#### Load required libraries

library(data.table)

library(dplyr)

library(tidyr)

library(ggplot2)

library(readr)

library(magrittr)

library(tidyverse)

library(ggpubr)

View(netflix\_titles)

head(netflix\_titles)

summary(netflix\_titles)

# What is the distribution of TV shows vs. movies in your dataset?

# count the number and percentage of TV shows and movies

tv\_movie\_count <- netflix\_titles %>%

group\_by(type) %>%

summarise(count = n()) %>%

mutate(percent = count/sum(count)\*100)

# create a bar plot

ggplot(tv\_movie\_count, aes(x = type, y = count, fill = type)) +

geom\_bar(stat = "identity") +

scale\_fill\_manual(values = c("red", "lightblue")) +

labs(title = "Distribution of TV shows vs. movies", x = "Type", y = "Count")

Chart, bar chart

Description automatically generated

#Which countries have produced the most movies/shows in your dataset, and how has this changed over time?

# count the number of movies and shows produced by each country in each year

country\_count <- netflix\_titles %>%

group\_by(country, release\_year) %>%

summarise(count = n()) %>%

ungroup() %>%

arrange(desc(count))

# plot the top 10 countries with the highest count of movies/shows

ggplot(head(country\_count, n = 10), aes(x = country, y = count, fill = country)) +

geom\_bar(stat = "identity") +

scale\_fill\_manual(values = c(rep("blue", 10))) +

labs(title = "Top 10 Countries by Number of Movies/Shows Produced", x = "Country", y = "Count") +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1))

A picture containing timeline

Description automatically generated

# plot the trend of the number of movies/shows produced over time for a selected country

ggplot(filter(country\_count, country == "United States"), aes(x = release\_year, y = count)) +

geom\_line() +

labs(title = "Number of Movies/Shows Produced by the United States Over Time", x = "Year", y = "Count")

Chart, line chart

Description automatically generated

# What is the most common rating for movies/shows in your dataset, and how does this vary by country or genre?

# count the number of movies/shows with each rating by country and genre

rating\_count <- netflix\_titles %>%

group\_by(rating, country, listed\_in) %>%

summarise(count = n()) %>%

ungroup() %>%

arrange(desc(count))

# plot the top 10 rated movies/shows by country and genre

ggplot(head(rating\_count, n = 10), aes(x = rating, y = count, fill = country)) +

geom\_bar(stat = "identity") + facet\_wrap(~listed\_in) + scale\_fill\_manual(values = c(rep("red", 10))) +

labs(title = "Top 10 Rated Movies/Shows by Country and Genre", x = "Rating", y = "Count")

# plot the top 10 rated movies/shows by country and rating

ggplot(head(rating\_count, n = 10), aes(x = rating, y = count, fill = listed\_in)) + geom\_bar(stat = "identity") + facet\_wrap(~country) + scale\_fill\_manual(values = c("yellow", "red", "darkblue", "blue", "lightblue", "gray","pink", "cyan", "black"))

Chart, bar chart

Description automatically generated

#Which directors have the most titles?

director\_counts <- netflix\_titles %>%

group\_by(director) %>%

summarize(count = n()) %>%

arrange(desc(count)) %>%

top\_n(10)

ggplot(director\_counts, aes(x = reorder(director, count), y = count)) +

geom\_col(fill = "pink") +

coord\_flip() +

labs(title = "Directors with the most titles in the dataset", x = "Director", y = "Number of titles")

Chart, funnel chart

Description automatically generated

# Filter out rows with missing values in duration

netflix\_titles <- netflix\_titles %>%

filter(!is.na(duration2))

# Group the data by genre and calculate the average runtime

genre\_runtime <- netflix\_titles %>%

group\_by(type, listed\_in) %>%

summarise(avg\_runtime = mean(duration2)) %>%

arrange(desc(avg\_runtime)) %>%

group\_by(type) %>%

top\_n(10, avg\_runtime) %>%

ungroup()

# Create the plot

ggplot(genre\_runtime, aes(x = avg\_runtime, y = fct\_reorder(listed\_in, avg\_runtime), fill = type)) +

geom\_col(position = "dodge") +

scale\_fill\_manual(values = c("#E50914", "#221F1F")) +

labs(title = "Top 10 Genres by Average Runtime on Netflix",

x = "Average Runtime (minutes)",

y = "Genre") +

theme\_minimal()

Chart

Description automatically generated

# Count the number of titles by listed\_in and country

titles\_by\_genre <- netflix\_titles %>%

group\_by(listed\_in, country) %>%

summarize(count = n()) %>%

ungroup()

# Create a stacked bar chart of the number of titles by genre and country

ggplot(titles\_by\_genre, aes(x = country, y = count, fill = listed\_in)) +

geom\_bar(stat = "identity") +

theme(legend.position = "bottom") +

labs(title = "Number of Titles by Genre and Country", x = "Country", y = "Count", fill = "Genre")

Table

Description automatically generated

#How does the number of cast members vary by movie/show type (i.e. movie vs. TV show)?

