Question:

Did the increase in the number of Covid-19 cases between June 2019 and June 2021 increase the number of subscribers for online streaming platforms for Hulu, Netflix, in the US?

Hypothesis:

> *We believe that the increase in the number of COVID-19 cases has resulted in the increase in the number of subscribers for Netflix, and Hulu, and we also believe that Netflix will have the greatest increase in the number of subscribers due to the increase in COVID-19 cases.*

The **Justification** for our hypothesis is based on the fact that most of our group members, from personal experience when we decided to subscribe to Netflix it was during COVID-19. Our instinct is that the decision to subscribe to an online streaming platform is due to Covid-19 lockdown which forced us and people living in the US and Canada to stay at home more often and much longer when compared to precovid within the 2019 year specifically. Staying at home resulted in people tending to have more time at home hence resulting in an increase in the number of subscribers to online streaming platforms.

Background Information:

Our project wants to explore if the increase in the number of COVID-19 cases has an impact on the number of subscribers of the listed streaming platforms: Netflix, Hulu.

We decided to make the time period between 2019 and 2021, because given the relative lack of coronavirus cases in the US in 2019, it can be deemed as the “Pre-COVID” period, whereas the subsequent years - 2020 and 2021 - can be considered as during COVID given the increase in number of positive cases.

We decided to use the number of coronavirus cases because the number of cases is a measurement of the severity of the coronavirus in the US . Consequently, the more severe the coronavirus the stricter the policies for a contagious disease that has caused this pandemic. Most of the policies - Lockdown policies - force people to stay at home, which means they’re likely to indulge in indoor leisure activities such as watching movies or TV shows found on streaming platforms.

We plan to include data on the subscribers of that respective streaming platform. By including these variables, we can see if the increase in the number of coronavirus cases correlates with an increase in the number of subscribers of the streaming platform. We will create data visualizations based on linear regression. The data table will consist of values such as number of subscribers of each streaming platform, number of COVID-19 cases, the time period (For eg: June, 2019), etc that we will have found from structured datasets prepared by others or from web-scraping websites. We will perform various data wrangling techniques to compose our own data table from other datasets or web-scraped data.

Ultimately, we will be expanding on the works of other people by combining their datasets so that we can see if the number of covid cases correlates strongly with the increase in the number of subscribers of these streaming platforms.

Links to sources for the variables:

https://ir.netflix.net/financials/financial-statements/default.aspx

www.statnews.com/feature/coronavirus/covid-19-tracker

https://www.statista.com/statistics/258014/number-of-hulus-paying-subscribers/#:~:text=Number%20of%20Hulu's%20paying%20subscribers,U.S.%202019%2D2021%2C%20by%20quarter&text=In%20the%20fourth%20quarter%20of,of%20the%20previous%20fiscal%20year

Data:

The ideal Dataset would keep track of the total number of subscribers of each specific platform in the year of 2019, 2020, and 2021 and would include the following observations; **Time-Period (2019, 2020, and 2021)** which will include the months and years that is relevant to the question we are answering and this observation will help limit the scope of the study and simplify our data and analysis in order to draw meaningful insight that will contribute to answering the question of study. The variables that we would collect are **Number of Subscribers for Netflix, Number of Subscribers, and Number of Subscribers for Hulu** for each month of each year (2019, 2020, 2021) this would track the total number of users of each month of each year during the two periods “Precovid” (2019) and during “Covid” (2020, 2021), and our final variable is the **Number of Covid cases** which will track the spread/severity of Covid-19 within the domestic region**.** With regards to the sources that we are using for our Datasets we were able to find a multitude of data tables and graphs on statista that portray the total number of subscribers on each given year for each streaming platform that we are considering and also, we found data sets that convey the number of COVID

cases per month of the year 2020-2021. The second source are financial reports of the streaming platforms that we are considering. These financial reports depict accurate data regarding the total number of subscribers and total revenue for each quarter of each year within the time period that we are interested in examining. We were able to find Financial reports for Netflix, however we couldn't find a report for Hulu therefore we will be using the web scraping method of different articles/websites to pull the financial information for Hulu. The third source that we plan on using will be through the use of web scraping in which we will use several websites (found below) that provide much more descriptive data regarding the spread of Covid-19 and the number of cases in domestically within a quarterly time period from 2019 to present day (2021).

