

STEP 1: INITIAL EXPLORATORY ANALYSIS

1. LOADING DATA FROM THE WEB USING SYNTAX `df<- read.csv()` AND VIEWING THE DF FILE IN ENVIRONMENT.

R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

```
> 'hello world'
[1] "hello world"
> #LOAD DATA
> df<-read.csv("https://public.tableau.com/app/sample-data/HollywoodsMostProfitableStories.csv")
Error: unexpected input in "df<-read.csv("")"
> df<- read.csv('https://public.tableau.com/app/sample-data/HollywoodsMostProfitableStories.csv')
> |
```

Environment History Connections Tutorial

Data

- df 74 obs. of 8 variables
 - \$ Film : chr "27 Dresses" "(500) Days of Summer" "A Dangerous Method" ...
 - \$ Genre : chr "Comedy" "Comedy" "Drama" "Drama" ...
 - \$ Lead.Studio : chr "Fox" "Fox" "Independent" "Universal" ...
 - \$ Audience..score.: int 73 81 89 64 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 ...

Files Plots Packages Help Viewer Presentation

New Folder New Blank File Delete Rename More

Home Applications Desktop Documents Downloads Library Movies Music OneDrive Pictures Public

R introduction.RData 4.5 KB Mar 17, 2023, 4:22 PM

staff.phone.numbers.txt 708 B Mar 23, 2020, 10:00 PM

N/B: INCLUDING A // HELPED TO ELIMINATE THE ERROS THAT CAME UP.

2. VIEWING, IMPORTING AND LOADING LIBRARY

	Film	Genre	Lead.Studio	Audience..score..	Profitability
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
14	Good Luck Chuck	Comedy	Lionsgate	61	2.3676851

Showing 1 to 14 of 74 entries, 8 total columns

Console Terminal Background Jobs

R 4.2.2 - ~/

```
kages'
>
> #IMPORT LIBRARY
> View(df)
> View(df)
> install.packages('tidyverse')
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.2/tidyverse_2.0.0.tgz'
Content type 'application/x-gzip' length 422995 bytes (413 KB)
=====
downloaded 413 KB
```

N/B USED SYNTAX `View(df)` TO SEE THE TABLE AT THE TOP OF THE CONSOLE IN A TABLE

STEP 2: CLEAN DATA

3. CHECKING FOR MISSING VALUES USING colSums(is.na(df))

The screenshot shows the RStudio interface with the following components:

- Data View:** A grid showing movie data with columns: Film, Genre, Lead.Studio, Audience..score., Profitability, and R.
- Console View:** Displays R code and its output. The code includes `#CHECK FOR MISSING VALUES` and `colSums(is.na(df))` which shows the count of missing values for each column.

Film	Genre	Lead.Studio	Audience..score..	Profitability	R
61 The Ugly Truth	Comedy	Independent		68	5.4026316
62 Twilight	Romance	Summit		82	10.1800270
63 Twilight: Breaking Dawn	Romance	Independent		68	6.3833636
64 Tyler Perry's Why Did I Get Married	Romance	Independent		47	3.7241924
65 Valentine's Day	Comedy	Warner Bros.		54	4.1840385
66 Waiting For Forever	Romance	Independent		53	0.0050000
67 Waitress	Romance	Independent		67	11.0897415
68 WALL-E	Animation	Disney		89	2.8960191
69 Water For Elephants	Drama	20th Century Fox		72	3.0814211
70 What Happens in Vegas	Comedy	Fox		72	6.2676470
71 When in Rome	Comedy	Disney		44	NA
72 You Will Meet a Tall Dark Stranger	Comedy	Independent		35	1.2118182
73 Youth in Revolt	Comedy	The Weinstein Company		52	1.0900000
74 Zack and Miri Make a Porno	Romance	The Weinstein Company		70	1.7475417

Showing 61 to 74 of 74 entries, 8 total columns

```

Console Terminal Background Jobs
R 4.2.2 . ~/ 
> #CHECK FOR MISSING VALUES
> colSums(is.na(df))
  Film          Genre
  0            0
  Lead.Studio Audience..score..
  0            1
  Profitability Rotten.Tomatoes..
  3            1
  Worldwide.Gross      Year
  0            0
> View(colSums)
> colSums(is.na(df))
  Film          Genre
  0            0
  Lead.Studio Audience..score..

