The purpose of this data visualization project is to analyze data gathered from Netflix in order to find out about Netflix's business practices as well as users' preferences. The data that was utilized to create the data visualizations involved a country's country code, GDP and Income Disparity index, the top 10 movies and TV shows Netflix users watched, Netflix subscription prices, Netflix library size, Netflix revenues, Netflix subscriptions, and the IMDB movie data. These data were gathered from many different sources, including Kaggle, IMDB, World Bank, Harvard Database, Netflix, and Comparitech. The audience of this data visualization is for those who are interested in the workings of a big company like Netflix, as well as those looking to learn something from the business practices of one of the most successful companies in the world.

There are 4 different questions that these data visualizations strive to answer. Firstly, "What changes do Netflix's business model encounter when entering a poorer country's market?". To answer this question, a scatter plot that compares GDP to many variables was created, with the conclusion that, surprisingly, there actually seems to be very little correlation between Netflix's business model and how rich a country is. This brings up the question, does Netflix only care about the richer customers in a country, ignoring the poor? Furthermore, are the majority of the films and TV shows that Netflix acquires internationally available, considering how the total library size of Netflix in each country doesn't seem to show any correlation to how much money Netflix obtains in any given country. This data visualization, while insightful in showing how there was no correlation, was visually unappealing and could've been more colorful and eye-catching.

The second question then naturally becomes, "Is Netflix really only targeting the rich customers?". The data visualization that answers this question is the stacked bar graph that compares the differences in the three prices (Basic, Standard, Premium) as the income disparity increases. And this time, the graph does show some correlation between the decreasing income disparity and the increasing prices. This disproves the hypothesis of Netflix targeting rich customers and seems to suggest that Netflix does, in fact, wish to attract a wide range of customers. While a correlation could be seen on the graph, the graph's scale made it such that the bars were very thin, making it hard to clearly see a trend as there were often large gaps between bars. Next time, it may be better to do some form of a histogram, or perhaps a triple line graph.

The third question was "What are the popular genres of entertainment in different countries and the world?". By utilizing a tree graph, the graph easily showed that Drama, Comedy, and Action were the top 3 genres that people enjoyed watching, adding up to over 1/3rd of the entire graph. Interestingly, Musicals were in last place, with Brazil making up a significant amount of the percentage that enjoyed watching Musicals. The tree graph easily showed the distribution of the genre's popularity and worked very well. An unfortunate thing that could be improved upon is that due to the scale being so large, a genre, "News" was unable to be seen at all as it was only one exactly 1 country's top 10 for 1 week. Additionally, because the dataset was so large, it lagged a bit.

The fourth question was "Where does Netflix get their money from?". By graphing it onto a Choropleth Map, the clear winner was the North American Continent, with the U.S. contributing to an overwhelming amount, with Canada contributing quite decently as well. Interestingly, Netflix was seen to be quite prominent in Brazil, while Europe was contributing a significant amount as expected. Additionally, the subscription numbers and revenue were graphed to one another to clarify that there was indeed a positive correlation between the two. The Choropleth Map was very insightful in showing which regions contributed how much, but because there were so many countries that didn't have data regarding Netflix revenue and subscription, it was hard to fully grasp the extent of what this visualization could've shown. Next time, it may be best to perhaps use a bubble chart or something similar to show the regions.

In the process of making these visualizations, the package "plotly" was utilized. Most of the creation process was straightforward thanks to plotly's easy documentation and usability; however, the treemap/sunburst chart's data frame requirement was very tedious to get right. What was required was the count, the label, the unique id, and the parent nodes it should branch out of. Furthermore, aggregates of the parent nodes were required. While the count and the label were already accounted for, the parent node and the unique id had to be created from scratch, involving a variety of string concatenations that required a very precise formatting which took a while to perfect. The aggregation process involved even more processing due to the difference in its id and parent node requirements. To add to that, the worldwide treemap and country-specific treemaps had different parents, an entirely new dataset needed to be created. There were also the occasional changes in which numerical columns became character columns that needed to be guarded against.

