Handbook for Bootcamp. JustIT Learning outcome. Main features on R and Power Bi (Dashboard)

Data Visualizations

Author : Farhan Khan (Learner)

Farhan Khan

Contents

Dashboard Using R and Power Bi

First Task

Policies and Procedures

A Data Protection Policy is **a statement that sets out how your organisation protects personal data**. It is a set of principles, rules and guidelines that informs how you will ensure ongoing compliance with data protection law.

The data remains confidential and should not be given to any unauthorised, user. Unless stated to do so.

Following rules must be adhered while using data and liabilities of Data analyst.

- Lawfulness, fairness and transparency.
- Purpose limitation.
- Data minimisation.
- Accuracy.

Second (Using R software) Task

The data can be retrieved from the following link.

https://public.tableau.com/app/sample-data/HollywoodsMostProfitableStories.csv

Steps involved (Storytelling):

The raw data was downloaded and ready to go through for the process. R software is useful for data analysing and it's precise tuning for better presentation. User can see it in better format with lowest chance of data being wrongly analysed.

Step1: Initial Exploratory Analysis

Following command is to load data in R (software)

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

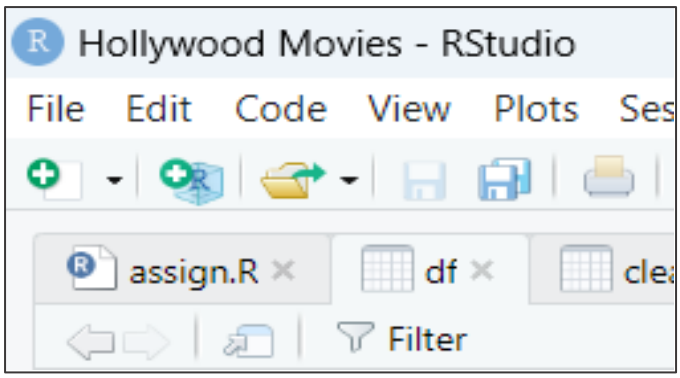


FIGURE 1

Data Frame (Df) :

This diagram represents the df (data frame) in R software

	Film	Genre	Lead.Studio	Audience..score..	Profitability	Rotten.Tomatoes..	Worldwide.Gross	Year
62	Twilight	Romance	Summit	82	10.1800270	49	376	column 7: numeric
63	Twilight: Breaking Dawn	Romance	Independent	68	6.3833636	26	702.170000	2011
64	Tyler Perry's Why Did I get Married	Romance	Independent	47	3.7241924	46	55.862886	2007
65	Valentine's Day	Comedy	Warner Bros.	54	4.1840385	17	217.570000	2010
66	Waiting For Forever	Romance	Independent	53	0.0050000	6	0.025000	2011
67	Waitress	Romance	Independent	67	11.0897415	89	22.179483	2007
68	WALL-E	Animation	Disney	89	2.8960191	96	521.283432	2008
69	Water For Elephants	Drama	20th Century Fox	72	3.0814211	60	117.094000	2011
70	What Happens in Vegas	Comedy	Fox	72	6.2676470	28	219.367646	2008
71	When in Rome	Comedy	Disney	44	NA	15	43.040000	2010
72	You Will Meet a Tall Dark Stranger	Comedy	Independent	35	1.2118182	43	26.660000	2010
73	Youth in Revolt	Comedy	The Weinstein Company	52	1.0900000	68	19.620000	2010
74	Zack and Miri Make a Porno	Romance	The Weinstein Company	70	1.7475417	64	41.941000	2008

Showing 62 to 74 of 74 entries, 8 total columns

FIGURE 2 # TAKE A LOOK AT THE DATA: **VIEW(DF)**

This snapshot represents the 74 rows and 7 Columns. Some rows missing values. I will ascertain and delete the rows which are incomplete and doing data violation.

