|  |  |
| --- | --- |
| **Netflix Movies and TV Shows** | **TEAM:**  M. Sri Vedahitha (004)  m. s tarun (005)  p. reethu akanksha (062)  m. prasad (098)  **bba business analytics**  **(3rd Semesteer)**  **2022-23** |

**1.INTRODUCTION TO DATA SET**

**Netflix Movies and TV Shows**

About this Dataset:

[*Netflix*](https://en.wikipedia.org/wiki/Netflix)*is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally.*

*This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.*

**Interesting Task Ideas**

1. Understanding what content is available in different countries
2. Identifying similar content by matching text-based features
3. Network analysis of Actors / Directors and find interesting insights
4. Does Netflix have more focus on TV Shows than movies in recent years.

**Questions/Problems**

**Analysis tasks**

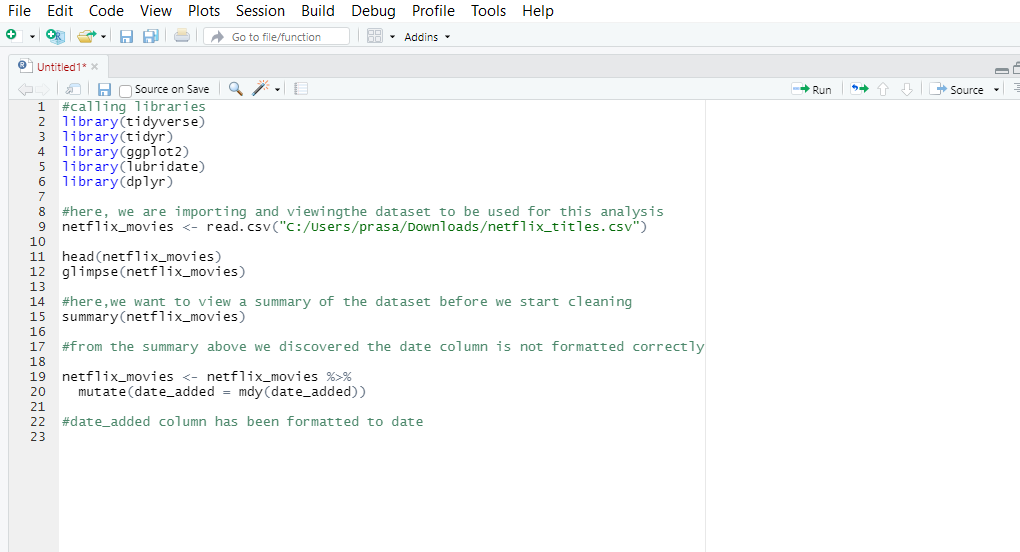
This is an analysis of the movies and TV Shows on the streaming platform "Netflix". In this analysis I will try to answer the following questions;

1. Understanding what content is available in different countries.
2. What year had the most Movies and TV show releases in Netflix.
3. Network analysis of Actors / Directors and find interesting insights.
4. Does Netflix have more focus on TV Shows than movies in recent years.
5. Has Duration of movies over the years increased or decreased.
6. Compare ratings of TV Shows and Movies and determine how they are rated.

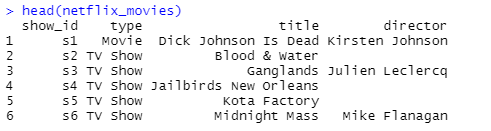
**Statistical techniques required for Analysis**

