How GDP and Social Inequality Affect Netflix's Growth? (Complete Report)

Caio di Felice Cunha

Analyzing the temperature in R

The main objective is to present graphics in R. Therefore, I will not spend too much time in cleaning and organizing the data (that is why I choose a simple dataset).

Thesis: For that I will compare the number of subscriptions in Netflix with the country's wage, to prove that as the salary condition increases, the number of subscribers also increases.

Stage 1 - Collecting the Data

Links for the data: IMDB data: https://datasets.imdbws.com/ Netflix Data: https://www.comparitech.com/blog/vpn-privacy/countries-netflix-cost/ GDP: https://data.worldbank.org/indicator/ Wage Inequality Dataset: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LM4OWF TOP 10 Netflix shows: https://top10.netflix.com/ Netflix subscribers and revenue by country: https://www.comparitech.com/tv-streaming/netflix-subscribers/ ISO Country Codes - Global: https://www.kaggle.com/datasets/andradaolteanu/iso-country-codes-global

```
# Necessary packages
options(warn = -1)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(readxl)
library(readr)
library(ggplot2)
```

```r
# Importing the data
netflix_data_dec_2021 <-read.csv("datasets\\dados_netflix_Dec_2021.csv")

netflix_subscriptions_jul_2021 <-read.csv("datasets\\as_net_2021.csv")

wage_inequality_data <-read.csv("datasets\\dados_desig_soc_harvard.csv")

world_bank_data <-read.csv("datasets\\dados_world_bank.csv", header = FALSE)

top_10_shows_netflix <-read_excel("C:\\Users\\Caio\\OneDrive\\Desktop PC\\Desktop\\Portfolio\\GDP X NETF

wikipedia_iso_country_codes <- read.csv("datasets\\iso-country-codes.csv")

IMDB_Data <- read_tsv("datasets\\data.tsv")
```

```
## Rows: 9512109 Columns: 9

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (8): tconst, titleType, primaryTitle, originalTitle, startYear, endYear,...
## dbl (1): isAdult
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Stage 2 - Mapping & Cleaning

Analyzing each dataset I have used this commands bellow to analyze the quality of the data and to understand the datasets

dim(x) colnames(x) rownames(x) head(x) str(x) summary(x)

```r
### Cleaning and Preparing the First Combined Dataset ###

# Create a column with the difference data for the bar chart
# (standard plan - basic plan)
netflix_data_dec_2021$basic_standard_diff <- (
  netflix_data_dec_2021$Cost.Per.Month...Standard....
  - netflix_data_dec_2021$Cost.Per.Month...Basic....)

# Create a column with the data difference for the bar chart
# (premium plan - standard plan)
netflix_data_dec_2021$premium_standard_diff <- (
  netflix_data_dec_2021$Cost.Per.Month...Premium....
  - netflix_data_dec_2021$Cost.Per.Month...Standard....)


# Combine previous data with GDP data
names(world_bank_data)[
  names(world_bank_data) == 'V1'] <- 'Country'

netflix_data_GDP <- merge(
```

```
  netflix_data_dec_2021,
  world_bank_data,
  by = "Country")

# Extracts the 2020 GDP
netflix_data_GDP2020 <- netflix_data_GDP[-c(11:72, 74, 75)]

names(netflix_data_GDP2020)[names(netflix_data_GDP2020) == 'V64'] <-
  "2020 GDP (World Bank)"

# Cleanup of wage inequality dataframe
wage_inequality_data <- wage_inequality_data[, c(1:3)]

wage_inequality_data_year <-
  wage_inequality_data %>%
  group_by(country) %>%
  summarise(max = max(year, na.rm = TRUE))

# Combine the dataframes
wage_inequality_data <- merge(
  wage_inequality_data,
  wage_inequality_data_year,
  by.x = c("country", "year"),
  by.y = c("country", "max"))

netflix_data_GDP_wage2020 <- merge(
  netflix_data_GDP2020,
  wage_inequality_data,
  by.x=c("Country"),
  by.y=c("country"))

# Clear the billing and subscription dataset
# and combine it with the previous dataframe
netflix_subscriptions <- netflix_subscriptions_jul_2021[,c(1, 23,24)]

complete <- merge(
  netflix_data_GDP_wage2020,
  netflix_subscriptions,
  by=c("Country"))

# Merge the countrycode into the choropleth map
countrycode <- wikipedia_iso_country_codes[,c(1, 3)]

complete <- merge(
  complete,
  countrycode,
  by.x=c("Country"),
  by.y=c("English.short.name.lower.case"))


# Save the dataframe produced so far
#write.csv(complete, "clean_datasets\\complete.csv", row.names = FALSE)
```

Stage 3 - Cleaning and Preparing the Second Combined Dataset

```
# Clear and filter the IMDB dataframe
genre <- IMDB_Data[,-c(1, 4:8)]
names(genre)[names(genre) == 'primaryTitle'] <- 'show_title'

# Associate the genre with the Top 10 shows
topgenre <- merge(
  top_10_shows_netflix,
  genre,
  by = "show_title")
# View(topgenre)

# Clear the previous dataframe to keep only 1 entry for each top 10
topgenre <- topgenre[
  (topgenre$category == "Films" & topgenre$titleType == "movie") |
    (topgenre$category == "TV" & topgenre$titleType == "tvSeries"), ]

topgenre <- distinct(
  topgenre,
  show_title,
  week,
  country_name,
  category,
  titleType,
  cumulative_weeks_in_top_10,
  .keep_all= TRUE)
# View(topgenre)

