Data Analysis on Netflix datasets

Vybhav Achar Bhargav
University of California, Davis
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Seongwoo Choi
University of California, Davis
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

STA 220 Final Project - Data Analysis on Netflix datasets

Vybhav Achar Bhargav, Seongwoo Choi, David Haddad

Abstract

Exploring datasets of Netflix for Future Release of TV shows and Movies on the Platform. In this project, we are going to explore the dataset from Kaggle and we would like to find out how long the Netflix platform takes a movie or a TV show to release on its platform, how many movies and TV shows are released in specific time frame, how many movies and TV shows are release in the recent ten years on the platform, and what were the top 10 genres that the audience of the Netflix platform liked the most. From here, we would like to apply a machine learning approach to understand the data fully and provide a great solution where the platform should be headed to. From our data analysis we conducted on R markdown, we have discovered that there were a wide variety of genres that movie directors produced worldwide and we have observed many cast members and genres they were in.

Data Analysis on the Netflix Datasets

Motivation

Netflix is the largest online movie and TV show streaming service on the planet. Its service is widely available in many countries including but not limited to the United States, India, South Korea, Japan, and many more. The service was first introduced as a DVD rental service on the Internet and later, the founder and CEO of the company Reed Hastings transitioned to a revolutionary way of delivering movies and TV shows through its website allowing many users to directly stream their favorite contents on their Internet-enabled devices including desktop computers, laptops, tablet PCs, mobile phones, and many more. With its a whole new approach of delivering shows and movies, the sales of Netflix went up exponentially. Since then, the platform created its own recommender system to understand what types of movies and TV shows the users would like to watch, what kind of style of cinematography they liked the most, and how they consume their favorite TV shows. With such analysis, the company released 'The House of Cards', which became a huge success in the history of streaming service providers. With the power of data analysis, more users were attracted to the platform, and many users tend to spend most of their time watching shows and movies on Netflix. With this approach, we would like explore the dataset to understand the trend of movies and TV shows on Netflix.

Introduction

We first wanted to get an overview of the dataset that we were dealing with. First we loaded up tidyverse for a simple data anlaysis purpose. We got the dataset from Kaggle and we are going to utilize data that the Kaggle website provides to understand the trend of movies and TV shows released on the platform. This dataset consists of

From the code, we could see the column names that the CSV file contains. We will utilize the following columns to understand what movies and TV shows were released in specific year, what genres they were, date when they were released and the rating the audience gave and so on.

From the column names, we could observe that there are twelve columns: show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description.

We will first demonstrate the overview of the dataset. Each column contains 8807 in length except for release year in interquartile range.

```
summary(df_netflix)
```

```
##    show_id              type               title             director
##  Length:8807        Length:8807        Length:8807        Length:8807
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      cast              country            date_added          release_year
##  Length:8807        Length:8807        Min.   :2008-01-01   Min.   :1925
##  Class :character   Class :character   1st Qu.:2018-04-20   1st Qu.:2013
##  Mode  :character   Mode  :character   Median :2019-07-12   Median :2017
##                                        Mean   :2019-05-23   Mean   :2014
##                                        3rd Qu.:2020-08-26   3rd Qu.:2019
##                                        Max.   :2021-09-25   Max.   :2021
##                                        NA’s   :98
##     rating            duration          listed_in         description
##  Length:8807        Length:8807        Length:8807        Length:8807
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

From the overview of the dataset, we could tell that the earliest year recorded as the movie or the TV show is 1925 and the latest content that is on the platform is 2021. With the information of directors, actors/actresses, movies, TV shows, rating, duration of each content and more, we can understand how long does Netflix take to upload the content on the platform, which genre is most popular on the platform, which actor/actresses are most popular and how genre popularity changed the movie or TV shows that an actor/actress appear on the genre over time. We will explore the trend of Netflix using R to figure it out.

Project Background

The report presents a data analysis research and coordination activity one in the Netflix dataset found on Kaggle to establish a new data exploration on the trend of movies on the platform. When Netflix was first launched, it did not have any data analysis to understand the trend of audience/users using the platform as previously mentioned. As time passes by, the importance of utilizing data analysis started to emerge, and the biggest hit shows were released on the platform after scrutinizing the data the company collected from the users. In the project, the team of three is working on a same procedure to understand the trend of the platform and attempt to understand the average time the platform takes to upload the contents, the top ten film directors in top ten countries where the platform is streaming, and the top genres that the top actors/actresses are in. The team will, then, attempt to understand the overall trend of the Netflix platform according to the dataset the team got from Kaggle.

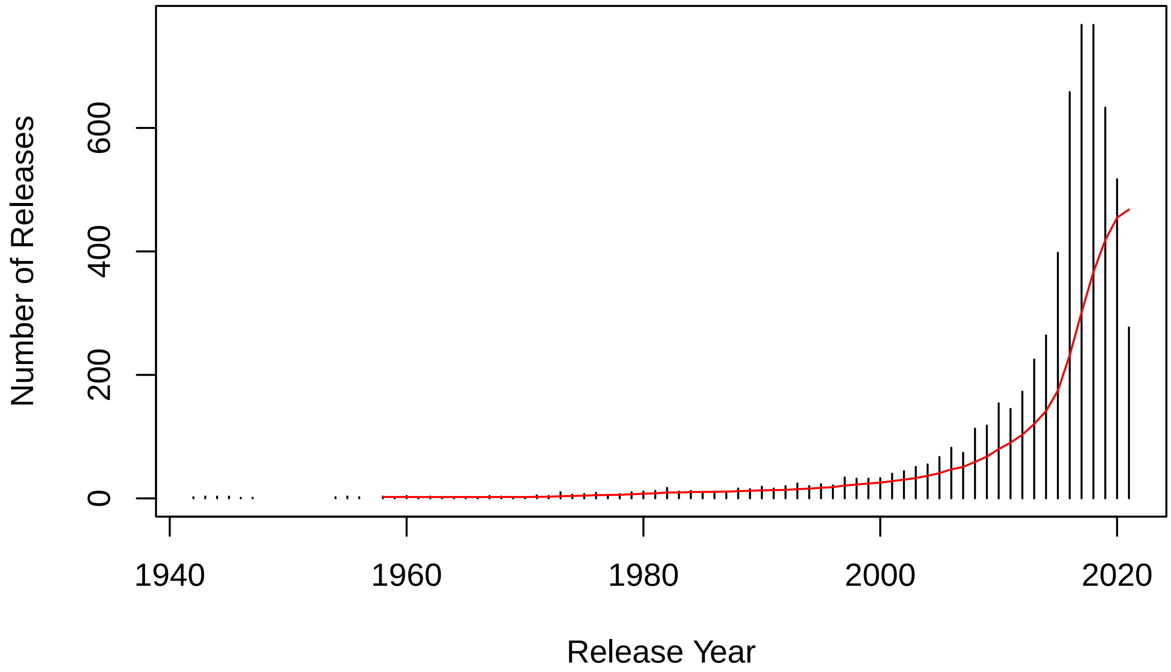
Methodology

Before initiating the data analysis, the dataset has been separated into two different sets: one is for movies only and the another dataset is TV shows only. Every date on the dataset is transformed to day and year format.

```
df_netflix <- transform(df_netflix,
                        date_added=as.Date(
                          date_added, tryFormats =
                            c("%B %d,%Y")))
movies <- subset(df_netflix, type=="Movie")
tvs <- subset(df_netflix, type == "TV Show")
```

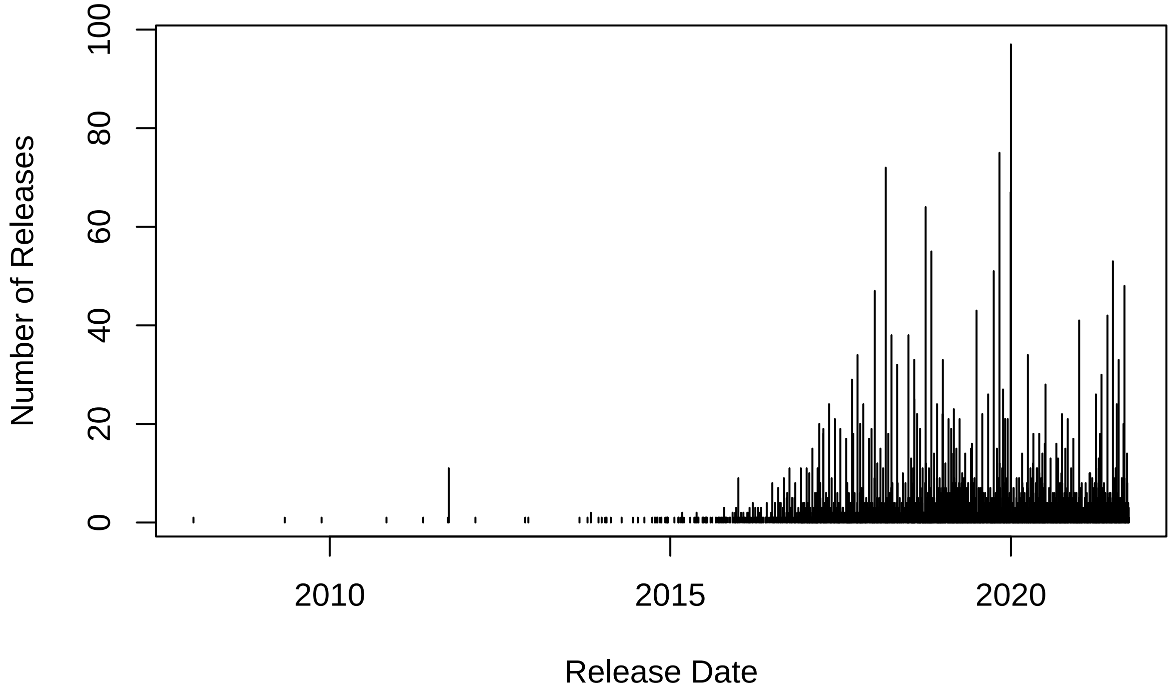
Release year is very important aspect of learning when a content on the platform is released and the average time the platform takes to upload the content.

