribers. The test was performed to either reject the null or alternative hypothesis. See below for the results of the Pearson correlation test.



```
x=df.subscribers
y=df.views
pearsonr(x, y)
```

```
PearsonRResult(statistic=0.7954610418418476, pvalue=1.6725298255153722e-177)
```

The test resulted in a correlation coefficient of 0.79, due to value being a positive number and the value is relatively close to 1 we can conclude that subscribers and views have a strong positive relationship. The p-value provided from the test resulted in a value of 1.67, since this value is greater than the 0.5 threshold, we can conclude that there is statistical significance. Due to the results of the correlation coefficient and p-value, there is sufficient evidence to reject the null hypothesis and validate our alternative hypothesis proving that there is a positive correlation between views and subscribers. This is significant analysis for the marketing company as views are a major aspect of this project as more views equate to a larger audience. Validation of the alternative hypothesis means that more subscribers also equate to a larger audience, and this can be used by the marketing company when evaluating which YouTubers have the largest audiences.

F.2 Practical Significance

The practical significance of this project is to identify which YouTuber/categories generate the largest audience. Finding answers to this through analysis is vital to the success of this project. Understanding which YouTubers and categories have the largest audiences is important to the scope of this project as the marketing company is looking to get their ads out to the largest audiences, so their advertisements can have the impact they are hoping for.

Practical significance was applied by evaluating relationships between specific variables to gain insight of what generates the largest audience. Using the results of the Pearson correlation test explained in F.1, which proved that views and subscribers have a positive relationship validating that these variables both relate to generating a large audience, I evaluated these relationships with the YouTuber field. I compiled code that resulted in the top 10 YouTubers in Views and subscribers. This is practically significant as the results provide the YouTubers that have the most views and subscribers, which equates to which YouTubers can generate the largest viewing audience. The analysis resulted in the YouTuber T-Series having both more views and subscribers than any other channel. This information vital for the marketing company to make the best decision on which YouTuber to invest their ads in because the larger the audience means the more attention a videos advertisements are going to receive. Based on the results of this analysis T-Series has the largest audience and has the best potential to generate the most views to potential advertisements.

While performing other exploratory analysis of this dataset I decided to look at which categories generate the most views and subscribers. This resulted in a top 5 categories for each of these variables. The number one category by views and subscribers resulted in the Music category for both variables and was the leader by large margin. This is practically significant as the category with the most subscribers and views is interpreted as having the largest audience. This analysis can be used by the marketing company to evaluate which niche/category they would like to target. Based on these results the Music category would be the most beneficial for them to target as it is the category with the most views and subscribers therefore generating the largest audience the marketing company can take advantage of.

F.3 Overall Success

Overall, I view this project as a success. The goal of the project was to provide a marketing company analysis of a YouTube dataset that can help them make an informed business decision based of the analysis. Through exploratory analysis and descriptive statistics, I was able to create visualizations and perform a correlation test to answer the business question lined out while proving my hypothesis correct. All the criteria outlined in project goals have been met and I was able to provide the marketing company with sufficient evidence from both statistical and practical significance. I was able to provide the top YouTuber and category in terms of subscribers and views that have been outlined in F1 and F2vi. This analysis will help the marketing company make the most informed decision of what YouTuber to sell their advertisements to.

G. Conclusion

G.1 Summary of Conclusions

The project's main objective was to provide a marketing company with analysis of a dataset that would be used to make a key business decision of which YouTuber to sell their advertisements to. To do this, I gather and explored a dataset by scripting python in Jupyter Notebooks using Pandas, NumPy, Matplotlib, Seaborn, and Pearson Libraries to complete the analysis. I proved statistical significance by performing a Pearson correlation test between two variables: subscribers and views. The results of this test proved a strong positive relationship by the correlation coefficient being a positive number that was close to one. It also proved statistically significant as the p-value was greater than the 0.5 threshold. This validated my hypothesis that subscribers and views have a positive correlation with one another, and it can be interpreted that these two variables both relate to a large audience. Through exploratory analysis I evaluated relationships using the results of the correlation test in mind. I created bar charts that portrayed the top YouTubers with the most views and subscribers as well as which categories have the most views and subscribers. These results provide the marketing company with sufficient evidence of which YouTuber would be the best to invest and sell their advertisements to.

G.2 Effective Storytelling

The visualizations used in this project were the best methods to use as they provide a clear correlation between variables as well as provide a visually clear answer to the business question. The visual representations created for this project were created by using Matplotlib and Seaborn python Libraries. The scatterplot that evaluated the relationships between the views and subscriber's variables painted a clear picture that there was a positive relationship between the two. This visualization was used to further validate the hypothesis from a visual standpoint that there was a positive correlation between the two variables. I created a series of bar charts that portrayed the top YouTuber and category by the number of subscribers and views that these variables achieve. These bar charts support effective storytelling relevant to the scope of this project because they visually show which YouTuber has the largest audience and which category has the largest audience. These visual representations will effectively give the marketing company evidence that answers the main business question of which YouTuber to sell their ads to. The YouTuber with the most subscribers and views within the highest viewed and subscribed to category would likely be the most informed choice of who they should target to sell their advertisements to. This YouTuber will have the largest audience and therefore their advertisements will have the largest impact potential.

G.3 Recommended Courses of Action

The analysis performed throughout this project was to answer the question of: What YouTuber should a marketing company sell their advertisements to. Based off this analysis we learned through both exploratory analysis and descriptive statistics that there is a correlation between views and subscribers that suggests these variables relate to having a large viewing audience. Through exploratory analysis I created bar plots that demonstrate the top YouTubers and Categories in terms of views and subscribers. These are important variables that will be used to make an informed recommendation.

My first recommendation for the marketing company to consider targeting for the sale of their advertisements to is the YouTuber T-Series. Based on the results of these bar charts and analysis T-Series is the top option. It has been observed that T-Series not only has the most subscribers and the highest amount of video views, but their niche/category which is Music is also the highest viewed and subscribed to category, these results mean that T-Series has the largest audience and will provide the marketing company with the best potential impact. For those reasons I recommend that the marketing company should attempt to sell their advertisements to T-Series.

The second recommendation I am going to make is not as clear cut as the first. I have decided to favor the number of views over the number of subscribers as views is the clearest way to determine the largest audience. I am going to recommend that the marketing company consider Cocomelon-Nursery Rhymes to sell their advertisements to as a second option. I have decided this based on the number of views that Cocomelon-Nursery Rhymes has generated. Although Cocomelon-Nursery Rhymes has the 4th highest number of subscribers they have the second most views than any other YouTuber, coming in at 1.64 billion views where the next highest is 1.48 billion. I have also concluded based on the results of my analysis that Cocomelon-Nursery Rhymes's niche/category is Education, which is the second most viewed category meaning there is a large audience of people searching and viewing education-based videos. Based on this evaluation I am going to recommend that the marketing company attempt to sell their ads to Cocomelon-Nursery Rhymes.

H Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=[redacted]

References

Nidula Elgiriyewithana, 2023, Global Youtube Statistics 2023, https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023

2022, How to Perform a Correlation Test in Python, https://www.statology.org/correlation-test-in-python/

Appendix A

Additional Files Provided in submission:

- Dataset: https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023
- Jupyter Notebook PDF file displaying python code performed for completion of work.