```

4. DROPPED MISSING VALUES USING df<- na.omit(df)

The screenshot shows the RStudio interface with the following components:

- Data View:** A grid showing movie data with columns: Film, Genre, Lead.Studio, Audience..score., Profitability, and R.
- Environment View:** Shows the global environment with a data frame `df` containing 70 observations and 8 variables.
- Console View:** Displays R code and its output. The code includes `View(df)` and `df<- na.omit(df)` followed by `colSums(is.na(df))` which shows the count of missing values for each column.
- File Explorer:** Shows the local file system with various folders and files.

Film	Genre	Lead.Studio	Audience..score..	Profitability	R
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	
14 Good Luck Chuck	Comedy	Lionsgate	61	2.3676851	

Showing 1 to 14 of 70 entries, 8 total columns

```

Console Terminal Background Jobs
R 4.2.2 . ~/ 
> View(df)
> 
> #DROP MISSING VALUES
> df<- na.omit(df)
> colSums(is.na(df))
  Film          Genre
  0            0
  Lead.Studio Audience..score..
  0            0
  Profitability Rotten.Tomatoes..
  0            0
  Worldwide.Gross      Year
  0            0
> df<- na.omit(df)
> 

```

N/B I NOTICED 4 OUT OF 74 OBS WERE OMITTED FROM THE DATA LEAVING 70 OBS

5. CONFIRMED ROWS HAVE BEEN REMOVED `colSums(is.na(df))`

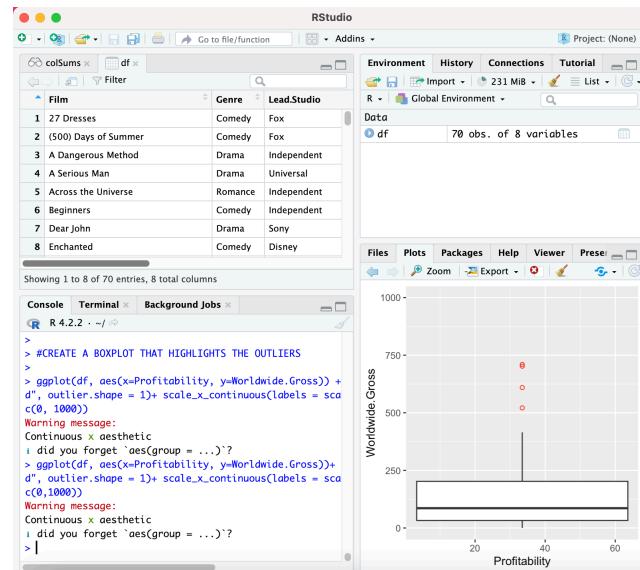
```
R 4.2.2 - ~/Desktop
> #CHECK TO MAKE SURE THE ROWS HAVE BEEN REMOVED
> colSums(is.na(df))
  Film      Genre      Lead.Studio Audience..score.. Profitability Rotten.Tomatoes.. Worldwide.Gross      Year
0       0        0           0            0             0                 0          0            0        0
  Lead.Studio Audience..score.. Profitability Rotten.Tomatoes.. Worldwide.Gross      Year
0       0        0           0             0            0            0            0        0
  Profitability Rotten.Tomatoes.. Worldwide.Gross      Year
0       0            0            0        0
  Worldwide.Gross      Year
0       0        0
```

6. CHECKING FOR DUPLICATES AND ROUNDING OFF VALUES TO 2 PLACES

```
R 4.2.2 - ~/Desktop
> #CHECK FOR DUPLICATES
> dim(df[duplicated(df$film),])[1]
[1] 0
>
> #CHECK FOR ROUND OFF VALUE TO 2 PLACES
> df$Profitability <- round(df$Profitability ,digit=2)
> df$Worldwide.Gross <- round(df$Worldwide.Gross ,digit=2)
>
> view(df)
Error in view(df) : could not find function "view"
> View(df)
> dim(df)
[1] 70 8
>
```