```
# Plot of number of movies release per year
numRelease <- movies %>% group_by(release_year)
tallyRelease <- tally(numRelease);

# 10 Year running avg of movie releases
tallyRelease$avg <- stats::filter(
  tallyRelease$n, rep(1/10, 10),
  method="convolution", sides=1);
plot(tallyRelease$release_year, tallyRelease$n,
     type = "h", xlab = "Release Year", ylab = "Number of Releases")
lines(tallyRelease$release_year, tallyRelease$avg, col="red");
```



The plotted graph indicates that most of the movies were released after the year 2000s. From 2000 to 2020, there is an exponential growth in the graph, which indicats that the number of releases that the movies were introduced to the platform started to grow higher.

```
# Plot of number of movies release per year
numDateAdded <- movies %>% group_by(date_added)
tallyDateAdded <- tally(numDateAdded);

plot(tallyDateAdded$date_added,
     tallyDateAdded$n, type = "h", xlab =
       "Release Date", ylab = "Number of Releases")
```



The graph above indicates the number of movies released in specif time frame. From 2010 to 2020, there are high increases in numbers of releases on the platform. Due to this reason, from the rest of the data analysis, the team will dissect the time frame from 2008 to 2020 to understand the overall trend of releases on the platform.

```r
movies$year_added_to_netflix <- strtoi(
  format(movies$date_added, format="%Y"))
movies$diff = movies$year_added_to_netflix - movies$release_year

v <- movies %>% select(diff, year_added_to_netflix, release_year)
#pick the variable

# set up cut-off values
breaks <- c(0,2,5,10,15,20,25,30,35,40)
# specify interval/bin labels
tags <- c(
  "[0-2)",
  "[2-5)",
  "[5-10)",
  "[10-15)",
  "[15-20)",
  "[20-25)",
  "[25-30)",
  "[30-35)",
  "[35-40)",
  ">40"
)

vgroup <- as_tibble(v) %>%
  mutate(tag = case_when(
    diff < 2 ~ tags[1],
    diff >= 2 & diff < 5 ~ tags[2],
    diff >= 5 & diff < 10 ~ tags[3],
    diff >= 10 & diff < 15 ~ tags[4],
    diff >= 15 & diff < 20 ~ tags[5],
    diff >= 20 & diff < 25 ~ tags[6],
    diff >= 25 & diff < 30 ~ tags[7],
    diff >= 30 & diff < 35 ~ tags[8],
    diff >= 35 & diff < 40 ~ tags[9],
    diff >= 40 ~ tags[10],
    ))

vgroup$tag <- factor(vgroup$tag,
                       levels = tags,
                       ordered = FALSE)
```

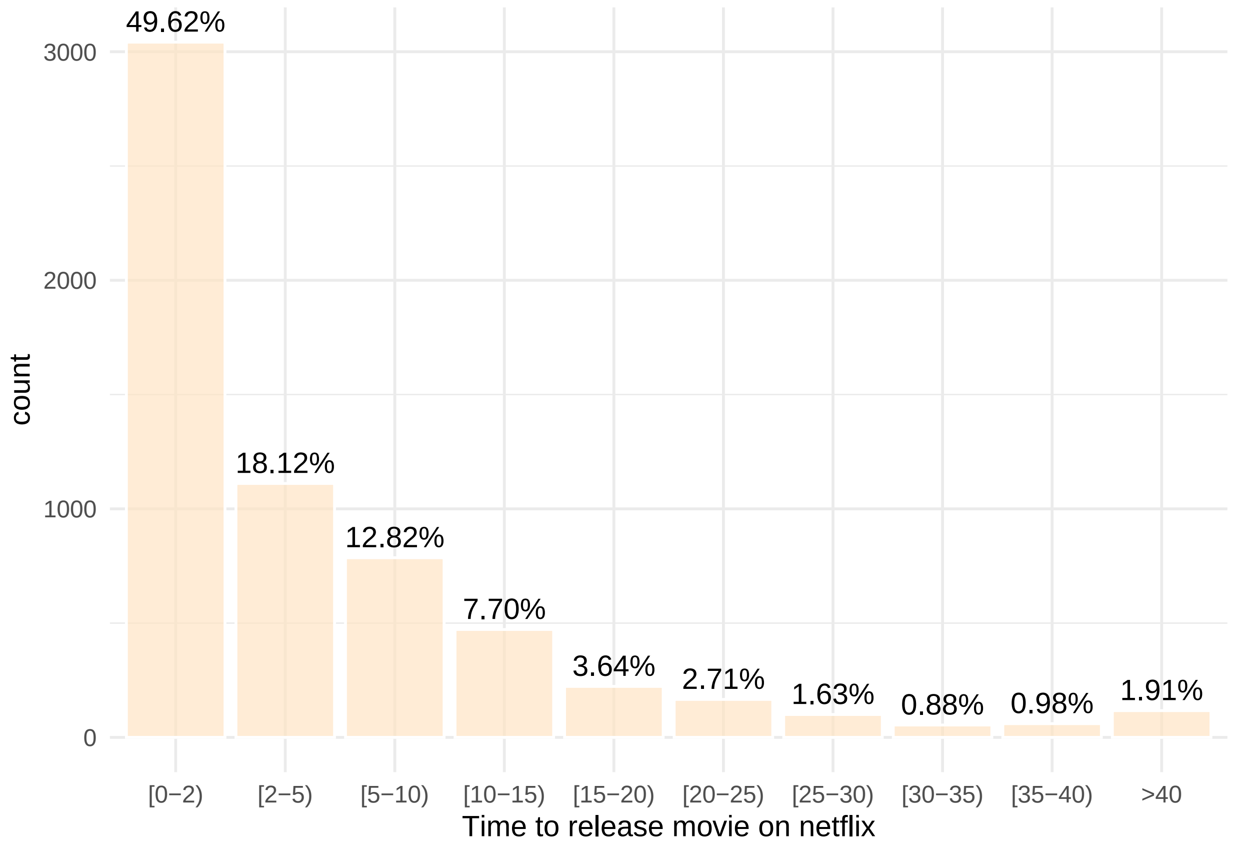
We need to understand and separate the group of releases in specific year. We also need to understand the average time the platform releases the contents on the website and how much would it need to release the content.

```
movies %>% count(listed_in, country, sort=TRUE) %>% slice(1:10)
```

```
##                                            listed_in       country   n
## 1                                      Documentaries United States 249
## 2                                    Stand-Up Comedy United States 209
## 3             Comedies, Dramas, International Movies         India 120
## 4                       Dramas, International Movies         India 118
## 5   Dramas, Independent Movies, International Movies         India 108
## 6                 Children & Family Movies, Comedies United States  90
## 7                                             Dramas United States  88
## 8                                           Comedies United States  84
## 9                           Children & Family Movies United States  80
## 10                        Dramas, Independent Movies United States  74
```

The top 10 genres that are released in this time frame indicates that the United States released the most documentaries on the platform and the naiton also produced stand-up comedies. India came the second place when it comes to which international country released the most contents on the platform.

```
ggplot(data = as_tibble(vgroup$tag),
       mapping = aes(x=value)) +
  geom_bar(fill="bisque",color="white",alpha=0.7) +
  stat_count(
    geom="text",
    aes(label=sprintf("%.2f%%",
                      (..count../length(vgroup$tag))*100)),
    vjust=-0.5) +
  labs(x='Time to release movie on netflix') +
  theme_minimal()
```



The graph above indicates time to release movies on Netflix. Many of the contents were uploaded within 0-2 years on the platform after they are released to the public. The graph also indicates that many Netflix original movies are released and uploaded simultaneously on the platform upon their debut.

```
movies_country_seperated = movies %>%
  separate_rows(country, sep=",") %>%
  mutate(across(where(is.character), str_trim))

movies_genre_seperated <- movies %>%
  separate_rows(listed_in, sep=",") %>%
  mutate(across(where(is.character), str_trim))
```

```r
top_genre_per_country <- movies_country_seperated %>%
  filter(!is.na(country)) %>%
  distinct(show_id, .keep_all = TRUE) %>%
  count(country, sort=TRUE) %>% head(n=10)

tg <- top_genre_per_country %>%
pull(country) %>%
sapply(function(cnt){
 movies_genre_seperated %>%
    filter(country == cnt) %>%
    count(listed_in, sort=TRUE) %>%
    slice(1:5) %>%
    pull(listed_in)
})

as.data.frame(tg)
```

```
##              United States                India       United Kingdom
## 1                   Dramas International Movies        Documentaries
## 2                 Comedies               Dramas International Movies
## 3            Documentaries             Comedies               Dramas
## 4       Independent Movies   Independent Movies             Comedies
## 5 Children & Family Movies   Action & Adventure   Independent Movies
##                     Canada               France                Spain
## 1                 Comedies International Movies International Movies
## 2 Children & Family Movies               Dramas               Dramas
## 3     International Movies        Documentaries             Comedies
## 4                   Dramas             Comedies            Thrillers
## 5            Documentaries   Action & Adventure        Documentaries
##                  Egypt              Nigeria               Mexico
## 1 International Movies International Movies International Movies
## 2             Comedies               Dramas               Dramas
## 3               Dramas             Comedies      Stand-Up Comedy
## 4   Action & Adventure      Romantic Movies             Comedies
## 5      Romantic Movies            Thrillers   Independent Movies
##                      Japan
## 1     International Movies
## 2           Anime Features
## 3       Action & Adventure
## 4                   Dramas
## 5 Children & Family Movies
```

The result above demonstrates the top ten countries with top five genres that the audience watch. If we explore only the United States, we can observe that top five genres are Drama, Comedies, Documentaries, Independent Movies, and Family movies. Since we have observed which genres are popular in each country on the list, we will then explore the top genre leased in each year from 2008 to 2020.

