Developing YouTube Matrices

by

Nitish Reddy Gaddam

This project is submitted to the Gannon University graduate faculty in

partial fulfillment for the degree Master of Science in Computer and Information Science.

Option: Information Technology

Approved:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | |  |
| Dr. Yunkai Liu, |  | Dr. Mei-Huei Tang, | | |
| Advising Professor in Charge of Research |  | | Committee Member |
|  |  | |  |
|  |  | |  |
| Dr. Madhusudan Singh, |  | | Dr. Mei-Huei Tang, |
| Committee Member |  | | Chair, Computer, and Information Science Department |
|  |  | |  |

Gannon University

Erie, Pennsylvania 16541

Dec 2022

Acknowledgements

I would like to express my special appreciation and thanks to my advisor Professor **Dr. Yunkai Liu**, you have been a tremendous mentor for me. I would like to thank you for encouraging my project and for allowing me to grow. I am very grateful for advice and knowledge and many insightful discussions and suggestions, as you were my primary resource for getting my project complete. Your advice on both project as well as on my career have been invaluable. I would also like to thank my committee members, **DR.Mei-Huei Tang**, **DR.Madhusudhan Singh** for serving as my committee members even with their busy schedules.

Table of Contents

[Acknowledgments ii](#_Toc120231168)

[Abstract iv](#_Toc120231169)

[List of Figures v](#_Toc120231170)

[List of Tables v](#_Toc120231171)

[1. Introduction 1](#_Toc120231172)

[1.1 Overview 1](#_Toc120231173)

[1.2 Project Impact 1](#_Toc120231175)

[2. Data Exploration 1](#_Toc120231177)

[2.1 Data Description 1](#_Toc120231179)

[2.2 Data Details 1](#_Toc120231180)

[2.2.1 Data Overview 2](#_Toc120231181)

[2.2.2 Exploration Statistics 4](#_Toc120231182)

[3. Methodologies 6](#_Toc120231183)

[3.1 Introduction **Error! Bookmark not defined.**](#_Toc120231185)

[3.2 Methodologies details **Error! Bookmark not defined.**](#_Toc120231186)

[3.2.1 What is your method? 6](#_Toc120231187)

[3.2.2 Flow chart 11](#_Toc120231188)

[4. Verification and Validation 11](#_Toc120231189)

[4.1 Test Method 11](#_Toc120231191)

[4.2 Test Results and Explanations: 12](#_Toc120231192)

[5. Feature work 17](#_Toc120231193)

**6. Conclusion...**…………………………………………………………………………………**17**

[7. Bibliography 19](#_Toc120231194)

[Appendix A: Source Code](#_Toc120231195) 20

Abstract

According to the YouTube 2022 report, YouTube has more than 51 million channels. The number of channels has continued to grow, with a 36% increase last year. Many people from all over the world have contributed to the creation of a YouTube channel, which is actively uploading 500 hours of video every minute. YouTube has over 40 million active gaming channels, but only 80,000 of them have more than 100,000 subscribers, while other creators have failed to achieve growth/popularity (YouTube). The main goal of this project is to improve the YouTube channel in terms of increasing subscribers, views, likes, and comments through technical analysis and comparison with other successful Youtubers. This project was created in order to perfectly access and analyze the channel's entire YouTube data. This is a YouTube Gaming content creators friendly project. This project makes it simple for YouTube Gaming Streamers to perform various tasks such as predicting future views, comparing their channel stats to other youtubers, and determining where they need to improve. It also assists in determining which videos will help them reach a larger audience and gain more views and subscribers. The technique used here is to find the probability values of the datasets and then perform linear regression on them before plotting them on a scatterplot to visualize the data. This method is chosen after extensive testing of various regression methods.

# List of Figures

[Figure 1. Mortal channel top ten videos 5](#_Toc1377698)

[Figure 2. 8bit Thug channel top ten videos 5](#_Toc1377699)

[Figure 3. Regaltos channel top 10 videos. 6](#_Toc1377697)

[Figure 4. Data preparation for three data set’s ”View” “Subscribers” ”Playlist” 6](#_Toc1377698)

[Figure 5. Video data 7](#_Toc1377698)

[Figure 6. Mortal 7](#_Toc1377699)

[Figure 7. Regaltos 8](#_Toc1377697)

[Figure 8. Thug 8](#_Toc1377699)

[Figure 9. Process diagram 11](#_Toc1377698)

[Figure 10. Mortal channel statistics 15](#_Toc1377699)

[Figure 11. Regaltos channel statistics 16](#_Toc1377697)

[Figure 12. Thug channel statistics 16](#_Toc1377698)

# List of Tables

**List of Tables**

[Table 1. Position Statement. 1](#_Toc1377697)

[Table 2. Channel Data with subscriber count 2](#_Toc1377698)

[Table 3. Channel name with view count 2](#_Toc1377698)

[Table 4. Channel name with videos count 3](#_Toc1377699)

[Table 5. Exploration statistics 4](#_Toc1377697)

[Table 6. Mortal channel two sample for mean calculation 9](#_Toc1377699)

[Table 7. Regaltos two samples for mean calculation 9](#_Toc1377698)

[Table 8. Thug channel two samples mean calculation 10](#_Toc1377699)

[Table 9. Three channels with two variable mean calculation 10](#_Toc1377697)

[Table 10. Data sets with two variables 10](#_Toc1377698)

[Table 11.Mortal two samples for mean 13](#_Toc1377699)

[Table 12.Regaltos two samples for mean 14](#_Toc1377697)

[Table 13. Thug channel two variables for mean 14](#_Toc1377698)

[Table 14. 3 Channel two variables for mean 14](#_Toc1377698)

.

1. Introduction

Overview

This project was created in order to access the entire YouTube data of the channel and then analyze and visualize it in the best way possible so that in the future channel performance can be calculated using the trend line. The project consists of a Python Notebook that uses Python Libraries to perform various types of analysis (scatterplot, line plot) in a visualization tool such as Power BI. Pandas is widely used for allocating Data frames to Datasets. Linear Regression is the Machine Learning Algorithm used to predict future video statistics. The method was chosen as a result of continuous testing of various regression methods. This can be a daunting task, but there are tools available to help.