Data validation is utmost important, when we build the graphs and charts in order to visualise. We want to make sure that user receive accurate information.

```
package 'tidyverse' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\khanf\AppData\Local\Temp\Rtmps5LYFA\downloaded_packages
>
```

FIGURE 3 #LOAD LIBRARY:

INSTALL.PACKAGES("TIDYVERSE")

Install Packages in Software

Above command install the package the Tidyverse . This command uses in the R software.

```
#Import library

library(tidyverse)
```

```
Console   Terminal   Background Jobs
R 4.2.2 · ~/Hollywood Movies/
Content type 'application/zip' length 428944 bytes (418 KB)
downloaded 418 KB

package 'tidyverse' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\khanf\AppData\Local\Temp\Rtmps5LYFA\downloaded_packages
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
✓ ggplot2 3.4.0     ✓ purrr   1.0.1
✓ tibble  3.1.8     ✓ dplyr   1.1.0
✓ tidyr   1.3.0     ✓ stringr 1.5.0
✓ readr   2.1.3     ✓ forcats 1.0.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✗ dplyr::filter() masks stats::filter()
✗ dplyr::lag()    masks stats::lag()
>
```

FIGURE 4 #IMPORT LIBRARY LIBRARY(TIDYVERSE)

This command loads some useful packages e.g ggplot which plot graphs later on stages.

```
# Check data types

str(df)
?dim(df)
```

```
Conflicts ------------------------------------------------------------ tidyverse_conflicts() -
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> str(df)
'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 ...
> dim(df)
[1] 74  8
> ?dim(df)
>
```

FIGURE 5 # CHECK DATA TYPES: STR(DF)

Check Data Types in R software.

The above command helps us see data structure of the current table. This data has 8 variables.

```
# Check for missing values

colSums(is.na(df))
```

```
 $ 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 ...
> dim(df)
[1] 74  8
> ?dim(df)
> colSums(is.na(df))
            Film            Genre     Lead.Studio Audience..score..     Profitability Rotten.Tomatoes..   Worldwide.Gross
               0                0                0               1                 3                 1                 0
            Year
               0
>
```

FIGURE 6 # CHECK FOR MISSING VALUES: COLSUMS(IS.NA(DF))

By looking into the tables, I can see that 5 rows have missing values. This is shown in the above snapshot.

73	Youth in Revolt	Comedy	The Weinstein Company	52	1.0900000	68	19.620000	2010
74	Zack and Miri Make a Porno	Romance	The Weinstein Company	70	1.7475417	64	41.941000	2008

Showing 55 to 70 of 70 entries, 8 total columns

```
            Film            Genre     Lead.Studio Audience..score..     Profitability Rotten.Tomatoes..   Worldwide.Gross
               0                0                0               1                 3                 1                 0
            Year
               0
> # Check for missing values and drop rows from the table
> df <- na.omit(df)
> View(df)
> colSums(is.na(df))
            Film            Genre     Lead.Studio Audience..score..     Profitability Rotten.Tomatoes..   Worldwide.Gross
               0                0                0               0                 0                 0                 0
            Year
               0
>
```

FIGURE 7

Check for missing values and drop rows from the table df <- na.omit(df). This command removes the rows which has Null values.

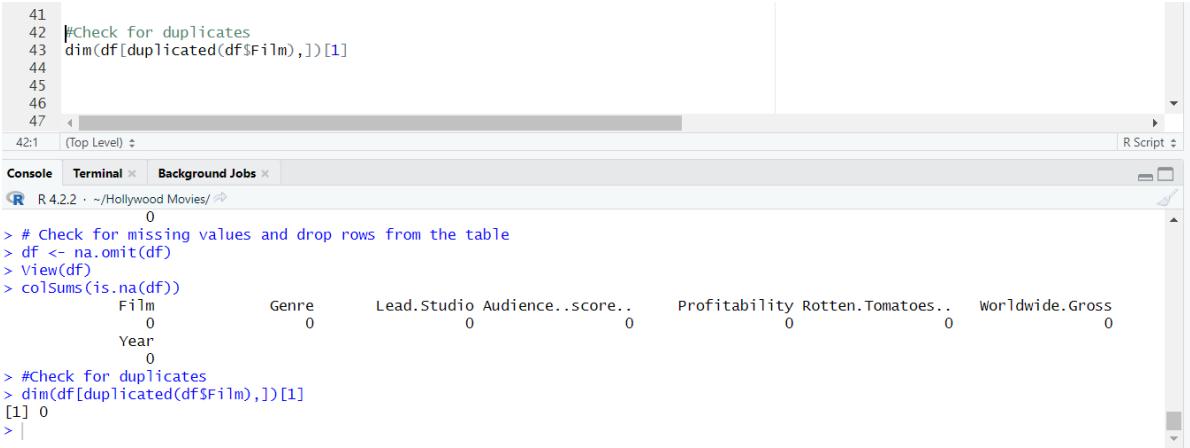


FIGURE 8

Check for Duplicates

Check for duplicates dim(df[duplicated(df$Film),])[1]

After running this command, I found NO duplicates.

