

R and power bi project

Rafal Szymanski



May 11, 2023

Contents

[1. Data source and aim 2](#_Toc135058304)

[2. Requirements of this assignments 2](#_Toc135058305)

[3. Importing data to R: 2](#_Toc135058306)

[4. Viewing the data that was added to df, data frame variable 2](#_Toc135058307)

[5. Adding (installing) tidyverse package to R Studio 3](#_Toc135058308)

[6. Loading installed tidyverse package to R Studio for using 4](#_Toc135058309)

[7. Checking data types in df 4](#_Toc135058310)

[8. Checking for missing values in df 4](#_Toc135058311)

[9. Removing missing values from df 5](#_Toc135058312)

[10. Checking to make sure that the missing values were removed 5](#_Toc135058313)

[11. Checking for duplicates 5](#_Toc135058314)

[12. Rounding off Profitability and Worldwide.Gross values to 2 decimal places 6](#_Toc135058315)

[13. Checking df variable for rows (70) and columns/variables (8) 6](#_Toc135058316)

[14. Checking for outliers using a boxplot 6](#_Toc135058317)

[Loading ggplot2 package to R Studio 6](#_Toc135058318)

[Creating a boxplot that highlights the outliers 7](#_Toc135058319)

[Checking no\_outliers data set (variable) for the current row and column count 8](#_Toc135058320)

[Removing outliers in 'Worldwide.Gross' data set 8](#_Toc135058321)

[Checking df1 data set (variable) for the current row and column count 8](#_Toc135058322)

[15. Summary Statistics/Univariate Analysis 9](#_Toc135058323)

[16. Bivariate analysis 9](#_Toc135058324)

[Scatterplot 9](#_Toc135058325)

[Bar Chart 10](#_Toc135058326)

[17. Exporting clean data 10](#_Toc135058327)

[18.Power BI Dashboard 11](#_Toc135058328)

# Data source and aim

The data was used from:

<https://public.tableau.com/app/sample-data/HollywoodsMostProfitableStories.csv>

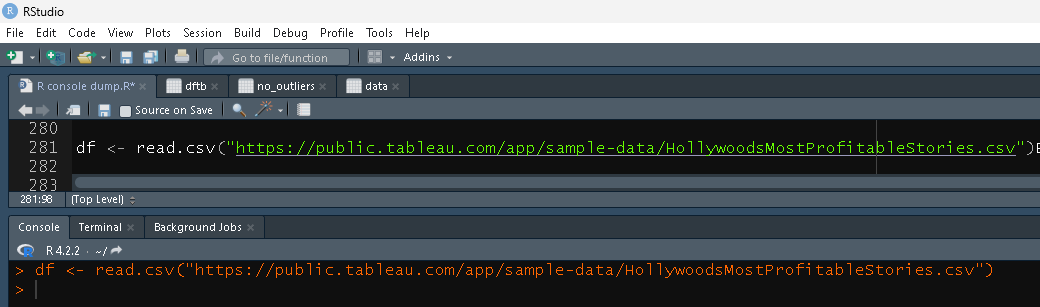
and the aim is to analyse the performance of Hollywood movies released between 2007 and 2012.

# Requirements of this assignments

* Create a word file and include screenshots of source code from R.
* Include 3 charts from R:
  + Box plot
  + Bar chart
  + Scatter plot