Project Impact

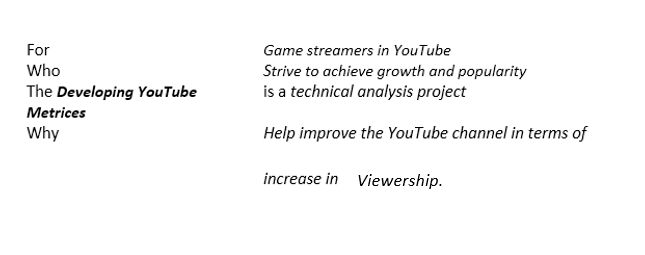


TABLE 1: Position Statement

1. Data Exploration

Data Description

The origin of the datasets for the three YouTube channels is being extracted using the google YouTube API V3.

The YouTube Data API lets you incorporate functions normally executed on the YouTube website into your own website or application. The distinct types of resources that you can retrieve using the API. The API also supports methods to insert, update, or delete many of these resources.

Data Details

The data for each channel consists of 5 columns, i.e.

1) Published date

2)Title of the video

3)Total number of views

4) likes

5)Published Month

Data Overview

Each dataset varies from one another as different YouTube channels upload different types of videos and each of them has different statistics.

**Subscriber Count**

**Views**

**Total Videos**

**Top 10 Videos**

**SUBSCRIBER COUNT–** In this table there will be an overview of the Subscriber count for the particular YouTube Channel. This table will target the total number of subscribers for each channel

Table

Description automatically generated

Table 2 Channel Data with subscriber count

* **VIEWS -**  In this table there will be an overview of the total number of views for the specified channel. This table will target views for each video and gives an overview of total number of views for a channel.

Graphical user interface, table

Description automatically generated

Table Channel name with views count

* **TOTAL VIDEOS –** This particular table will provide a detailed overview of the total number of videos present in the channel. This table will target the details of a particular video.

Graphical user interface, application

Description automatically generated

Table Channel name with videos count

**Benefits and Supporting Features**

**Subscriber Count**

Subscriber Count table is used to analyze the total number of subscribers for the particular YouTube channel. In this project there are 3 YouTube channels which are to be analyzed and processed for the data visualization. As the main aim of this project is to develop the YouTube metrices, subscriber count plays a vital role for the great reach of one’s channel.

**Views**

Total number of views changes from one video to the other video. Views table is the backbone for this project. To analyze the reach of a particular video, views table come in hand for the analyzation. Statistics are being calculated according to the views. The revenue for a particular video is also generated by total number of views. The views will increase if the channel has more subscribers, so to analyze views there has to be the subscriber count for the channel for better understanding.

**Total Videos**

For every channel there has to be N number of videos that needs to be analyzed. For better understanding whether the channel is actively uploading videos or not, there has to be a detailed description of date and time they uploaded. While comparing the properties of the videos which they uploaded gets a better understanding of the current time and day to upload the video. This will not only analyze the day, date and time but will also analyze regularity of the uploaded videos.

**Dependent Data and Independent Data**

Subscriber count and Views are the independent data which is based on the particular YouTube channel. Total videos, published date, day and month are the dependent data

1. **Channel name** – Mortal

The Youtuber Mortal is being uploading videos from 2016. So, this dataset has 1561 records. These are classified into 4 different columns (Likes are disabled for this channel) i.e. Published date, Title of video, Views and Published Month.

1. **Channel name –** Soul Regaltos

The Youtuber Soul Regaltos is being uploading videos from 2018. So, this dataset has 963 records. These are classified into 5 different columns i.e. Published date, Title of video, Views, Likes and Published Month.

1. **Channel name** – Thug

The Youtuber Thug is being uploading videos from 2018. So, this dataset has 471 records. These are classified into 5 different columns i.e. Published date, Title of video, Views, Likes and Published Month.

Exploration Statistics

Each dataset consists of to all the video titles with its published dates, views and likes and the dataset is cleaned accordingly, the data is cleaned based on the year which it is uploaded and arranged in descending order to get the perfect statistics. As the below figure represents the dataset which is arranged accordingly with the year i.e. 2022(recent statistics).

