R AND POWER BI

An analysis of Hollywood movies

Alex Thompson
Just IT Data Technician Bootcamp 03/12/2024

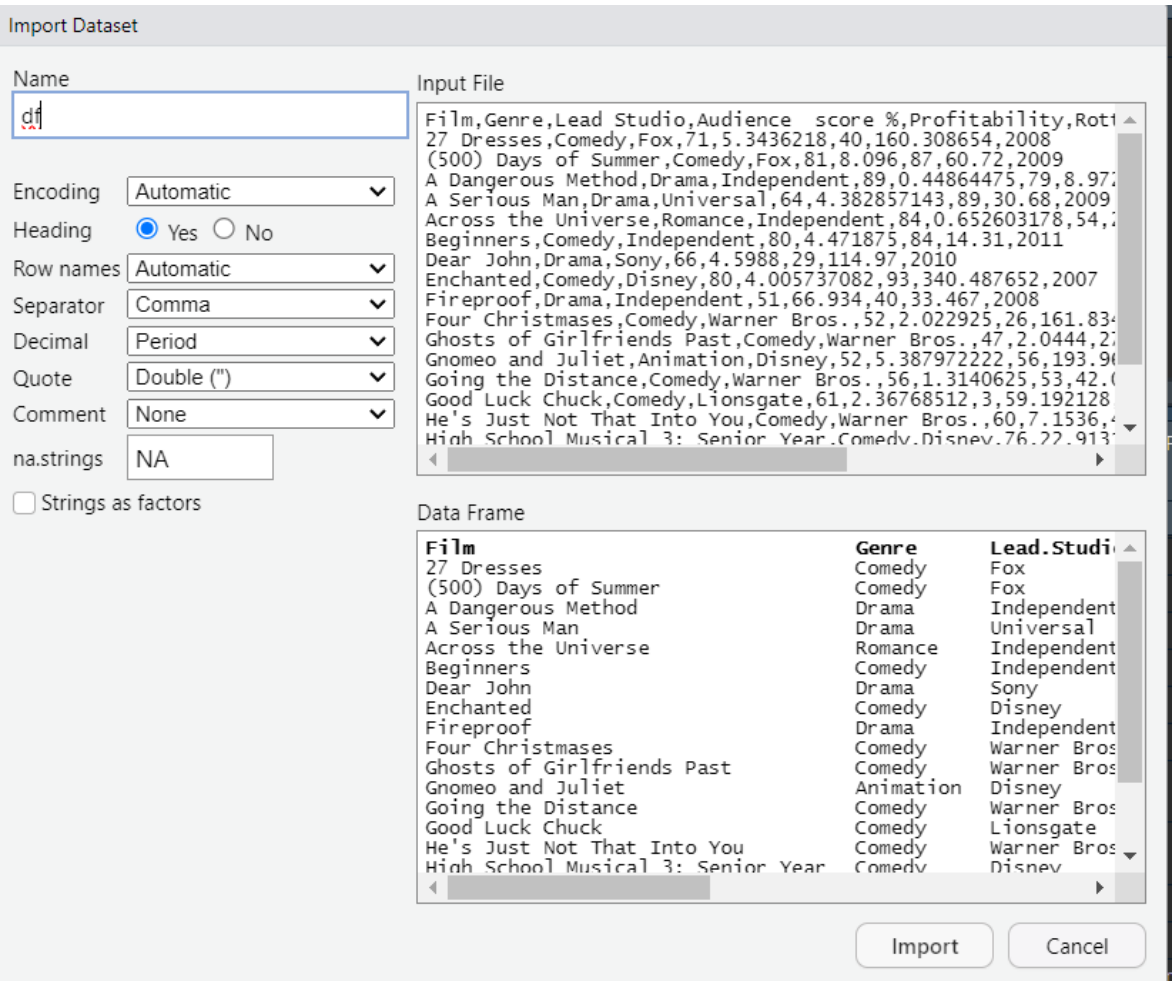
Table of Contents

Upload Data into R

- Import tidyverse library package

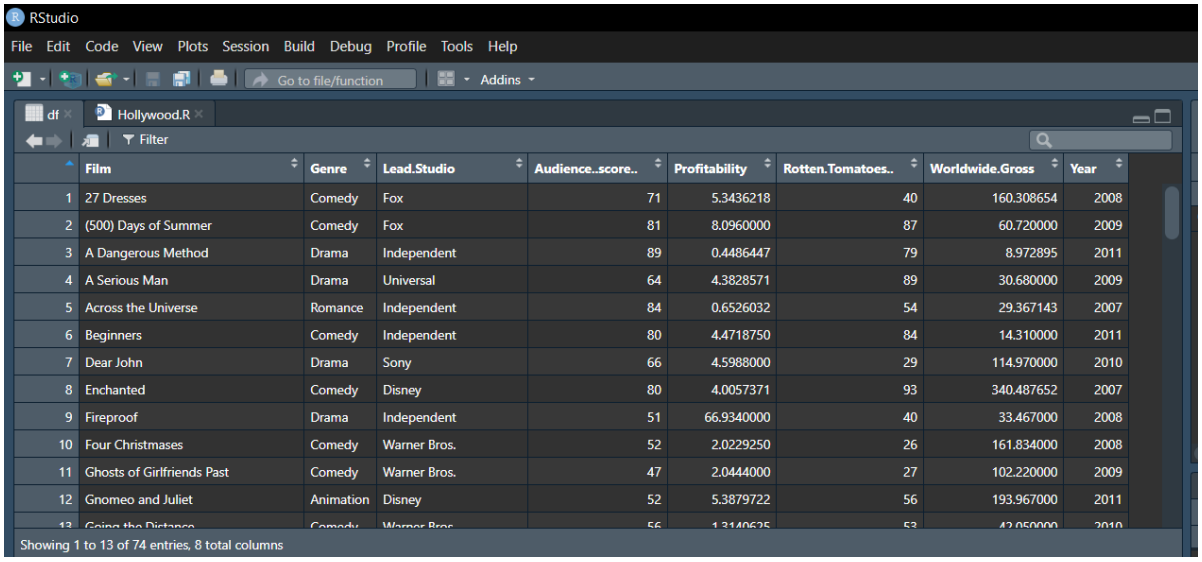
```
Console   Terminal   Background Jobs
R 4.4.2 · ~/
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(tidyverse)
── Attaching core tidyverse packages ──────────────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────────────── tidyverse_conflicts() ──
✗ dplyr::filter() masks stats::filter()
✗ dplyr::lag()    masks stats::lag()
```

- Import HollywoodsMostProfitableStories database



Check Database

- View(df)

