3/22/2023

R and Power BI

Analyse the performance of Hollywood movies

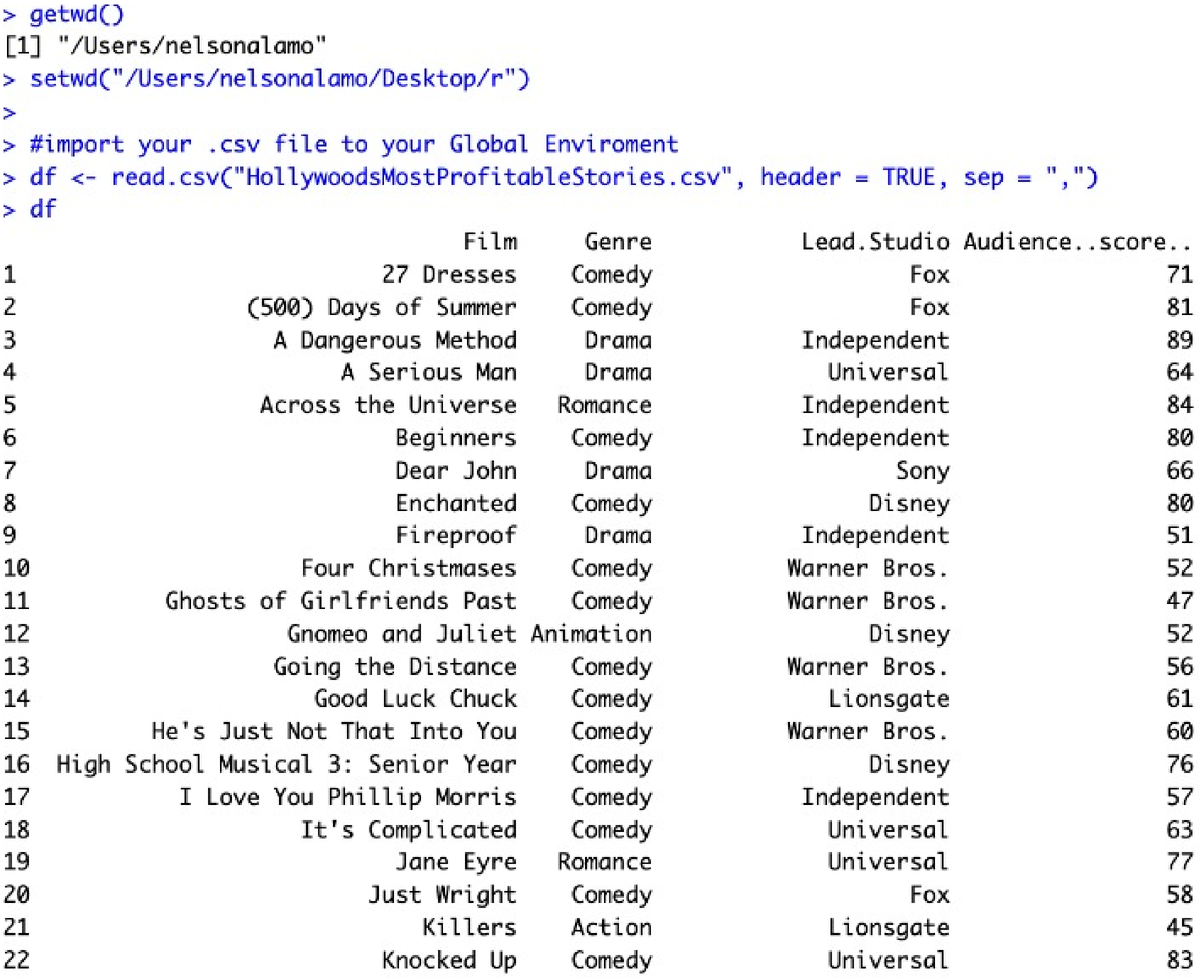
Just IT

Nelson da Graça Leite Alamô

Contents

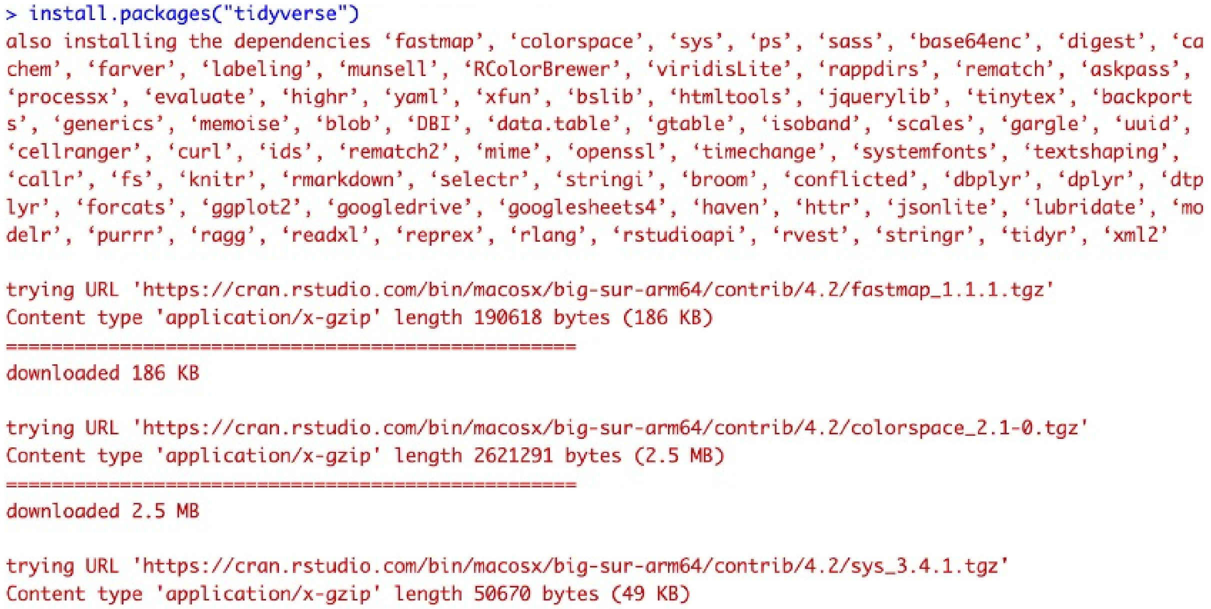
R language

Load data



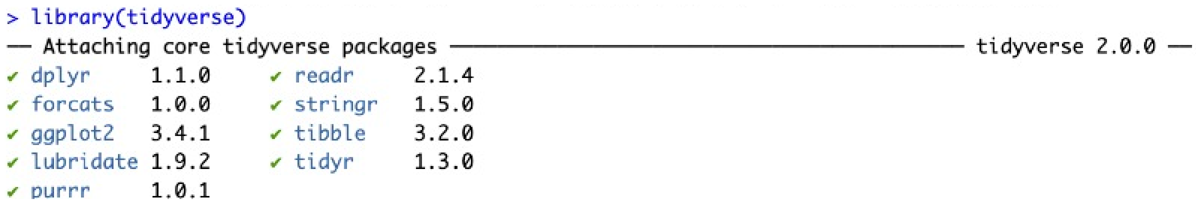
```
> getwd()
[1] "/Users/nelsonalamo"
> setwd("/Users/nelsonalamo/Desktop/r")
>
> #import your .csv file to your Global Enviroment
> df <- read.csv("HollywoodsMostProfitableStories.csv", header = TRUE, sep = ",")
> df
                                Film     Genre          Lead.Studio Audience..score..
1                         27 Dresses    Comedy                  Fox                71
2               (500) Days of Summer    Comedy                  Fox                81
3                 A Dangerous Method     Drama          Independent                89
4                      A Serious Man     Drama            Universal                64
5                Across the Universe   Romance          Independent                84
6                          Beginners    Comedy          Independent                80
7                          Dear John     Drama                 Sony                66
8                          Enchanted    Comedy               Disney                80
9                          Fireproof     Drama          Independent                51
10                   Four Christmases    Comedy         Warner Bros.                52
11         Ghosts of Girlfriends Past    Comedy         Warner Bros.                47
12                  Gnomeo and Juliet Animation               Disney                52
13                 Going the Distance    Comedy         Warner Bros.                56
14                    Good Luck Chuck    Comedy            Lionsgate                61
15         He's Just Not That Into You    Comedy         Warner Bros.                60
16  High School Musical 3: Senior Year    Comedy               Disney                76