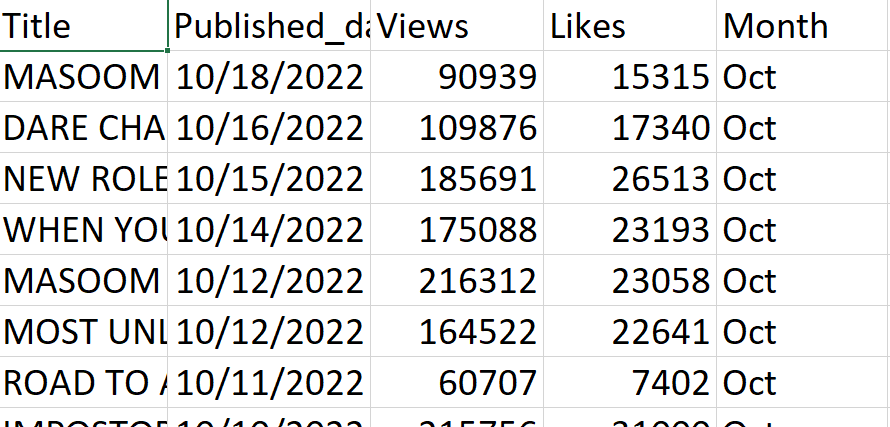
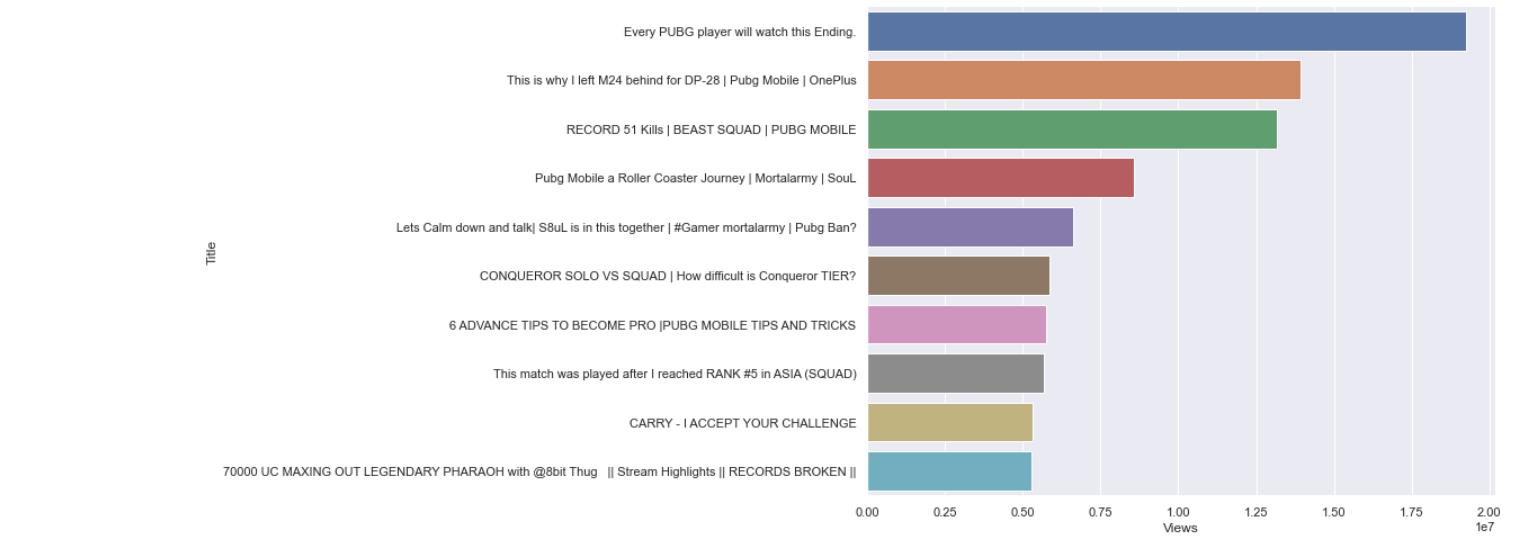


Table 5 Exploration statistics

The dataset is also classified based on the top 10 videos for every youtuber.

Mortal Top 10 Videos

Figure Mortal channel top ten videos

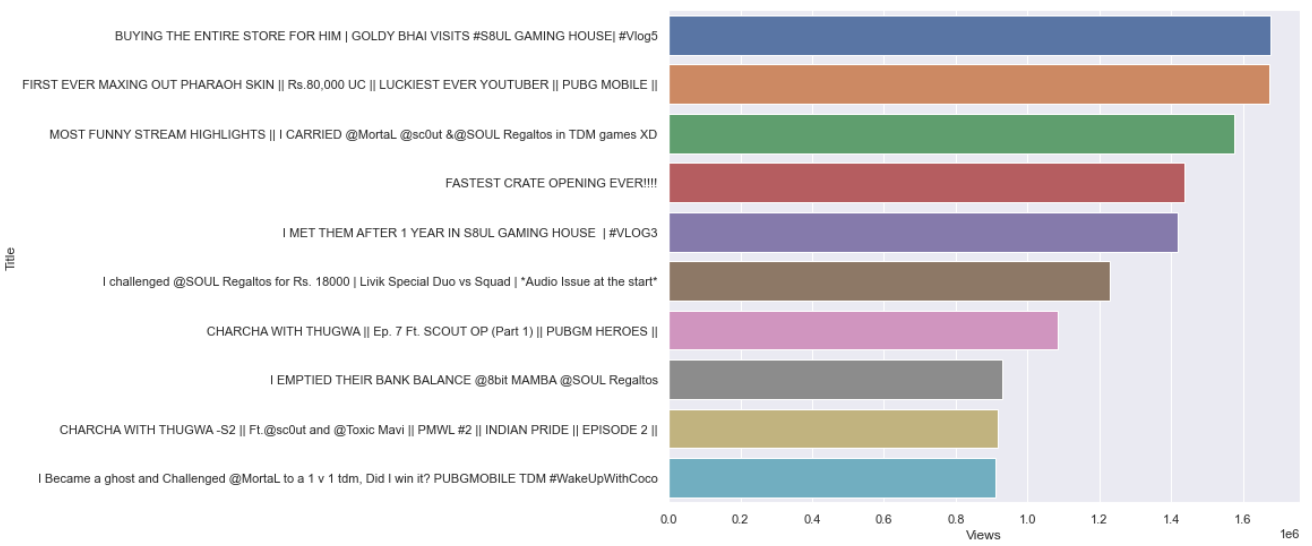
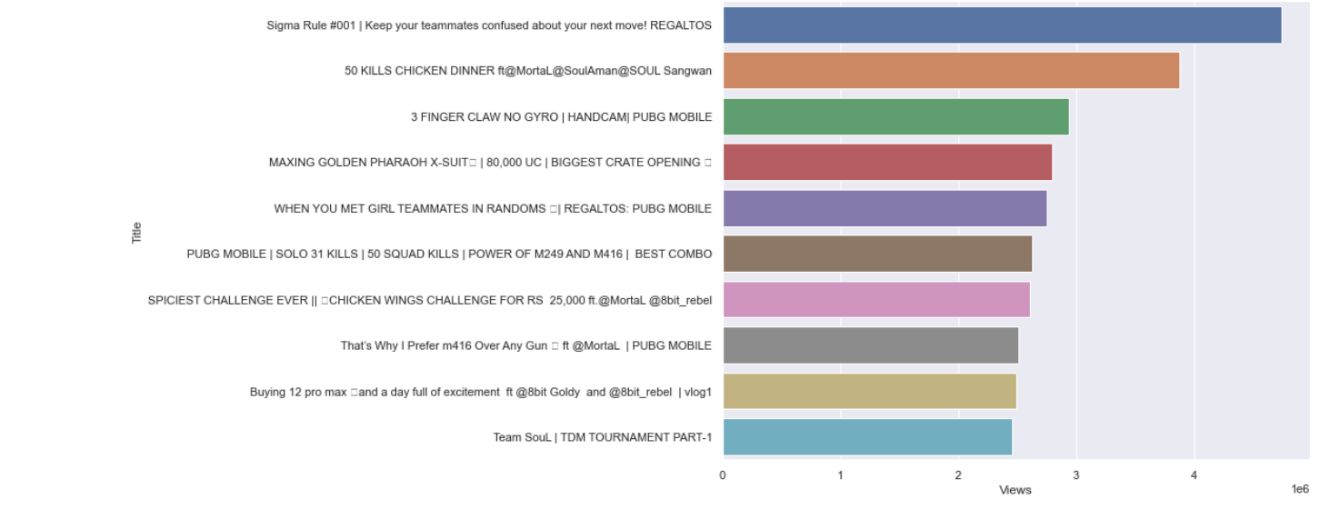