# Keep only movie genre information by country
topgenrecountry <- topgenre[,-c(1, 3:9)]
#View(topgenrecountry)

# Dataframe pivot
topgenrecountry <- separate(
  topgenrecountry,
  c("genres") ,
  c("genre1", "genre2", "genre3"),
  sep = ",")

topgenrecountry <- pivot_longer(
  topgenrecountry,
  c("genre1", "genre2", "genre3"),
  names_to = "genre123",
  values_to = "genres")
# View(topgenrecountry)

# Count the number of genres
genrecount <- count(topgenrecountry, country_name, genres)
genrecount <- na.omit(genrecount)
genrecount <-subset(genrecount, genres!="\\N")
genrecount$n <- as.numeric(genrecount$n)
```

```
# Save to disk
#write.csv(genrecount, "clean_datasets/genrecount.csv", row.names = FALSE)
```

Stage 4 - Cleaning and Preparing the Third Combined Dataset

```
# Rename the previous dataframe
sunburst <- rename(genrecount, label = country_name)

# Remove the dashes
sunburst$genres = sub("-", " ", sunburst$genres)

# Set the name
sunburst$parent = c("total  - ")
sunburst$parent <- paste(sunburst$parent, sunburst$genres)
sunburst$id = c(" - ")
sunburst$id <- paste(sunburst$parent, sunburst$id)
sunburst$id <- paste(sunburst$id, sunburst$label)
sunburst$n <- as.numeric(sunburst$n)
#View(sunburst)

# Aggregate
added <- aggregate(sunburst$n, list(sunburst$genres), FUN=sum)
added <- rename(added, label = Group.1)
added <- rename(added, n = x)
added$n <- as.numeric(added$n)
added$genres <- c(NA)
added$parent <- c("total")
added$id <- c(" - ")
added$id <- paste(added$parent, added$id)
added$id <- paste(added$id, added$label)
#View(added)

# calculate sum
total = sum(added$n)


# Combine everything into the final dataframe
sunburst <- rbind(added, sunburst)
sunburst <- rbind(c("total", total, NA, NA, "total"), sunburst)
sunburst <- sunburst[,-c(3)]
sunburst$n <- as.numeric(sunburst$n)
#View(sunburst)

# Save to disk
#write.csv(sunburst, "clean_datasets/sunburst.csv", row.names = FALSE)
```

Stage 5 - Cleaning and Preparing the Fourth Combined Dataset

```
## Macro View
# Let's work with top 10 to avoid performance issues in graphics
top10sunburst <- sunburst[-c(1:28),]
top10sunburst$n <- as.numeric(top10sunburst$n)
#View(top10sunburst)

# Top 10 genres by country
top10sunburst <- top10sunburst %>%
  group_by(label) %>%
  top_n(10,n)
#View(top10sunburst)

# Recalculate the totals, adjust and match the dataframe
top10add <- aggregate(
  top10sunburst$n,
  list(top10sunburst$parent),
  FUN = sum)
top10add <- rename(top10add, id = Group.1)
top10add <- rename(top10add, n = x)

top10add$label = sub("total  -  ", "", top10add$id)
top10add$parent = c("total")
top10add$n <- as.numeric(top10add$n)

total = sum(top10add$n)
top10sunburst <- rbind(top10add, top10sunburst)
top10sunburst <- rbind(c("total", total, NA, NA, "total"), top10sunburst)
top10sunburst$n <- as.numeric(top10sunburst$n)
#View(top10sunburst)

# Save to disk
#write.csv(top10sunburst, "clean_datasets/top10sunburst.csv", row.names = FALSE)
```

Stage 6 - Cleaning and Preparing the Fifth Combined Dataset

```
# Filter the previous dataframe and create a new one
in_total <- sunburst[-c(1),]
in_total$parent = sub("total  -  ", "", in_total$parent)
in_total$parent = sub("total", NA, in_total$parent)
in_total$id = sub("total  -  ", "", in_total$id)
#View(in_total)

# Salva em disco
#write.csv(in_total, "clean_datasets/in_total.csv", row.names = FALSE)
```

Stage 7 - Sixth Combined Dataset Cleanup and Preparation

```
# Filter the previous dataframe and create a new one
countrytree <- in_total[-c(1:28),]
```

```r
countrytree <- rename(countrytree, parents = label)
countrytree <- rename(countrytree, labels = parent)
countrytree$id = c(" - ")
countrytree$id <- paste(countrytree$parent, countrytree$id)
countrytree$id <- paste(countrytree$id, countrytree$label)
countries <- aggregate(countrytree$n, list(countrytree$parents), FUN = sum)
countries <- rename(countries, labels = Group.1)
countries <- rename(countries, n = x)
countries$n <- as.numeric(countries$n)
countries$id <- countries$label
countries$parents <- c(NA)
countrytree <- rbind(countrytree, countries)
#View(countrytree)

# Save to disk
#write.csv(countrytree, "clean_datasets/countrytree.csv", row.names = FALSE)
```

Stage 8 - Adjusting Data Type and excluding Outliers

```r
# Load the first clean dataset
complete <- read.csv("clean_datasets/complete.csv")

# Set the data type of some columns
complete$X..of.Subscribers.Q4.2021..Estimate. <-
  as.numeric(
    gsub(
      ",",
      "",
      complete$X..of.Subscribers.Q4.2021..Estimate.))

complete$Q4.2021.Revenue....Estimate. <-
  as.numeric(
    gsub(
      ",",
      "",
      complete$Q4.2021.Revenue....Estimate.))