17            I Love You Phillip Morris    Comedy          Independent                57
18                   It's Complicated    Comedy            Universal                63
19                          Jane Eyre   Romance            Universal                77
20                       Just Wright    Comedy                  Fox                58
21                           Killers    Action            Lionsgate                45
22                        Knocked Up    Comedy            Universal                83
```

Load library



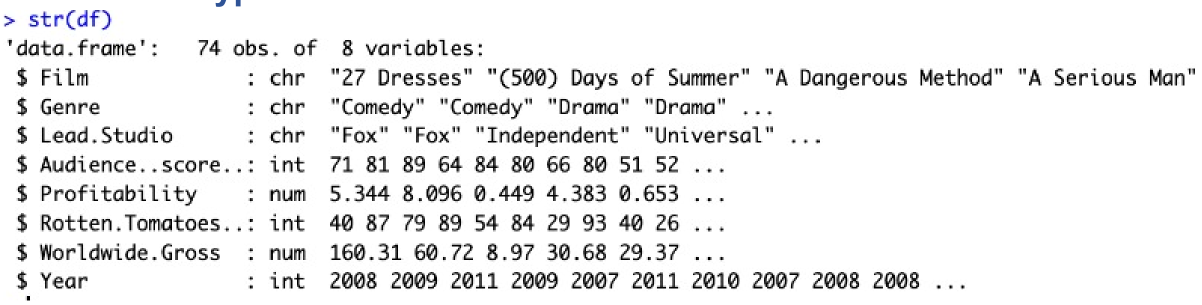
```
> install.packages("tidyverse")
also installing the dependencies ‘fastmap’, ‘colorspace’, ‘sys’, ‘ps’, ‘sass’, ‘base64enc’, ‘digest’, ‘ca
chem’, ‘farver’, ‘labeling’, ‘munsell’, ‘RColorBrewer’, ‘viridisLite’, ‘rappdirs’, ‘rematch’, ‘askpass’,
‘processx’, ‘evaluate’, ‘highr’, ‘yaml’, ‘xfun’, ‘bslib’, ‘htmltools’, ‘jquerylib’, ‘tinytex’, ‘backport
s’, ‘generics’, ‘memoise’, ‘blob’, ‘DBI’, ‘data.table’, ‘gtable’, ‘isoband’, ‘scales’, ‘gargle’, ‘uuid’,
‘cellranger’, ‘curl’, ‘ids’, ‘rematch2’, ‘mime’, ‘openssl’, ‘timechange’, ‘systemfonts’, ‘textshaping’,
‘callr’, ‘fs’, ‘knitr’, ‘rmarkdown’, ‘selectr’, ‘stringi’, ‘broom’, ‘conflicted’, ‘dbplyr’, ‘dplyr’, ‘dtp
lyr’, ‘forcats’, ‘ggplot2’, ‘googledrive’, ‘googlesheets4’, ‘haven’, ‘httr’, ‘jsonlite’, ‘lubridate’, ‘mo
delr’, ‘purrr’, ‘ragg’, ‘readxl’, ‘reprex’, ‘rlang’, ‘rstudioapi’, ‘rvest’, ‘stringr’, ‘tidyr’, ‘xml2’

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/fastmap_1.1.1.tgz'
Content type 'application/x-gzip' length 190618 bytes (186 KB)
==================================================
downloaded 186 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/colorspace_2.1-0.tgz'
Content type 'application/x-gzip' length 2621291 bytes (2.5 MB)
==================================================
downloaded 2.5 MB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.2/sys_3.4.1.tgz'
Content type 'application/x-gzip' length 50670 bytes (49 KB)
```

Import library.