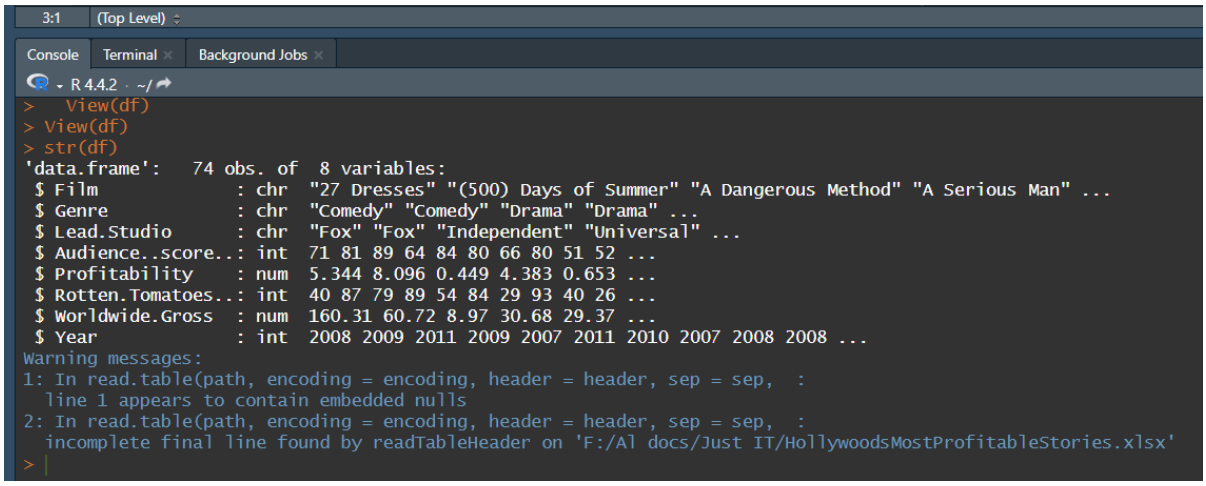


df | Hollywood.R

	Film	Genre	Lead.Studio	Audience..score..	Profitability	Rotten.Tomatoes..	Worldwide.Gross	Year
1	27 Dresses	Comedy	Fox	71	5.3436218	40	160.308654	2008
2	(500) Days of Summer	Comedy	Fox	81	8.0960000	87	60.720000	2009
3	A Dangerous Method	Drama	Independent	89	0.4486447	79	8.972895	2011
4	A Serious Man	Drama	Universal	64	4.3828571	89	30.680000	2009
5	Across the Universe	Romance	Independent	84	0.6526032	54	29.367143	2007
6	Beginners	Comedy	Independent	80	4.4718750	84	14.310000	2011
7	Dear John	Drama	Sony	66	4.5988000	29	114.970000	2010
8	Enchanted	Comedy	Disney	80	4.0057371	93	340.487652	2007
9	Fireproof	Drama	Independent	51	66.9340000	40	33.467000	2008
10	Four Christmases	Comedy	Warner Bros.	52	2.0229250	26	161.834000	2008
11	Ghosts of Girlfriends Past	Comedy	Warner Bros.	47	2.0444000	27	102.220000	2009
12	Gnomeo and Juliet	Animation	Disney	52	5.3879722	56	193.967000	2011

Showing 1 to 13 of 74 entries, 8 total columns

- str(df)



```
>   View(df)
> View(df)
> 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 ...
Warning messages:
1: In read.table(path, encoding = encoding, header = header, sep = sep,  :
  line 1 appears to contain embedded nulls
2: In read.table(path, encoding = encoding, header = header, sep = sep,  :