# Importing data to R:

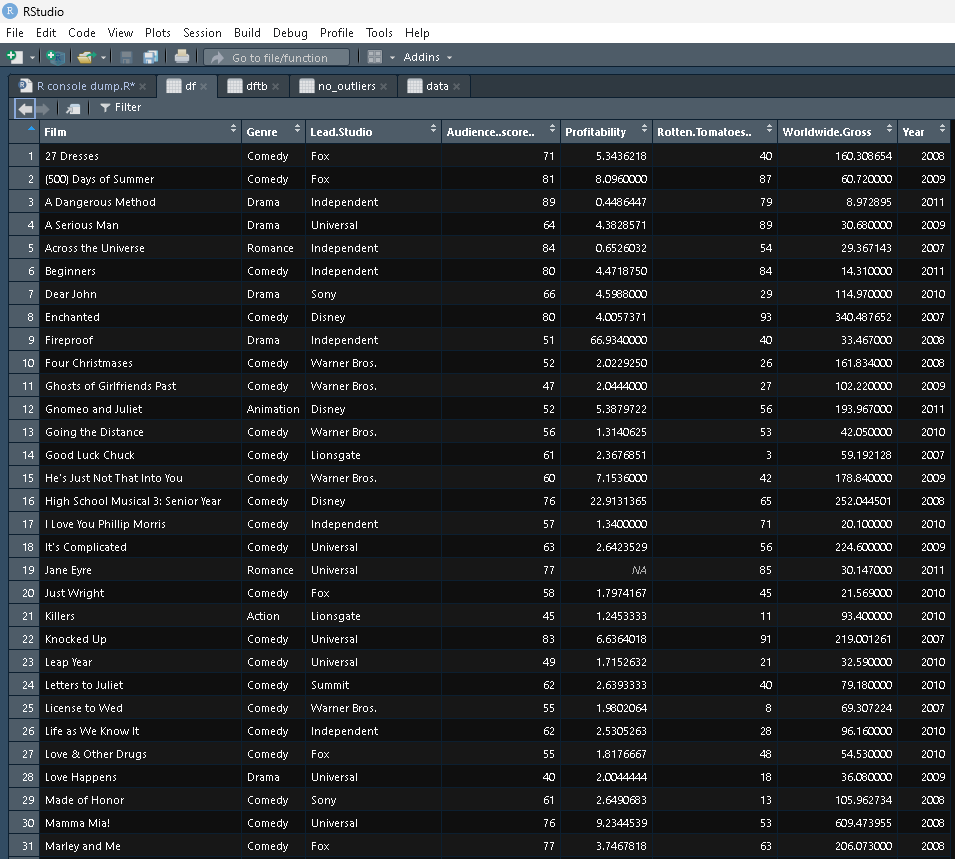
df<- read.csv(“<https://public.tableau.com/app/sample-data/HollywoodsMostProfitableStories.csv>")



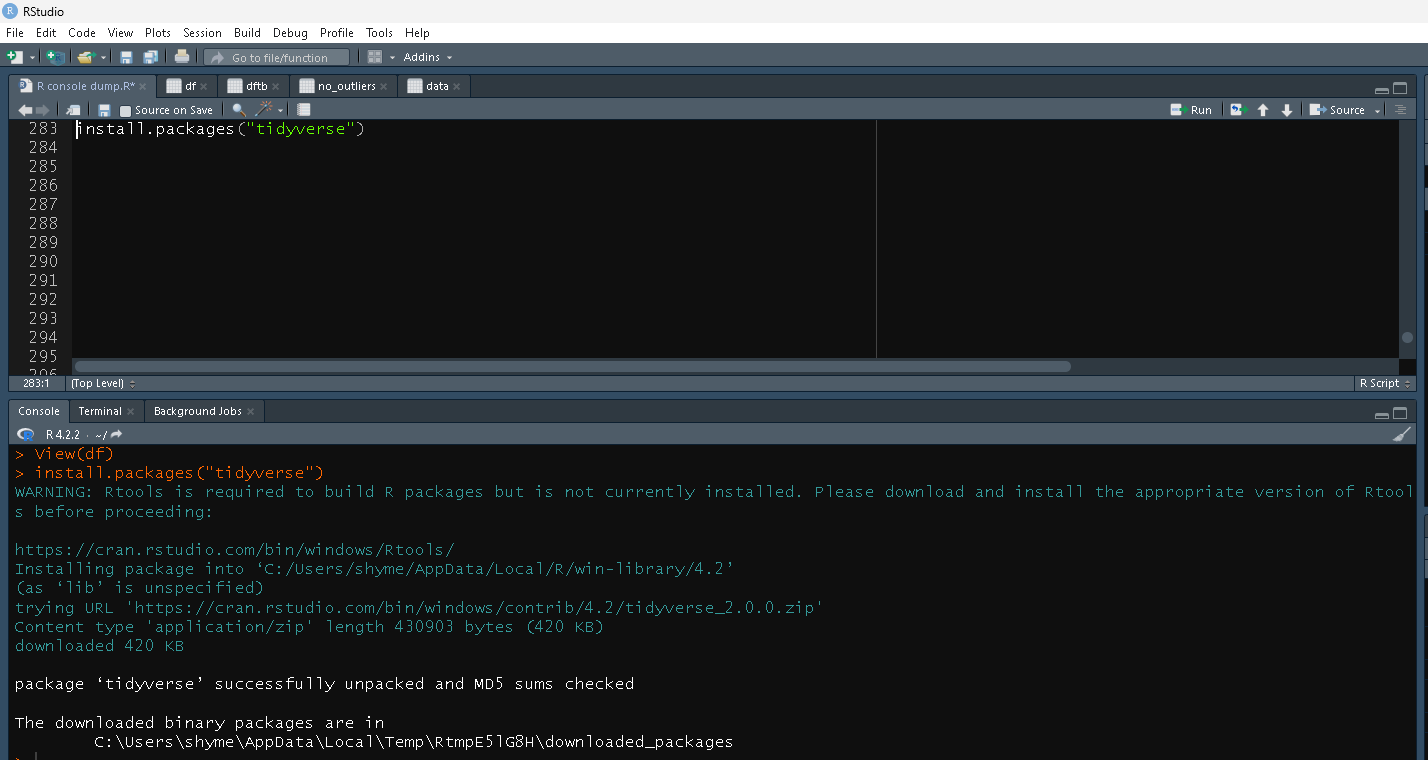
# Viewing the data that was added to df, data frame variable

