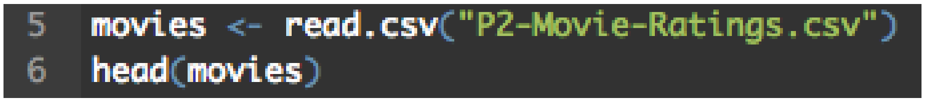
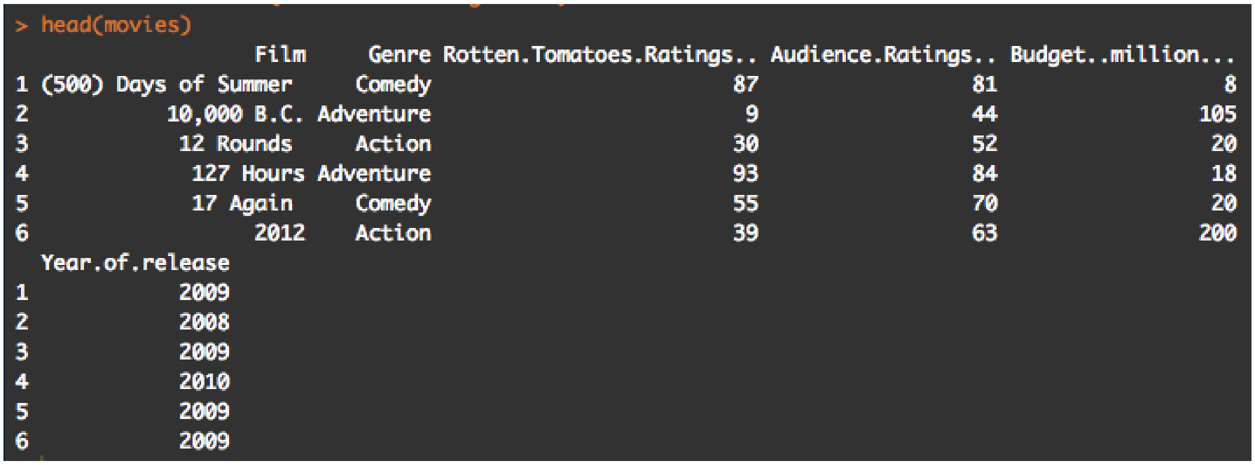
Movie Trends in Ratings and Budgets 2007-2011

First step in the analysis is to read the csv file



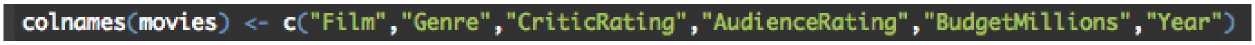
```r
movies <- read.csv("P2-Movie-Ratings.csv")
head(movies)
```

Let’s see what kind of data we have.



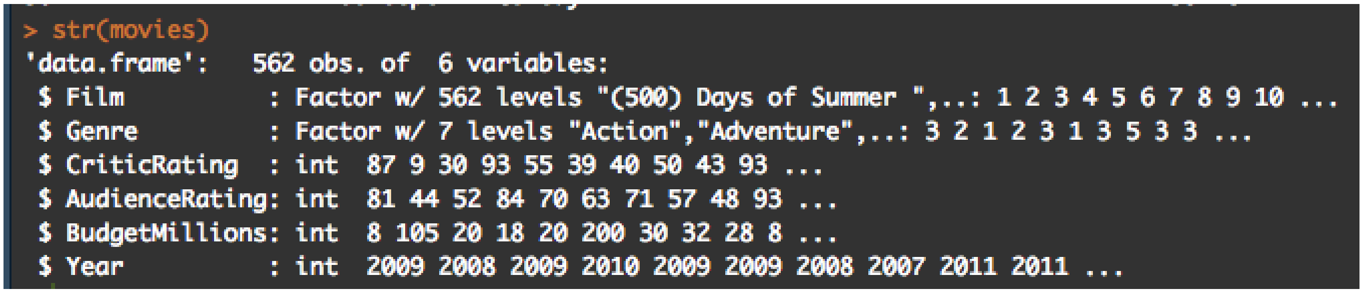
```
> head(movies)
                   Film     Genre Rotten.Tomatoes.Ratings.. Audience.Ratings.. Budget..million...
1 (500) Days of Summer     Comedy                        87                 81                  8
2          10,000 B.C. Adventure                          9                 44                105
3            12 Rounds     Action                        30                 52                 20
4            127 Hours Adventure                         93                 84                 18
5             17 Again     Comedy                        55                 70                 20
6                 2012     Action                        39                 63                200
  Year.of.release
1            2009
2            2008
3            2009
4            2010
5            2009
6            2009
```

Let’s rename “Rotten.Tomatoes.Ratings” to “Critic Ratings” and “Year.of.release” to just “Year”



```r
colnames(movies) <- c("Film","Genre","CriticRating","AudienceRating","BudgetMillions","Year")
```

Using the structure function in R reveals to us how R is interpreting the format of these variables and instances. Let’s take a look



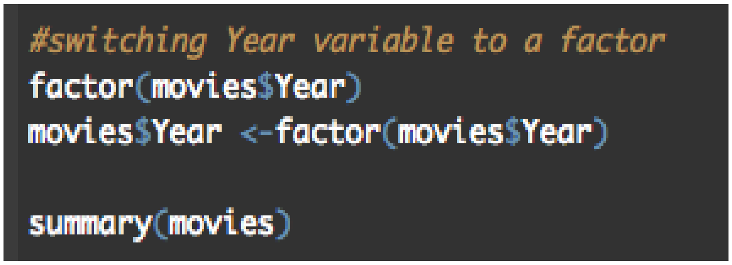
```
> str(movies)
'data.frame':	562 obs. of  6 variables:
 $ Film          : Factor w/ 562 levels "(500) Days of Summer ",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Genre         : Factor w/ 7 levels "Action","Adventure",..: 3 2 1 2 3 1 3 5 3 3 ...
 $ CriticRating  : int  87 9 30 93 55 39 40 50 43 93 ...
 $ AudienceRating: int  81 44 52 84 70 63 71 57 48 93 ...
 $ BudgetMillions: int  8 105 20 18 20 200 30 32 28 8 ...
 $ Year          : int  2009 2008 2009 2010 2009 2009 2008 2007 2011 2011 ...
```

We have 6 variables and 562 rows of data.