# Create dataframes by filtering outliers
complete_scat_out <-
  filter(
    complete,
    Country != "United States")
#Because the volume of signatures and views are overwhelmingly high

complete_bar <-
  filter(
    complete,
    Country != "Switzerland") #Because social inequality is very low

complete_bar_out <-
  filter(complete_bar,
         Country != "South Africa") #Because social inequality is very high
```

Stage 9 - NA Values

```
# Load datasets 2, 3 and 6
genre <- read.csv("clean_datasets/genrecount.csv")
tree <- read.csv("clean_datasets/sunburst.csv")
countries <- read.csv("clean_datasets/countrytree.csv")

# Filter the list of countries by removing NA values
country_list <- filter(countries, is.na(parents))
```

Stage 9.1 - Loading and Renaming

```
# Load the first clean dataset
complete <- read.csv("clean_datasets/complete.csv")

# Set the data type of some columns
complete$X..of.Subscribers.Q4.2021..Estimate. <-
  as.numeric(
    gsub(
      ",",
      "",
      complete$X..of.Subscribers.Q4.2021..Estimate.))

complete$Q4.2021.Revenue....Estimate. <-
  as.numeric(
    gsub(
      ",",
      "",
      complete$Q4.2021.Revenue....Estimate.))

## Chanhging the columns names
names(complete)[
  names(complete) == 'Q4.2021.Revenue....Estimate.'] <-
  'Netflix.Q42021.Revenue'

names(complete)[
  names(complete) == 'X..of.Subscribers.Q4.2021..Estimate.'] <-
  'Netflix.Subscriptions.Q42021'

names(complete)[
  names(complete) == 'Total.Library.Size'] <-
  'Total.Catalog.Size'

names(complete)[
  names(complete) == 'Cost.Per.Month...Basic....'] <-
  'Basic.Subscription.Price'

names(complete)[
```

```
  names(complete) == 'Cost.Per.Month...Standard....'] <-
  'Standard.Subscription.Price'

names(complete)[
  names(complete) == 'Cost.Per.Month...Premium....'] <-
  'Premium.Subscription.Price'

# Create dataframes by filtering outliers
complete_scat_out <-
  filter(
    complete,
    Country != "United States")
#Because the volume of signatures and views are overwhelmingly high

complete_bar <-
  filter(
    complete,
    Country != "Switzerland") #Because social inequality is very low

complete_bar_out <-
  filter(complete_bar,
         Country != "South Africa") #Because social inequality is very high

# Load datasets 2, 3 and 6
genre <- read.csv("clean_datasets/genrecount.csv")
tree <- read.csv("clean_datasets/sunburst.csv")
countries <- read.csv("clean_datasets/countrytree.csv")

# Filter the list of countries by removing NA values
country_list <- filter(countries, is.na(parents))
```

Stage 10 - Visualization

We are going to analyze this mainly variables:

Netflix.Q42021.Revenue Netflix.Subscriptions.Q42021, Total.Catalog.Size, Basic.Subscription.Price, Standard.Subscription.Price, Premium.Subscription.Price

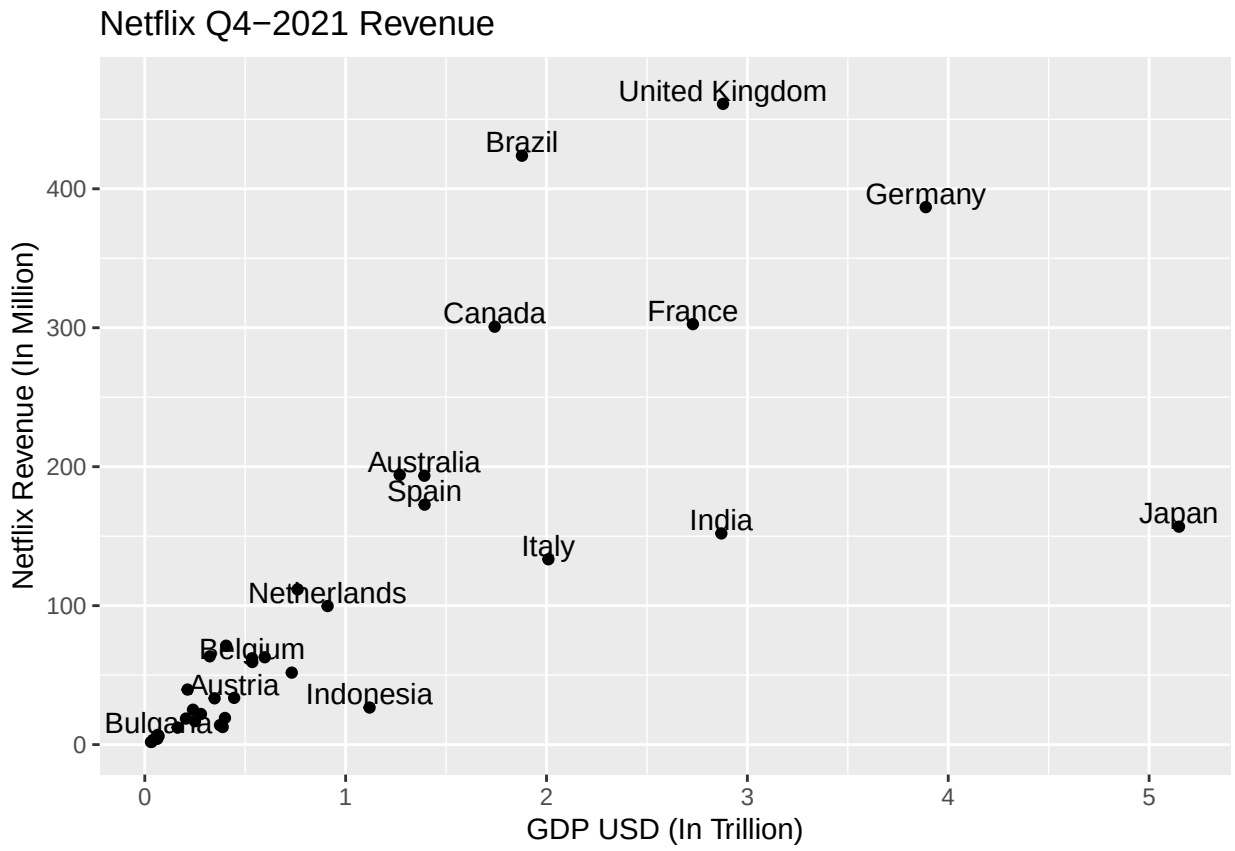
```
# Scatter Plot