View(df)



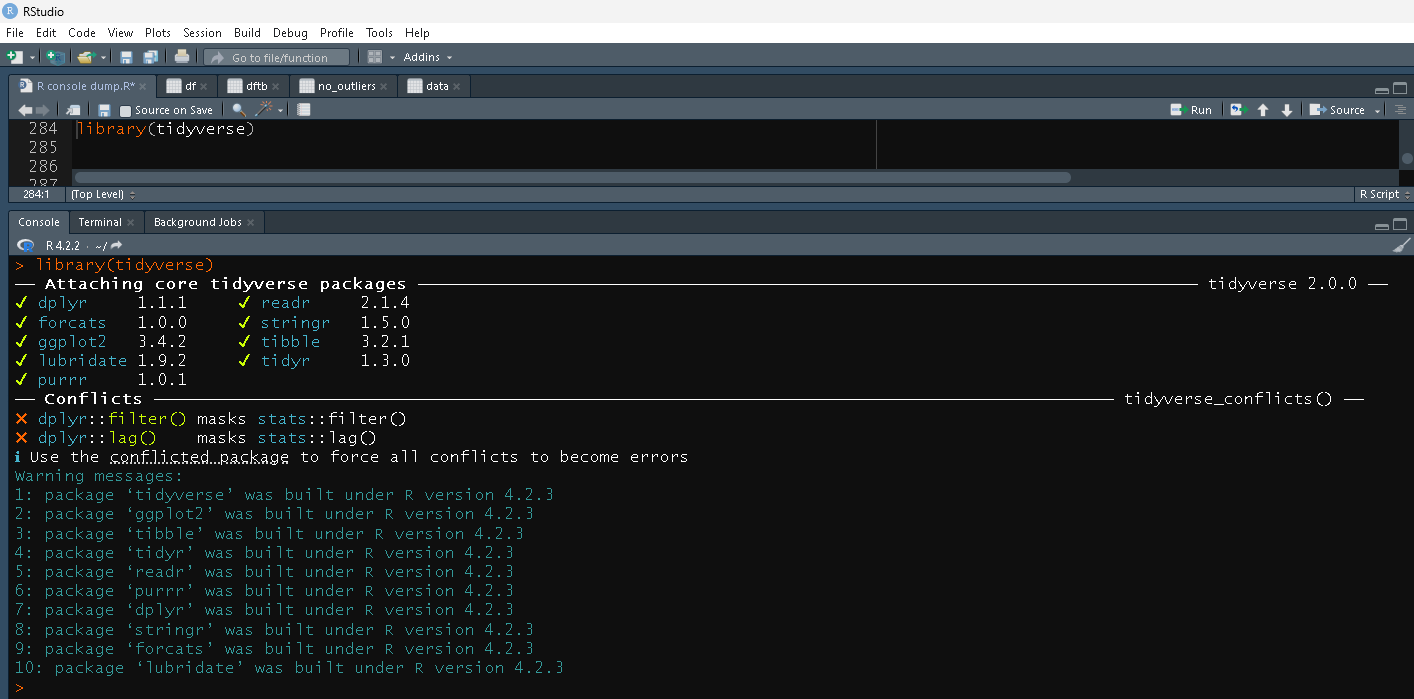
And Hello

# Adding (installing) tidyverse package to R Studio

install.packages("tidyverse")