- Film is represented as a factor; you can see on the row of $ Film there are numbers from 1-10 etc. That is how R is interpreting the values of the Films, we don’t really mind this configuration.
- Genre is also a factor

- CriticRating is an integer
- AudienceRating is an integer
- BudgetMillions is an integer
- And Year is an integer

Let’s change the Year structure into a factor so it is easily represented.



```
#switching Year variable to a factor
factor(movies$Year)
movies$Year <-factor(movies$Year)

summary(movies)
```

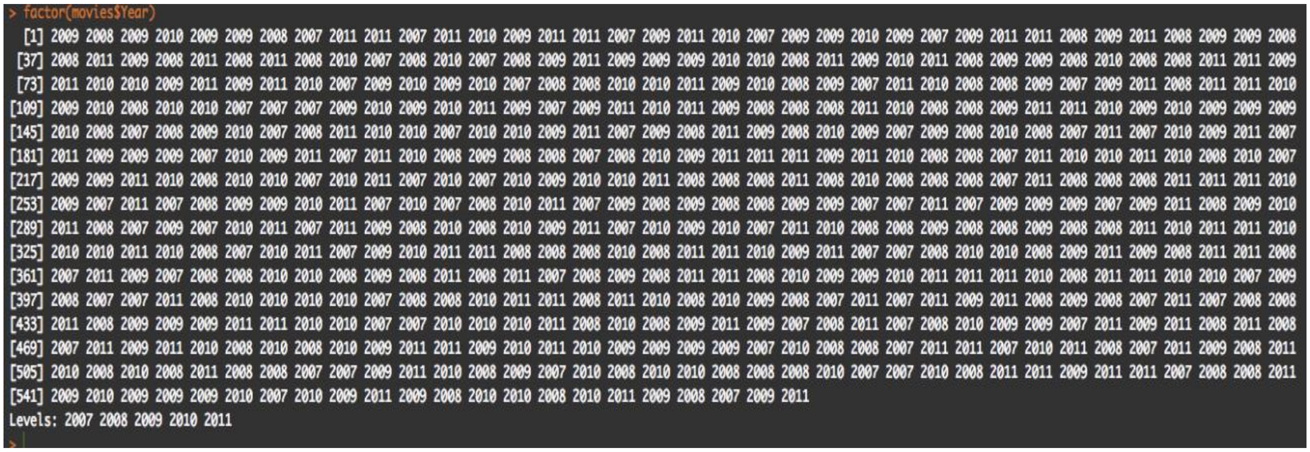
In this code, we tell R that we want to create the variable “Year” into a factor.

To create a factor we explicitly state it like this - “factor(---)”

Within the function we specify the data frame “movies” and the variable “Year”
To refer to the specific column within a data frame we state it with a dollar sign in between data frame and column.

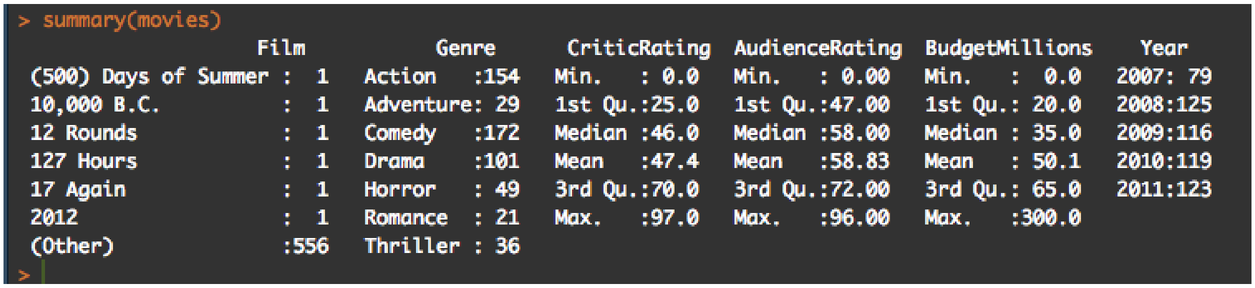
After that we have to explicitly again state the (data frame $ column) is the new factor(movies$Year)