**DATA PREPARATION:**

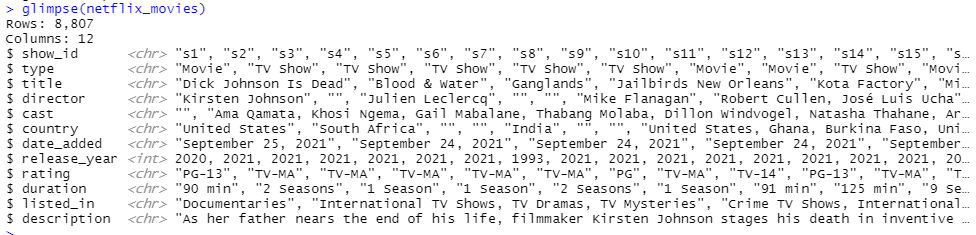
Here we imported the netflix dataset and named it netflix. We first viewed the data its summary and we used str() in R language to identify the internal structure of the objects in the data set



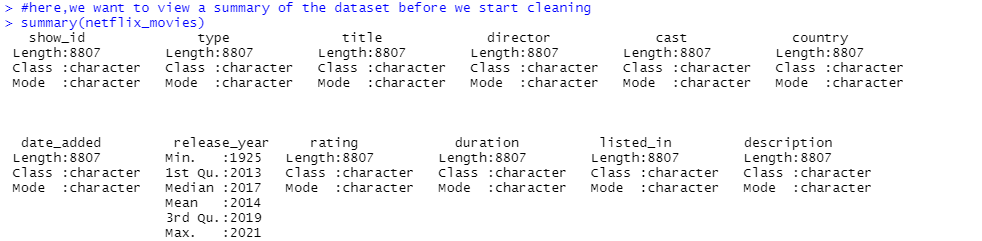
UNDERSTANDING THE DATA:



Glimpse of data:



SUMMARY OF DATA



**Summarisation of data:**

Exploratory Data Analysis of the movies and Tv shows on the streaming platform netflix.