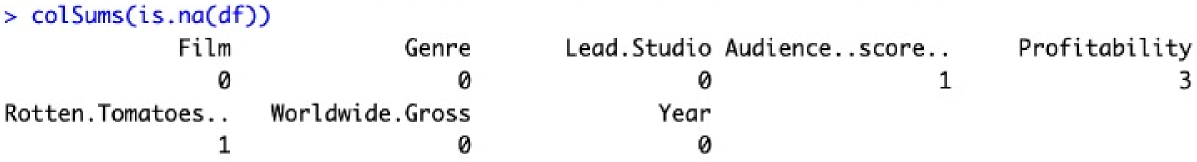
```
> library(tidyverse)
── Attaching core tidyverse packages ──────────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.0     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.1     ✔ tibble    3.2.0
✔ lubridate 1.9.2     ✔ tidyr     1.3.0
✔ purrr     1.0.1
```

Check data types.



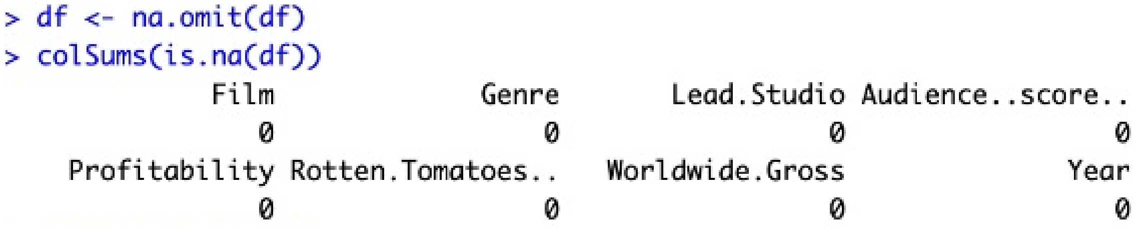
```
> str(df)
'data.frame':   74 obs. of  8 variables:
 $ Film             : chr  "27 Dresses" "(500) Days of Summer" "A Dangerous Method" "A Serious Man"
 $ Genre            : chr  "Comedy" "Comedy" "Drama" "Drama" ...
 $ Lead.Studio      : chr  "Fox" "Fox" "Independent" "Universal" ...
 $ Audience..score..: 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 ...
```

Check for missing values.



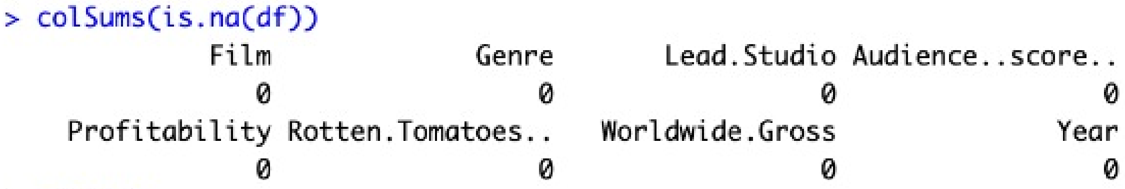
```
> colSums(is.na(df))
             Film             Genre       Lead.Studio Audience..score..     Profitability
                0                 0                 0                 1                 3
Rotten.Tomatoes..   Worldwide.Gross              Year
                1                 0                 0
```

Drop missing values



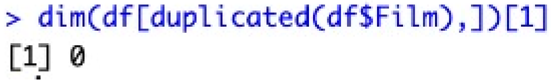
```
> df <- na.omit(df)
> colSums(is.na(df))
             Film             Genre       Lead.Studio Audience..score..
                0                 0                 0                 0
    Profitability Rotten.Tomatoes..   Worldwide.Gross              Year
                0                 0                 0                 0
```

Check to make sure that the rows have been removed.



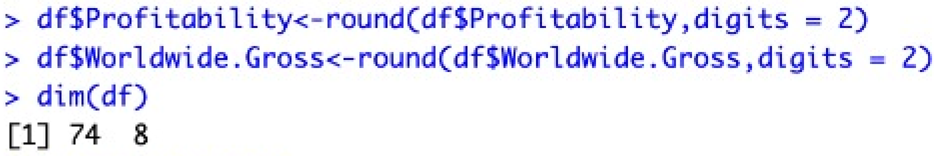
```
> colSums(is.na(df))
             Film             Genre       Lead.Studio Audience..score..
                0                 0                 0                 0
    Profitability Rotten.Tomatoes..   Worldwide.Gross              Year
                0                 0                 0                 0
```

Check for duplicates.



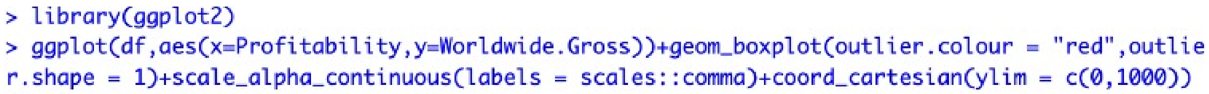
```
> dim(df[duplicated(df$Film),])[1]
[1] 0
```

Round off values to 2 places.

