11/9/2023

Nisanth Mahalingam

JUSTIT DATA TECHNICIAN TRAINEE

Assignment 3

R & Power BI

Contents

[Hollywood Film Performance in RStudio 0](#_Toc150453203)

[Initial Exploratory Data Analysis 0](#_Toc150453204)

[Cleaning Data 1](#_Toc150453205)

[Exploratory Data Analysis 2](#_Toc150453206)

[Power BI Data Visualization 9](#_Toc150453207)

[Further Transformation of Data 9](#_Toc150453208)

[Data Visualization 9](#_Toc150453209)

[Reflection 9](#_Toc150453210)

# Hollywood Film Performance in RStudio

## Initial Exploratory Data Analysis

**# Setting Work Directory**

setwd("~/Data Technician Bootcamp/Assignment 3 - R & Power BI")

**# Load data**

df<- read.csv("Hollywood Movies.csv")

**# Take a look at the data:**

View(df)

**# Load library:**

install.packages("tidyverse")

**# Import library**

library(tidyverse)

**# Check data types:**

str(df)

**# Check Column Names**

names(df)

df$Lead.Studio

## Cleaning Data

**# Check for missing values:**

colSums(is.na(df))

**# Drop missing values**

df<-na.omit(df)

**# or df <- df %>% drop\_na()**

**# Check to make sure that the rows have been removed**

colSums(is.na(df))

**# Declare as df1**

df1<-df

## Exploratory Data Analysis

**# Summary statistics:**

summary(df1)

**# Size of df1**

dim(df1)

**# When viewing here we can identify that the row index values aren't in the logical sequence missing values**

view(df1)

**# Conducting a check on the value which helps realise it is an empty string ""**

typeof(df1[38,3])

**# This resets the row numbering**

rownames(df1) <- NULL

**# This checks the row numbering**

view(df1)

**# This is to manually check for any strings**

df1$Lead.Studio

**# This is to add in an additional restriction to remove any empty strings**

df1\_filtered <- df1 %>% filter(Lead.Studio != "")

**# Check**

df1\_filtered$Lead.Studio

**# Renaming the variable**

df1f<-df1\_filtered

**# Check**

view(df1f)

**# Check size**

dim(df1f)

**# Generating the mean values for the profitability to add the red average point on the scatter graph**

means <- df1f %>% group\_by(Lead.Studio) %>% summarise(mean = mean(Profitability))

mean\_df1f <- data.frame(means)

mean\_df1f

**# Checking the columns**

names(df1f)

**# Generating the mean values for the rotten tomatoes score to add the red average point on the scatter graph**

means1 <- df1f %>% group\_by(Lead.Studio) %>% summarise(mean = mean(Rotten.Tomatoes..))

mean1\_df1f <- data.frame(means1)

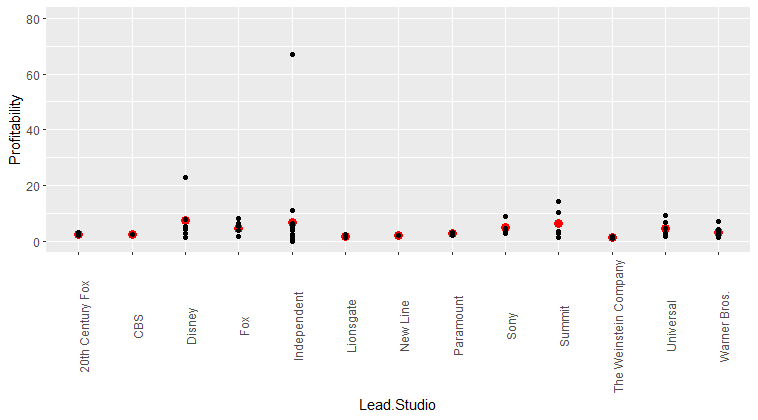
mean1\_df1f

**# Scatterplots**

**#Plotting Profitability against Lead Studio with the mean value for profitability for each lead studio shown in red**

ggplot(df1f, aes(x=Lead.Studio, y=Profitability)) + geom\_point( mapping = aes(y = mean ), data = mean\_df, colour = 'red', size = 3)+ geom\_point()+scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 80))+theme(axis.text.x = element\_text(angle = 90))

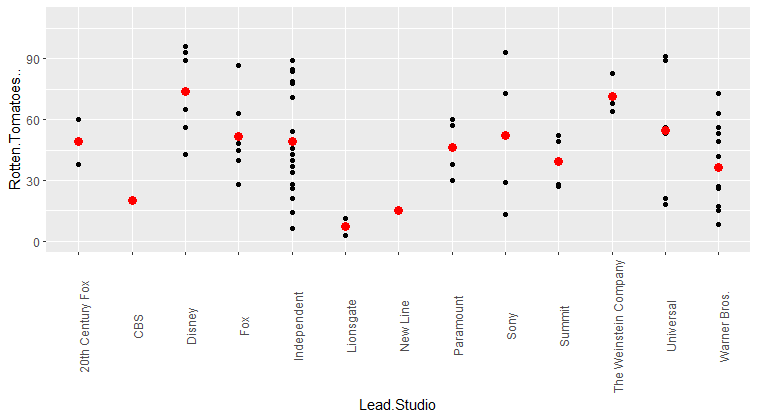
plot1=ggplot(df1f, aes(x=Lead.Studio, y=Profitability)) + geom\_point( mapping = aes(y = mean ), data = mean\_df1f, colour = 'red', size = 3)+ geom\_point()+scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 80))+theme(axis.text.x = element\_text(angle = 90))