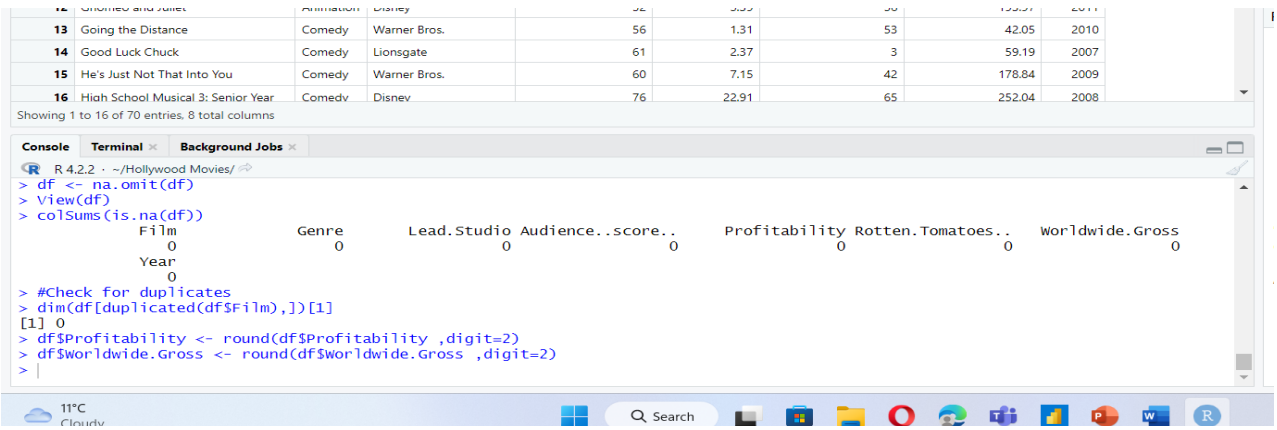


FIGURE 9

Round off values to 2 places

This is important part of analysis. Round up to two decimal values (Profitability --- column and Worldwide.Gross----- Column). This helps us to analyze the data.

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

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

The above commands round the decimal point to two. This helps us to visualize the charts and graphs.

