Assignment 0

2025-01-31

```
summary(netflix_data)
```

```
##    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        Length:8807        Min.   :1925
##  Class :character   Class :character   Class :character   1st Qu.:2013
##  Mode  :character   Mode  :character   Mode  :character   Median :2017
##                                                           Mean   :2014
##                                                           3rd Qu.:2019
##                                                           Max.   :2021
##     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
##
##
##
```

```
glimpse(netflix_data)
```

```
## Rows: 8,807
## Columns: 12
## $ show_id      <chr> "s1", "s2", "s3", "s4", "s5", "s6", "s7", "s8", "s9", "s1~
## $ type         <chr> "Movie", "TV Show", "TV Show", "TV Show", "TV Show", "TV ~
## $ title        <chr> "Dick Johnson Is Dead", "Blood & Water", "Ganglands", "Ja~
## $ director     <chr> "Kirsten Johnson", NA, "Julien Leclercq", NA, NA, "Mike F~
## $ cast         <chr> NA, "Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Mola~
## $ country      <chr> "United States", "South Africa", NA, NA, "India", NA, NA,~
## $ date_added   <chr> "September 25, 2021", "September 24, 2021", "September 24~
## $ release_year <dbl> 2020, 2021, 2021, 2021, 2021, 2021, 2021, 1993, 2021, 202~
## $ rating       <chr> "PG-13", "TV-MA", "TV-MA", "TV-MA", "TV-MA", "TV-MA", "PG~
## $ duration     <chr> "90 min", "2 Seasons", "1 Season", "1 Season", "2 Seasons~
## $ listed_in    <chr> "Documentaries", "International TV Shows, TV Dramas, TV M~
## $ description  <chr> "As her father nears the end of his life, filmmaker Kirst~
```

we can see our data set has 8 thousand rows, including 12 columns, all of this describing the type, title, director, cast, etc of a movie

```r
colSums(is.na(netflix_data))
```

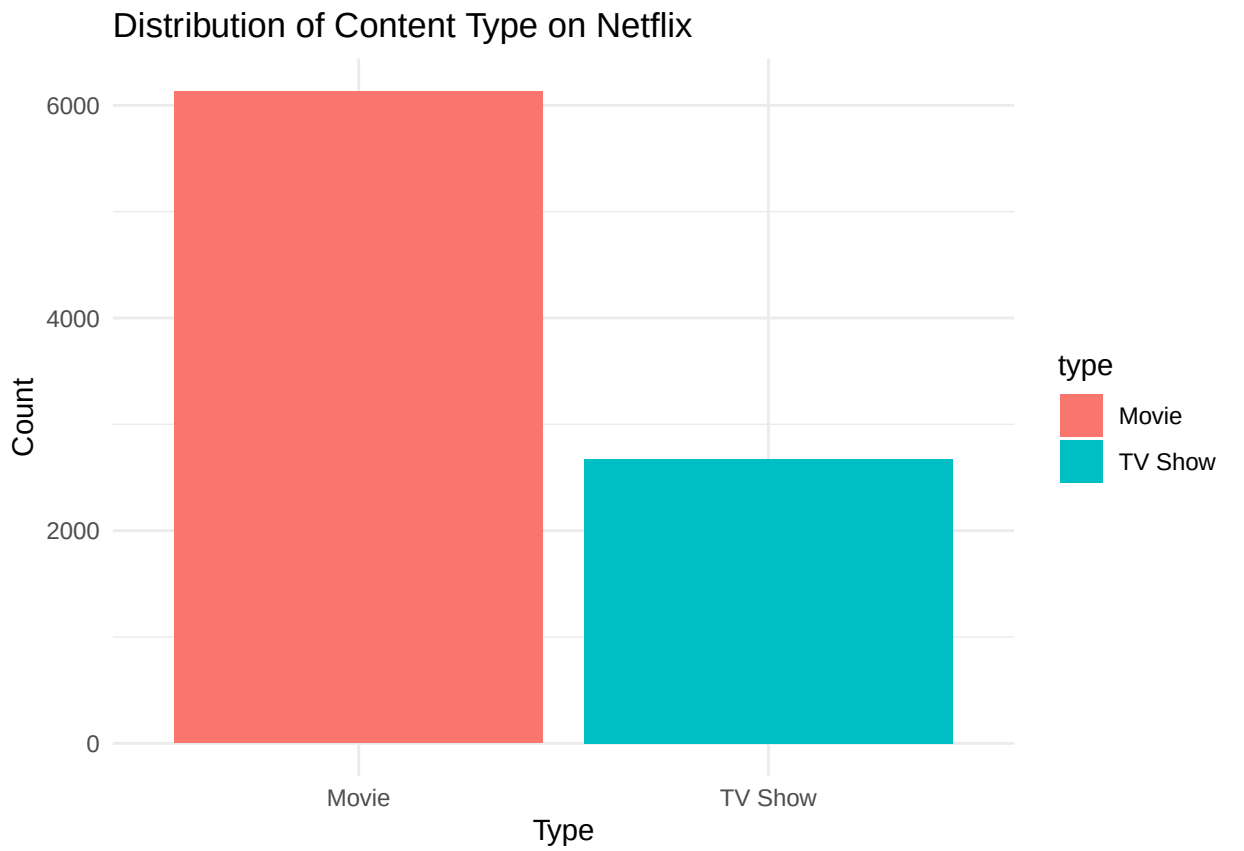
```
##      show_id         type        title     director         cast      country
##            0            0            0         2634          825          831
##   date_added release_year       rating     duration    listed_in  description
##           10            0            4            3            0            0
```