When we do that, we have each instance



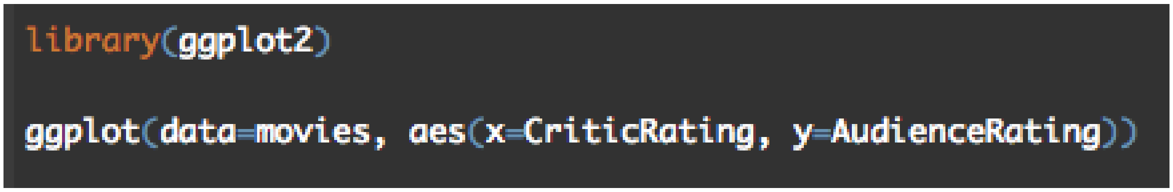
```
> factor(movies$Year)
  [1] 2009 2008 2009 2010 2009 2009 2008 2007 2011 2011 2007 2011 2010 2009 2011 2011 2007 2009 2011 2010 2007 2009 2009 2010 2009 2007 2009 2011 2011 2008 2009 2011 2008 2009 2009 2008
 [37] 2008 2011 2009 2008 2011 2008 2011 2008 2010 2007 2008 2010 2007 2008 2009 2011 2009 2009 2009 2010 2010 2008 2011 2009 2010 2011 2008 2009 2009 2008 2010 2008 2008 2011 2011 2009
 [73] 2011 2010 2010 2009 2011 2009 2011 2010 2007 2009 2010 2009 2010 2007 2008 2008 2010 2010 2011 2009 2010 2008 2009 2007 2011 2010 2008 2008 2009 2007 2009 2011 2008 2011 2011 2010
[109] 2009 2010 2008 2010 2010 2007 2007 2007 2009 2010 2009 2010 2011 2009 2007 2009 2011 2010 2011 2009 2008 2008 2008 2011 2010 2008 2008 2009 2011 2011 2010 2009 2010 2009 2009 2009
[145] 2010 2008 2007 2008 2009 2010 2007 2008 2011 2010 2010 2007 2010 2010 2009 2011 2007 2009 2008 2011 2009 2008 2010 2009 2007 2009 2008 2010 2008 2007 2011 2007 2010 2009 2011 2007
[181] 2011 2009 2009 2009 2007 2010 2009 2011 2007 2011 2010 2008 2009 2008 2008 2007 2008 2010 2009 2011 2011 2011 2009 2011 2010 2008 2008 2007 2011 2010 2010 2011 2010 2008 2010 2007
[217] 2009 2009 2011 2010 2008 2010 2010 2007 2010 2011 2007 2010 2007 2010 2009 2010 2010 2011 2008 2008 2008 2011 2008 2010 2008 2008 2008 2007 2011 2008 2008 2008 2011 2011 2011 2010
[253] 2009 2007 2011 2007 2008 2009 2009 2010 2011 2007 2010 2007 2008 2010 2011 2007 2009 2008 2009 2008 2008 2009 2009 2007 2007 2011 2007 2009 2009 2009 2007 2009 2011 2008 2009 2010
[289] 2011 2008 2007 2009 2007 2010 2011 2007 2011 2009 2008 2010 2008 2010 2009 2011 2007 2010 2009 2010 2007 2011 2010 2008 2008 2009 2008 2008 2009 2008 2008 2011 2010 2011 2011 2010
[325] 2010 2010 2011 2010 2008 2007 2010 2011 2007 2009 2010 2011 2011 2008 2008 2008 2010 2008 2011 2011 2010 2009 2011 2007 2007 2008 2010 2010 2008 2009 2011 2009 2008 2011 2011 2008
[361] 2007 2011 2009 2007 2008 2008 2010 2010 2008 2009 2008 2011 2008 2011 2007 2008 2009 2008 2011 2011 2008 2010 2009 2009 2010 2011 2011 2011 2010 2008 2011 2011 2010 2010 2007 2009
[397] 2008 2007 2007 2011 2008 2010 2010 2010 2010 2007 2008 2008 2010 2011 2011 2008 2011 2010 2008 2010 2009 2008 2007 2011 2007 2011 2009 2011 2008 2009 2008 2007 2011 2007 2008 2008
[433] 2011 2008 2009 2009 2009 2011 2011 2010 2010 2007 2007 2010 2010 2010 2011 2008 2010 2008 2009 2011 2009 2007 2008 2011 2007 2008 2010 2009 2009 2007 2011 2009 2011 2008 2011 2008
[469] 2007 2011 2009 2011 2010 2008 2010 2008 2010 2009 2011 2011 2009 2010 2011 2010 2009 2009 2009 2009 2007 2010 2008 2008 2007 2011 2011 2007 2010 2011 2008 2007 2011 2009 2008 2011
[505] 2010 2008 2010 2008 2011 2008 2008 2007 2007 2009 2011 2010 2008 2009 2007 2010 2008 2010 2010 2008 2008 2008 2010 2007 2007 2010 2008 2011 2011 2009 2011 2011 2007 2008 2008 2011
[541] 2009 2010 2009 2009 2009 2010 2007 2010 2009 2011 2009 2008 2010 2010 2008 2010 2011 2009 2008 2007 2009 2011
Levels: 2007 2008 2009 2010 2011
>
```

The summary of the new table “Year” is below



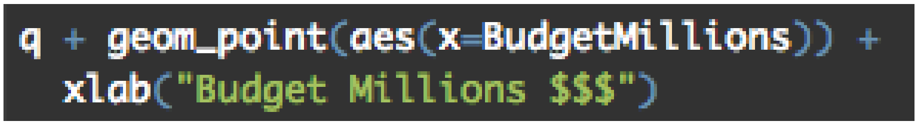
```
> summary(movies)
                        Film              Genre        CriticRating   AudienceRating   BudgetMillions      Year    
 (500) Days of Summer :  1   Action   :154   Min.   : 0.0   Min.   : 0.00   Min.   :  0.0   2007: 79  
 10,000 B.C.          :  1   Adventure: 29   1st Qu.:25.0   1st Qu.:47.00   1st Qu.: 20.0   2008:125  
 12 Rounds            :  1   Comedy   :172   Median :46.0   Median :58.00   Median : 35.0   2009:116  
 127 Hours            :  1   Drama    :101   Mean   :47.4   Mean   :58.83   Mean   : 50.1   2010:119  
 17 Again             :  1   Horror   : 49   3rd Qu.:70.0   3rd Qu.:72.00   3rd Qu.: 65.0   2011:123  
 2012                 :  1   Romance  : 21   Max.   :97.0   Max.   :96.00   Max.   :300.0             
 (Other)              :556   Thriller : 36                                                             
> 
```