Figure 8bit thug channel top ten videos



1. Methodologies

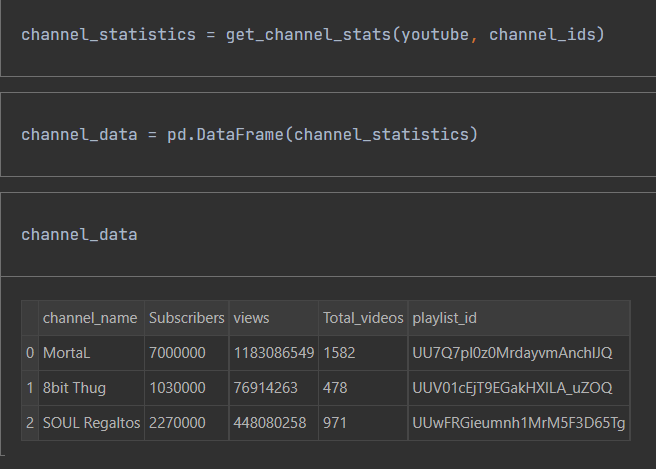
Figure Regaltos channel top 10 Videos



What is your method?

**Data Preparation**: To begin the project, the dataset was loaded into a pandas data frame, which is a very standard tool to work with.

Figure Data preparation for three data set’s "View" "Subscribers" "Playlist"



It includes the total views, subscribers, and playlist identifiers for the three datasets that are considered.

After fetching the data, it was identified that the dataset required more features. As a result, the Data column was split into four features—Title, Published date, Views and Month for all the three datasets.

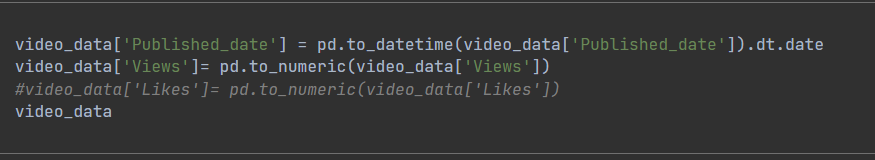


Figure Video data

Figure Data preparation for three data set’s "View" "Subscribers" "Playlist"

**Channel name – Mortal**



Figure 6 Mortal

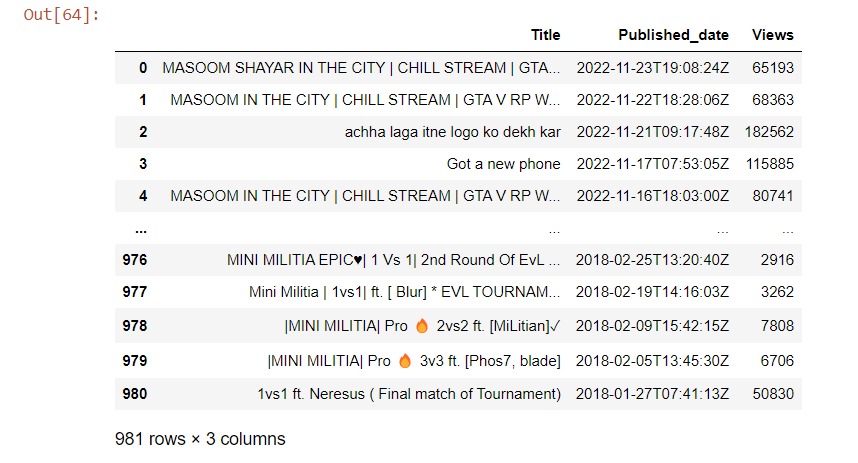
**Channel name – Regaltos**

Figure Regaltos

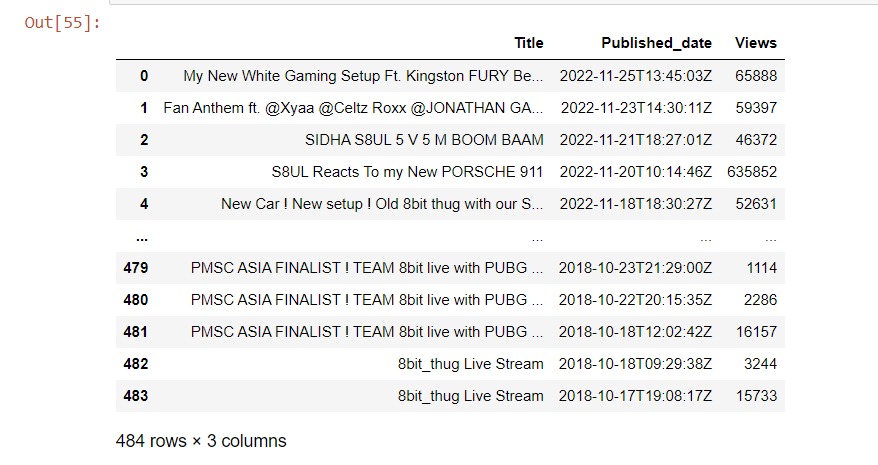
**Channel name – Thug**

Figure Thug

Later, the datasets are converted into .xlsx file to perform T-test and Linear Regression.