```
movies_genre_seperated %>%
    filter(country=="United States") %>%
    group_by(year_added_to_netflix) %>%
    slice_max(listed_in, n = 1) %>%
    distinct(listed_in) %>%
    rename(top_genre=listed_in) %>%
    rename(year=year_added_to_netflix)
```

```
## # A tibble: 14 x 2
## # Groups:   year [14]
##    top_genre        year
##    <chr>           <int>
##  1 Thrillers        2008
##  2 Horror Movies    2009
##  3 Horror Movies    2010
##  4 Thrillers        2011
##  5 Documentaries    2012
##  6 Stand-Up Comedy  2013
##  7 Stand-Up Comedy  2014
##  8 Thrillers        2015
##  9 Thrillers        2016
## 10 Thrillers        2017
## 11 Thrillers        2018
## 12 Thrillers        2019
## 13 Thrillers        2020
## 14 Thrillers        2021
```

Most of the times, thriller seemed to be the most popular genre from year 2008 to 2020 in the United States despite the fact that they are not one of top five genres that the United States audience like to watch. One thing to note is that the thriller genre may mixed up with the other genres. For instance thriller, comedy could be the genre that the United States audience watched in 2008.

From now on, we will see which directors produced the most genres in the United States from 2008 to 2020.

Top 10 Directors that produced movies in the United States

1) **What is the top genre that the top 10 directors produced in the United States?**

The top 10 directors that produced movies in the United States have made mostly stand-up comedies. This makes sense due to the little time it takes to record a live performance show and edit them as compared to hiring cast and writing scripts for movies that could take years.

2) **How many movies did the top 10 directors make that were produced in the United States?**

The top director produced 15 movies, and the top genre was stand-up comedy which makes sense because of the same reason in the question above

```r
directors_produced_movies = movies %>%
    separate_rows(director, sep=",") %>%
    mutate(across(where(is.character), str_trim)) %>%
    filter(country=="United States")

top_directors <-  directors_produced_movies %>%
                    filter(!is.na(director)) %>%
                    group_by(director) %>%
                    count(sort=TRUE) %>%
                    head(n=10)

temp <- top_directors %>%
pull(director) %>%
sapply(function(d){
    directors_produced_movies %>%
    filter(director == d) %>%
    slice_max(listed_in, n=1) %>%
    pull(listed_in) %>% head(1)
})

director_top_genre = as.data.frame(temp)
final_table <- cbind(top_directors, genre=director_top_genre$temp)
final_table <- final_table %>%
  rename(total_movies_produced=n) %>%
  rename(top_directors=director)
final_table = as.data.frame(final_table)
final_table
```

```
##       top_directors total_movies_produced                              gen
## 1         Jay Karas                    15                    Stand-Up Come
## 2      Marcus Raboy                    14                    Stand-Up Come
## 3       Jay Chapman                    12                    Stand-Up Come
## 4   Shannon Hartman                     9                    Stand-Up Come
## 5   Martin Scorsese                     8                          Thrille
## 6       Troy Miller                     8                    Stand-Up Come
## 7       Lance Bangs                     7                    Stand-Up Come
## 8      Leslie Small                     7                    Stand-Up Come
## 9       Ryan Polito                     7                    Stand-Up Come
## 10 Robert Rodriguez                     6 Children & Family Movies, Comedie
```

The chart indicates that Jay Karas made 15 Stand-Up comedy movies from 2008 to 2020. Marcus Raboy made fourteen stand up comedy movies and so on. The top ten directors are from the United States. The following shows the international directors that produced the movies in specific time frame.

Top 10 Directors that produced movies in international countries

1) **What is the top genre that the top 10 directors produced in international countries?**

The top 10 directors that produced movies in international countries have made mostly stand-up comedies. This makes sense due to the little time it takes to record a live performance show and edit them as compared to hiring cast and writing scripts for movies that could take years. But many international movies have taken second spot for most movies directed unlike in the United States.

2) **How many movies did the top 10 directors make that were produced in international countries?**

The top director produced 21 movies, and the top genre was stand-up comedy. Third top director made 13 movies that had the most genre that is international movies and romantic movies. So more diversity internationally compared to US produced movies top directors.

Print only International in 2008-2021 and display names of directors who produced the most movies and TV shows

This will answer in which country a director produced the most releases of movies and TV shows

```
directors_produced_movies = movies %>%
    separate_rows(director, sep=",") %>%
    mutate(across(where(is.character), str_trim)) %>%
    filter(country!="United States")

top_directors <-  directors_produced_movies %>%
                    filter(!is.na(director)) %>%
                    group_by(director) %>%
                    count(sort=TRUE) %>%
                    head(n=10)

temp <- top_directors %>%
pull(director) %>%
sapply(function(d){
    directors_produced_movies %>% filter(director == d) %>% slice_max(listed_
})

director_top_genre = as.data.frame(temp)
final_table <- cbind(top_directors, genre=director_top_genre$temp)
final_table <- final_table %>% rename(total_movies_produced=n) %>% rename(top_
final_table = as.data.frame(final_table)
final_table
```

```
##          top_directors total_movies_produced
## 1            Jan Suter                    21
## 2          Raúl Campos                    19
## 3  Cathy Garcia-Molina                    13
## 4      Youssef Chahine                    12
## 5         David Dhawan                     9
## 6       Yılmaz Erdoğan                     9
## 7       Anurag Kashyap                     8
## 8          Hakan Algül                     8
## 9     Hanung Bramantyo                     8
## 10          Johnnie To                     8
##                                                genre
## 1                                  Stand-Up Comedy
## 2                                  Stand-Up Comedy
## 3              International Movies, Romantic Movies
## 4      Dramas, International Movies, Romantic Movies
## 5   Comedies, International Movies, Sci-Fi & Fantasy
## 6      Dramas, International Movies, Romantic Movies
## 7                    International Movies, Thrillers
## 8      Comedies, International Movies, Sports Movies
## 9      Dramas, International Movies, Romantic Movies
## 10                      Dramas, International Movies
```

Jan Suter made twenty-one stand-up comedy movies and Raúl Campos followed after that with nineteen movies of stand-up comedy genre. Cathy Garcia-Molina produced international movies mostly romantic movies. The top ten international movie directors produced mostly dramas, and comedies throughout 2008 to 2020. The following explores which cast members appeared in movied produced in United States between 2008-2021.

Cast Members United States

Top Cast Members that appeared in movies produced in United States between 2008-2021

1) Who are the top 10 cast members that appeared in movies produced in USA?
2) What genres of movies do the top cast members appear in?
3) What is number of movies for each genre that the top cast members appear in?
4) What is are the movie genres that the top movie stars appear in over the years?

Answers to those questions are below in tables and plots

```
usa_produced_movies = movies %>%
    separate_rows(cast, sep=",") %>%
    mutate(across(where(is.character), str_trim)) %>%
    filter(country=="United States")

top_cast <- usa_produced_movies %>%
                    filter(!is.na(cast)) %>%
                    distinct(show_id, .keep_all = TRUE) %>%
                    group_by(cast) %>%
                    count(sort=TRUE) %>%
                    head(n=10) %>%
                    rename(num_movies=n)