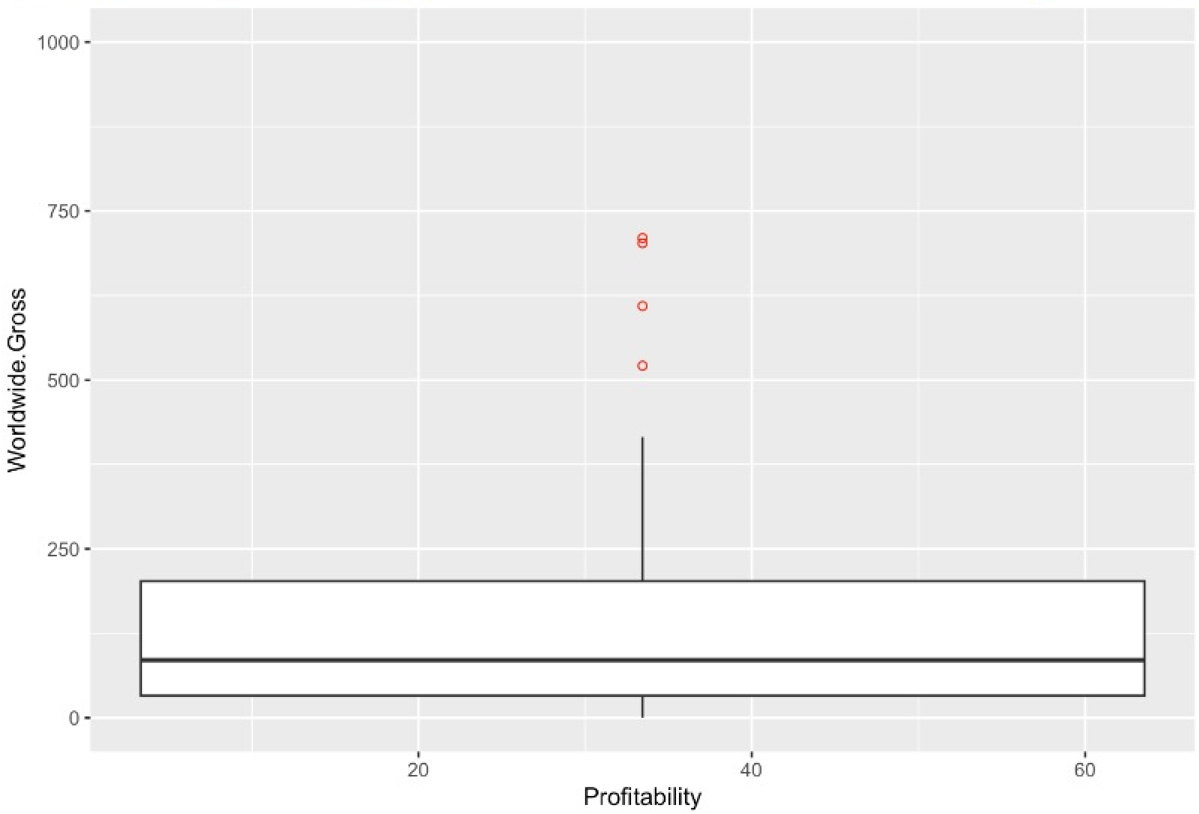


```
> df$Profitability<-round(df$Profitability,digits = 2)
> df$Worldwide.Gross<-round(df$Worldwide.Gross,digits = 2)
> dim(df)
[1] 74  8
```

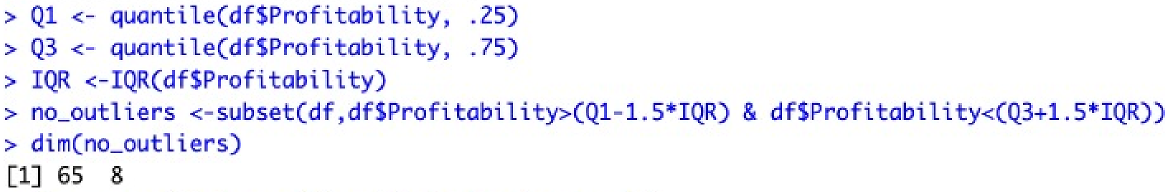
Create a boxplot that highlights the outliers, and check.



```
> library(ggplot2)
> ggplot(df,aes(x=Profitability,y=Worldwide.Gross))+geom_boxplot(outlier.colour = "red",outlie
r.shape = 1)+scale_alpha_continuous(labels = scales::comma)+coord_cartesian(ylim = c(0,1000))
```

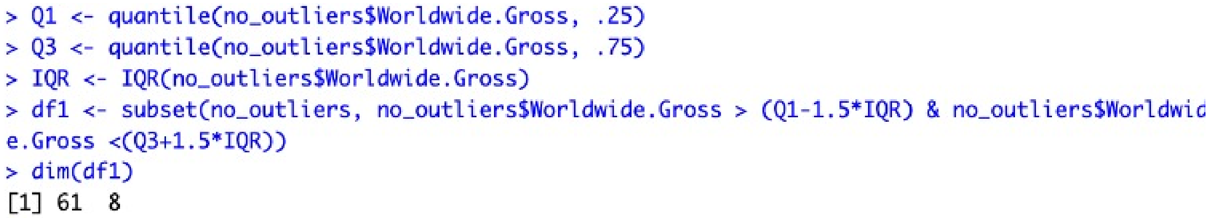


Remove outliers in 'Profitability’.



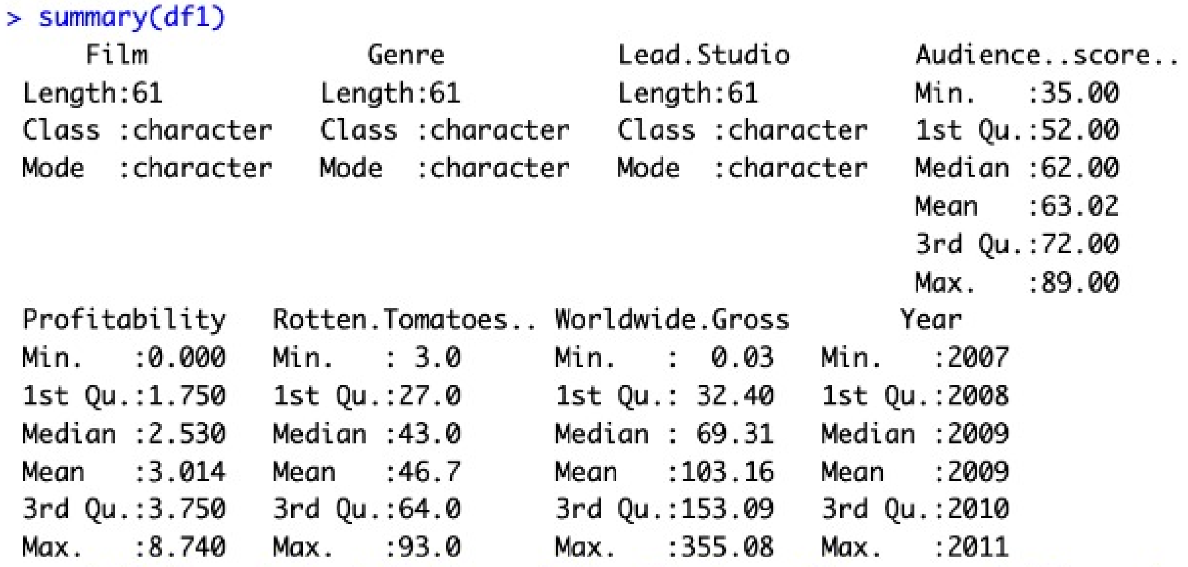
```
> Q1 <- quantile(df$Profitability, .25)
> Q3 <- quantile(df$Profitability, .75)
> IQR <-IQR(df$Profitability)
> no_outliers <-subset(df,df$Profitability>(Q1-1.5*IQR) & df$Profitability<(Q3+1.5*IQR))
> dim(no_outliers)
[1] 65  8
```