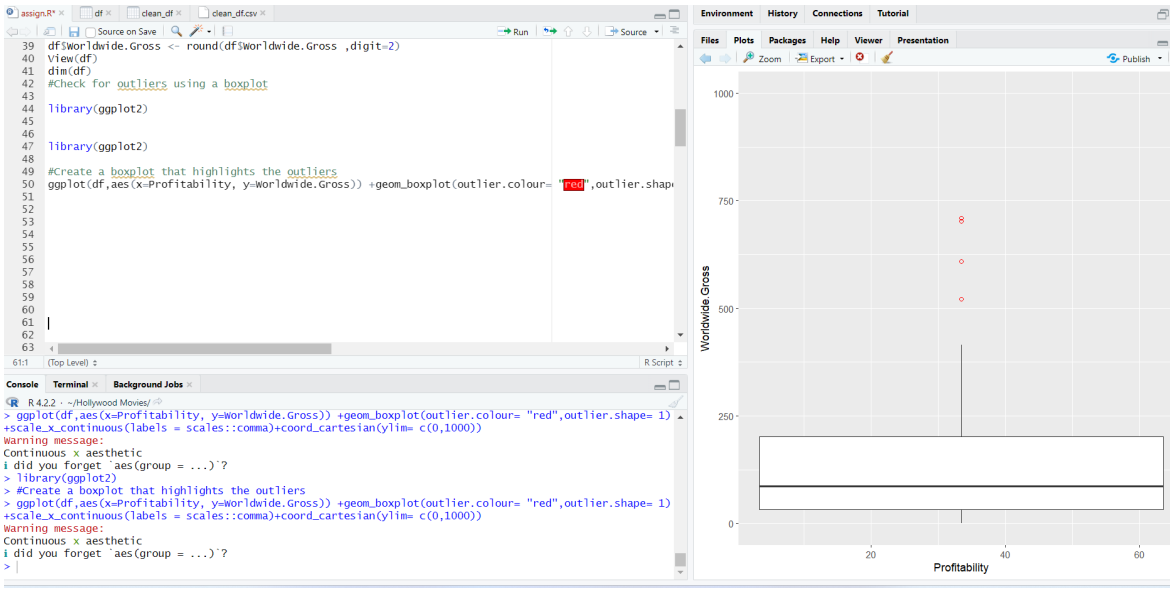


FIGURE 10

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.colour = "red", outlier.shape = 1)+ scale_x_continuous(labels = scales::comma)+coord_cartesian(ylim = c(0, 1000))

the above commands use to plot a graph as shown in above graphs.

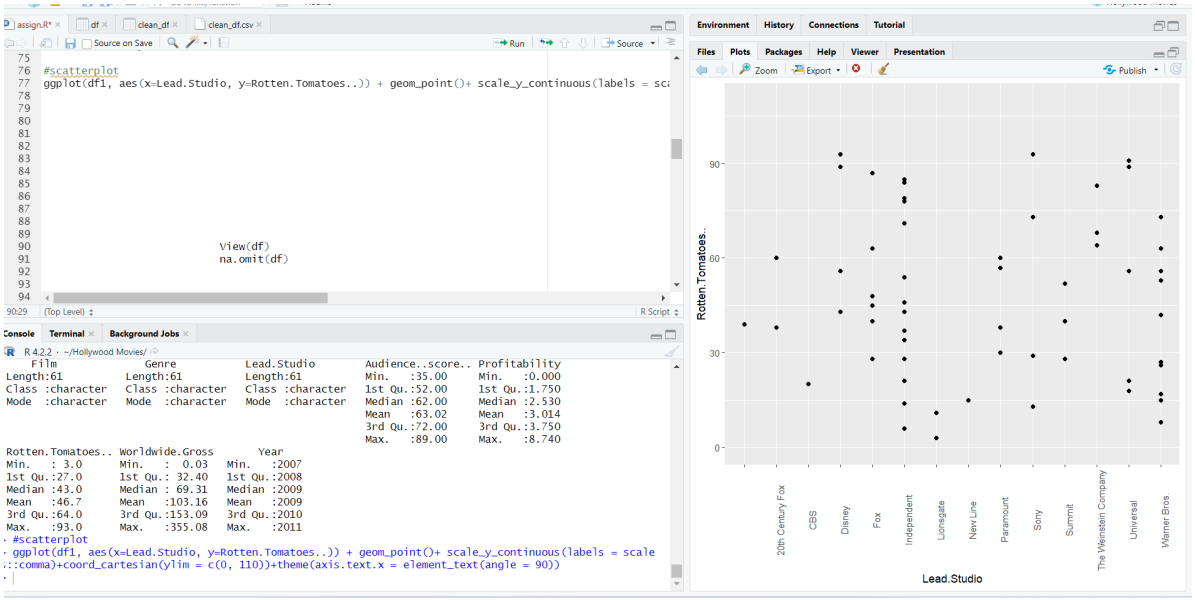


FIGURE 11

Birvariate 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))

