Assignment 3: DATA ANALYSIS

R AND POWERBI

NOORULAIN FAHAD

Table of Contents

INTRODUCTION

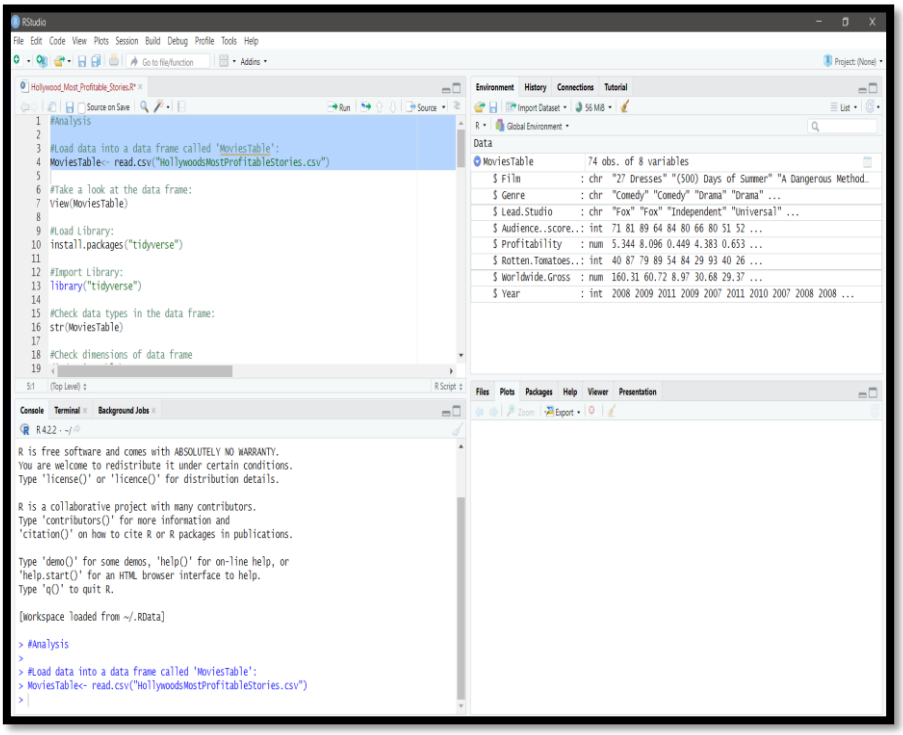
The aim of this project is to analyse the performance of Hollywood movies. The dataset used for this project is 'Hollywood's Most Profitable Stories' which contains the Title, Genre, Lead Studio, Profitability and Rotten Tomatoes Ratings for movies released between 2007 and 2012 in CSV format. The source of this dataset is given below :

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

TASK – 1 : DATA ANALYSIS IN R

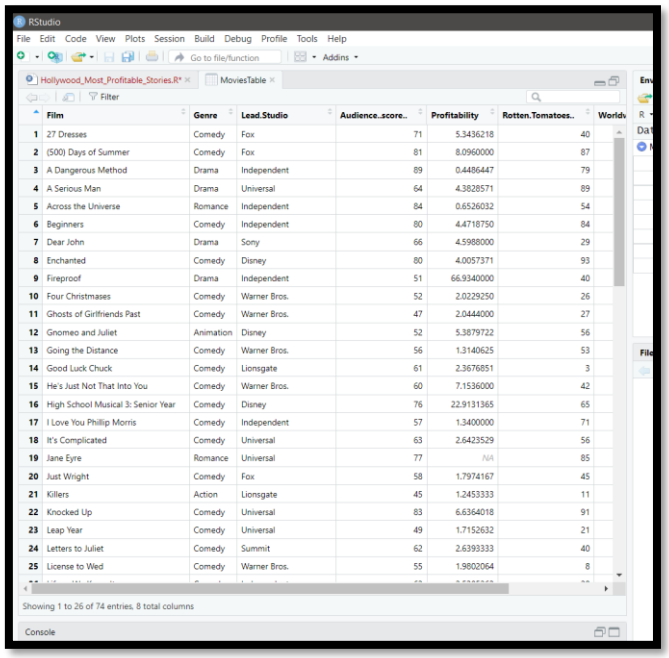
To perform data analysis in R, the workflow consists of importing the data, removing missing values, removing duplicates, removing outliers, feature engineering (creating new column using existing data), summary statistics, EDA(Exploratory Data Analysis) and export the cleaned data.

1. IMPORT DATASET To import the dataset in R studio, "read.csv()" command is used. It reads the file and stores it into a data frame. In current scenario, the dataset is imported and stored into a data frame called "MoviesTable".



2. VIEWING THE DATA FRAME To view the data frame, "View(MoviesTable) " command is used. To run a command, you can either press "alt + Enter" or simply click the "Run" on top pane of the editor. The View command generated the data frame with column

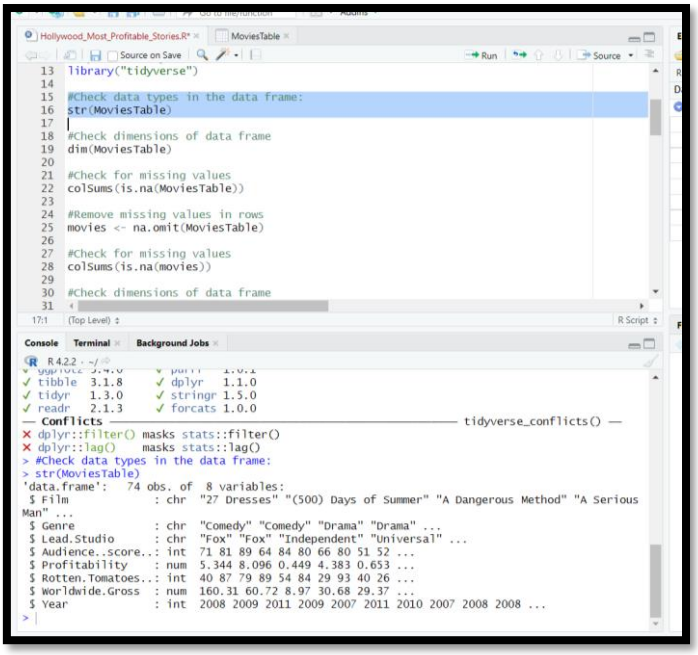
headers Film, Genre, Lead Studio, Audience score, Profitability, Rotten Tomatoes, Worldwide Gross and Year as shown below.



	Film	Genre	Lead.Studio	Audience..score..	Profitability	Rotten.Tomatoes..	Worldv
1	27 Dresses	Comedy	Fox	71	5.3436218	40	
2	(500) Days of Summer	Comedy	Fox	81	8.0960000	87	
3	A Dangerous Method	Drama	Independent	89	0.4486447	79	
4	A Serious Man	Drama	Universal	64	4.3828571	89	
5	Across the Universe	Romance	Independent	84	0.6526032	54	
6	Beginners	Comedy	Independent	80	4.4718750	84	
7	Dear John	Drama	Sony	66	4.5988000	29	
8	Enchanted	Comedy	Disney	80	4.0057371	93	
9	Fireproof	Drama	Independent	51	66.9340000	40	
10	Four Christmases	Comedy	Warner Bros.	52	2.0229250	26	
11	Ghosts of Girlfriends Past	Comedy	Warner Bros.	47	2.0444000	27	
12	Gnomeo and Juliet	Animation	Disney	52	5.3879722	56	
13	Going the Distance	Comedy	Warner Bros.	56	1.3140625	53	
14	Good Luck Chuck	Comedy	Lionsgate	61	2.3676851	3	
15	He's Just Not That Into You	Comedy	Warner Bros.	60	7.1536000	42	
16	High School Musical 3: Senior Year	Comedy	Disney	76	22.9131365	65	
17	I Love You Phillip Morris	Comedy	Independent	57	1.3400000	71	
18	It's Complicated	Comedy	Universal	63	2.6423529	56	
19	Jane Eyre	Romance	Universal	77	NA	85	
20	Just Wright	Comedy	Fox	58	1.7974167	45	
21	Killers	Action	Lionsgate	45	1.2453333	11	
22	Knocked Up	Comedy	Universal	83	6.6364018	91	
23	Leap Year	Comedy	Universal	49	1.7152632	21	
24	Letters to Juliet	Comedy	Summit	62	2.6393333	40	
25	License to Wed	Comedy	Warner Bros.	55	1.9802064	8	

3. INSTALLING TIDYVERSE R has integrated packages which are designed to work together to make common data science operations more user-friendly. The "Tidyverse" suite of integrated packages have functions used for data wrangling, tidying, reading/writing, parsing, and visualizing. The commands "install.pacakages(tidyverse)" and "library(tidyverse)" are used to install this package in R studio.

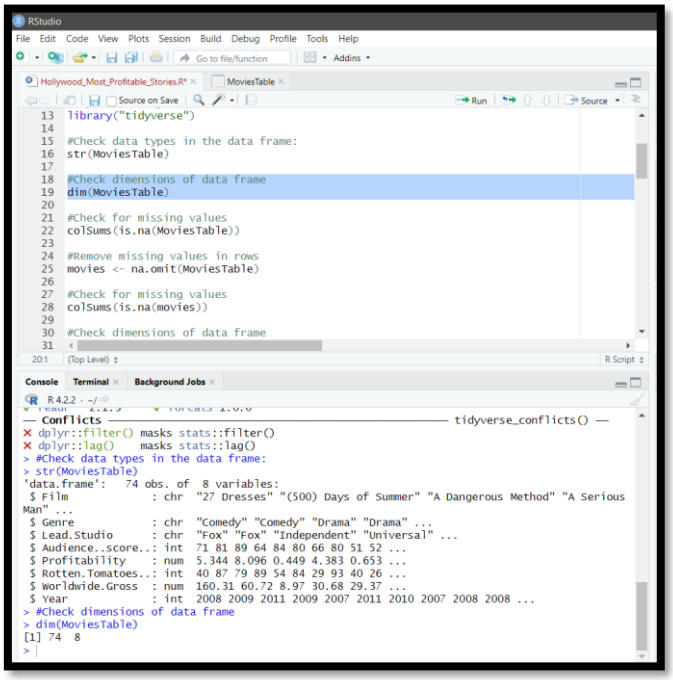
4. CHECK DATA TYPES Checking the data type of all the fields in the dataset is a foundation step in performing data analysis. It aligns with the principles of data validation and integrity. To perform this in R, "str(MoviesTable)" command is used.



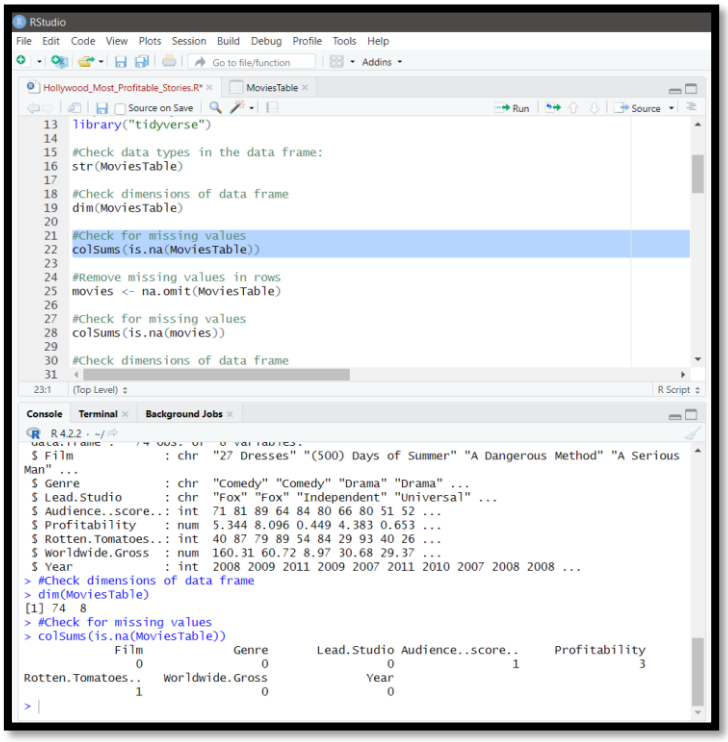
```
library("tidyverse")

#Check data types in the data frame:
str(MoviesTable)

#Check dimensions of data frame
dim(MoviesTable)

#Check for missing values
colSums(is.na(MoviesTable))

#Remove missing values in rows
movies <- na.omit(MoviesTable)

#Check for missing values
colSums(is.na(movies))

#Check dimensions of data frame
```

```
✓ 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()
> #Check data types in the data frame:
> str(MoviesTable)
'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 ...
```

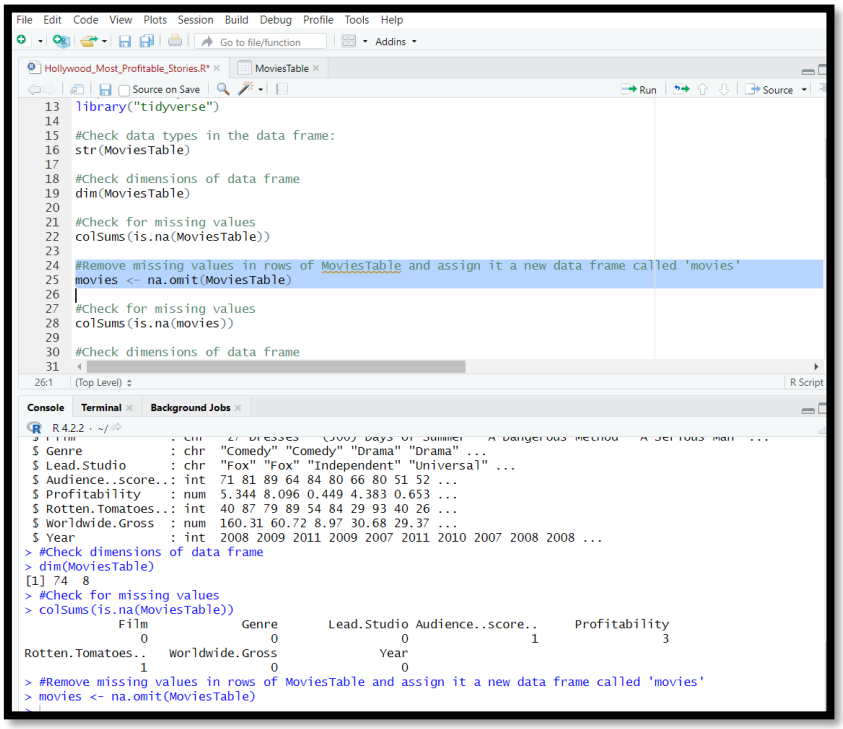
5. CHECKING DIMESIONS OF DATA FRAME To check the dimensions of the data frame "MoviesTable", the command "dim(MoviesTable)" is used. It shows that our data frame has 74 rows and 8 columns.