**2.Determining the P-Value:**

To determine whether there is a connection between the number of videos and views, research has been conducted. T-test is used because it is effective in determining whether the null hypothesis may be accepted or rejected for a specific dataset. The results of T-test for the three datasets are as follows-

**Channel name – Mortal**

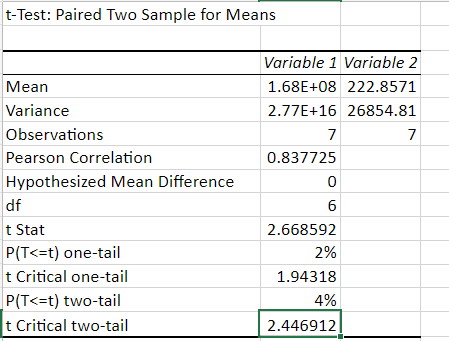


Table 6 Mortal channel two sample for mean calculation

**Channel name** – **Regaltos**

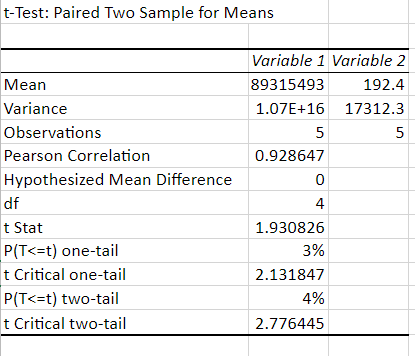


Table 7 Regaltos two samples for mean calculation

**Channel name – Thug**

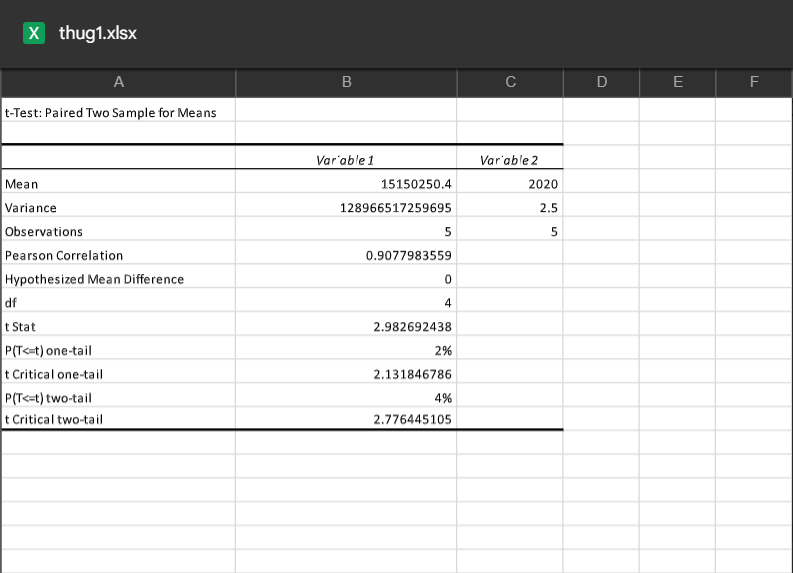


Table 8 Thug channel two samples mean calculation

From the above observations, it is evident that for all three datasets, the P-value, or probability value, which establishes statistical significance, is much lower. Statistics show that the outcome is substantial. In general, choose the alternative hypothesis over the null hypothesis. The Probability value for the three datasets are as follows-

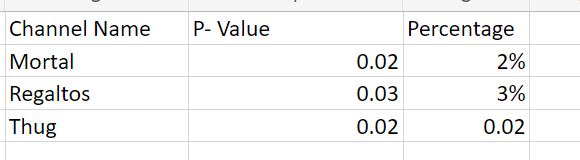


Table 9 Three channels with two variable mean calculation

From the aforementioned facts, it is noticeable that for all three datasets, the P-value is less than 0.05 or 5% which concludes that Alternative Hypothesis is accepted I.e., there is an evident connection between views uploaded and views.

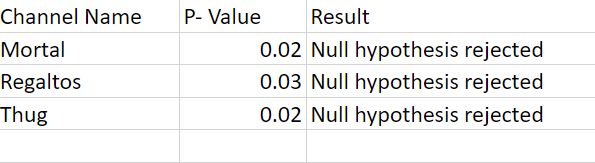


Table 10 3 Data sets with two variables

**Linear Regression and Data Visualization**

It is used to assess the strength and direction of a relationship between two variables. In other words, on a scatterplot, a correlation coefficient(R-value) is calculated between two variables. This can be used to determine whether there is a positive or negative correlation between the variables. The two variables chosen here are the number of videos and the number of views in a year. A scatterplot is plotted with Average Yearly Views on the X-axis and Year on the Y-axis. The trendline for each dataset has increased significantly year after year as the number of videos has increased. To the existing dataset with Title, Published date, and Views, Month, Few more features- “Year, Average Yearly views, No. Of videos” for all the three datasets are added for better analysis.