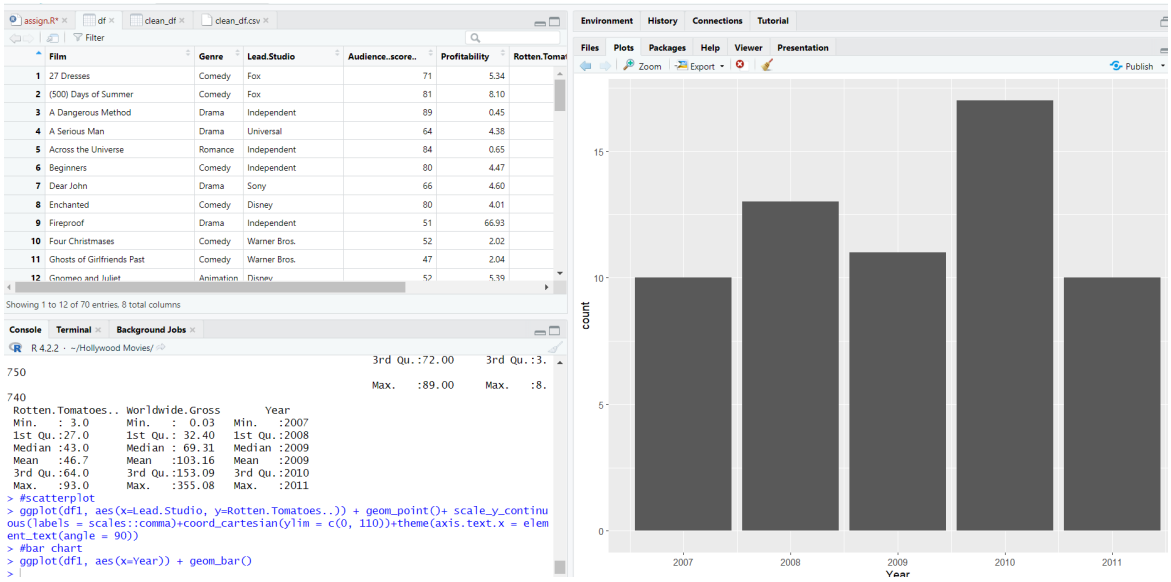


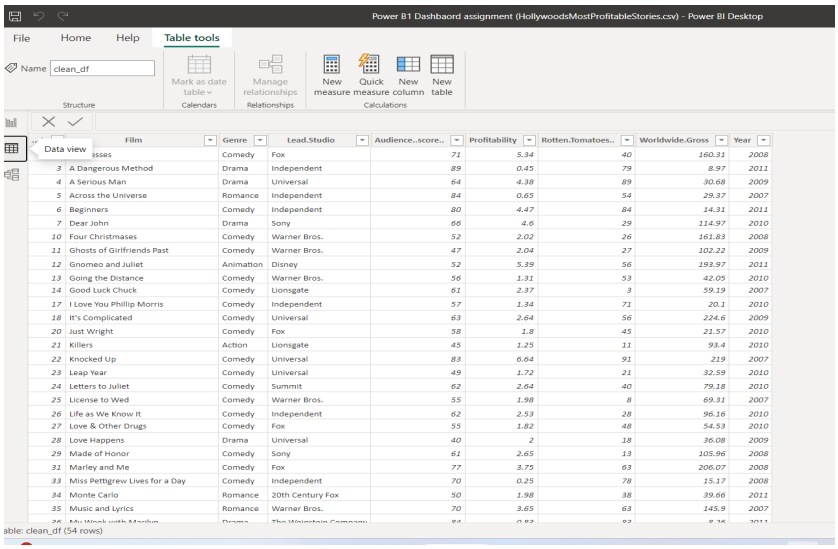
FIGURE 12

Summary and Bar graph in R Software

Power Bi:

The Power Bi has three main different views (Report view, Data view and Model view).

This will help us in order to visualize (in shape of graphs, charts and other related views).



	Film	Genre	Lead.Studio	Audience..score..	Profitability	Rotten.Tomatoes..	Worldwide.Gross	Year
	sses	Comedy	Fox	71	5.34	40	160.31	2008
3	A Dangerous Method	Drama	Independent	89	0.45	79	8.97	2011
4	A Serious Man	Drama	Universal	64	4.38	89	30.68	2009
5	Across the Universe	Romance	Independent	84	0.65	54	29.37	2007
6	Beginners	Comedy	Independent	80	4.47	84	14.31	2011
7	Dear John	Drama	Sony	66	4.6	29	114.97	2010
10	Four Christmases	Comedy	Warner Bros.	52	2.02	26	161.83	2008
11	Ghosts of Girlfriends Past	Comedy	Warner Bros.	47	2.04	27	102.22	2009
12	Gnomeo and Juliet	Animation	Disney	52	5.39	56	193.97	2011
13	Going the Distance	Comedy	Warner Bros.	56	1.31	53	42.05	2010
14	Good Luck Chuck	Comedy	Lionsgate	61	2.37	3	59.19	2007
17	I Love You Phillip Morris	Comedy	Independent	57	1.34	71	20.1	2010
18	It's Complicated	Comedy	Universal	63	2.64	56	224.6	2009
20	Just Wright	Comedy	Fox	58	1.8	45	21.57	2010
21	Killers	Action	Lionsgate	45	1.25	11	93.4	2010
22	Knocked Up	Comedy	Universal	83	6.64	91	219	2007
23	Leap Year	Comedy	Universal	49	1.72	21	32.59	2010
24	Letters to Juliet	Comedy	Summit	62	2.64	40	79.18	2010
25	License to Wed	Comedy	Warner Bros.	55	1.98	8	69.31	2007
26	Life as We Know It	Comedy	Independent	62	2.53	28	96.16	2010
27	Love & Other Drugs	Comedy	Fox	55	1.82	48	54.53	2010
28	Love Happens	Drama	Universal	40	2	18	36.08	2009
29	Made of Honor	Comedy	Sony	61	2.65	13	105.96	2008
31	Marley and Me	Comedy	Fox	77	3.75	63	206.07	2008
33	Miss Pettigrew Lives for a Day	Comedy	Independent	70	0.25	78	15.17	2008
34	Monte Carlo	Romance	20th Century Fox	50	1.98	38	39.66	2011
35	Music and Lyrics	Romance	Warner Bros.	70	3.65	63	145.9	2007