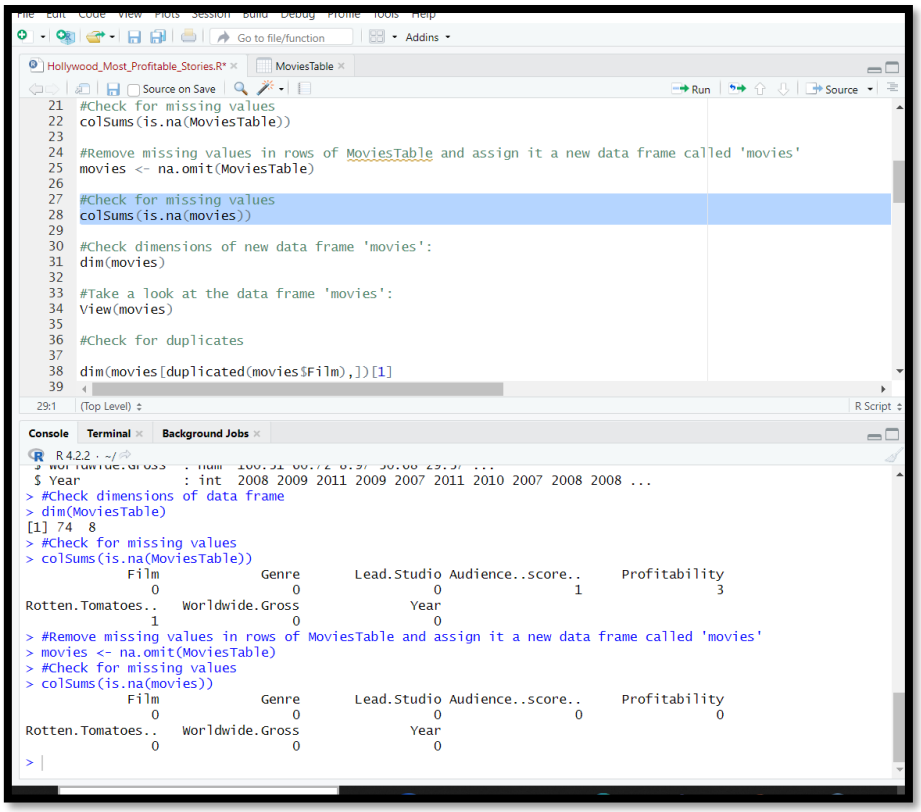
6. CHECK FOR MISSING VALUES The next step is to check for missing values in the dataset before performing the analysis. To check for missing values, "colSums(is.na(MoviesTable) " command is used. After running this command, you get the number of missing values in each column of the table. There are 1, 3 and 1 missing values in columns Audience score, Profitability and Rotten Tomatoes columns respectively, as shown below.



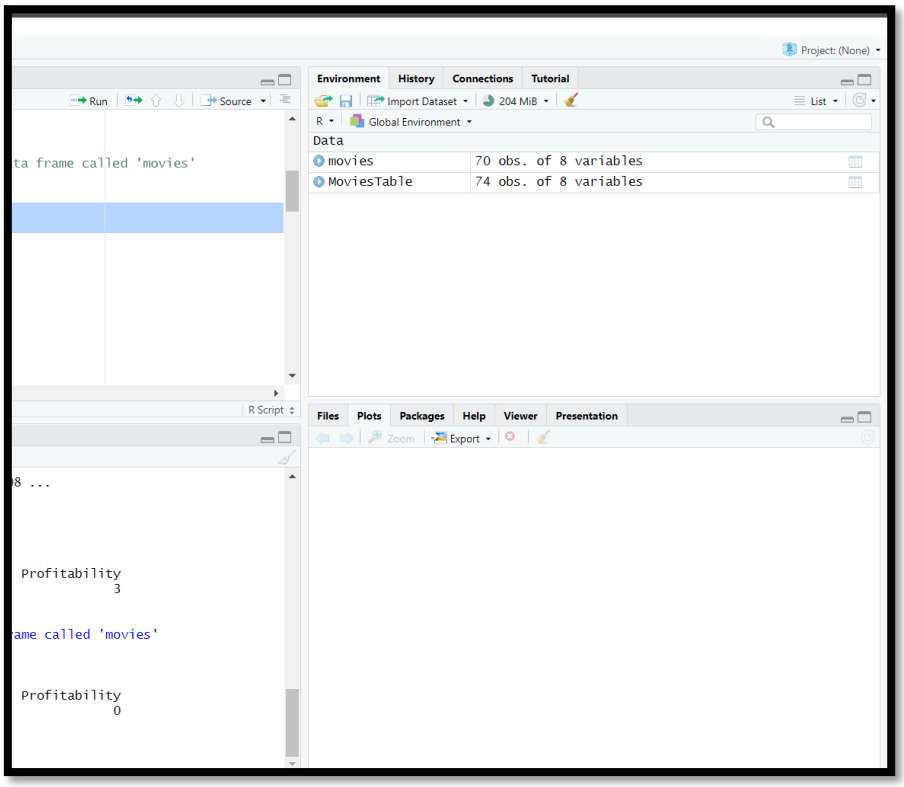
7. REMOVE MISSING VALUES The next step is to remove the missing values in the rows of the data frame MoviesTable and assign the name "Movies" to the new data frame. The command "na.omit()" is used.



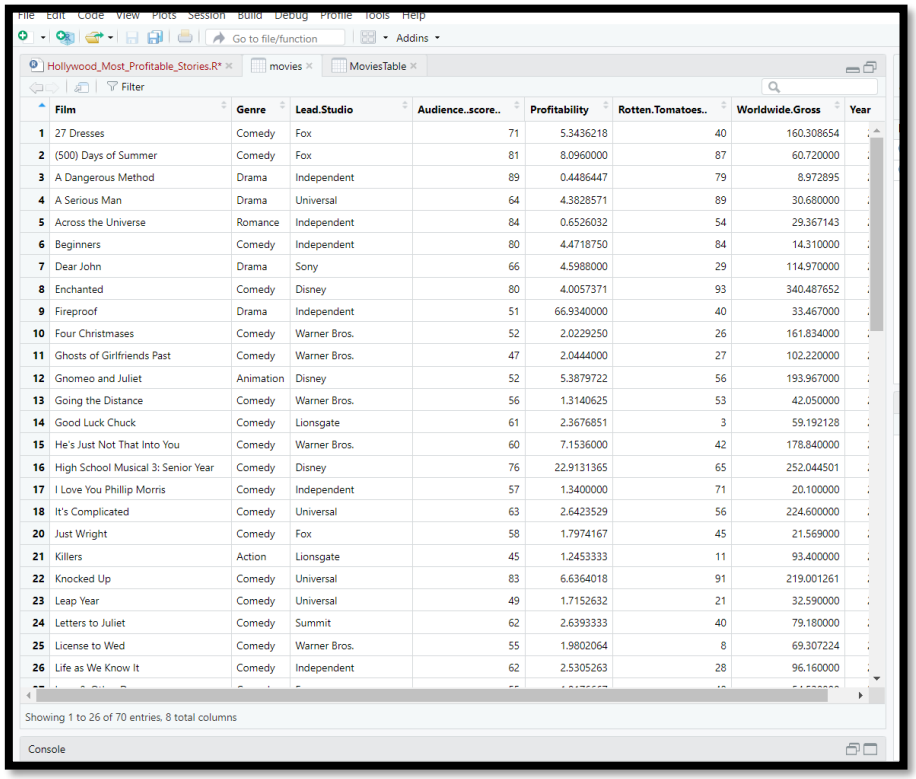
After this, we check for missing values again in the new data frame "Movies" and there are no missing values as shown below.



The dimensions of the new data frame "movies" are now 70 rows and 8 columns as shown below.

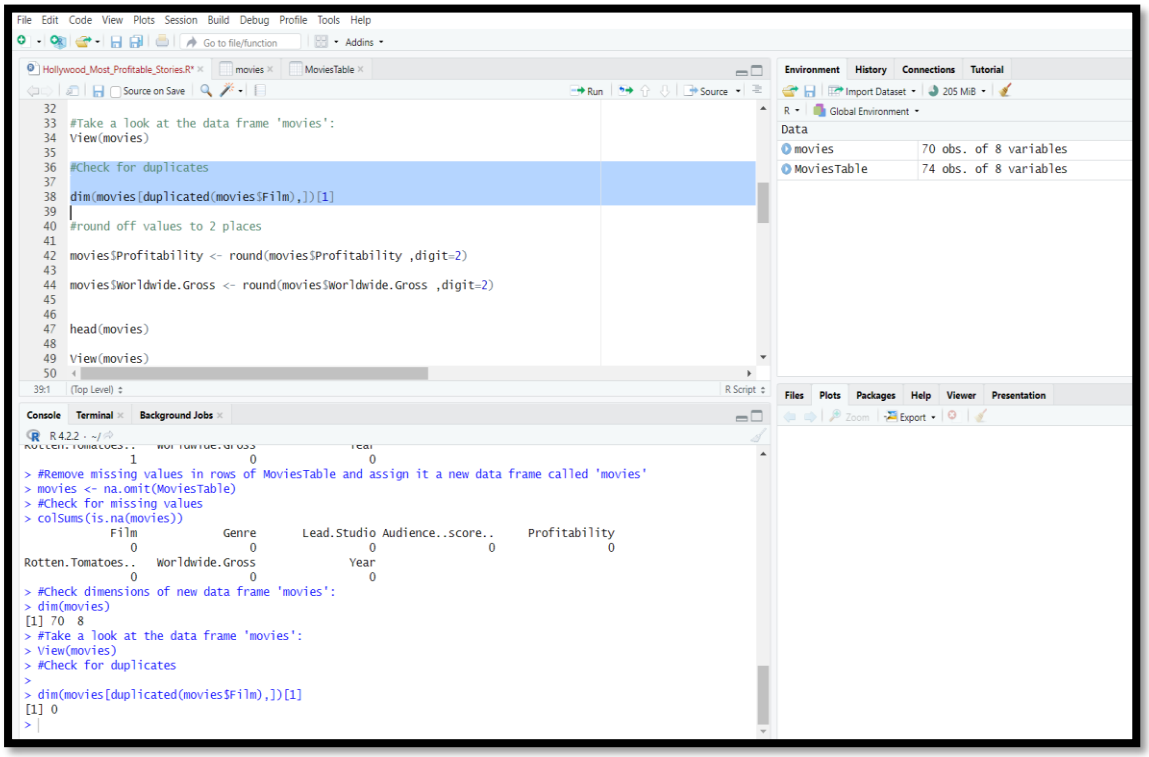


The new data frame movies with no missing values is shown as follows.

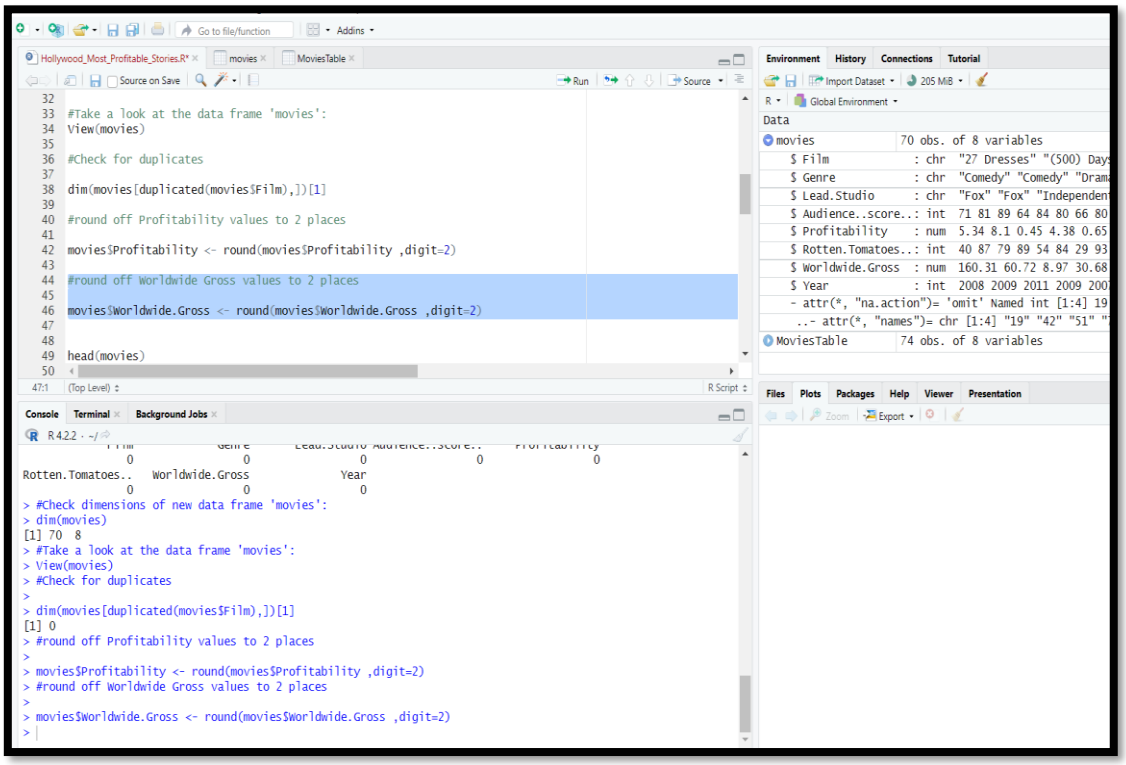


	Film	Genre	Lead.Studio	Audience..score..	Profitability	Rotten.Tomatoes..	Worldwide.Gross	Year
1	27 Dresses	Comedy	Fox	71	5.3436218	40	160.308654	
2	(500) Days of Summer	Comedy	Fox	81	8.0960000	87	60.720000	
3	A Dangerous Method	Drama	Independent	89	0.4486447	79	8.972895	
4	A Serious Man	Drama	Universal	64	4.3828571	89	30.680000	
5	Across the Universe	Romance	Independent	84	0.6526032	54	29.367143	
6	Beginners	Comedy	Independent	80	4.4718750	84	14.310000	
7	Dear John	Drama	Sony	66	4.5988000	29	114.970000	
8	Enchanted	Comedy	Disney	80	4.0057371	93	340.487652	
9	Fireproof	Drama	Independent	51	66.9340000	40	33.467000	
10	Four Christmases	Comedy	Warner Bros.	52	2.0229250	26	161.834000	
11	Ghosts of Girlfriends Past	Comedy	Warner Bros.	47	2.0444000	27	102.220000	
12	Gnomeo and Juliet	Animation	Disney	52	5.3879722	56	193.967000	
13	Going the Distance	Comedy	Warner Bros.	56	1.3140625	53	42.050000	
14	Good Luck Chuck	Comedy	Lionsgate	61	2.3676851	3	59.192128	
15	He's Just Not That Into You	Comedy	Warner Bros.	60	7.1536000	42	178.840000	
16	High School Musical 3: Senior Year	Comedy	Disney	76	22.9131365	65	252.044501	
17	I Love You Phillip Morris	Comedy	Independent	57	1.3400000	71	20.100000	
18	It's Complicated	Comedy	Universal	63	2.6423529	56	224.600000	
20	Just Wright	Comedy	Fox	58	1.7974167	45	21.569000	
21	Killers	Action	Lionsgate	45	1.2453333	11	93.400000	
22	Knocked Up	Comedy	Universal	83	6.6364018	91	219.001261	
23	Leap Year	Comedy	Universal	49	1.7152632	21	32.590000	
24	Letters to Juliet	Comedy	Summit	62	2.6393333	40	79.180000	
25	License to Wed	Comedy	Warner Bros.	55	1.9802064	8	69.307224	
26	Life as We Know It	Comedy	Independent	62	2.5305263	28	96.160000	

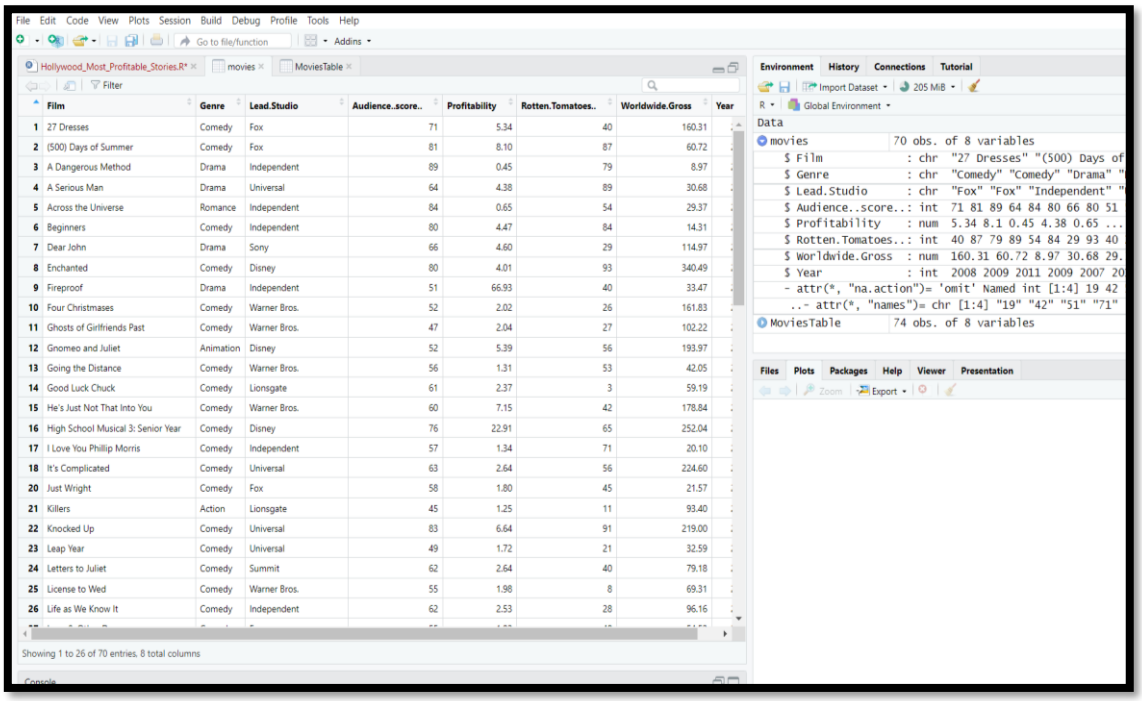
8. CHECK DUPLICATES The next step is to check for duplicate records in the data frame. The command used is shown below and you can see there are no duplicate records in the dataset.



