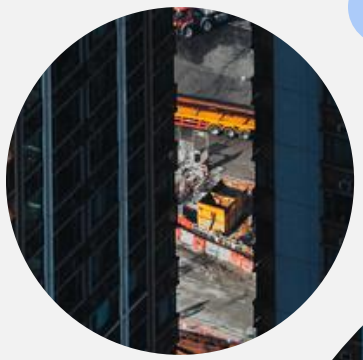

Dixon + Ratz
Final Project
August 10, 2021
Intro to Data Science Summer Course

Prerequisites

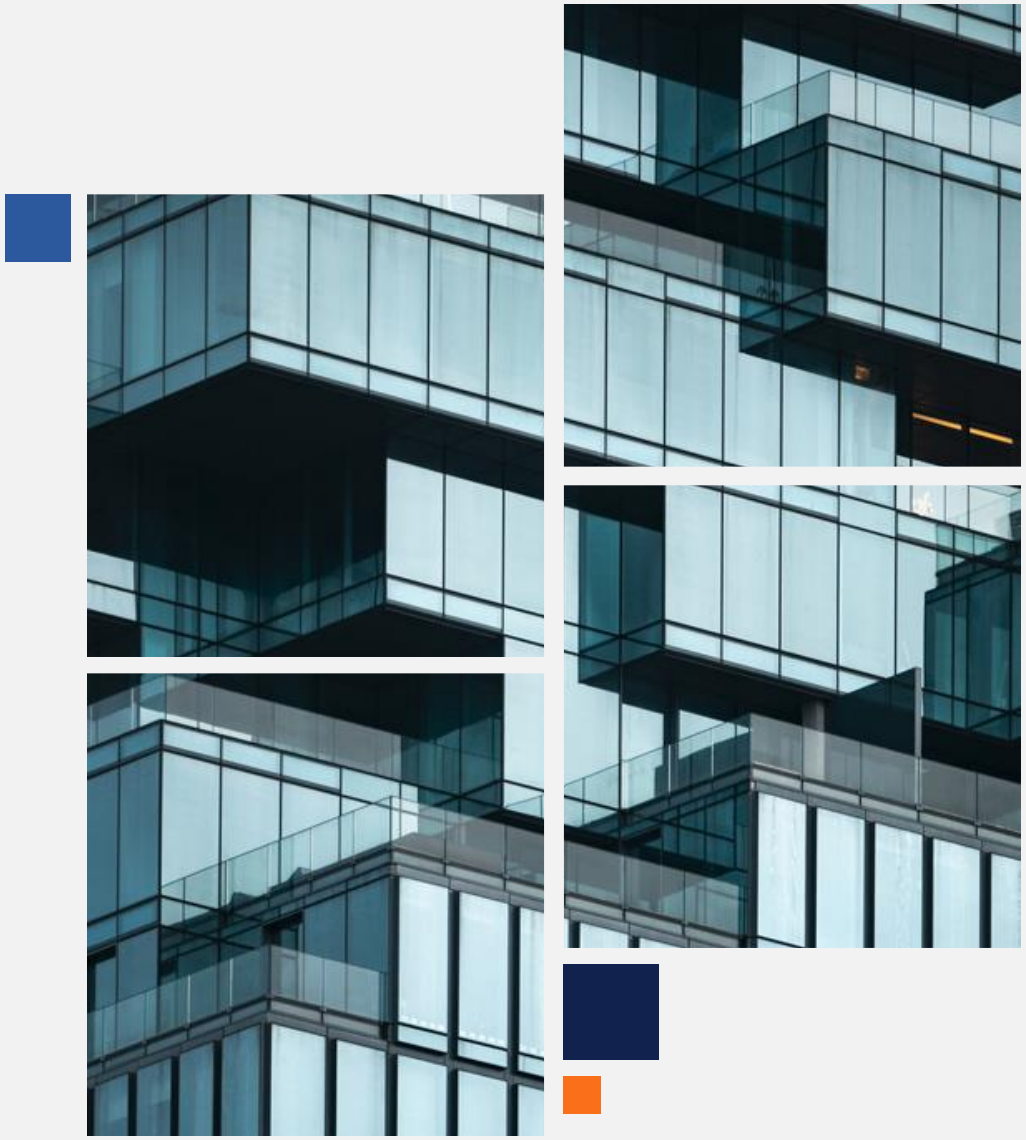


- Library:
  - Tidyverse
  - Dplyr
  - Ggplot2
  - Scales
  - lubridate

Introduction

Netflix provide its users a plethora of shows to watch. The data set we're using comes from Shivam Bansal. This data set is broken down into multiple columns, some of which are removed to irrelevance to the data set. The columns are the following:

- Type
- Title
- Director
- Country
- Release Year
- Rating
- Duration
- Listed_In