FIGURE 13

Importing Data(Hollywoods Most Profitable Stories) :

This picture shows after importing data from the excel in (csv) format.

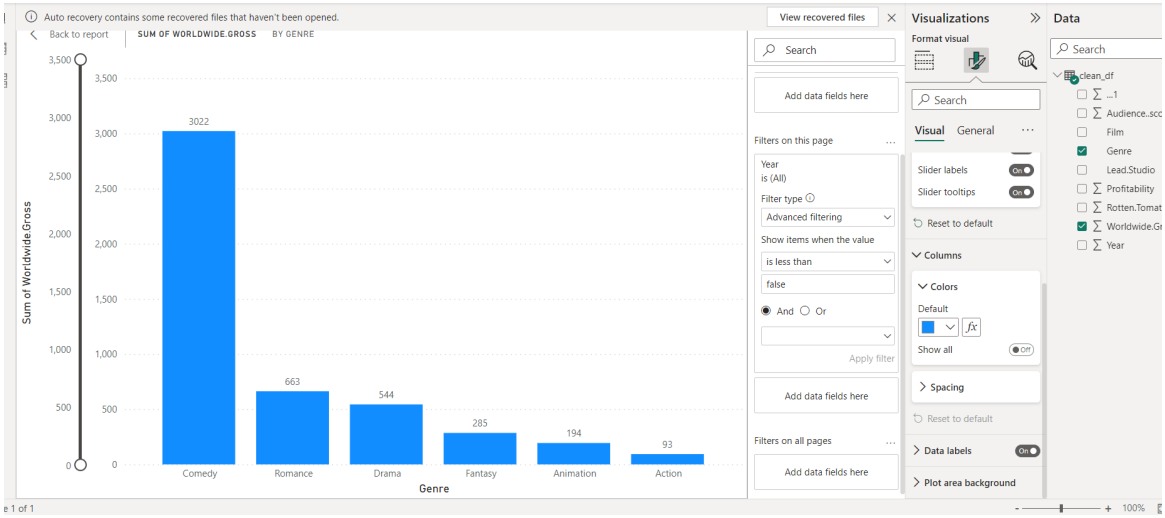


FIGURE 14

Bar Graph in Focus mode

This is comprehensive diagram showing all the labels, with tooltip, zoom slider and the numbers showing on the graphs. With the aid of zoom slider, we can lift the bar chart (up/down) in order to make it more better.

- This graph aims to analyze the the average Rotten Tomatoes ratings of each genre