Financial reports:

- "Netflix - Financials - Financial Statements." *Netflix*, 2021, https://ir.netflix.net/financials/financial-statements/default.aspx
- "The Walt Disney Company - Fiscal year 2020 Annual Financial Report." Walt Disney Company, 2021 https://thewaltdisneycompany.com/app/uploads/2021/01/2020-Annual-Report.pdf

DataSets:

- Statista. "Number of Hulu's Paying Subscribers in the U.S. 2019–2021, by Quarter." *Statista*, 15 Nov. 2021, https://www.statista.com/statistics/258014/number-of-hulus-paying-subscribers/#:~:text=Number%20of%20Hulu's%20paying%20subscribers,U.S.%202019%2D2021%2C%20by%20quarter&text=In%20the%20fourth%20quarter%20of,of%20the%20previous%20fiscal%20year
- https://www.statista.com/statistics/1095077/hulu-average-revenue-per-subscriber-us/

Web Scraping:

- https://www.comparitech.com/tv-streaming/netflix-subscribers/
- https://www.theverge.com/2020/7/16/21326434/netflix-second-quarter-earnings-tv-shows-movies-originals-subscribers-adds-ted-sarandos

- STAT. “The Covid-19 Tracker.” STAT, 22 Sept. 2020, www.statnews.com/feature/coronavirus/covid-19-tracker.

Ethical Consideration:

1. Data Collection

The ethical consideration must be made for the Data Collection process is to inform users’ consents and make sure the data is unbiased. The sources that we will be using are from Statnews, New York Times articles, Statista analysis,etc., which consists of information about the monthly new Covid-19 cases globally and for the United States only, also the amount of subscribers of the streaming platforms increases after Covid-19 hits. Since we are not collecting the data directly from the users but comparing the overall results from the articles specifically on our chosen variables provided from the sources: Covid-19 monthly cases since March 2020,Netflix and Hulu subscribers as data points. We should also **check on the Data Collection bias of our source during the process of collecting data as part of our ethical considerations**. The data of Covid-19 cases is unbiased since we had found sources from statnews and it does not contain any kind of personal information, hence, we do not need to worry about users’ consents either. Also, we need to **ensure that our data would be appropriately usable for future reproduction and replication**.

2.Data Storage

For Data Storage, the ethical consideration we have to make is to **decide a plan for removing data if we no longer need it**. Our plan to protect our data to ensure data security is to limit the access to our dataset within our group members only. Since the dataset we use is originally provided from Statnews, New York Times articles, Statista analysis,etc., not directly from the users, so we do not have to worry about including sensitive information because the above listed source will not ask nor provide their users for it. Also, the above listed sources might have already checked for their own ethical considerations before publishing.

3.Data Analysis

For Data Analysis, the ethical consideration we must make is to **ensure that our analysis process (interpreting) and results stay transparent and unbiased**. For missing perspective and data set bias, we face the fact that we are analyzing our data of Covid-19 cases quarterly instead of monthly, the missing value for the gap, which is

in between quarter by quarter, might affect our data analysis. However, We do not have to worry about honest representation since our input data for visualizations, summary statistics, and reports is precisely the underlying data. Additionally, we need to keep in mind that our **analysis is reproducible and replicable for further analysis**. To ensure reproducibility, our data is not a big data, which resolves the problem of having big data and streaming pipelines which cause the failure of reproducibility, also keeping our documentation organized, and making sure our analysis is transparent should avoid failure of reproducibility as well. For replicability, since our data is “real” and should not have measurement error, we should be able to avoid the failure of replicability. But as the time goes, the analysis might give us completely different results which is uncontrollable since we could not predict the tendency of Covid-19 cases in the future, which might cause failure of replicability.

4.Modeling

For Modeling, the ethical consideration consists of Modeling **regards fairness across the data we are using for the visualizations**. Making unbiased model visualizations that are also easy to understand for the users is our top priority. Since our source for Covid-19 cases is provided by the Statnews and New York Times, its accuracy and fairness are ensured. We paid attention to the **displaying color to avoid messy visualization**. Hence, we used strong contrast colors such as red, blue and yellow and indicated the corresponding variables to avoid possible confusion on our visualizations.