**# Plotting Rotten Tomatoes against Lead Studio with the mean value for Rotten Tomatoes for each lead studio shown in red**

ggplot(df1f, aes(x=Lead.Studio, y=Rotten.Tomatoes..)) + geom\_point()+ geom\_point( mapping = aes(y = mean ), data = mean1\_df1f, colour = 'red', size = 3)+ scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 110))+theme(axis.text.x = element\_text(angle = 90))

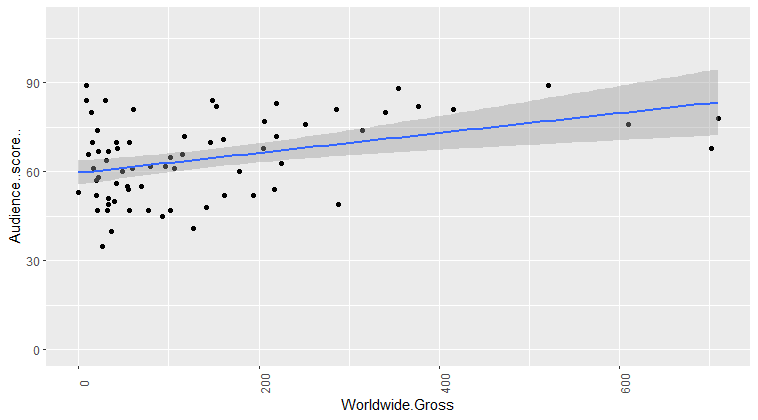
plot2=ggplot(df1f, aes(x=Lead.Studio, y=Rotten.Tomatoes..)) + geom\_point()+ geom\_point( mapping = aes(y = mean ), data = mean1\_df1f, colour = 'red', size = 3)+ scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 110))+theme(axis.text.x = element\_text(angle = 90))



**# Plotting Audience Score against Worldwide Gross with the trend line shown with the shaded region showing its confidence interval**

ggplot(df1f, aes(x=Worldwide.Gross, y=Audience..score..)) + geom\_point()+ geom\_smooth(method = "lm")+scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 110))+theme(axis.text.x = element\_text(angle = 90))

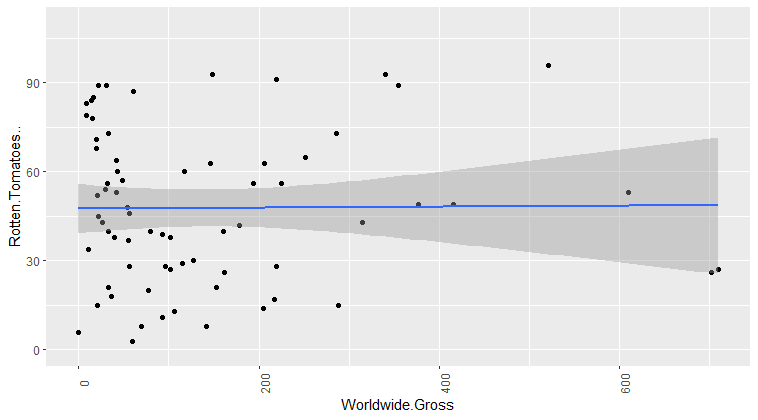
plot3=ggplot(df1f, aes(x=Worldwide.Gross, y=Audience..score..)) + geom\_point()+ geom\_smooth(method = "lm")+scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 110))+theme(axis.text.x = element\_text(angle = 90))



**# Plotting Rotten Tomatoes against Worldwide Gross with the trend line shown with the shaded region showing its confidence interval**

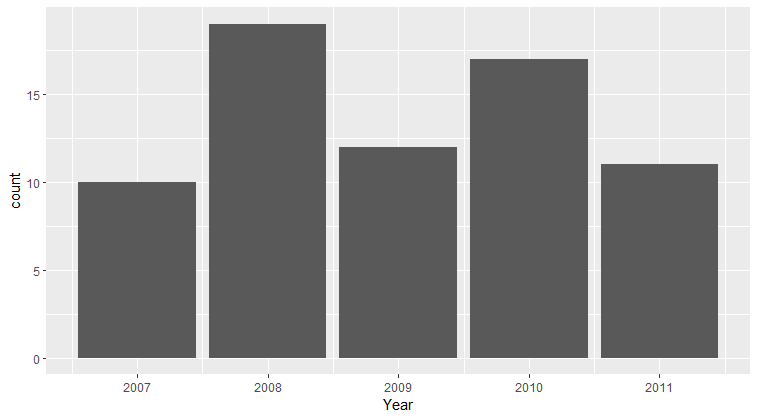
ggplot(df1, aes(x=Worldwide.Gross, y=Rotten.Tomatoes..)) + geom\_point()+ geom\_smooth(method = "lm")+ scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 110))+theme(axis.text.x = element\_text(angle = 90))