top_cast
```

```
## # A tibble: 10 x 2
## # Groups:   cast [10]
##    cast            num_movies
##    <chr>                <int>
##  1 Adam Sandler            20
##  2 Nicolas Cage            10
##  3 Kevin Hart               9
##  4 Eddie Murphy             8
##  5 Jeff Dunham              8
##  6 Will Smith               8
##  7 Harrison Ford            7
##  8 John Travolta            7
##  9 Kristen Stewart          7
## 10 Ben Stiller              6
```

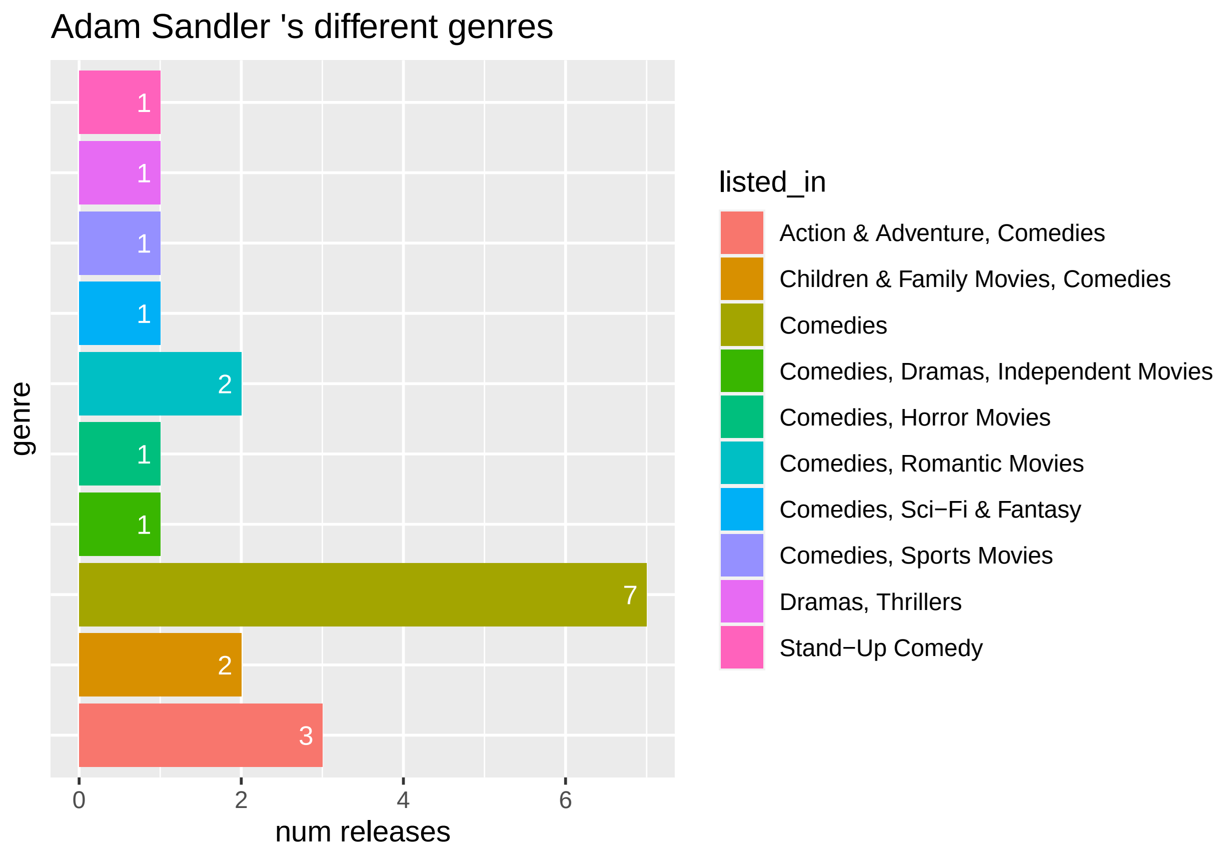
```
all_genres_top_cast<- usa_produced_movies %>%
                    filter(!is.na(cast)) %>%
                    filter(!is.na(listed_in)) %>%
                    mutate(across(where(is.character), str_trim))  %>%
                    group_by(cast)

# Get a count of each genre associated with cast member
get_genres_for_cast <- function(c) {
   count_genres_top_cast <- all_genres_top_cast %>%
    filter(cast == c) %>%
    group_by(listed_in) %>%
    count(sort=TRUE)

    return (count_genres_top_cast)
}

get_genres_with_years_for_cast <- function(c) {
   genre_year <- all_genres_top_cast %>%
    select(cast, listed_in, release_year)  %>%
    filter(cast == c) %>%
    group_by(release_year)

    return (
      genre_year[order(genre_year$release_year), ] %>%
        group_by(listed_in, release_year))
}
```

Adam Sandler, one of the movie stars in the United States, was in twenty movies in this time frame, and Nicolas Cage followed after that with ten movies. Kevin Hart were nine movies and so on. We will plot those three movie stars to understand their appearance in those movies and then we will analyze if they change their genre over time.

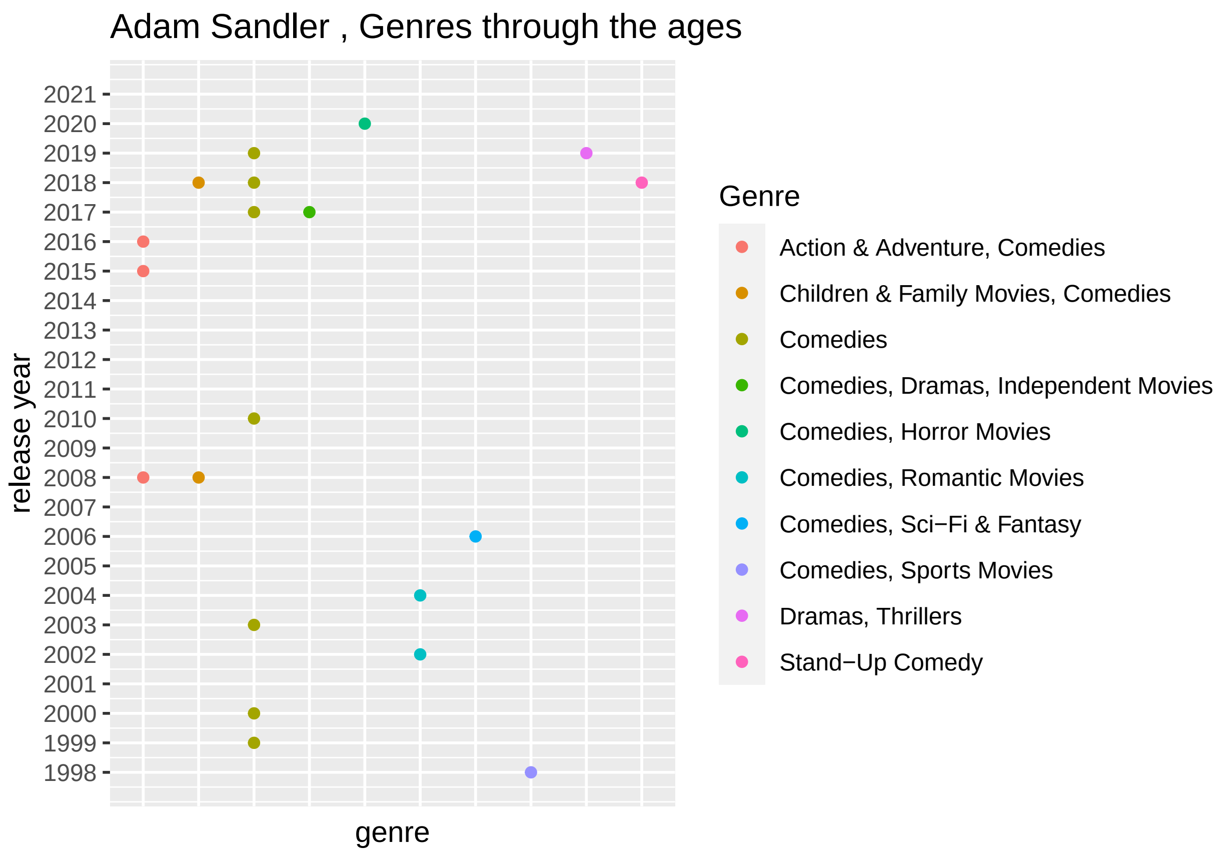
```r
plot_genre_box <- function(c) {
    data <- get_genres_for_cast(c)

    graph <- ggplot(data, aes(x=n, y=listed_in, fill=listed_in)) +
        geom_bar(stat="identity") +
        ggtitle(paste(c,"'s different genres")) +
        labs(x="num releases", y="genre", col = "Genre") +
        theme(axis.text.y=element_blank(),axis.ticks.y=element_blank())+
        geom_text(aes(label=n), hjust=1.6, color="white", size=3.5)

    graph
}
```

```r
plot_release_genre_points <- function(c, s, e) {
    data <- get_genres_with_years_for_cast(c)
    g <- ggplot(data, aes(x=listed_in, y=release_year, color=listed_in)) +
        ggtitle(paste(c, ", Genres through the ages")) +
        labs(x="genre", y="release year", col = "Genre") +
        geom_point() +
        scale_y_continuous(limits=c(s, e), breaks=seq(s,e,1)) +
        theme(axis.text.x=element_blank(),axis.ticks.x=element_blank())

    g
}
```

```r
cast_member="Adam Sandler"
plot_genre_box(cast_member)
```

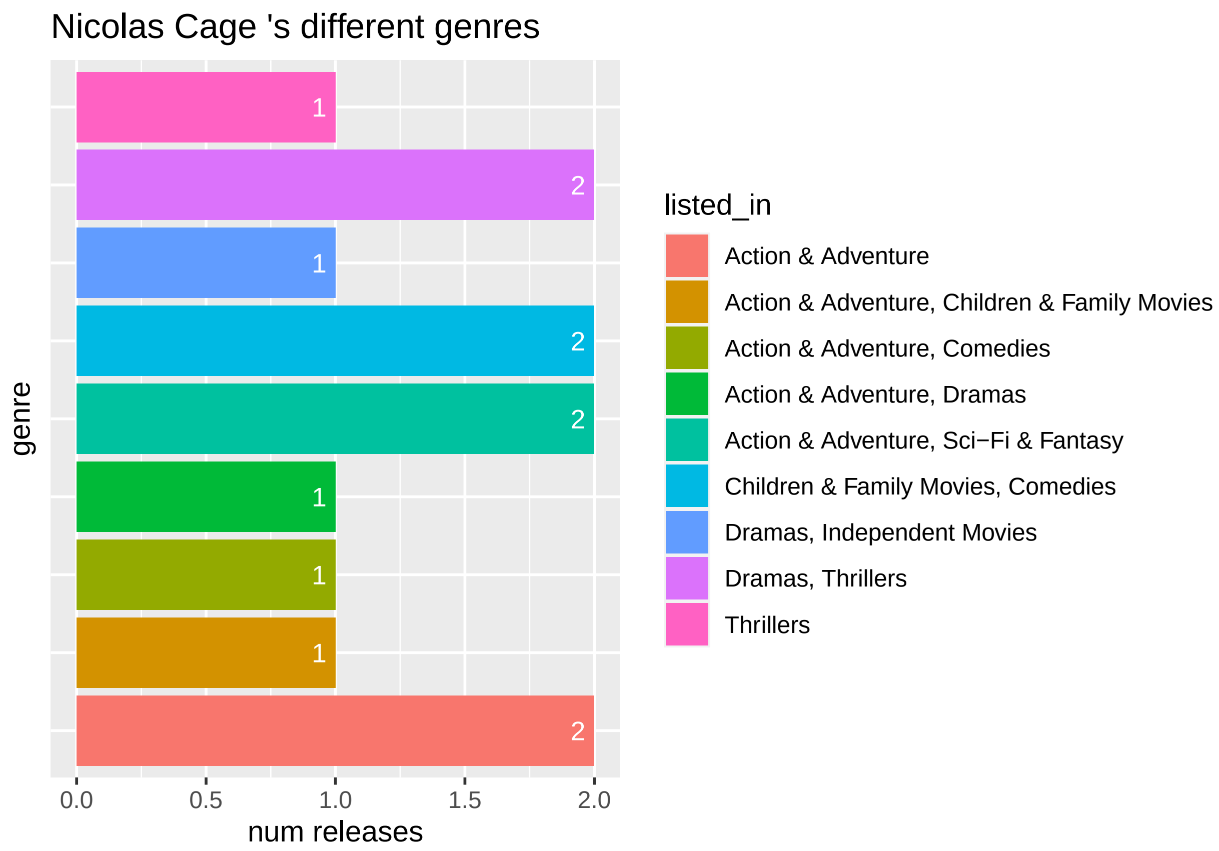