5.Deployment

If we lost our data by any instance, we could re-do web scraping from different sources from the internet, since the information about Covid-19 cases is a great concern which people are keeping their eye on, so restoring the Covid-19 cases data should not be an issue. For the data of Netflix and Hulu subscriber counts after Covid-19 hit, there aren't many available sources online, but what we could do is to save a copy of the data to ensure there is a copy to back-up. To address unintended use, we could consider that streaming platforms subscribers might abuse the results by possible cases such as A whole family is sharing the same account, which our group might result in inaccurate data. Hence, **we can identify password sharing as one of the limitations of our project because it could potentially distort our analysis slightly. But we are not strictly focused on the amount of subscribers for Netflix and Hulu, we are focusing on the overall correlation of the increasing or decreasing tendency of streaming platforms subscribers after Covid-19 hit with**

respect to the number of Covid-19 cases, which will not affect our general analysis conclusion.

Analysis Proposal:

Method 1: Data Collection (web scraping, APIs, etc.)

The first step to analyzing data is to collect a sufficient amount of valid data to reach a valid conclusion with a reasonable analysis. Too few data can decrease the precision of our analysis and it might be more biased due to bigger effects from the outliers. The best way to minimize the sampling error bias is to collect more data. Other than the quantity of data, the quality of the data is important too. By having reliable quality data we avoid non-sampling errors. Hence, we collected sufficient data from reliable sources before moving on to data wrangling techniques to then prepare tidy data. The sources that we will be using are a mix of web scraping, and Data Sets, including financial reports:

- Statista. “Number of Hulu's Paying Subscribers in the U.S. 2019–2021, by Quarter.” *Statista*, 15 Nov. 2021, https://www.statista.com/statistics/258014/number-of-hulus-paying-subscribers/#:~:text=Number%20of%20Hulu's%20paying%20subscribers,U.S.%202019%2D2021%2C%20by%20quarter&text=In%20the%20fourth%20quarter%20of,of%20the%20previous%20fiscal%20year
- https://www.statista.com/statistics/1095077/hulu-average-revenue-per-subscriber-us/
- https://www.comparitech.com/tv-streaming/netflix-subscribers/
- https://www.theverge.com/2020/7/16/21326434/netflix-second-quarter-earnings-tv-shows-movies-originals-subscribers-adds-ted-sarandos
- STAT. “The Covid-19 Tracker.” STAT, 22 Sept. 2020, www.statnews.com/feature/coronavirus/covid-19-tracker.
- “Netflix - Financials - Financial Statements.” *Netflix*, 2021, https://ir.netflix.net/financials/financial-statements/default.aspx
- “The Walt Disney Company - Fiscal year 2020 Annual Financial Report.” Walt Disney Company, 2021 https://thewaltdisneycompany.com/app/uploads/2021/01/2020-Annual-Report.pdf

Method 2: Data Wrangling

Web scraping and Data wrangling:

- https://www.comparitech.com/tv-streaming/netflix-subscribers/
- https://www.theverge.com/2020/7/16/21326434/netflix-second-quarter-earnings-tv-shows-movies-originals-subscribers-adds-ted-sarandos
- STAT. “The Covid-19 Tracker.” STAT, 22 Sept. 2020, www.statnews.com/feature/coronavirus/covid-19-tracker
- Statista. “Number of Hulu's Paying Subscribers in the U.S. 2019–2021, by Quarter.” *Statista*, 15 Nov. 2021, https://www.statista.com/statistics/258014/number-of-hulus-paying-subscribers/#:~:text=Number%20of%20Hulu's%20paying%20subscribers,U.S.%202019%2D2021%2C%20by%20quarter&text=In%20the%20fourth%20quarter%20of,of%20the%20previous%20fiscal%20year
- https://www.statista.com/statistics/1095077/hulu-average-revenue-per-subscriber-us/

For the number of new COVID-19 cases and the time period, we scraped from the statnews website. For the number of Netflix Subscribers, we scraped from the Compaitech website and the number of Hulu subscribers we scraped from the Statista website. We gathered quarterly data from March 2019 to September 2021. We had encountered a difficulty that we could not find the data for globally new COVID-19 cases before December 2019. Also, we could not find monthly basis data. However, we found data quarterly, which might cause sampling bias due to not having enough (insufficient) data points, since a monthly data would include more data points. Also, the financial statements for Netflix and Hulu are organized on a quarterly basis, so we decided to collect all of our data quarterly.