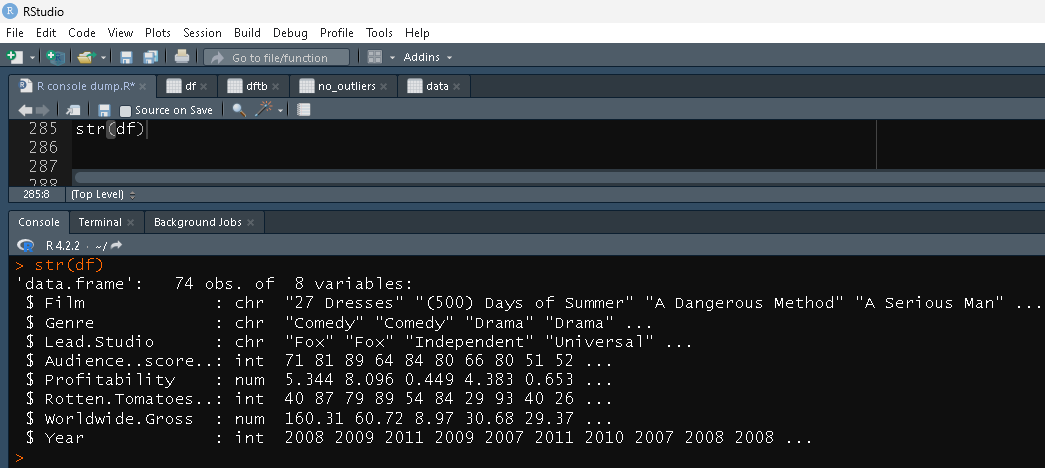
# Loading installed tidyverse package to R Studio for using

library(tidyverse)



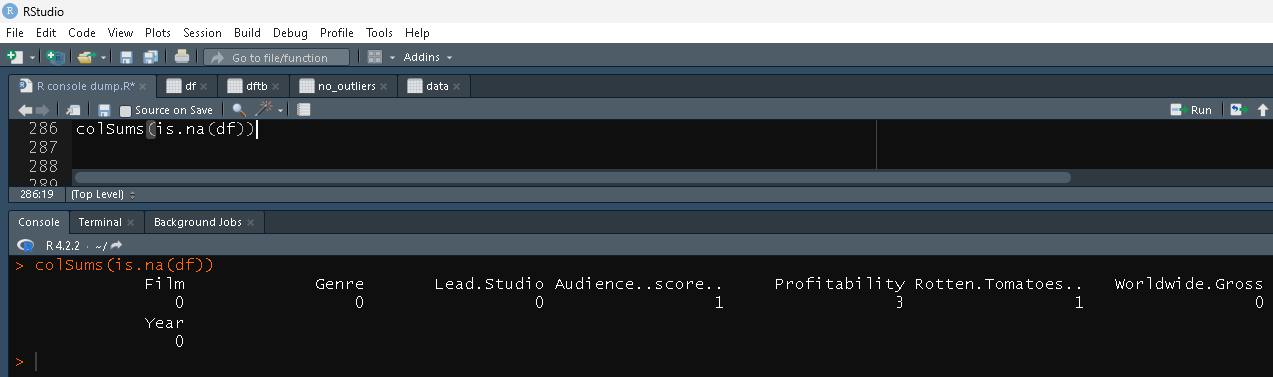
# Checking data types in df

str(df)



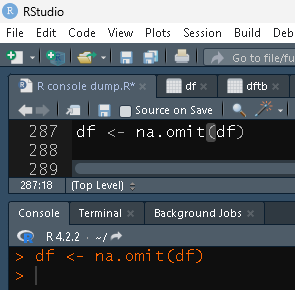
# Checking for missing values in df

colSums(is.na(df))