To create the dataset for the scatter plot, the Netflix subscription fee Dec-2021.csv file was used, combining only the 'V64' (2019 GDP) of the dataset API_NY.GDP.MKTP.CD_DS2_en_csv_v2_3628616.csv and the revenue and subscription values of column 23, 24 of the dataset Netflix Subscribers by Country - Update July 2021 - Subsciber Figures by Country.csv. Those two columns had to be converted into numerical code from character mode, removing the "," in order to do so. For the stacked bar graph, the GIDI of the countries were taken from swiid9_2_summary.csv and appended as a column to the Netflix subscription fee Dec-2021.csv. Then, the differences between the basic and standard and standard and premium were taken and placed into columns. For the Choropleth Map, the country codes from wikipedia-iso-country-codes.csv (due to the fact the official iso websites requires you to pay for the dataset) were appended into the Netflix subscription fee Dec-2021.csv. All this was then exported to a file called Netflix subscription fee Dec-2021 with GDP and Income Dis 2019.csv.

To deal with the TreeMap dataset, the IMDB file, data.tsv, was read with the package readr, and the Netflix file, all-weeks-countries.xlsx, was read with the readxl package. The two sets were combined, after which the IMDB's false duplicate entries were cleaned by finding only the "Films" and "movies" or "TV" and "tvSeries" category and titletype pairing respectively. Finally, a distinct was called one more time to completely combine the duplicate entries. Then, all other information besides genre and country name was thrown out, after which the dataset was pivoted to become longer, separating the genres so that multiple genres couldn't be in one column. Finally, the count function was called, counting up the genres according to the country. To format this dataset for the treemap graph, the "-" was removed and replaced with a space to prevent formatting errors. Then, a parent column was created with the format "root - parent". Then, an id column was created by adding on the label (Country) to the parent column. After this, a count was aggregated to find the total number of genres. For the treemap of the countries, the previously made dataset was changed such that the parent column became the label and the label column became the parent column. Then, the same format was required to be created, just flipped, which was done. Finally, this was exported as No Center Sunburst.csv and Country Tree Map.csv respectively. Then, all the dataset was graphed accordingly with plotly's functions. Finally, the R Shiny UI was created to store the plots, creating a navbar and a tileset for each of the 4 questions and the associated plots.

The scatter plot is a table that tries to use all data to try and find trends and outliers. The scatter plot encoded the data by arranging them to order, then manipulated the plot by changing the y-axis, and reduced the information visual noise by embedding the values and country so that it would appear only after a mouse hover.

The stacked bar graph is a table that tries to use all data to try and find trends and outliers, as well as try and find attributes of one distribution. The stacked bar graph encoded the data by arranging them in alignment, with mappings of color hue as well to show differences in the three price categories. The encoding of mapping with sizes was there to show differences in prices as well. Essentially, the graph was faceted to superimpose the three price categories on top of each other. Additionally, the data was reduced by embedding the country, price, and GINI index so that the information would appear after a mouse hover.

The treemap is a network data of trees that attempt to find the attributes of one distribution as well as the extremes. It encodes the data by mapping it with different sizes according to the proportion as well as color to differentiate the different genres. It manipulates data by changing when the genre is clicked upon, showing a focused version of that genre distribution. It also reduces data by filtering when a genre is clicked on, focusing on the distribution of that specific genre, as well as aggregating data by showing the total amount of people that like a specific genre. It embeds as well, showing the precise number when hovered over.

The Choropleth Map is a geometry(spatial) data that tries to find the attributes of one distribution of where Netflix earns its money at. It encodes the data by arranging them and using it on a world map, mapping by coloring in saturation to show how much that country contributes. The data can be manipulated by changing between revenue and subscription. Furthermore, it reduces the data by embedding the precise number and the country name, showing the precise number when hovered over.

Ultimately, while the data visualizations aren’t perfect, they answer the 4 questions imposed upon them well.