Aim: To analyze the performance of Hollywood movies

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

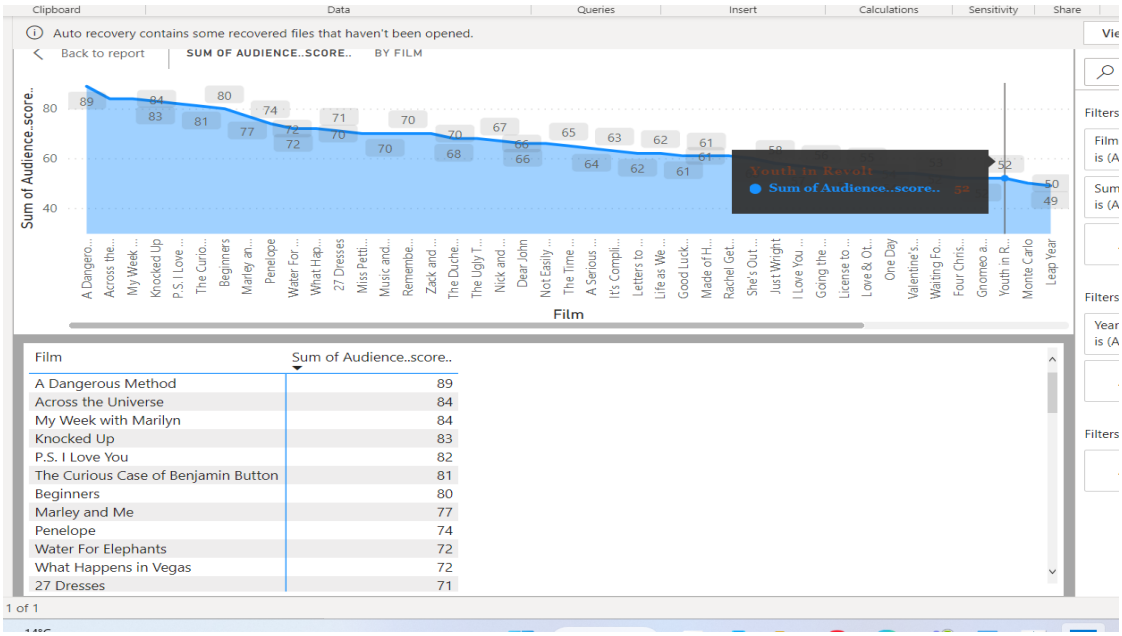


FIGURE 15 area plot graph

Area Graph with table

This picture shows the Focus mode in Power Bi . Focus mode is one the powerful tool along with Table . It really helps us to visualize the graph as well as the graph. Text also shown in the graph as a number to show actual Sum of audience score by films.

- This graph aims to analyze the audience score for each film

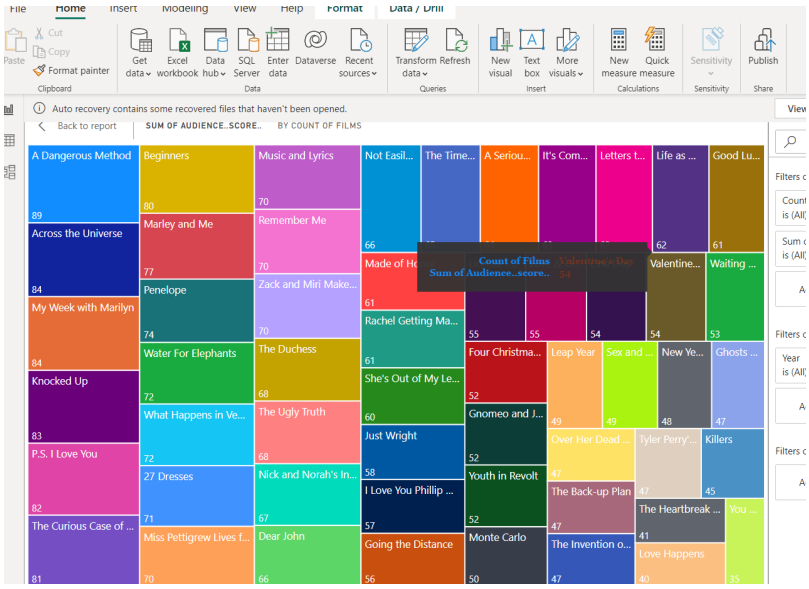


FIGURE 16

This Matrix shows Audience score for each film.