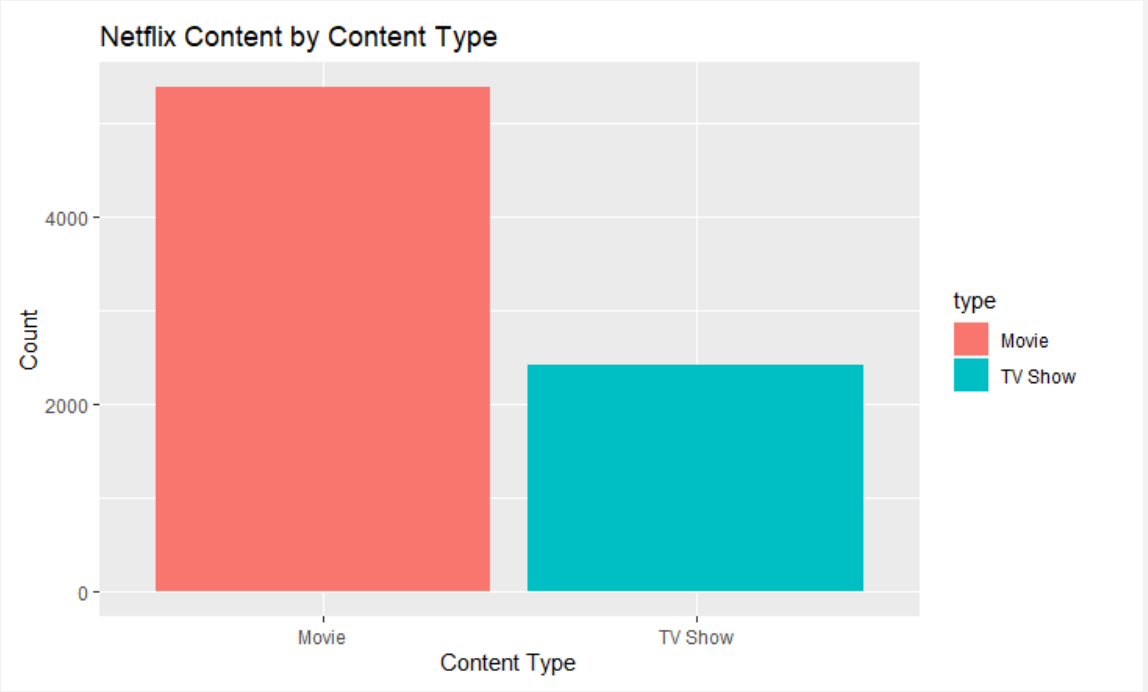
Data Exploration & Analysis

Let’s dive in

Content Types

- The visualization below graphically represents the distribution of the content released on Netflix's platform. They can be either labeled as 'Movie' or a 'TV Show'.

```
netflixshows%>%
  group_by(type)%>%
  filter(type != is.na(type))%>%
  ggplot()+
  geom_bar(mapping = aes(
    x = type,
    fill = type))+
  labs(
    title = "Netflix Content by Content Type",
    x = "Content Type",
    y = "Count"
  )
```