Flow chart

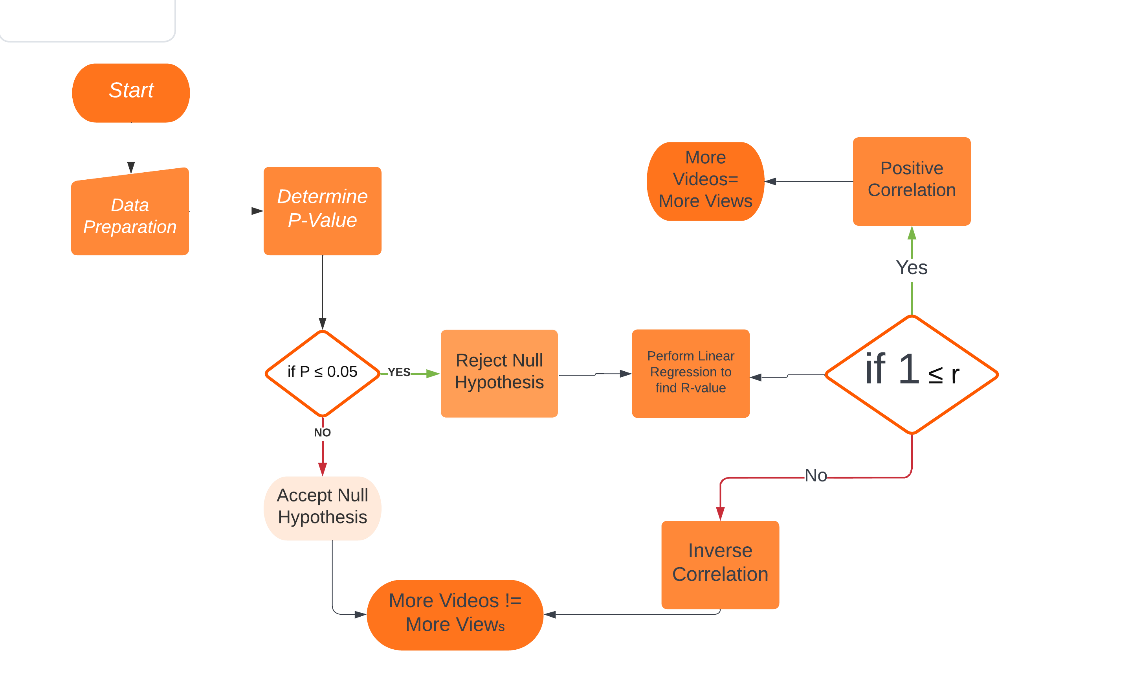


Figure Process diagram

1. Verification and Validation

Test Method

To precisely assess the relationship between the number of videos and annual views Paired To determine whether the P value accepts or rejects the null hypothesis, T-test Paired Two Sample for Means analysis was conducted. The Coefficient of determination (R-value), which is useful to determine the strength of a relationship, has been predicted later using linear regression for average yearly views and the year.

**T-test Paired Two Sample for Means**

The t-Test Paired Two Sample for Means tool uses a paired two-sample test to determine whether or not the null hypothesis (means of two populations are equal) can be accepted. The variances of both samples are not assumed to be equal in this test.This utility generates a calculated t-value as its output. Depending on the facts, this value may be either negative or positive. Considering the population means to be equal

a.) The likelihood that a value of the t-Statistic will be detected that is more negative than t is known as P(T = t) one-tail if t 0.

b.) If t > 0, P(T=t) one tail is the likelihood that an t-Statistic value greater than t will be detected.

c.) P(T <=t) The likelihood that an t-Statistic value bigger in absolute value than t will be observed is known as the "two tail" probability.

**P-Test**

In hypothesis testing, a p value is used to help you support or reject the null hypothesis. The p value is proof against the null hypothesis. The lower the p-value, the stronger the evidence that the null hypothesis should be rejected. A low p value ( 0.05) would rule out the null hypothesis. The validity of the null hypothesis is clearly demonstrated by this. A high p value (> 0.05) indicates that the alternative hypothesis is tenuous, hence the null is not rejected.

**Null Hypothesis:** **The null hypothesis considered here is there is no evident relationship between the number of videos uploaded and the views.**

According to the null hypothesis, there is no correlation between the independent and dependent variables of a population. If the results support the hypothesis and there is a correlation between the two parameters, an experimental or sampling error may have occurred. The measured phenomena does, however, have a relationship if the null hypothesis is incorrect.

**Alternative Hypothesis:** **The alternative hypothesis provided here is that it shows a significant increase in the views as the more number of video uploads increases.**

The opposite of a null hypothesis is an alternative hypothesis. Only one of the two hypotheses may be true because an alternative hypothesis and a null hypothesis are mutually exclusive. Between the two variables, there is a statistically significant relationship. Statistical significance exists between the two variables and the alternate hypothesis is true if samples used to test the null hypothesis yield a misleading result.

**Linear Regression**: By applying a linear equation to observed data, linear regression can predict the relationship between two variables. The term "variable" refers to one of two types of variables: independent variables and dependent variables. Linear regression is a popular method for performing predictive analysis. The main idea behind regression is to look at two things. First, can a set of predictor variables accurately predict an outcome (dependent) variable? The second question concerns which variables are significant predictors of the outcome variable.

**R-value**: It is a Pearson's correlation coefficient Or, to put it another way, R is a correlation coefficient that assesses both the strength and the direction of a relationship between two variables on a scatterplot. R's value is always between a -1 and a +1 (negative and positive).

**Positive Correlation**: The dependent variable rises as the independent variable rises in a positive correlation. The scatter plot visually slopes higher due to the creation of a positive slope. This indicates a strong relationship.

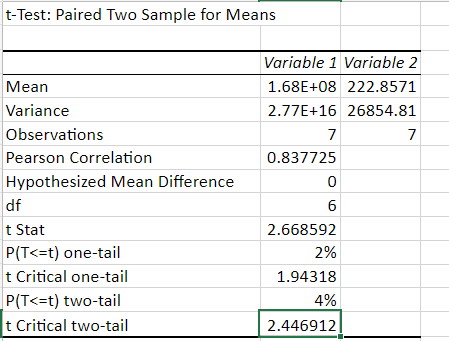
**Negative Correlation:** The dependent variable declines when the independent variable rises in a negative correlation. This results in a downward slope. This is represented by downward-pointing points. This indicates a strained connection.

Test Results and Explanations:

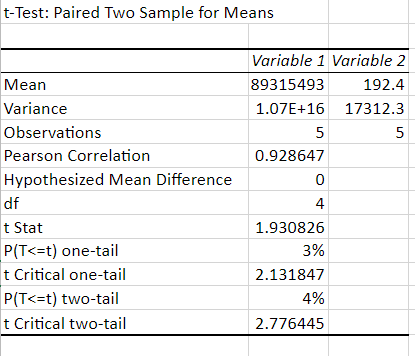
**T-test Paired Two Sample for Means Results**

This test was performed in order to determine the probability value of the given sample. The T-test was calculated for all three datasets, and it significantly indicates that the p-value for all three datasets is less than 0.05 or 5%. This implies that there is a strong relationship between the number of videos and the number of views. The outcomes are as follows.