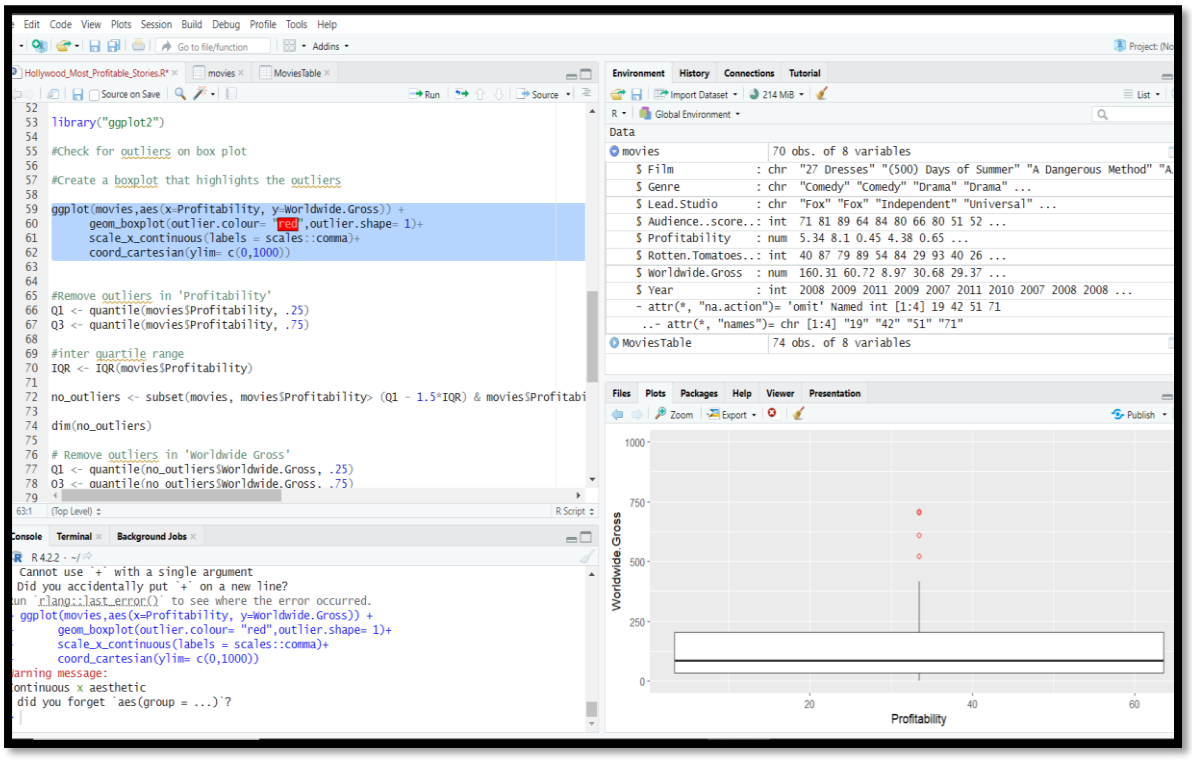
9. ROUND-OFF VALUES The next step is to round off the values of variable Profitability and Worldwide Gross to two decimal places to make the numerical data consistent.



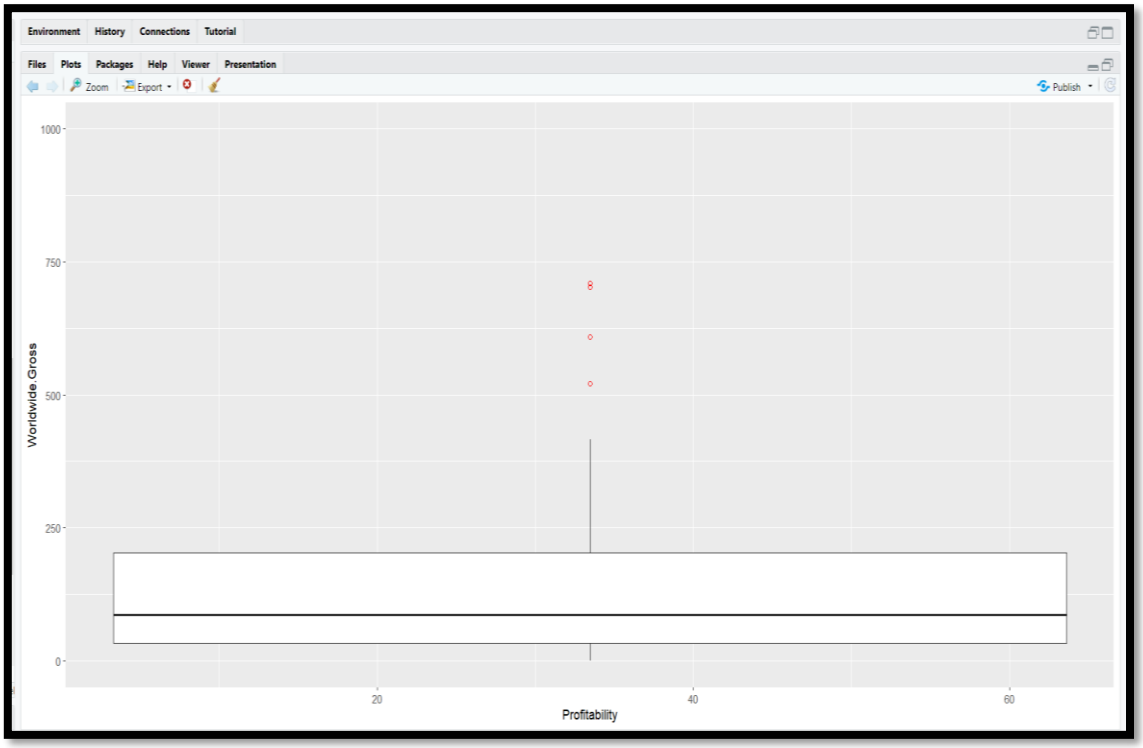
You can now see that the Profitability and Worldwide gross columns both have values with two decimal places.



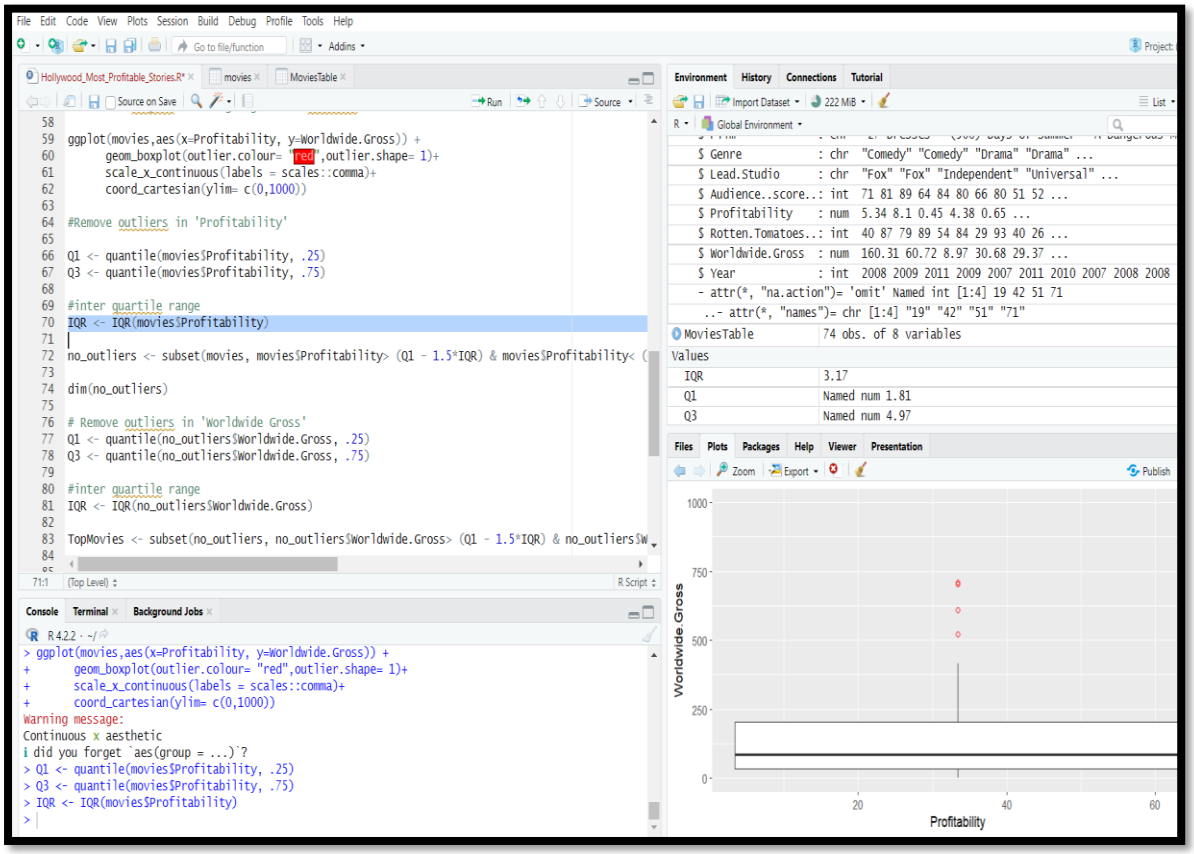
10. CHECK FOR OUTLIERS ON BOX PLOT To check for outliers (a data point that lie away from the other points of the dataset), we first need to make a box plot using the following command. A box plot in R, also known as box and whisker plot, is a graphical representation which allows you to summarize the main characteristics of the data (position, dispersion, skewness etc) and identify the presence of outliers.



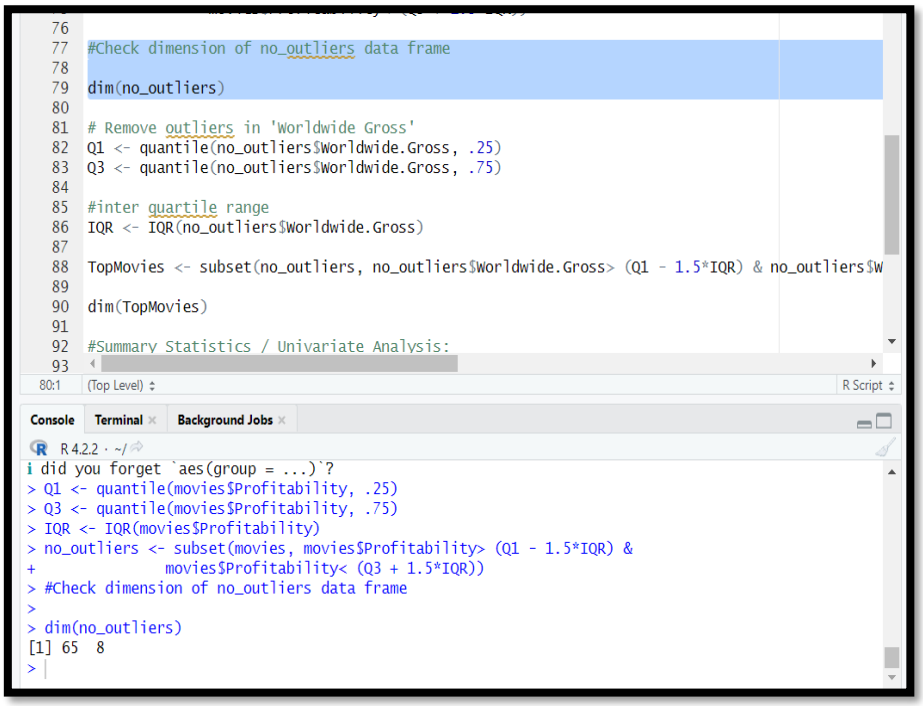
The following images shows the box plot for the movies data frame.



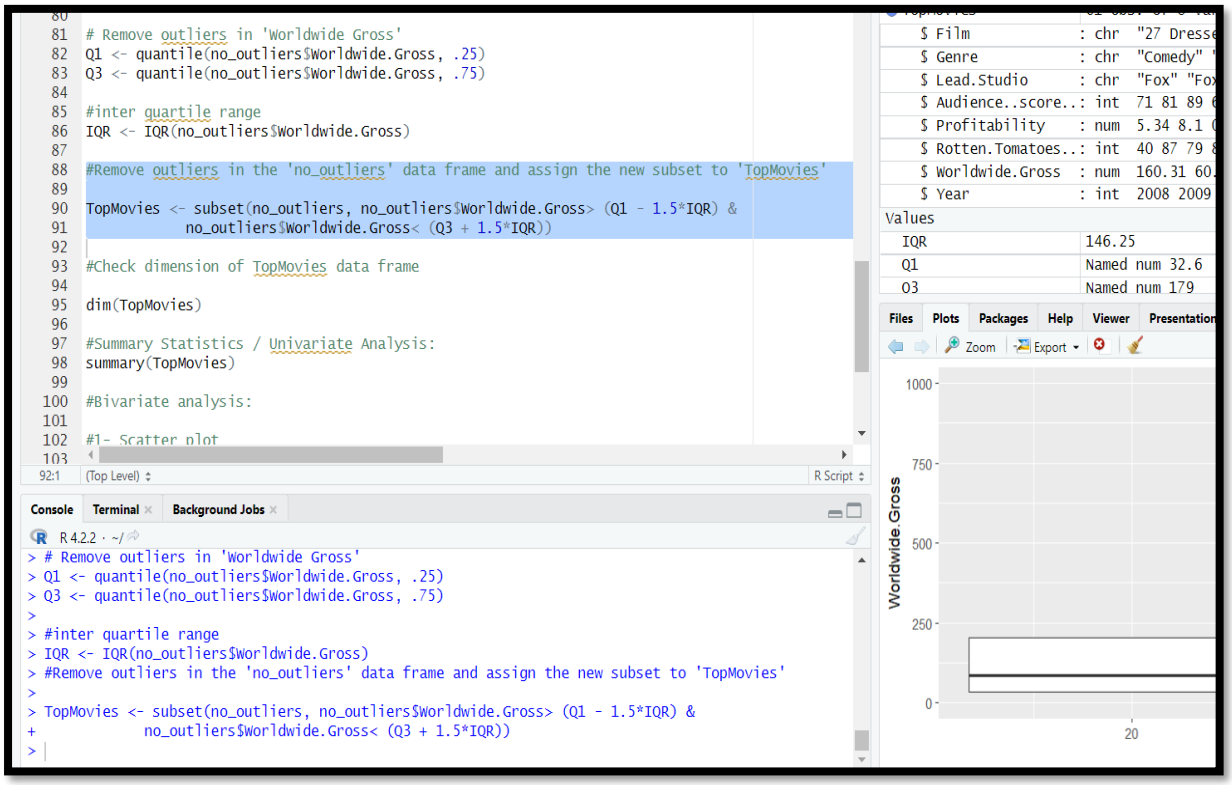
The next step is to remove the outliers in the "Profitability" variable and assign the subset of movies data frame with no outliers to a new data frame called "no_outliers".



