8. Use R-Project for data visualization of social media data

Data set from Kaggle - *SocialMedia.csv*

Install packages (if not already installed)

```
install.packages(c("tidyverse", "ggplot2"))

# Load necessary libraries
library(tidyverse)
library(ggplot2)

# Set working directory (modify the path accordingly)
setwd("D:/")

# Set the path to the uploaded CSV file
file_path <- "SM.csv"

# Read the CSV file
social_data <- read.csv(file_path)

head(social_data)
```

	N	PROFILE	FOLLOWERS	POSTS	Platform
1	1	Instagram\xa0	539446645	7202	Instagram\xa0
2	2	Cristiano Ronaldo\xa0	473864939	3338	Instagram\xa0
3	3	Kylie <U+0001F90D>\xa0	364542529	6935	Instagram\xa0
4	4	Leo Messi\xa0	355790796	890	Instagram\xa0
5	5	Selena Gomez\xa0	341579063	1828	Instagram\xa0
6	6	Dwayne Johnson\xa0	333221596	6738	Instagram\xa0

```
# Check for missing values
colSums(is.na(social_data))
#Output
```

N	PROFILE	FOLLOWERS	POSTS	Platform
0	0	0	100	0

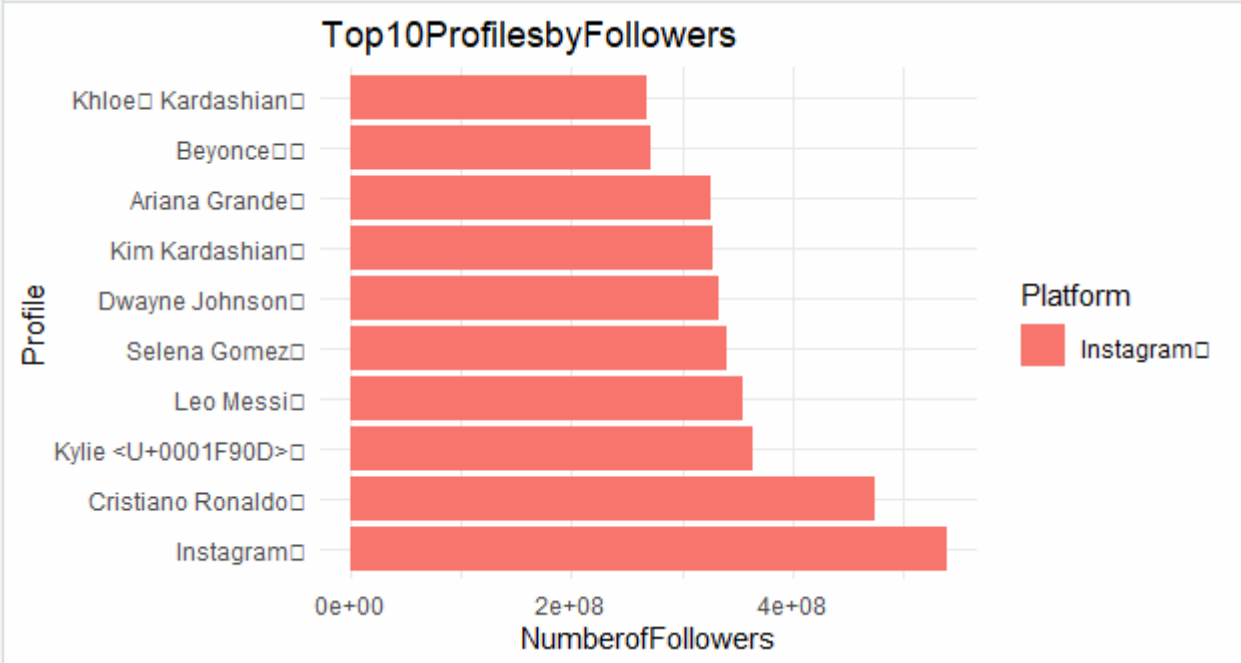
```
 social_data<-social_data%>%
 mutate(PROFILE=str_replace_all(PROFILE,"\\xa0|\\<U\\+0002F90D>",""),
              Platform=str_replace_all(Platform,"\\xa0",""))
head(social_data)
#Ouptut
```

	N	PROFILE	FOLLOWERS	POSTS	Platform
1	1	Instagram<U+FFFD>	539446645	7202	Instagram<U+FFFD>
2	2	Cristiano Ronaldo<U+FFFD>	473864939	3338	Instagram<U+FFFD>
3	3	Kylie <U+0001F90D><U+FFFD>	364542529	6935	Instagram<U+FFFD>
4	4	Leo Messi<U+FFFD>	355790796	890	Instagram<U+FFFD>
5	5	Selena Gomez<U+FFFD>	341579063	1828	Instagram<U+FFFD>
6	6	Dwayne Johnson<U+FFFD>	333221596	6738	Instagram<U+FFFD>

```
social_data<-social_data%>%
  mutate(PROFILE=str_replace_all(PROFILE,"<U+FFFD>",""),
            Platform=str_replace_all(Platform,"<U+FFFD>",""))
head(social_data,10)
#Output
```

