Sunitha Pandirla

[sunithapandirla.bootcamp@justit.co.uk]

R & Power BI

Data cleaning & Visualization

*This document explains data cleaning with R program commands and data visualization with Power BI tool*

Contents

[**Step1:** InitialExploratory Analysis 3](#_Toc159873491)

[a) Load data in R studio as new R script file 3](#_Toc159873492)

[b) Take a look at the data 3](#_Toc159873493)

[c) Load library: 3](#_Toc159873494)

[**Step 2:** Clean data 5](#_Toc159873495)

[a) Check for missing Values: 5](#_Toc159873496)

[b) Drop missing (Null) values 5](#_Toc159873497)

[**Step 3:** Exploratory Data Analysis 8](#_Toc159873498)

[1) Summary Statistics 8](#_Toc159873499)

[2) Scatterplot 9](#_Toc159873500)

[3) Bar Chart 10](#_Toc159873501)

[**Step 4:** Export clean data 10](#_Toc159873502)

[a) Writing dataframe.R file to csv file 10](#_Toc159873503)

[**Step 5**: Import “clean\_df.csv” into Power BI Desktop 12](#_Toc159873504)

[a) Loading CSV File to Power BI Desktop 12](#_Toc159873505)

[**Step 6:** Create Power BI Dashboard 13](#_Toc159873506)

[a) Film Ratings, Profit and Worldwide Gross 13](#_Toc159873507)

[b) Top 10 Films (Profit, worldwide gross, Rotten Tomatoes, Audience score) 14](#_Toc159873508)

[c)Profit, Worldwide Gross, Rotten Tomatoes and Audience Score by Genre of the Film 15](#_Toc159873509)

[4) Profit, worldwide gross, Rotten Tomatoes and Audience score by Lead Studio 16](#_Toc159873510)

[5) Films Per Year and Their Audience Score 17](#_Toc159873511)

[**Advantages :** PowerBI Data reporting 18](#_Toc159873512)

**Aim:** This Assignment is to clean the data in ‘R’ and convert refined data into a csv file and

Visualize the data and analyze the performance of Hollywood movies in Power BI .

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

**Source**: [*InformationIsBeautiful.net*](https://www.informationisbeautiful.net/data/)

# **Step1:** InitialExploratory Analysis

## Load data in R studio as new R script file

*df1<- read.csv("https://public.tableau.com/app/sample-data/HollywoodsMostProfitableStories.csv")*

## Take a look at the data

*View(df1)*

A screenshot of a computer

Description automatically generated

## c) Load library:

*install.packages("tidyverse")*

*library(tidyverse)*

d)check data Types

*dim(df1)*

It shows that there are 74 rows and 8 columns in the data frame df1

*str(df1)*

This command gives the Data types of The different columns

A screenshot of a computer program

Description automatically generated

# **Step 2:** Clean data

## Check for missing Values:

*colSums(is.na(df1))*

This command checks for total number of null values per column in the data frame df1

A screenshot of a computer

Description automatically generated

## b) Drop missing (Null) values

check for number of Rows before dropping the null values.

*nrow(df1)*

There are 74 rows before dropping the null values.

A screenshot of a computer

Description automatically generated

*df<-na.omit(df1)*

This R command removes the rows which contain Null values from dataframe.

*nrow(df1)*

This command now shows 70 rows after removing the null values

A screenshot of a computer program

Description automatically generated

# **Step 3:** Exploratory Data Analysis

## Summary Statistics

*summary(df1)*

This gives summary of statistics of all columns with int data type (min, max, mean, median,1st quantile,3rd quantile).

A screenshot of a computer

Description automatically generated

## Scatterplot

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

A screenshot of a graph

Description automatically generated

## Bar Chart

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

A screenshot of a computer

Description automatically generated

# **Step 4:** Export clean data

## Writing dataframe.R file to csv file

*write.csv(df1, "clean\_df1.csv")*

This command writes the cleaned data frame in to csv file in your local computer folder.

A screenshot of a computer program

Description automatically generated

# **Step 5**: Import “clean\_df.csv” into Power BI Desktop

## a) Loading CSV File to Power BI Desktop

1. Use get (csv file) option in Power BI and load (cleaned\_df1.csv) file in power BI
2. Transform Year column data type from int to Year type in power query of Power BI

A screenshot of a computer

Description automatically generated

1. click close & apply and click on Report View
2. The data visualized graphically by using measures from clean\_df1 Table under “**Data**” pane where the numerical columns are aggregated automatically.
3. The Report View also has Dash board pages, Filters and Visualizations pane

# **Step 6:** Create Power BI Dashboard

## **Film Ratings, Profit and Worldwide Gross**

A screenshot of a computer

Description automatically generated

The above dashboard consists of 5 slides.