  incomplete final line found by readTableHeader on 'F:/Al docs/Just IT/HollywoodsMostProfitableStories.xlsx'
>
```

Clean Database

- colSums(is.na(df))
- df<-na.omit(df)
- colSums(is.na(df))

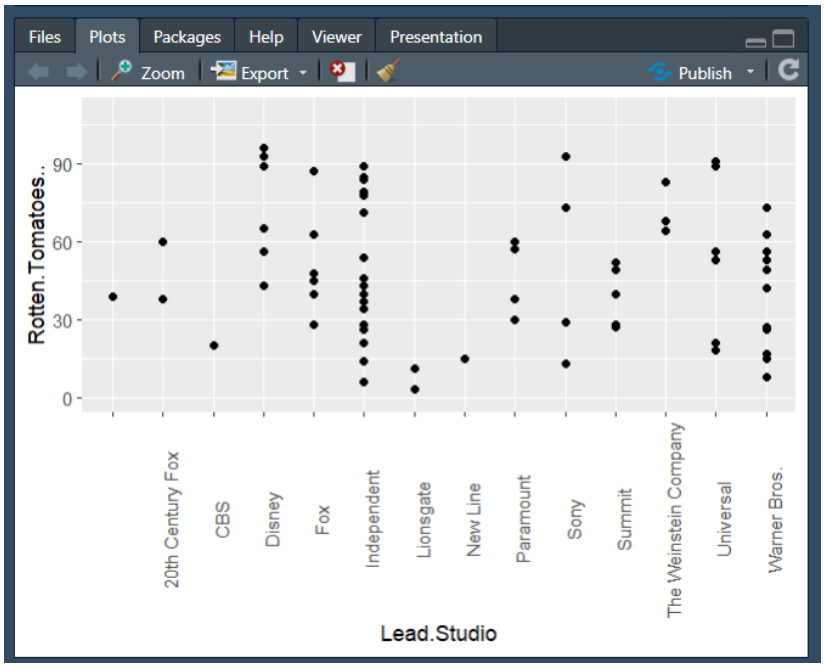
```
6:1   (Top Level)
Console   Terminal   Background Jobs
R 4.4.2 · ~/
 $ Year            : int  2008 2009 2011 2009 2007 2011 2010 2007 2008 2008 ...
Warning messages:
1: In read.table(path, encoding = encoding, header = header, sep = sep,  :
  line 1 appears to contain embedded nulls
2: In read.table(path, encoding = encoding, header = header, sep = sep,  :
  incomplete final line found by readTableHeader on 'F:/Al docs/Just IT/HollywoodsMostProfitableStories.xlsx'
> colSums(is.na(df))
             Film             Genre       Lead.Studio Audience..score..     Profitability Rotten.Tomatoes..
                0                 0                 0                1                 3                 1
  Worldwide.Gross              Year
                0                 0
> df<-na.omit(df)
> colSums(is.na(df))
             Film             Genre       Lead.Studio Audience..score..     Profitability Rotten.Tomatoes..
                0                 0                 0                0                 0                 0
  Worldwide.Gross              Year
                0                 0
>
```