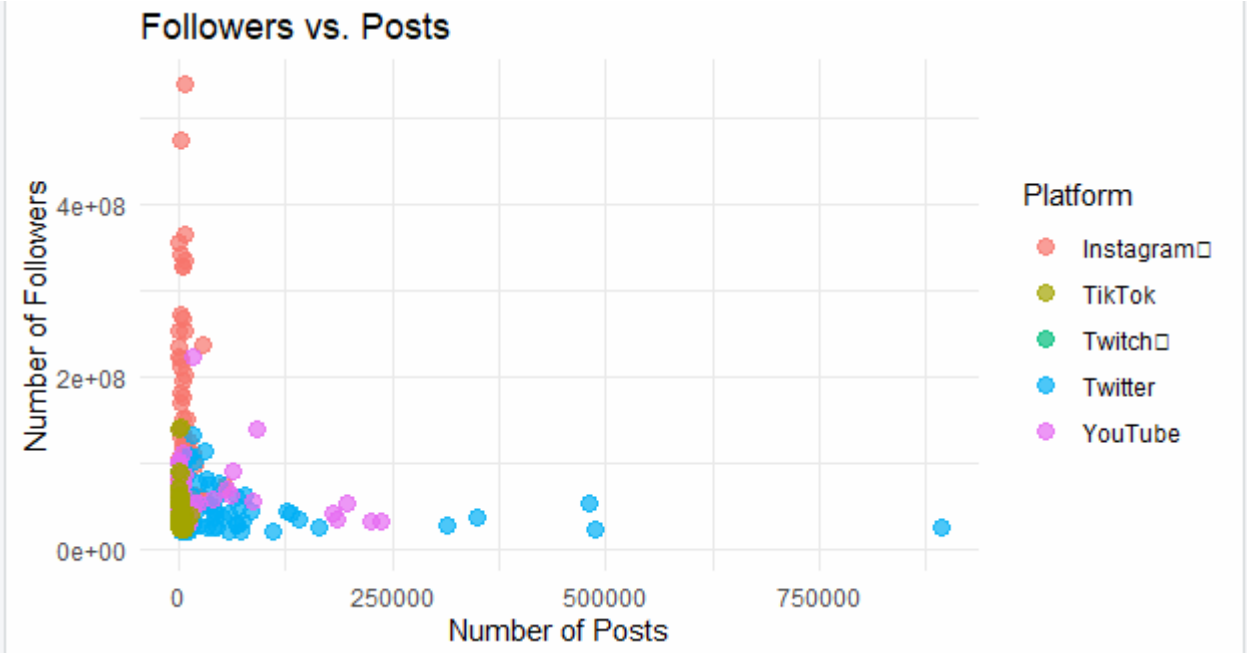
	N	PROFILE	FOLLOWERS	POSTS	Platform
1	1	Instagram<U+FFFD>	539446645	7202	Instagram<U+FFFD>
2	2	Cristiano Ronaldo<U+FFFD>	473864939	3338	Instagram<U+FFFD>
3	3	Kylie <U+0001F90D><U+FFFD>	364542529	6935	Instagram<U+FFFD>
4	4	Leo Messi<U+FFFD>	355790796	890	Instagram<U+FFFD>
5	5	Selena Gomez<U+FFFD>	341579063	1828	Instagram<U+FFFD>
6	6	Dwayne Johnson<U+FFFD>	333221596	6738	Instagram<U+FFFD>
7	7	Kim Kardashian<U+FFFD>	328323322	5598	Instagram<U+FFFD>
8	8	Ariana Grande<U+FFFD>	326473270	4963	Instagram<U+FFFD>
9	9	Beyonce<U+FFFD><U+FFFD>	272211534	2037	Instagram<U+FFFD>
10	10	Khloe<U+FFFD> Kardashian<U+FFFD>	267912338	4104	Instagram<U+FFFD>

```
top_profiles_followers <- social_data%>%
arrange(desc(FOLLOWERS))%>%
slice(1:10)
ggplot(top_profiles_followers,aes(x=reorder(PROFILE,-FOLLOWERS),
                y=FOLLOWERS, fill =Platform))+
        geom_bar(stat="identity")+
        coord_flip()+
        labs(title="Top10ProfilesbyFollowers",
        x="Profile",
        y="NumberofFollowers",
        fill="Platform")+
        theme_minimal()
```