Next we create the parameters

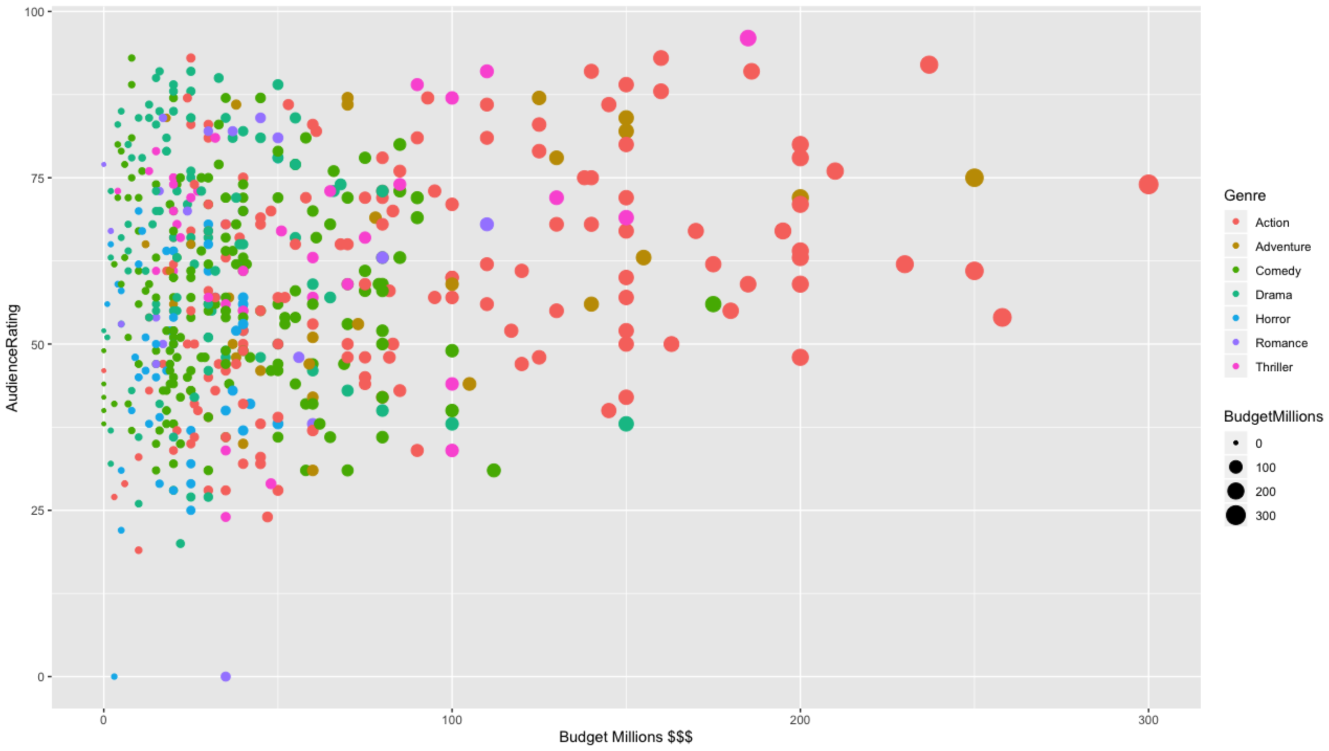


```
library(ggplot2)

ggplot(data=movies, aes(x=CriticRating, y=AudienceRating))
```

The first chart is a simple scatter plot that showcases CriticRating and AudienceRating
With the size of the points for “BudgetMillions” and the different colours for “Genre”

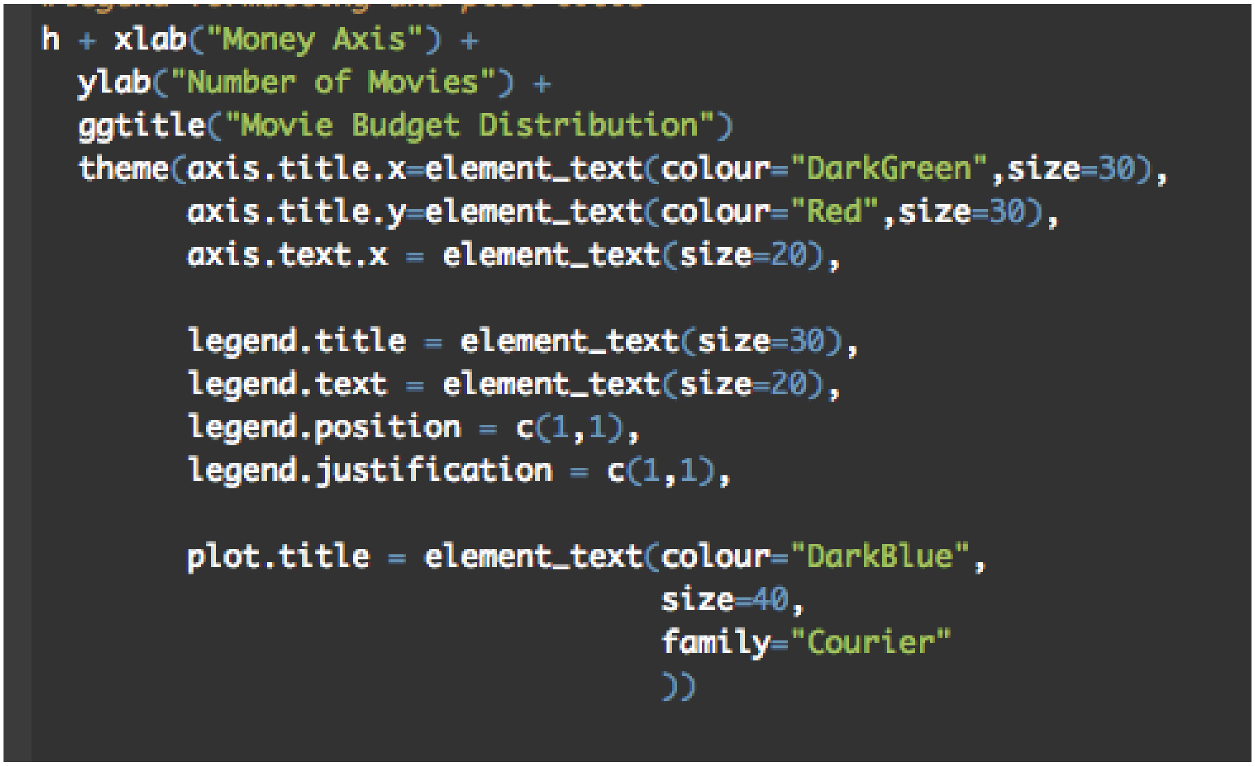


```
q + geom_point(aes(x=BudgetMillions)) +
  xlab("Budget Millions $$$")
```