- summary(df)

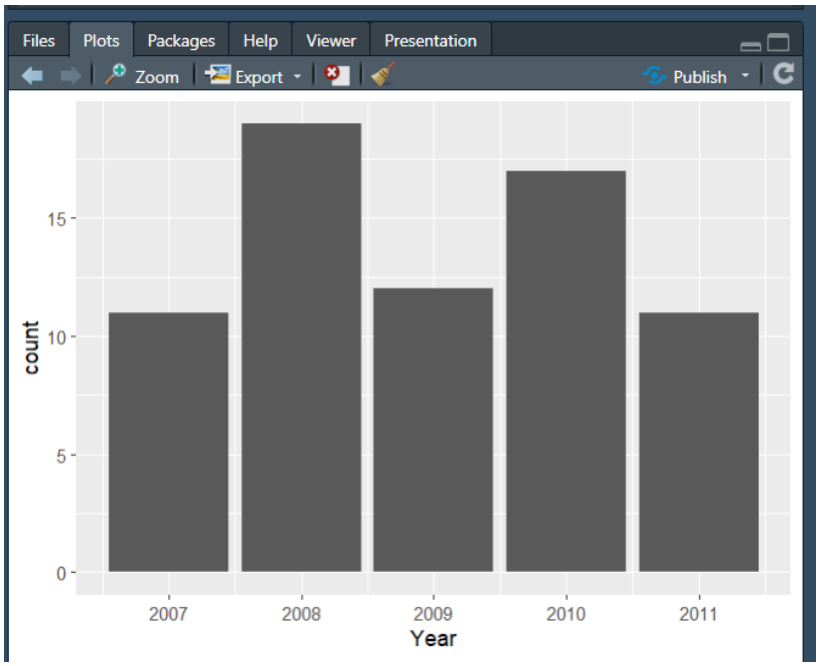
```
7:1   (Top Level)
Console   Terminal   Background Jobs
R 4.4.2 · ~/
  Worldwide.Gross              Year
                0                 0
> summary(df)
     Film             Genre           Lead.Studio        Audience..score..  Profitability     Rotten.Tomatoes..
 Length:70          Length:70          Length:70          Min.   :35.00     Min.   : 0.005   Min.   : 3.00
 Class :character   Class :character   Class :character   1st Qu.:53.25     1st Qu.: 1.802   1st Qu.:27.25
 Mode  :character   Mode  :character   Mode  :character   Median :64.50     Median : 2.646   Median :45.50
                                                          Mean   :64.46     Mean   : 4.785   Mean   :47.76
                                                          3rd Qu.:75.50     3rd Qu.: 4.977   3rd Qu.:64.75
                                                          Max.   :89.00     Max.   :66.934   Max.   :96.00

 Worldwide.Gross        Year
 Min.   :  0.025   Min.   :2007
 1st Qu.: 32.809   1st Qu.:2008
 Median : 85.891   Median :2009
 Mean   :141.933   Mean   :2009
 3rd Qu.:202.467   3rd Qu.:2010
 Max.   :709.820   Max.   :2011
>
```