1. Average Rotten Tomatoes score of the film genre (Pie chart)
2. Profitability by Lead studio produced the film (Clustered column chart)

-Highest for Independent, followed by Disney studios

1. Audience scores for each film (Multi row card)
2. Sum of worldwide gross per genre (Treemap)

-Highest for comedy, followed by Romance.

1. Count of film by Year (Donut chart)

## **Top 10 Films (Profit, worldwide gross, Rotten Tomatoes, Audience score)**

A screenshot of a computer

Description automatically generated

## **c)Profit, Worldwide Gross, Rotten Tomatoes and Audience Score by Genre of the Film**

A screenshot of a computer

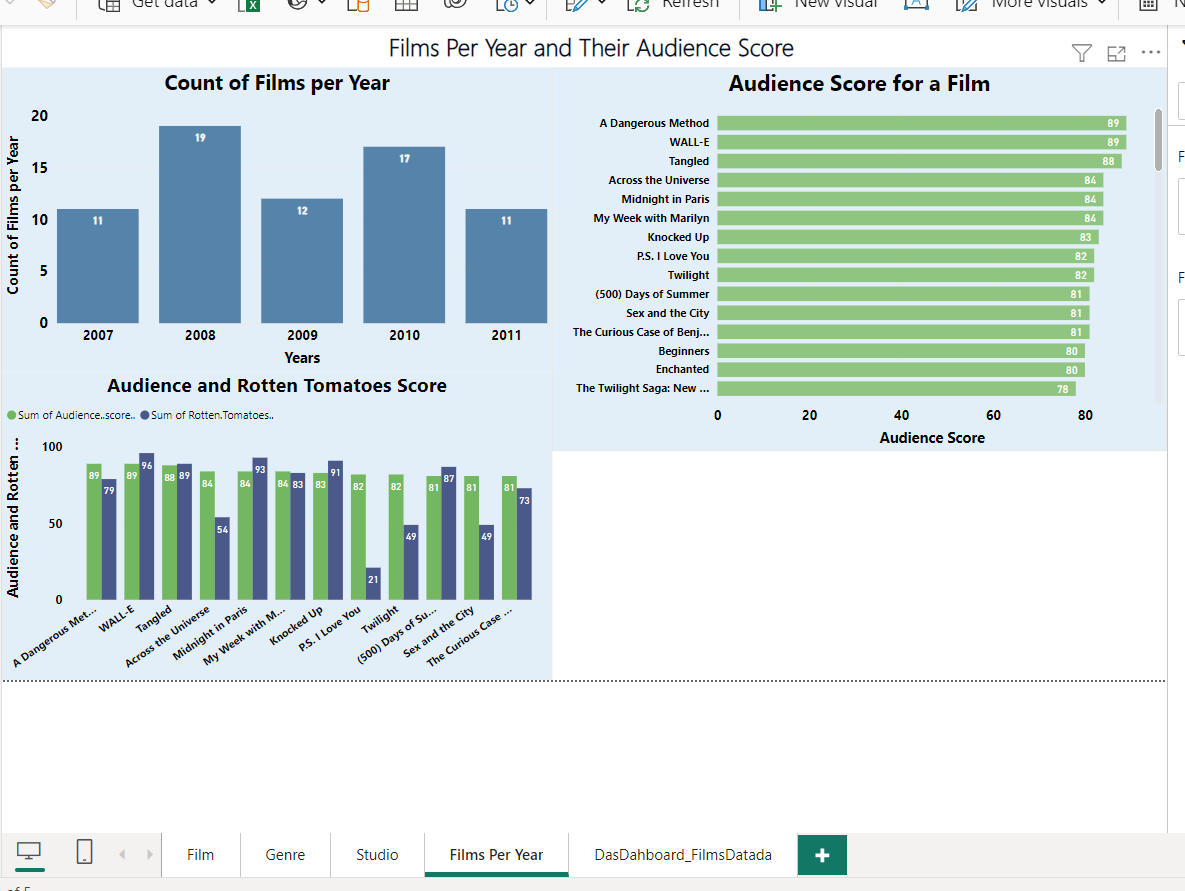
Description automatically generated

## **Profit, worldwide gross, Rotten Tomatoes and Audience score by Lead Studio**

A screenshot of a computer

Description automatically generated

## **Films Per Year and Their Audience Score**



# **Advantages :** PowerBI Data reporting

**Data Visualization**: Power BI allows users to create interactive and visually appealing reports and dashboards, making it easier to understand complex data at a glance.

**Integration**: It seamlessly integrates with various data sources, including Excel, SQL databases, cloud services like Azure, and many others, allowing users to pull data from multiple sources into a single report.

**Real-Time Analytics**: With its ability to connect to live data sources, Power BI enables users to perform real-time analytics, providing insights into the latest trends and patterns