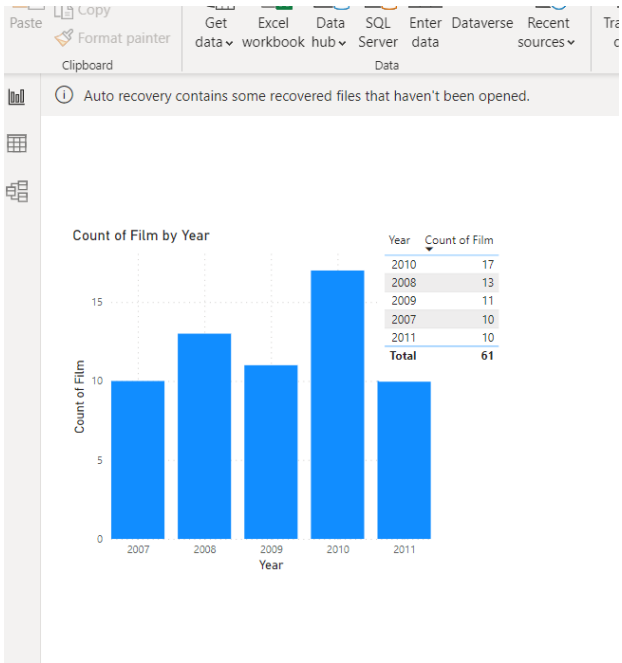


FIGURE 17 CLUSTERED COLUMN CHART

Clustered Column Chart table

This picture shows the Focus mode in Power Bi . Focus mode is one the powerful tool along with Table. It really helps us to visualize the graph as well as the graph. Text also shown in the graph as a number to show actual Years and Count of Films/year.

- This graph aims to analyze the number of movies produced per year.

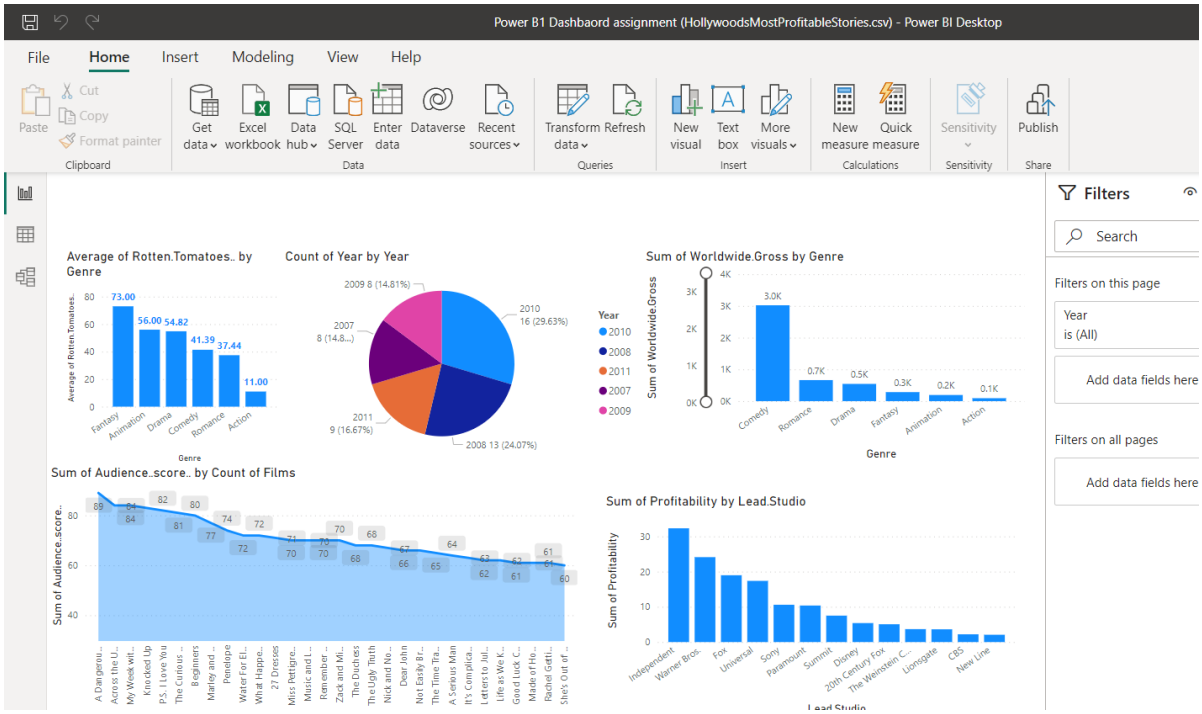


FIGURE 18 DASHBOARD IN POWER BI

Dashbaord in Power Bi

This picture shows the overall view in Power Bi . These are total of 5 graphs. These graphs are to analyze the performances of Hollywood movies according to client's requirements.

Aim: To analyze the performance of Hollywood movies

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

Assignment 2 (R and Power Bi Data Visualisation)

https://app.powerbi.com/groups/me/reports/51e1f119-7aa0-49df-8bd7-65c271316c43/ReportSection