Data Visualization

Nachiket Hinge

11/8/2021

```
knitr::opts_chunk$set(fig.width=12, fig.height=8)
```

```
movie.ratings <- read.csv("P2-Movie-Ratings.csv")
head(movie.ratings)
```

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

```
colnames(movie.ratings) <- c("Film", "Genre", "CriticRating", "AudienceRating", "BudgetMillions", "Year
```

```
head(movie.ratings)
```

```
##                    Film     Genre CriticRating AudienceRating BudgetMillions
## 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
## 1 2009
## 2 2008
## 3 2009
## 4 2010
## 5 2009
## 6 2009
```

```
tail(movie.ratings)
```

```
##                          Film    Genre CriticRating AudienceRating
## 557             Your Highness   Comedy           26             36
## 558           Youth in Revolt   Comedy           68             52
```

```
## 559 Zack and Miri Make a Porno  Romance           64             70
## 560                     Zodiac Thriller           89             73
## 561                Zombieland    Action           90             87
## 562                  Zookeeper   Comedy           14             42
##     BudgetMillions Year
## 557             50 2011
## 558             18 2009
## 559             24 2008
## 560             65 2007
## 561             24 2009
## 562             80 2011
```

```
str(movie.ratings)
```

```
## 'data.frame':    562 obs. of  6 variables:
##  $ Film          : chr  "(500) Days of Summer " "10,000 B.C." "12 Rounds " "127 Hours" ...
##  $ Genre         : chr  "Comedy" "Adventure" "Action" "Adventure" ...
##  $ 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 ...
```

```
summary(movie.ratings)
```

```
##      Film              Genre            CriticRating  AudienceRating
##  Length:562         Length:562         Min.   : 0.0   Min.   : 0.00
##  Class :character   Class :character   1st Qu.:25.0   1st Qu.:47.00
##  Mode  :character   Mode  :character   Median :46.0   Median :58.00
##                                        Mean   :47.4   Mean   :58.83
##                                        3rd Qu.:70.0   3rd Qu.:72.00
##                                        Max.   :97.0   Max.   :96.00
##  BudgetMillions       Year
##  Min.   :  0.0   Min.   :2007
##  1st Qu.: 20.0   1st Qu.:2008
##  Median : 35.0   Median :2009
##  Mean   : 50.1   Mean   :2009
##  3rd Qu.: 65.0   3rd Qu.:2010
##  Max.   :300.0   Max.   :2011
```

```
factor(movie.ratings$Year)
```

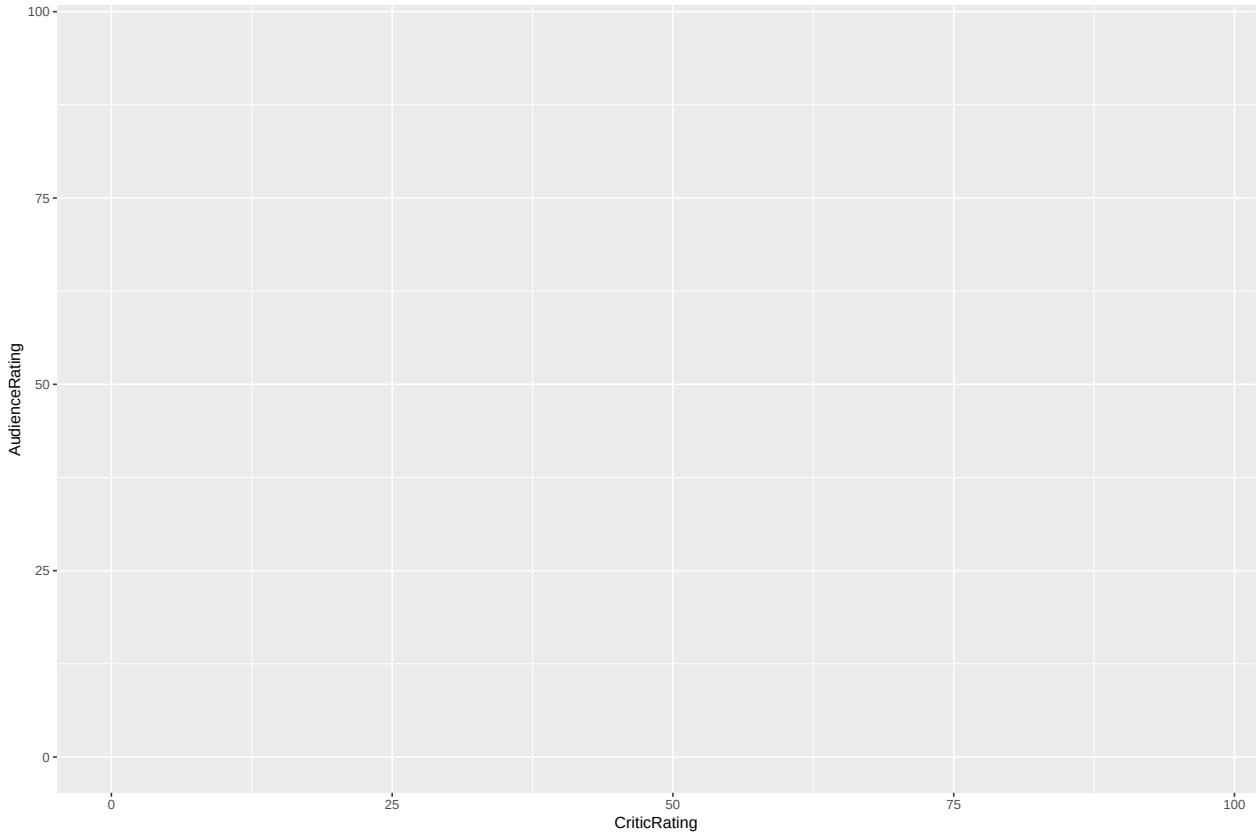
```
##   [1] 2009 2008 2009 2010 2009 2009 2008 2007 2011 2011 2007 2011 2010 2009 2011
##  [16] 2011 2007 2009 2011 2010 2007 2009 2009 2010 2009 2007 2009 2011 2011 2008
##  [31] 2009 2011 2008 2009 2009 2008 2008 2011 2009 2008 2011 2008 2011 2008 2010
##  [46] 2007 2008 2010 2007 2008 2009 2011 2009 2009 2009 2010 2010 2008 2011 2009
##  [61] 2010 2011 2008 2009 2009 2008 2010 2008 2008 2011 2011 2009 2011 2010 2010
##  [76] 2009 2011 2009 2011 2010 2007 2009 2010 2009 2010 2007 2008 2008 2010 2010
##  [91] 2011 2009 2010 2008 2009 2007 2011 2010 2008 2008 2009 2007 2009 2011 2008
## [106] 2011 2011 2010 2009 2010 2008 2010 2010 2007 2007 2007 2009 2010 2009 2010
## [121] 2011 2009 2007 2009 2011 2010 2011 2009 2008 2008 2008 2011 2010 2008 2008
## [136] 2009 2011 2011 2010 2009 2010 2009 2009 2009 2010 2008 2007 2008 2009 2010
## [151] 2007 2008 2011 2010 2010 2007 2010 2010 2009 2011 2007 2009 2008 2011 2009
## [166] 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
## [196] 2007 2008 2010 2009 2011 2011 2011 2009 2011 2010 2008 2008 2007 2011 2010
## [211] 2010 2011 2010 2008 2010 2007 2009 2009 2011 2010 2008 2010 2010 2007 2010
```

```
## [226] 2011 2007 2010 2007 2010 2009 2010 2010 2011 2008 2008 2008 2011 2008 2010
## [241] 2008 2008 2008 2007 2011 2008 2008 2008 2011 2011 2011 2010 2009 2007 2011
## [256] 2007 2008 2009 2009 2010 2011 2007 2010 2007 2008 2010 2011 2007 2009 2008
## [271] 2009 2008 2008 2009 2009 2007 2007 2011 2007 2009 2009 2009 2007 2009 2011
## [286] 2008 2009 2010 2011 2008 2007 2009 2007 2010 2011 2007 2011 2009 2008 2010
## [301] 2008 2010 2009 2011 2007 2010 2009 2010 2007 2011 2010 2008 2008 2009 2008
## [316] 2008 2009 2008 2008 2011 2010 2011 2011 2010 2010 2010 2011 2010 2008 2007
## [331] 2010 2011 2007 2009 2010 2011 2011 2008 2008 2008 2010 2008 2011 2011 2010
## [346] 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
## [376] 2008 2009 2008 2011 2011 2008 2010 2009 2009 2010 2011 2011 2011 2010 2008
## [391] 2011 2011 2010 2010 2007 2009 2008 2007 2007 2011 2008 2010 2010 2010 2010
## [406] 2007 2008 2008 2010 2011 2011 2008 2011 2010 2008 2010 2009 2008 2007 2011
## [421] 2007 2011 2009 2011 2008 2009 2008 2007 2011 2007 2008 2008 2011 2008 2009
## [436] 2009 2009 2011 2011 2010 2010 2007 2007 2010 2010 2010 2011 2008 2010 2008
## [451] 2009 2011 2009 2007 2008 2011 2007 2008 2010 2009 2009 2007 2011 2009 2011
## [466] 2008 2011 2008 2007 2011 2009 2011 2010 2008 2010 2008 2010 2009 2011 2011
## [481] 2009 2010 2011 2010 2009 2009 2009 2009 2007 2010 2008 2008 2007 2011 2011
## [496] 2007 2010 2011 2008 2007 2011 2009 2008 2011 2010 2008 2010 2008 2011 2008
## [511] 2008 2007 2007 2009 2011 2010 2008 2009 2007 2010 2008 2010 2010 2008 2008
## [526] 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
## [556] 2010 2011 2009 2008 2007 2009 2011
## Levels: 2007 2008 2009 2010 2011
```

```
movie.ratings$Year <- factor(movie.ratings$Year)
```

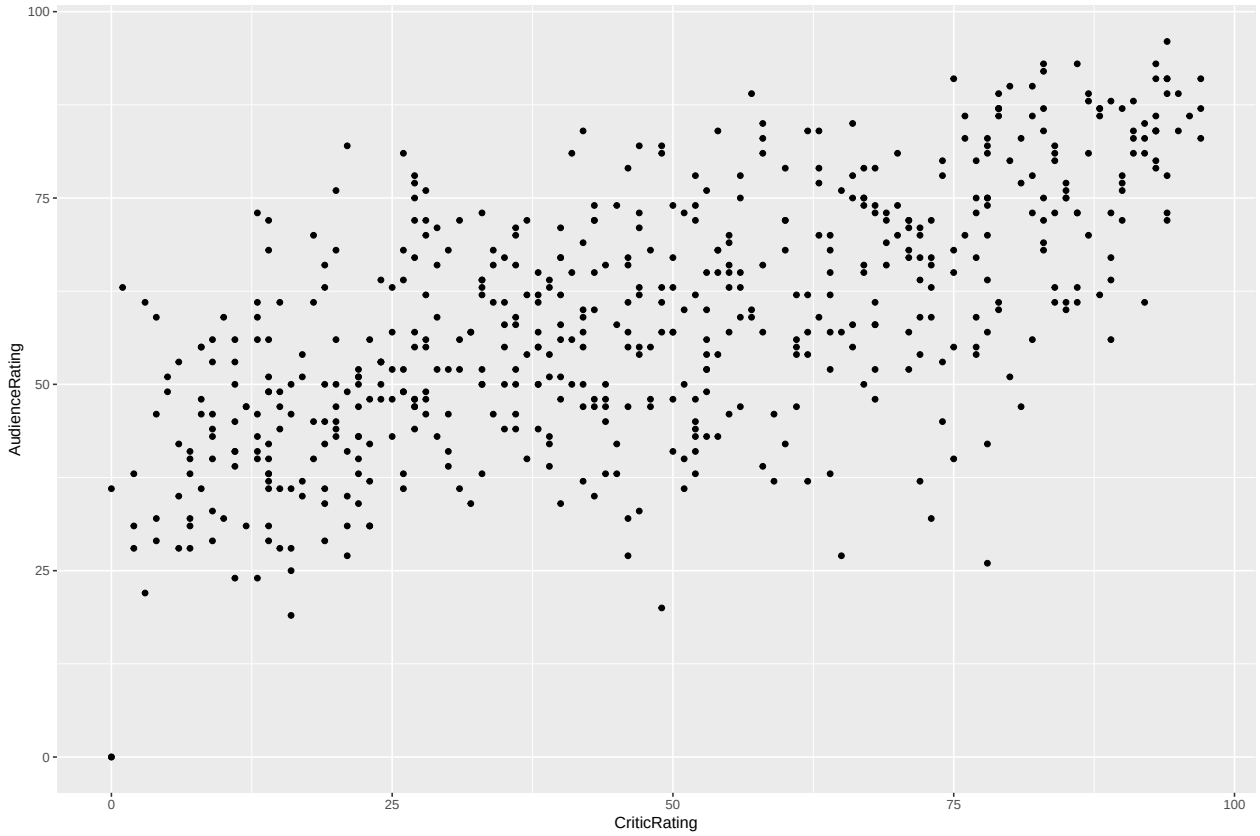
#Aesthetics

```
library(ggplot2)
ggplot(data = movie.ratings, aes(x = CriticRating, y = AudienceRating))
```



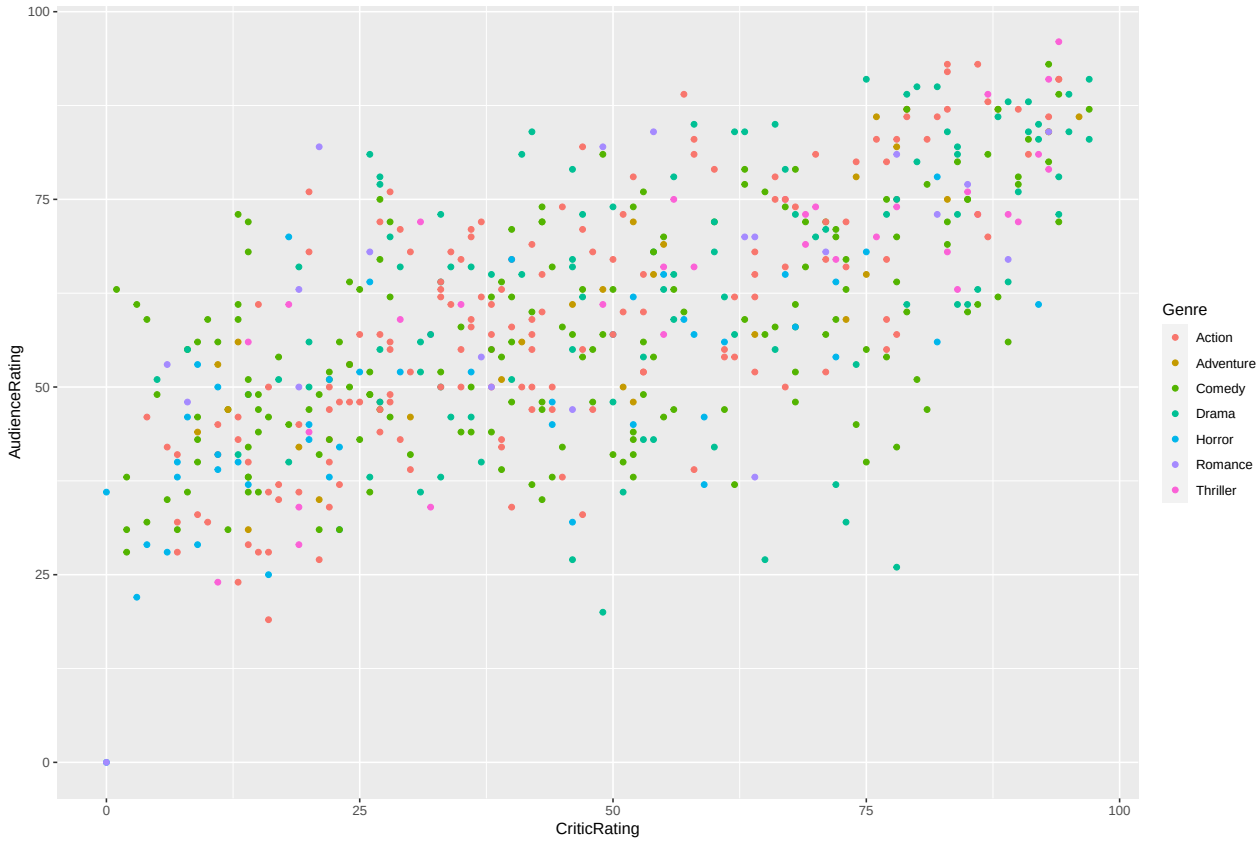
#Add Geometry

```
ggplot(data = movie.ratings, aes(x = CriticRating, y = AudienceRating)) + geom_point()
```