Analyse Database in R

- ggplot(df, 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))



- ggplot(df, aes(x=Year)) + geom_bar()



Import cleaned Database into Power BI

- write.csv(df, "clean_df.csv")

```
9:1   (Top Level) ÷                                                                                R Script
Console   Terminal ×   Background Jobs ×
R 4.4.2 · ~/
 Mode  :character   Mode  :character   Mode  :character   Median :64.50     Median : 2.646   Median :45.50
                                                          Mean   :64.46     Mean   : 4.785   Mean   :47.76
                                                          3rd Qu.:75.50     3rd Qu.: 4.977   3rd Qu.:64.75
                                                          Max.   :89.00     Max.   :66.934   Max.   :96.00
 Worldwide.Gross        Year
 Min.   :  0.025   Min.   :2007
 1st Qu.: 32.809   1st Qu.:2008
 Median : 85.891   Median :2009
 Mean   :141.933   Mean   :2009
 3rd Qu.:202.467   3rd Qu.:2010
 Max.   :709.820   Max.   :2011
> ggplot(df, aes(x=Lead.Studio, y=Rotten.Tomatoes..)) + geom_point()+ scale_y_continuous(labels = scales::comma)+coord_c
artesian(ylim = c(0, 110))+theme(axis.text.x = element_text(angle = 90))
> ggplot(df, aes(x=Year)) + geom_bar()
> write.csv(df, "clean_df.csv")
> view(clean_df)
```

- Import into Power BI

clean_df.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

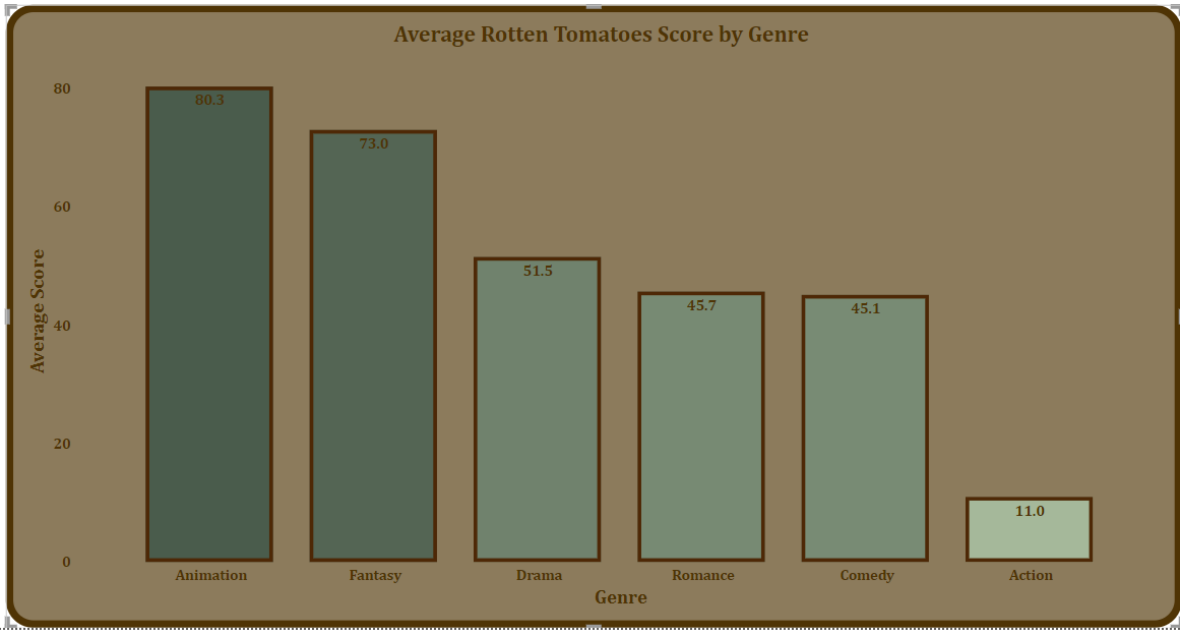
Data Type Detection: Based on first 200 rows

