R and Power BI Assignment Hollywood Movies

ABSTRACT

Analyzing the performance of Hollywood movies.

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

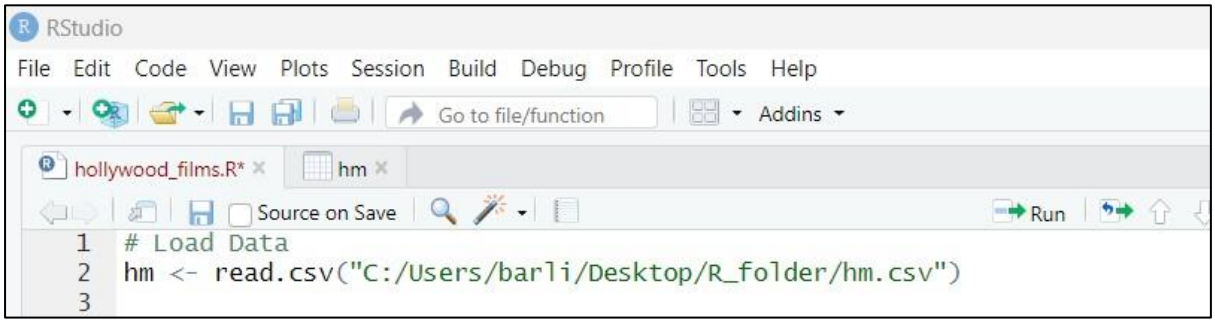
Busra Arlier

GLA 16 Data Technician

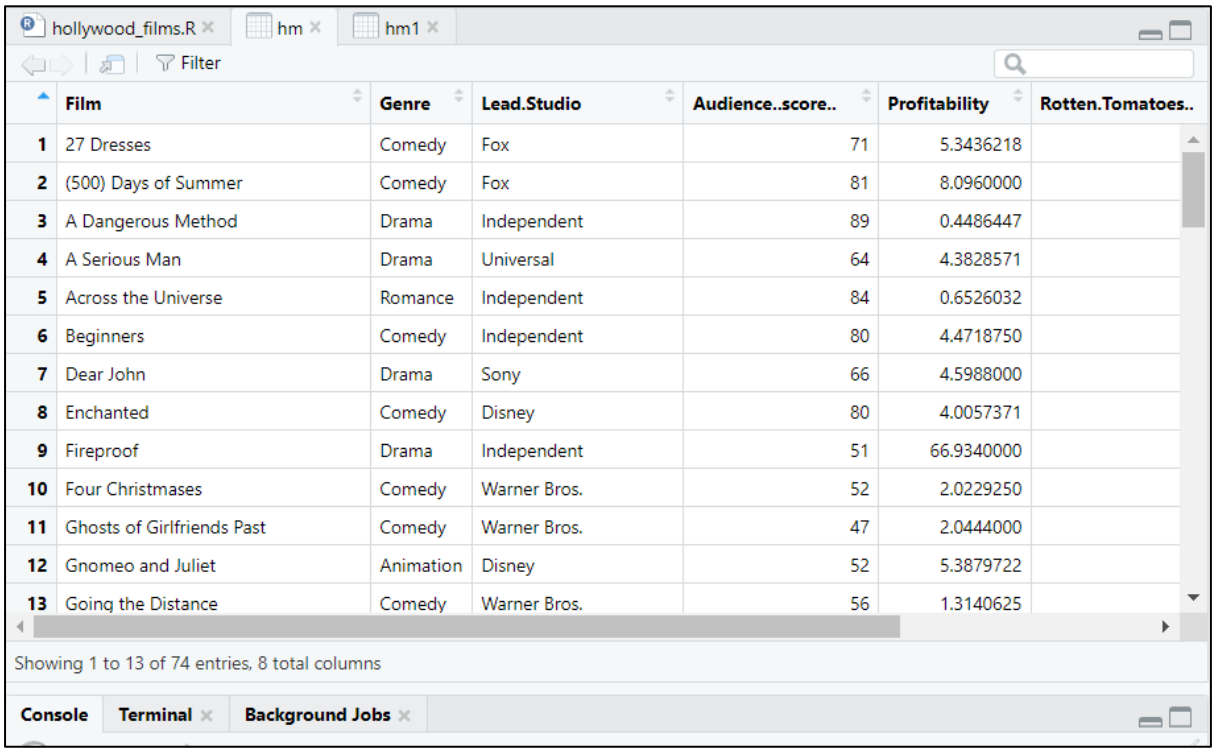
Contents

Exploratory Analysis

Task 1 – Load Data



Task 2- Take a look at the data



	Film	Genre	Lead.Studio	Audience..score..	Profitability	Rotten.Tomatoes..
1	27 Dresses	Comedy	Fox	71	5.3436218	
2	(500) Days of Summer	Comedy	Fox	81	8.0960000	
3	A Dangerous Method	Drama	Independent	89	0.4486447	
4	A Serious Man	Drama	Universal	64	4.3828571	
5	Across the Universe	Romance	Independent	84	0.6526032	
6	Beginners	Comedy	Independent	80	4.4718750	
7	Dear John	Drama	Sony	66	4.5988000	
8	Enchanted	Comedy	Disney	80	4.0057371	
9	Fireproof	Drama	Independent	51	66.9340000	
10	Four Christmases	Comedy	Warner Bros.	52	2.0229250	
11	Ghosts of Girlfriends Past	Comedy	Warner Bros.	47	2.0444000	
12	Gnomeo and Juliet	Animation	Disney	52	5.3879722	
13	Going the Distance	Comedy	Warner Bros.	56	1.3140625	

Task 3- Load Library and Import Library

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

hollywood_films.R hm hm1

Source on Save Run Source

```
#Take a look at the data
View(hm)

#Load Library
install.packages("tidyverse")

#Import library
library(tidyverse)
```

13:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.1 · ~/

```
> install.packages("tidyverse")
WARNING: Rtools is required to build R packages but is not currently installed. Please
download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/barli/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/tidyverse_2.0.0.zip'
Content type 'application/zip' length 430894 bytes (420 KB)
downloaded 420 KB

package 'tidyverse' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\barli\AppData\Local\Temp\RtmpaAkrDx\downloaded_packages
> #Import library
> library(tidyverse)