**Channel Name- Mortal**

 Table 11 Mortal two samples for mean

**Channel Name- Regaltos:**



**Table 12 Regaltos two samples for means**

**Channel Name- Thug:**

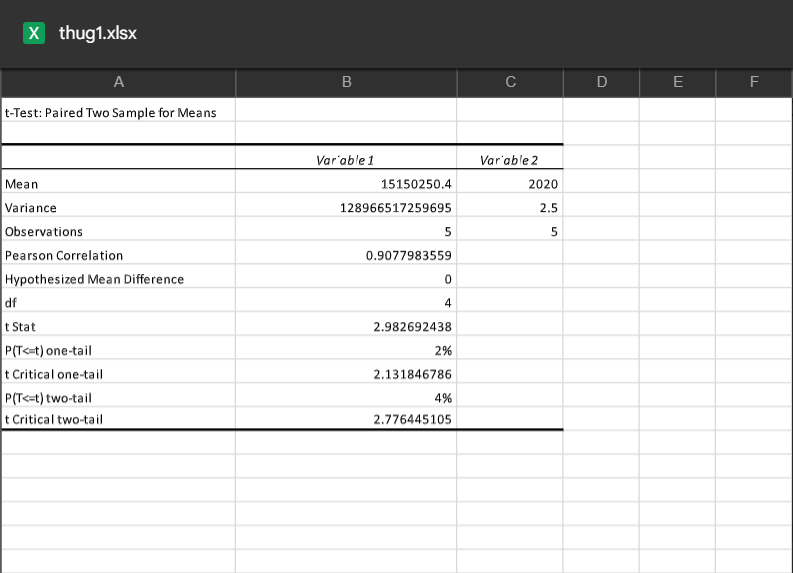


Table 13 Thug channel two variable for means

More precise table containing the probability values of all three datasets is given below-

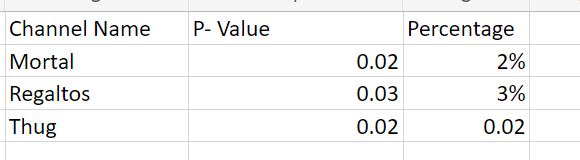


Table 14 3 Channel two variables for mean

1. **Linear Regression and Visualization**

Once the partial result is obtained (that is, the probability value analyzed using the statistical hypothesis T-test), the next step is to calculate the R-value, also known as the Correlation Coefficient. Positive correlation for the two variables was found in all three datasets. This strongly implies that the independent and dependent variables have a strong relationship. This linear regression has also been plotted on a scatterplot, with the Y-axis representing Average Yearly Views and the X-axis representing the number of videos and the year in which the video was uploaded. The scatterplots for all the three datasets are

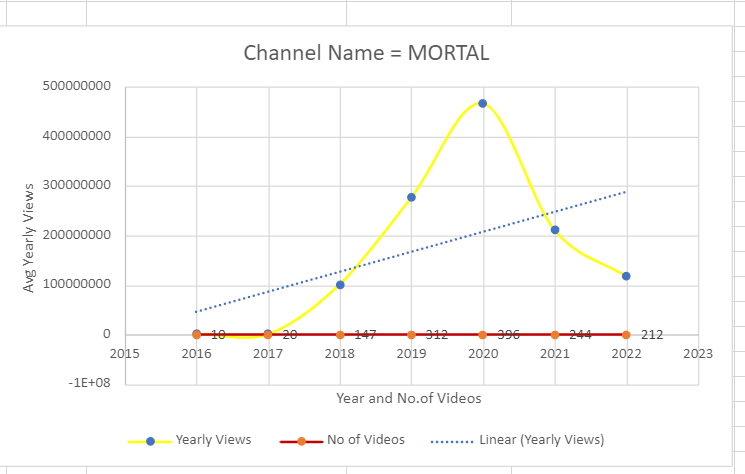
**Channel Name- Mortal:**

Figure 10 Mortal channel statistics

The scatterplots above show that the number of videos increases the number of views. As it is clearly proved that in the year 2019 the content creator have uploaded 312 videos and likely 396 videos were uploaded in the year 2020 and the viewership has a significant rise with the video uploads increased. This, in turn, leads to an increase in the number of subscribers. This conclusion was reached due to the obvious strong relationship between the independent variable - Year - and the dependent variable Views. Later on the video uploads were decreased in the year 2021 and also in the year 2022 and the viewership significantly dropped down w.r.t to the video uploads. The correlation coefficient values obtained indicate that the scatterplots above are highly correlated.

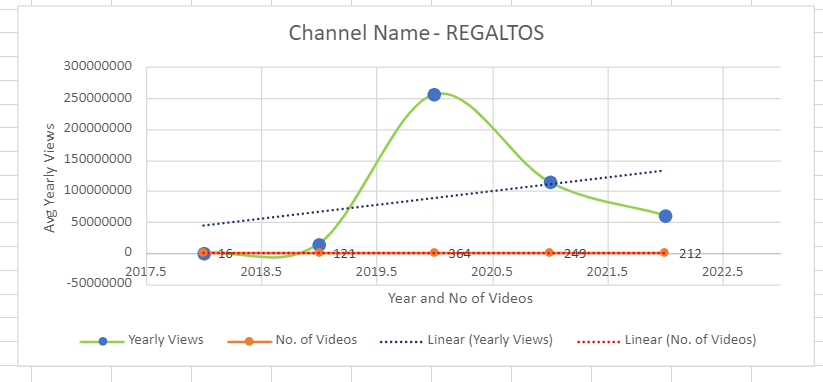
**Channel Name- Regaltos**

Figure 11 Regaltos statistics

The total number of videos uploaded in the initial stage were quite low and the similar is shown with respect to views. As the number of videos uploaded raised to 364 in the year 2020, A drastic change in the trendline is shown and the same is followed for the later video uploads and views. This, in turn, leads to an increase in the number of subscribers. This conclusion was reached due to the obvious strong relationship between the independent variable - Year - and the dependent variable Views. The correlation coefficient values obtained indicate that the scatterplots above are highly correlated.