Remove outliers in 'Worldwide. Gross'



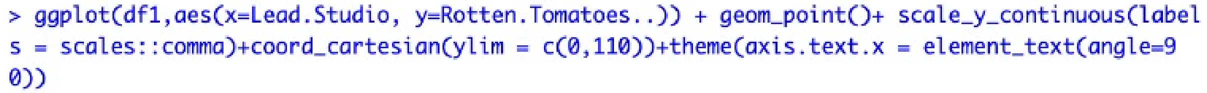
```
> Q1 <- quantile(no_outliers$Worldwide.Gross, .25)
> Q3 <- quantile(no_outliers$Worldwide.Gross, .75)
> IQR <- IQR(no_outliers$Worldwide.Gross)
> df1 <- subset(no_outliers, no_outliers$Worldwide.Gross > (Q1-1.5*IQR) & no_outliers$Worldwid
e.Gross <(Q3+1.5*IQR))
> dim(df1)
[1] 61  8
```

Summary Statistics/Univariate Analysis

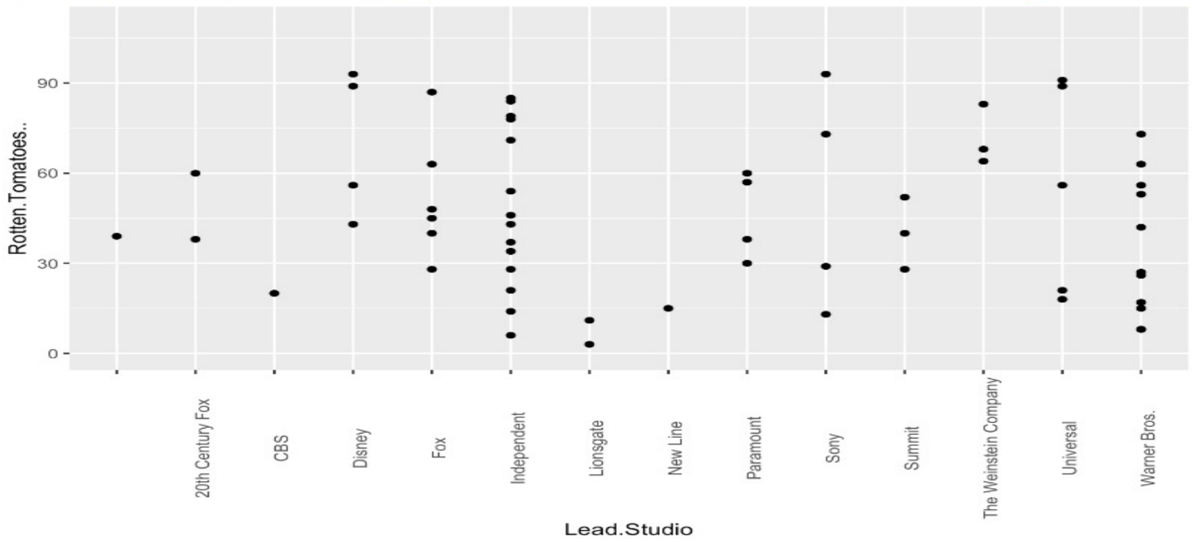


```
> summary(df1)
     Film              Genre            Lead.Studio        Audience..score..
 Length:61          Length:61          Length:61          Min.   :35.00
 Class :character   Class :character   Class :character   1st Qu.:52.00
 Mode  :character   Mode  :character   Mode  :character   Median :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
```

scatterplot

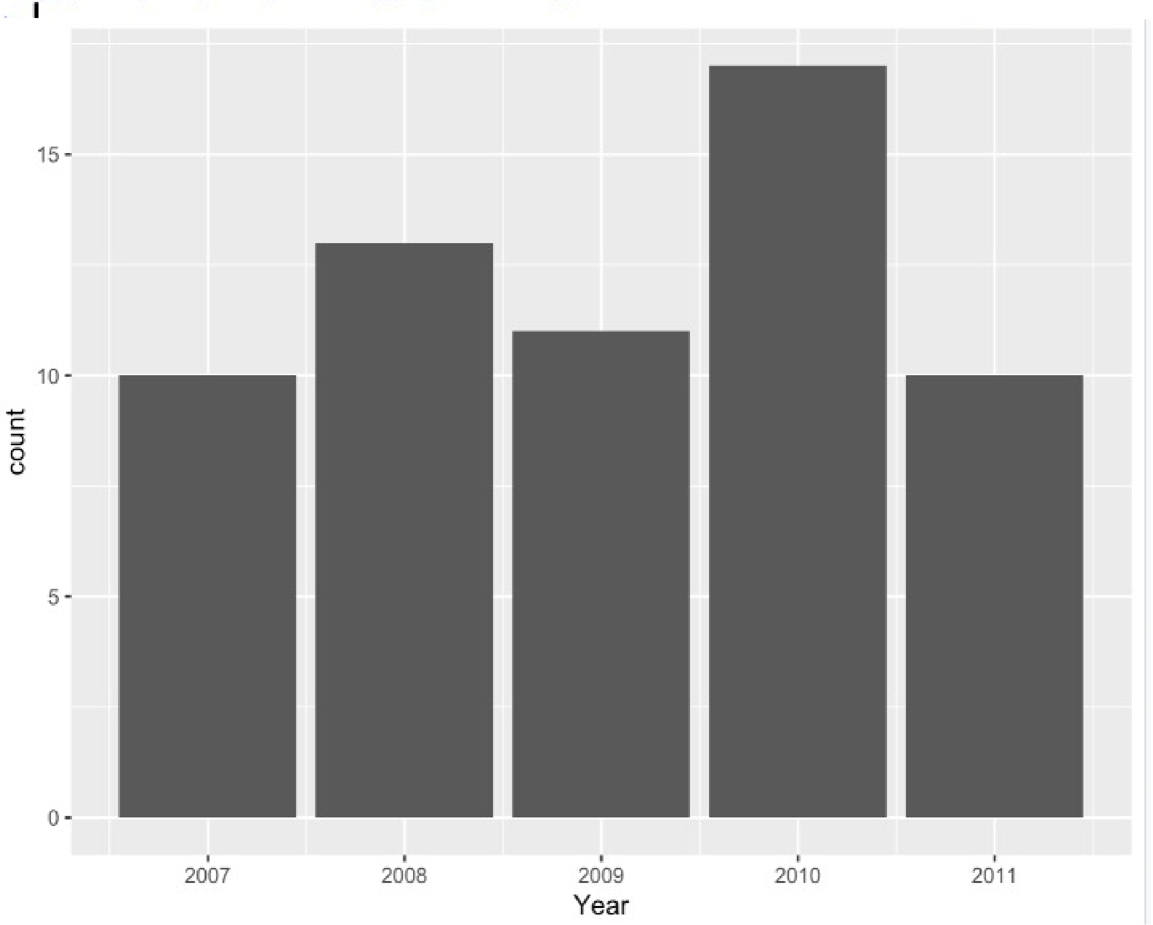


```
> ggplot(df1,aes(x=Lead.Studio, y=Rotten.Tomatoes..)) + geom_point()+ scale_y_continuous(label
s = scales::comma)+coord_cartesian(ylim = c(0,110))+theme(axis.text.x = element_text(angle=9
0))
```