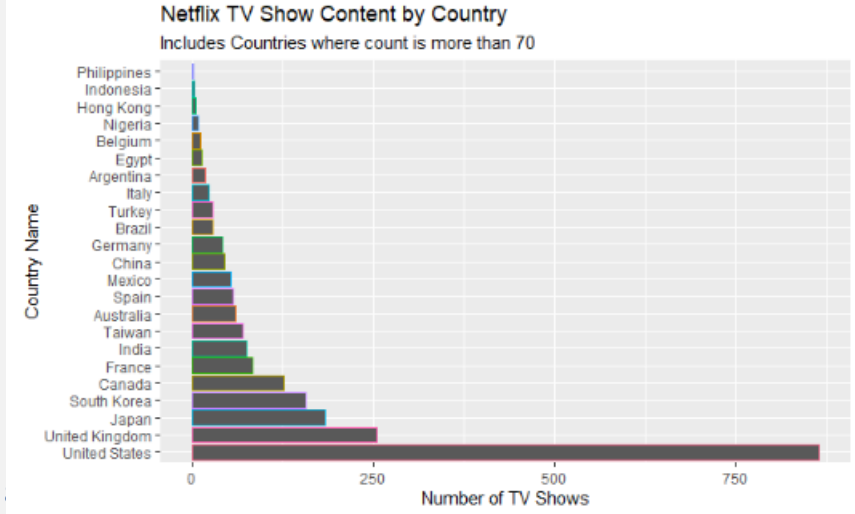


Content Type by Country

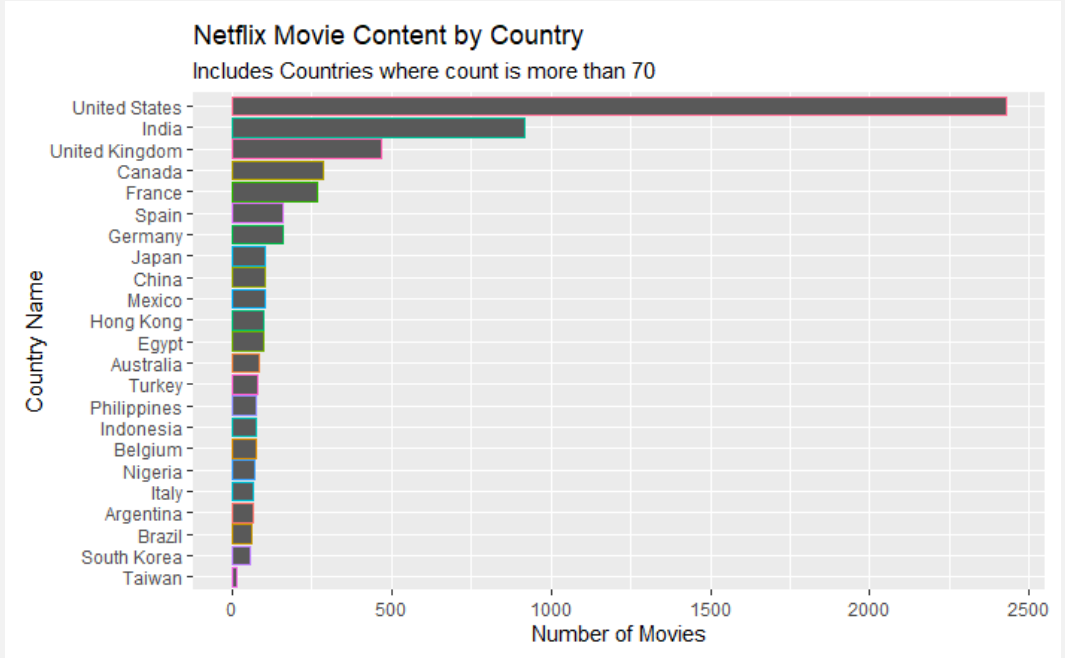
Netflix TV Show Content by Country

Hide

```
netflixshows%>%
  separate_rows(country, sep = ", ")%>%
  group_by(country)%>%
  filter(n()> 70 & type == "TV Show" & country != "")%>%
  count(.)%>%
  ggplot()+
    geom_bar(stat = "identity",mapping = aes(
      y = reorder(country, -n),
      x = n,
      color = country),
      na.rm = T,show.legend = F)+
        labs(
          title = "Netflix TV Show Content by Country",
          subtitle = "Includes Countries where count is more tha
n 70",
          y = "Country Name",
          x = "Number of TV Shows",

        )
```




```
#Because countries are sometimes grouped together, separating them is required.
  #Due to the size of the dataset, we had to filter out countries that had under 70 Movie entries.
netflixshows%>%
  separate_rows(country, sep = ", ")%>%
  group_by(country)%>%
  filter(n()> 70 & type == "Movie" & country != "")%>%
  count(.)%>%
  ggplot()+
    geom_bar(stat = "identity",mapping = aes(
      y = reorder(country,n),
      x = n,
      color = country),
      na.rm = T, show.legend = F)+
        labs(
          title = "Netflix Movie Content by Country",
          subtitle = "Includes Countries where count is more than 70",
          y = "Country Name",
          x = "Number of Movies"
        )
```



Analyzing potential Data Relationships

How does Data A have a relationship to Data B?

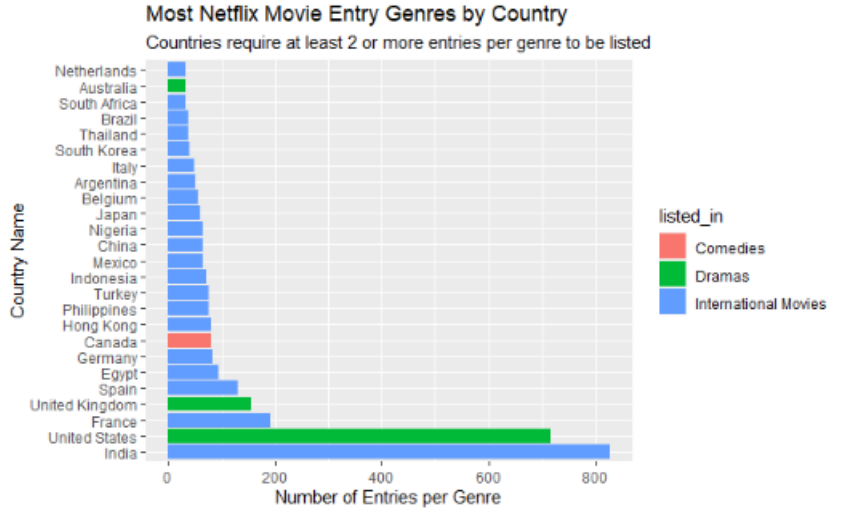
Genres

Each Country has their own favorite type of Genre. Below is a graphical representation of the relationship between Genres and associating countries. The data is separated by content type, either “Movie” or “TV Show”. A graph like this is helpful to answer questions such as:

- What Genre does each country prefer?
- Which country has the most of that specific genre?
- What is the target audience per country?
- Is Netflix showing appropriate content genres per country?