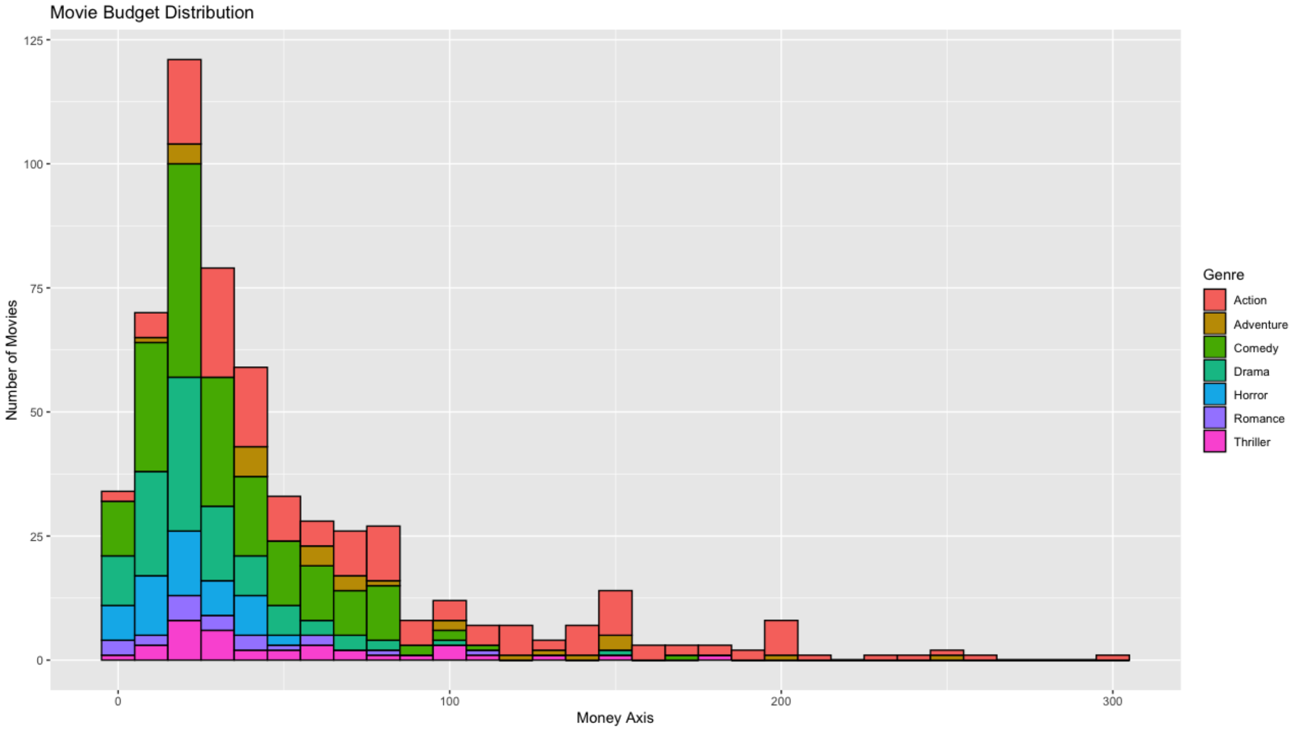


From this colorful, intuitive scatter plot we can tell that audience ratings tend to stay past a score of 50 after the 150 million budget mark, but with few films of that range of money the majority of points are in the 0-100 budget million range. Below 150 million budget mark the range of audience ratings ranges from below 25 to a sky-rocketing 90-95.

Next, I will display a histogram of Movie Budget Distribution the y axis is the number of movies the x axis is the money axis. The colour differences are the different genres

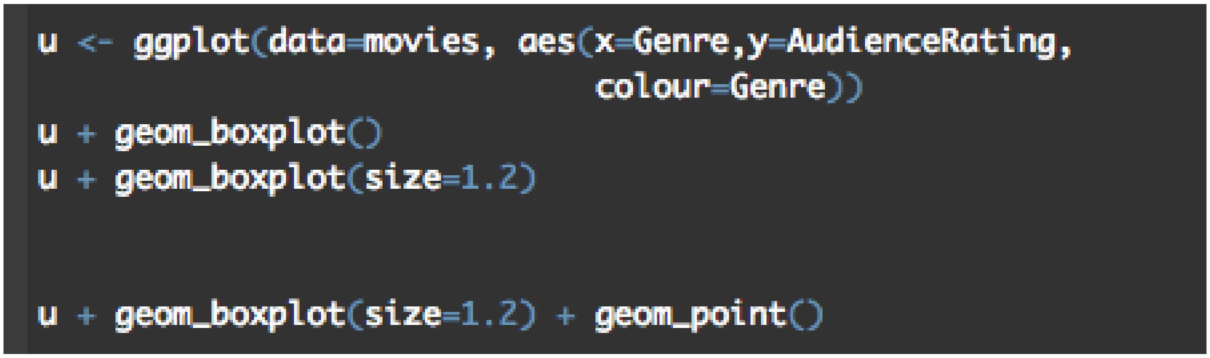


```
h + xlab("Money Axis") +
  ylab("Number of Movies") +
  ggtitle("Movie Budget Distribution")
  theme(axis.title.x=element_text(colour="DarkGreen",size=30),
        axis.title.y=element_text(colour="Red",size=30),
        axis.text.x = element_text(size=20),

        legend.title = element_text(size=30),
        legend.text = element_text(size=20),
        legend.position = c(1,1),
        legend.justification = c(1,1),

        plot.title = element_text(colour="DarkBlue",
                                  size=40,
                                  family="Courier"
                                  ))
```