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.3     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.3     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.0
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✗ dplyr::filter() masks stats::filter()
✗ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package to force all conflicts to become errors
>
```

Task 4- Check Data Types

```r
# Load Data
hm <- read.csv("C:/Users/barli/Desktop/R_folder/hm.csv")

#Take a look at the data
View(hm)

#Load Library
install.packages("tidyverse")

#Import library
library(tidyverse)


# Check data types
str(hm)
```

```
✓ dplyr     1.1.3     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.3     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.0
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✗ dplyr::filter() masks stats::filter()
✗ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package to force all conflicts to become errors
> # Check data types
> str(hm)
'data.frame':   74 obs. of  8 variables:
 $ Film             : chr  "27 Dresses" "(500) Days of Summer" "A Dangerous Method" "A
Serious Man" ...
 $ Genre            : chr  "Comedy" "Comedy" "Drama" "Drama" ...
 $ Lead.Studio      : chr  "Fox" "Fox" "Independent" "Universal" ...
 $ Audience..score..: int  71 81 89 64 84 80 66 80 51 52 ...
 $ Profitability    : num  5.344 8.096 0.449 4.383 0.653 ...
 $ Rotten.Tomatoes..: int  40 87 79 89 54 84 29 93 40 26 ...
 $ Worldwide.Gross  : num  160.31 60.72 8.97 30.68 29.37 ...
 $ Year             : int  2008 2009 2011 2009 2007 2011 2010 2007 2008 2008 ...
>
```

Task 5- Check for Missing Values

```
> # Check missing values
> colSums(is.na(hm))
             Film            Genre       Lead.Studio Audience..score..
                0                0                 0                 1
    Profitability Rotten.Tomatoes..   Worldwide.Gross             Year
                3                1                 0                 0
>
```

Task 6- Drop Missing Values

```
> # Drop missing values
> hm1 <- hm %>% drop_na()
>
```