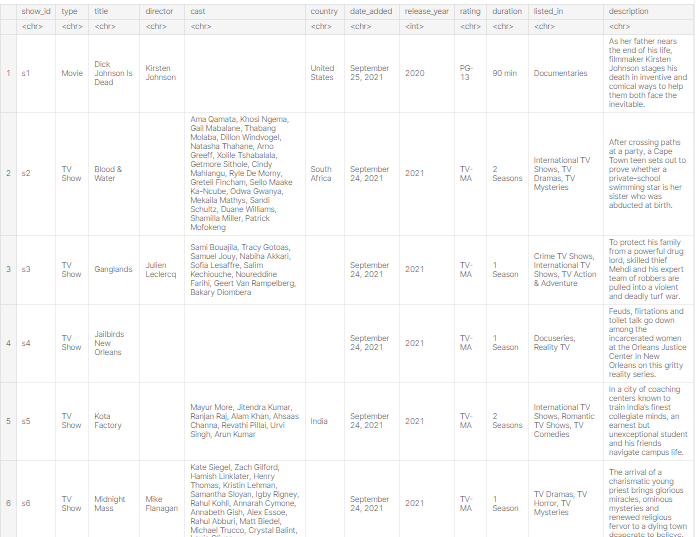
Industry : Streaming platform

Dimensions: 8807 rows and 12 columns

Missing Data / Null Values: No

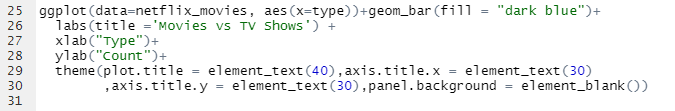
Duplicated Values: No

Structure: Data frame



**Analysis and Insights**

**1.Let us determine whether Netflix has been more focused on TV shows or movies.**



**CODE:**

ggplot(data=netflix\_movies, aes(x=type))+geom\_bar(fill = "dark blue")+

labs(title ='Movies vs TV Shows') +

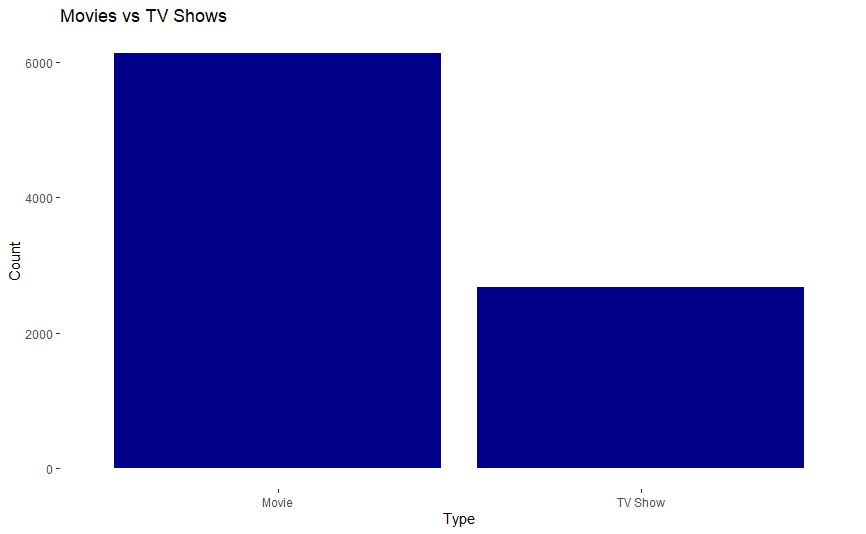
xlab("Type")+

ylab("Count")+

theme(plot.title = element\_text(40),axis.title.x = element\_text(30)

,axis.title.y = element\_text(30),panel.background = element\_blank())

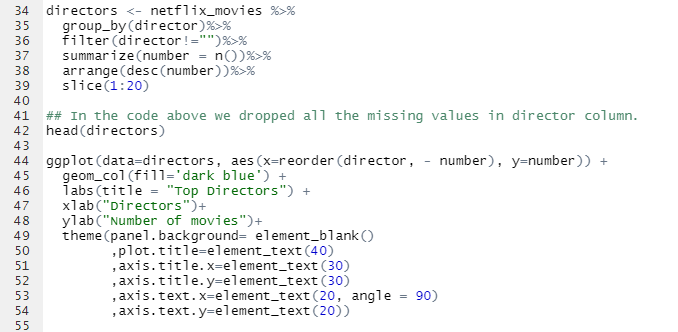
**Visualization:**

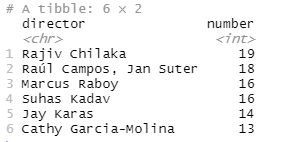


Insights:

* Movies have highest viewing count (i.e 6000)
* Shows viewing count is lesser than movies (i.e 2500)
* Movies have more priority than tv shows in netflix
* **Netflix is clearly more focused on MOVIES**

**2.Determine who are the TOP DIRECTORS for netflx movies and TV shows**





**CODE:**

directors <- netflix\_movies %>%

group\_by(director)%>%

filter(director!="")%>%

summarize(number = n())%>%

arrange(desc(number))%>%

slice(1:20)

*## In the code above we dropped all the missing values in director column.*

head(directors)

ggplot(data=directors, aes(x=reorder(director, - number), y=number)) +

geom\_col(fill='dark blue') +

labs(title = "Top Directors") +

xlab("Directors")+

ylab("Number of movies")+

theme(panel.background= element\_blank()

,plot.title=element\_text(40)