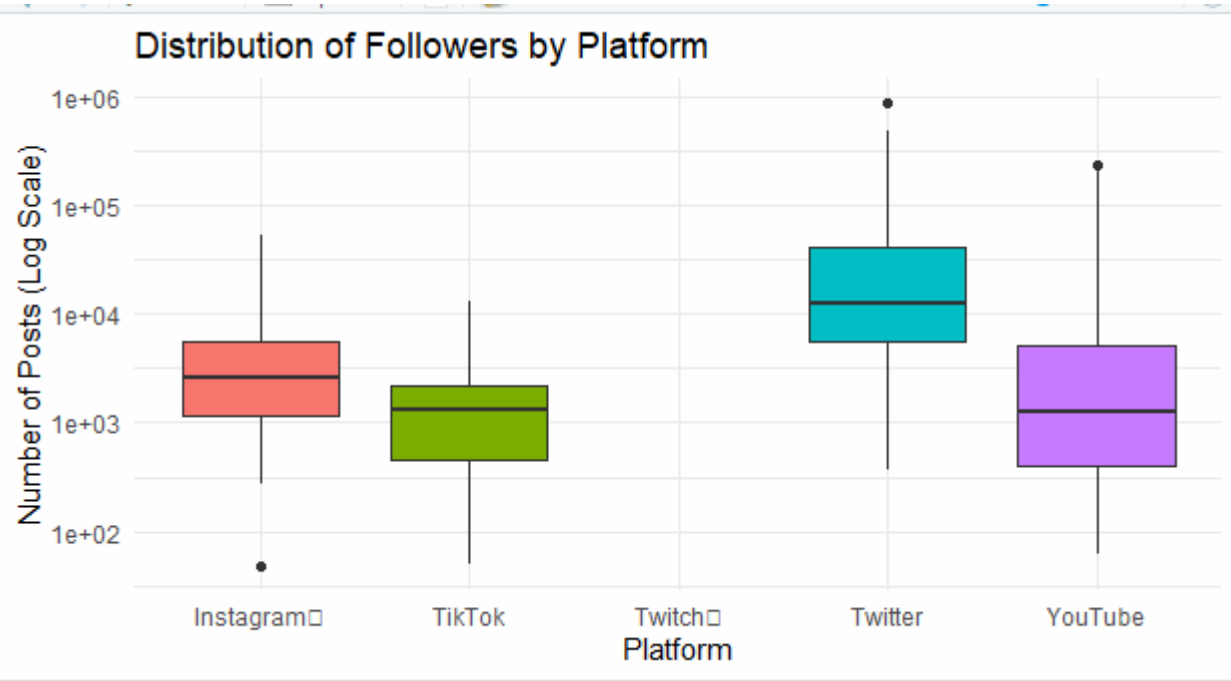
Scatter plot of followers vs. posts

```
ggplot(social_data, aes(x = POSTS, y = FOLLOWERS, color = Platform)) +
                geom_point(size = 3, alpha = 0.7) +
        labs(title = "Followers vs. Posts",
                x = "Number of Posts",
                y = "Number of Followers",
                color = "Platform") +
        theme_minimal()
```