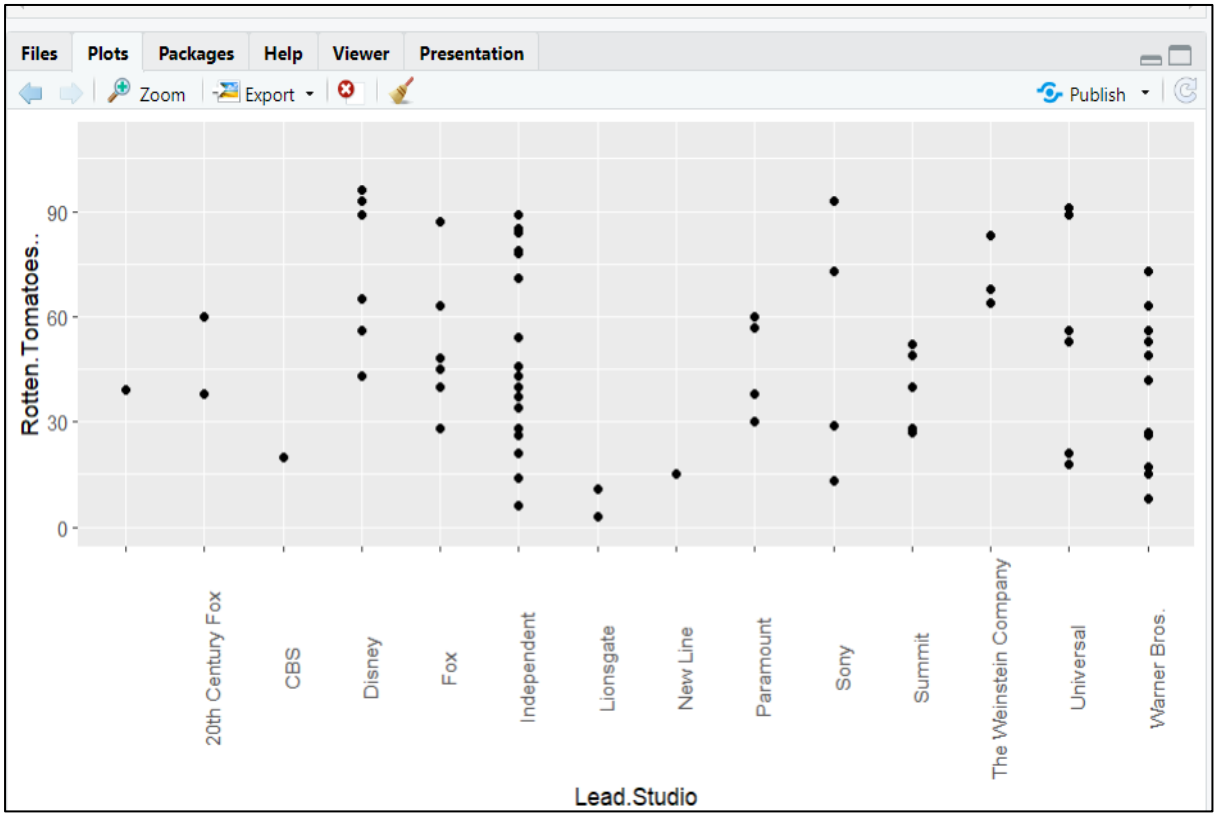
```
R ▾   Global Environment ▾
Data
hm      74 obs. of 8 variables
hm1     70 obs. of 8 variables
```

Task 7- Summary Statistics

```
> #Summary Statistics
> summary(hm1)
     Film              Genre           Lead.Studio        Audience..score..
 Length:70          Length:70          Length:70          Min.   :35.00
 Class :character   Class :character   Class :character   1st Qu.:53.25
 Mode  :character   Mode  :character   Mode  :character   Median :64.50
                                                          Mean   :64.46
                                                          3rd Qu.:75.50
                                                          Max.   :89.00
 Profitability    Rotten.Tomatoes.  Worldwide.Gross        Year
 Min.   : 0.005   Min.   : 3.00     Min.   :  0.025   Min.   :2007
 1st Qu.: 1.802   1st Qu.:27.25     1st Qu.: 32.809   1st Qu.:2008
 Median : 2.646   Median :45.50     Median : 85.891   Median :2009
 Mean   : 4.785   Mean   :47.76     Mean   :141.933   Mean   :2009
 3rd Qu.: 4.977   3rd Qu.:64.75     3rd Qu.:202.467   3rd Qu.:2010
 Max.   :66.934   Max.   :96.00     Max.   :709.820   Max.   :2011
>
```