Below is a sample of the data we collected:

22.8 (in millions) 22,800 (in thousands)

Time period	No. of new COVID-19 cases	No. of Netflix Subscribers	No. of Hulu Subscribers
03/2019	0	66.63	22.8
06/2019	0	66.50	25.2
09/2019	0	67.11	27.9
12/2019	0	67.66	28.5
03/2020	5,391	69.96	30.4

06/2020	26,627	72.90	32.1
09/2020	39,762	73.08	35.5
12/2020	205,087	73.93	36.6
03/2021	57,949	74.38	41.6
06/2021	12,796	73.95	42.8
09/2021	142,987	74.02	43.8

Method 3: Descriptive & Exploratory Data Analysis (summary stats, correlation, etc.)

The descriptive and exploratory analysis portion of our research paper will focus on exploring the quarterly rates of covid cases and no. of subscribers from each streaming platform, that is, no. of cases/month and no. of subscribers/month. If an increase in rate of number of covid cases corresponds to an increase in rate of number of subscribers, then it signifies a direct relationship (# of covid cases directly affect # of subscribers) between the number of covid cases and the number of subscribers of the streaming platforms and a correlation too. However, there can still be correlation if the both the rates for both number of subscribers and number of covid cases are merely positive for a particular time period (3 months). We also have a table that shows the maximum and minimum values for number of subscribers and number of cases.

	No. of COVID-19 cases	No. of Netflix Subscribers (in millions)	No. of Hulu Subscribers (in millions)
Maximum	205,087	74.384	43.8
Minimum	0	66.501	22.8

The maximum no. of covid cases is not in the same month as the maximum no. of Netflix subscribers and no. of Hulu subscribers, which suggests that there is no direct correlation. However, in the month, where there were 205,087 COVID-19 cases, the no. of Netflix subscribers was 73.9 million which is the 3rd highest relative to the maximum, and the no. of Hulu subscribers was 36.6 million which is the fourth highest relative to the maximum.

So, while there is no direct correlation, there still seems to be a relationship between the number of covid cases and the number of subscribers, albeit a slightly wayward one.

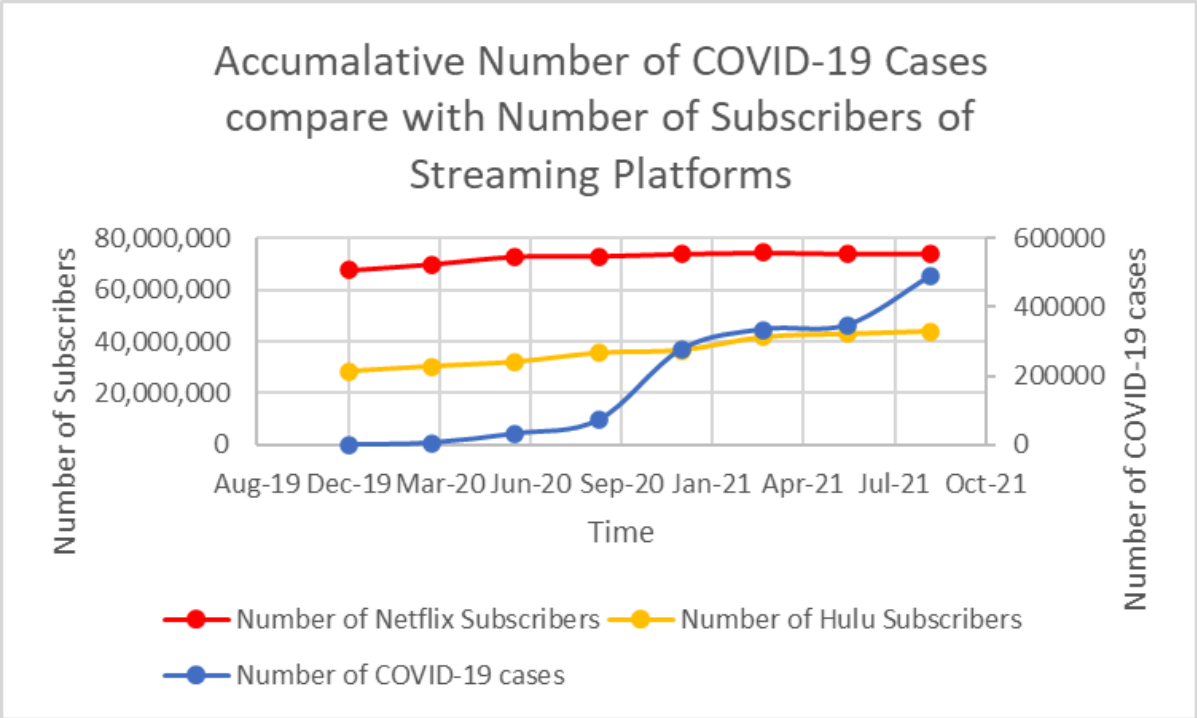
Quarterly rate:

Time period	No. of covid cases/month	No. of Netflix subscribers(in millions) /month	No of Hulu subscribers (in millions)/ month
03/2019-06/2019	0	+0.0333	+0.8
03/2019-06/2019	0	+0.1713	+0.9
06/2019-09/2019	0	+0.2046	+0.9
09/2019-12/2019	0	+0.1827	+0.2
12/2019-03/2020	+1597	+0.769	+0.6
03/2020-06/2020	+7078	+0.9783	+0.6
06/2020-09/2020	+4378	+0.0600	+1.1
09/2020-12/2020	+55108	+0.2850	+0.4
12/2020-03/2021	-49046	+0.1493	+1.7
03/2021-06/2021	-15051	-0.1443	+0.4
06/2021-09/2021	+43397	+0.0243	+0.3

The increase in the rates of subscribers for each of the streaming platforms once there is an increase in the number of covid-19 cases shows that there is a relationship between the two variables. The rates of the number of Netflix subscribers and Hulu subscribers sit at 0.1827 and 0.2 respectively during the “pre-pandemic” era since the number of covid cases are zero, and there is no change in covid cases from March, 2019 to December 2019. From December, 2019 to March, 2020, there is a positive rate in the number of covid cases, where the rates of number of subscribers increase drastically: Netflix’s rate increased from +0.1827 to +0.769, and Hulu’s rate increased from +0.2 to +0.6. The biggest rate in COVID cases may not correspond to the biggest rate in number of subscribers of the streaming platforms, but for the most part, when there’s a positive rate in no. of COVID cases, there’s a positive rate in no. of subscribers for the streaming platforms, which shows that there is a correlation. However, the correlation is not ever-present in our results. For instance from

03/2021-06/2021, even though the rate of COVID-19 has decreased, Hulu still experiences a positive rate of subscribers which contradicts the hypothesis since the hypothesis states that the increase in number of covid cases leads to an increase in number of subscribers. Based on the Quarterly Rate data table, the sudden increase in rate of COVID cases does seem to have an effect on the rate of the number of subscribers, but there seems to be a COVID rate threshold, that once met, breaks away from the correlation. This is reasonable because once the COVID rate threshold has been reached, the lockdown policies have taken hold, which is what prompts people from watching streaming platforms.