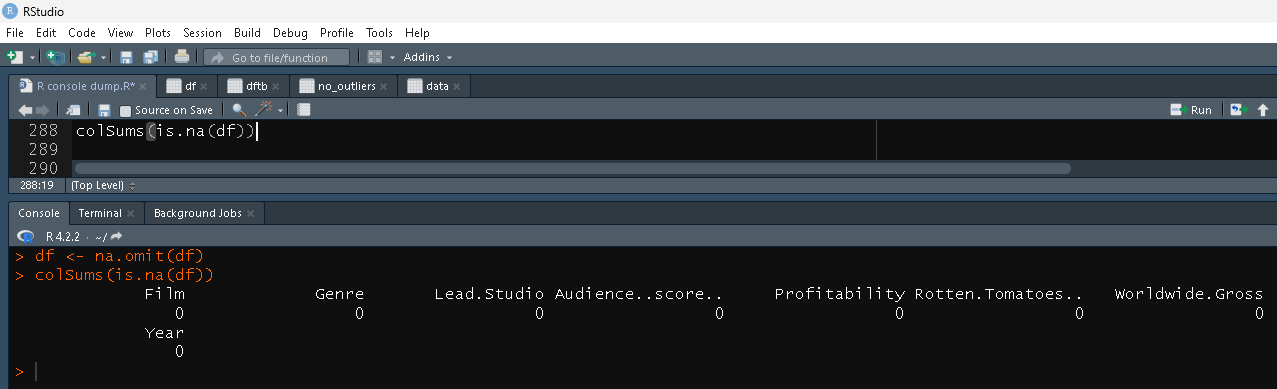
# Removing missing values from df

df <- na.omit(df)



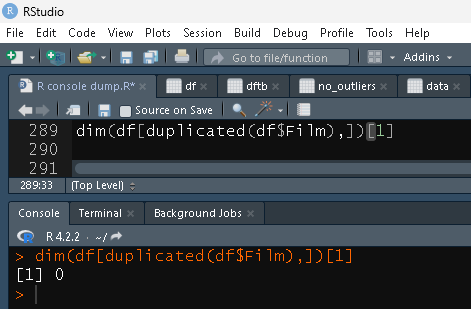
# Checking to make sure that the missing values were removed

colSums(is.na(df))



# Checking for duplicates

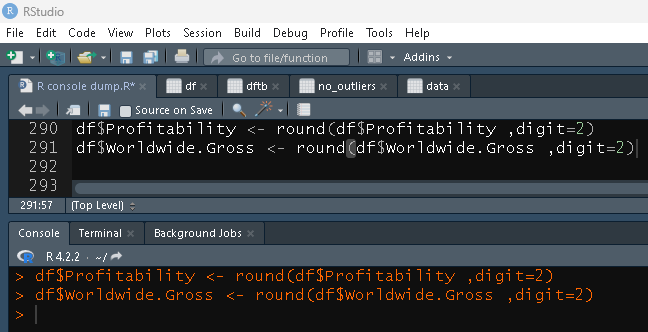
dim(df[duplicated(df$Film),])[1]



# Rounding off Profitability and Worldwide.Gross values to 2 decimal places

df$Profitability <- round(df$Profitability ,digit=2)

df$Worldwide.Gross <- round(df$Worldwide.Gross ,digit=2)



# 13. Checking df variable for rows (70) and columns/variables (8)

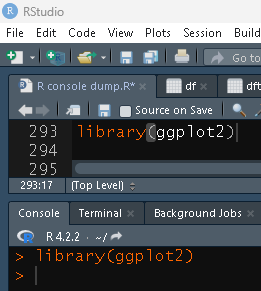
dim(df)



# 14. Checking for outliers using a boxplot

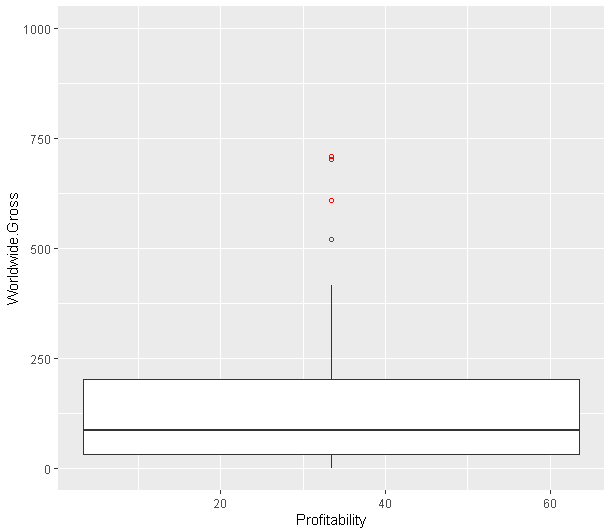
## Loading ggplot2 package to R Studio

library(ggplot2)