Task 8- Scatterplot Chart

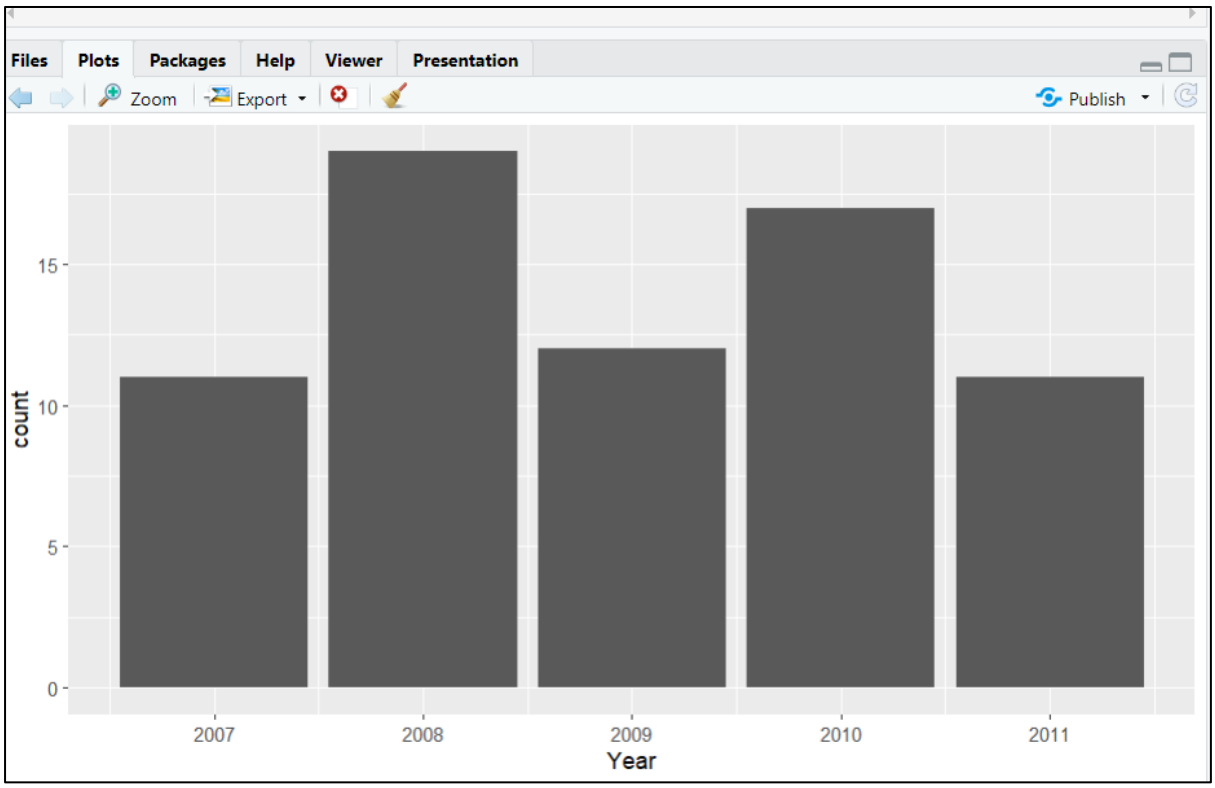
```
> ggplot(hm1,aes(x=Lead.Studio, y=Rotten.Tomatoes..))+geom_point()+scale_y_continuous(l
abels=scales::comma)+coord_cartesian(ylim=c(0,110))+theme(axis.text.x=element_text(angl
e=90))
>
```