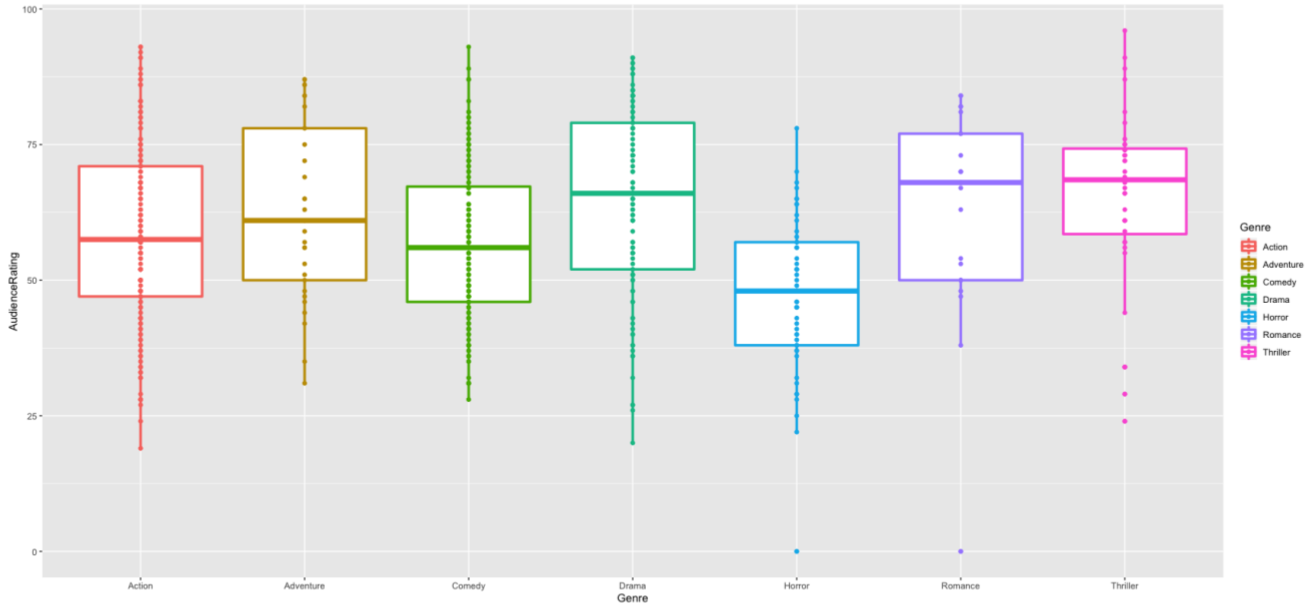
The peak of the number of movies that have similar money budget expenditures lies below 50 million dollars. With few 300 million dollar movies this chart directs our attention at the not-so expensive budgeted movies.
The next interactive graph I will showcase is a box and whiskers plot.
With the Genre as the X axis, the AudienceRating as the Y axis, and the difference in colours as Genre as well.

Lets start the code as below



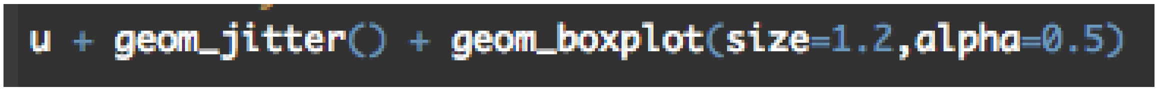
```
u <- ggplot(data=movies, aes(x=Genre,y=AudienceRating,
                              colour=Genre))
u + geom_boxplot()
u + geom_boxplot(size=1.2)


u + geom_boxplot(size=1.2) + geom_point()
```

Look at how engaging and well-documented the different genres are in colour and side by side. The genre Thriller has many points that are outliers well above the 75 Audience rating mark. But the genre that seems to have the highest average of Audience Rating seems to be between Romance and Thriller. The genre that has well-proportioned ratings is Drama.