	Film	Genre	Lead.Studio	Audience..score..	Profitability	Rotten.Tomatoes..	Worldwide.Gross	Yea
1	27 Dresses	Comedy	Fox	71	5.3436218	40	160.308654	
2	(500) Days of Summer	Comedy	Fox	81	8.096	87	60.72	
3	A Dangerous Method	Drama	Independent	89	0.44864475	79	8.972895	
4	A Serious Man	Drama	Universal	64	4.382857143	89	30.68	
5	Across the Universe	Romance	Independent	84	0.652603178	54	29.367143	
6	Beginners	Comedy	Independent	80	4.471875	84	14.31	
7	Dear John	Drama	Sony	66	4.5988	29	114.97	
8	Enchanted	Comedy	Disney	80	4.005737082	93	340.487652	
9	Fireproof	Drama	Independent	51	66.934	40	33.467	
10	Four Christmases	Comedy	Warner Bros.	52	2.022925	26	161.834	
11	Ghosts of Girlfriends Past	Comedy	Warner Bros.	47	2.0444	27	102.22	
12	Gnomeo and Juliet	Animation	Disney	52	5.387972222	56	193.967	
13	Going the Distance	Comedy	Warner Bros.	56	1.3140625	53	42.05	
14	Good Luck Chuck	Comedy	Lionsgate	61	2.36768512	3	59.192128	
15	He's Just Not That Into You	Comedy	Warner Bros.	60	7.1536	42	178.84	
16	High School Musical 3: Senior Year	Comedy	Disney	76	22.91313646	65	252.044501	
17	I Love You Phillip Morris	Comedy	Independent	57	1.34	71	20.1	
18	It's Complicated	Comedy	Universal	63	2.642352941	56	224.6	
20	Just Wright	Comedy	Fox	58	1.797416667	45	21.569	
21	Killers	Action	Lionsgate	45	1.245333333	11	93.4	

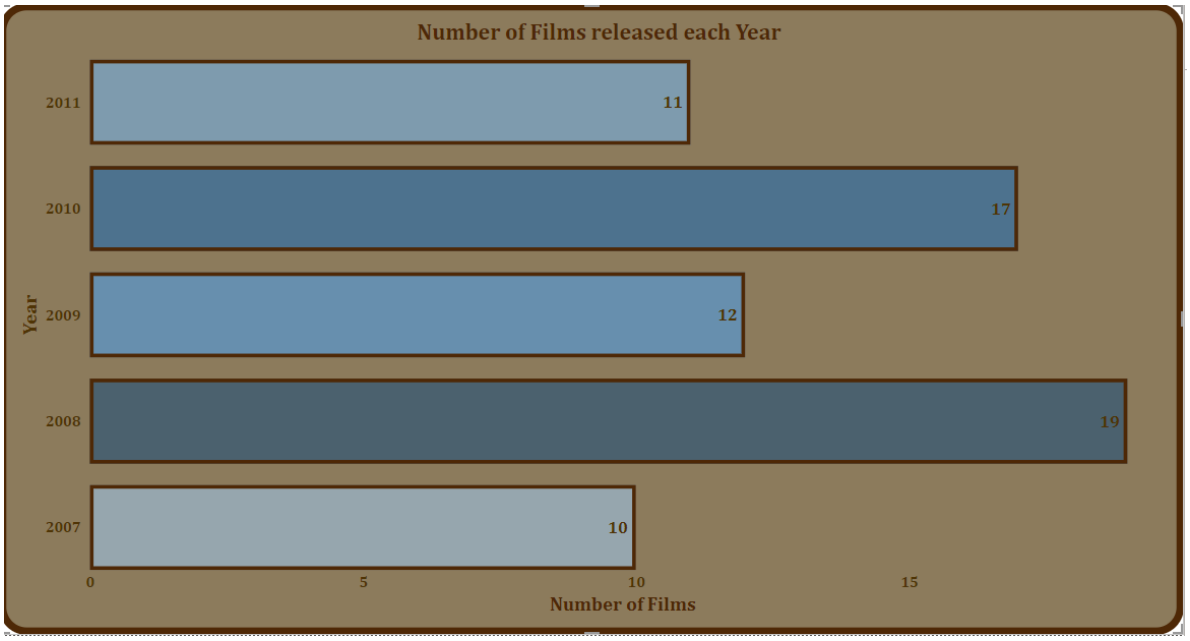
Extract Table Using Examples | Load | Transform Data | Cancel