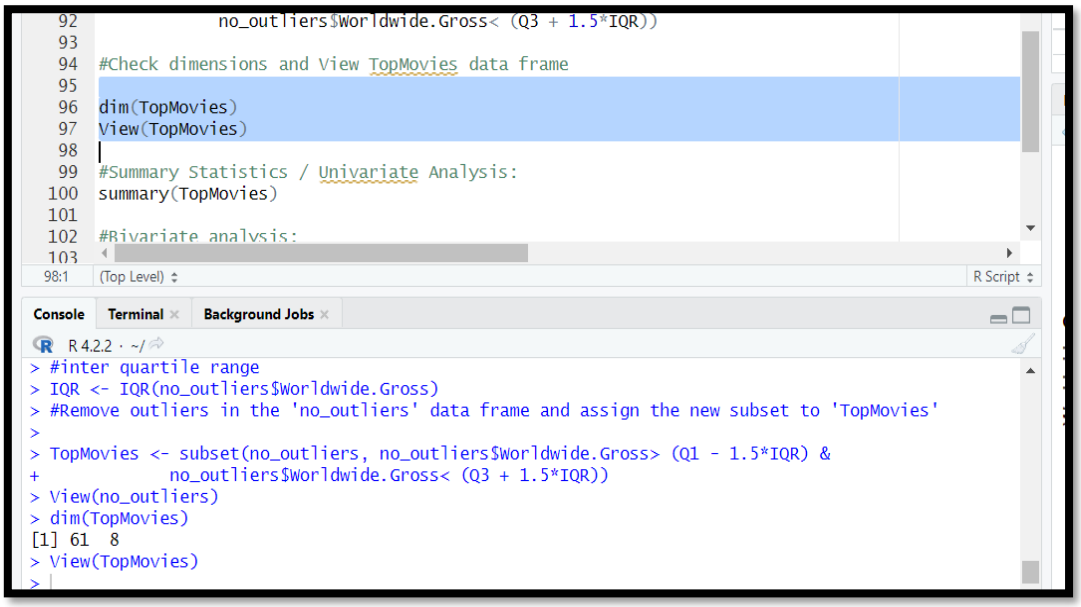
The dimensions of new data frame "no_outliers" after the removal of outliers from the profitability variable are now 65 rows and 8 columns.



Repeat the same procedure to remove the outliers from the Worldwide Gross variable and assign the subset of data frame no_outliers to a new data frame "TopMovies".



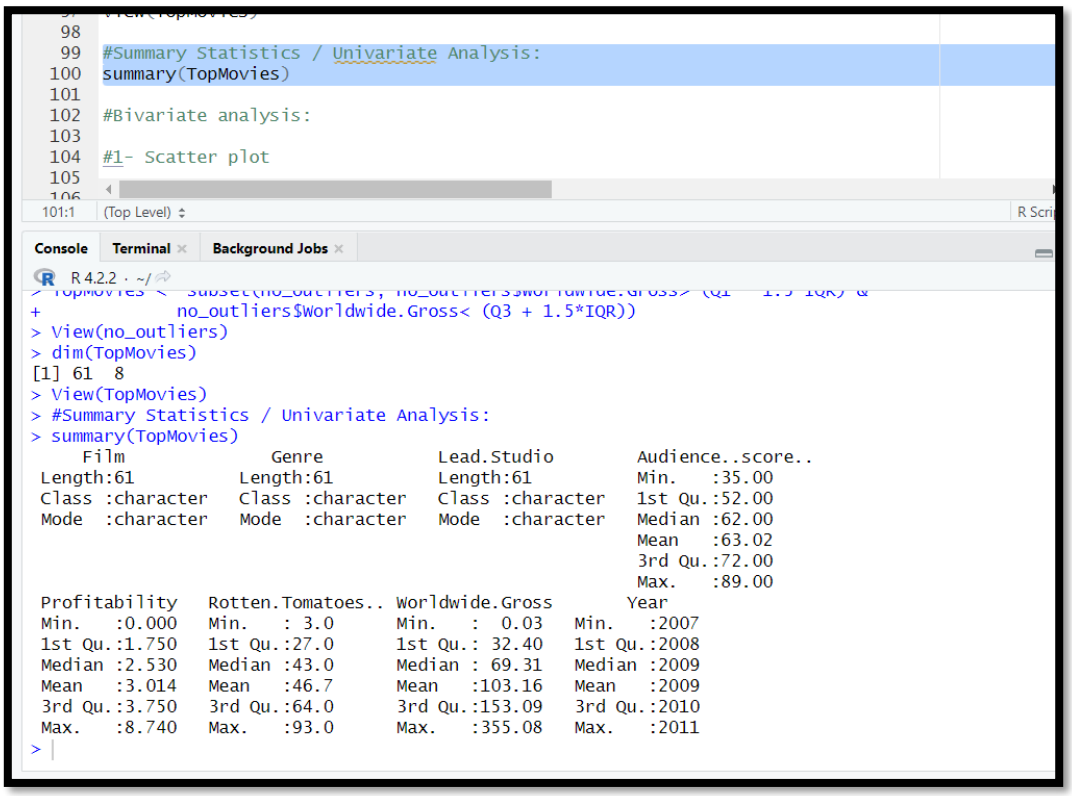
The dimensions of new data frame "TopMovies" after the removal of outliers from the Worldwide Gross variable are now 61 rows and 8 columns.



```
            no_outliers$Worldwide.Gross< (Q3 + 1.5*IQR))

#Check dimensions and View TopMovies data frame

dim(TopMovies)
View(TopMovies)

#Summary Statistics / Univariate Analysis:
summary(TopMovies)

#Bivariate analysis:
```

```
> #inter quartile range
> IQR <- IQR(no_outliers$Worldwide.Gross)
> #Remove outliers in the 'no_outliers' data frame and assign the new subset to 'TopMovies'
>
> TopMovies <- subset(no_outliers, no_outliers$Worldwide.Gross> (Q1 - 1.5*IQR) &
+             no_outliers$Worldwide.Gross< (Q3 + 1.5*IQR))
> View(no_outliers)
> dim(TopMovies)
[1] 61  8
> View(TopMovies)
>
```

11. SUMMARY STATISTICS (UNIVARIATE ANALYSIS) R provides a wide range of functions for obtaining summary statistics or single variable analysis. One method of obtaining descriptive statistics of a data set is to use the summary() function. The summary statistics for each of the 8 Variables in the dataset are shown below.



```
#Summary Statistics / Univariate Analysis:
summary(TopMovies)

#Bivariate analysis:

#1- Scatter plot
```

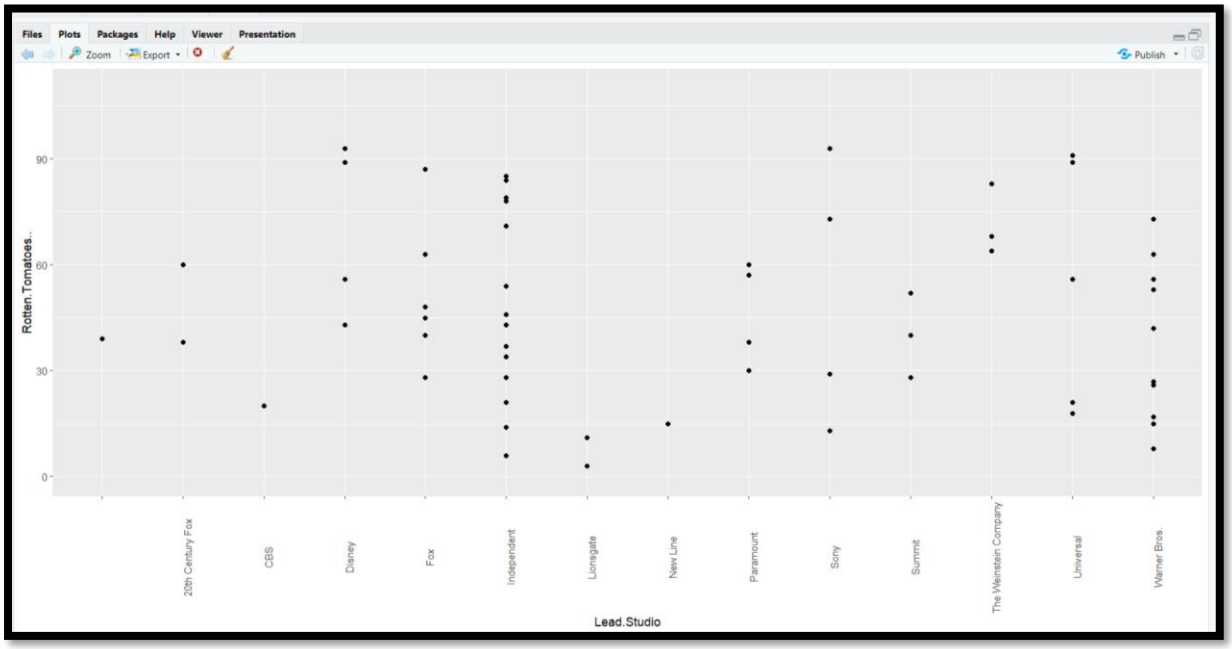
```
+             no_outliers$Worldwide.Gross< (Q3 + 1.5*IQR))
> View(no_outliers)
> dim(TopMovies)
[1] 61  8
> View(TopMovies)
> #Summary Statistics / Univariate Analysis:
> summary(TopMovies)
     Film              Genre           Lead.Studio        Audience..score..
 Length:61          Length:61          Length:61          Min.   :35.00
 Class :character   Class :character   Class :character   1st Qu.:52.00
 Mode  :character   Mode  :character   Mode  :character   Median :62.00
                                                          Mean   :63.02
                                                          3rd Qu.:72.00
                                                          Max.   :89.00
 Profitability   Rotten.Tomatoes..  Worldwide.Gross       Year
 Min.   :0.000   Min.   : 3.0      Min.   :  0.03   Min.   :2007
 1st Qu.:1.750   1st Qu.:27.0      1st Qu.: 32.40   1st Qu.:2008
 Median :2.530   Median :43.0      Median : 69.31   Median :2009
 Mean   :3.014   Mean   :46.7      Mean   :103.16   Mean   :2009
 3rd Qu.:3.750   3rd Qu.:64.0      3rd Qu.:153.09   3rd Qu.:2010
 Max.   :8.740   Max.   :93.0      Max.   :355.08   Max.   :2011
>
```

12. BIVARIATE ANALYSIS The term bivariate refers to the analysis of two variables. It is one of the simplest forms of statistical analysis. It is generally used to find out if there is a relationship between two sets of values or two variables. In present case, we have used scatter plot and bar chart to do the bivariate analysis.

    a. SCATTER PLOT A scatter plot is used to display the relationship between two variables Lead Studio and Rotten Tomatoes Ratings.

```
101
102 #Bivariate analysis:
103
104 #1- Scatter plot
105
106 ggplot(TopMovies, aes(x=Lead.Studio, y=Rotten.Tomatoes..)) +
107        geom_point()+ scale_y_continuous(labels = scales::comma)+
108        coord_cartesian(ylim = c(0, 110))+
109        theme(axis.text.x = element_text(angle = 90))
110
111 #2- Bar chart
112 ggplot(TopMovies, aes(x=Year)) + geom_bar()
113
114 #Export clean data
115
116
110:1  (Top Level)                                                    R Script

Console  Terminal  Background Jobs
R 4.2.2 · ~/

   Genre              Lead.Studio        Audience..Score..
Length:61          Length:61          Length:61          Min.   :35.00
Class :character   Class :character   Class :character   1st Qu.:52.00
Mode  :character   Mode  :character   Mode  :character   Median :62.00
                                                         Mean   :63.02
                                                         3rd Qu.:72.00
                                                         Max.   :89.00
 Profitability   Rotten.Tomatoes.. Worldwide.Gross       Year
 Min.   :0.000   Min.   : 3.0      Min.   :  0.03   Min.   :2007
 1st Qu.:1.750   1st Qu.:27.0      1st Qu.: 32.40   1st Qu.:2008
 Median :2.530   Median :43.0      Median : 69.31   Median :2009
 Mean   :3.014   Mean   :46.7      Mean   :103.16   Mean   :2009
 3rd Qu.:3.750   3rd Qu.:64.0      3rd Qu.:153.09   3rd Qu.:2010
 Max.   :8.740   Max.   :93.0      Max.   :355.08   Max.   :2011
> #Bivariate analysis:
>
> #1- Scatter plot
>
> ggplot(TopMovies, 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))
>
```

Following is the image of the scatter plot between Rotten Tomatoes Ratings and Lead Studios.

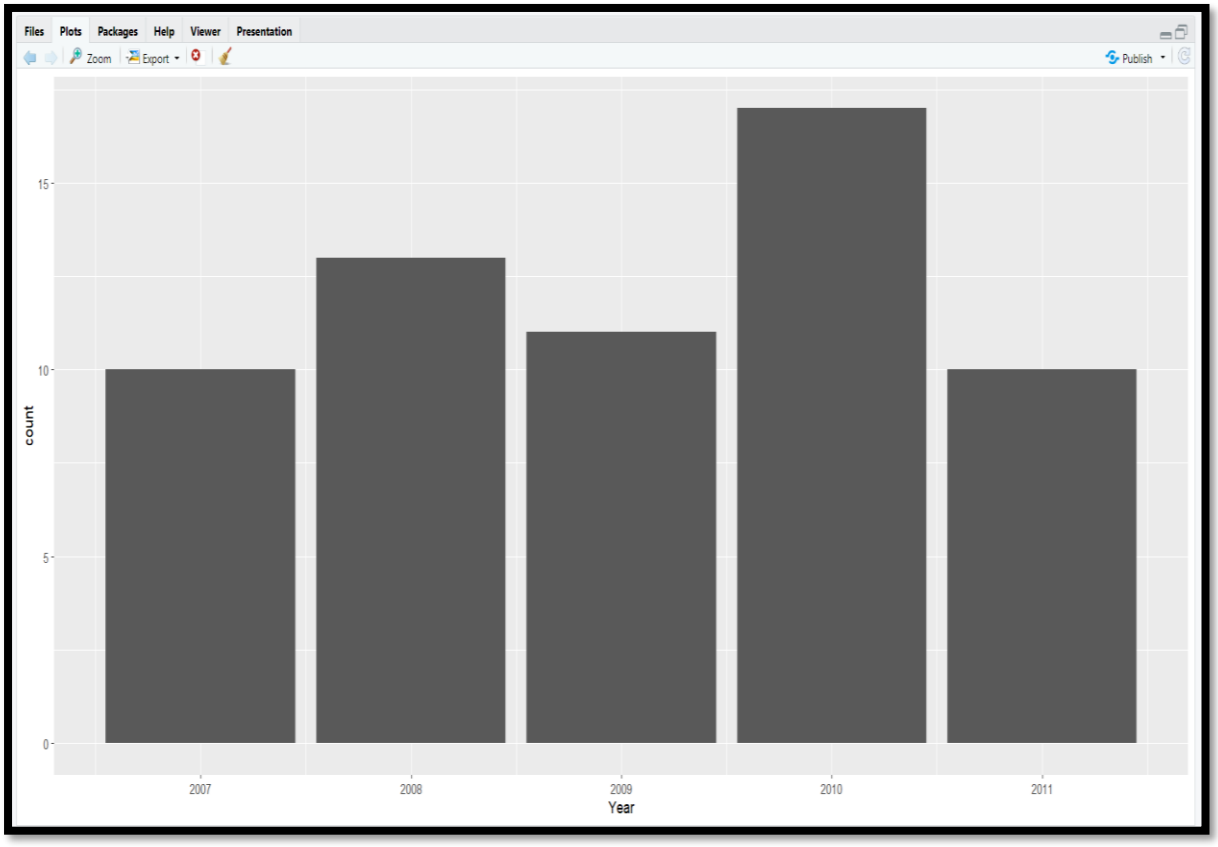


b. BAR CHART A bar chart is used to display the relationship between two variables Year and Count (Number of Films produced each year).

