Task 2

A1. Research Question/ Organizational Need

For this project, the 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 company's advertisements and what factors influence the subscribers and views that the channel creates? This is an important organizational need for a 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.

A2. Context/ Background

A marketing company has hired a data analyst to analyze a recent Global YouTube dataset containing information about YouTubers and YouTube channels. The marketing company sells ads for a variety of businesses. The dataset contains fields such as channel name, category, number of subscribers, views, rank, uploads and many more that the analyst will use to gain statistical and logical insight about the dataset. They will be able to offer the marketing company with the analysis results to help them make the most informed decision on which channels they should be targeting to sell their advertisements to. The analyst should look specifically at which channels are the most popular and generate the most views and subscribers. This will give insight to which channels will provide the largest audience. Another field the analyst is tasked to gain insight on is which categories/niche fields offer the most benefit, by

allowing them to match the audience of the channels they decide to audience the company is targeting.

A3. Summary of Published Works

In 2014, Gary W Randazzo's book Developing Successful Marketing Strategies was published. Randazzo speaks to many aspects of successful marketing strategies for businesses and breaks each chapter into specific categories, some of which validate the thought process used to complete the project of analyzing YouTube data. Randazzo states in chapter XX p64: "*There is abundant information available on media use by demographic and psychographic characteristics. Matching each of the advertiser's customer groups using these characteristics with the appropriate media improves the chances that the targeted advertising message will reach the right audience."* (Randazzo, 2014).

This book informs the development of the project by showing the importance of reaching the right audience. It speaks to how you can match the characteristics of the advertisement target audience with where you plan to share your advertisement. I will use this information when developing my project of analyzing Global YouTube Data by looking into what categories/niche channels to target based on the advertisements available to promote. It will be a focus to ensure that I am catering similar content to my target audience as that will get the message or product to the viewers most likely to explore the promoted product, service, or message.

In AUDIENCE: Marketing in the age of subscribers, fans & followers written by Jefferey K, Rohrs, published in 2014, Rohrs speaks to many important aspects of marketing that apply to this project. In AUDIENCE Increase what matters: Size, Engagement and Value, Rohrs speaks specifically about the importance of size when it comes to marketing breaking size up into three

categories; Relative Size, Database size and Reach. *"1. Relative Size: the size of your audience compared to direct competitors. 2. Database size: volume and quality of the data you have gathered about your audiences to improve message relevance. 3. Reach: the percentage of your audience that sees your message."* (Rohrs, p.71). He speaks to the relevance of these three types of size and why they are important for marketing.

This source informs development of this project by giving guidance on what are the important factors to look for when completing the work of analyzing the YouTube dataset. This source will help decide the right factors I should be analyzing throughout this project. It has made it very clear that the audience size is an important factor to look at, I will use this information and apply it to my project by making sure to compare channels to the audience size by looking at the number of subscribers and views in relation to specific YouTube channels.

In YouTube marketing: A complete guide for your brand written by Mahnoor Sheikh, Sheikh talks about the important factors of YouTube marketing hitting on many factors that I will be using in my analysis. **"***70% of consumers say they've bought a product after seeing it on YouTube. Running ads on the platform can deliver high returns for your brand*." (YouTube Marketing, Sheikh). This shows the importance and high return advertisements on YouTube can generate, which supports the entire basis of this project proposal. *"YouTube has an incredibly large, active user base. Publishing quality content consistently and partnering with influential creators in your niche can get you massive exposure and traffic."* (YouTube Marketing, Sheikh). This quote speaks to the vast userbase of YouTube and further validates that partnering with creators in your niche can provide large exposure for your ads. This information further informs development of this project by providing the basis for which factors are important to consider, in order to find the best avenue for high exposure and traffic of your advertisements. This helps me

use this knowledge and apply it to my analysis by comparing certain niches with higher traffic to ultimately make the most informed decision of which YouTube channel to sell our ads to maximize profit.