checking to look for NA's in our data set

```r
netflix_data$director[is.na(netflix_data$director)] <- "none"
netflix_data$cast[is.na(netflix_data$cast)] <- "none"
netflix_data$country[is.na(netflix_data$country)] <- "none"
```

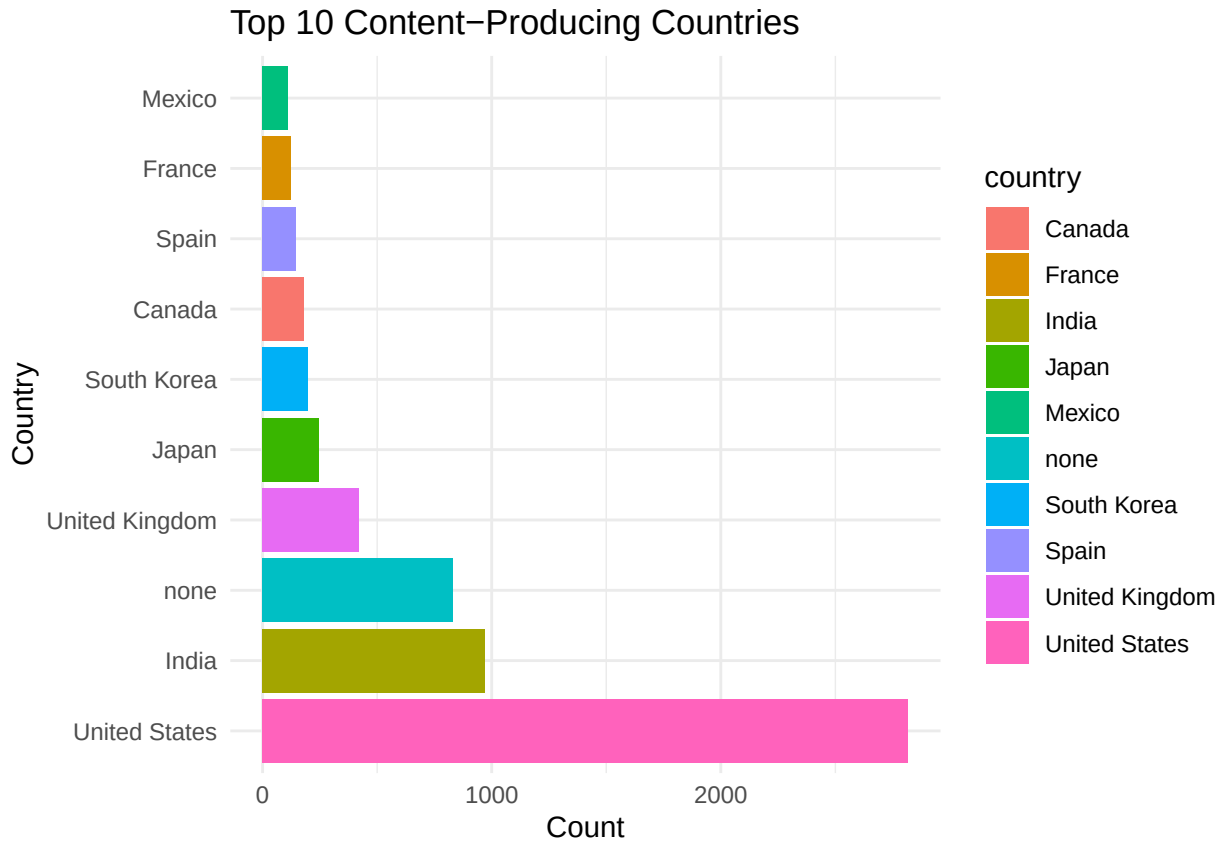
fill in those NA's with none values

```r
ggplot(netflix_data, aes(x = type, fill = type)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Distribution of Content Type on Netflix",
       x = "Type", y = "Count")
```