Power BI Data Visuals

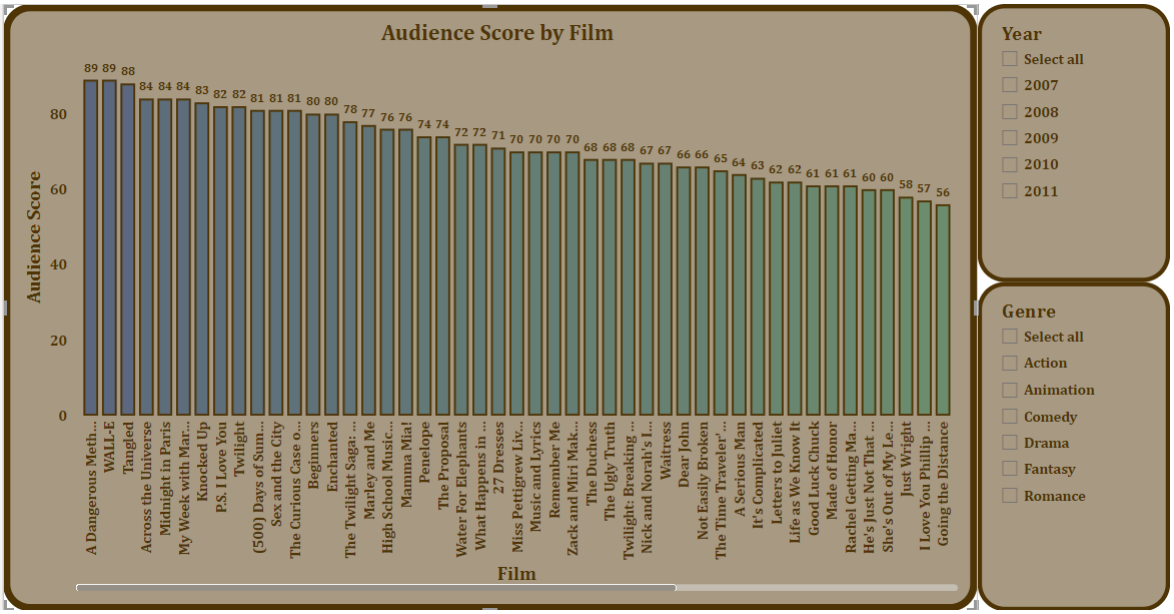
- Average Rotten Tomatoes score by genre