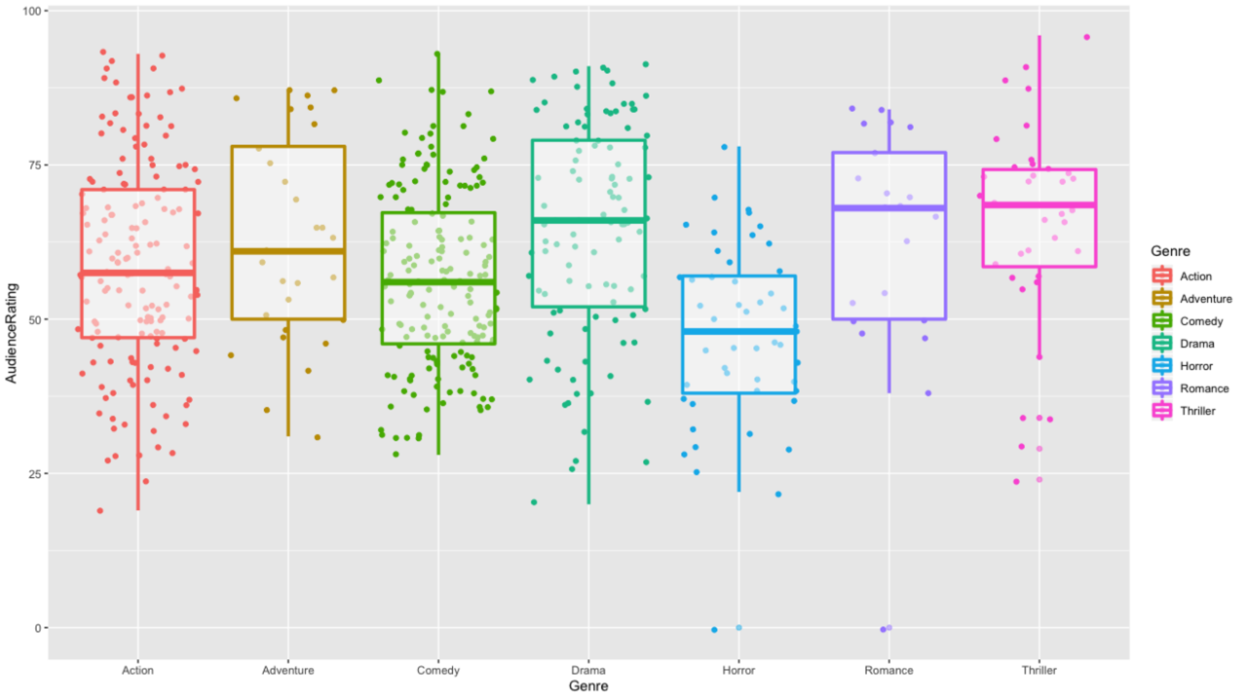
That last line of code doesn’t really help us in visualizing how the data is segmented, it just seems normalized. It isn’t.

Let’s modify the code a tad.



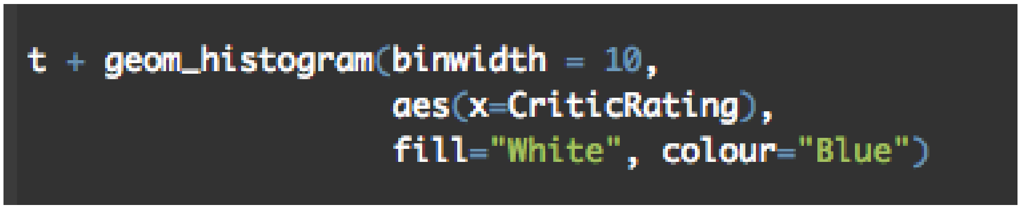
```
u + geom_jitter() + geom_boxplot(size=1.2,alpha=0.5)
```

Using the geom_jitter () function really helps us view the scattered data alongside the boxplot. Look below for the newly improved graph.

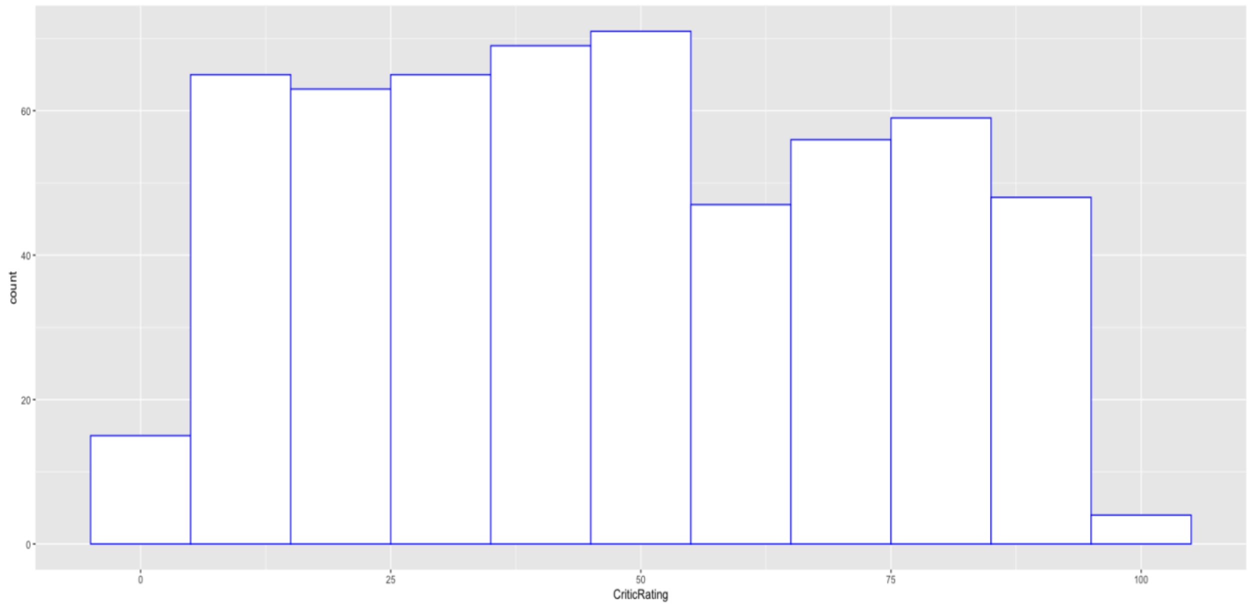


So now with the new function added on top of the box plot we can see that the Thriller category has very few outliers above the 75 Audience ratings section. It actually shows that Thrillers are not too well critiqued, just look at how many points there are in Drama, Comedy, and Action. These 3 genres are all scattered within the boxplot. The action genre seems to have the most points within the box average.