#Add Colors

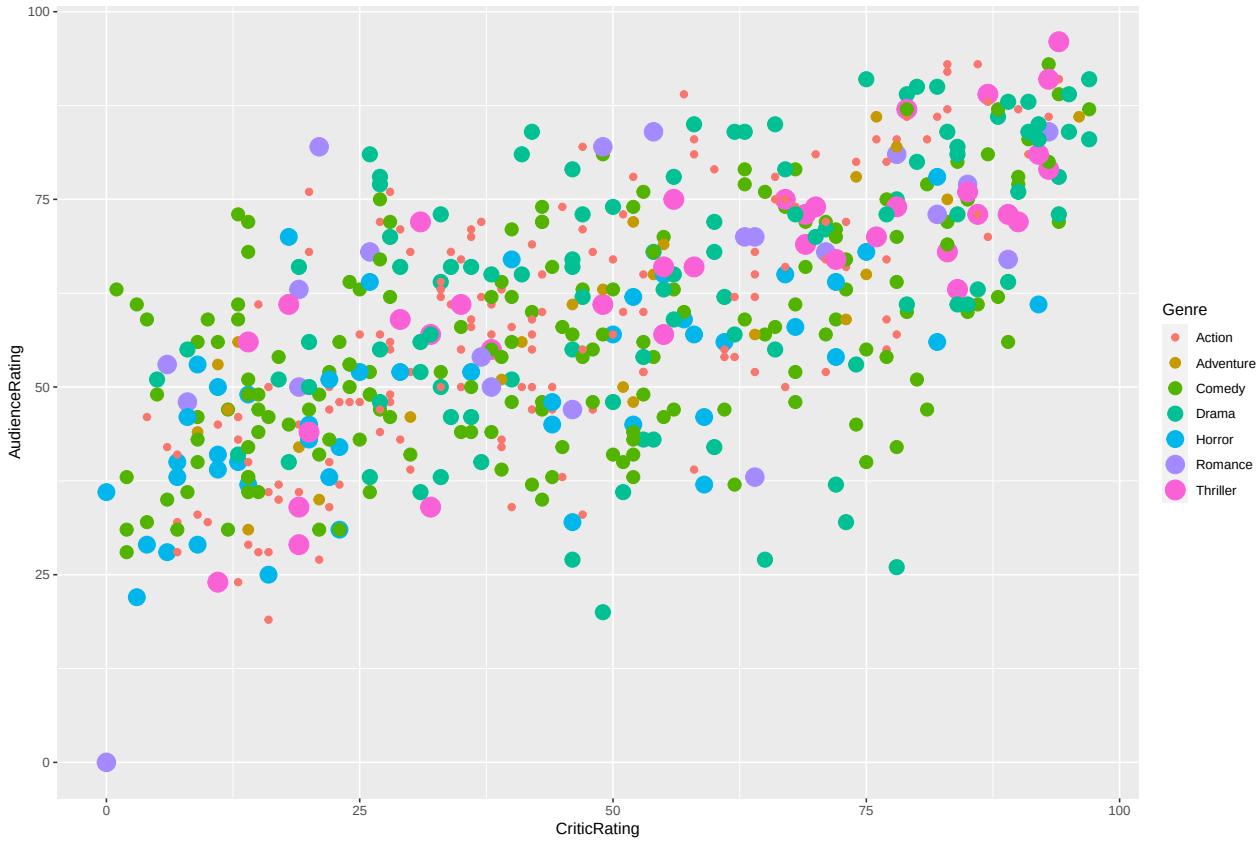
```
ggplot(data = movie.ratings, aes(x = CriticRating, y = AudienceRating,
                                 color=Genre)) + geom_point()
```



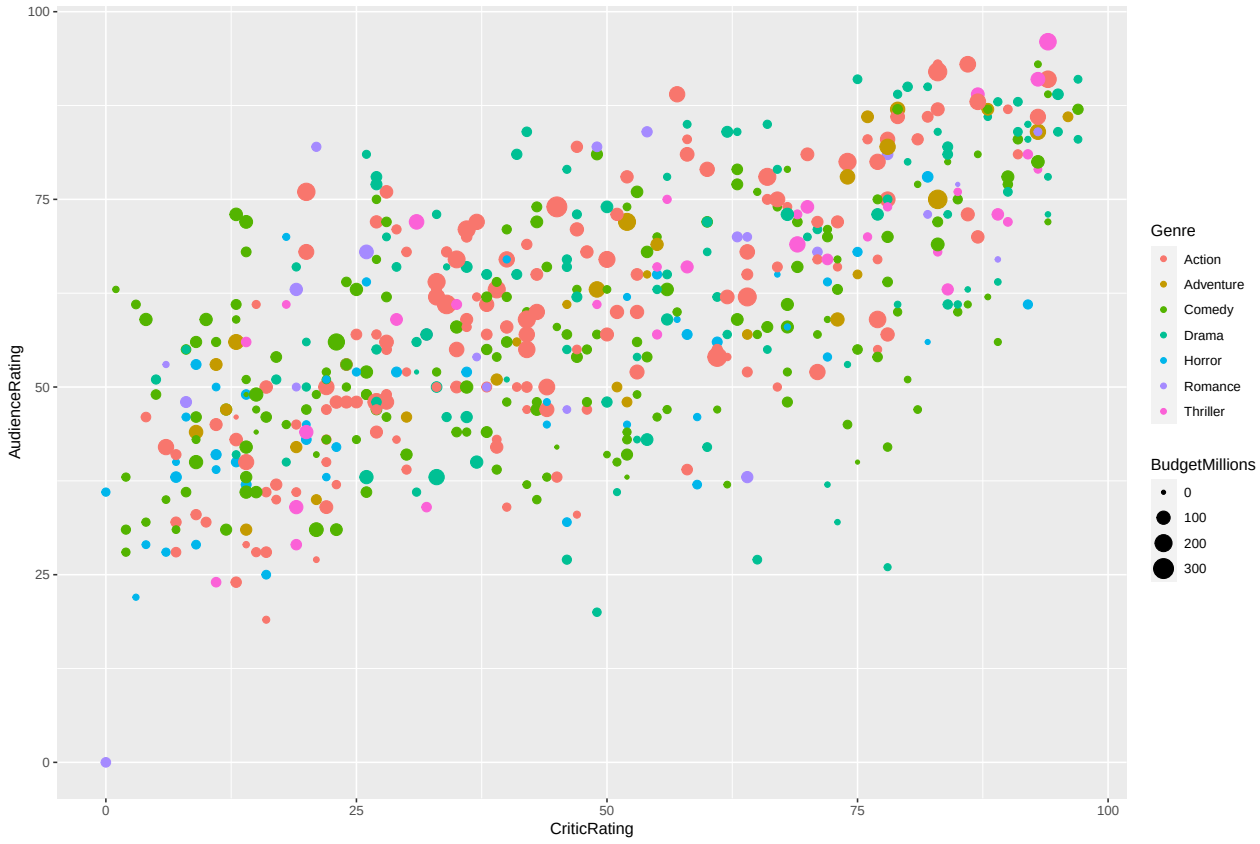
#Add Size

```
ggplot(data = movie.ratings, aes(x = CriticRating, y = AudienceRating,
                                 color=Genre, size=Genre)) + geom_point()
```

```
## Warning: Using size for a discrete variable is not advised.
```



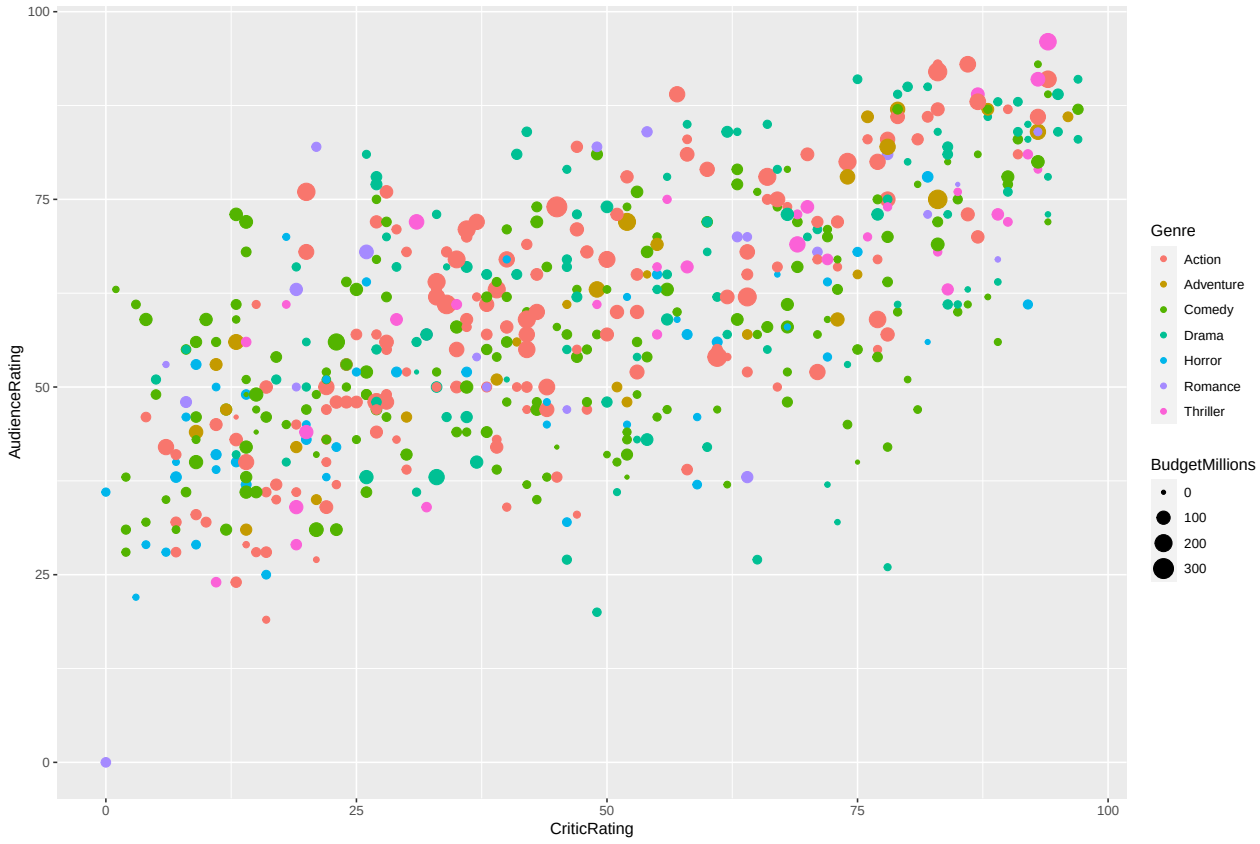

```
ggplot(data = movie.ratings, aes(x = CriticRating, y = AudienceRating,
                                 color=Genre, size=BudgetMillions)) + geom_point()
```



#Plotting with Layers

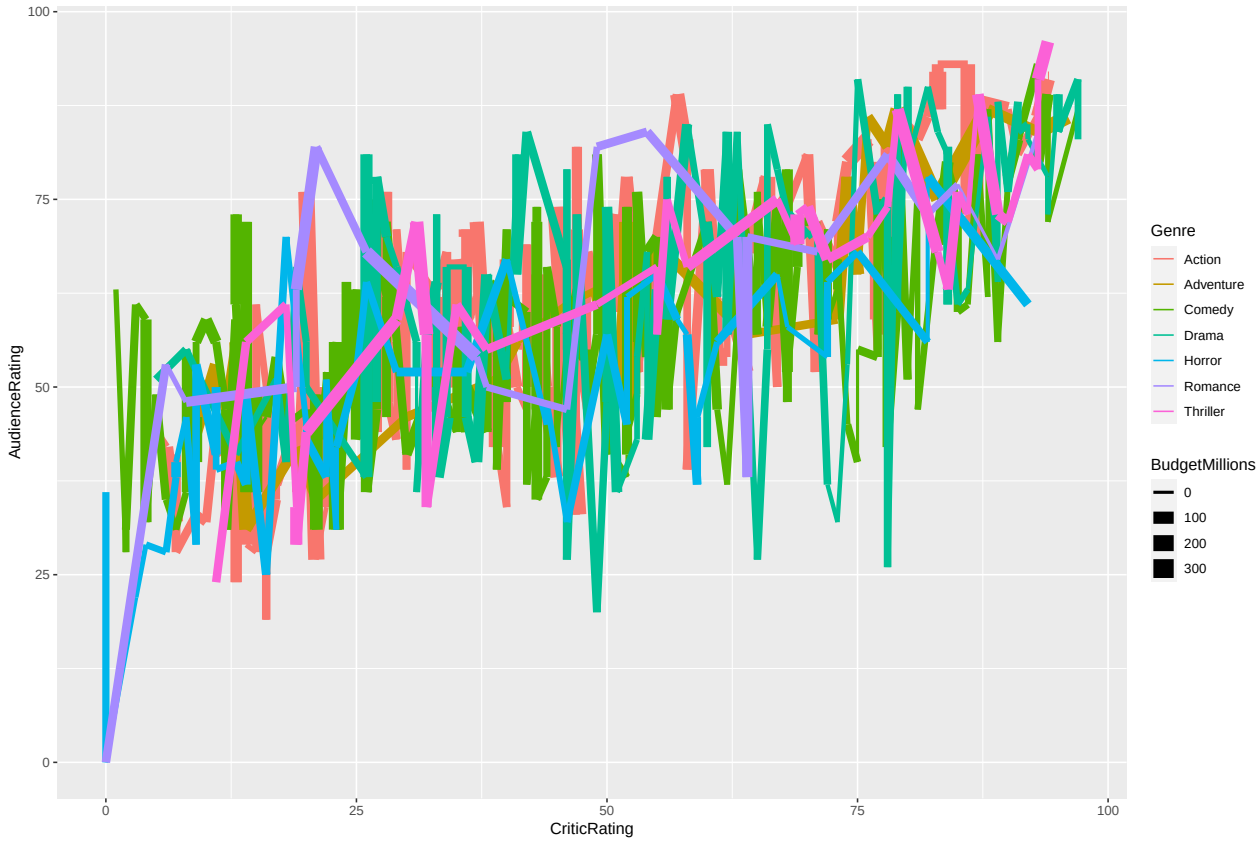
```
object <- ggplot(data = movie.ratings, aes(x = CriticRating, y = AudienceRating,
                                           color=Genre, size=BudgetMillions))
```

```
object + geom_point()
```