Bar chart

```
> ggplot(df1,aes(x=Year))+geom_bar()
```

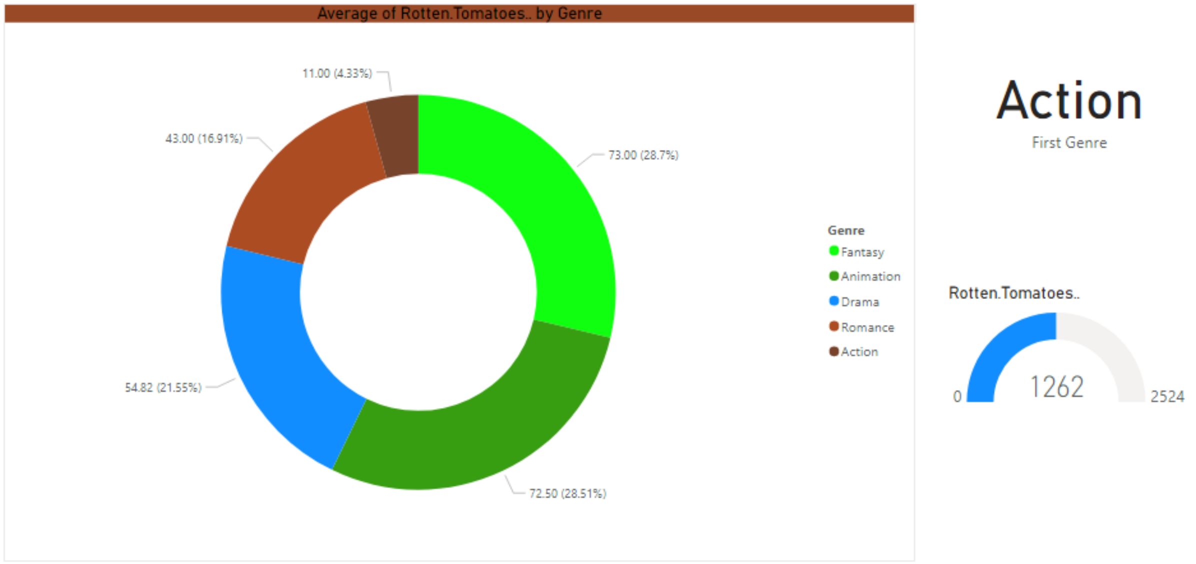


Export clean data

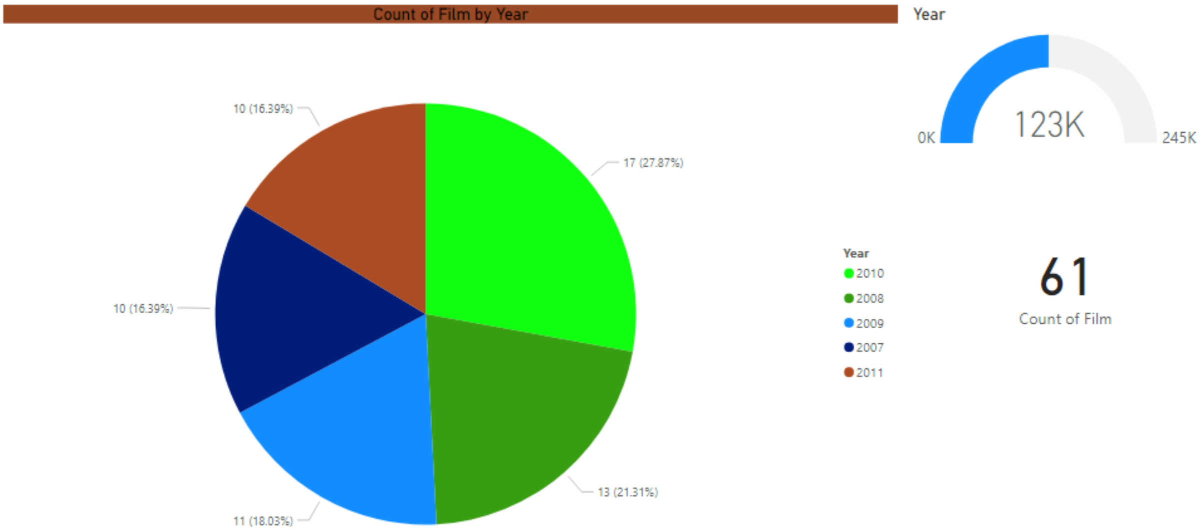
```
> write.csv(df1,"clean_df.csv")
```

Power BI

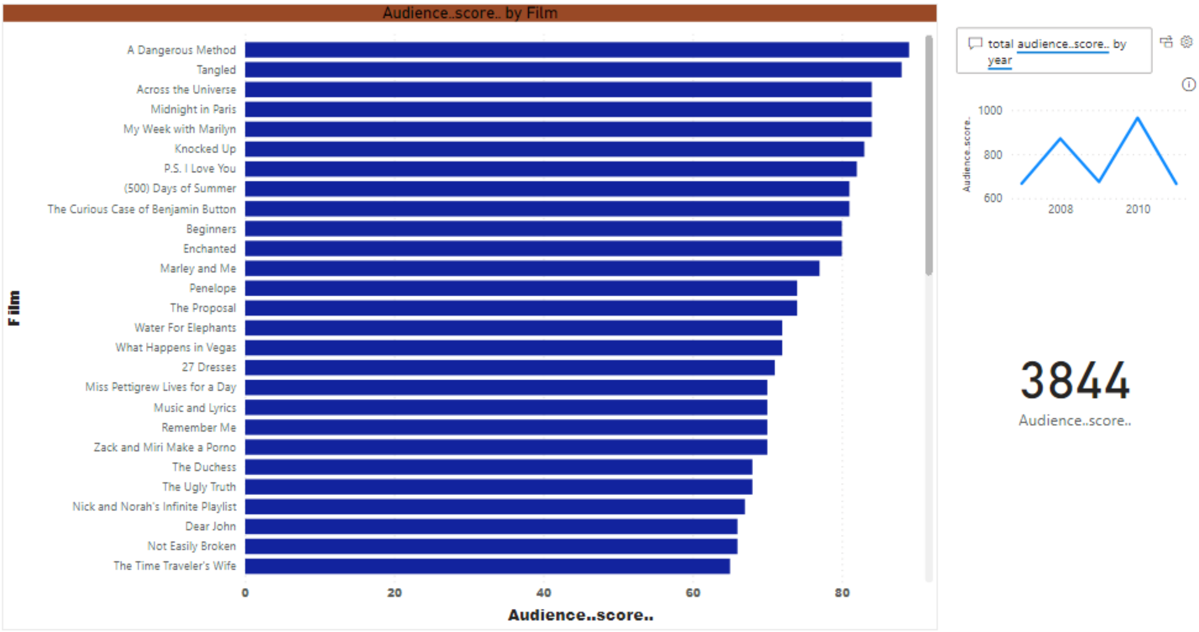