A4. Summary of Data Analytics Solution

My solution for this project is to use a dataset of Global YouTube Data that has been provided as a downloadable dataset from Kaggle.com. This data set consists of twenty-seven columns and 996 rows that contains the necessary fields such as: category, number of views, subscribers, uploads, and rank to answer my business question. The dataset I will be using will be fully downloadable as a zip file. Upon downloading the dataset, I plan to use Jupyter notebooks using python as the language to gather, clean, and analyze the dataset to gain vital insight. The insight obtained will then be used to make final decisions on which YouTube channel to sell advertisements to. I will be using Pandas Library within my Jupyter Notebook to import the CSV file using read.csv. This will allow me to fulfill the next steps of the project where I can clean and analyze the dataset. I will be using NumPy, Matplotlib and Seaborn Libraries to compare relationships 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, rank, and uploads that will provide further insight to the dataset. To prove statistical significance, I will use Pearson library to conduct a correlation test using two quantitative variables (views and subscribers) to reject or validate my hypothesis. In performing this analysis, I will be able to provide the necessary information needed to make the most informed decision of which channels to target for selling advertisements to.

A5. Benefits of Analysis

The benefits of the analysis of this dataset are to provide a marketing business with statistical and practical significance from the dataset which will help the marketing business make the most informed decision that will be able to benefit from most. The analysis will provide insight about which YouTubers have the most subscribers, views, and rank. This is a benefit as it will help the business see which YouTubers to target based on their following as the YouTubers with the most subscribers and views are going to provide the largest audience for the marketing company to take advantage of. I will be looking at which categories are most popular in terms of views as well. This will provide the company with vital data to help them decide which categories/ niches to target for the sales of their ads. These analysis tactics befit and support the projects main task which is to provide the marketing company with sufficient data analysis that ultimately helps them make the most informed and beneficial business decision.

B1. Project Goals, Objectives, and Deliverables

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 will consist 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 end result of this project will be a slideshow presentation of the data analyst's findings that clearly shows the benefit certain YouTubers. This will steer the business in the direction to make the best decision of which YouTuber to sell their advertisements to.

Objectives and Deliverables:

- Determine if there is a correlation between subscribers and views.
    - The deliverables of this objective will be to create a scatterplot that visually depict the relationship between the two variables and perform a correlation test between the two variables that will result in statistical significance.
- Evaluate relationships between fields to determine which YouTubers are most successful and which categories get the most views.
    - The deliverables of this objective will be creating bar charts that depict the top YouTuber by subscribers and views. This will show which YouTubers will be worth investing in.
    - Creating bar charts that produce the top categories by views and subscribers will represent which categories are generation the most attention.

B2. Project Scope

The scope of the project is to gather and analyze a dataset using python in Jupyter Notebooks using Pandas, NumPy, Matplotlib and Seaborn 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 a hypothesis. The analyst will be looking to analyze comparisons between subscribers & views, subscribers & YouTuber, views & YouTuber, categories & subscribers, and subscribers & uploads. Using these fields to compose statistical analysis using a correlation test as well as creating appealing data visualizations by creating scatter plots and bar charts that provide the marketing company with data insight. The project will end with a slideshow presentation where

the analyst will show off their findings to key decision makers in the company that will help the marketing company make an informed decision in the end.

Evaluating other relationships such as in the dataset such as population, latitude, longitude, unemployment rate and others are outside of the scope of this project. The scope of this project is also limited to YouTube statistical data from 2023 only and data from outside the year 2023 is out of scope.

B3. Project Methodology

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. Agile development follows a 5-stage cycle Meet & Plan, design, develop, release, and feedback. Below are how these five stages will be implemented to complete this project under Agile methodology:

Meet & Plan

During this stage in the project, it will be vital to meet and communicate thoroughly the requirements of the project. This will encompass a project timeline, budget, requirements, potential scope changes. During this phase I will meet with the marketing company and retain all

the requirements and details that pertain to the project to ensure I am on the same page as the customer so that I can fulfill their expectations.

Design Phase

During the design phase I will gather and import the dataset to review. I will be reviewing the dataset with the details and requirements from the marketing company in mind. In this phase I will be looking at the dataset and planning how I will go about performing analysis on the dataset to fit the scope and requirements laid out in phase one. In this phase I will also determine which analytical methods to use to complete the project following the specified guidelines.

Develop Phase

During the develop phase I will be performing the python scripting to complete the work for the project. I will start cleaning the dataset and get it to a workable place where I can then begin evaluating and analyzing the dataset following the plan outlined in the design phase by performing statistical and exploratory analysis.

Release Phase

During the release phase I will be communicating with the marketing company, sharing my work and the results of the project in an iterative manner. This phase begins after the completion of the develop phase.

Feedback Phase

During this phase, the marketing company can provide feedback about the work that I have shared to them in the release phase. If there are changes to the scope or other requirements, they would like me to investigate during my analysis they will offer that input during this phase. Based on the results of the feedback phase the cycle may start over through another iteration.