#Lines

```
object + geom_line()
```



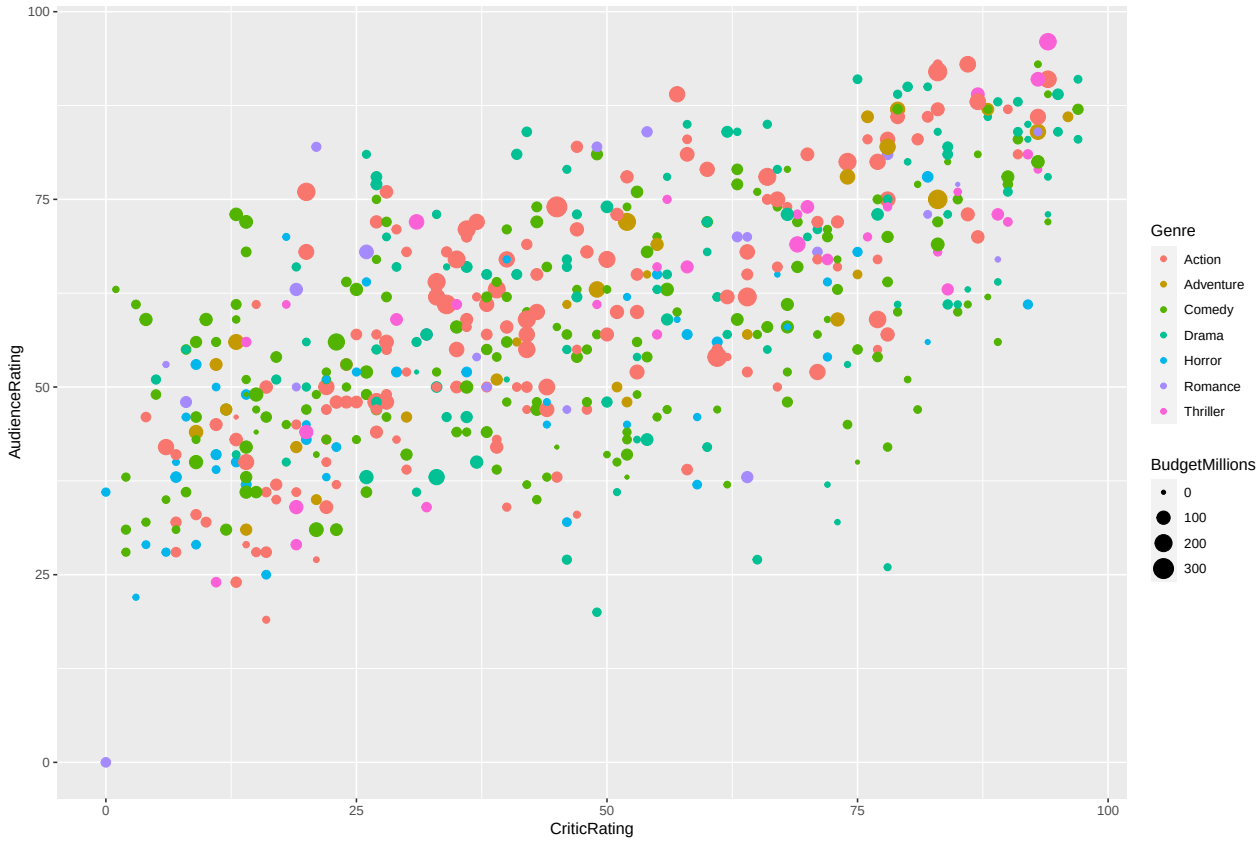
#Multiple Layers

```
object + geom_tile()
```

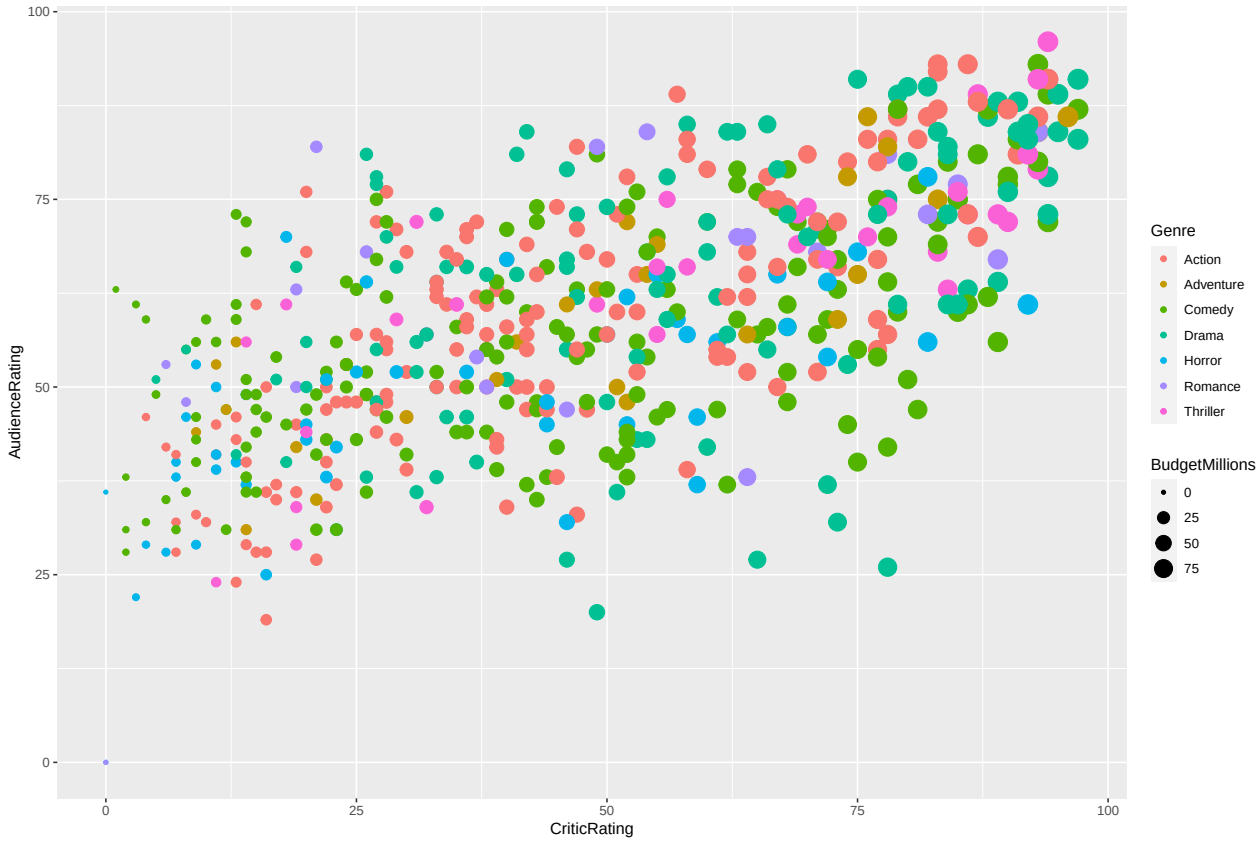
#Overriding Aesthetics

```
q <- ggplot(data = movie.ratings, aes(x=CriticRating, y=AudienceRating, color=Genre, size=BudgetMillions
```

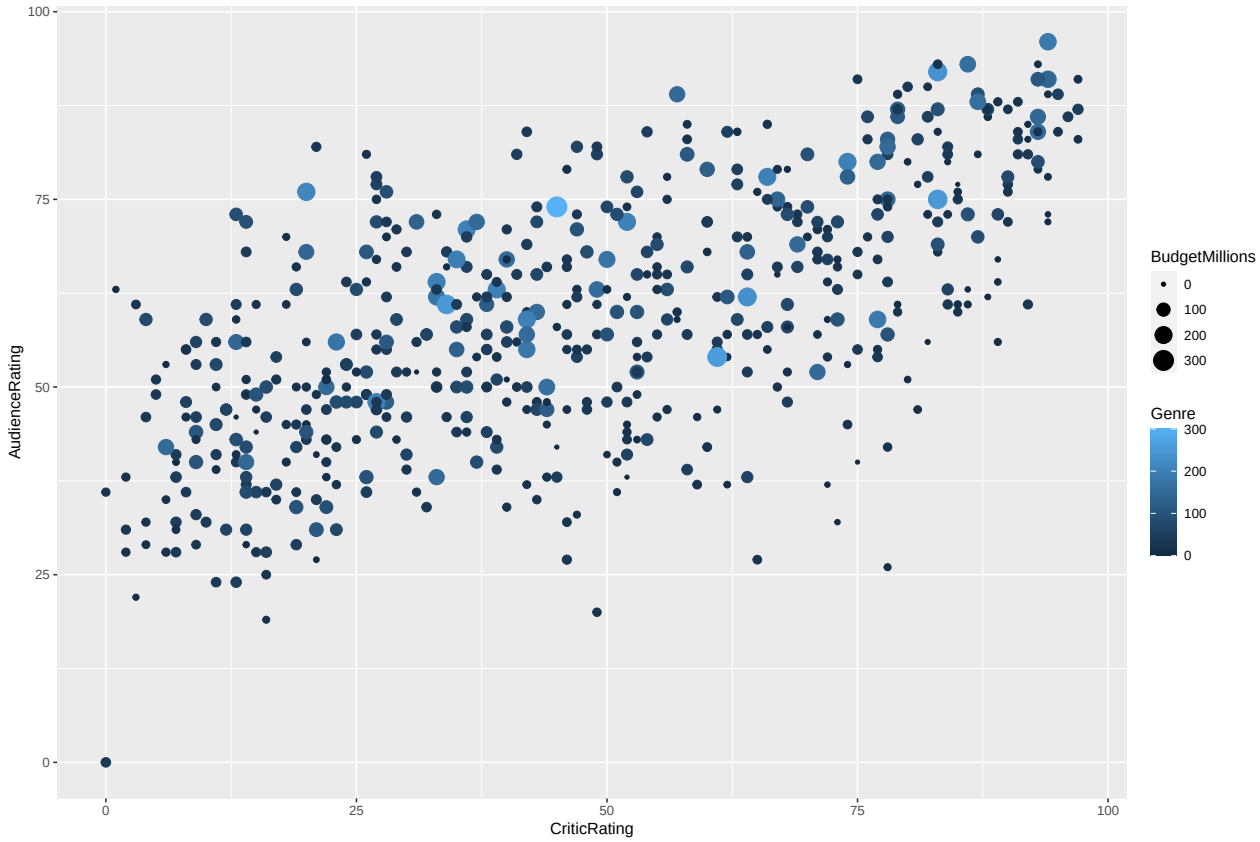
```
q + geom_point()
```



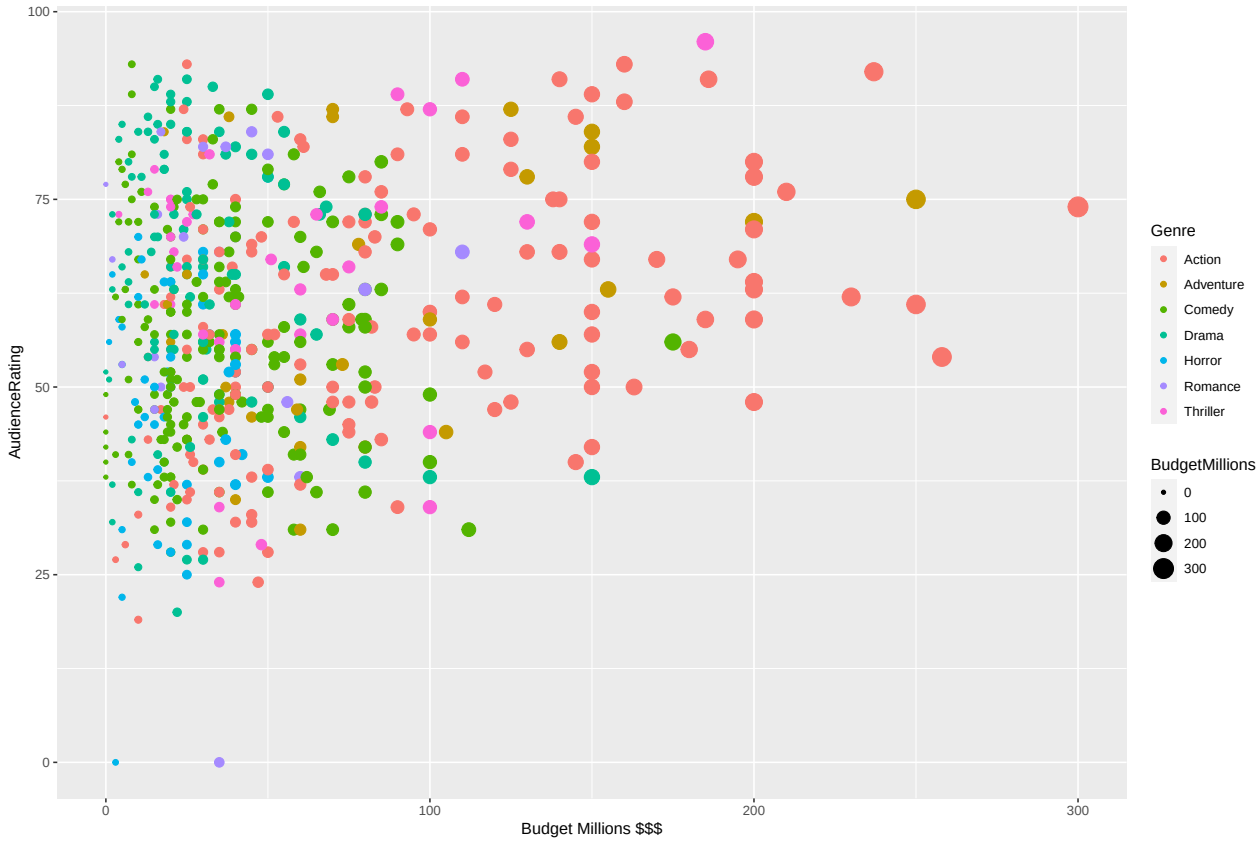

```
q + geom_point(aes(size=CriticRating))
```




```
q + geom_point(aes(color=BudgetMillions))
```



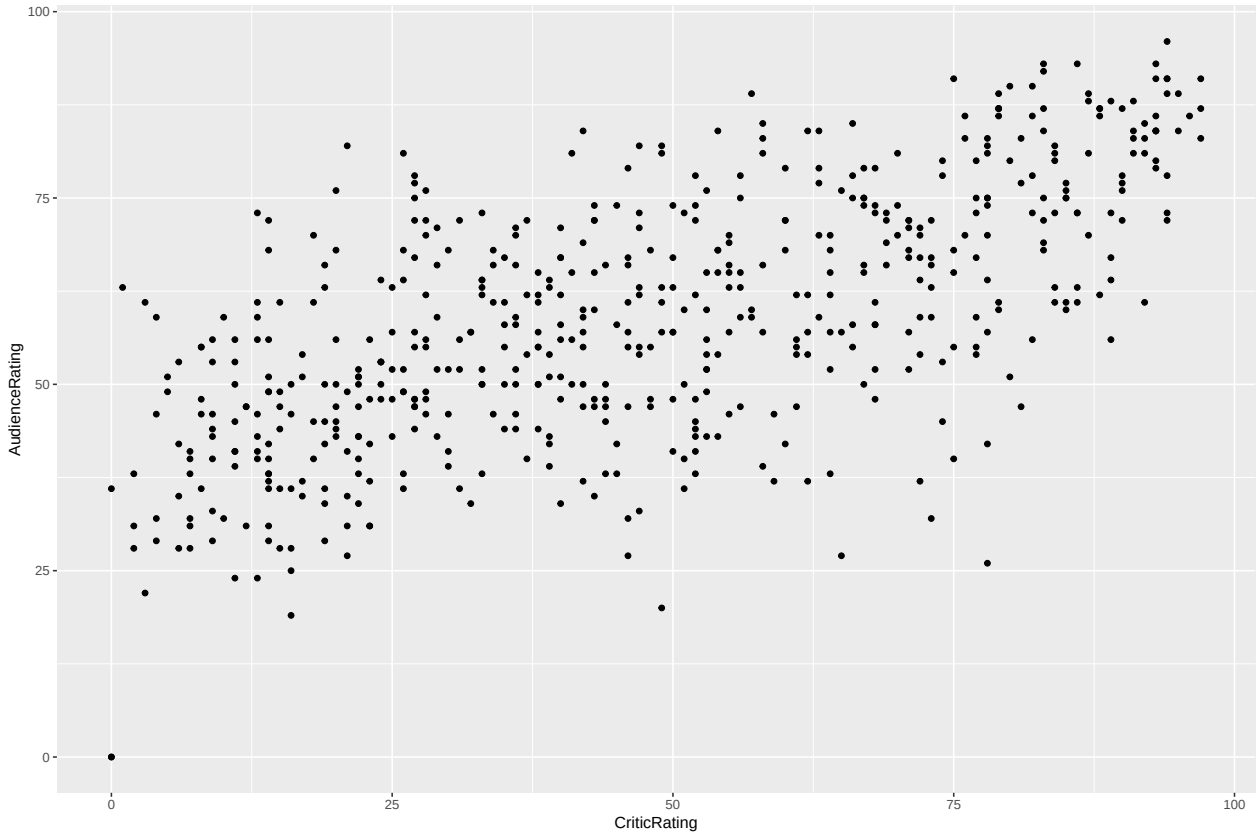

```
q + geom_point(aes(x=BudgetMillions)) + xlab("Budget Millions $$$")
```




```
q + geom_line(size=1) + geom_point()
```