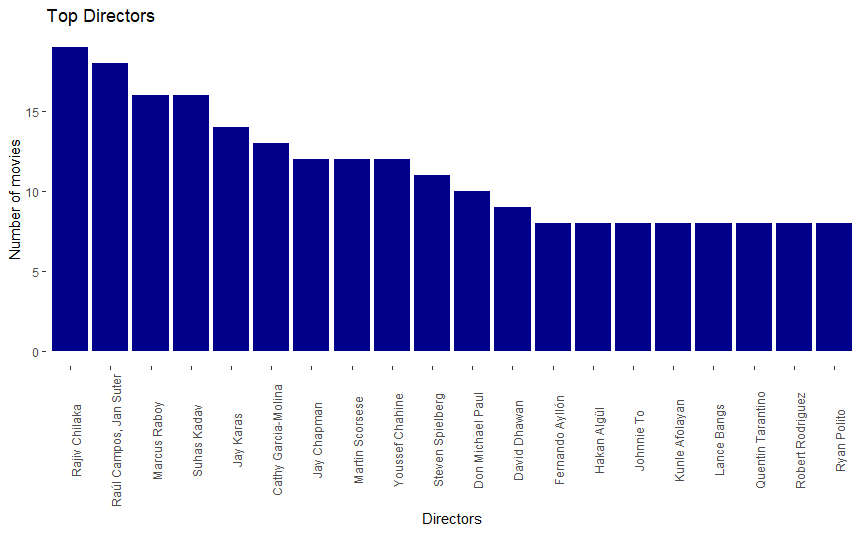
,axis.title.x=element\_text(30)

,axis.title.y=element\_text(30)

,axis.text.x=element\_text(20, angle = 90)

,axis.text.y=element\_text(20))

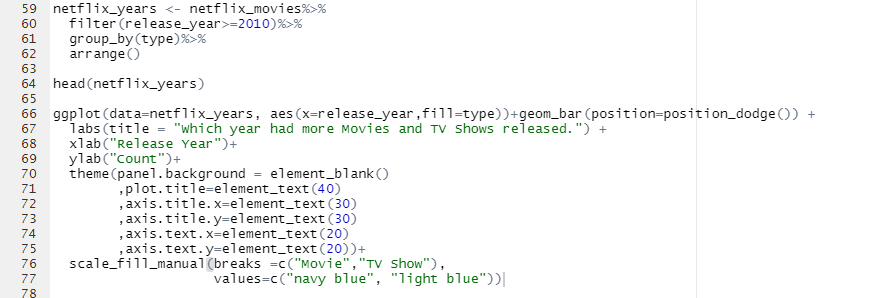
**Visualization:**

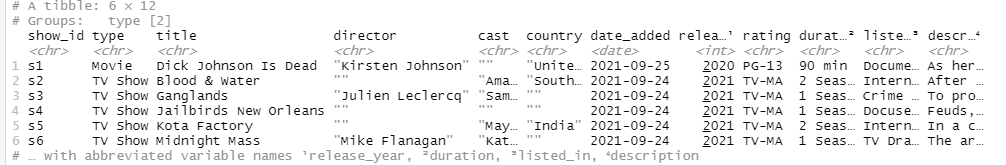


**Insights:**

* Rajiv chilaka is top director with over 19 movies
* Raul campos, jan suter are the second most director with 18 movies
* Ryan polito is the director with the least movies 9 movies

**3. Determine which year(in recent years) had the most releases on Netflix.**



****

**CODE:**

netflix\_years <- netflix\_movies%>%

filter(release\_year>=2010)%>%

group\_by(type)%>%

arrange()

head(netflix\_years)

ggplot(data=netflix\_years, aes(x=release\_year,fill=type))+geom\_bar(position=position\_dodge()) +

labs(title = "Which year had more Movies and TV Shows released.") +

xlab("Release Year")+

ylab("Count")+

theme(panel.background = element\_blank()

,plot.title=element\_text(40)

,axis.title.x=element\_text(30)

,axis.title.y=element\_text(30)