Creating a ggplot to see which type of Netflix content is movies or tv-shows where we can see the majority of our data sets are movies, more than tv-shows made

```r
netflix_data %>%
  group_by(country) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(country, -count), y = count, fill = country)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Top 10 Content-Producing Countries", x = "Country", y = "Count")
```



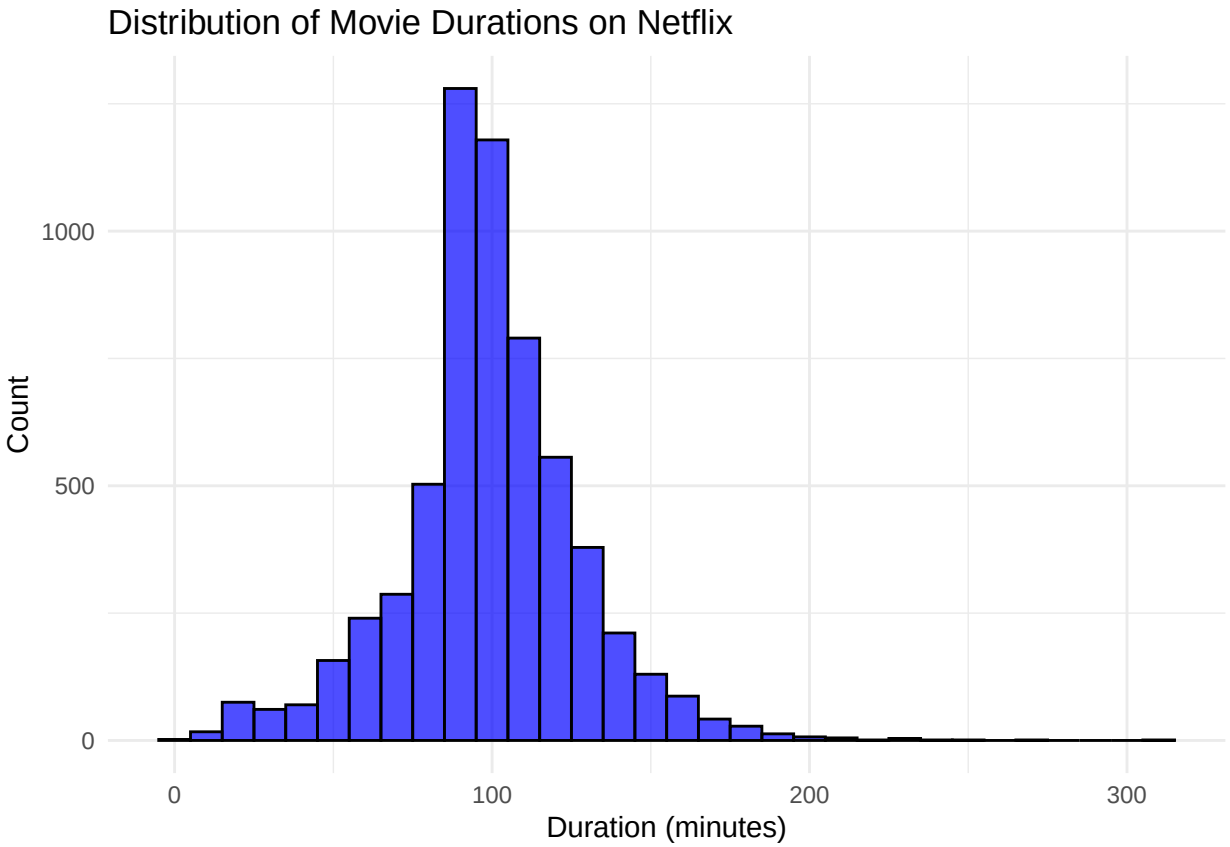
we notice our top countries with the most content being generated being the US, India, UK, Japan, South Korea, Canada, Spain, France and Mexico. I notice we still have some Na's to fill but these are our top countries making the most content for Netflix so far.

```r
netflix_data <- netflix_data %>%
  filter(type == "Movie") %>%
  mutate(duration_num = as.numeric(gsub(" min", "", duration)))

ggplot(netflix_data, aes(x = duration_num)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black", alpha = 0.7) +
  theme_minimal() +
  labs(title = "Distribution of Movie Durations on Netflix",
       x = "Duration (minutes)", y = "Count")
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
```

```
## (`stat_bin()`).
```



After converting Duration as numeric and only focusing on Movies I see that the majority of our movies duration are in the average of 100 minutes for most of our content.