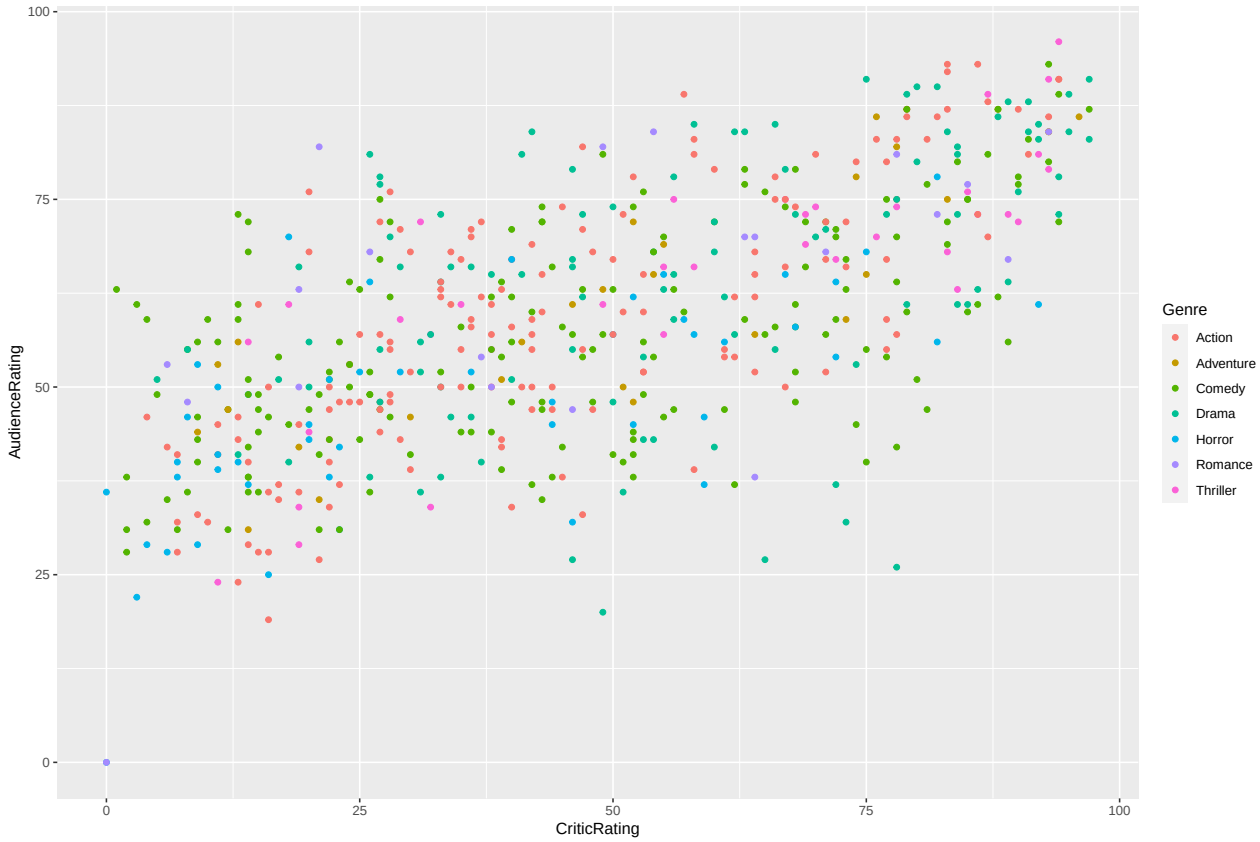
#Mapping vs Setting

```
r <- ggplot(data=movie.ratings, aes(x=CriticRating, y=AudienceRating))
r + geom_point()
```



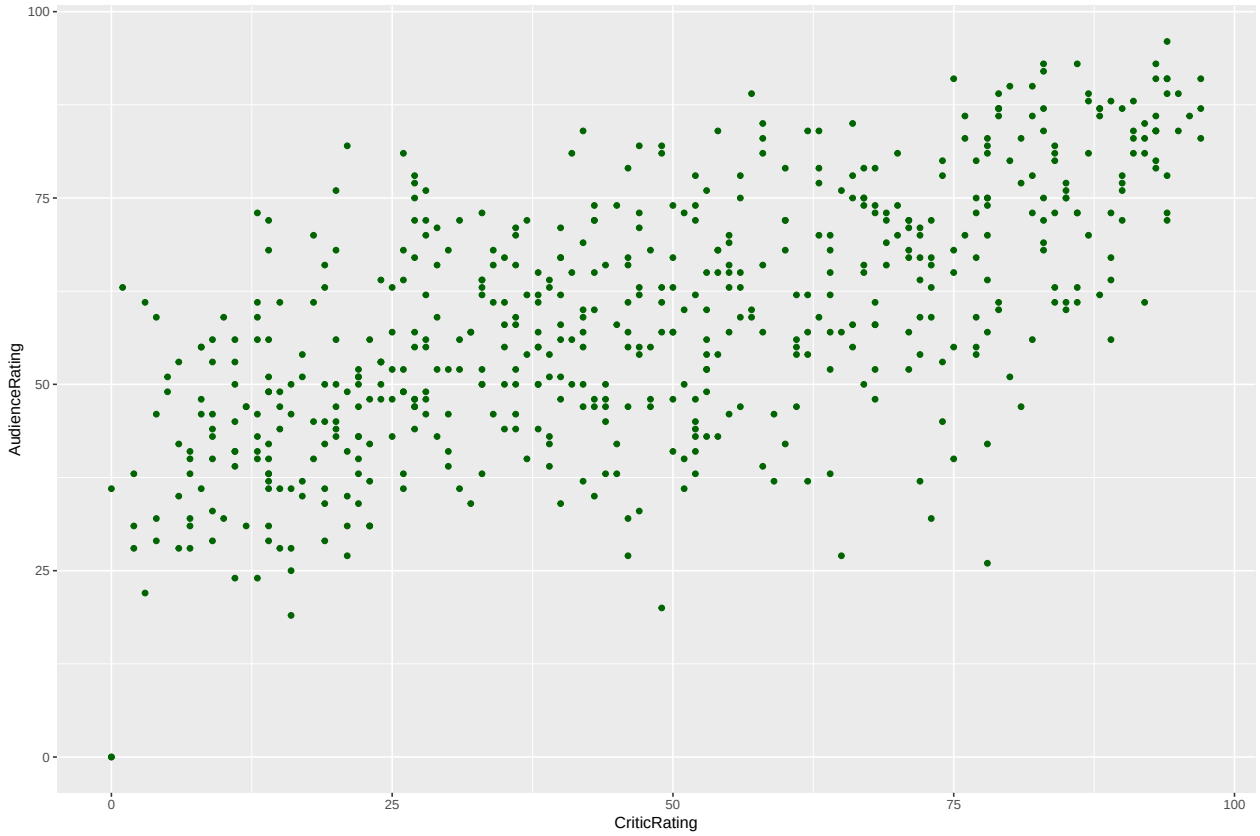
#Add Color by Mapping

```
r + geom_point(aes(color=Genre))
```



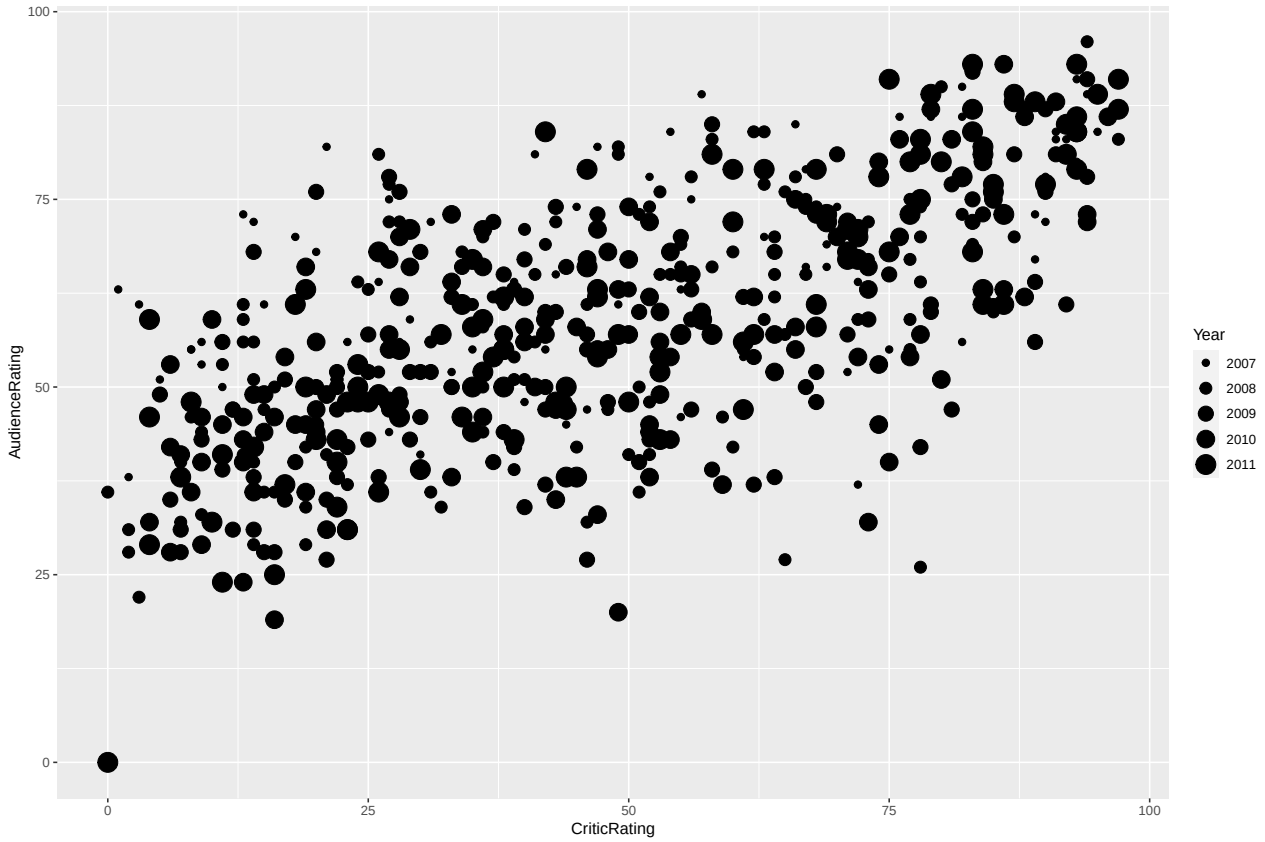
#Add Color by Setting

```
r + geom_point(color="DarkGreen")
```



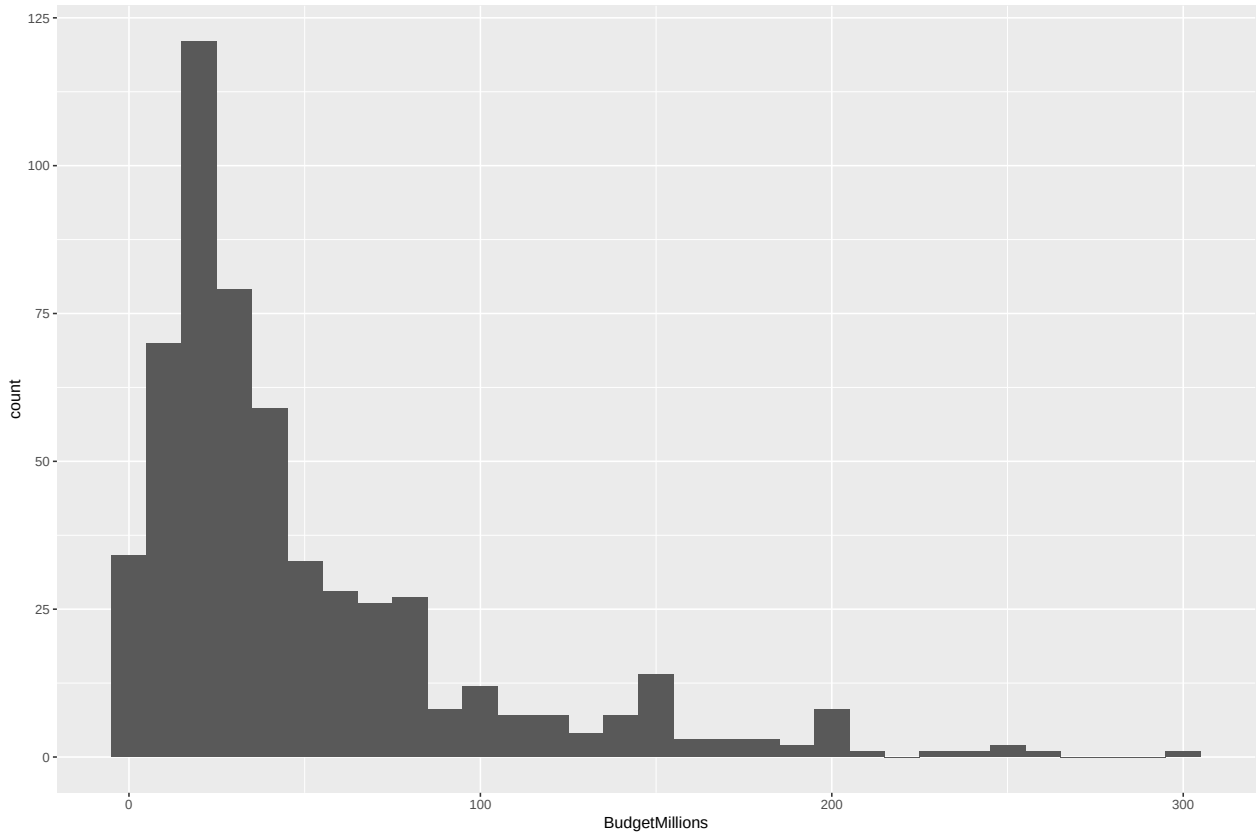

```
r + geom_point(aes(size=Year))
```

```
## Warning: Using size for a discrete variable is not advised.
```