The average Rotten Tomatoes ratings of each genre.



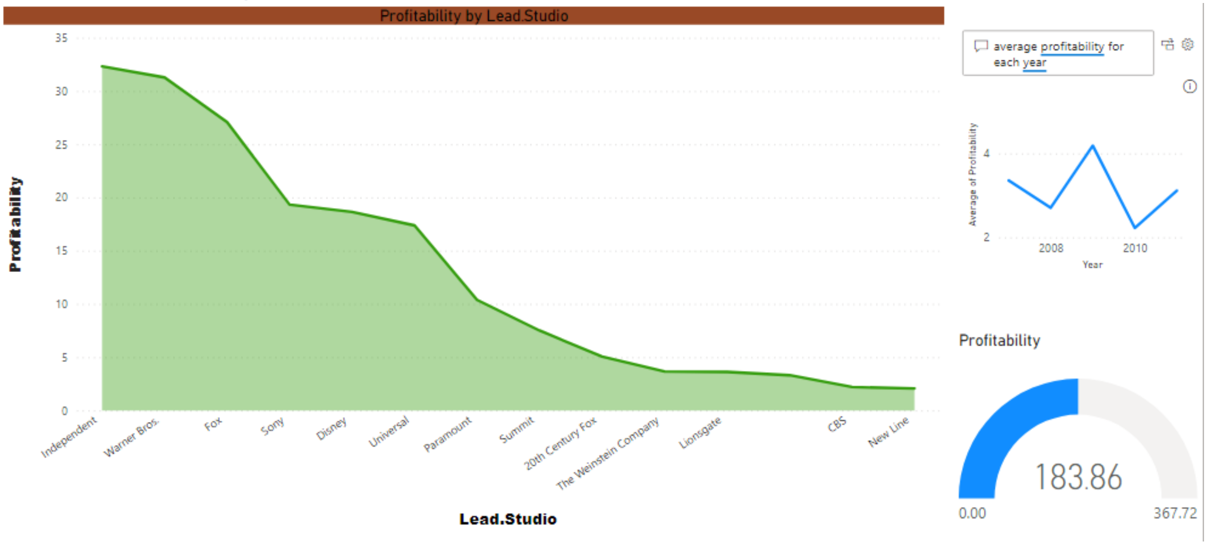
The number of movies produced per year.



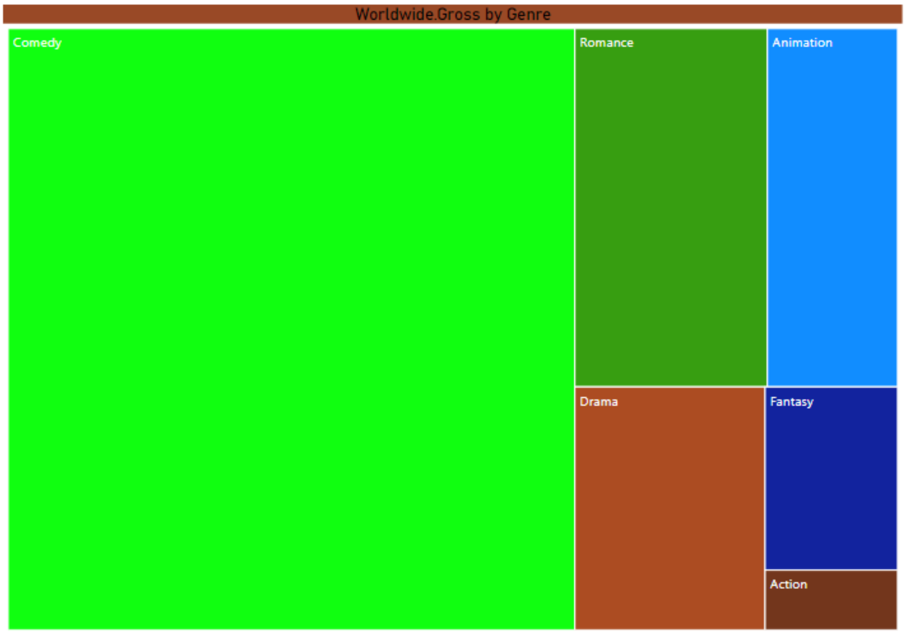
The audience score for each film.