```
plot_release_genre_points(cast_member,1998, 2021)
```

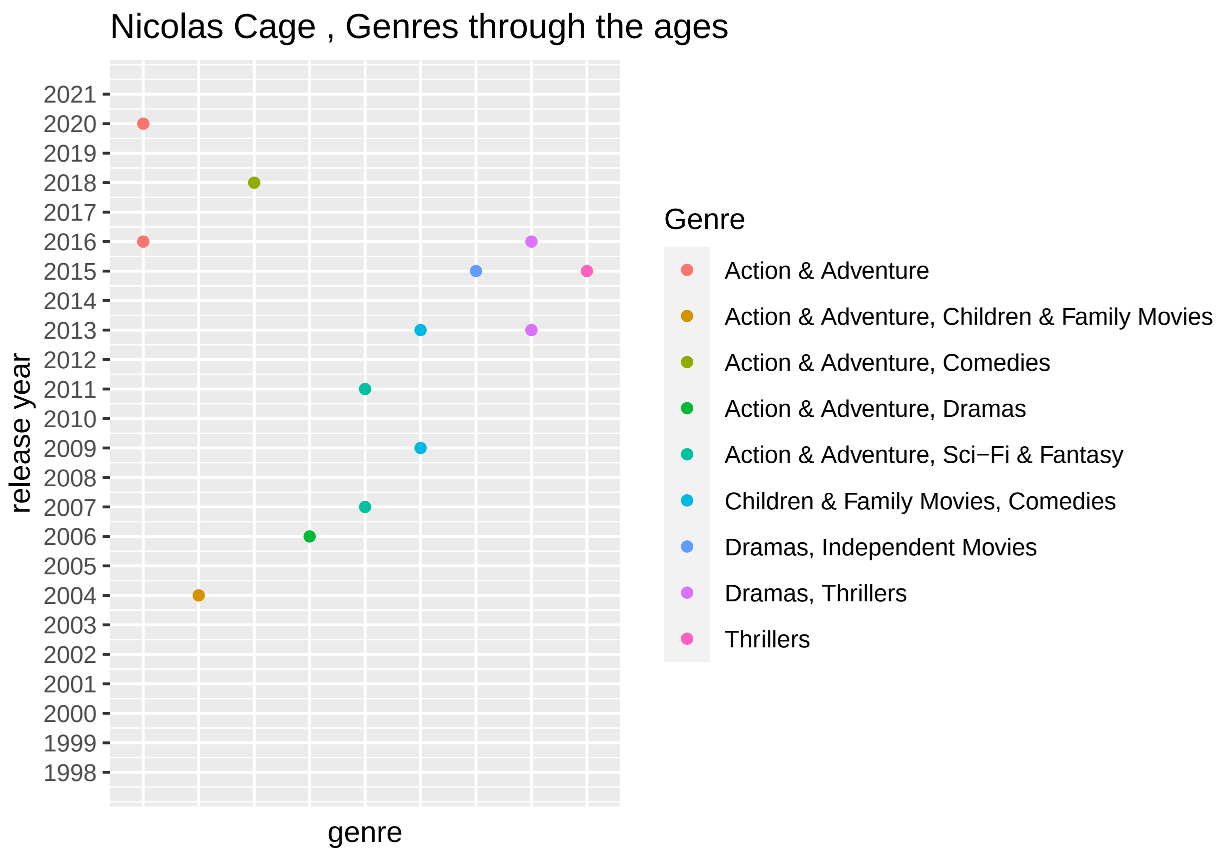


Adam Sandler first appeared in 1998 with a comedy genre movie and then he continued to appear many comedy movies after that. He did not seem to change his genre over time and the bar chart indicates that he was in thirteen comedy related movies.

```
cast_member="Nicolas Cage"
plot_genre_box(cast_member)
```

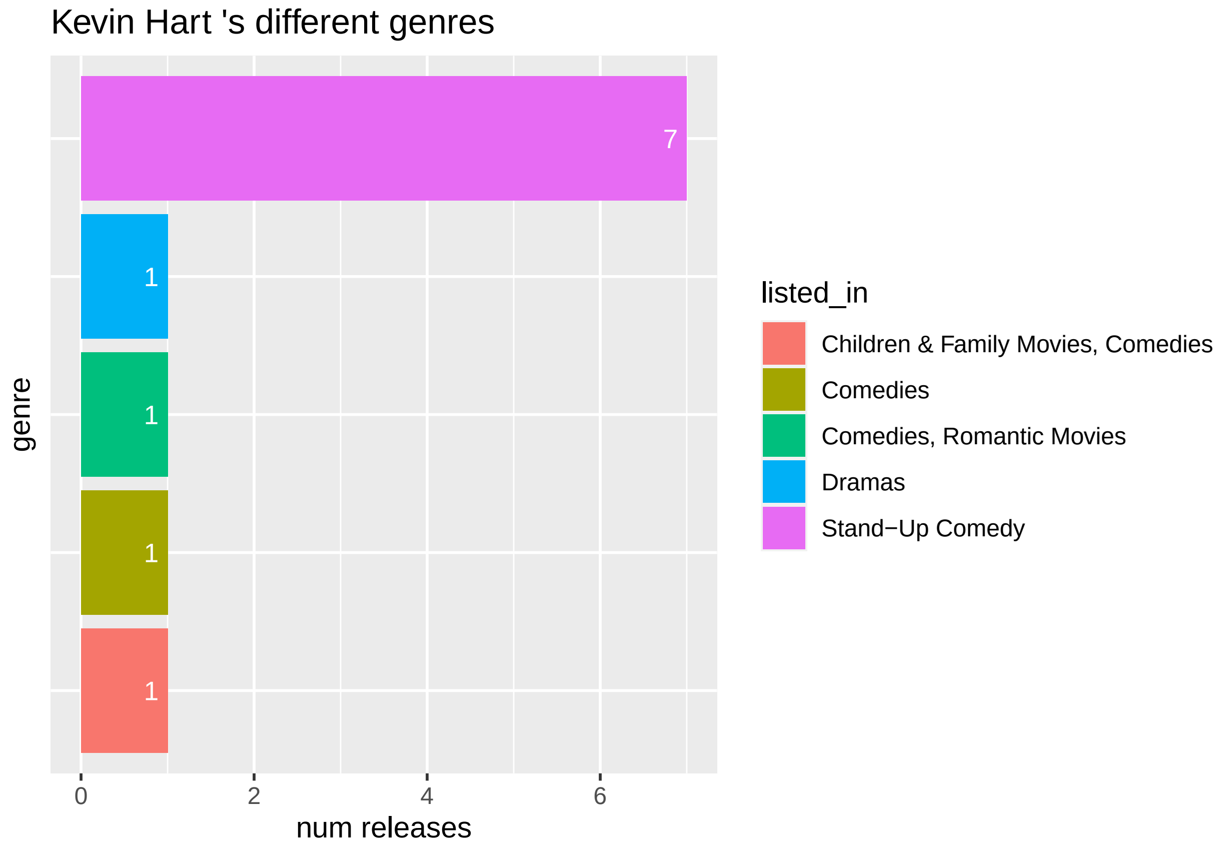