## We are going to analyze this mainly variables:
## Netflix.Q42021.Revenue, Netflix.Subscriptions.Q42021,
## Total.Catalog.Size, Basic.Subscription.Price,
## Standard.Subscription.Price, Premium.Subscription.Price

## without outliers

## Netflix.Q42021.Revenue
ggplot(complete_scat_out,
       aes(x=X2020.GDP..World.Bank. / 1000000000000, #trillion
           y=Netflix.Q42021.Revenue / 1000000)) + #Million
  geom_point() + # Show dots
  geom_text(
```

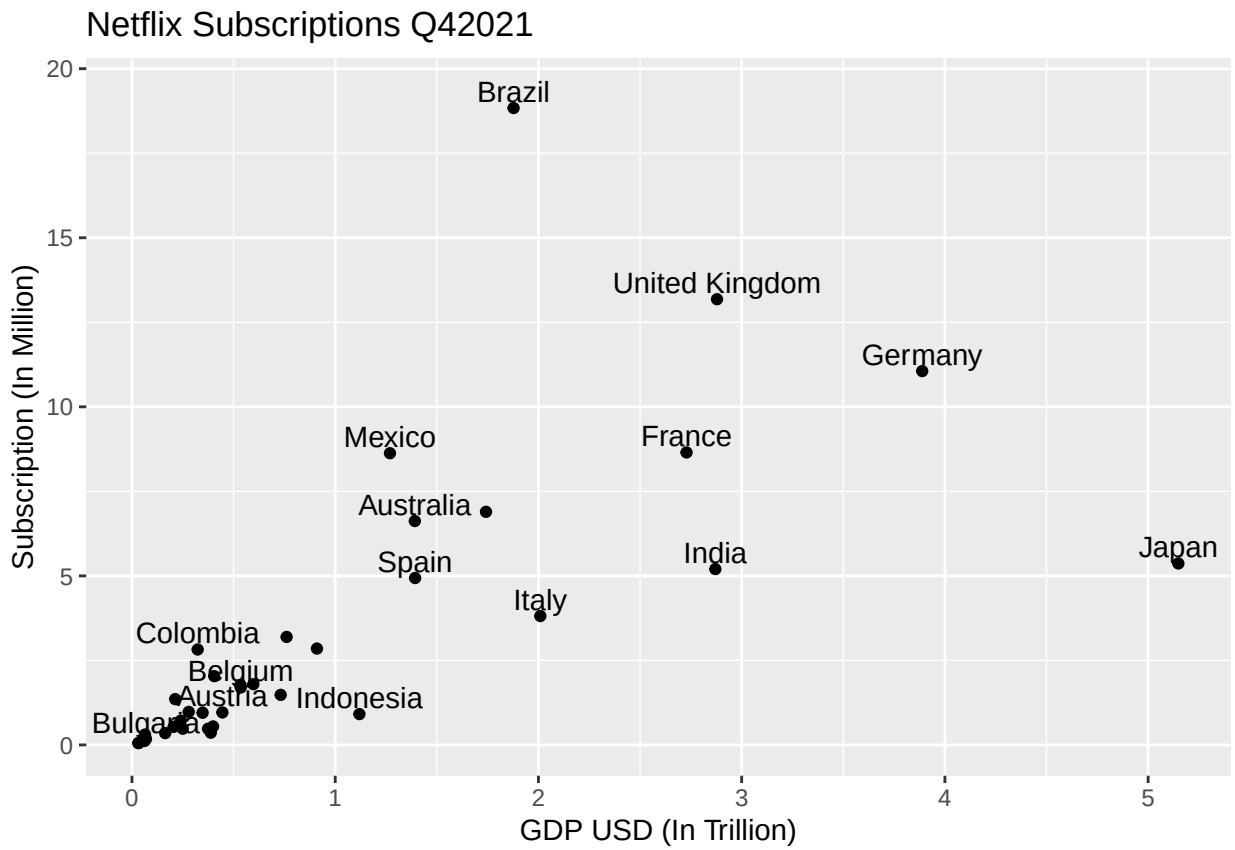
```
    label=complete_scat_out$Country,
    nudge_x = 0, nudge_y = 10,
    check_overlap = T
  ) +
  labs(y= "Netflix Revenue (In Million)", x = "GDP USD (In Trillion)") +
  ggtitle("Netflix Q4-2021 Revenue")
```



Analysis: In the chart above, which shows Netflix's revenue in relation to the country's GDP, we see that until Canada there is a clear positive relationship between Revenue and GDP, that is, the higher the GDP, the higher the Revenue. However, after that, we have two countries that have a big difference between them. The first is Brazil, which seems to follow the standardization of the previous countries, but the second country, Italy, deviates from this pattern, because, despite having a high GDP, Netflix Revenue is lower than countries like Spain, which It has a much smaller GDP. From Italy, countries begin to dissipate, implying that GDP influences Netflix's revenue, but to a certain extent. Perhaps, the number of shows characteristic of that region will help in a bigger Revenue. For example, Italian Shows to increase Revenue in Italy, Indian Shows to increase Revenue in India.

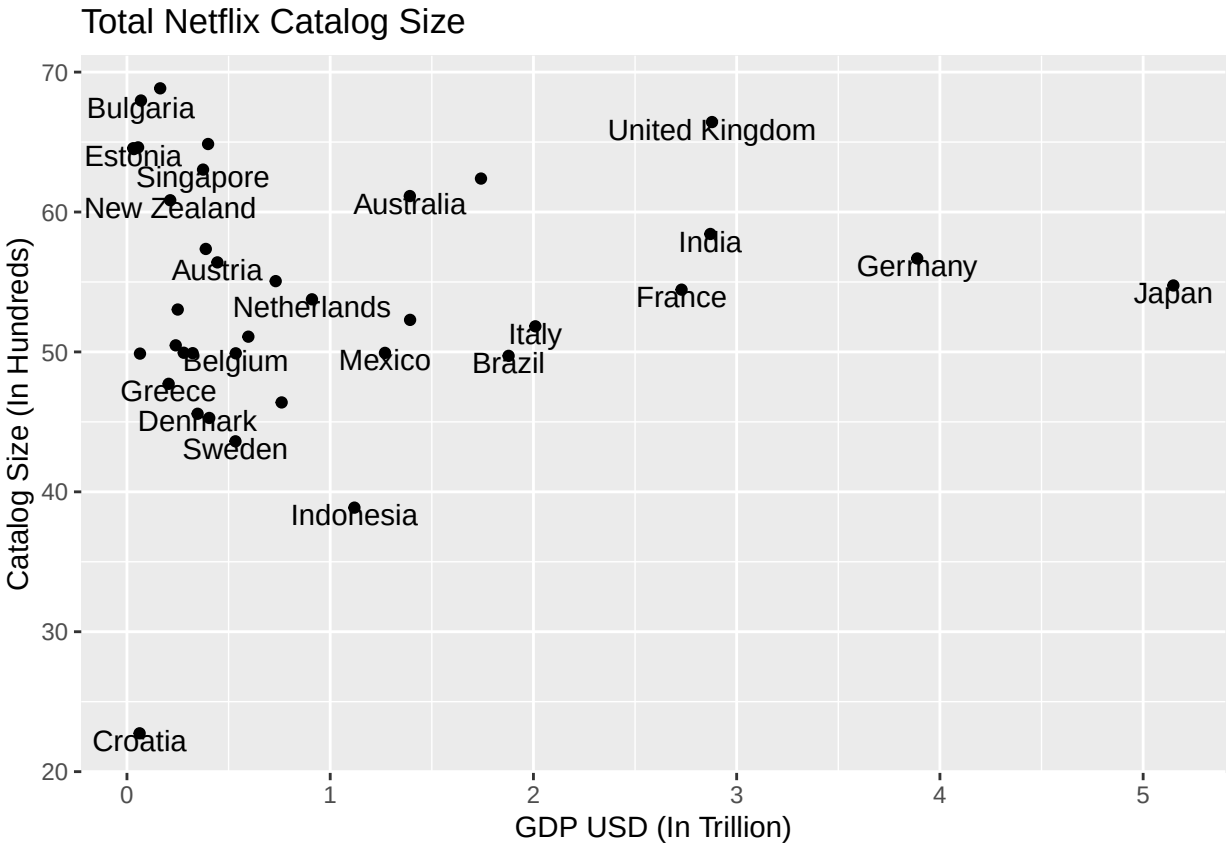
```
## Netflix.Subscriptions.Q42021
ggplot(complete_scat_out,
       aes(x=X2020.GDP..World.Bank. / 1000000000000, #trillion
           y=Netflix.Subscriptions.Q42021 / 1000000)) + #Million
  geom_point() + # Show dots
  geom_text(
    label=complete_scat_out$Country,
    nudge_x = 0, nudge_y = 0.5,
    check_overlap = T
```