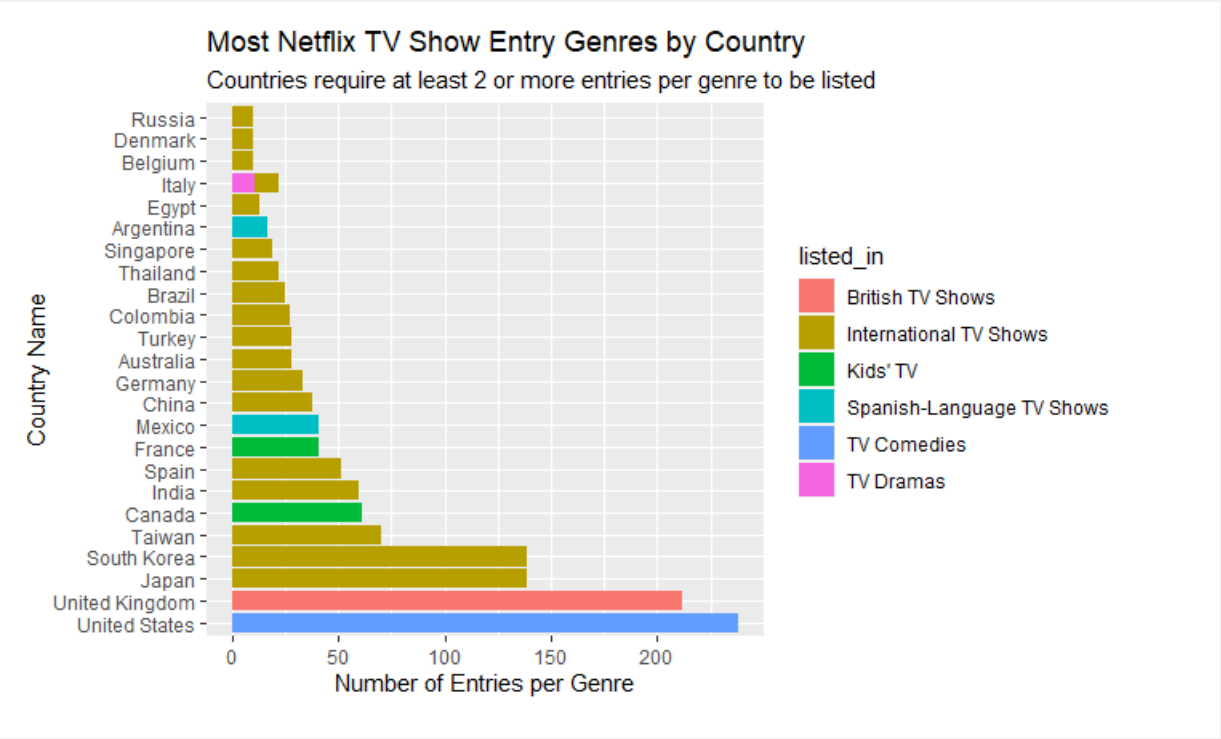
```
netflixshows%>%
  separate_rows(country, sep = ", ")%>%
  separate_rows(listed_in, sep = ", ")%>%
  filter(type == "Movie" & country != "" & country != is.na(coun
try))%>%
  group_by(country)%>%
  count(listed_in, name = "Genres")%>%
  filter(Genres == max(Genres) & Genres >= 2)%>%
  arrange(desc(Genres))%>%
  head(25)%>%
  ggplot(mapping = aes(
    x = Genres,
    y = reorder(country, -Genres),
    fill = listed_in
  ))+
  geom_histogram(stat = "identity")+
  labs(
    title = "Most Netflix Movie Entry Genres by Country",
    subtitle = "Countries require at least 2 or more entries per
 genre to be listed",
    x = "Number of Entries per Genre",
    y = "Country Name"
  )
```

```
Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Genres pt.2

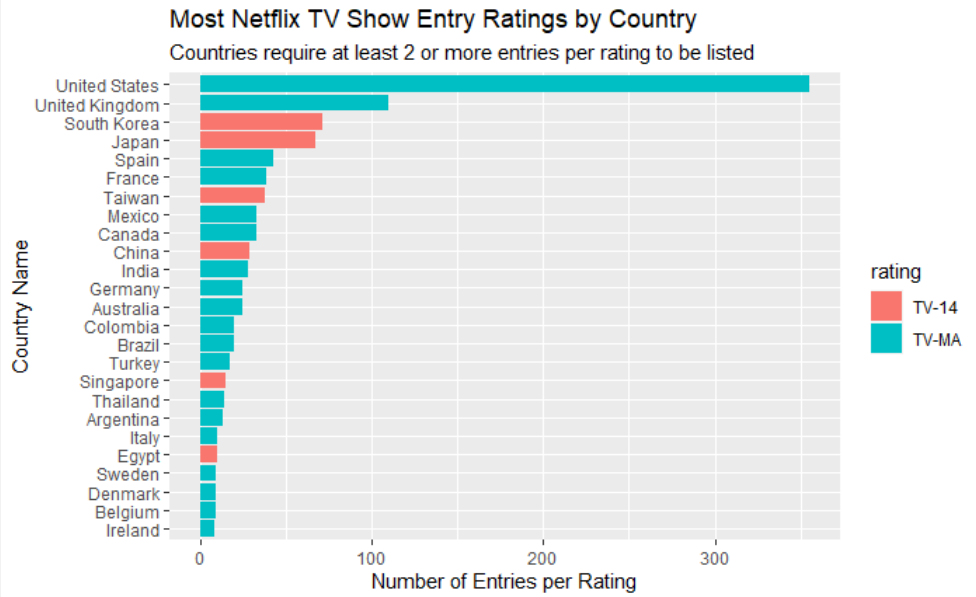
```
netflixshows%>%
  separate_rows(country, sep = ", ")%>%
  separate_rows(listed_in, sep = ", ")%>%
  filter(type == "TV Show" & country != "" & country != is.na(country))%>%
  group_by(country)%>%
  count(listed_in, name = "Genres")%>%
  filter(Genres == max(Genres) & Genres >= 2)%>%
  arrange(desc(Genres))%>%
  head(25)%>%
  ggplot(mapping = aes(
    x = Genres,
    y = reorder(country, -Genres),
    fill = listed_in
  ))+
  geom_histogram(stat = "identity")+
  labs(
    title = "Most Netflix TV Show Entry Genres by Country",
    subtitle = "Countries require at least 2 or more entries per genre to be listed",
    x = "Number of Entries per Genre",
    y = "Country Name"
  )
```



Ratings

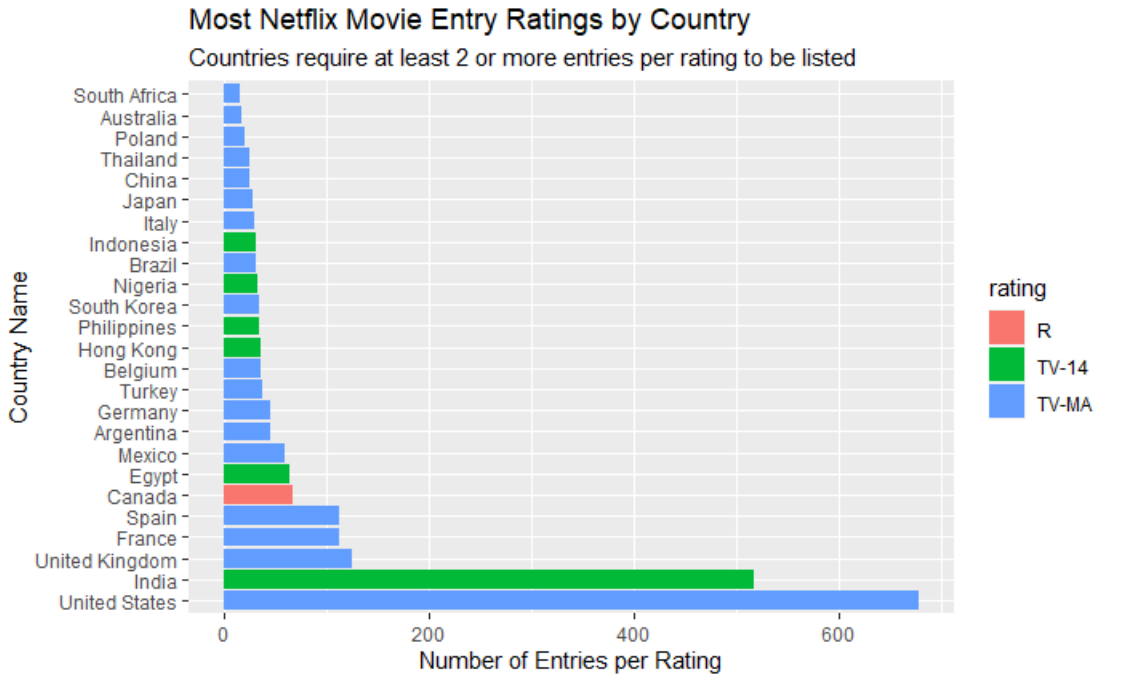
```
netflixshows%>%
  separate_rows(country, sep = ", ")%>%
  filter(type == "TV Show" & country != "" & country != is.na(country))%>%
  group_by(country)%>%
  count(rating, name = "RatingN")%>%
  filter(RatingN == max(RatingN) & RatingN >= 2)%>%
  arrange(desc(RatingN))%>%
  head(25)%>%
  ggplot(mapping = aes(
    x = RatingN,
    y = reorder(country,RatingN),
    fill = rating
  ))+
  geom_histogram(stat = "identity")+
  labs(
    title = "Most Netflix TV Show Entry Ratings by Country",
    subtitle = "Countries require at least 2 or more entries per rating to be listed",
    x = "Number of Entries per Rating",
    y = "Country Name")
```