```
plot_release_genre_points(cast_member,1998, 2021)
```

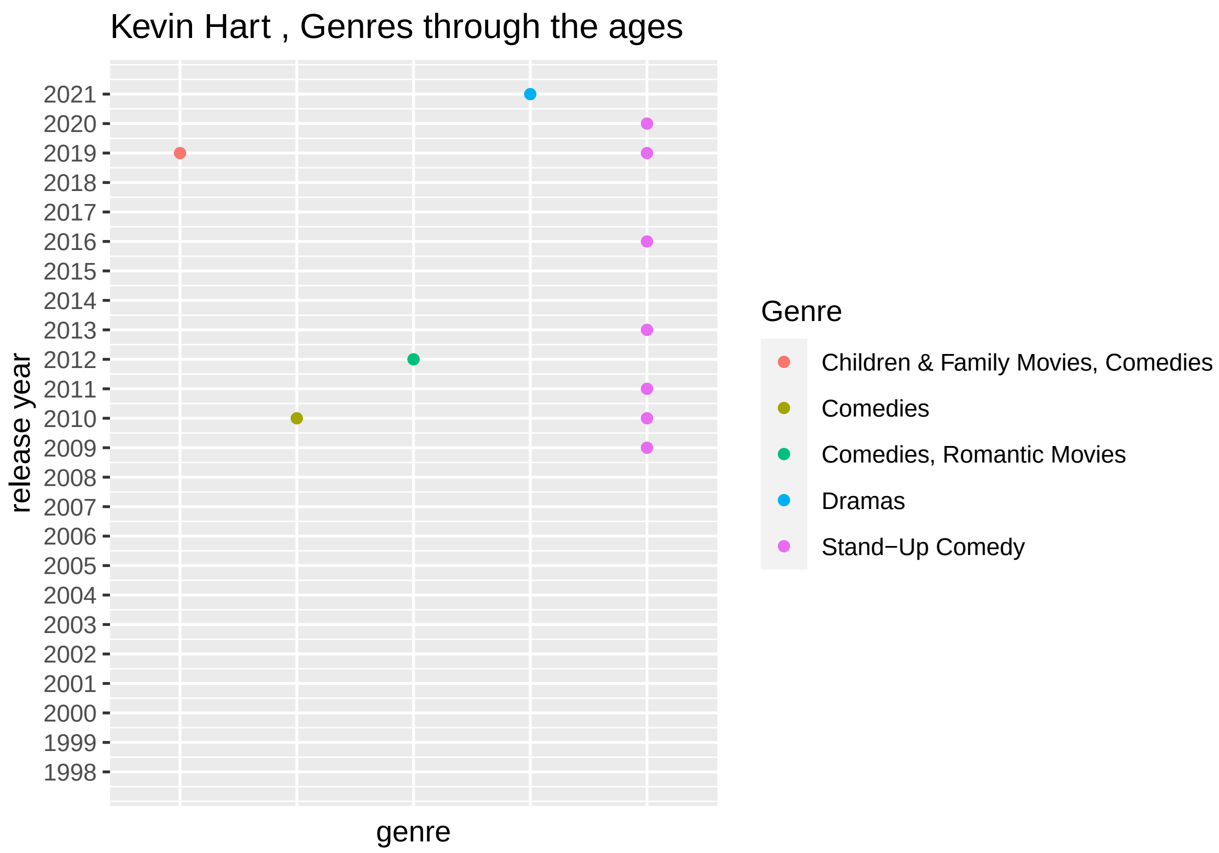


According to the dataset, Nicolas Cage was in many different genres through the years 2000s to 2020. He played mostly in action and adventure movies and his tendency of apperance changed from Dramas and Thriller to action later on. In 2004, he was in an action movie and then his latest appearance was in 2020 with an action and adventure movie.

```
cast_member="Kevin Hart"
plot_genre_box(cast_member)
```



```
plot_release_genre_points(cast_member,1998, 2021)
```



Kevin Hart did not change his appearance in a comedy movie in the time frame. The first appearance in 2009 in a stand-up comedy made his position as a great comedian in the history of the United States film industry. He appeared in other genres such as dramas and family movies, but his style of comedy did not change over time.

The following explores the same procedure, but with international cast members.

Cast Member International

```
intl_produced_movies = movies %>%
    separate_rows(cast, sep=",") %>%
    mutate(across(where(is.character), str_trim)) %>%
    filter(country!="United States")

top_cast_intl <- intl_produced_movies %>%
                    filter(!is.na(cast)) %>%
                    distinct(show_id, .keep_all = TRUE) %>%
                    group_by(cast) %>%
                    count(sort=TRUE) %>%
                    head(n=10) %>%
                    rename(num_movies=n)
```

```
top_cast_intl
```

```
## # A tibble: 10 x 2
## # Groups:   cast [10]
##    cast             num_movies
##    <chr>                 <int>
##  1 Shah Rukh Khan           25
##  2 Akshay Kumar             23
##  3 Amitabh Bachchan         20
##  4 Ajay Devgn               16
##  5 Aamir Khan               14
##  6 Ahmed Helmy              13
##  7 Anil Kapoor              13
##  8 Salman Khan              12
##  9 Naseeruddin Shah         11
## 10 Sanjay Dutt              11
```

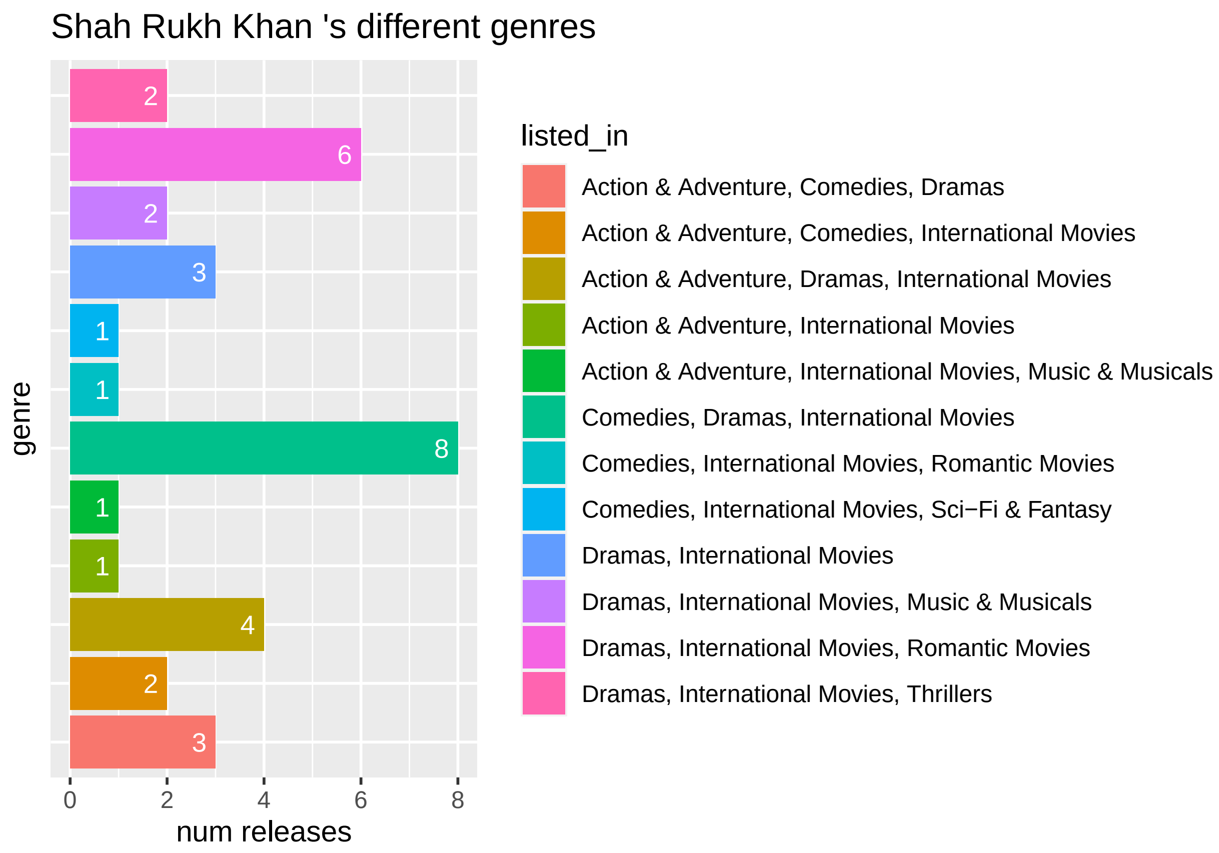
```
all_genres_top_cast_intl<- intl_produced_movies %>%
                    filter(!is.na(cast)) %>%
                    filter(!is.na(listed_in)) %>%
                    filter(!is.na(release_year)) %>%
                    mutate(across(where(is.character), str_trim))  %>%
                    group_by(cast)

get_genres_for_cast_intl <- function(c) {
   count_genres_top_cast_intl <- all_genres_top_cast_intl %>%
    filter(cast == c) %>%
    group_by(listed_in) %>%
    count(sort=TRUE)

    return (count_genres_top_cast_intl)
}

get_genres_with_years_for_cast_intl <- function(c) {
   genre_year_intl <- all_genres_top_cast_intl %>%
    select(cast, listed_in, release_year)  %>%
    filter(cast == c) %>%
    group_by(release_year)