## Creating a boxplot that highlights the outliers

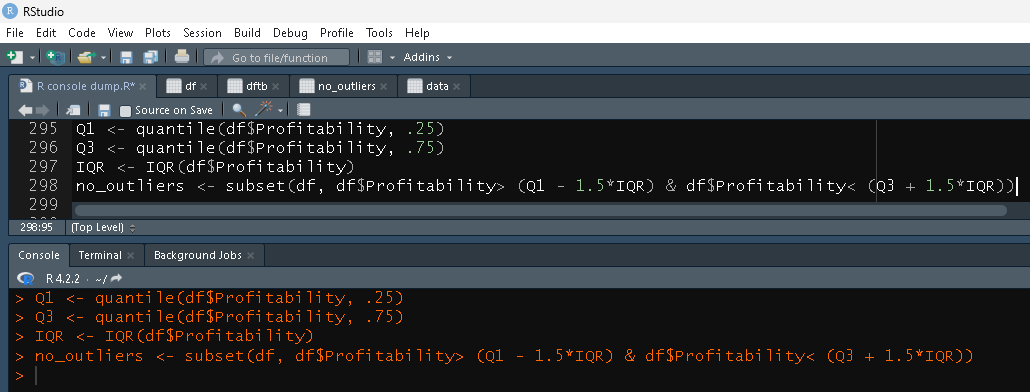
ggplot(df, aes(x=Profitability, y=Worldwide.Gross)) + geom\_boxplot(outlier.colour = "red", outlier.shape = 1)+ scale\_x\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 1000))

  
Removing outliers in 'Profitability' data set

Q1 <- quantile(df$Profitability, .25)

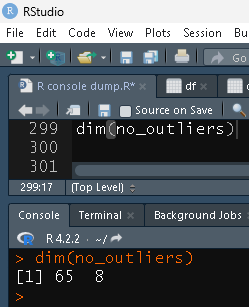
Q3 <- quantile(df$Profitability, .75)

IQR <- IQR(df$Profitability)

no\_outliers <- subset(df, df$Profitability> (Q1 - 1.5\*IQR) & df$Profitability< (Q3 + 1.5\*IQR))

## Checking no\_outliers data set (variable) for the current row and column count

dim(no\_outliers)



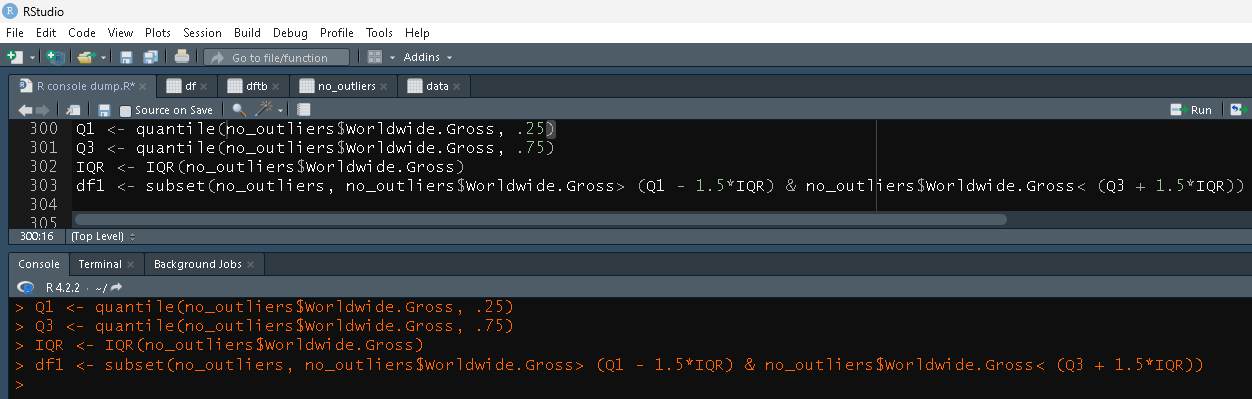
## Removing outliers in 'Worldwide.Gross' data set

Q1 <- quantile(no\_outliers$Worldwide.Gross, .25)

Q3 <- quantile(no\_outliers$Worldwide.Gross, .75)

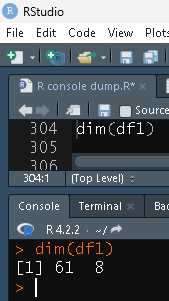
IQR <- IQR(no\_outliers$Worldwide.Gross)

df1 <- subset(no\_outliers, no\_outliers$Worldwide.Gross> (Q1 - 1.5\*IQR) & no\_outliers$Worldwide.Gross< (Q3 + 1.5\*IQR))