```
Warning: Ignoring unknown parameters: binwidth, bins, pad
```



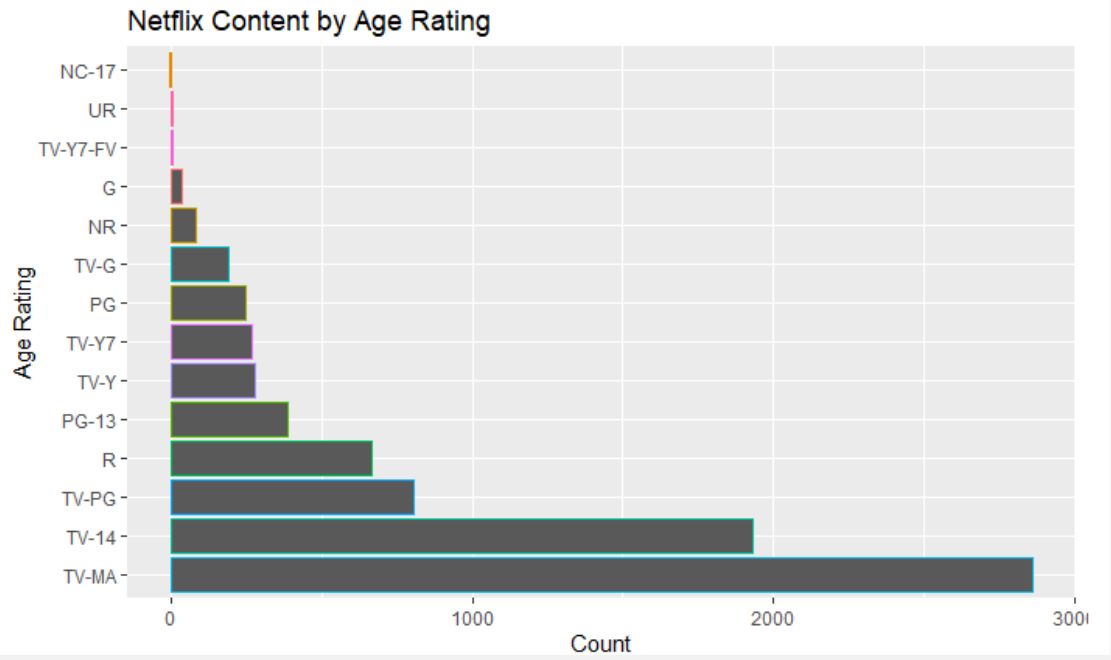

