Mamatha G

Abstract:

By using of R Language cleaning Data, Deletion of duplicate values. Importing cleaned CSV file into Power BI to build DashBoard.

R AND POWER BI Project

Dashboard

Contents

[Step1: Initial Exploratory Analysis 2](#_Toc130844602)

[Step 2: Clean Data 3](#_Toc130844603)

[Step 3: Exploratory Data Analysis 7](#_Toc130844604)

[Step 4: Export data 8](#_Toc130844605)

[Step 5: Create Power BI Dashboard 10](#_Toc130844606)

**Objective:** To analyze the performance of Hollywood movies Site

**Data:** Title, genre, studio, profitability, and ratings for movies released 2007-2012.

# Step1: Initial Exploratory Analysis

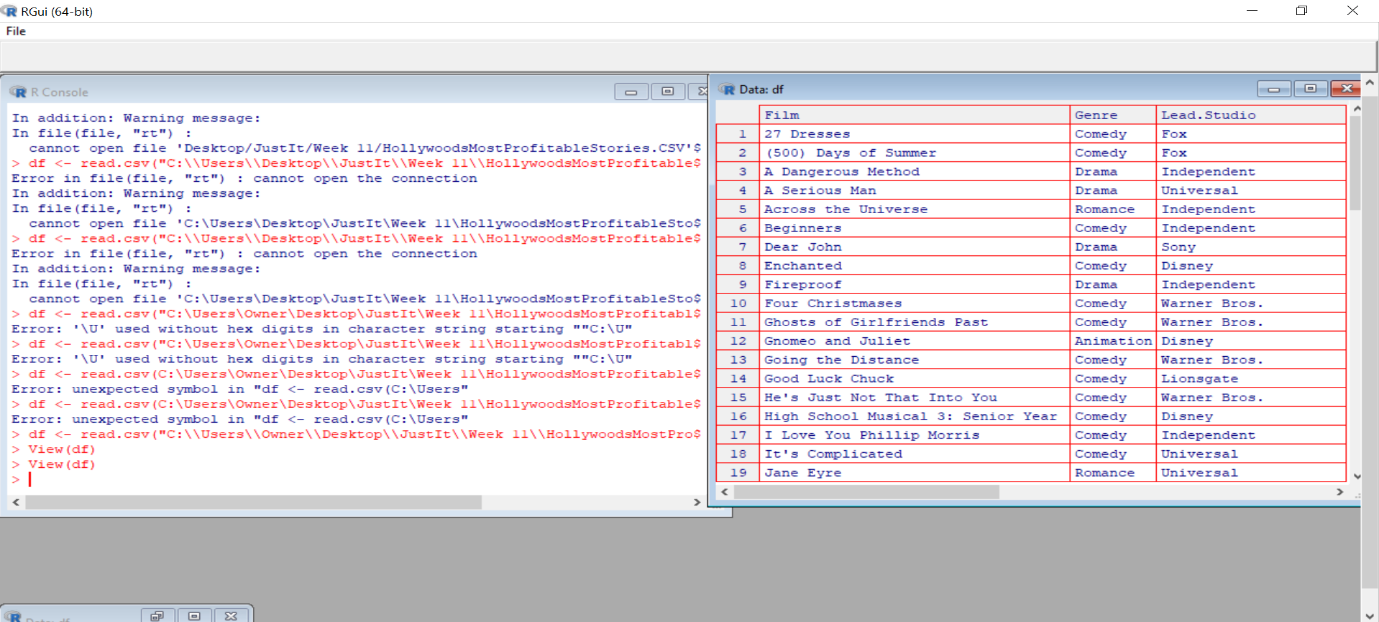
​ #Load data ​

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) ​​

Graphical user interface, text, application

Description automatically generated

# Step 2: Clean Data

# Check for missing values: ​

colSums(is.na(df)) ​

#Drop missing values ​

na.omit(df) ​

# check to make sure that the rows have been removed ​

colSums(is.na(df))​

​

​

Graphical user interface

Description automatically generated with medium confidence

#Check for duplicates ​

dim(df[duplicated(df$Film),])[1] ​

#round off values to 2 places ​

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

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

​#View(df) ​

​dim(df) ​

* Step 2.1: Outlier removal

#Check for outliers using a boxplot ​

​library(ggplot2) ​

#Create a boxplot that highlights the outliers ​

ggplot(df, aes(x=Profitability, y=Worldwide.Gross)) + geom\_boxplot(outlier.colou​

r = "red", outlier.shape = 1)+ scale\_x\_continuous(labels = scales::comma)+coord\_cartesian(ylim = c(0, 1000)) ​

​ Graphical user interface, application

Description automatically generated

**S #Remove outliers in 'Profitability' ​ Initial Exploratory Analysis**

#Remove outliers in 'Profitability' ​

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)) ​

​dim(no\_outliers)

Graphical user interface, text, application, email

Description automatically generated

# Remove outliers in 'Worldwide.Gross' ​

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)) ​ ​dim(df1) ​

**Step1: In** Graphical user interface, text, application

Description automatically generated

# Step 3: Exploratory Data Analysis

#Summary Statistics/Univariate Analysis: ​

summary(df1) ​

#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() ​

A picture containing graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

# Step 4: Export data

#Export clean data ​

write.csv(df1, "clean\_df.csv") ​

Graphical user interface, application

Description automatically generated

Import clean\_df in Power BI.

* Feel free to add more charts apart from the ones mentioned in the next slide.
* For the dashboard, the company would like you to use their brand colors which are blue, green and brown. You can use light or dark shades of each color. For example, light blue and dark blue are acceptable.

# Step 5: Create Power BI Dashboard

The average Rotten Tomatoes ratings of each genre​

Graphical user interface, application

Description automatically generated

* The number of movies produced per year ​

Graphical user interface, application, table, Excel, PowerPoint

Description automatically generated

The audience scores for each film.

Graphical user interface, chart, application, table

Description automatically generated

The profitability per studio

The worldwide gross per genre

Graphical user interface, application

Description automatically generated

Finally, Movies Analysis Report

Graphical user interface, application

Description automatically generated