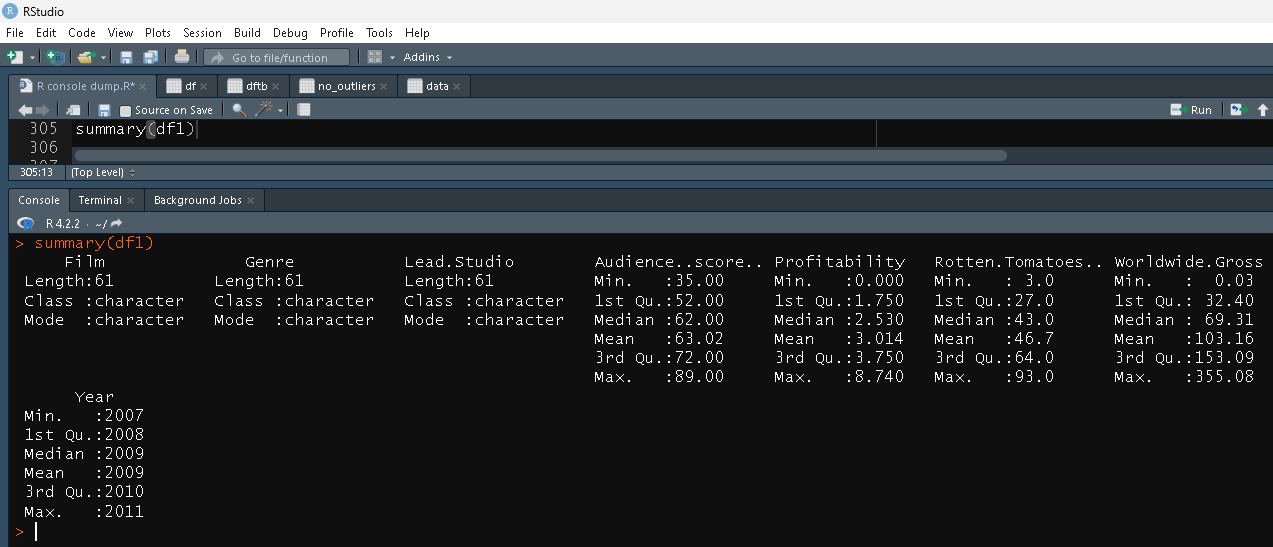
## Checking df1 data set (variable) for the current row and column count

dim(df1)



# 15. Summary Statistics/Univariate Analysis

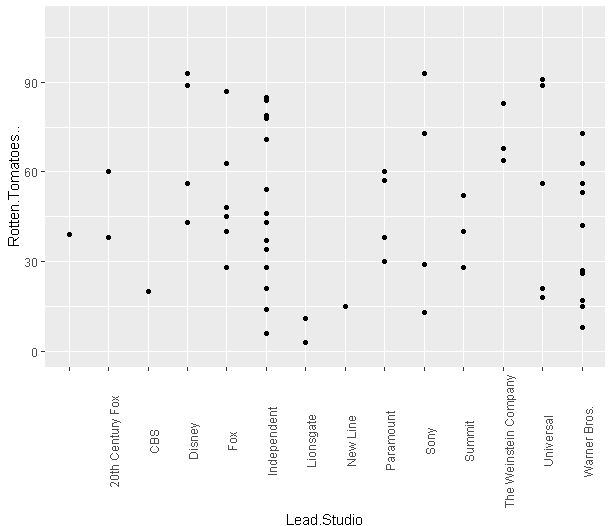
summary(df1)



# 16. Bivariate analysis

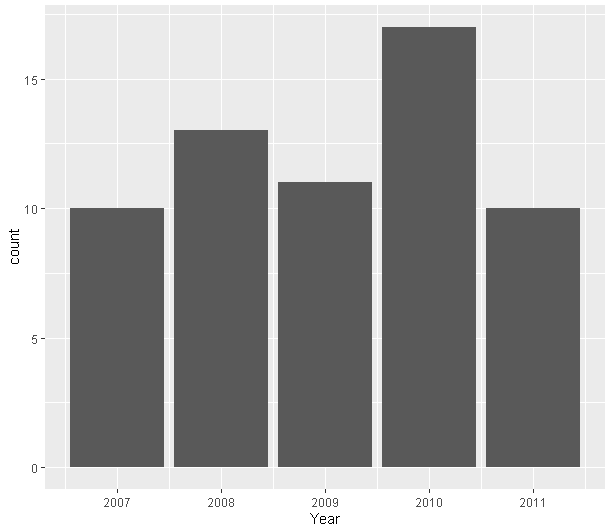
## Scatterplot

ggplot(df1, aes(x=Lead.Studio, y=Rotten.Tomatoes..)) + geom\_point()+ scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 110))+theme(axis.text.x = element\_text(angle = 90))