```
netflixshows%>%
  separate_rows(country, sep = ", ")%>%
  filter(type == "Movie" & country != "" & country != is.na(country))%>%
  group_by(country)%>%
  count(rating, name = "RatingN")%>%
  filter(RatingN == max(RatingN) & RatingN >= 2)%>%
  arrange(desc(RatingN))%>%
  head(25)%>%
  ggplot(mapping = aes(
    x = RatingN,
    y = reorder(country, -RatingN),
    fill = rating
  ))+
  geom_histogram(stat = "identity")+
  labs(
    title = "Most Netflix Movie Entry Ratings by Country",
    subtitle = "Countries require at least 2 or more entries per rating to be listed",
    x = "Number of Entries per Rating",
    y = "Country Name")
```

```
Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Average Rating per Country

```
netflixshows%>%
  group_by(rating)%>%
  filter(rating != "" & rating != is.na(rating))%>%
  count(.)%>%
  ggplot()+
  geom_bar(stat = "identity",mapping = aes(
    y = reorder(rating, -n),
    x = n,
    color = rating),show.legend = F)+
    labs(
        title = "Netflix Content by Age Rating",
        x = "Count",
        y = "Age Rating"
      )
```

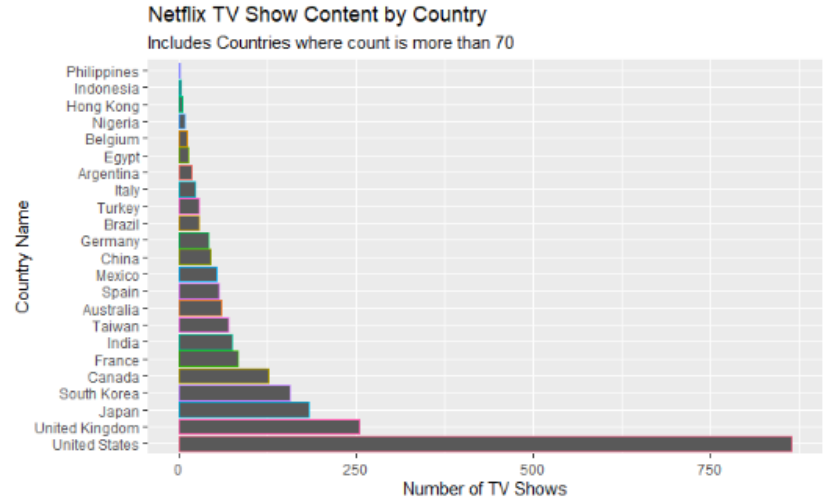


Potential connection?