The decrease in curve in the later year is a proof that the uploaded videos were lessas shown in the scatterplot.

**Channel Name- Thug**

Chart, line chart

Description automatically generated

Figure 12 Thug statistics

The scatterplots above show that the number of videos increases the number of views. Though there is a flat curve in the above graph but it still leads to the increase in the viewership as the Views are calculated averagely compared to the number of videos uploaded in that particular year. This, in turn, leads to an increase in the number of subscribers. This conclusion was reached due to the obvious strong relationship between the independent variable - Year - and the dependent variable Views. The correlation coefficient values obtained indicate that the scatterplots above are highly correlated.

# Future Work

This project is not only limited for the YouTube gaming content creators but it can also be developed for all type of content. In this project it is only limited to predict the future views with the help of linear regression trend line but in future this can be developed to predict the like count and also the comment count of the videos. The comment count and the like count of the videos are based on the viewership, to calculate the viewership of the channel this project might be helpful.

6.Conclusion

After performing the T-test to calculate the p Value, aforementioned facts, it is noticeable that for all three datasets, the P-value is less than 0.05 or 5% which concludes that Alternative Hypothesis is accepted I.e., there is an evident connection between videos uploaded and views. This implies that there is a strong relationship between the number of videos and the number of views.

After Performing Linear Regression for the three data sets, the scatterplots in the above document show’s that the number of videos increases the number of views. This, in turn, leads to an increase in the number of subscribers. This conclusion was reached due to the obvious strong relationship between the independent variable - Year - and the dependent variable Views. The correlation coefficient values obtained indicate that the scatterplots above are highly correlated.

This project will discuss a variety of actions that YouTube gaming streamers can do to raise the number of views on their videos and the typical audience of their streams. After much research and performing couple of tests to understand how exactly the YouTube algorithm works and how Gaming content creators can achieve popularity, it can be concluded that - "the more videos you publish, the more views you receive." This is only done for gaming streamers on YouTube. Although it is difficult to create a gaming channel that understands the YouTube algorithm, you will likely find more success as your channel continues to expand.

Git hub repository Link : *https://github.com/NititishGU/Nitish-final-project*

1. Bibliography

|  |  |
| --- | --- |
| [1] | J. Doe, "Strange papers are worth citing," in *IEEE Conference on Strange Report Writing*, TimbukTu, 1995. |

**Appendix A: Source Code**

pip install --upgrade google-api-python-client google-auth-httplib2 google-auth-oauthlib

from googleapiclient.discovery import build

import json

import pandas as pd

import seaborn as sns

api\_key='AIzaSyCGhvduj5m5WVrih73FVhQ2fOoYH6KbLcg'

channel\_ids=["UCwFRGieumnh1MrM5F3D65Tg", #Soul\_Regaltos

"UC7Q7pl0z0MrdayvmAnchlJQ", #Soul\_Mortal

"UCV01cEjT9EGakHXILA\_uZOQ" #8bit\_thug

]

youtube = build('youtube','v3', developerKey = api\_key)

#Function to get channel statistics

def get\_channel\_stats(youtube, channel\_ids):

all\_data = []

request = youtube.channels().list(

part='snippet,contentDetails,statistics',

id=','.join(channel\_ids))

response = request.execute()

for i in range(len(response['items'])):

data = dict(channel\_name = response['items'][i]['snippet']['title'],

Subscribers = response['items'][i]['statistics']['subscriberCount'],

views = response['items'][i]['statistics']['viewCount'],

Total\_videos = response['items'][i]['statistics']['videoCount'],

playlist\_id = response['items'][i]['contentDetails']['relatedPlaylists']['uploads'])

all\_data.append(data)

return all\_data

channel\_statistics = get\_channel\_stats(youtube, channel\_ids)

channel\_data = pd.DataFrame(channel\_statistics)

channel\_data

channel\_data['Subscribers']= pd.to\_numeric(channel\_data['Subscribers'])

channel\_data['views']= pd.to\_numeric(channel\_data['views'])

channel\_data['Total\_videos']= pd.to\_numeric(channel\_data['Total\_videos'])

channel\_data.dtypes

sns.set(rc={'figure.figsize':(10,8)})

ax = sns.barplot(x='channel\_name', y='Subscribers', data=channel\_data)

playlist\_id = channel\_data.loc[channel\_data['channel\_name']=='8bit Thug','playlist\_id'].iloc[0]

playlist\_id

def get\_video\_ids(youtube, playlist\_id):

request = youtube.playlistItems().list(

part='contentDetails',

playlistId = playlist\_id,

maxResults = 50)

response = request.execute()

video\_ids = []

for i in range(len(response['items'])):

video\_ids.append(response['items'][i]['contentDetails']['videoId'])

next\_page\_token = response.get('nextPageToken')

more\_pages = True

while more\_pages:

if next\_page\_token is None:

more\_pages = False

else:

request = youtube.playlistItems().list(

part='contentDetails',

playlistId = playlist\_id,

maxResults = 50,

pageToken = next\_page\_token)

response = request.execute()

for i in range(len(response['items'])):

video\_ids.append(response['items'][i]['contentDetails']['videoId'])

next\_page\_token = response.get('nextPageToken')

return video\_ids

video\_details = get\_video\_details(youtube, video\_ids)

video\_data = pd.DataFrame(video\_details)

video\_data['Published\_date'] = pd.to\_datetime(video\_data['Published\_date']).dt.date

video\_data['Views']= pd.to\_numeric(video\_data['Views'])

#video\_data['Likes']= pd.to\_numeric(video\_data['Likes'])

video\_data