# Count the number of cast members by type and title

cast\_counts <- netflix\_titles %>%

group\_by(type, title) %>%

summarize(num\_casts = n()) %>%

ungroup()

# Average the number of cast members by type

avg\_casts\_by\_type <- netflix\_titles %>%

group\_by(type) %>%

summarize(avg\_casts = mean(num\_casts))

# Create a bar chart of the average number of cast members by type

ggplot(avg\_casts\_by\_type, aes(x = type, y = avg\_casts)) +

geom\_bar(stat = "identity") +

labs(title = "Average Number of Cast Members by Type",

x = "Type",

y = "Average Number of Cast Members")

A picture containing text, screenshot, display, electronics

Description automatically generated

#What is the distribution of release years for action & adventure, anime, comedy, international, British TV, Independent, Reality movies/shows in your dataset?

# Filter the dataset for the selected genres

selected\_genres <- c("Action & Adventure", "Anime", "Comedy", "International", "British TV", "Independent", "Reality")

filtered\_netflix <- netflix\_titles %>% filter(listed\_in %in% selected\_genres)

# Group the data by release\_year and count the number of titles in each year

netflix\_by\_genre\_year <- filtered\_netflix %>% group\_by(release\_year) %>% summarize(count=n())

# Plot the distribution of release years for each genre

ggplot(netflix\_by\_genre\_year, aes(x=release\_year, y=count, color=release\_year)) + geom\_line() + scale\_color\_gradient(low="blue", high="red")

Chart, line chart, histogram

Description automatically generated

# Which actors or actresses have appeared in the most action & adventure, anime, comedy, international, British TV, Independent, Reality?

genres <- c("Action & Adventure", "Anime", "Comedy", "International", "British TV", "Independent", "Reality")

cast\_by\_genre <- netflix\_titles %>%

filter(str\_detect(listed\_in, str\_c(genres, collapse = "|"))) %>%

group\_by(cast) %>%

summarize(num\_appearances = n()) %>%

arrange(desc(num\_appearances))

cast\_by\_genre %>%

top\_n(10, num\_appearances) %>%

ggplot(aes(x = reorder(cast, num\_appearances), y = num\_appearances, fill = cast)) +

geom\_bar(stat = "identity") +

coord\_flip() +

labs(title = "Top Actors/Actresses by Genre", x = "Actor/Actress", y = "Number of Appearances", fill = "Actor/Actress")

# How does the average runtime of action & adventure, anime, comedy, international, British TV, Independent, Reality compare to the average runtime of movies/shows in general?

# Plot 1: Average Runtime of Movies/Shows in General

avg\_runtime\_plot <- ggplot(avg\_runtime, aes(x = "", y = avg\_total\_runtime, fill = "Total")) +

geom\_bar(width = 1, stat = "identity") +

geom\_text(aes(label = round(avg\_total\_runtime, 1)), vjust = 1.5, size = 5) +

coord\_polar(theta = "y") +

ggtitle("Average Runtime of Movies/Shows in General")

# Plot 2: Average Runtime of Action & Adventure, Anime, Comedy, International, British TV, Independent, Reality Movies/Shows

avg\_genre\_runtime\_plot <- ggplot(avg\_runtime, aes(x = "", y = avg\_genre\_runtime, fill = "Genre")) +

geom\_bar(width = 1, stat = "identity") +

geom\_text(aes(label = round(avg\_genre\_runtime, 1)), vjust = 1.5, size = 5) +

coord\_polar(theta = "y") +

ggtitle("Average Runtime of Action & Adventure, Anime, Comedy, International, British TV, Independent, Reality Movies/Shows")

# Arrange and render the plots

ggarrange(avg\_runtime\_plot, avg\_genre\_runtime\_plot, ncol = 2)

Chart, pie chart

Description automatically generated

# What are the most common sub-genres of action & adventure, anime, comedy, international, British TV, Independent, Reality movies/shows, and how do these vary by country or release year?

subgenres <- netflix\_titles %>%

filter(str\_detect(listed\_in, "Action & Adventure|Anime|Comedy|International|British TV|Independent|Reality")) %>%

separate\_rows(listed\_in, sep = ", ") %>%

group\_by(listed\_in, country, release\_year) %>%

summarise(num\_titles = n()) %>%

ungroup()

subgenres\_plot <- subgenres %>%

filter(listed\_in %in% c("Action & Adventure", "Anime", "Comedy", "International", "British TV", "Independent", "Reality")) %>%

ggplot(aes(x = country, y = num\_titles, fill = listed\_in)) +

geom\_bar(stat = "identity", position = position\_dodge()) +

facet\_wrap(~listed\_in, scales = "free") +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1)) +

ggtitle("Number of Titles by Subgenre, Country and Release Year")

subgenres\_plot

A picture containing timeline

Description automatically generated