Netflix TV Show Content by Country

Hide

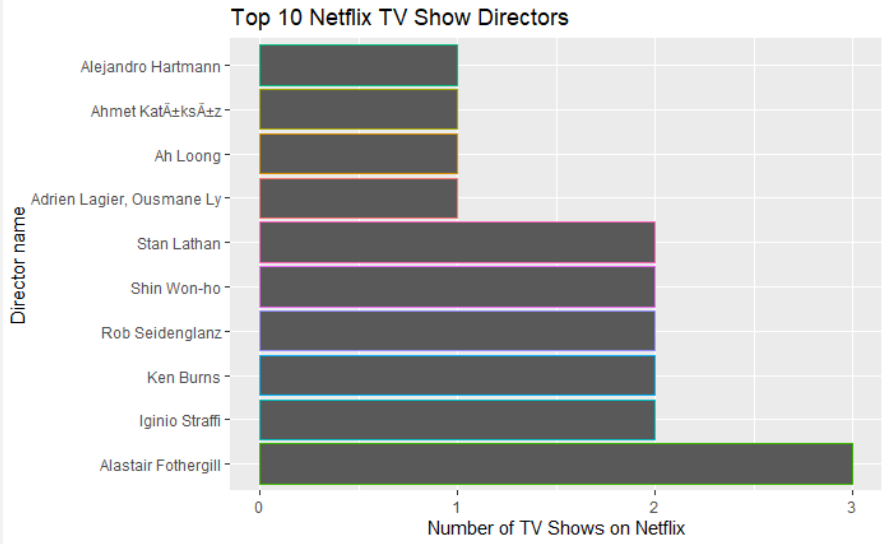
```
netflixshows%>%
  separate_rows(country, sep = ", ")%>%
  group_by(country)%>%
  filter(n()> 70 & type == "TV Show" & country != "")%>%
  count(.)%>%
  ggplot()+
    geom_bar(stat = "identity",mapping = aes(
      y = reorder(country, -n),
      x = n,
      color = country),
      na.rm = T,show.legend = F)+
        labs(
          title = "Netflix TV Show Content by Country",
          subtitle = "Includes Countries where count is more tha
n 70",
          y = "Country Name",
          x = "Number of TV Shows",

        )
```



Top Directors

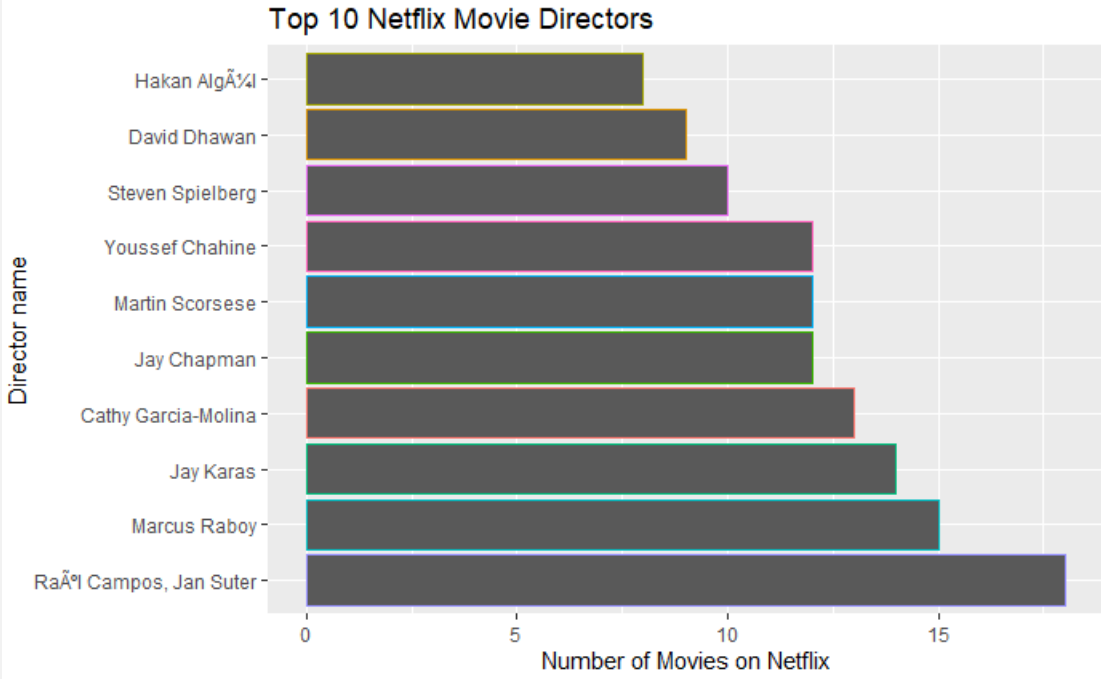
```
netflixshows%>%
  filter(director != is.na(director) & director != "" & type == "TV Show")%>%
  group_by(director)%>%
  summarize(director)%>%
  count(.)%>%
  arrange(desc(n))%>%
  head(10)%>%
  ggplot(mapping = aes(
    y = reorder(director, -n),
    x = n,
    color = director))+
  geom_bar(stat = "identity",show.legend = F)+
  labs(
    title = "Top 10 Netflix TV Show Directors",
    y = "Director name",
    x = "Number of TV Shows on Netflix")
```

```
`summarise()` has grouped output by 'director'. You can override using the `.groups` argument.
```




```
netflixshows%>%
  filter(director != is.na(director) & director != "" & type == "Movie")%>%
  group_by(director)%>%
  summarize(director)%>%
  count(.)%>%
  arrange(desc(n))%>%
  head(., n = 10)%>%
  ggplot(mapping = aes(
    y = reorder(director, -n),
    x = n,
    color = director))+
  geom_bar(stat = "identity",show.legend = F)+
  labs(
    title = "Top 10 Netflix Movie Directors",
    y = "Director name",
    x = "Number of Movies on Netflix")
```