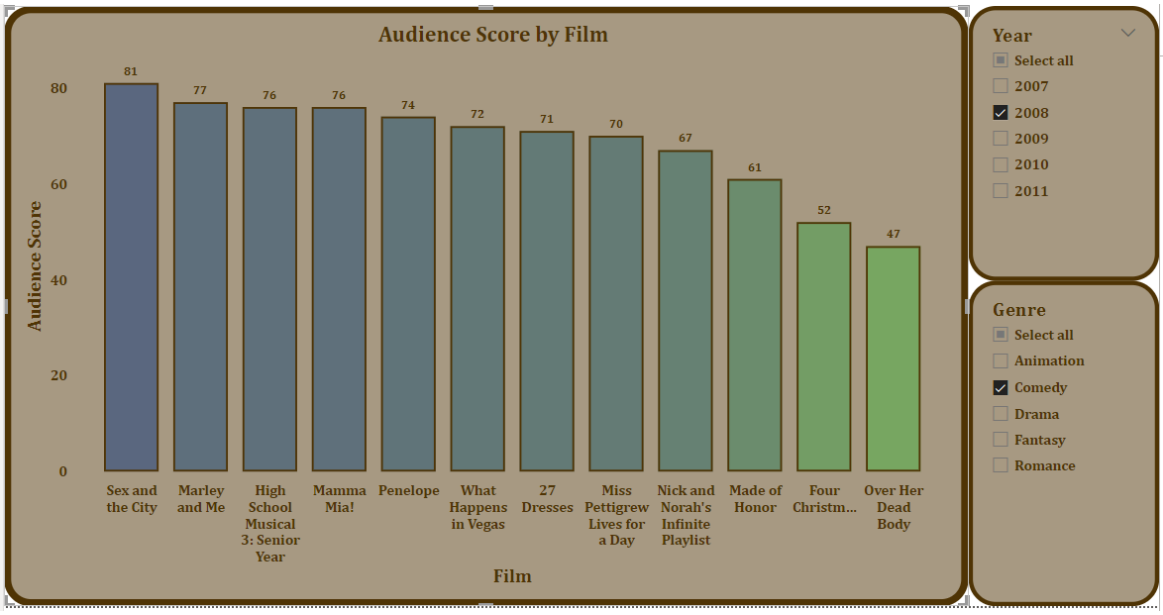
- Number of films released each year



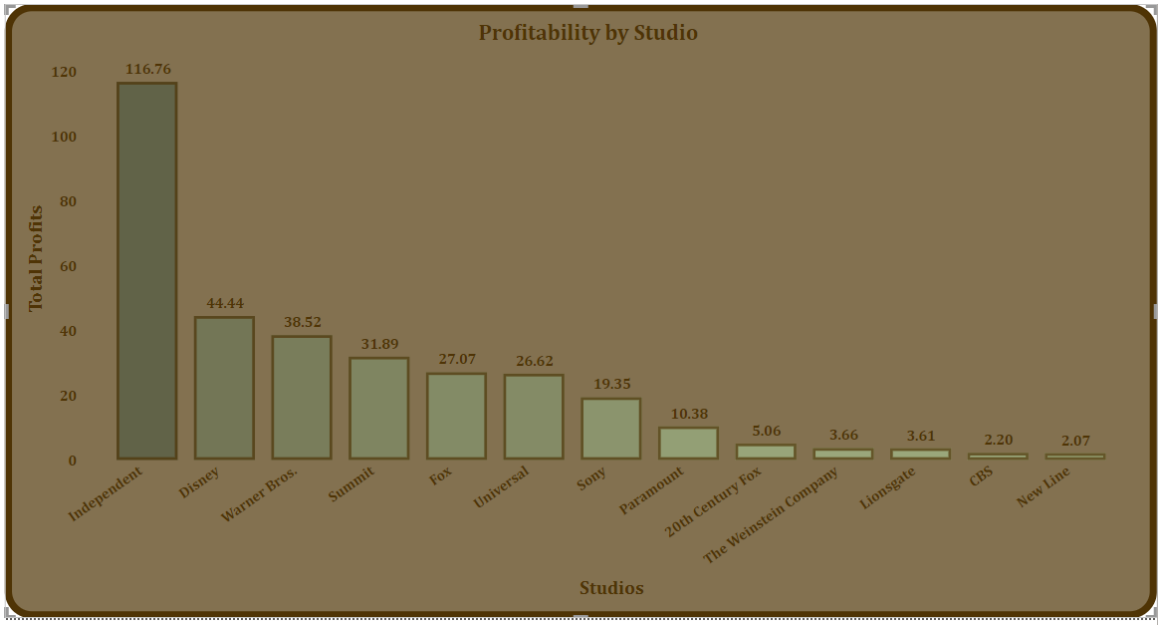
- Audience scores for each film



- 2 slicers included for Year and Genre



- Profitability per Studio



- Worldwide Gross Profit by Genre