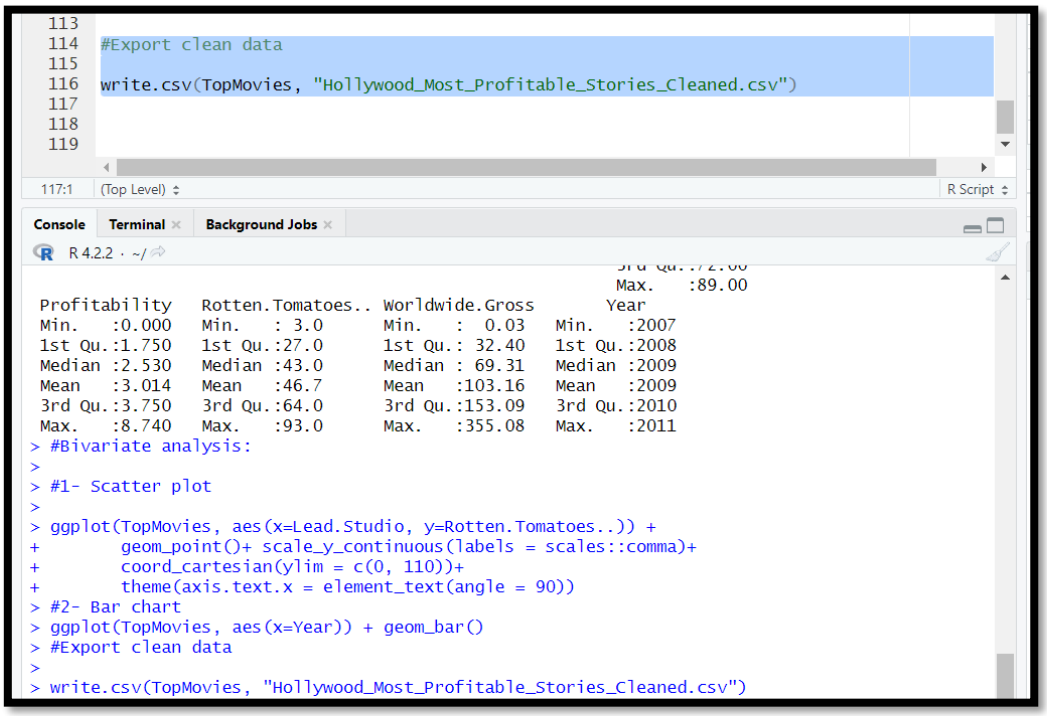
```
ggplot(TopMovies, aes(x=Lead.Studio, y=Rotten.Tomatoes..)) +
       geom_point()+ scale_y_continuous(labels = scales::comma)+
       coord_cartesian(ylim = c(0, 110))+
       theme(axis.text.x = element_text(angle = 90))

#2- Bar chart
ggplot(TopMovies, aes(x=Year)) + geom_bar()

#Export clean data

write.csv(TopMovies, "Hollywood Most Profitable Stories Cleaned.csv")
```

```
Mode  :character   Mode  :character   Mode  :character   Median :62.00
                                                         Mean   :63.02
                                                         3rd Qu.:72.00
                                                         Max.   :89.00

 Profitability   Rotten.Tomatoes.. Worldwide.Gross       Year
 Min.   :0.000   Min.   : 3.0      Min.   :  0.03   Min.   :2007
 1st Qu.:1.750   1st Qu.:27.0      1st Qu.: 32.40   1st Qu.:2008
 Median :2.530   Median :43.0      Median : 69.31   Median :2009
 Mean   :3.014   Mean   :46.7      Mean   :103.16   Mean   :2009
 3rd Qu.:3.750   3rd Qu.:64.0      3rd Qu.:153.09   3rd Qu.:2010
 Max.   :8.740   Max.   :93.0      Max.   :355.08   Max.   :2011
> #Bivariate analysis:
>
> #1- Scatter plot
>
> ggplot(TopMovies, 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))
> #2- Bar chart
> ggplot(TopMovies, aes(x=Year)) + geom_bar()
>
```

Following is the image of the scatter plot between Rotten Tomatoes Ratings and Lead Studios.



13. EXPORT CLEANED DATA Finally, the cleaned data is exported to csv file named "Hollywood_Most_Profitable_Stories_Cleaned.csv" and then it can be used in any platform like Tableau or Power BI for data visualisation and story-telling.

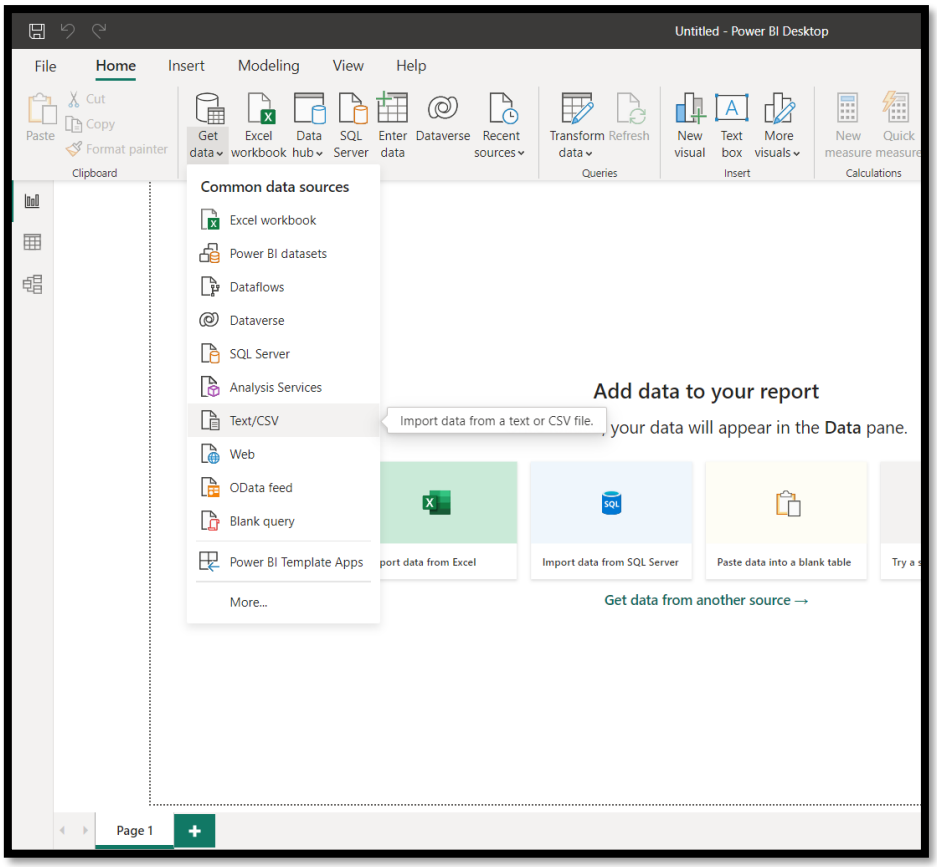


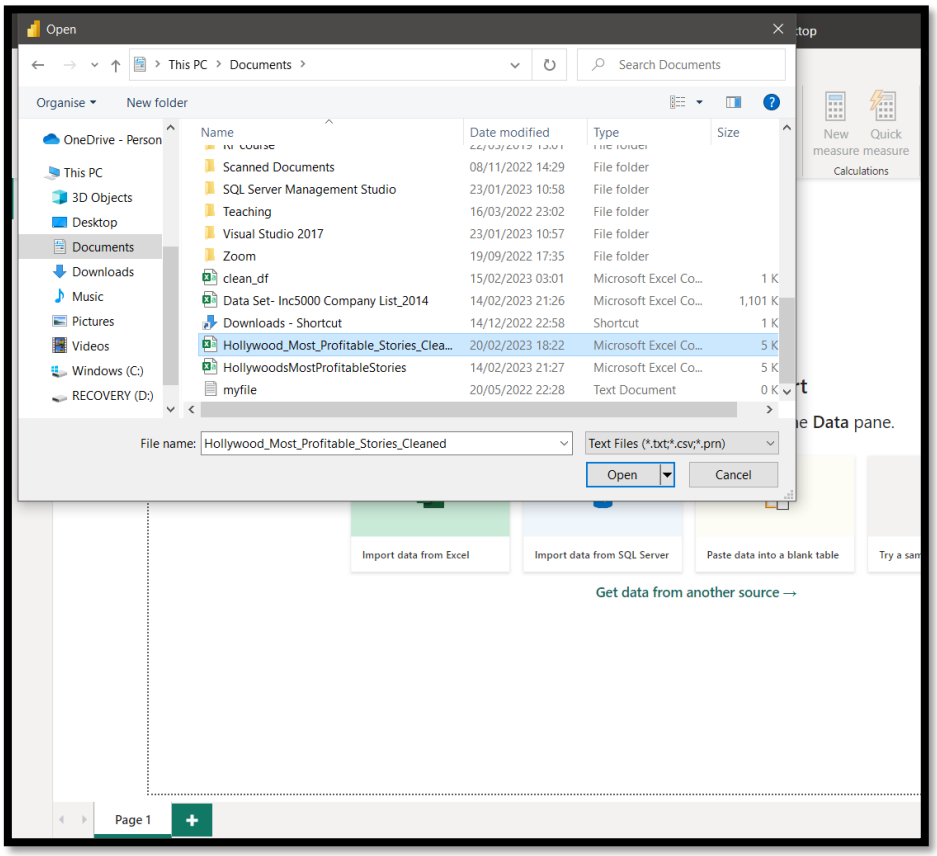
TASK – 2 : DASHBOARD IN POWERBI

In this task, "Hollywood_Most_Profitable_Stories_Cleaned.csv" dataset is visualised in a dashboard using PowerBI. Following are the charts required by the client, but additional charts can be used for effective story telling:

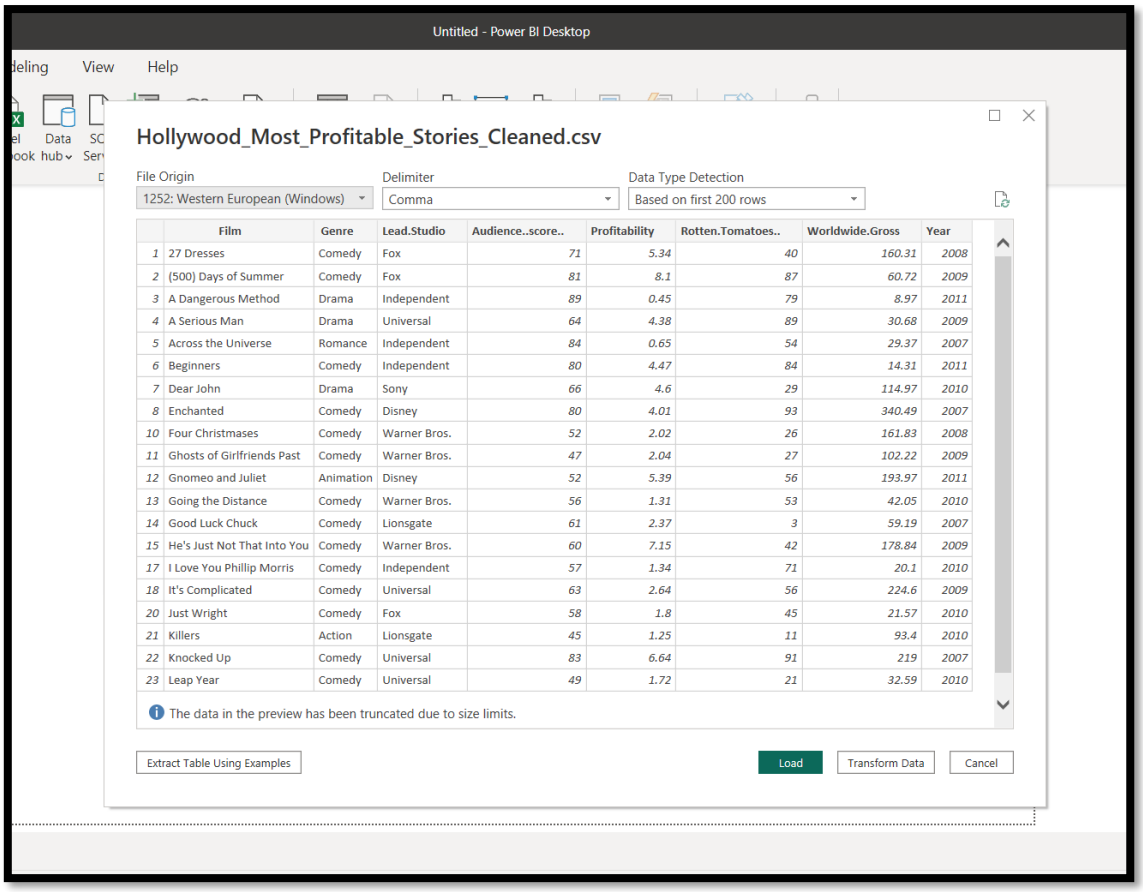
- The average Rotten Tomatoes Ratings of each Genre
- The number of Movies produced per Year
- The Audience Scores for each Movie
- The Profitability per Studio
- The Worldwide Gross per Genre

1. IMPORT DATA Open PowerBI desktop and click on "Get Data" in "Text/CSV" header under the "Common Data Sources" section. Select the required file and click Open. This will import the "Hollywood_Most_Profitable_Stories_Cleaned.csv" dataset in PowerBI.

