Assignment 3: R + Power Bi

Abdulaziz Mohamed Nur

GLA 1 2022

[Task 1: R 3](#_Toc303068980)

[Task 2: PowerBi 6](#_Toc664550735)

# Task 1: R

Here is the code I used to clean and organise the data in R:

#Load data

df<- read.csv("/cloud/project/HollywoodsMostProfitableStories.csv")

#Take a look at the data:

View(df)

#Load library:

install.packages("tidyverse")

#Import library

library(tidyverse)

# Check data types:

str(df)

# Check Dimensions (number of rows and columns)

dim(df)

# Check for missing values:

colSums(is.na(df))

#Drop missing values in row

df <- df[rowSums(is.na(df)) == 0, ]

view(df)

df1 <- na.omit(df)

view(df1)

#check to make sure that the rows have been removed

head(df1)

#Check for duplicates

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

#round off values to 2 places

df1$Profitability <- round(df$Profitability ,digit=2)

df1$Worldwide.Gross <- round(df$Worldwide.Gross ,digit=2)

view(df1)

#Check Dimensions

dim(df1)

#Check for outliers using a boxplot

library(ggplot2)

#Create a boxplot that highlights the outliers

ggplot(df1, 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))

#Remove outliers in 'Profitability'

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

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

IQR <- IQR(df$Profitability)

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

dim(no\_outliers)

WGQ1 <- quantile(no\_outliers$Worldwide.Gross, .25)

WGQ3 <- quantile(no\_outliers$Worldwide.Gross, .75)

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

df2 <- subset(no\_outliers, no\_outliers$Worldwide.Gross> (WGQ1 - 1.5\*WGIQR) & no\_outliers$Worldwide.Gross< (WGQ3 + 1.5\*WGIQR))

dim(df2)

#Summary Statistics/Univariate Analysis:

summary(df2)

#scatterplot

ggplot(df2, 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()

#Export clean data

write.csv(df2, "clean\_df.csv")

# Task 2: Power Bi

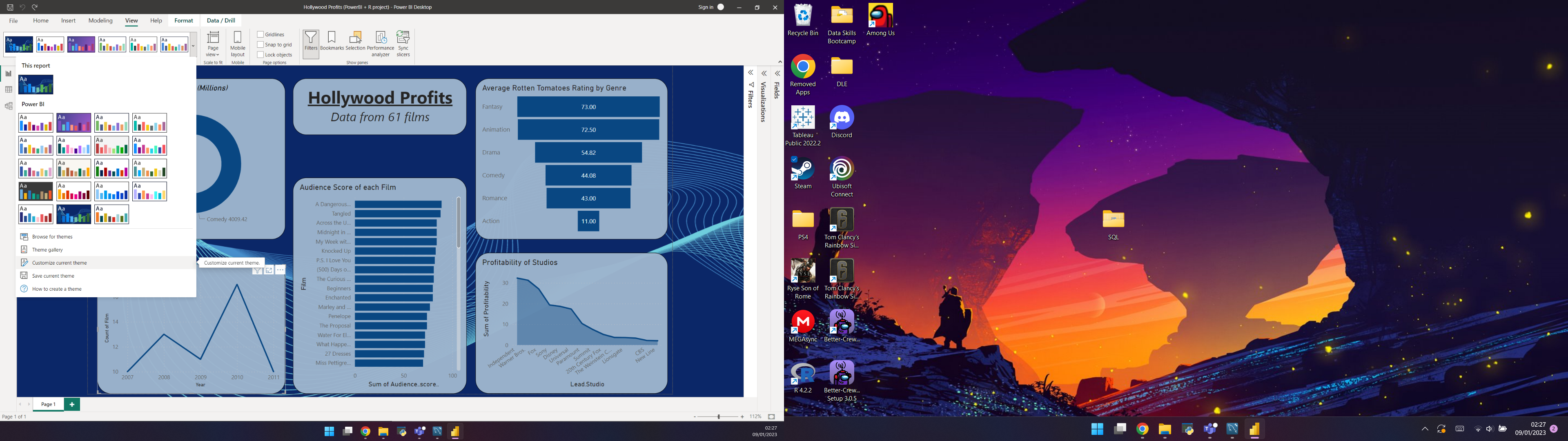
In this scenario, the company would like you to use their brand colors which are blue, green and brown. The use of light or dark shades of each colour is acceptable. For example, light blue and dark blue are acceptable.

I was allowed combine these colours any way that you like. For example, you can use only blue and green if you want to.

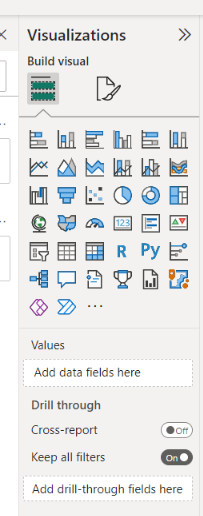
The company wanted to see the following items on the dashboard:

* The average Rotten Tomatoes ratings of each genre
* The number of movies produced per year
* The audience score for each film
* The profitability per studio
* The worldwide gross per genre

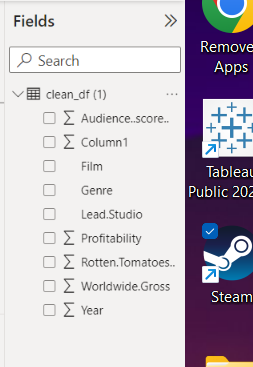
To begin with I changed the colour palette of the dashboard to meet the client’s requirements. I did this by changing the theme of the dashboard through the ‘View’ tab at the top of the screen. There I was able to customise my own theme:



I then created various visuals using the tab on the right side of the screen:



I then used the Fields tab to add variables to said visuals making them into a working chart.



Here are the 5 visuals I created and which variables were used:

* Doughnut Chart – Genre and Worldwide Gross
* Line Graph – Film and Year
* Clustered Bar Chart – Audience Score and Film
* Funnel – Genre and Rotten Tomatoes Rating
* Area Chart – Lead Studio and Profitability.

These 5 visuals show all the information that the client requested.