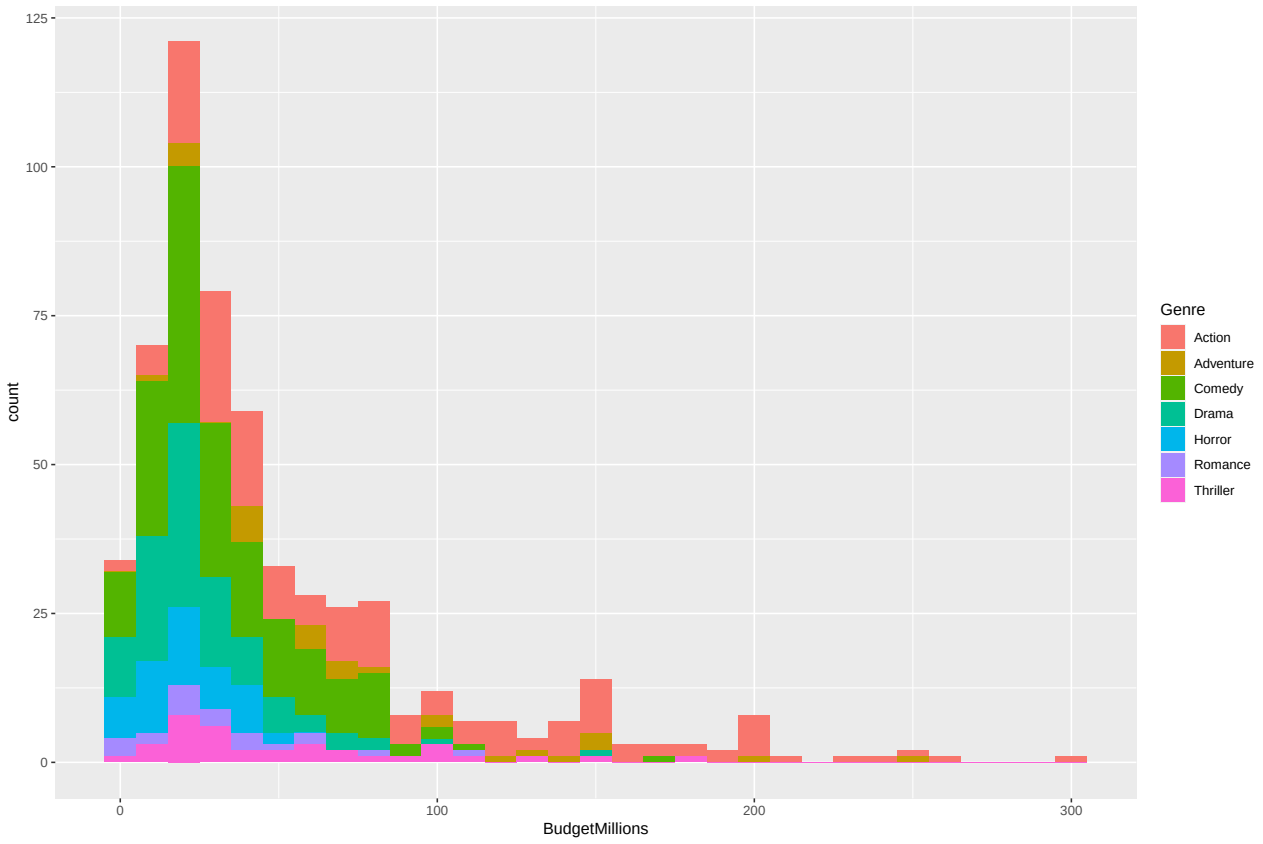
#Histograms and Density Charts

```
s <- ggplot(data=movie.ratings, aes(x=BudgetMillions))
s + geom_histogram(binwidth = 10)
```



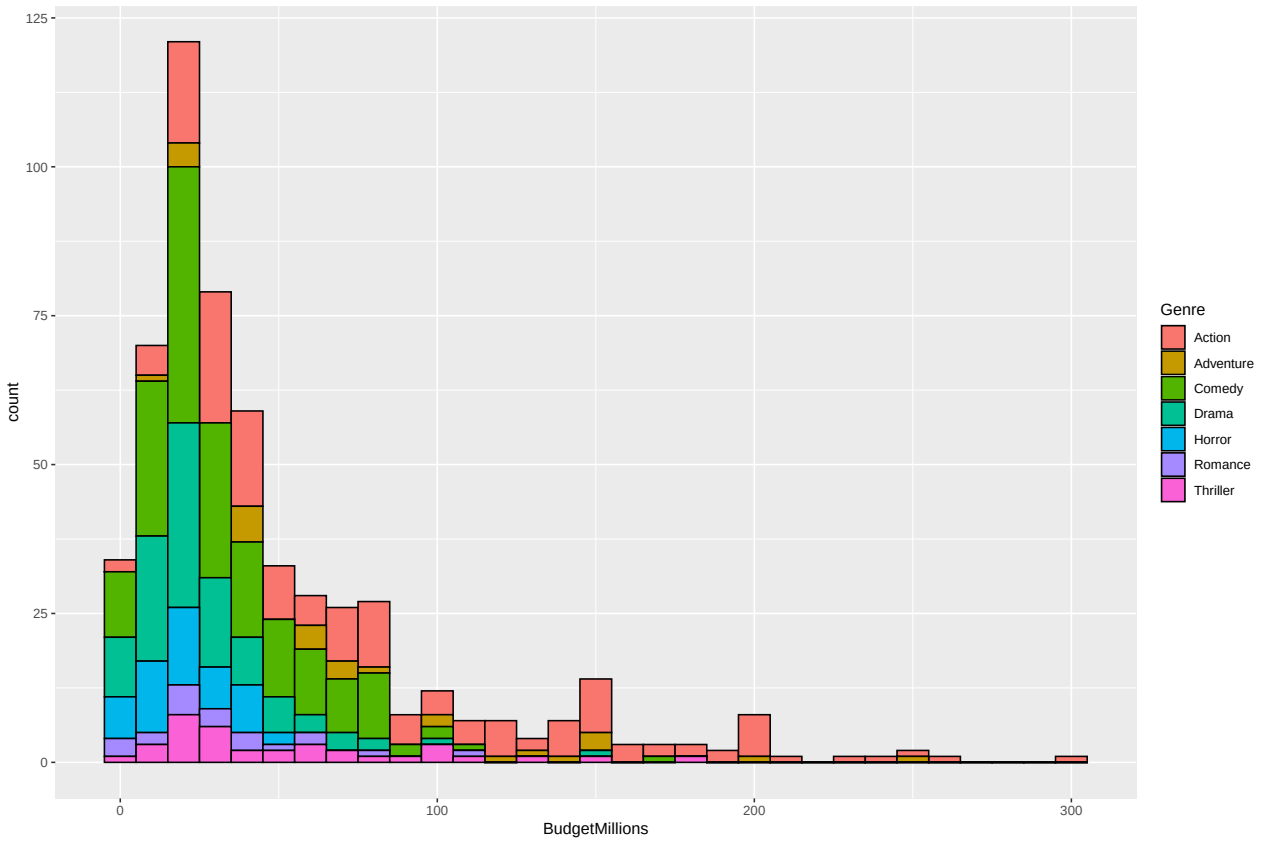
#Adding color

```
s + geom_histogram(binwidth = 10, aes(fill=Genre))
```



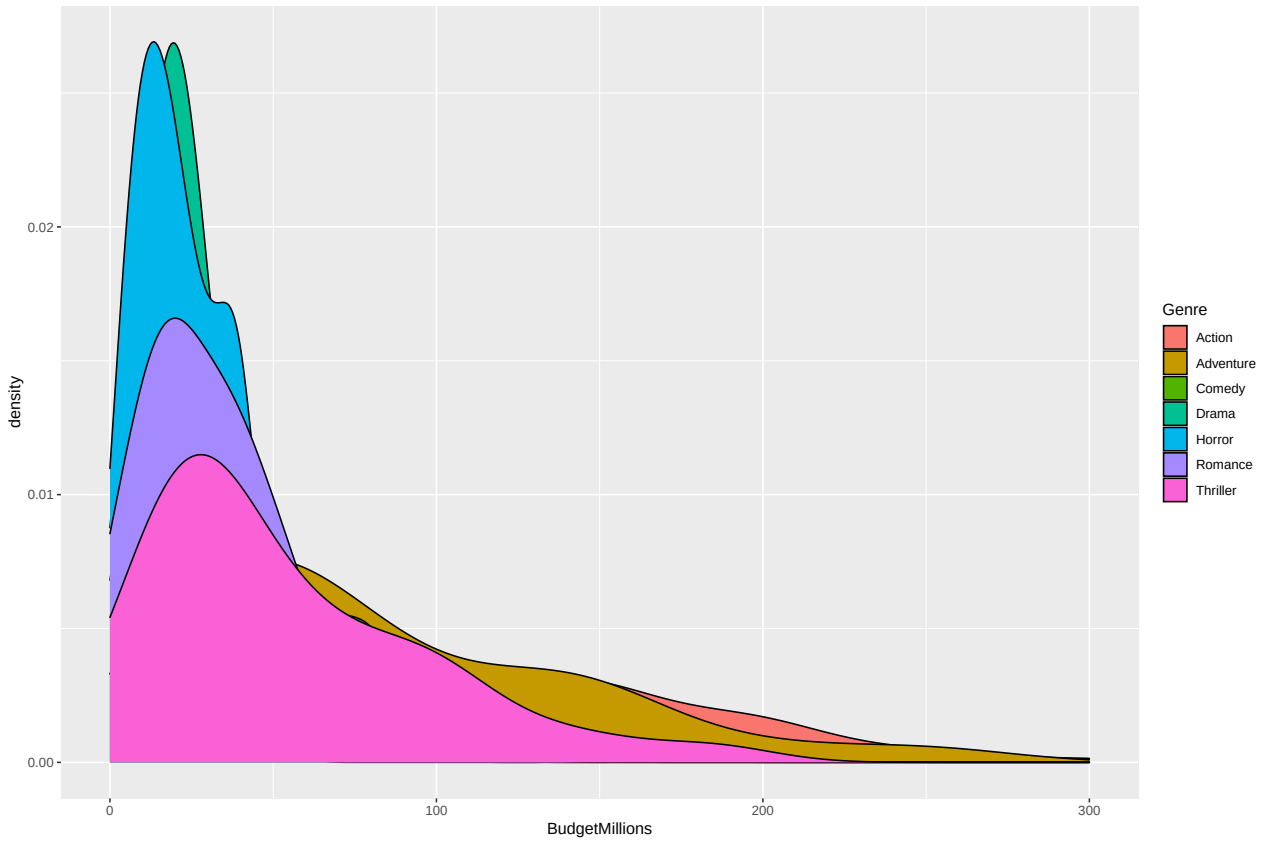
#Adding a border

```
s + geom_histogram(binwidth = 10, aes(fill=Genre), color = "Black")
```

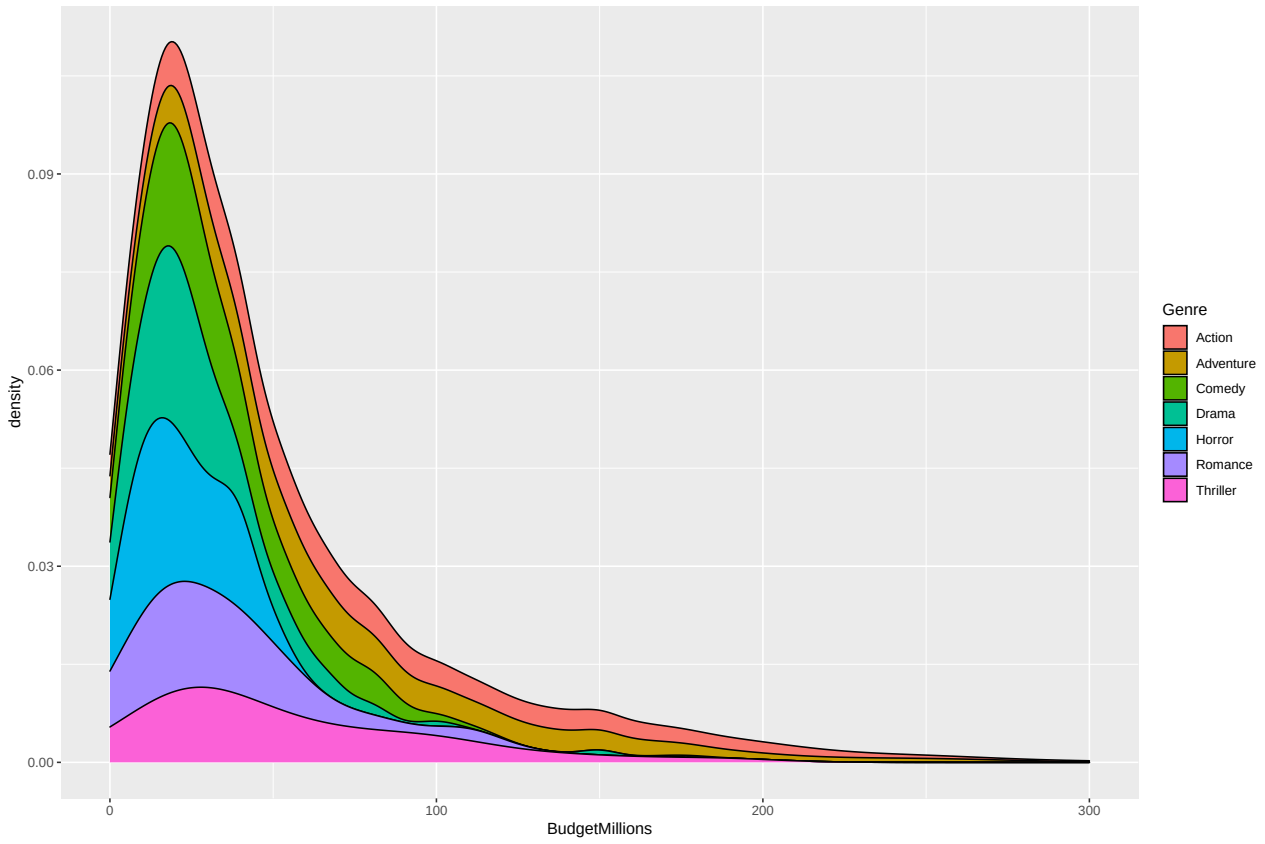


#Density Charts

```
s + geom_density(aes(fill=Genre))
```

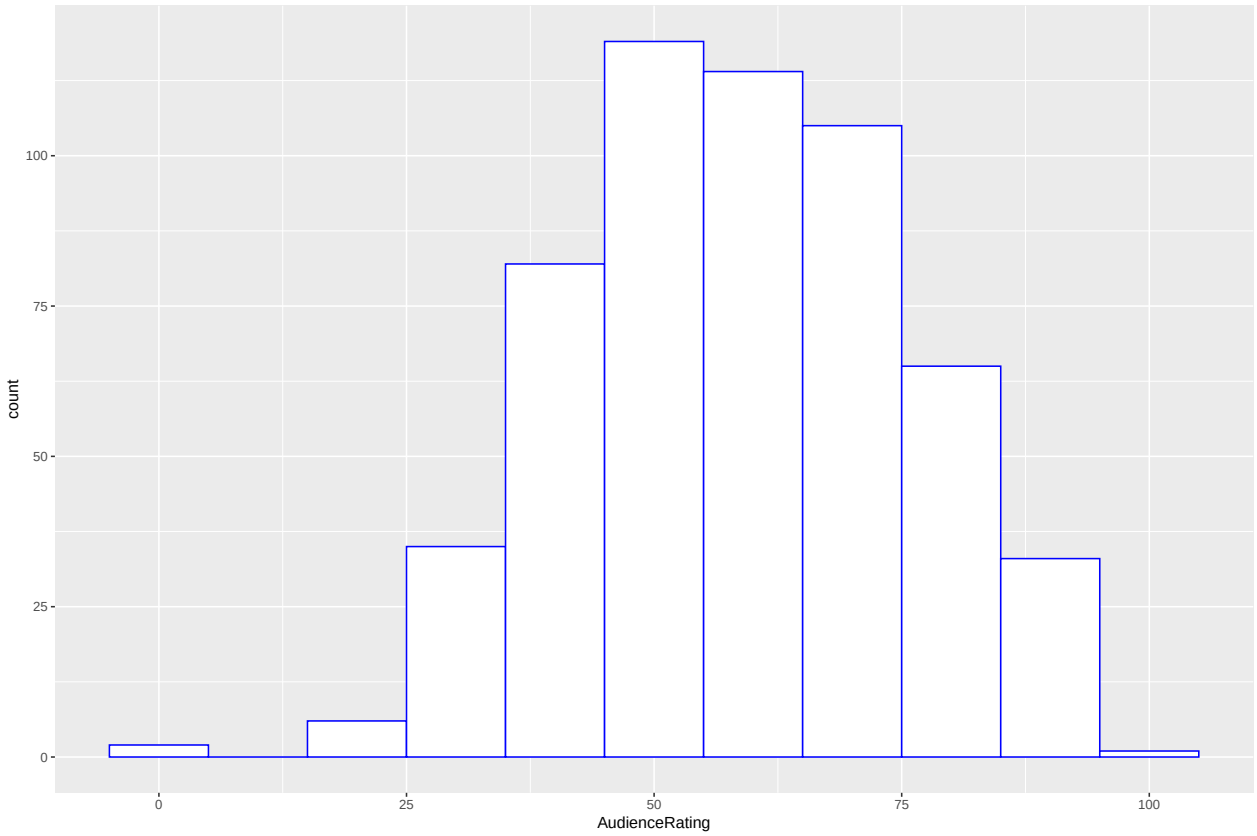



```
s + geom_density(aes(fill=Genre), position = "stack")
```