Task 9- Bar Chart

```
#Bar Chart

ggplot(hm1,aes(x=Year))+geom_bar()
```

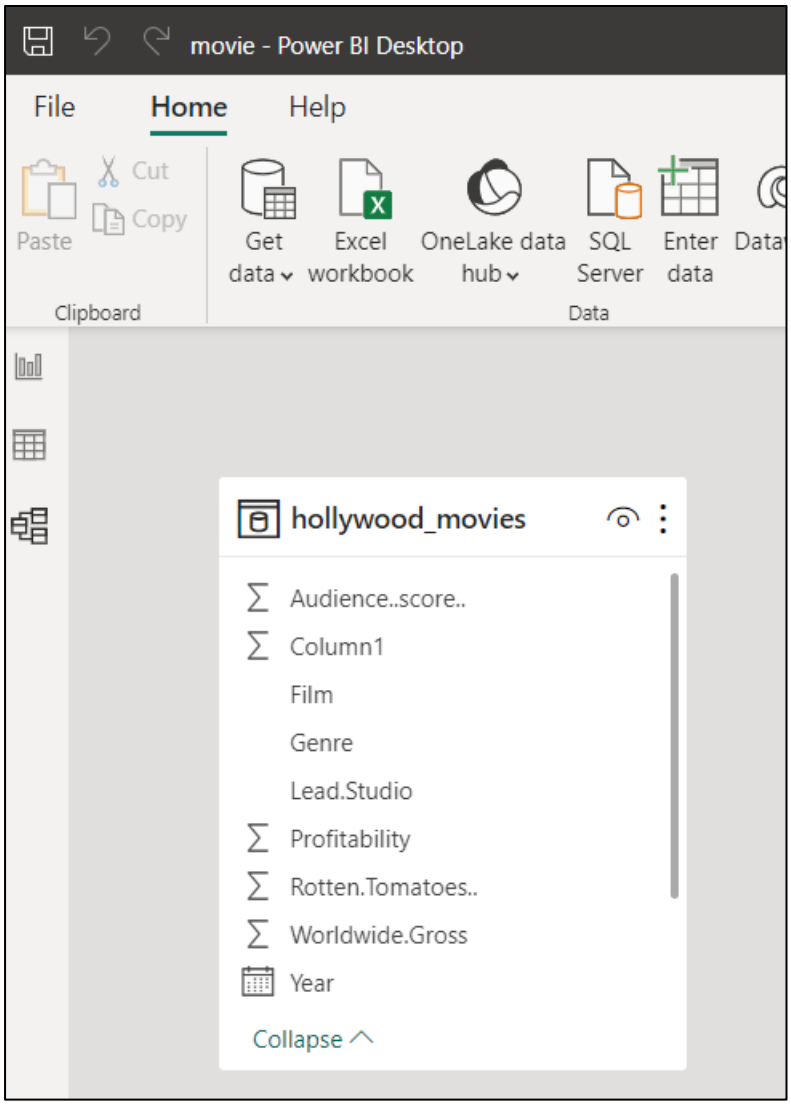


Task 10- Export Clean Data

```
#Export clean data
write.csv(hm1, "clean_hm.csv")
```

Create Power BI Dashboard

Task 1- Import clean_df in Power BI



Task 2- Power BI Dashboard

The company would like to use its brand colours which are blue, green, and brown.

The client would like to see the below analysis in the dashboard:

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

Reflection

In this project, R language is used for data preprocessing. R has a wide range of packages (e.g., dplyr, tidyr) that are useful for data preprocessing tasks. Data preprocessing involves cleaning, transforming, and organizing the data to make it suitable for analysis.

After preprocessing the data, I performed EDA to gain insights into the Hollywood movies dataset. R offers various visualization libraries like ggplot2 and Plotly that help to create informative charts and graphs.

Once I have performed the analysis in R and generated insights, I moved on to creating a Power BI dashboard. Power BI is a user-friendly tool for building interactive and visually appealing dashboards. I started by importing my pre-processed and analyzed data from R into Power BI. Power BI supports various data sources, including Excel, SQL databases, and web services. I like creating visuals using Power BI because it allows you to add interactive features like slicers, filters, and drill-through actions, enabling users to explore the data further.

In conclusion, combining R for data analysis with Power BI for data visualization and dashboard creation can be a powerful approach for gaining insights into Hollywood movies' performance and trends. It allowed me to explore the data, perform statistical analysis, and present the results in an easily understandable and interactive format.