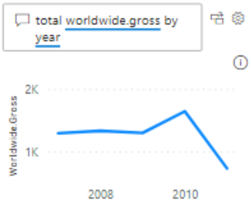


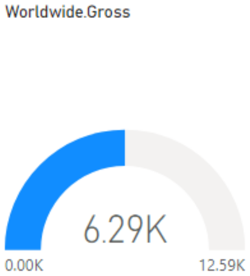
The profitability per studio.



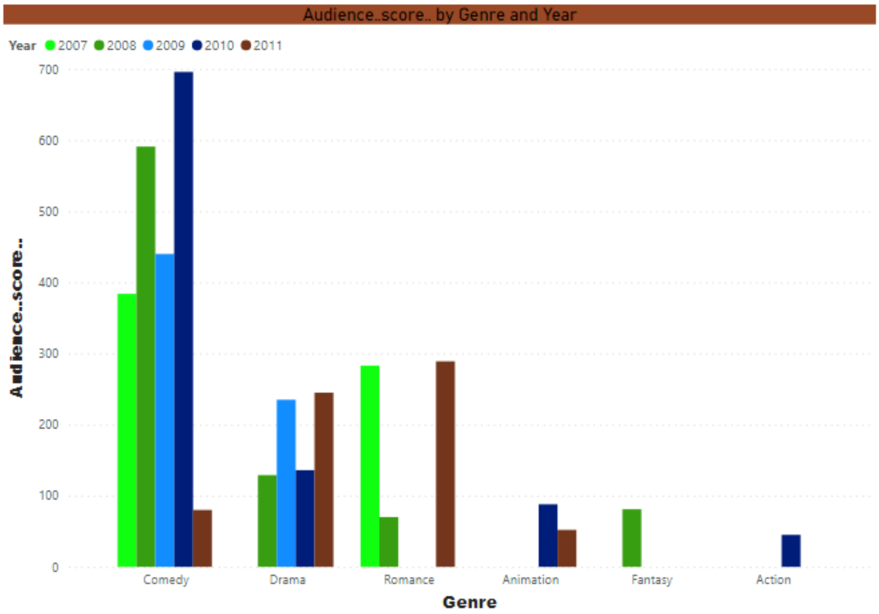
The worldwide gross per genre.

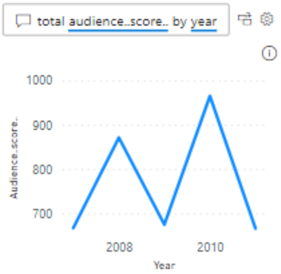


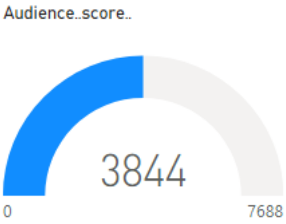




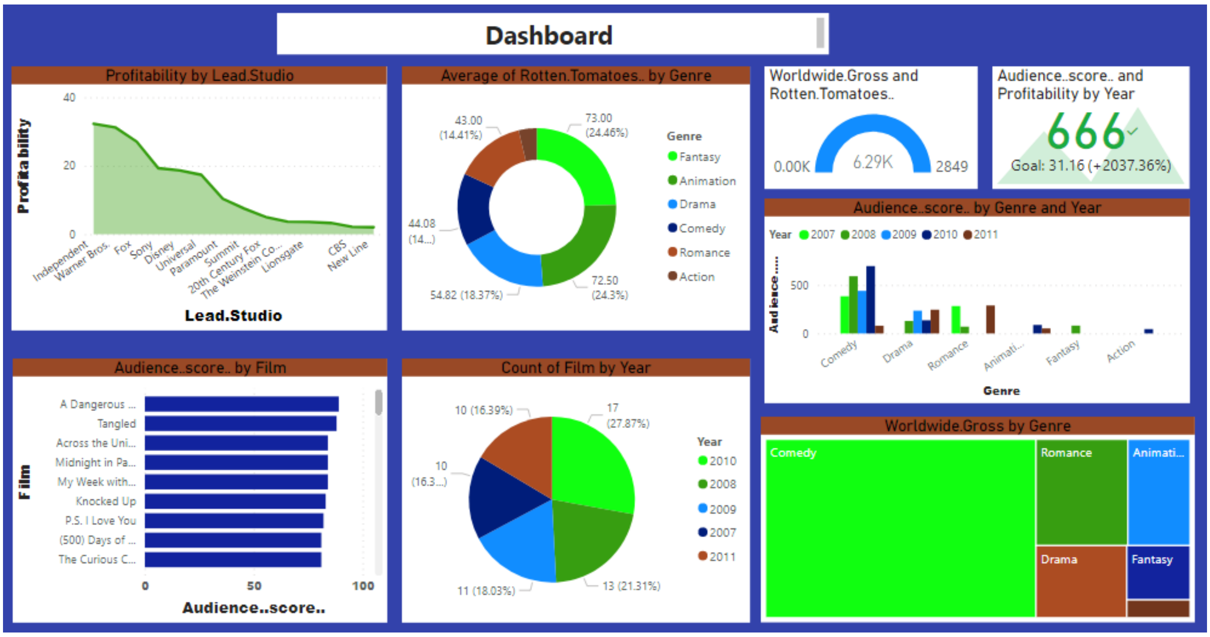
The audience score genre per year.







Dashboard



Reference

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

https://informationisbeautiful.net/