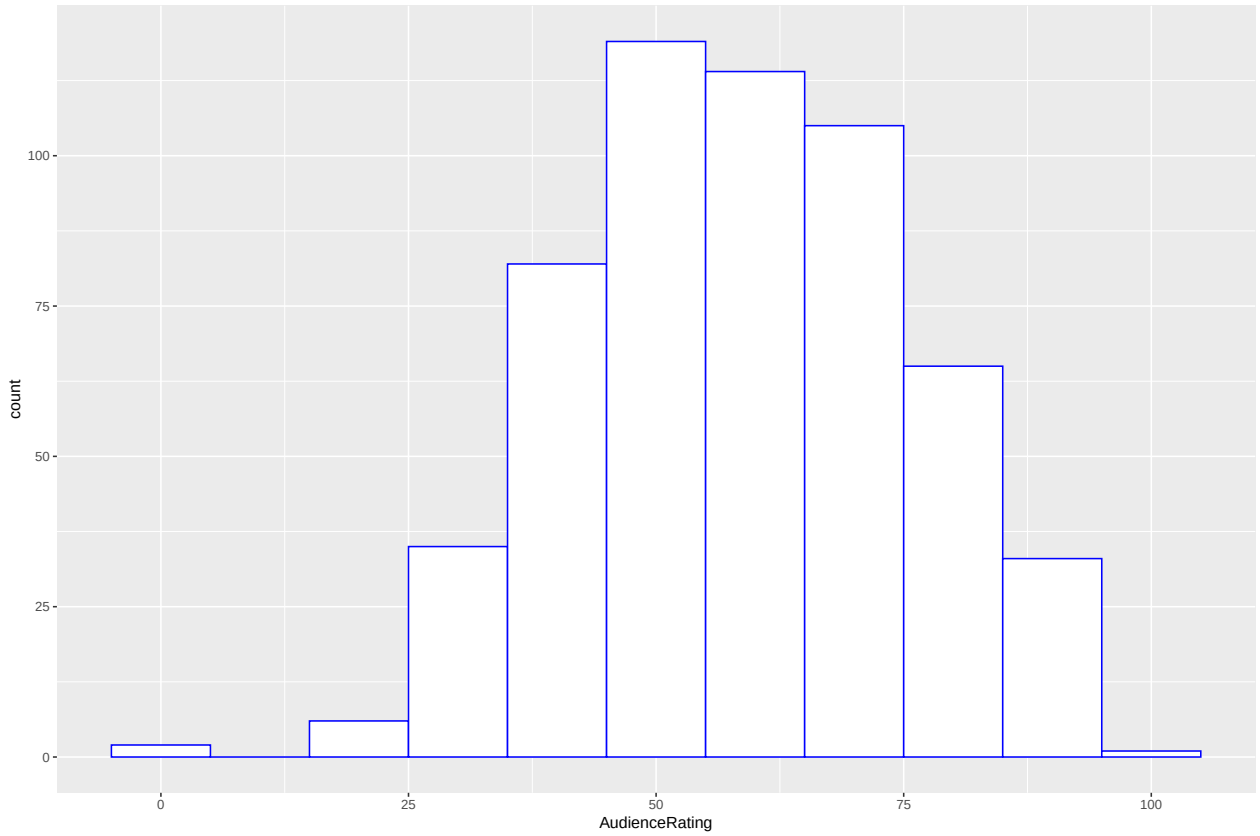
#Starting Layer Tips

```
t <- ggplot(data = movie.ratings, aes(x=AudienceRating))
t + geom_histogram(binwidth = 10, fill = "White", color = "Blue")
```

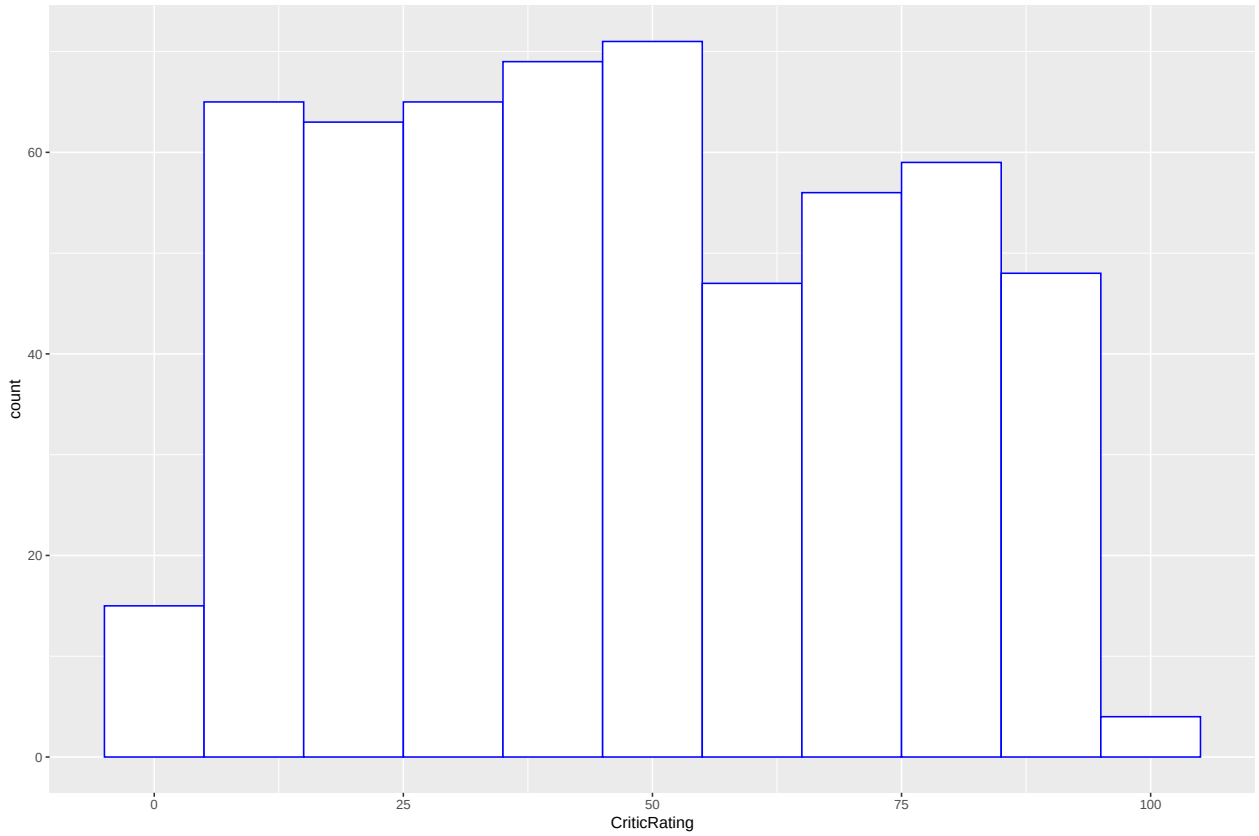


#Another Way

```
t <- ggplot(data=movie.ratings)
t + geom_histogram(binwidth = 10, fill = "White", color = "Blue", aes(x=AudienceRating))
```



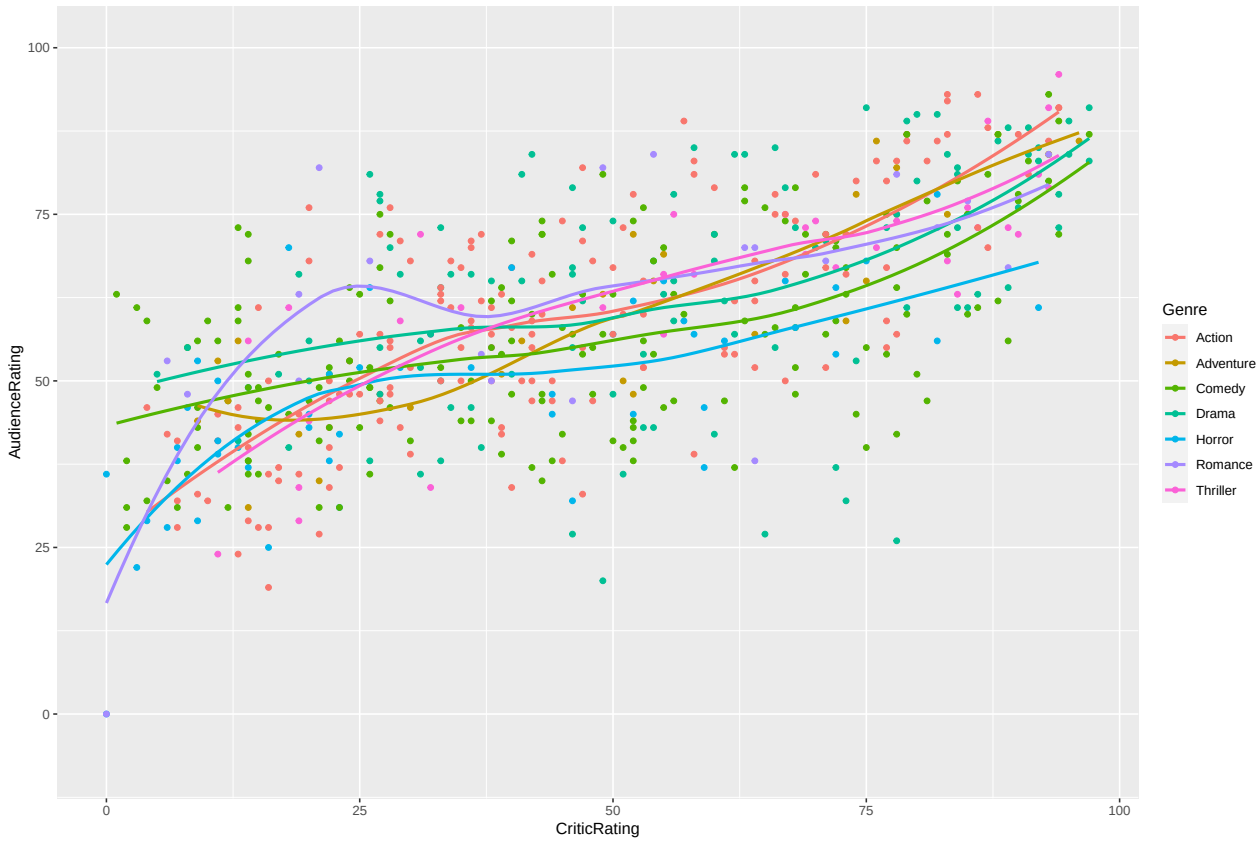

```
t + geom_histogram(binwidth = 10, fill = "White", color = "Blue", aes(x=CriticRating))
```



#Statistical Transformations

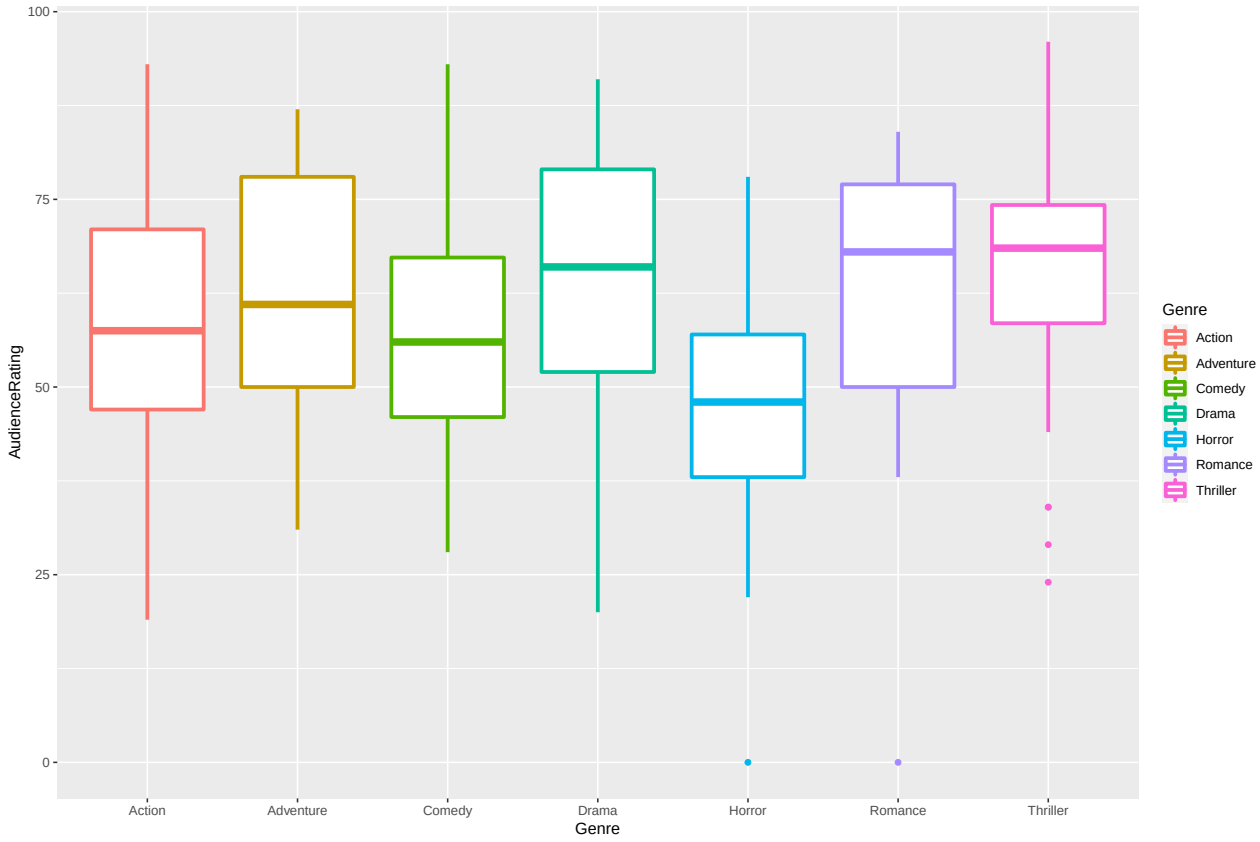
```
#?geom_smooth()

u <- ggplot(data=movie.ratings, aes(x=CriticRating, y = AudienceRating, color = Genre))
u + geom_point() + geom_smooth(fill=NA)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

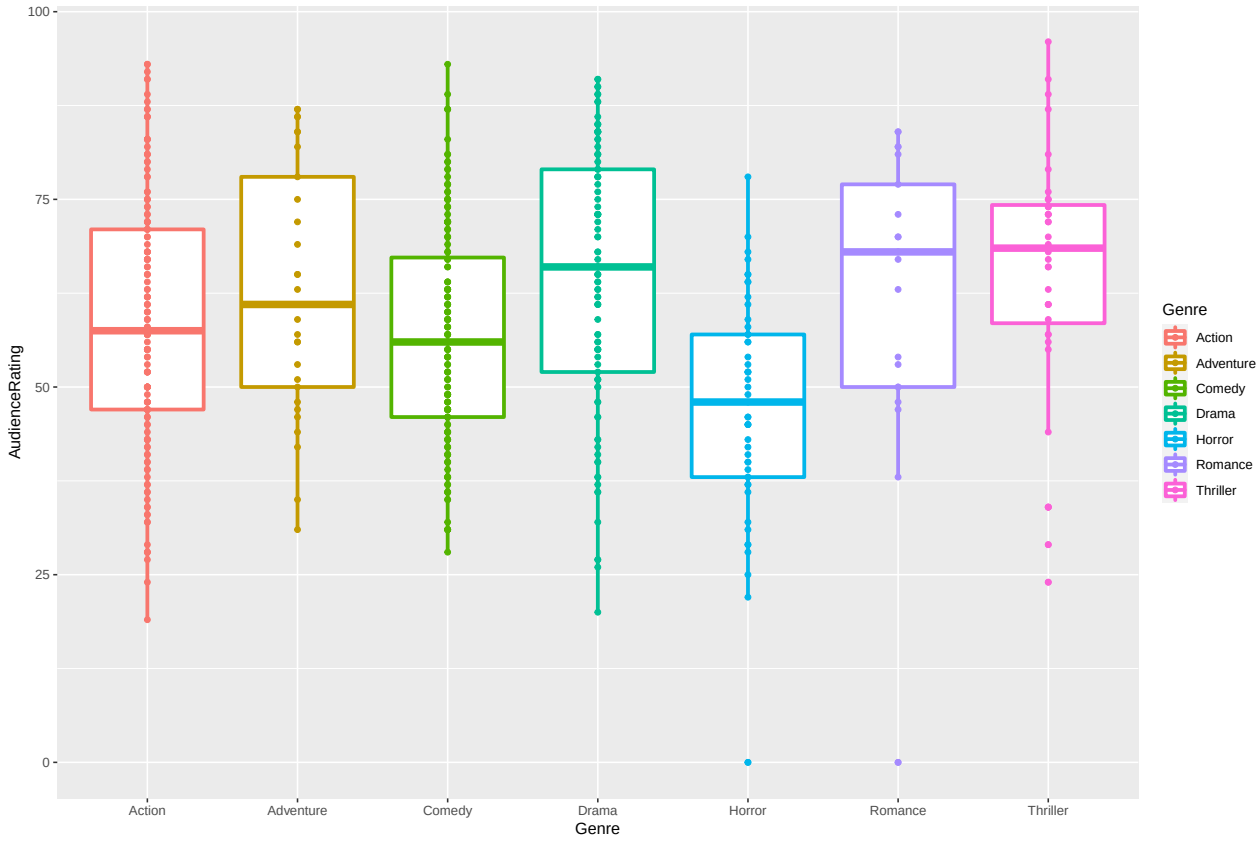


#boxplots