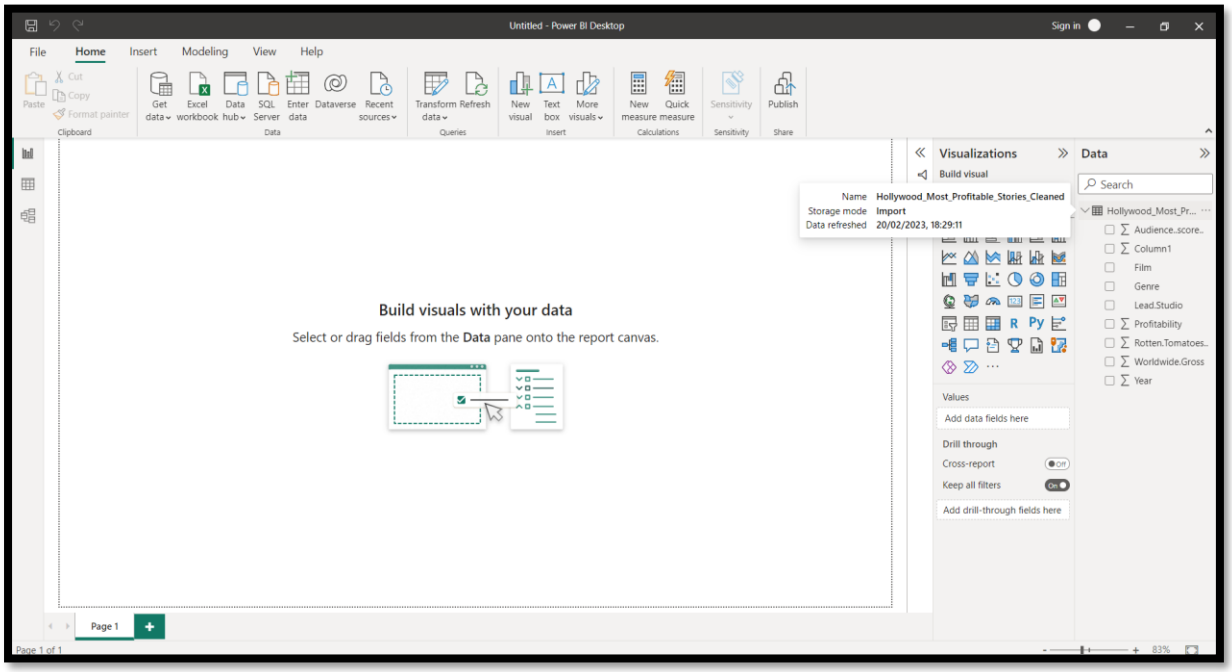




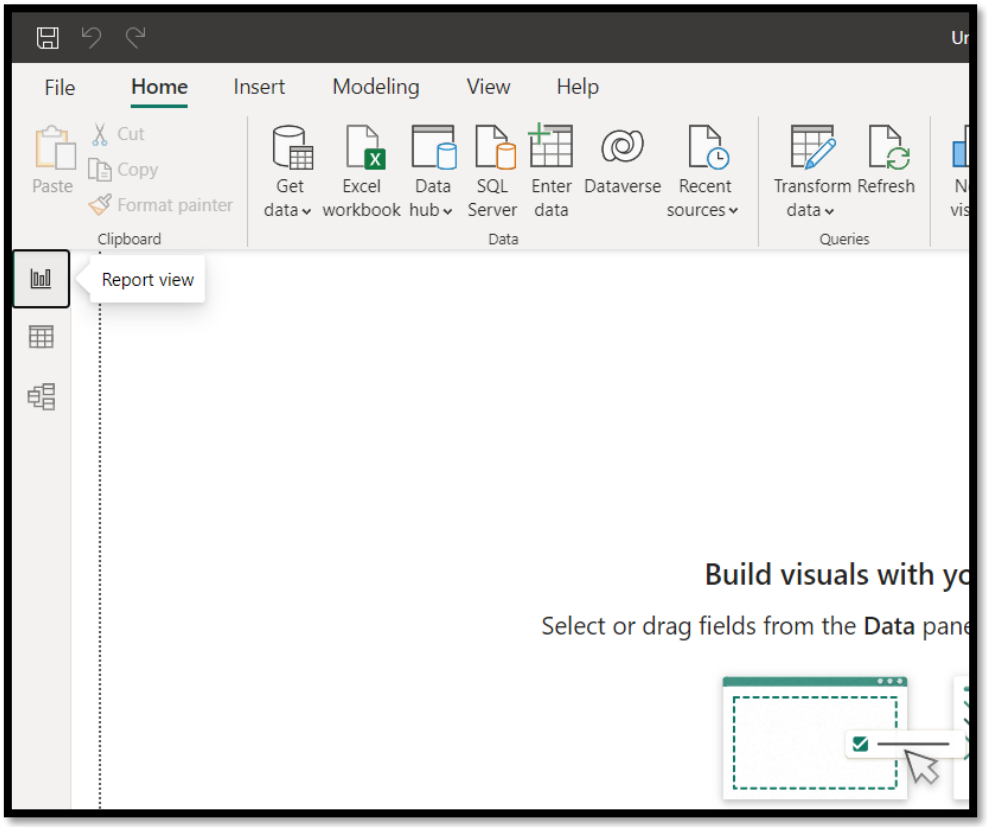
After opening the file, click Load and it will open the csv file in a table format as shown below.

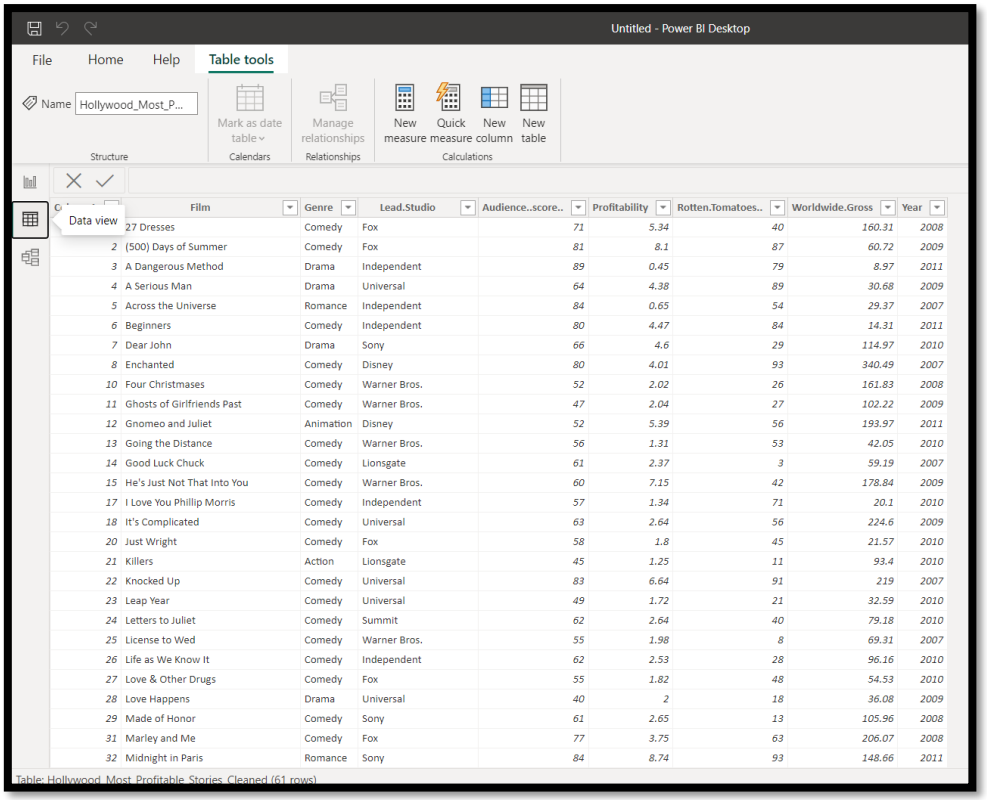


The dataset is imported and ready for building the dashboard. You can check the header names on the right hand side of the PowerBI screen.

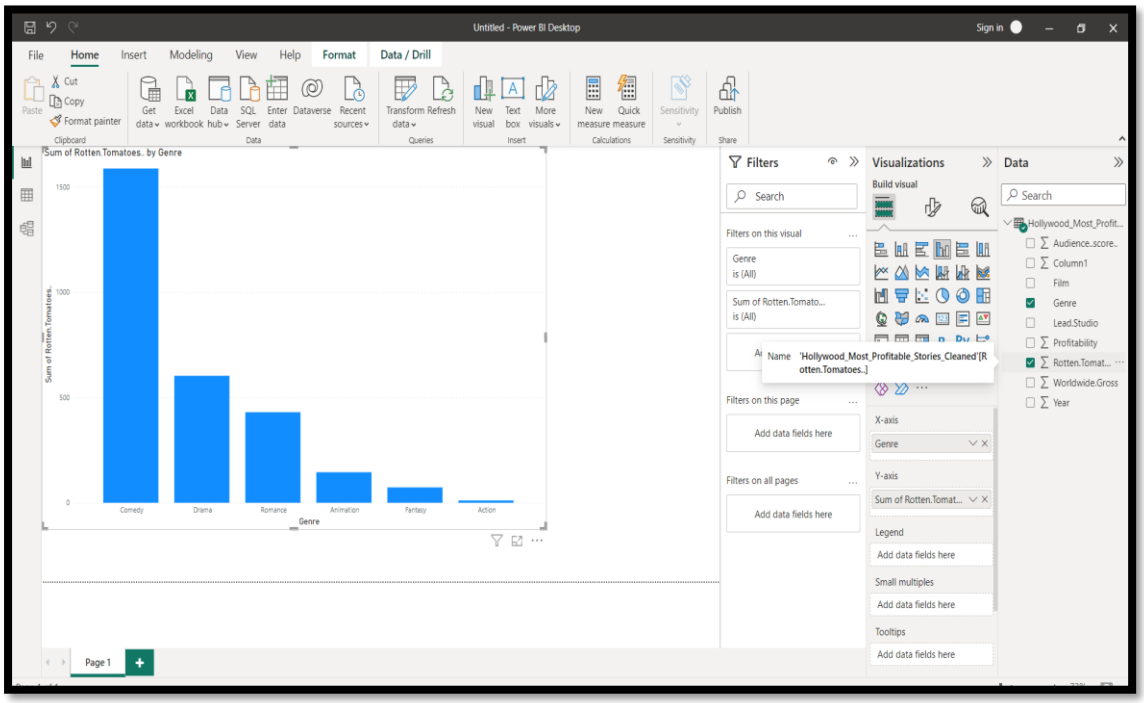


You can also switch between the report view and data view on the left hand side of the screen as shown below.

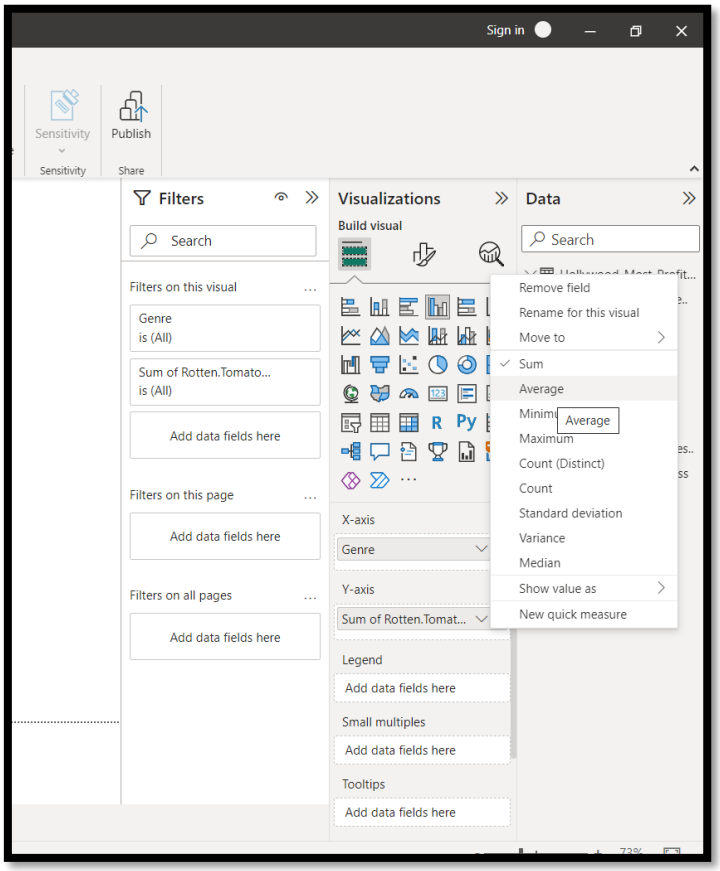




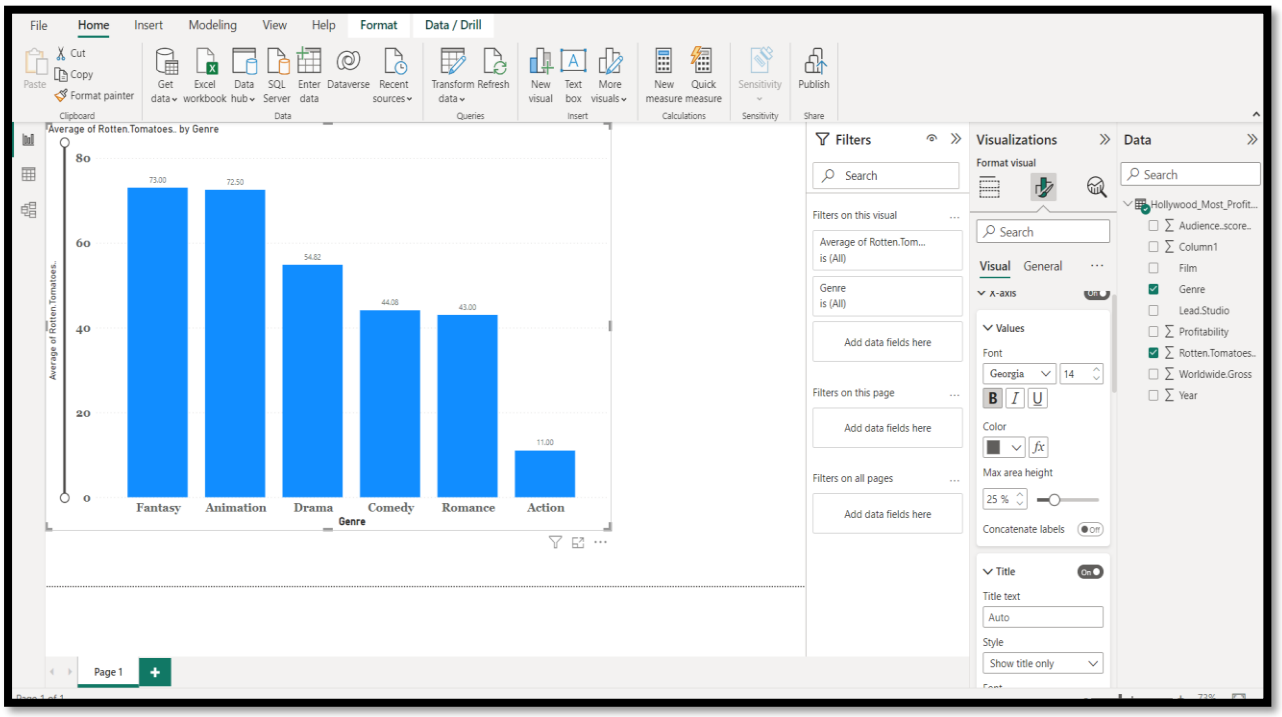
2. BUILD VISUALS To start building , visualisations, you can select or drag and drop the required variables to data field placeholder and then select the visual type.

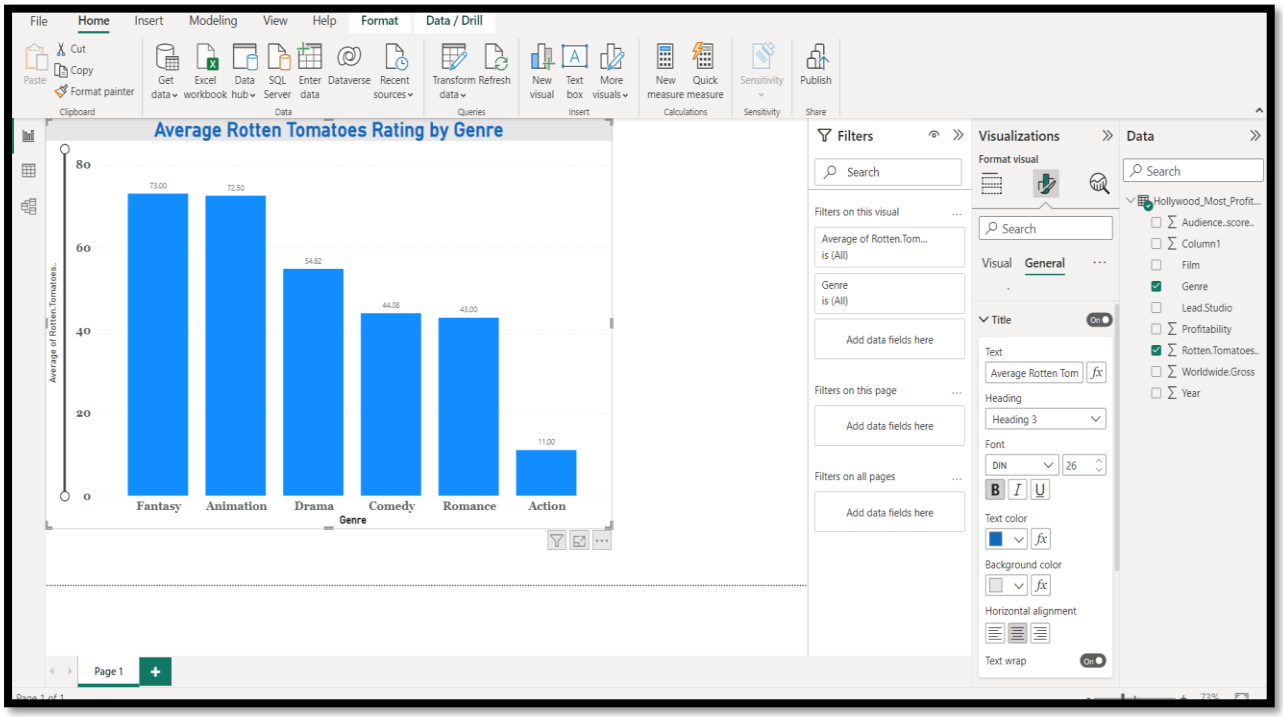


After building the chart, you can change the value to be shown as sum, average etc. For the first chart, which requires Average Rotten Tomatoes Ratings by Genre. We will display the variable as an average.