```
  ) +
  labs(y= "Subscription (In Million)", x = "GDP USD (In Trillion)") +
  ggtitle("Netflix Subscriptions Q42021")
```



Analysis: Seeing the graph of subscribers, we noticed that it follows the same reasoning as the previous graph, with a few changes. When combining the two views, it is clear that the number of registrations does not necessarily mean a higher Revenue. This is very evident when we look at the UK, which provides almost 500 million Revenue, while having 13 million subscribers, while Brazil needs 19 million subscribers to give a return of 420 million. One suggestion would be to invest in new shows according to the country's growth potential in relation to new subscriptions. For example, would it be better to invest in Italy or Japan? One answer could be: It depends on which country has fewer subscribers as the reach of new subscribers is greater

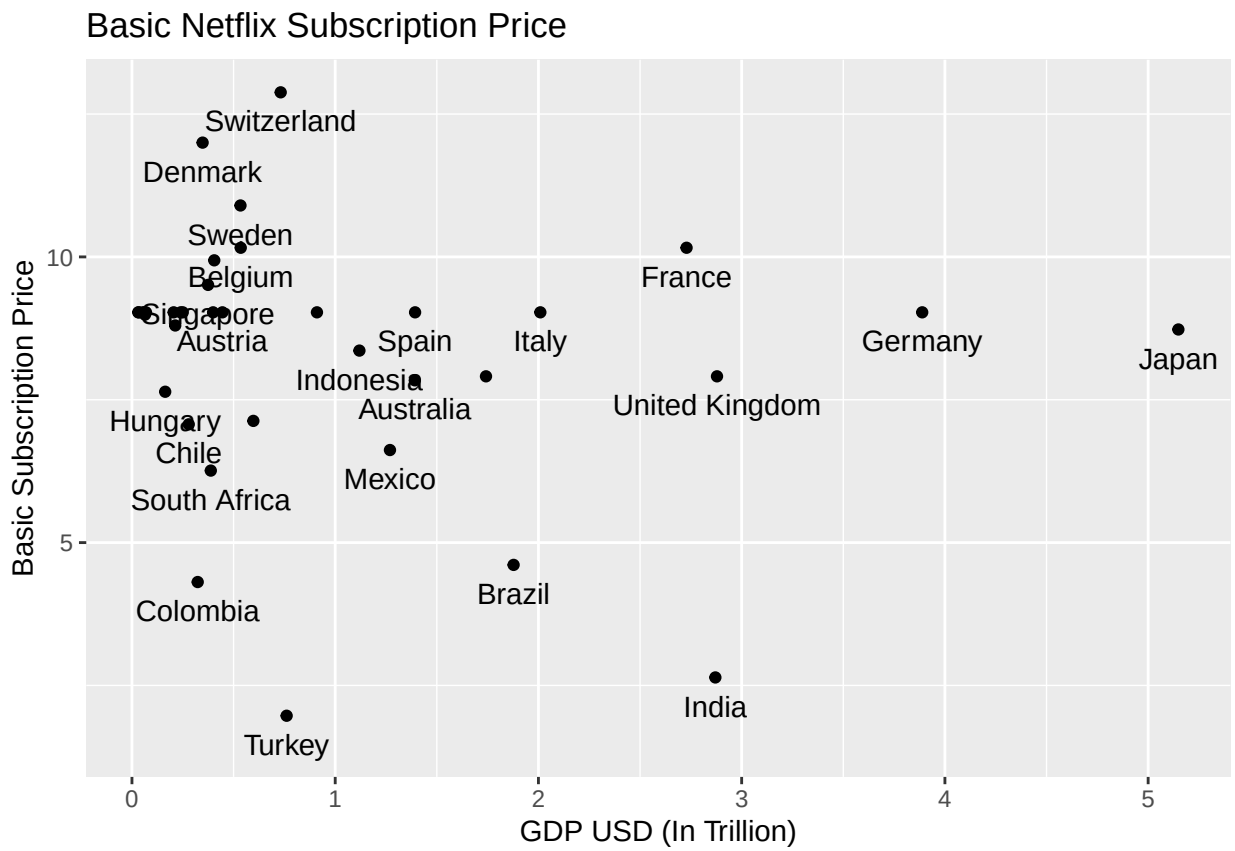
```
## Total.Catalog.Size
## Total.Catalog.Size
ggplot(complete_scat_out,
       aes(x=X2020.GDP..World.Bank. / 1000000000000, #trillion
           y=Total.Catalog.Size / 100)) + #Million
  geom_point() + # Show dots
  geom_text(
    label=complete_scat_out$Country,
    nudge_x = 0, nudge_y = - 0.5,
    check_overlap = T
  ) +
  labs(y= "Catalog Size (In Hundreds)", x = "GDP USD (In Trillion)") +
  ggtitle("Total Netflix Catalog Size")
```



Analysis: Here we see something that helps us when interpreted with the other graphs. Despite the very low return that some countries give, such as Hungary, Bulgaria, Lithuania, among others, these are the countries that have a larger catalog. Perhaps, it would be better to invest in countries that already give a greater financial return, but that do not have so many subscribers or such a vast catalog, as is the case of Spain, Mexico, Brazil, Italy and Japan

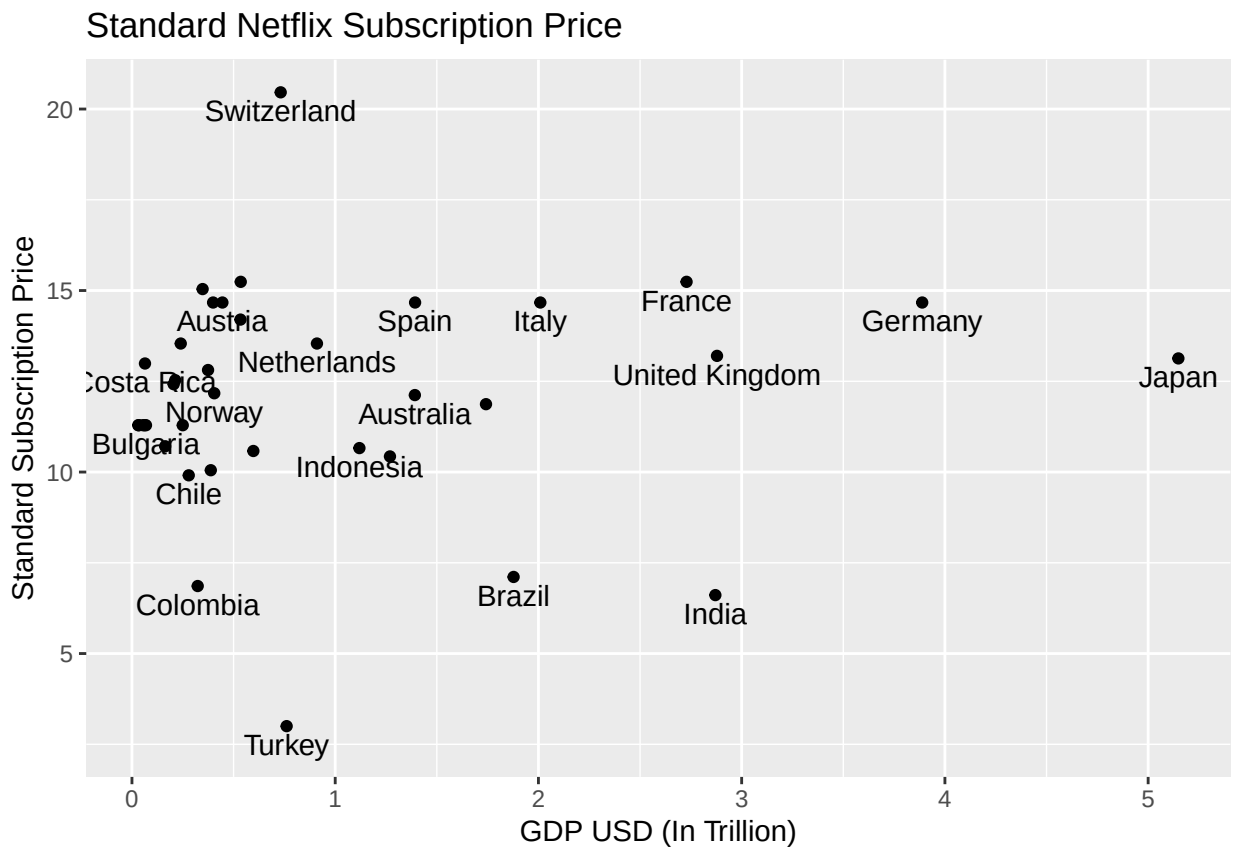
```
## Basic.Subscription.Price