```
u <- ggplot(data=movie.ratings, aes(x=Genre, y=AudienceRating, color=Genre))
u + geom_boxplot(size=1.2)
```

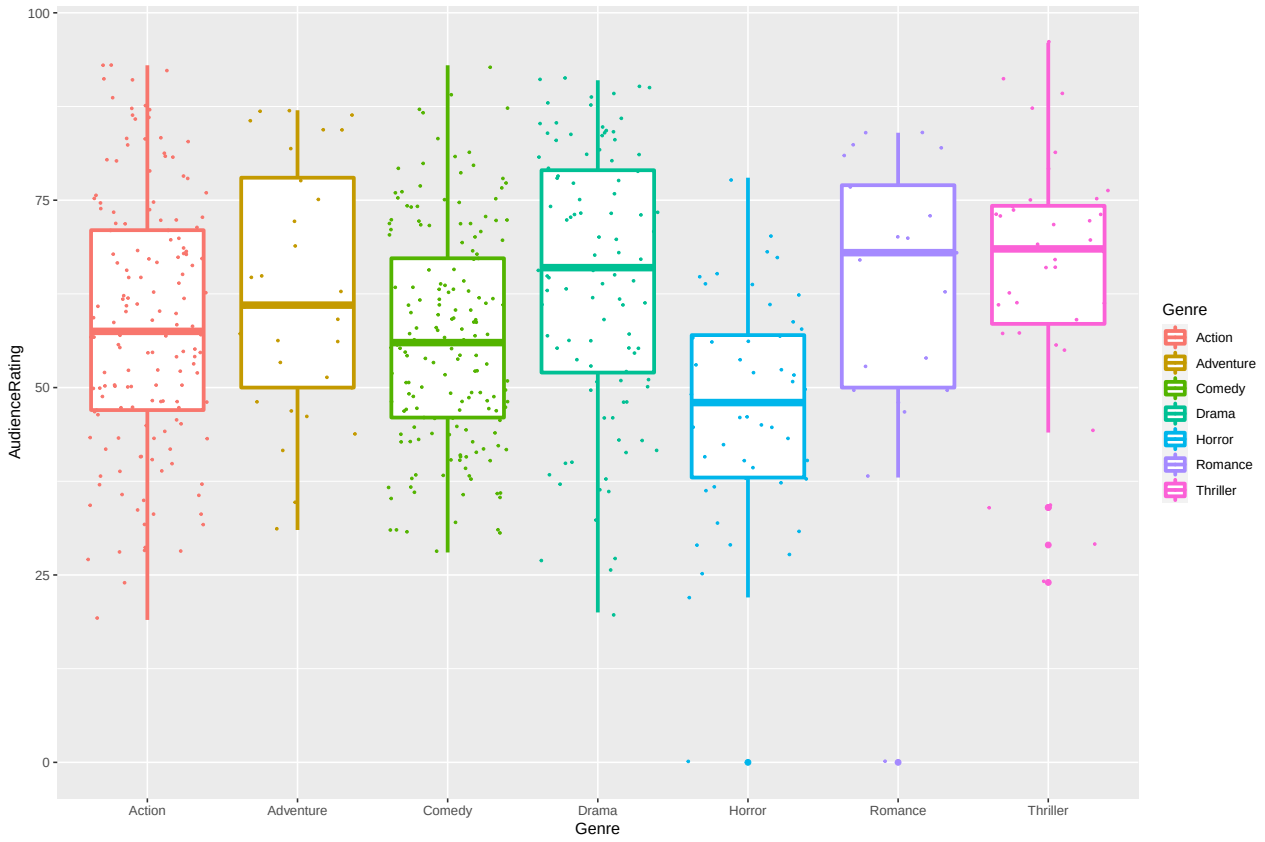



```
u + geom_boxplot(size=1.2) + geom_point()
```

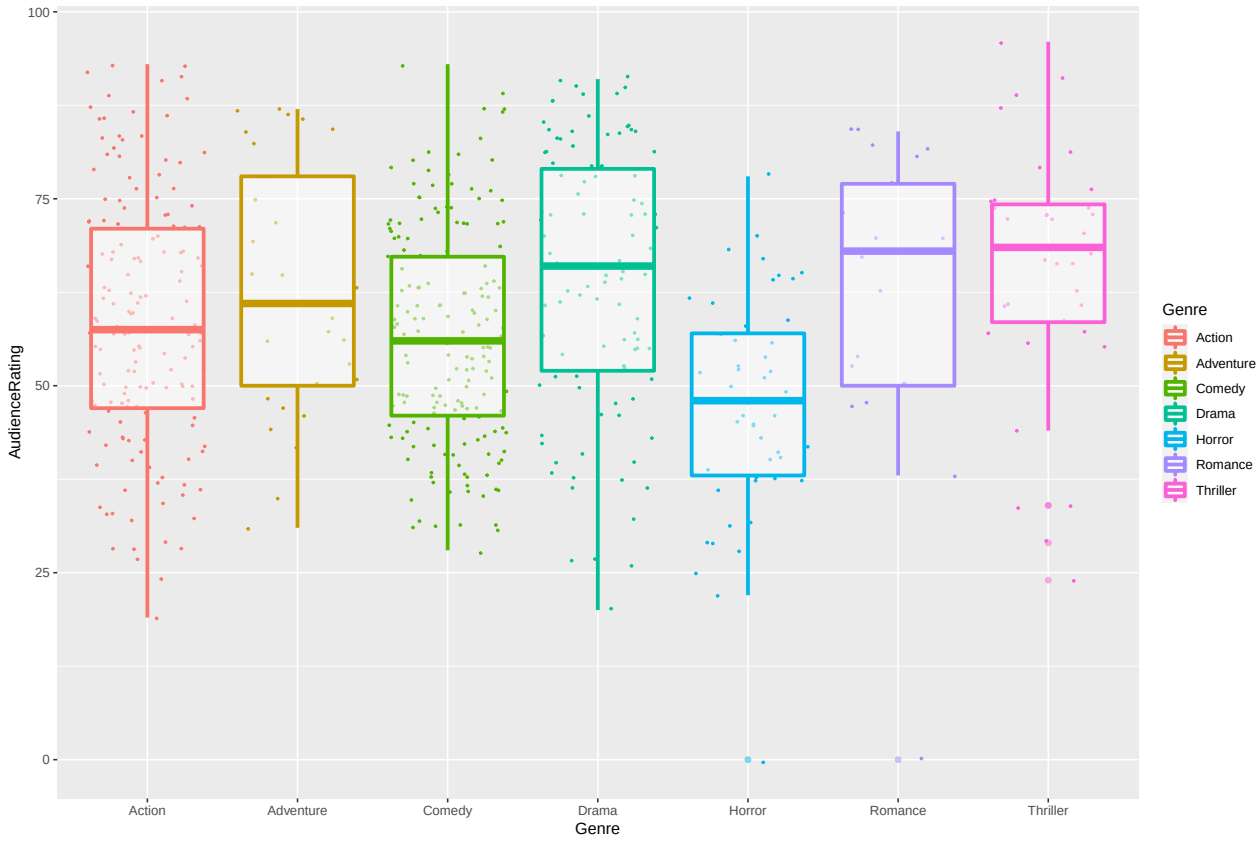


#tip

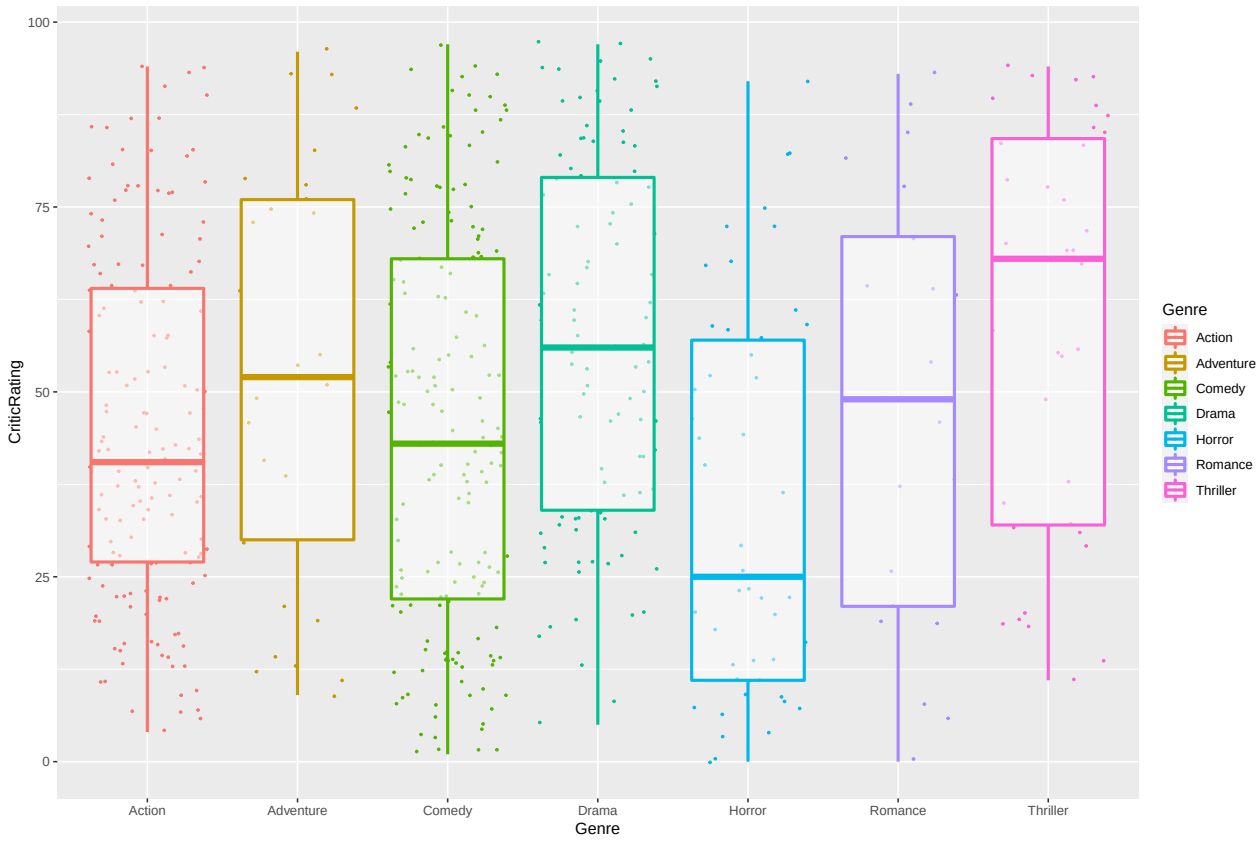
```
u + geom_boxplot(size=1.2) + geom_jitter(size=0.5)
```




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

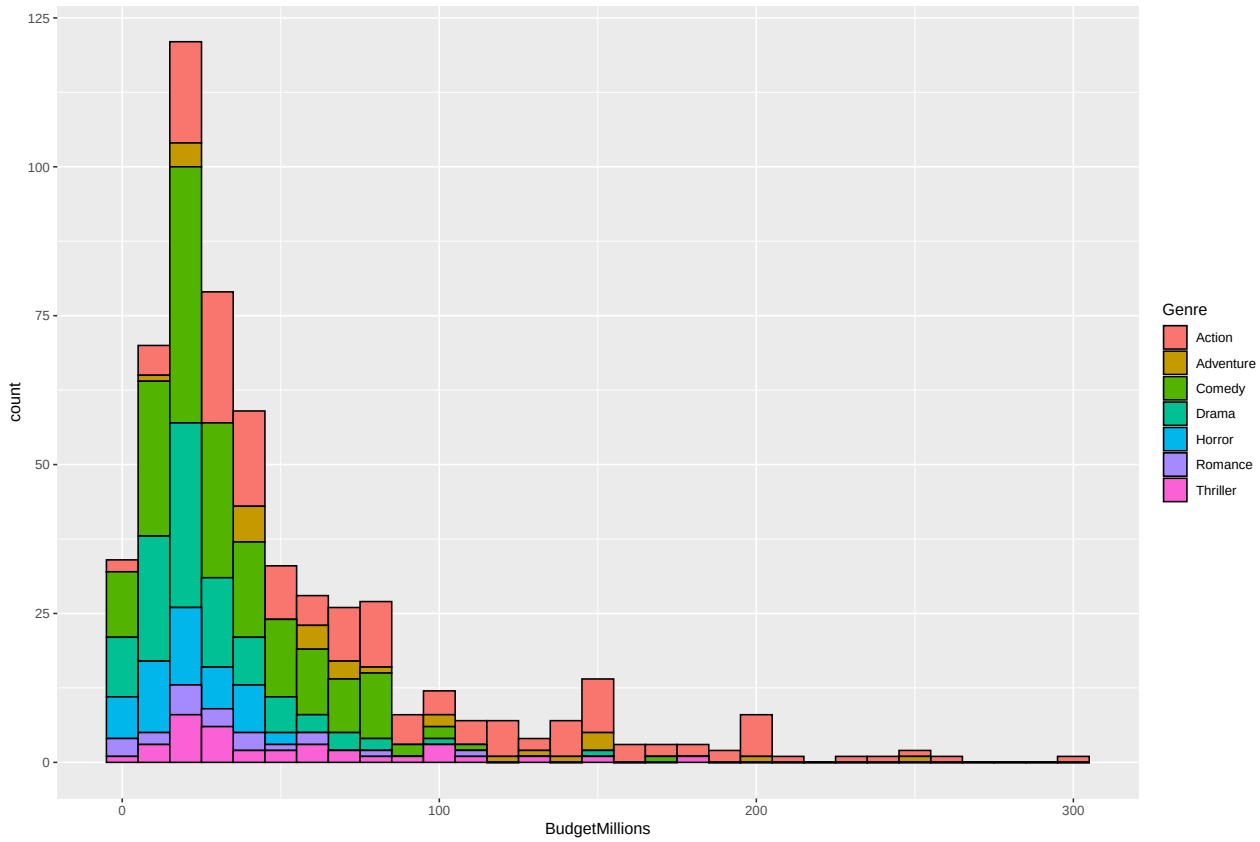