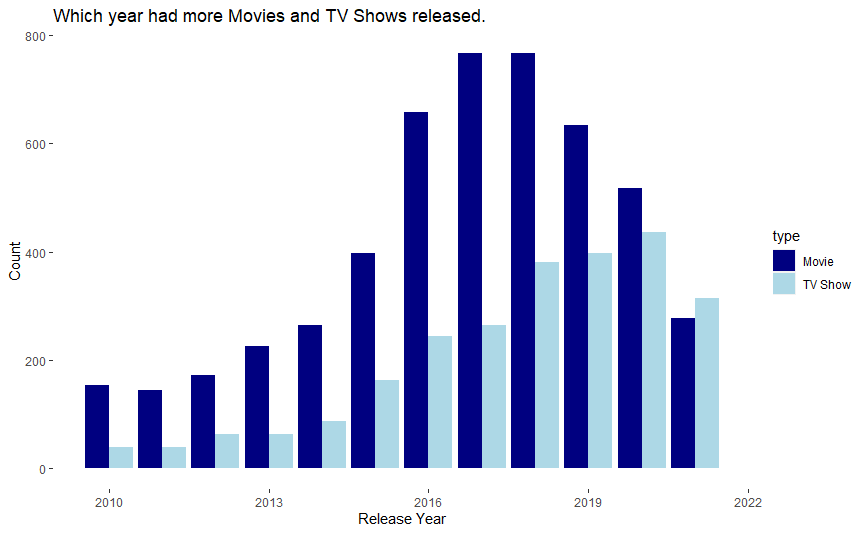
,axis.text.x=element\_text(20)

,axis.text.y=element\_text(20))+

scale\_fill\_manual(breaks =c("Movie","TV Show"),

values=c("navy blue", "light blue"))

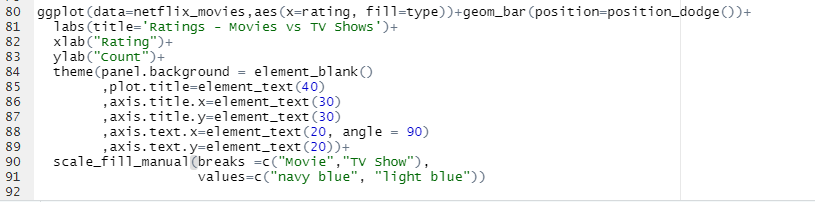
**Visualization:**



**Insights:**

* **The years with the most movie releases are 2017 and 2018, while 2020 has the most television show launches.**
* **The year 2011 is the year with least movie releases and tv shows.**
* **The year 2010 is the second least year with less releases.**

**4. Compare ratings of TV Shows and Movies and determine how they are rated.**



**CODE:**

ggplot(data=netflix\_movies,aes(x=rating, fill=type))+geom\_bar(position=position\_dodge())+

labs(title='Ratings - Movies vs TV Shows')+

xlab("Rating")+

ylab("Count")+

theme(panel.background = element\_blank()

,plot.title=element\_text(40)

,axis.title.x=element\_text(30)

,axis.title.y=element\_text(30)

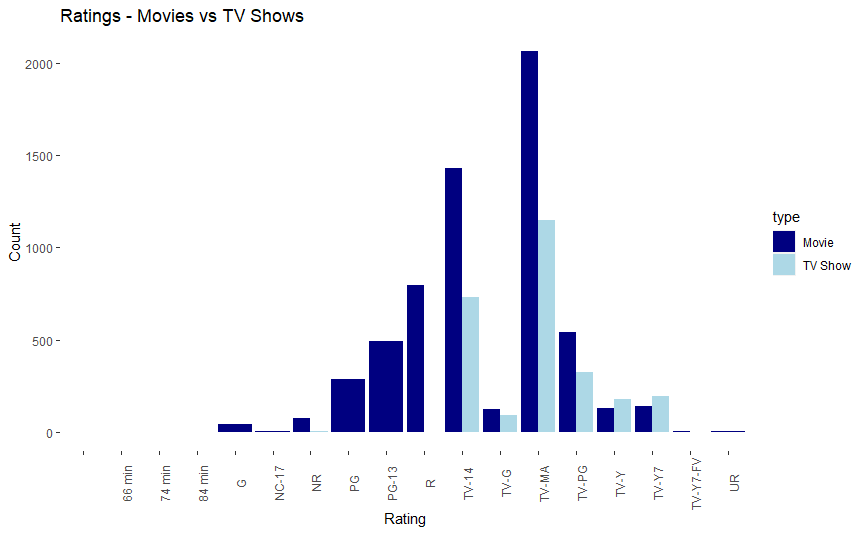
,axis.text.x=element\_text(20, angle = 90)