ggplot(complete_scat_out,
       aes(x=X2020.GDP..World.Bank. / 1000000000000, #trillion
           y=Basic.Subscription.Price)) + #Million
  geom_point() + # Show dots
  geom_text(
    label=complete_scat_out$Country,
    nudge_x = 0, nudge_y = - 0.5,
    check_overlap = T
  ) +
  labs(y= "Basic Subscription Price", x = "GDP USD (In Trillion)") +
  ggtitle("Basic Netflix Subscription Price")
```



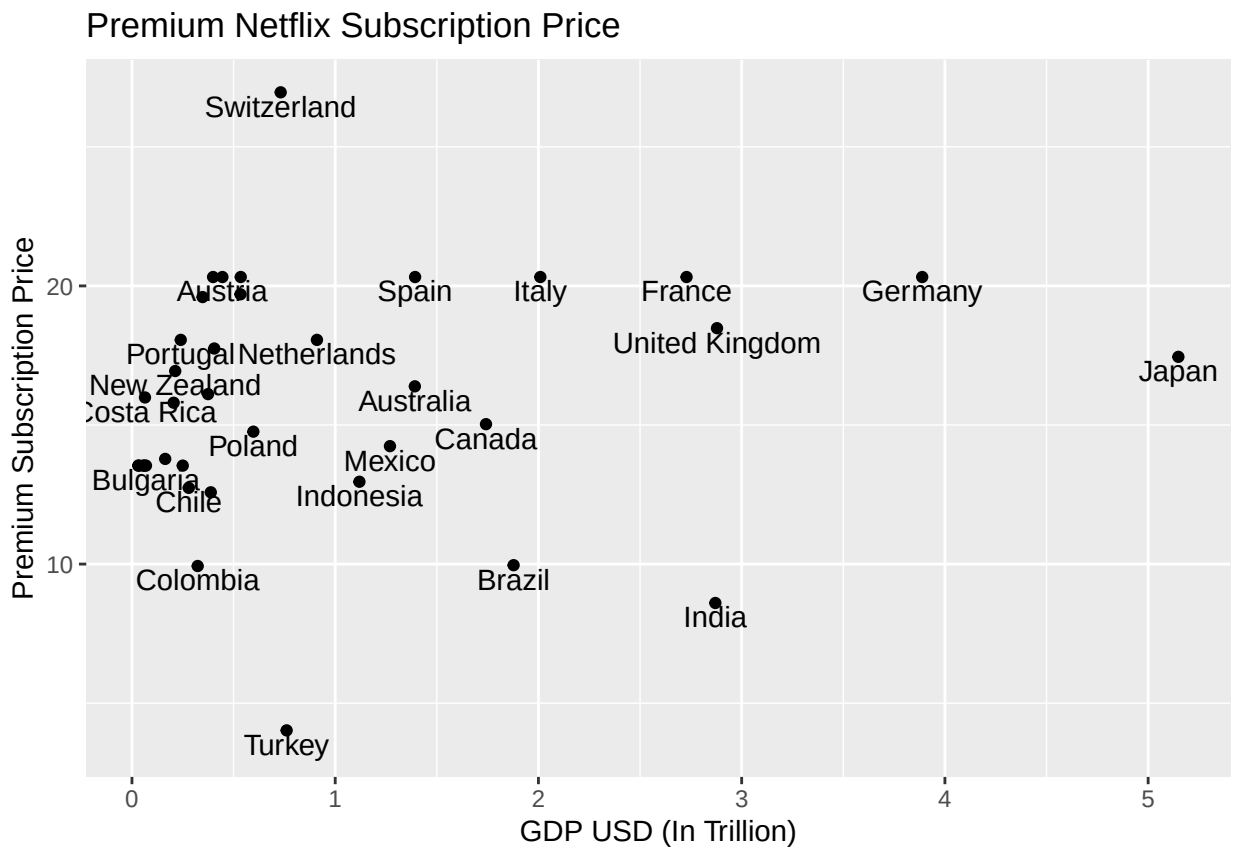
Analysis: For basic level subscribers, it is also noticeable that the higher the value, the less subscribers, even more so in developing countries. We see that in cases like Lithuania and Hungary, the catalog is very large, but has few subscribers and little revenue. Therefore, it is better, for the basic level, to reduce the catalog in order to reduce the subscription price and thus have more subscribers