B4. Timeline and Milestones

Milestone	Projected Time	Start Date	End Date
Gather/Import Dataset	**2 Hrs**	**9/20/2023**	**9/20/2023**
Clean Dataset	**8 Hrs**	**9/20/2023**	**9/21/2023**
Evaluate and Analyze Dataset	**10 Hrs**	**9/21/2023**	**9/23/2023**
Create Visualizations	**6 Hrs**	**9/23/2023**	**9/24/2023**
Create Slideshow of Results	**5 Hrs**	**9/25/2023**	**9/26/2023**
Present Slideshow of Analysis Results	**1 Hr**	**9/31/2023**	**9/31/2023**

B5. Resources and Costs

Resources needed and used for this project and associated costs:

- Anaconda
    - No cost
- Jupyter Notebook
    - No cost
- PowerPoint
    - Provided from WGU (No cost)
- Work Hours
    - No cost

The are no costs of the resources used for this project as the dataset was free to download, Anaconda and Jupyter Notebook are free to use, PowerPoint is provided through Microsoft 365 through WGU at no cost and I am using my own labor which is no cost to me.

B6. Criteria for Project Success

For this project to be deemed successful the following criteria must be met:

- Analysis must provide statistical significance using a correlation test to validate or disprove hypothesis.
- Visualizations must be made to provide insight on analysis.
- PowerPoint Slideshow must be created to be used to present to the marketing company that answers the primary business question and fulfills the project goals and requirements.

C1. Hypothesis

The purpose of this project is to identify which YouTuber to sell a marketing companies ads to and this will be done by analyzing relationships in the dataset. The goal is to understand what quantitative relationships are relevant to success and relate to the most views. Views are viewed as a major part of this project analysis as if a video generates the most views, they are generating the largest audience. That data will be used as a guide to make the final decision of what YouTuber to sell ads to.

Null Hypothesis: There is no correlation between the number of subscribers and the number of views generated.

Alternative Hypothesis: There is a positive relationship between number of subscribers and the number of views generated.

C2. Analytical Method

In this project I will be using both qualitative and quantitative data to observe and compare relationships between fields in the dataset. I will be using descriptive statistics by performing a Pearson correlation test to compare relationships between fields such as views and subscribers to determine the strength and direction of the relationship. This analytical method is appropriate for my project because I will be evaluating the correlation between two quantitative variables and correlation tests are used when determining linear relationships between two quantitative variables. This aligns with my project as I am trying to determine if there is positive correlation between these variables which will either validate or reject my hypothesis and the results can be used to prove statistical significance that will be used in proving the significance of subscribers and views.

C3. Tools and Environment

The tools that I plan to 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 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 will be using Pandas, NumPy, Matplotlib, Seaborn and Pearson libraries. Pandas Library will be used to import the dataset I will be using for this project into my Jupyter Notebook, which will allow me to do all the work needed to complete this

project. The NumPy Library allows for performing mathematical functions and help dealing with NaN values. Matplotlib and Seaborn Libraries will be used to create plots and charts that will help paint a visual picture of the dataset. Finally, Pearson Library will be used to perform descriptive statistics by performing a correlation test that will give me statistical information about the relationships between fields.

C4. Statistical Significance

I will be using the Pearson correlation test to evaluate the statistical significance of my data analytic solution. I will be comparing the relationships of two quantitative variables: views and subscribers from my dataset. The null hypothesis will be that there is no relationship between views and subscribers. While the alternative hypothesis will be that there is a positive relationship between views and subscribers. I will be performing a correlation test between these two variables in order to prove statistical significance. I will be looking at the correlation coefficient and the p-value results from this test to determine statistical significance. If the correlation coefficient is a positive number, we will know that there is a positive relationship between the two variables and the closer the number is to 1 the stronger that relationship is. If the correlation coefficient is a negative number, then we will know that the variables have a negative relationship. If the p-value is less than 0.5 then we will know the results are statistically significant and we can reject the null hypothesis. If the p-value is greater than 0.5 then it means, there was a deviation from the null hypothesis and the null hypothesis is not rejected. The justification for using this type of test to prove statistical significance is because of the types of variables I am using as a part of my hypothesis. Since I am using two quantitative variables using a correlation test is most appropriate. The results of this test are important and appropriate for

this project because views are an important aspect of this project as more views equate to a larger audience. If the alternative hypothesis proves true and the null is rejected it means that more subscribers also equate to a larger audience. This is important to know when evaluating which YouTubers to sell ads to as the largest audiences are the most important factor.