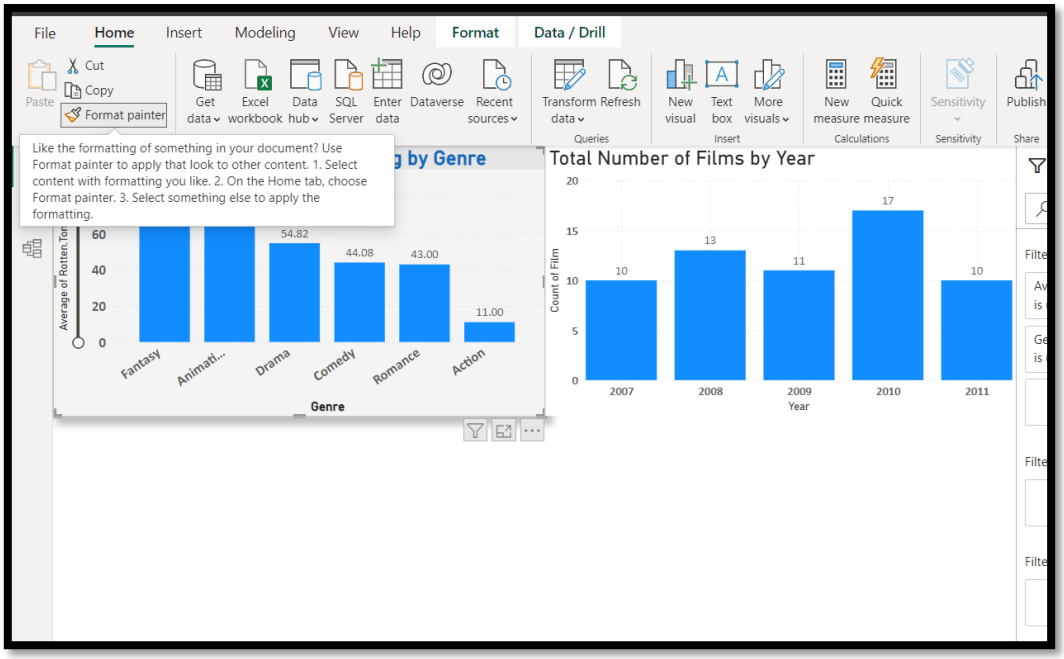


You can also add title, data labels, filters, fonts and colour attributes to your chart to make it more visually appealing and more understandable for the stakeholders.

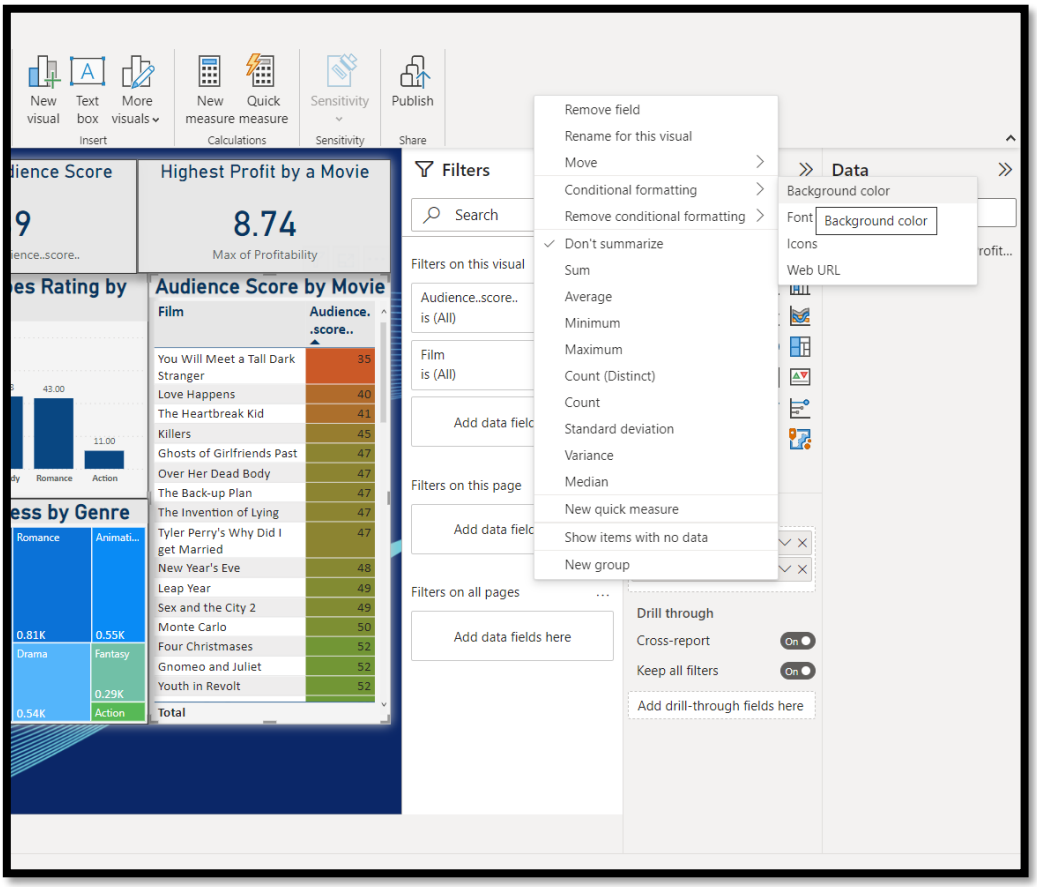




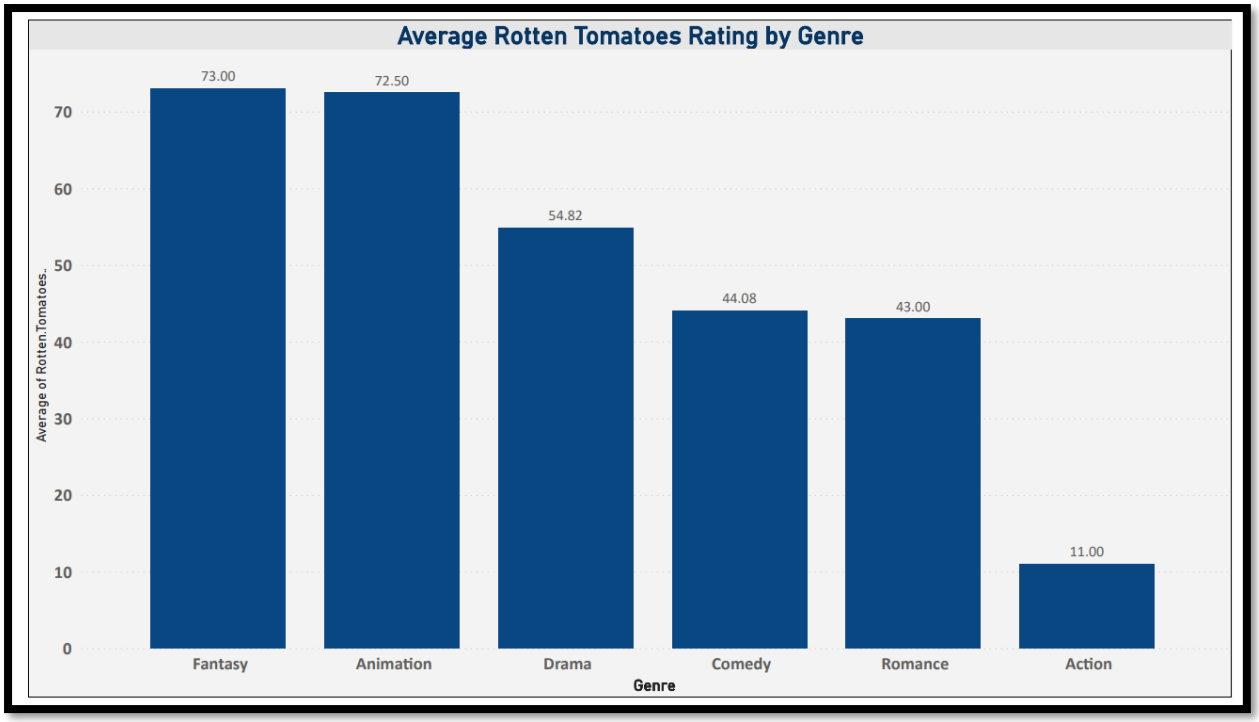
Once you have created one visualisation, you can simply select it and click format painter and paste it the rest of the visualisations to keep formatting consistent throughout your dashboard. It also makes the building process seamless and efficient.



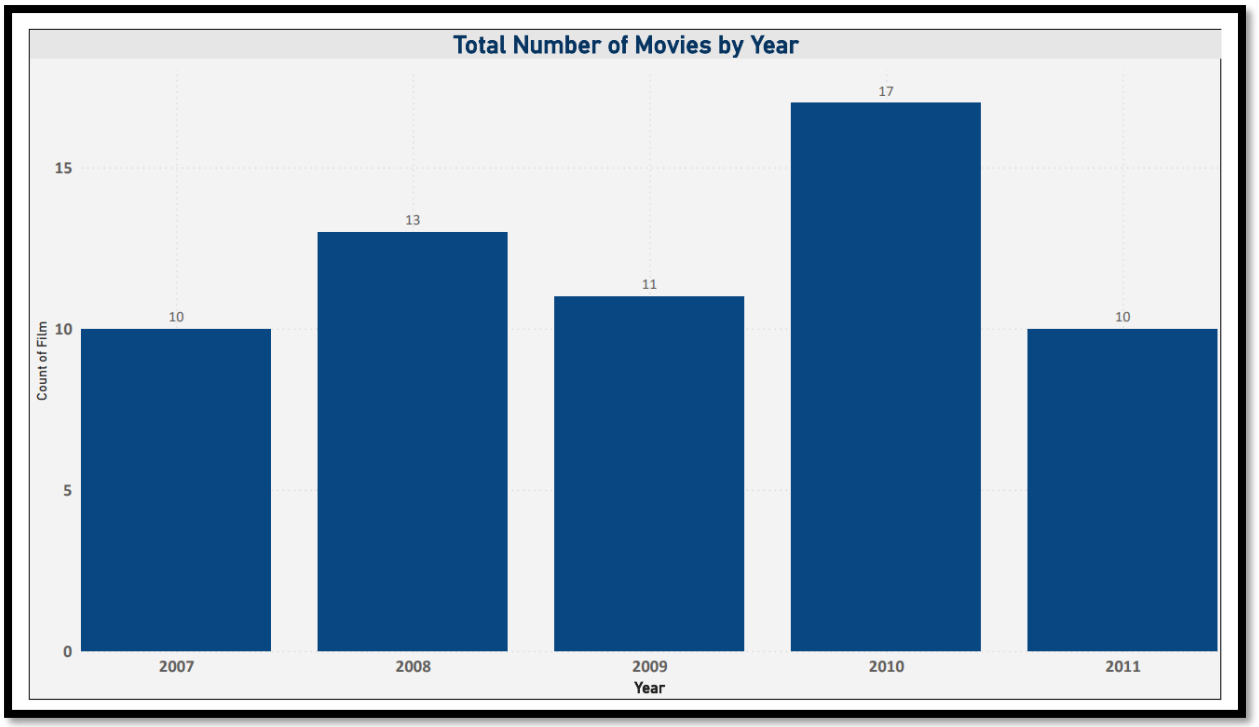
You can add features like conditional formatting to the table sorting the values from low to high or vice versa and assigning a colour palette to it.



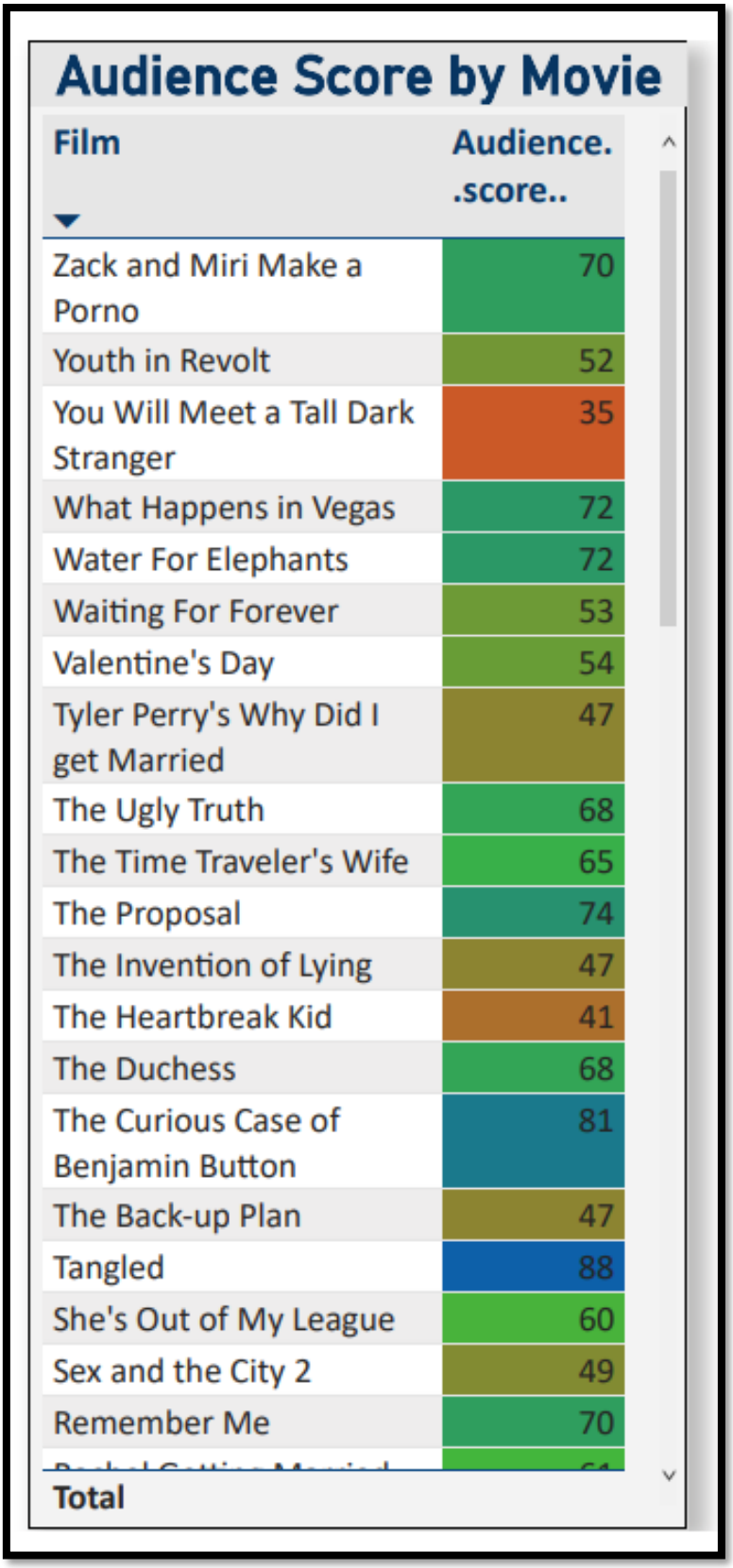
To bar chart displaying the Average Rotten Tomatoes Ratings by Genre is shown below. The Genre Fantasy has the highest rating with Animation as a close second while Action Genre has the lowest Rotte Tomatoes Ratings by user.