```
## Standard.Subscription.Price
ggplot(complete_scat_out,
       aes(x=X2020.GDP..World.Bank. / 1000000000000, #trillion
           y=Standard.Subscription.Price)) + #Million
  geom_point() + # Show dots
  geom_text(
    label=complete_scat_out$Country,
    nudge_x = 0, nudge_y = - 0.5,
    check_overlap = T
  ) +
  labs(y= "Standard Subscription Price", x = "GDP USD (In Trillion)") +
  ggtitle("Standard Netflix Subscription Price")
```



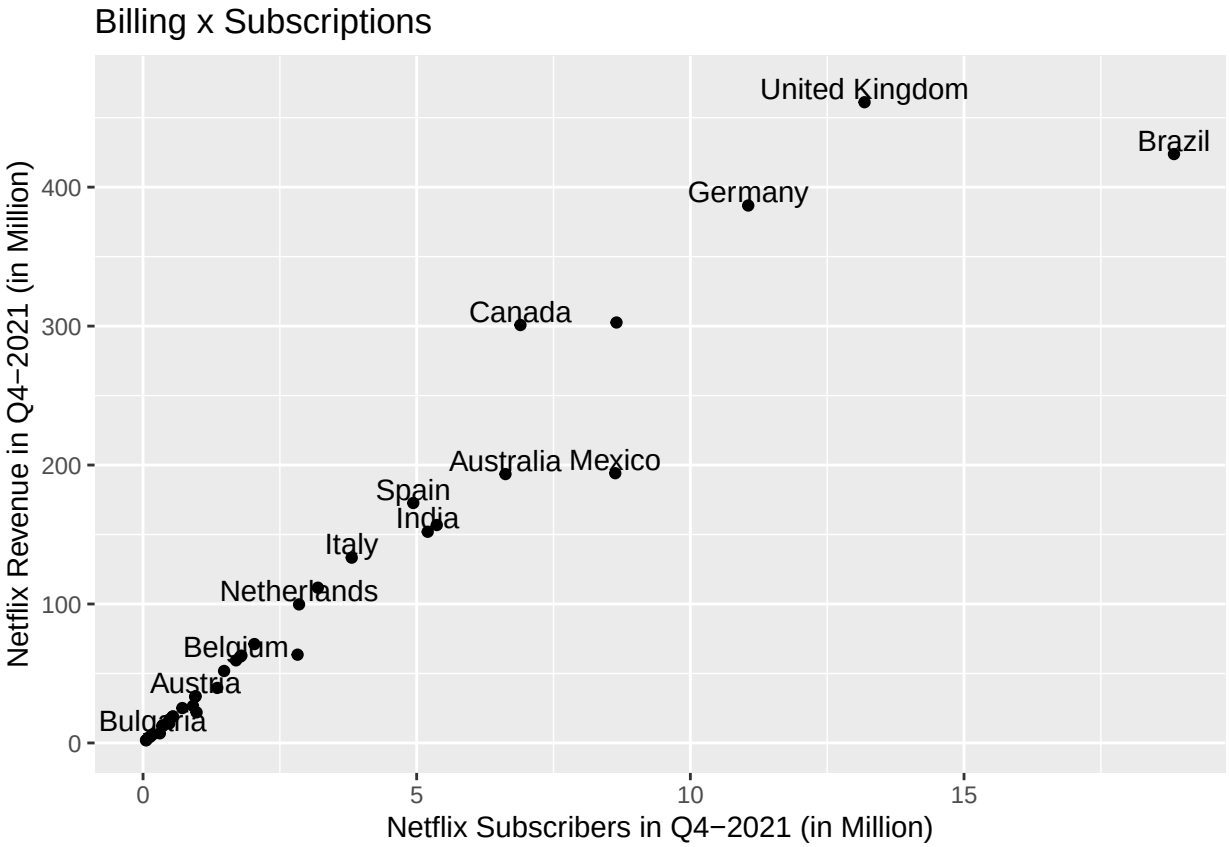
Analysis: For this plan, the same reasoning discussed above is followed. Perhaps a smaller catalog with a better price will help increase the number of subscribers and revenue