Method 4: Data Visualization

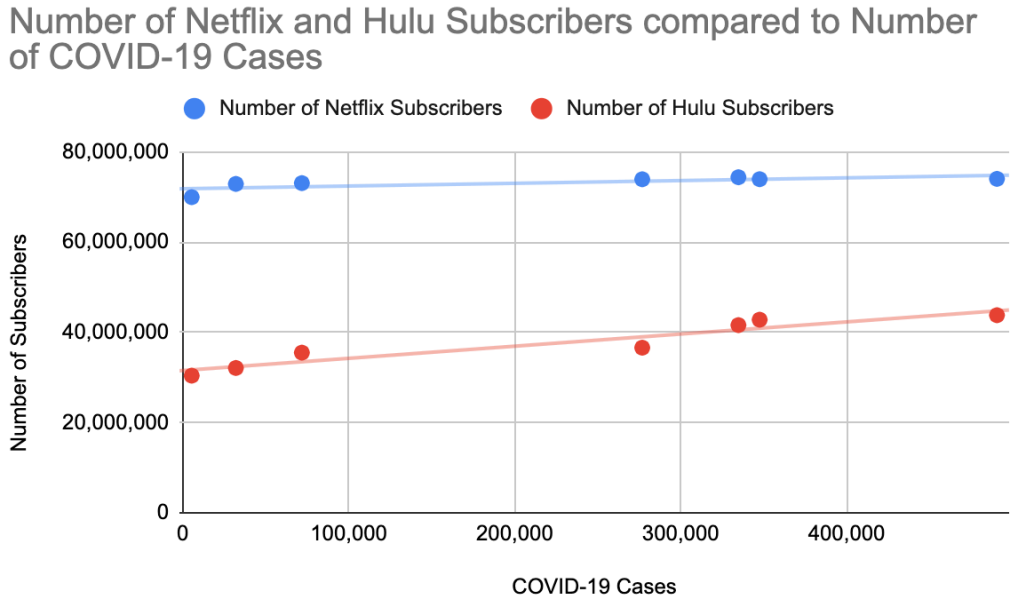


Our group decided that the best way to visualize our data is by creating a scatter plot. We learned from the lecture that the best way to visualize a data set with two or three numerical variables is by using a scatter plot, and as shown in our Data Table we are measuring two different online streaming platforms which are Hulu and Netflix and comparing those online streaming platforms to the number of Covid-19 cases for each given month of each represented year. After establishing the scatter plot we created a line for each variable that connected each value for that single variable that was measured in the scatter plot. We chose to do this because the line will clearly show either an increase or decrease of the slope for each variable that we measured. We also decided to use three different types of colors to represent each variable being measured, because as we learned from the lecture that the types of color combination

we choose are extremely important when making our visualization easier to interpret and analyze to the audience.

Method 5: Predictive Analysis

For our predictive analysis, we chose a multiple linear regression model to show whether there is a correlation between the number of Covid-19 cases and the number of subscribers of Netflix and Hulu. This correlation will allow us to predict the relationship between the two variables.



1

As shown by the data points, the line of best fit for both streaming platforms has a positive linear slope. Based on these linear slopes, the increase in the number of Covid cases also leads to a gradual increase in the number of subscribers of both Hulu and Netflix. We wanted to later on in the future include some of the Python libraries in order to create a more cultivated model that would better express our data.

Discussion:

After carefully collecting our Data from trusted sources, then cleaning our data in order to display it in a Tidy Data set form, then creating a visualization that followed the rules and standards of “what makes a good visualization”. As a result, we were then able to develop a multiple linear regression model that depicts the strength of the correlation and what the future relationship looks like between the number of Covid-19 cases and the number of subscribers for Netflix and the number of subscribers for Hulu.

From our analysis we observed that as the number of Covid-19 cases increased so did the number of Netflix subscribers and Hulu subscribers however, Netflix had the largest number of increase in subscribers when compared to the number of Covid-19 cases while Hulu had a very minimal increase in subscriber count within comparison to the number of of Covid-19 cases. Based on this our group can infer that there is a strong correlation with the number of Covid-19 cases and the number of Netflix subscribers. However we cannot infer that there is a correlation with the number of subscribers for Hulu and the number of Covid-19 cases because as the number of Covid-19 cases fluctuated the number of Hulu subscribers was growing steadily no matter how low or high the number of Covid cases there are. In conclusion our hypothesis was indeed supported by our analysis however, we don't believe that our analysis proposal is conclusive because the scope of our data is limited due to the fact that we didn’t have multiple sources that represented the amount of “full paid membership subscribers” for Hulu, nor were we able to find sufficient data for other streaming platforms other than just Netflix and Hulu. More specifically the subscriber count for Hulu didn’t factor in the amount of subscribers that have actually paid for a membership and were not on the free trial mode. Another issue we encountered is that although we were able to access the official financial reports for Netflix, for Hulu on the other hand we weren't able to find any sort of financial statements and we merely had to rely on other sources that discuss the financial reports for Hulu, consequently there is a possibility that their might be a slightly inaccurate subscriber count for Hulu. One way to improve our project is to incorporate the specific policies that may have been put in place when the Covid-19 cases were rising. It would allow us to make a clearer logical connection between the number of Covid-19 cases and the number of subscribers for the streaming platforms.