,axis.text.y=element\_text(20))+

scale\_fill\_manual(breaks =c("Movie","TV Show"),

values=c("navy blue", "light blue"))

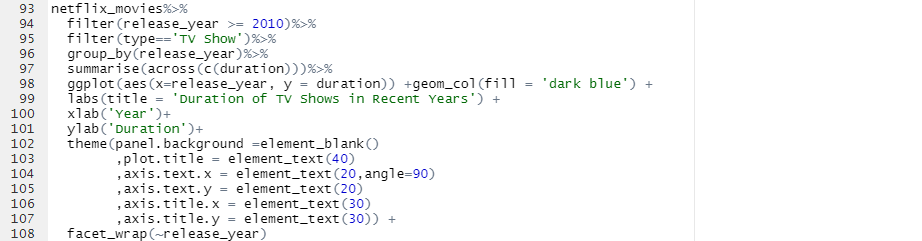
**Visualization:**



**Insights:**

* TV-MA has the highest rating for movies and tv shows.
* TV-14 has second highest rating for movies and tv shows.
* **We can tell from the plot above that the majority of Netflix movies and TV Shows are geared toward mature audiences, with movies and episodes not suitable for children under 14 coming in second**

**5.Has Duration of Tv shows over the years increased or decreased.**

****

**CODE:**

netflix\_movies%>%

filter(release\_year >= 2010)%>%

filter(type=='TV Show')%>%

group\_by(release\_year)%>%

summarise(across(c(duration)))%>%

ggplot(aes(x=release\_year, y = duration)) +geom\_col(fill = 'dark blue') +

labs(title = 'Duration of TV Shows in Recent Years') +

xlab('Year')+

ylab('Duration')+

theme(panel.background =element\_blank()

,plot.title = element\_text(40)

,axis.text.x = element\_text(20,angle=90)

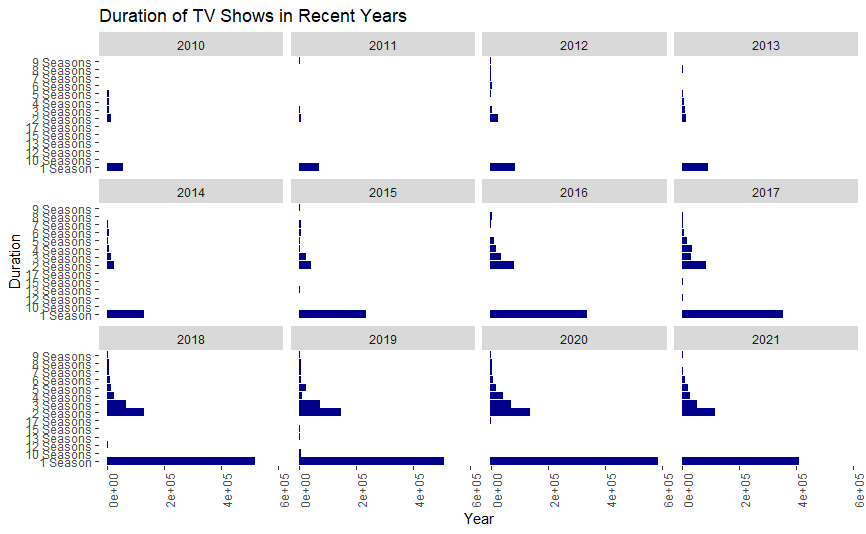
,axis.text.y = element\_text(20)

,axis.title.x = element\_text(30)

,axis.title.y = element\_text(30)) +

facet\_wrap(~release\_year)