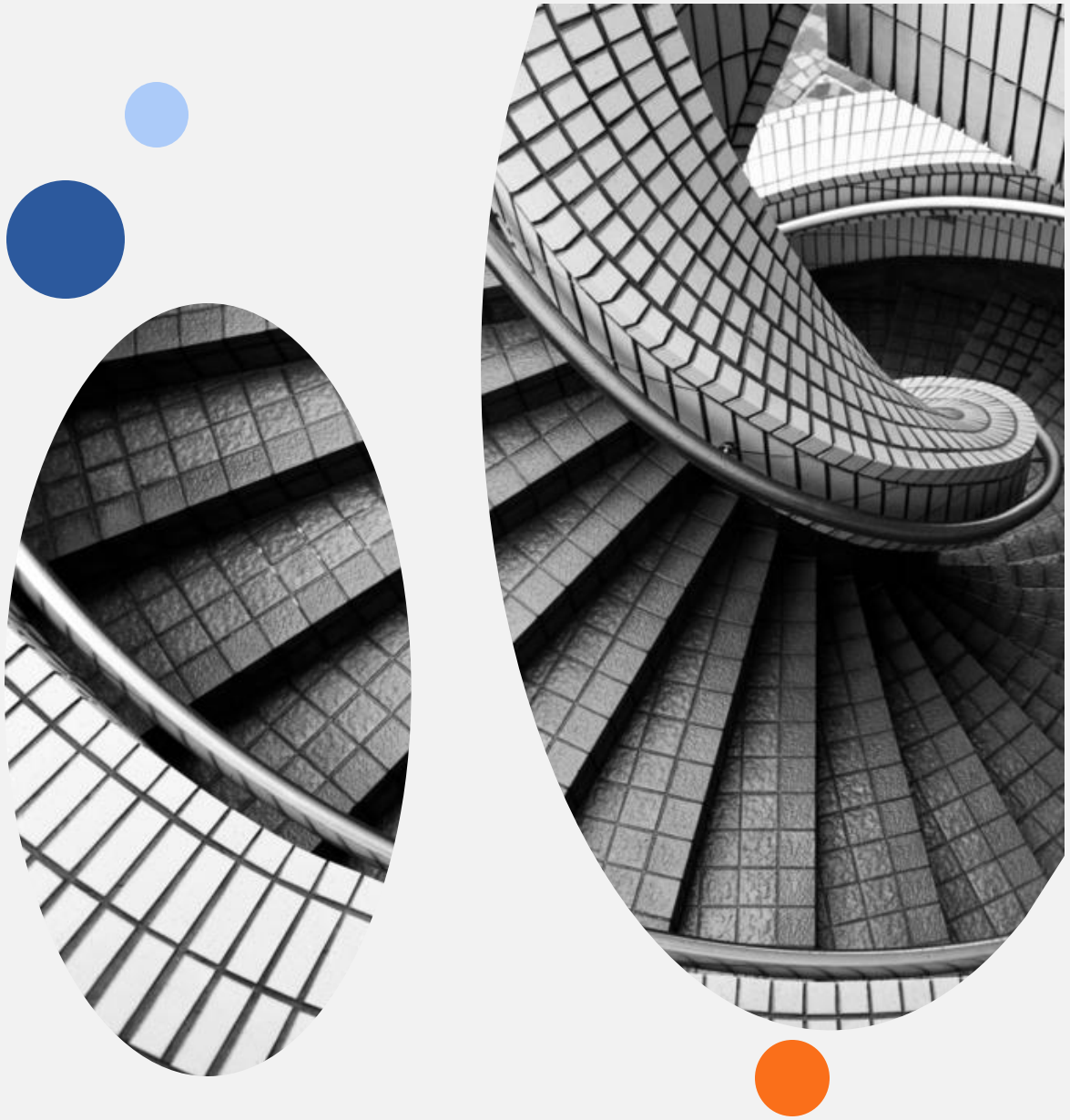
```
`summarise()` has grouped output by 'director'. You can override using the `.groups` argument.
```



Conclusion

After further analysis of the data set, we've gathered enough data to answer the questions that we asked.

The use of the Netflix Dataset, even though it wasn't supplied to us via github, made an excellent source of information. The abundance of data and variables made it a good visual representation of Netflix as a whole. Through the course of this project, we learned the importance of clear and concise visual representations to help convey an overall message. The use of our skills helped strengthen our knowledge of R as a language and R Studio as an effective IDE to use. We were able to apply data manipulating tools, such as 'mutate()', 'filter()', and others.



Citations

Special thanks to Shivam Bansal for providing the Netflix Dataset, which can be found here!

For more detailed report, please click here!