## Bar Chart

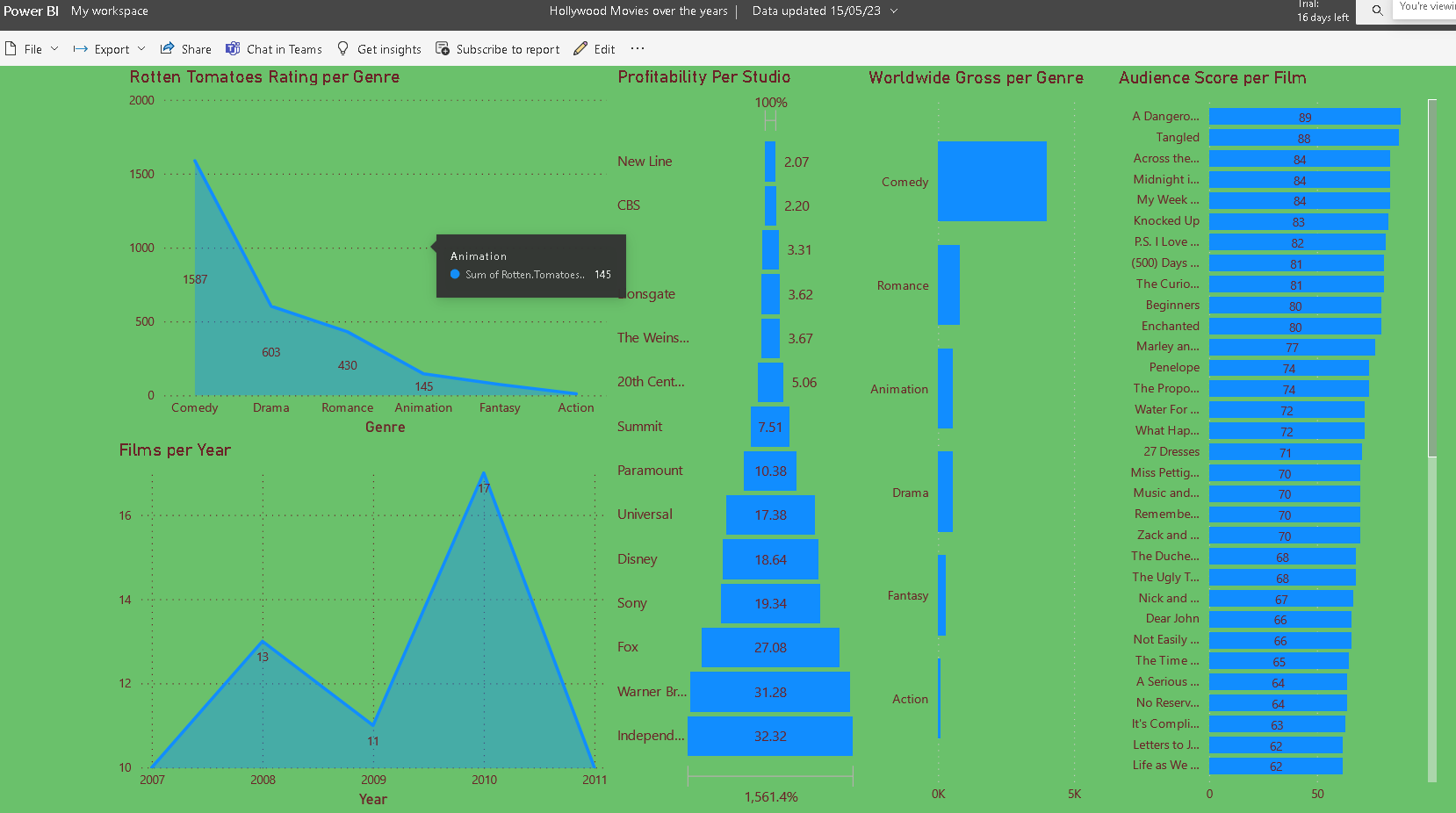
ggplot(df1, aes(x=Year)) + geom\_bar()



# 17. Exporting clean data

write.csv(df1, "clean\_df.csv")

# 18.Power BI Dashboard



<https://app.powerbi.com/groups/me/reports/5c45af6f-3824-4a10-82e8-20ac69693138/ReportSection>