    return (genre_year_intl[order(genre_year_intl$release_year), ] %>%
              group_by(listed_in, release_year))
}
```

The chart indicates that Shah Rukh Khan appeared in twenty five movies, Akshay Kumar appeared in twenty three movies, Amitabh Bachchan appeared in twenty movies, and so on. Most of them were Indian/Bollywood actors/actresses and that is due to the dataset contains second most data from Indian film industry. For each of those actors/actresses, the following shows the graphs of which genre of movies they appearted and how the genre changed over time.

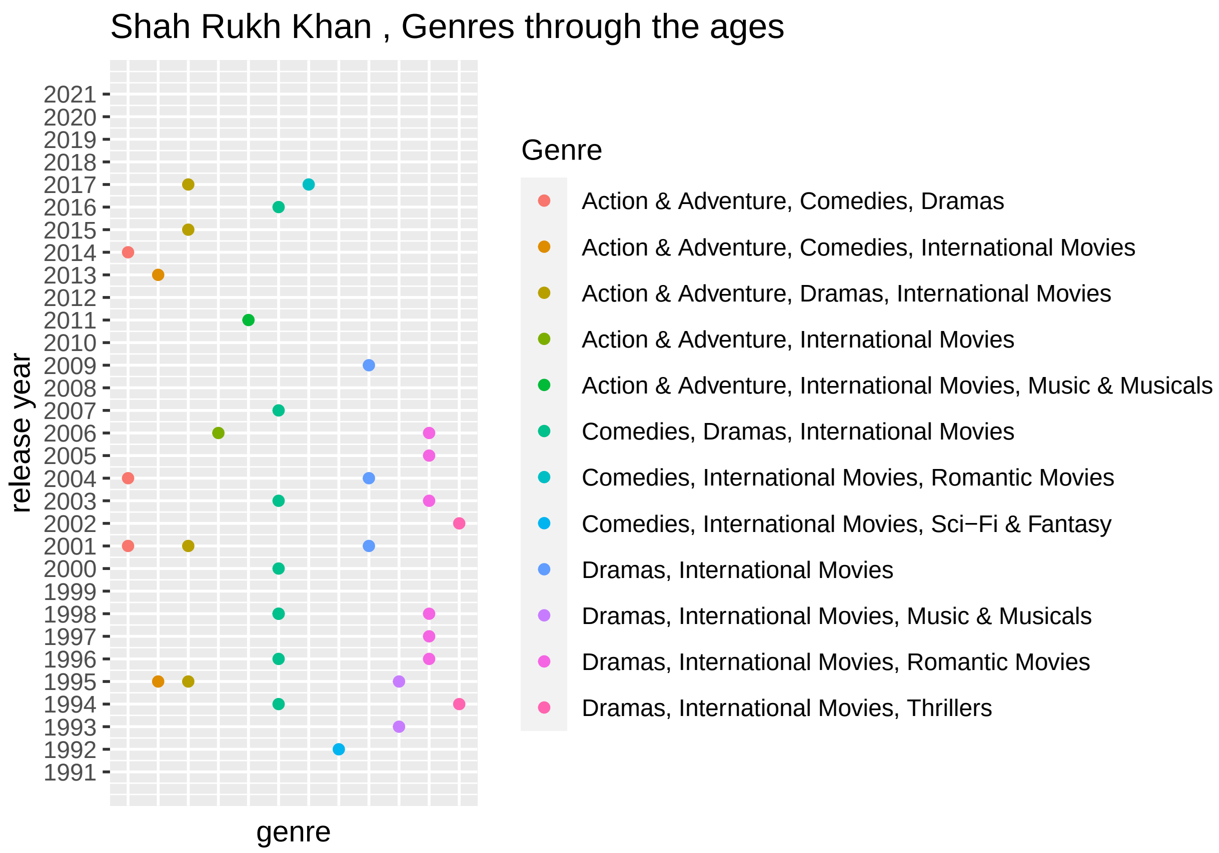
```r
plot_genre_box_intl <- function(c) {
    data <- get_genres_for_cast_intl(c)

    graph <- ggplot(data, aes(x=n, y=listed_in, fill=listed_in)) +
        geom_bar(stat="identity") +
        ggtitle(paste(c,"'s different genres")) +
        labs(x="num releases", y="genre", col = "Genre") +
        theme(axis.text.y=element_blank(),axis.ticks.y=element_blank())+
        geom_text(aes(label=n), hjust=1.6, color="white", size=3.5)

    graph
}
```

```r
plot_release_genre_points_intl <- function(c, s, e) {
    data <- get_genres_with_years_for_cast_intl(c)
    g <- ggplot(data, aes(x=listed_in, y=release_year, color=listed_in)) +
        ggtitle(paste(c, ", Genres through the ages")) +
        labs(x="genre", y="release year", col = "Genre") +
        geom_point() +
        scale_y_continuous(limits=c(s, e), breaks=seq(s,e,1)) +
        theme(axis.text.x=element_blank(),axis.ticks.x=element_blank())

    g
}
```

```r
cast_member="Shah Rukh Khan"
plot_genre_box_intl(cast_member)
```

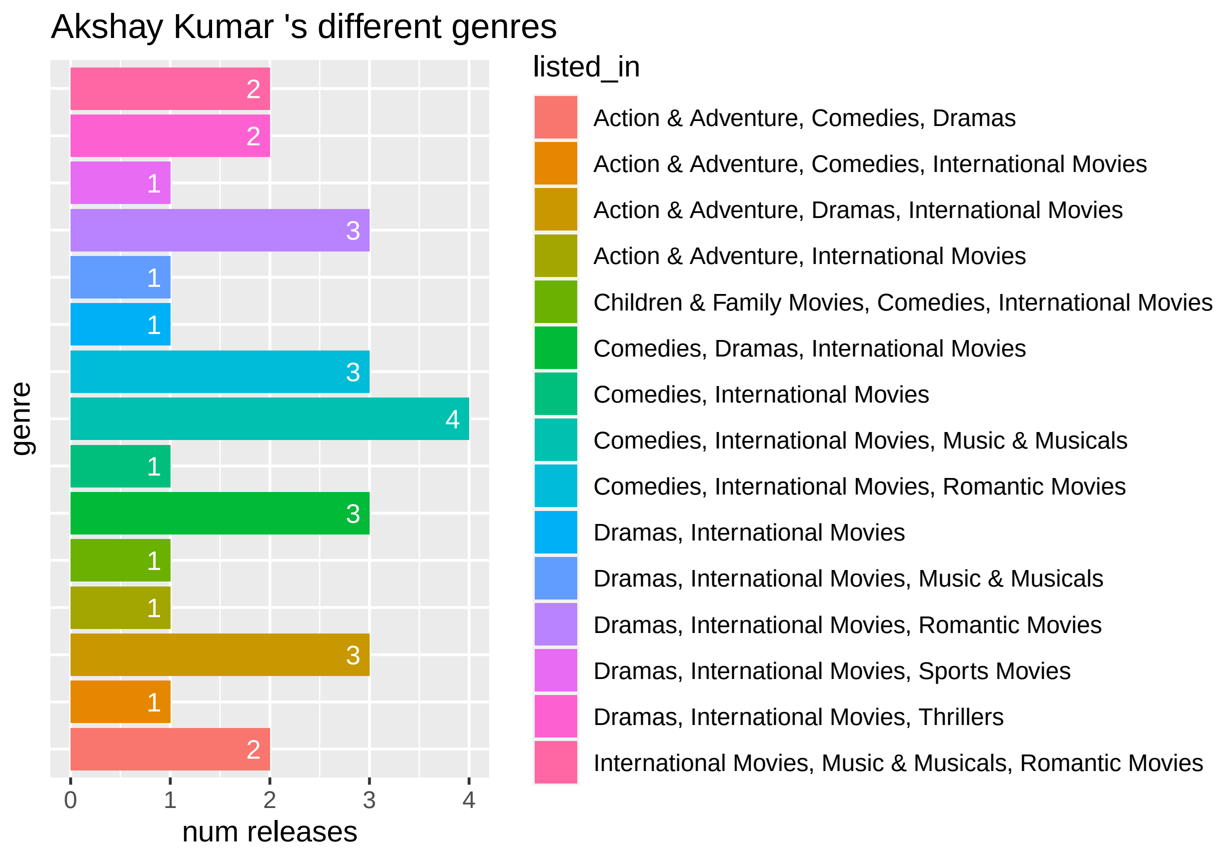


```
plot_release_genre_points_intl(cast_member, 1991, 2021)
```

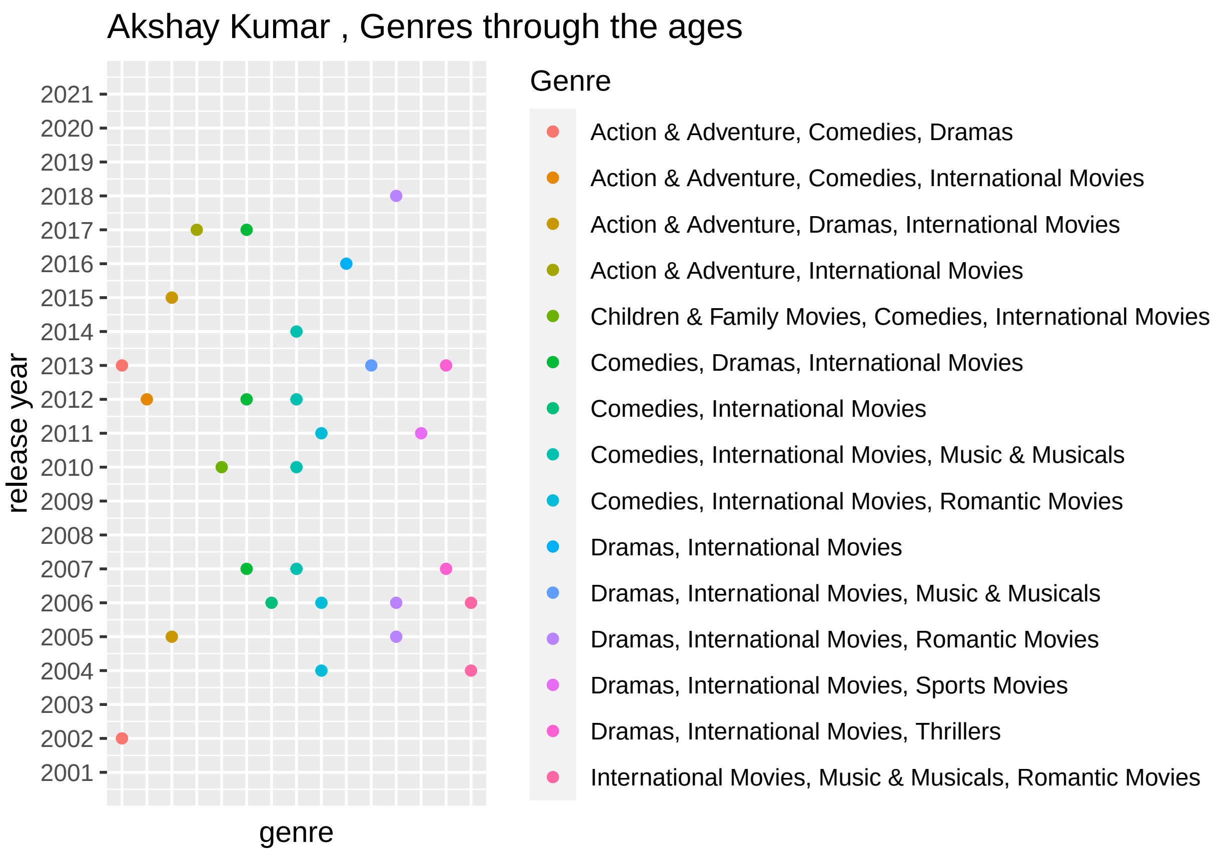


Shah Rukh Khan was in many different movie genres. First appeared in 1992, in a comedy related movie and the latest appearance was in action and adventure movie. The cast member has a wide variety of genre selection in 1992 to 2020 time frame.