```
## Premium.Subscription.Price
ggplot(complete_scat_out,
       aes(x=X2020.GDP..World.Bank. / 1000000000000, #trillion
           y=Premium.Subscription.Price)) + #Million
  geom_point() + # Show dots
  geom_text(
    label=complete_scat_out$Country,
    nudge_x = 0, nudge_y = - 0.5,
    check_overlap = T
  ) +
  labs(y= "Premium Subscription Price", x = "GDP USD (In Trillion)") +
  ggtitle("Premium Netflix Subscription Price")
```



Analysis: Premium subscription presumes to have many subscribers due to better services.

```
## Billing x Subscriptions
ggplot(complete_scat_out,
       aes(x=Netflix.Subscriptions.Q42021 / 1000000, #Million
           y=Netflix.Q42021.Revenue/ 1000000)) + #Million
  geom_point() + # Show dots
  geom_text(
    label=complete_scat_out$Country,
    nudge_x = 0, nudge_y = 10,
    check_overlap = T
  ) +
  labs(y= "Netflix Revenue in Q4-2021 (in Million)", x = "Netflix Subscribers in Q4-2021 (in Million)") +
  ggtitle("Billing x Subscriptions")
```



Analysis: With this view, we understand that, to some extent, the number of subscribers directly influences revenue. However, around 6.7M of subscribers, the scenario may undergo some changes, often involving developing countries such as Brazil and Mexico.

Conclusion

We concluded that Netflix could, through a relocation, obtain more subscribers and, consequently, increase its influence and profitability by taking into account the country in which it operates and the number of series and films it supplies to that region, as well as the value for signature and catalog size

Disclaimer:

Disclaimer: a good part of this project was largely done in the Data Science Academy, Big Data Analytics with R and Microsoft Azure Machine Learning course (part of the Data Scientist training)

End