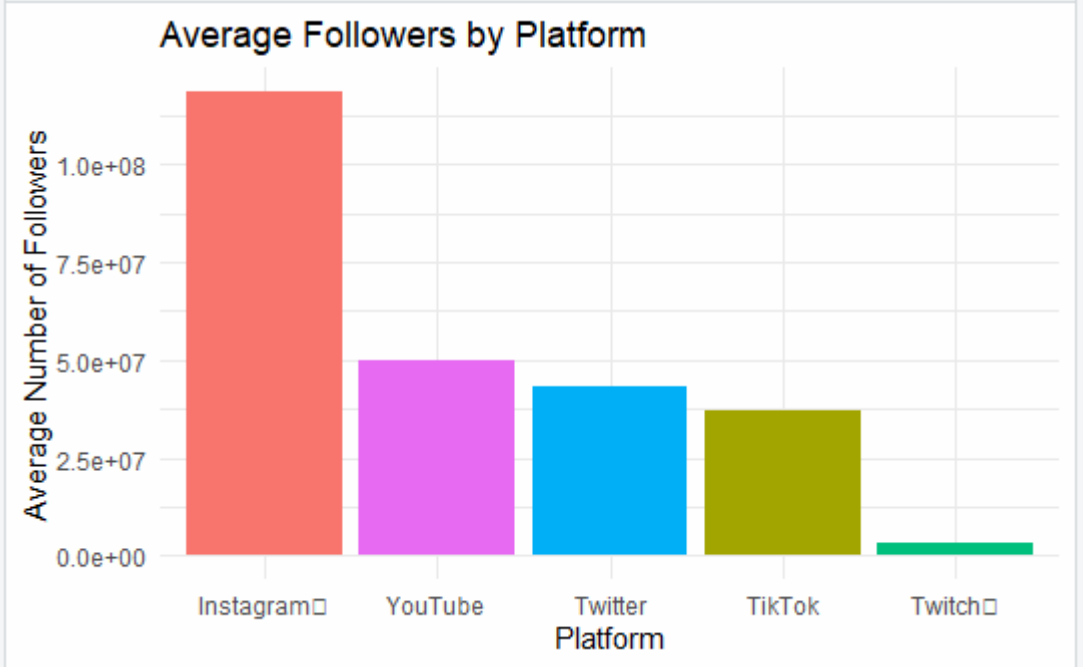


Boxplot of followers by platform

```
ggplot(social_data, aes(x = Platform, y = POSTS, fill = Platform)) +
  geom_boxplot() +
  scale_y_log10() + # Log scale to handle wide range in follower counts
  labs(title = "Distribution of Followers by Platform",
       x = "Platform",
       y = "Number of Posts (Log Scale)") +
  theme_minimal() +
  theme(legend.position = "none")
```

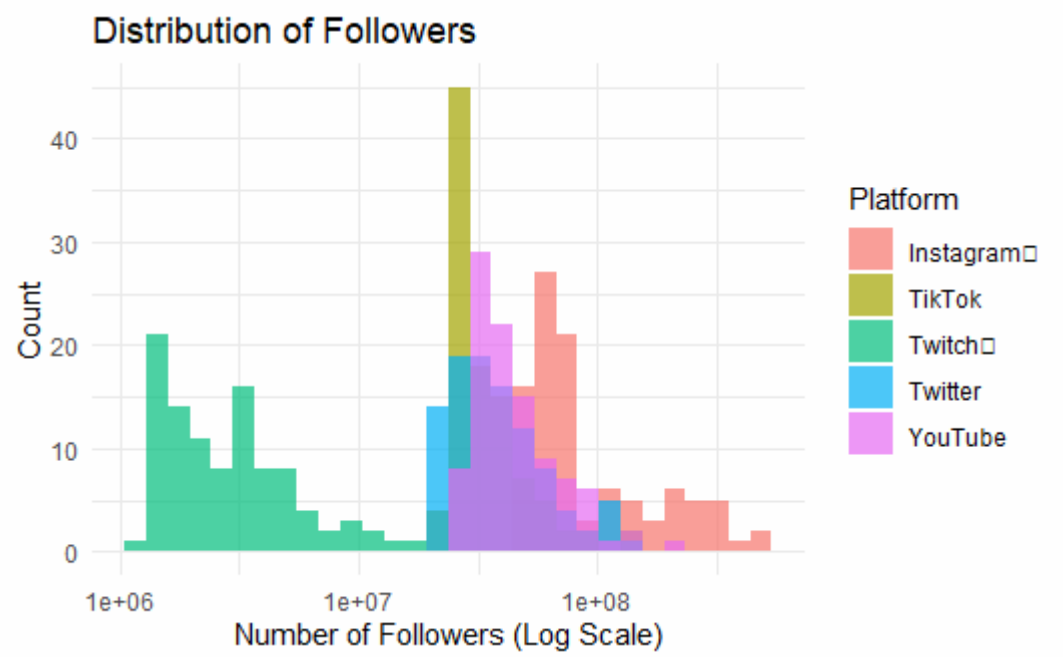


Calculate average followers by platform
avg_followers_platform <- social_data %>%
group_by(Platform) %>%
summarise(Average_Followers = mean(FOLLOWERS))
Plot the average followers by platform
ggplot(avg_followers_platform, aes(x = reorder(Platform,-Average_Followers),
y = Average_Followers, fill = Platform)) +
geom_bar(stat = "identity") +
labs(title = "Average Followers by Platform",
x = "Platform",
y = "Average Number of Followers") +
theme_minimal() +
theme(legend.position = "none")



Histogram of followers

```
ggplot(social_data, aes(x = FOLLOWERS, fill = Platform)) +
        geom_histogram(bins = 30, alpha = 0.7, position = "identity") +
         scale_x_log10() + # Log scale to handle large range of follower counts
        labs(title = "Distribution of Followers",
             x = "Number of Followers (Log Scale)",
             y = "Count",
             fill = "Platform") +
        theme_minimal()
```


Distribution of Followers
Count
40
30
20
10
0
1e+06
1e+07
1e+08
Number of Followers (Log Scale)
Platform
Instagram
TikTok
Twitch
Twitter
YouTube