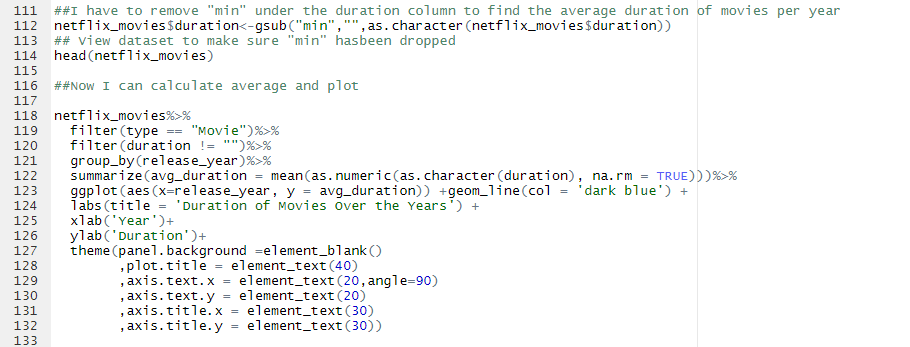
**Visualization:**



**Insights:**

* The majority of TV Shows on Netflix have reduced to one season in recent years

**6.Has Duration of movies over the years increased or decreased.**

****

**CODE:**

##I have to remove "min" under the duration column to find the average duration of movies per year

netflix\_movies$duration<-gsub("min","",as.character(netflix\_movies$duration))

## View dataset to make sure "min" hasbeen dropped

head(netflix\_movies)

##Now I can calculate average and plot

netflix\_movies%>%

filter(type == "Movie")%>%

filter(duration != "")%>%

group\_by(release\_year)%>%

summarize(avg\_duration = mean(as.numeric(as.character(duration), na.rm = TRUE)))%>%

ggplot(aes(x=release\_year, y = avg\_duration)) +geom\_line(col = 'dark blue') +

labs(title = 'Duration of Movies Over the Years') +

xlab('Year')+

ylab('Duration')+

theme(panel.background =element\_blank()

,plot.title = element\_text(40)

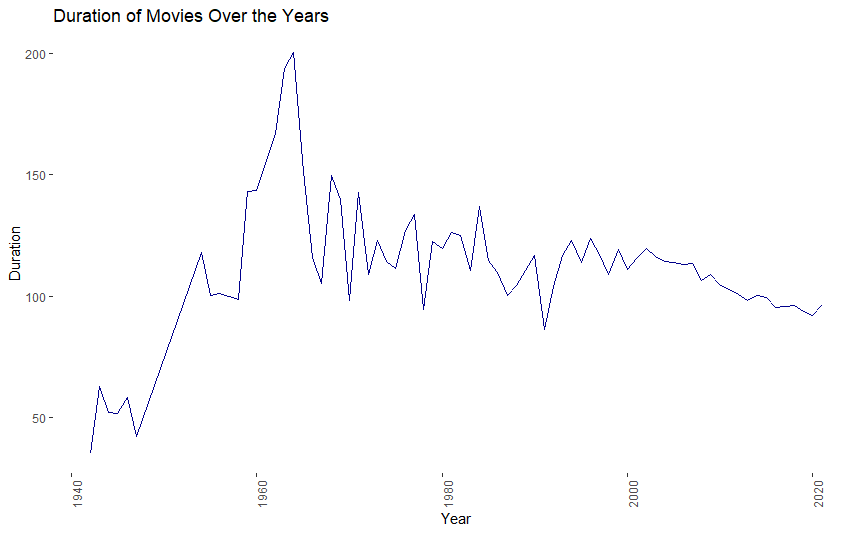
,axis.text.x = element\_text(20,angle=90)

,axis.text.y = element\_text(20)

,axis.title.x = element\_text(30)

,axis.title.y = element\_text(30))

**Visualization:**



**Insights:**

* **The plot above makes it quite evident that while Netflix movie lengths first climed, the duration has considerably decreased over time.**