Perhaps I would like to view the range of Critic Ratings and their counts. The code is below

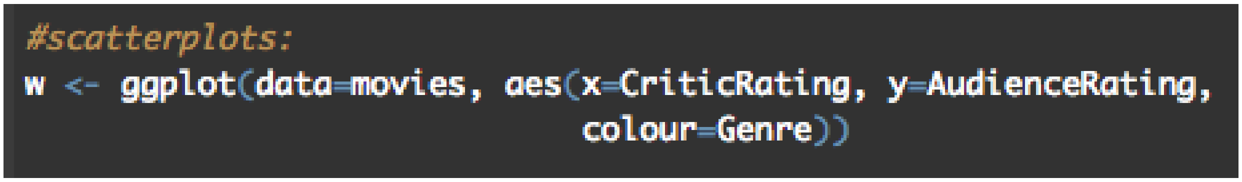


```
t + geom_histogram(binwidth = 10,
                   aes(x=CriticRating),
                   fill="White", colour="Blue")
```

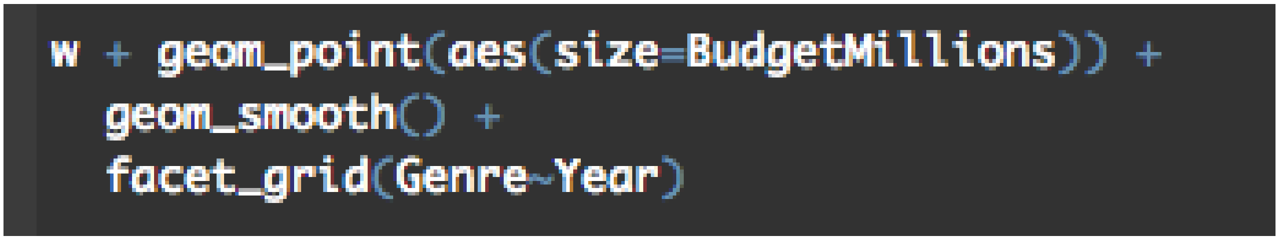


5 labels of the Critic Ratings are shown below ranging from 0 – 100. Very few Critic Ratings are in the 100 percentiles. The majority are between 25 and 50 ratings.

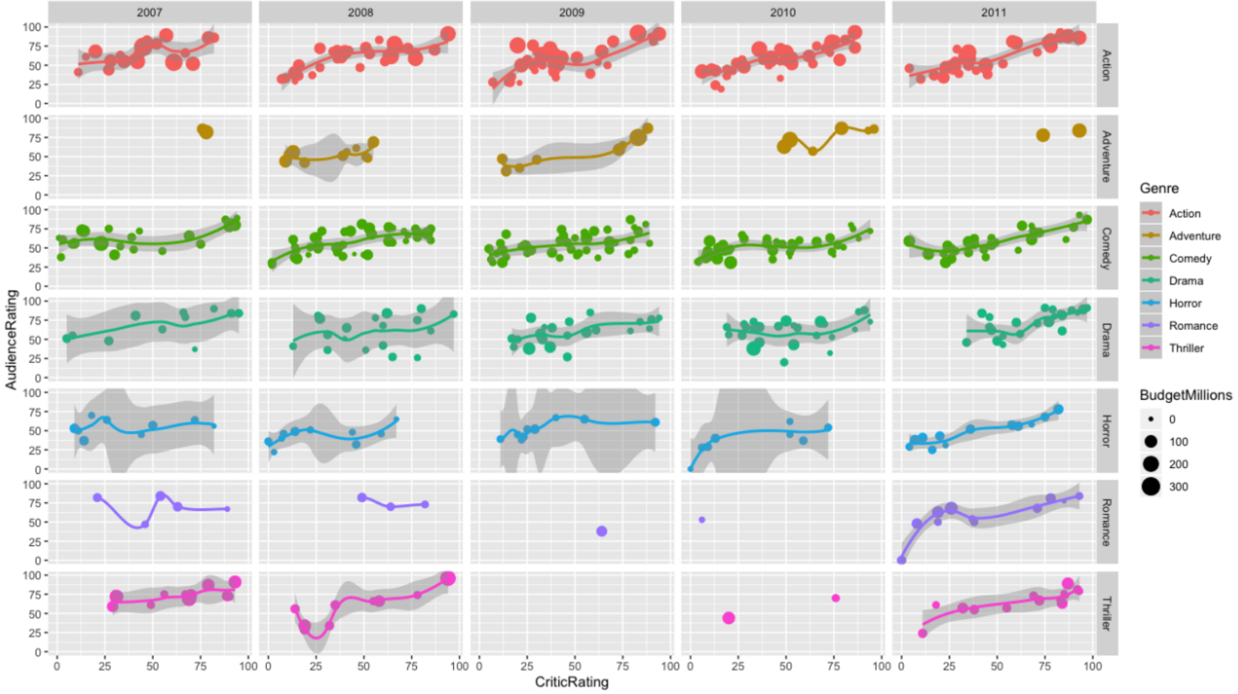
Lastly, let’s see a scatterplot of the different Genres, Critic Ratings, and Audience Ratings, as well as the Budgets Millions over the years.



```
#scatterplots:
w <- ggplot(data=movies, aes(x=CriticRating, y=AudienceRating,
                              colour=Genre))
```



```
w + geom_point(aes(size=BudgetMillions)) +
  geom_smooth() +
  facet_grid(Genre~Year)
```



The scatterplot displayed shows:

A surge of action movies that are increasingly more and more popular as the years go by. Adventure genres have their spots and a trend in 2009. Comedy seems to also have a growing reputation as the years go by they have a positive trend. Drama films are also slowly following suit but in 2011 the trend is abrupt. Horror films are all over the place until a steady trend in 2011.

In conclusion, I hope the graphs and charts that I displayed showed some insights as to how genres have been critiqued over time. Which genres are more popular, less popular, and the budgets that these films had.

It is always a pleasure to view data in different charts that give a different story about the transformation and the direction the data is headed.