C5. 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. Having an understanding of 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.

The practical significance of this project is to determine which YouTuber to sell ads to by evaluate relationships between variables that will help guide which YouTuber to sell ads to. The practical significance of this project is to identify which YouTuber/categories generate the largest audiences. 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 for which they are hoping. I will evaluate many variables in this process to help support this. It will be important to know which YouTubers have the most views, subscribers, and rank to evaluate which channels are getting the most attention. The more attention a video gets the more attention a potential advertisement played on that channel will generate. Another important criterion to evaluate is the categories

with the greatest number of videos, this is beneficial to know because it gives the marketing company intel on which niches may me overcrowded. I will also be evaluating the relationship between the number of subscribers, views, and the category. This will be extremely beneficial to the marketing company as it will be important to know which genres are the most popular and are generating the most viewers. This is very important as the marketing company will be able to look at the top viewed and subscribed to categories to make sure they are taking advantage of getting the most eyes on their ads as possible.

C6. Graphical Representations

Throughout this project I will be using Matplotlib on my Jupyter Notebook to create data visualizations that will portray the findings of my data analytics solution. I will be using bar graphs to portray some of my findings. These bar charts will visually communicate which YouTubers generate the most views, subscribers, and the highest rank. I will also be using bar charts to describe the top YouTube categories. Finally, to visually communicate the relationships between two quantitative variables, specifically views and subscribers I will be using scatterplots, which will further support my hypothesis. The data visualizations will help paint a clear picture of the results of my data analytics solution that will be used to explain the benefits of investing in certain YouTubers and categories.

D1. Dataset Source

The dataset that I will be using for the completion of this project is called Global YouTube Statistics 2023 and it is provided by Kaggle.com. This dataset consists of 27 columns and 996 rows that contain statistics from YouTube from 2023 and includes the necessary fields to answer my business question. It is provided by Kaggle.com and is accessible as unrestricted

data to Kaggle users and is downloadable as a zip file directly from the site. This dataset can be found at: https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023.

D2. Dataset Goals

The goal of this project is to analyze a dataset using YouTube data that will ultimately help a marketing company that hired you make an informed decision of which YouTuber to sell their ads to that would be most beneficial to them. The dataset that will be used for this project is a dataset that contains YouTube statistics from 2023. This dataset is appropriate to accomplish the goals of this project because it contains the necessary fields that are needed to make the evaluations that will help the company reach their goal of choosing the right YouTuber. Some fields that are vital to be used are the YouTubers, category, rank and number of subscribers and views. These variables will be able to be used to find statistical and practical significance that proves beneficial for the marketing company. It is also important that the dataset is composed of most recent data because the analyst needs to provide analysis that is current. Since the dataset is statistics from 2023 it is appropriate to use in order to answer the business question.

D3. Collection Methods

The dataset used for this project is accessible to the public from Kaggle.com for Kaggle members. The dataset is downloadable as a zip file directly from the sight. The data collection

methods used are to download the dataset as a CSV file, then upload the dataset into Jupyter Notebook using pandas read_csv.

D4. Data Quality

The quality and completeness of this dataset was relatively good. There is not too much cleaning that I anticipate being necessary to work with this dataset. The main issue with the dataset is the number of null values from certain fields, these null values will have to be evaluated and dealt with appropriately. A few field names could be changed for extra clarity and to make the dataset easier to work with and some columns may be dropped due to them not being needed to complete the scope of this project. Aside from that the entire dataset is complete is of decent quality straight from the source.

D5. Dataset Privacy and compliance

This dataset is provided from a public source on Kaggle.com, this dataset does not list any license and is encouraged to be used and explored to the public. Due to this dataset being freely available to the public and it does not contain any personal identifiable information (PII) there are no major risks of privacy, security, ethical, legal compliance, or governance issues.

References

2014, Rohrs, K Jefferey, AUDIENCE: Marketing in the age of subscribers, fans & followers, https://eds.p.ebscohost.com/eds/detail/detail?vid=0&sid=f28e3536-a3e6-4fdf-8f88-57d190bd13f8%40redis&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=ebc.EBC1547075&db=cat07141

2014, Randazzo, W Gary, Developing Successful Marketing Strategies https://ebookcentral.proquest.com/lib/westerngovernors-ebooks/reader.action?docID=1683379

2022, Mahnoor Sheikh, YouTube marketing: A complete guide for your brand https://sproutsocial.com/insights/youtube-marketing/