plot4=ggplot(df1, aes(x=Worldwide.Gross, y=Rotten.Tomatoes..)) + geom\_point()+ geom\_smooth(method = "lm")+ scale\_y\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 110))+theme(axis.text.x = element\_text(angle = 90))



**# Bar chart**

ggplot(df1f, aes(x=Year)) + geom\_bar()



## Exporting Data

**# Export clean data**

write.csv(df1f, "clean\_df1.csv")

# Power BI Data Visualization

## Further Transformation of Data

Changed the format of columns to:

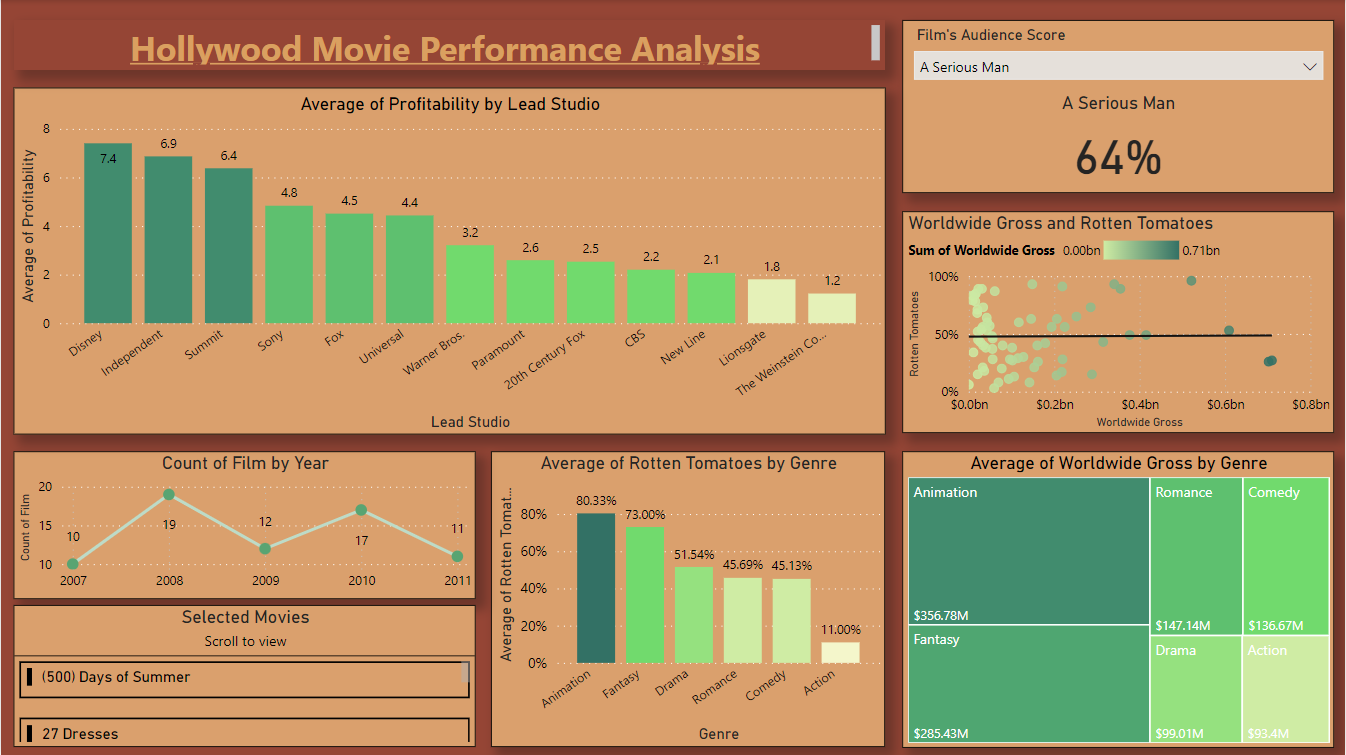
% : Audience Score, Rotten Tomatoes

$ : Worldwide Gross

## Data Visualization

Edit the interactions between the filters so that only the linked data is appropriately filtered

This includes isolating the Film’s Audience Score filter to the rest of the selections that will display the selected films based on the categories selected to filter in the interactive dashboard.



# Reflection

I really enjoyed doing this assignment as it allowed me to both explore the different colour Hex Codes combinations in Power BI as well as strengthening my understanding of how to utilise R to both clean data and create data visualizations. This task also helped grasp better understanding of DAX despite not being used in this particular dashboard as I had been experimenting.