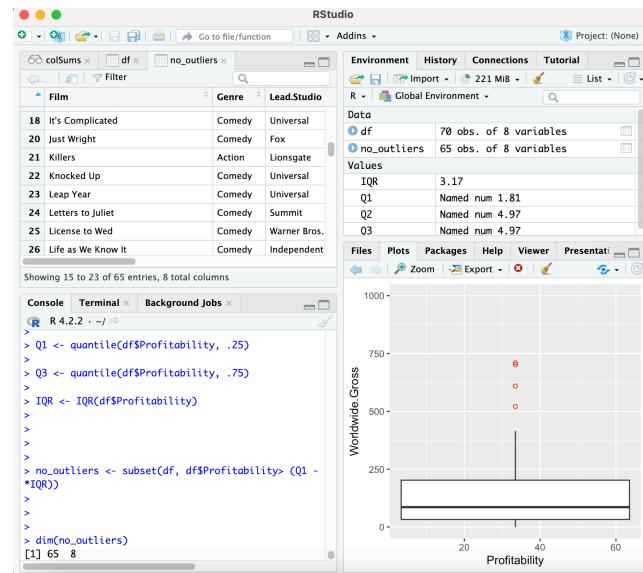
STEP 2.1 OUTLIER REMOVAL

7. CHECKING FOR OUTLIERS USING A BOXPLOT USING library(ggplot2)



Boxplots display asterisks or other symbols on the graph to indicate explicitly when datasets contain outliers. These graphs use the interquartile method with fences to find outliers, which I explain later. The boxplot below displays our example dataset. It's clear that the outlier is quite different than the typical data value.

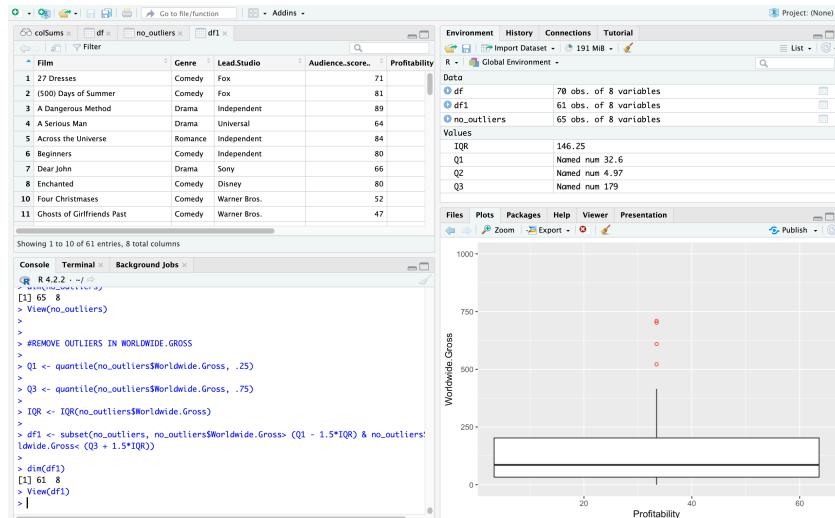
8. REMOVING OUTLIERS IN “PROFITABILITY”



The IQR is the middle 50% of the dataset. It's the range of values between the third quartile and the first quartile ($Q_3 - Q_1$). We can take the IQR, Q_1 , and Q_3 values to calculate the following outlier fences for our dataset: lower outer, lower inner, upper inner, and upper outer. These fences determine whether data points are outliers and whether they are mild or extreme.

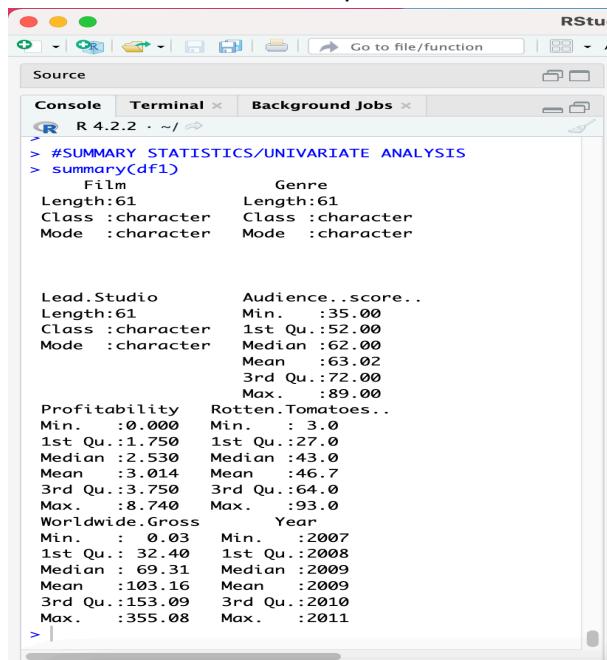
N/B: Values that fall inside the two inner fences are not outliers. Let's see how this method works using our example dataset.