```
cast_member="Akshay Kumar"
plot_genre_box_intl(cast_member)
```

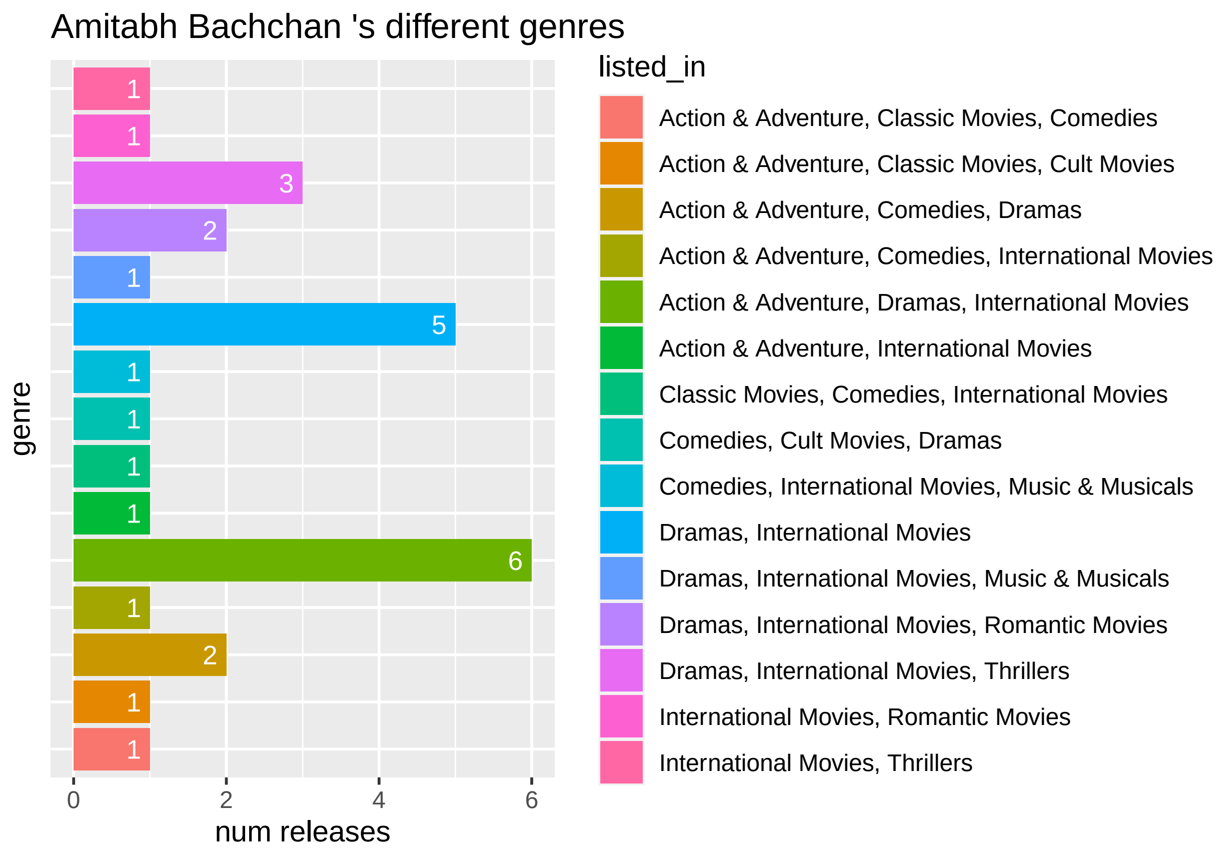


```
plot_release_genre_points_intl(cast_member, 2001, 2021)
```

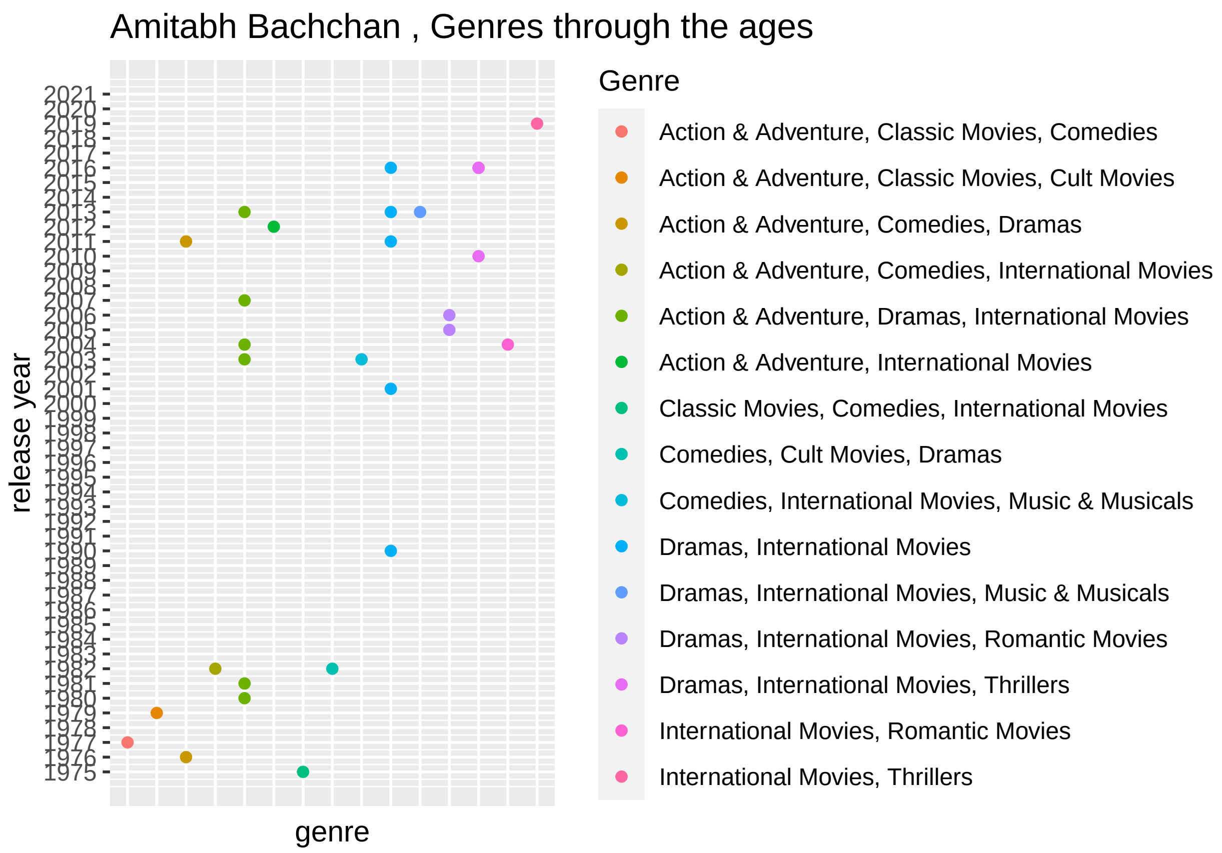


Akshay Kumar was the second most appeared cast member from the chart and the selection seemed to be very diverse in terms of genres of movies the cast member appeared in. First appeared in 2002 and the latest movie 2018 indicates that the cast member was in action and adventure movies. The genres seemed to be changed over time.

```
cast_member="Amitabh Bachchan"
plot_genre_box_intl(cast_member)
```



```
plot_release_genre_points_intl(cast_member, 1975, 2021)
```



Amitabh Bachchan is the cast member that appeared from 1975 to 2019 and probably the cast member that has been in the film industry for the longest in internationally according to the dataset. The genre that the cast member started was comedy and later on the cast member switched to other genres such as action or dramas. The wide variety of genres probably made the cast member to appear many movies for a long period of time.

Future Work

Given more time we would work to develop or improve on several ideas. Further work could also have been done with applying deep learning to this problem. A few viable methodologies might be to use techniques for unsupervised learning from the numerical data and maybe used in other neural networks to perform classification.

Conclusion

To conclude, we have constructed a relatively accurate data analysis to determine the genre of a given TV show and movie. Processing and narrowing down the features of the Netflix Dataset by identifying the top ten countries with top five genres the audience watch and feeding this data to figure out the trend over the years, we were able to identify which actors/actresses by which directors were popular in each country. Steps such as expanding the dataset and feature set and using logistic regression might have the potential to improve upon on results in the future.

Team Membership and Attestation of Work

Seongwoo Choi, Vybhab Achar Bhargav, and David Haddad agree to participate and work on this project with the understanding of the project.

Video Presentation Link

https://ucdavis.zoom.us/rec/share/upsU8l_s89iJ5lLC2Cz6a5OTJJZyz9W9uIUTd8rwnnWLOi41oG_WnIfrnGhTRQfK.Sgt0OhacFnJu6zyT

Reference

https://www.kaggle.com/shivanirana63/netflix-eda-movie-recommendation-system/data

https://www.kaggle.com/shivamb/netflix-shows

https://github.com/yihui/knitr-examples/blob/master/077-wrap-output.Rmd (for output formatting)