To bar chart displaying the Total Number of Movies produced each year is shown below. The highest number of movies (17) were produced in year 2010 while 2007 and 2010 has the same movie count of 10.



The table displaying Audience Score for each film is hown below, displaying the values in ascending order and with the conditional formatting in colour palette (green, blue and brown) as required by the client .

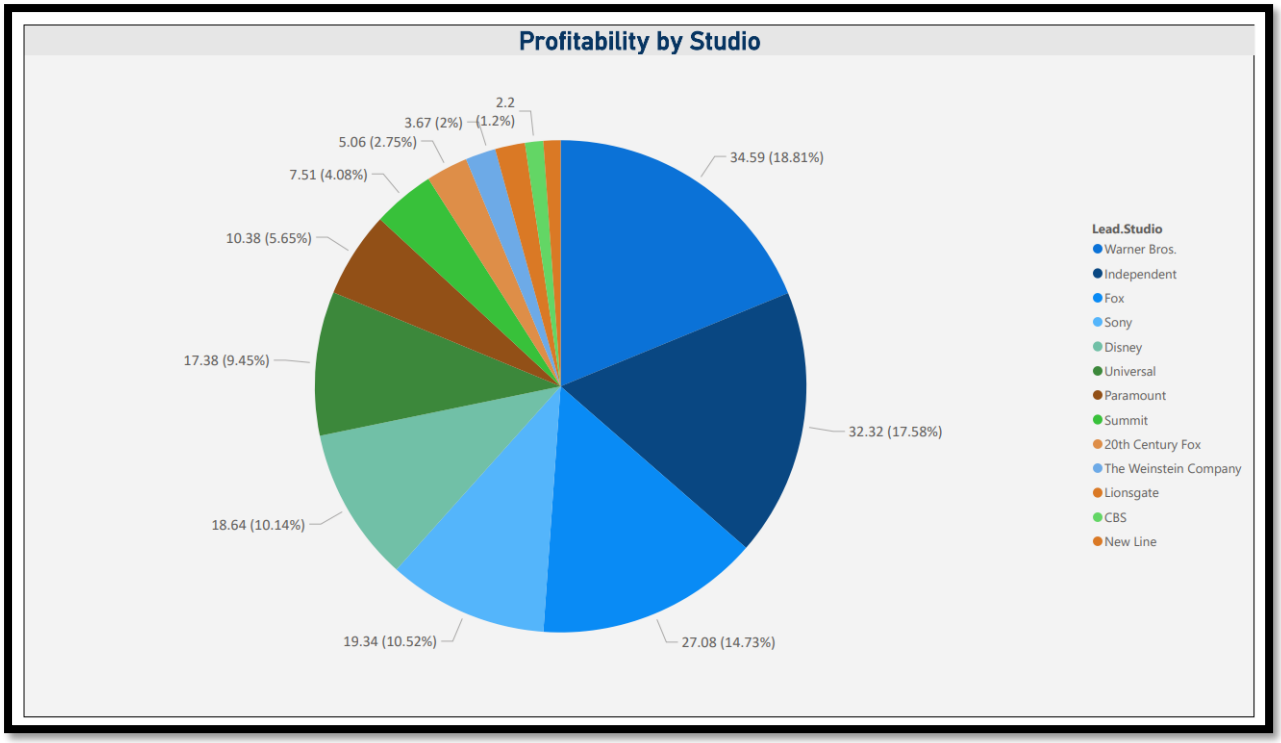


Audience Score by Movie

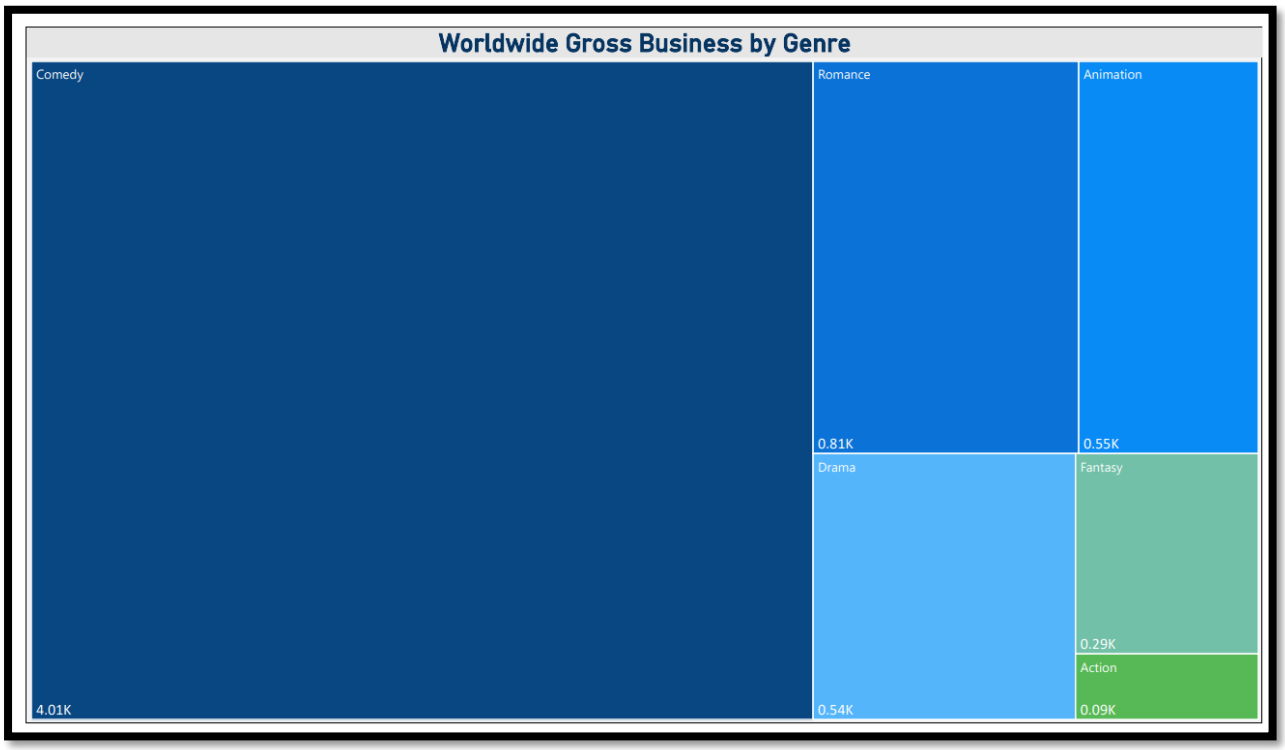
Film	Audience. .score..
Zack and Miri Make a Porno	70
Youth in Revolt	52
You Will Meet a Tall Dark Stranger	35
What Happens in Vegas	72
Water For Elephants	72
Waiting For Forever	53
Valentine's Day	54
Tyler Perry's Why Did I get Married	47
The Ugly Truth	68
The Time Traveler's Wife	65
The Proposal	74
The Invention of Lying	47
The Heartbreak Kid	41
The Duchess	68
The Curious Case of Benjamin Button	81
The Back-up Plan	47
Tangled	88
She's Out of My League	60
Sex and the City 2	49
Remember Me	70

Total

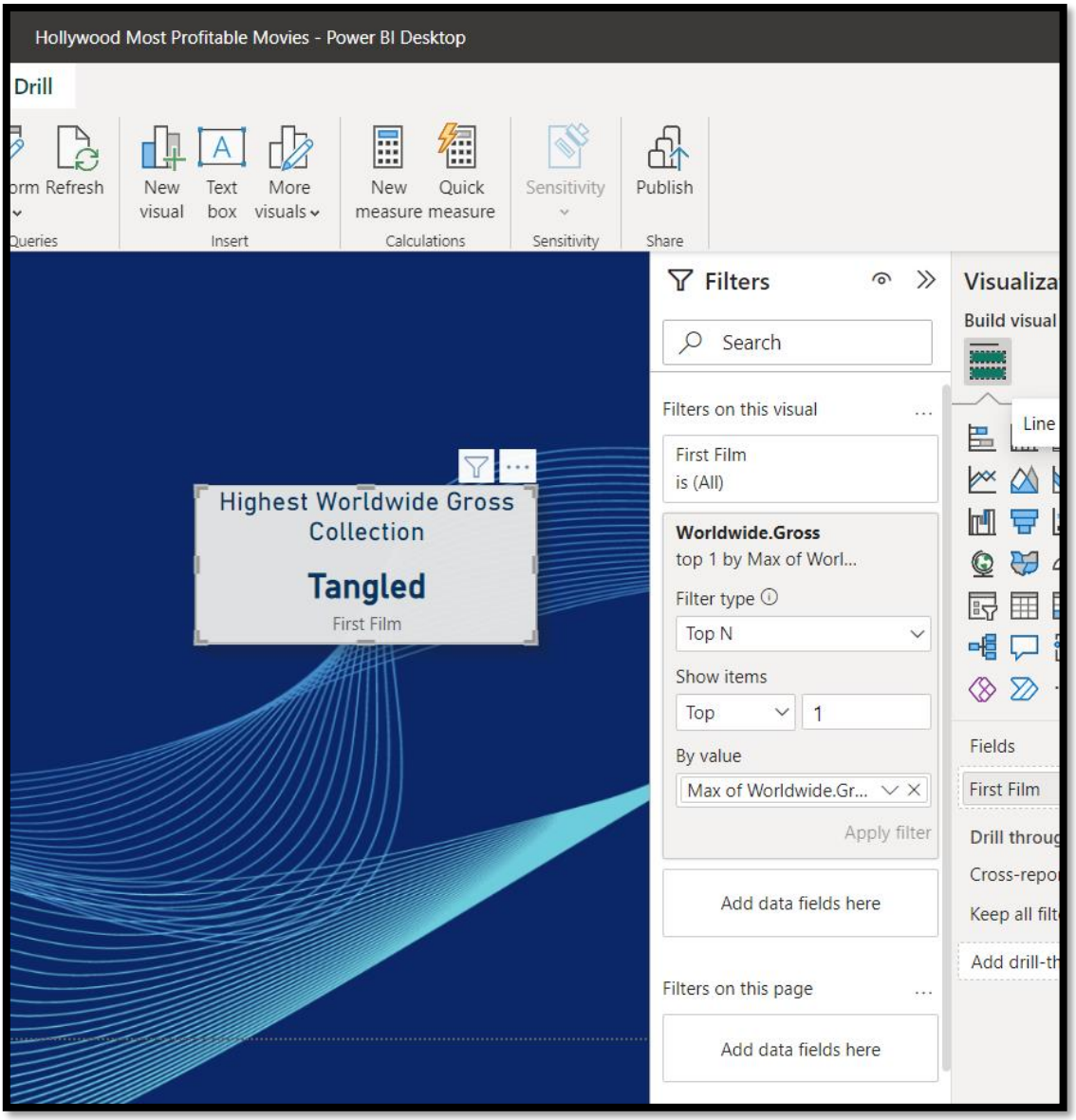
The pie chart displaying the Profitability by Studio is shown below. Warner Bros. is the studio with highest Profitability while New line studio has the lowest Profitability.



The tree map displaying the Worldwide Gross Business by Genre is shown below. The Genre Comedy has the highest worldwide gross business while action is the lowest.



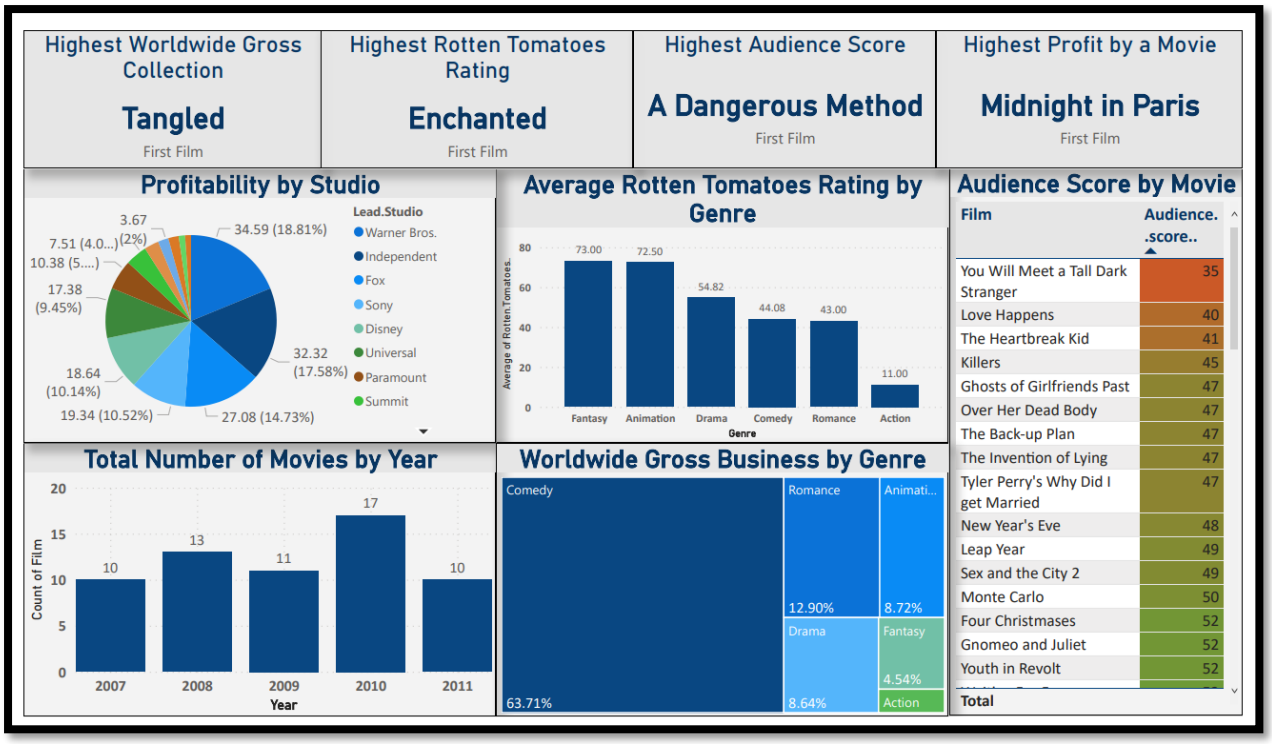
To display cards with highest values, you can use the filter and select Top N values as shown below.



3. BUILD DASHBOARD

Once you have completed building all your individual charts, you can re-arrange them in an appropriate and visually appealing order to build your dashboard. For this project, client asked for a colour-palette consisting of blue, green and brown colours. The dashboard shown below has been created following the guidelines and principles outlined by the client.

The final dashboard for the dataset is shown below. It gives insight to the Hollywood most Profitable Movies between the year 2007-2012.



REFLECTIVE

This project has helped me to utilise the knowledge of programming language R for data cleaning and analysis and Microsoft Power BI to visualise the data for effective storytelling. I found it highly interesting to learn to write the script in R as I enjoy coding. I particularly enjoyed using PowerBI to visualise the data and see the story behind the data set come to life. It provided me insights into dataset which didn't make much sense to begin with.