9. REMOVING OUTLIERS IN “WORLDWIDE.GROSS”

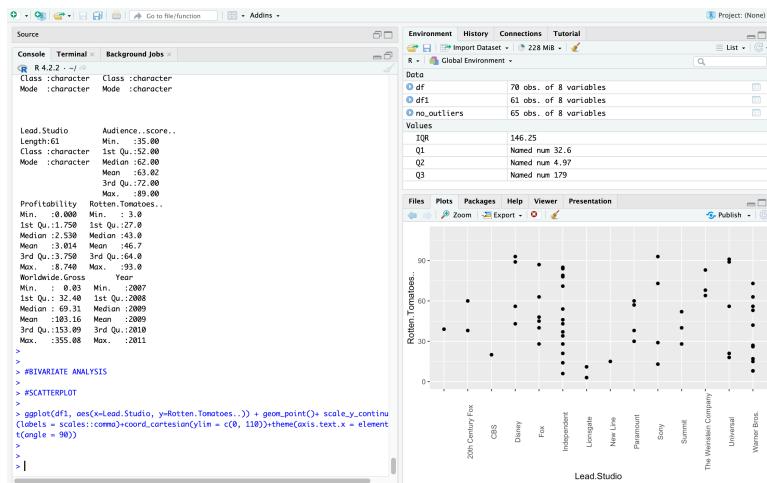


STEP 3: EXPLORATORY DATA ANALYSIS

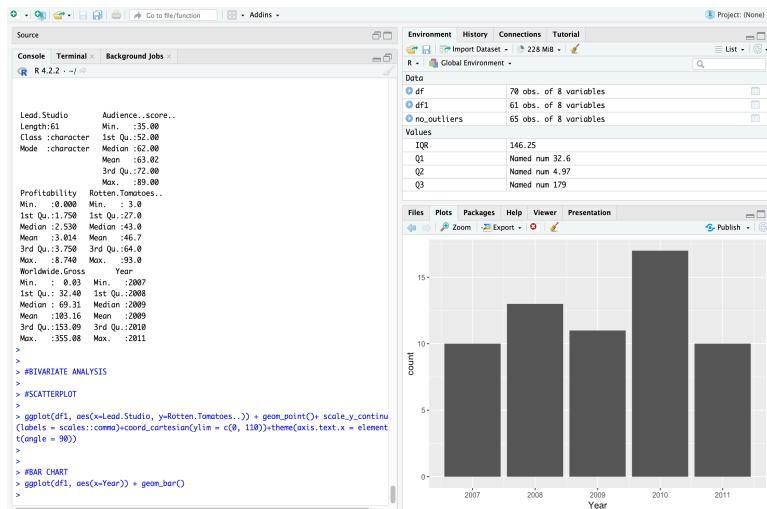
10. SUMMARY STATISTICS/UNIVARIATE ANALYSIS



11. BIVARIATE ANALYSIS USING SCATTER PLOT

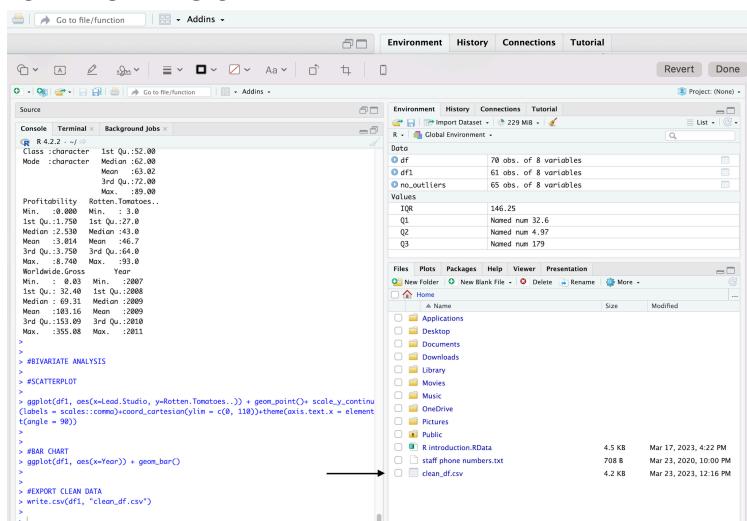


12. BIVARIATE ANALYSIS USING BAR CHART



STEP 4: EXPORT DATA

13. EXPORTING CLEAN DATA



STEP 5: CREATE POWER BI DASHBOARDS

14. THE AVERAGE ROTTEN TOMATOES RATING OF EACH GENRE
 15. THE PROFITABILITY PER STUDIO
 16. THE AUDIENCE SCORE FOR EACH FILM
 17. THE WORLDWIDE GROSS PER GENRE
 18. THE NUMBER OF MOVIES PRODUCED PER YEAR