```
u.critic <- ggplot(data = movie.ratings, aes(x = Genre, y = CriticRating, color = Genre))
u.critic + geom_jitter(size = 0.5) + geom_boxplot(size = 1.0, alpha = 0.5)
```

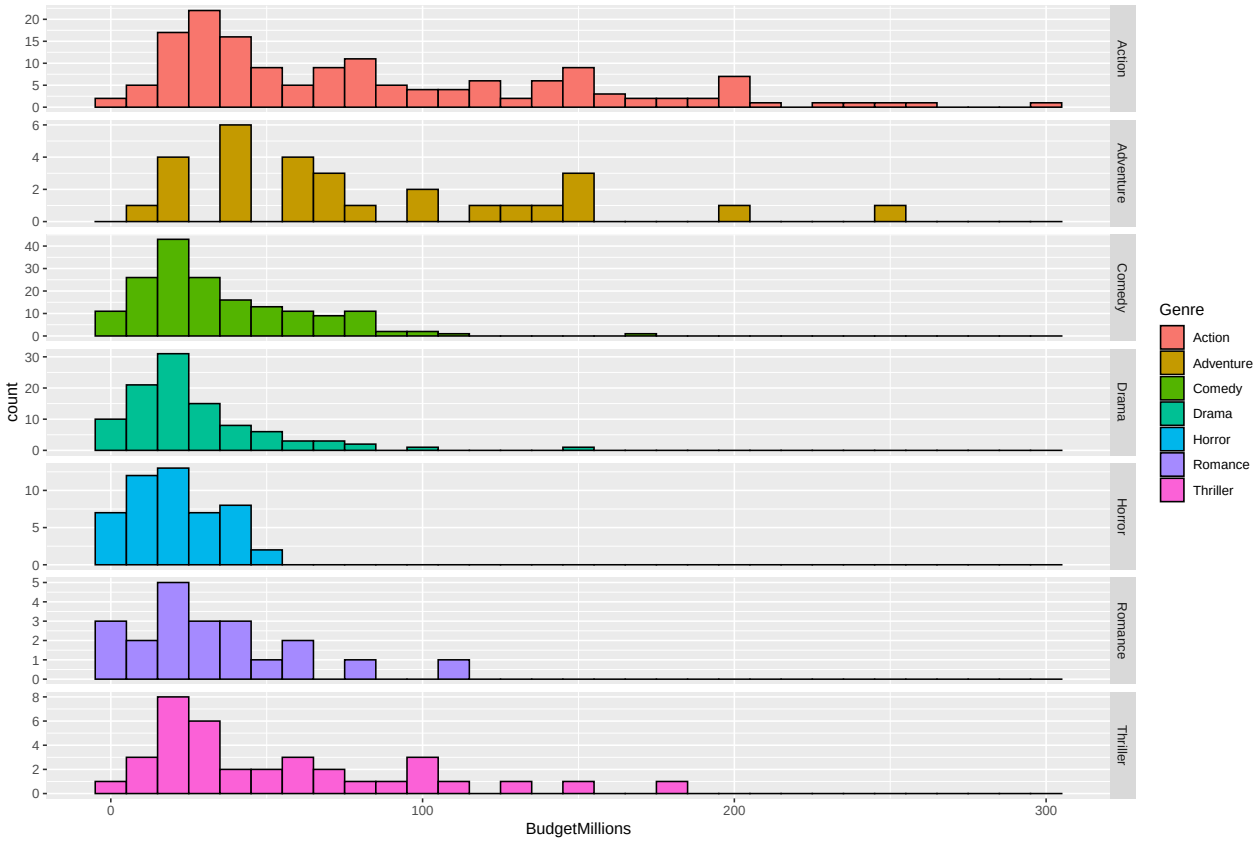


#Facets

```
v <- ggplot(data=movie.ratings, aes(x=BudgetMillions))
v + geom_histogram(binwidth = 10, aes(fill = Genre), color = "black")
```

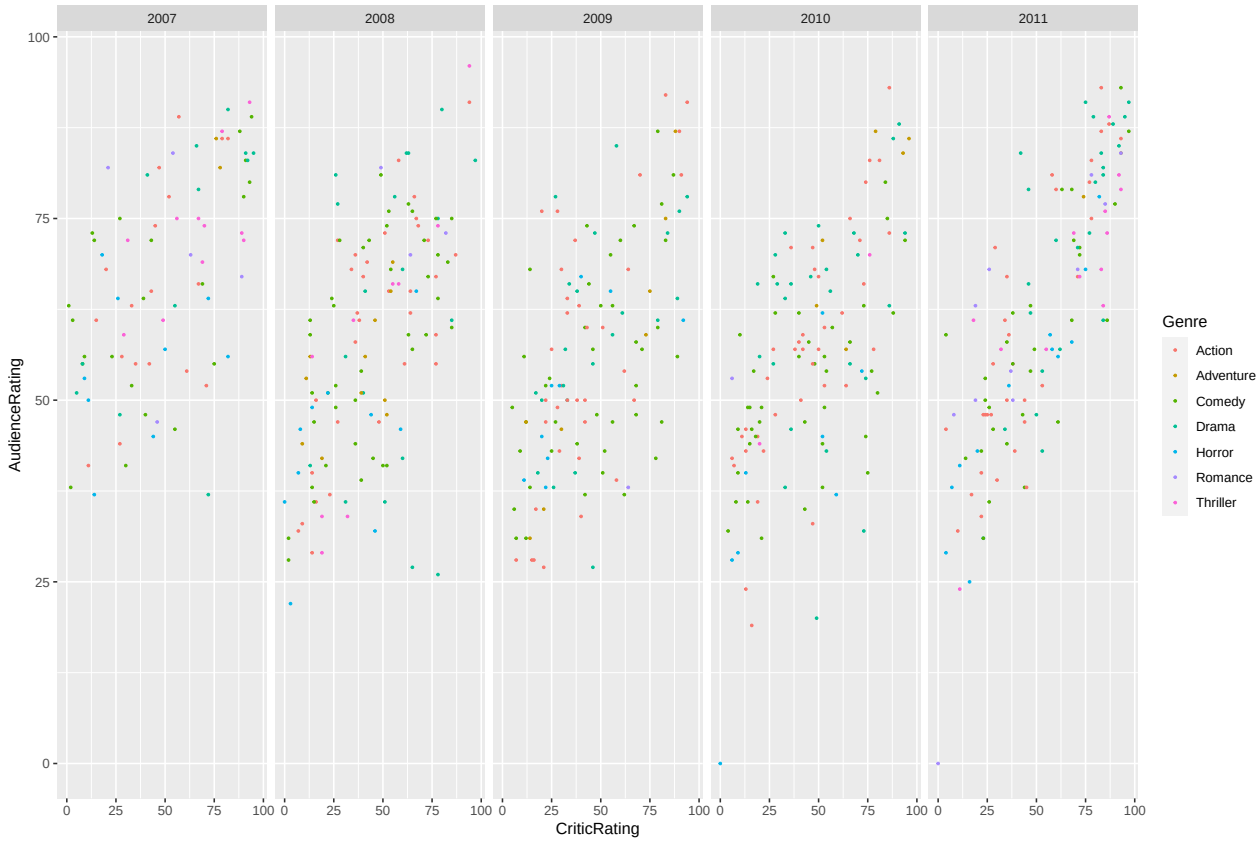



```
v + geom_histogram(binwidth = 10, aes(fill = Genre), color = "black") + facet_grid(Genre~., scale="free
```

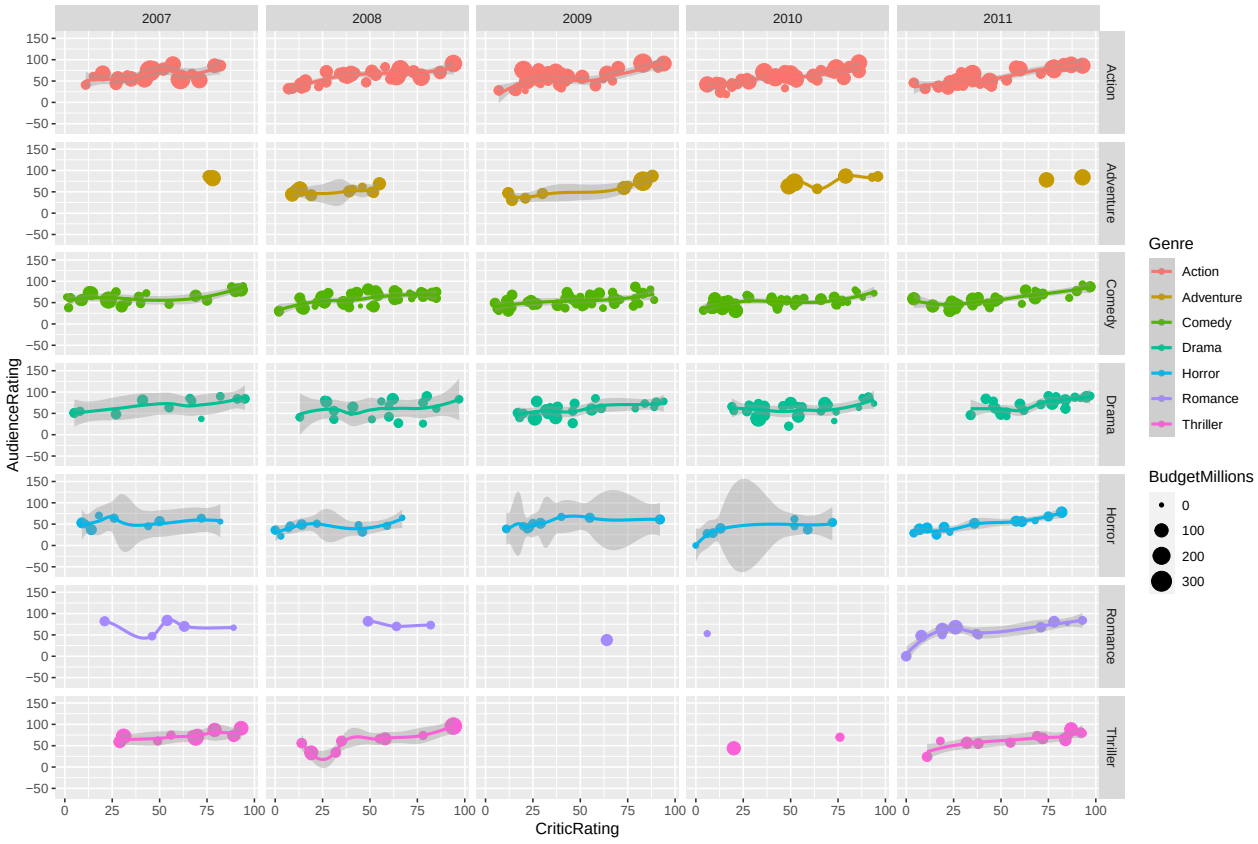


#Scatterplot: Facets

```
w <- ggplot(data=movie.ratings, aes(x=CriticRating, y = AudienceRating, color = Genre))
w + geom_point(size = 0.5) + facet_grid(.~Year)
```

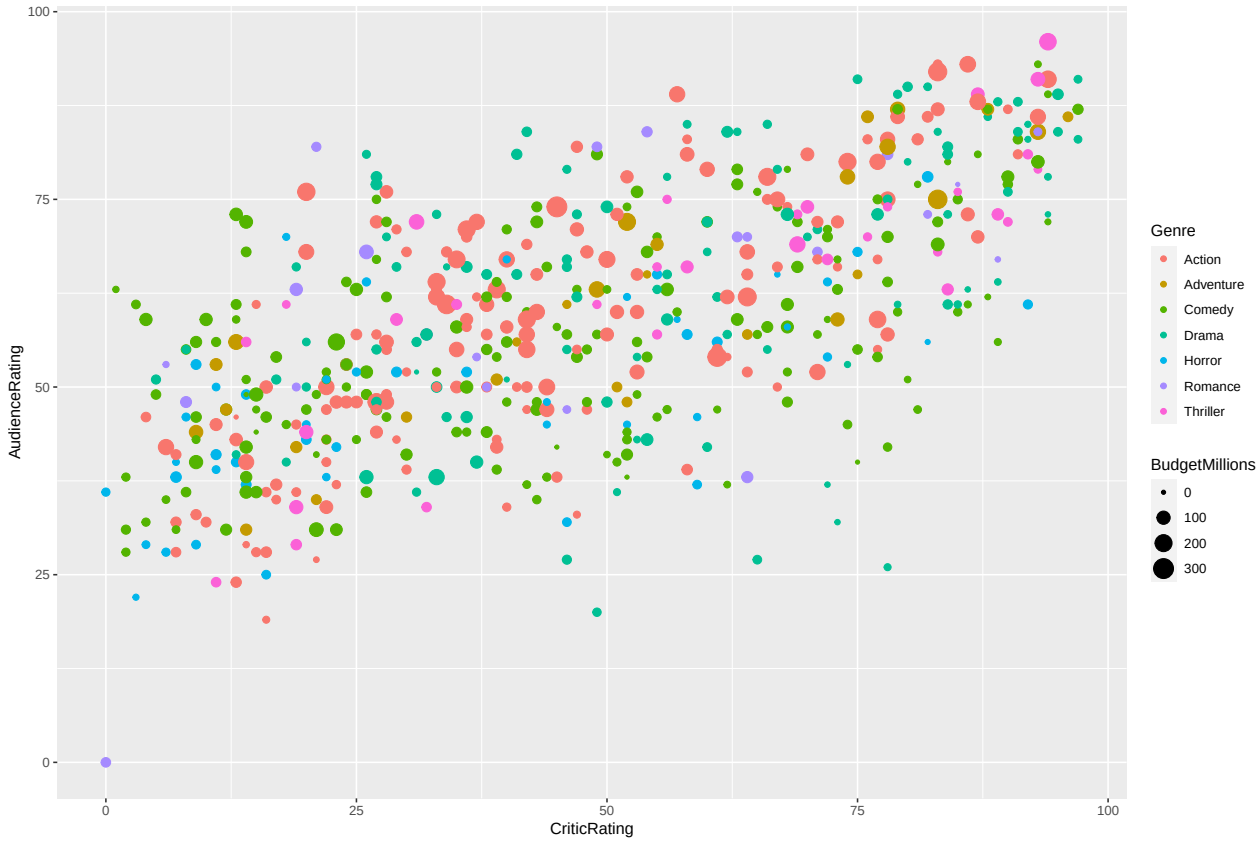



```
w + geom_point(aes(size=BudgetMillions)) + facet_grid(Genre~Year) + geom_smooth()
```



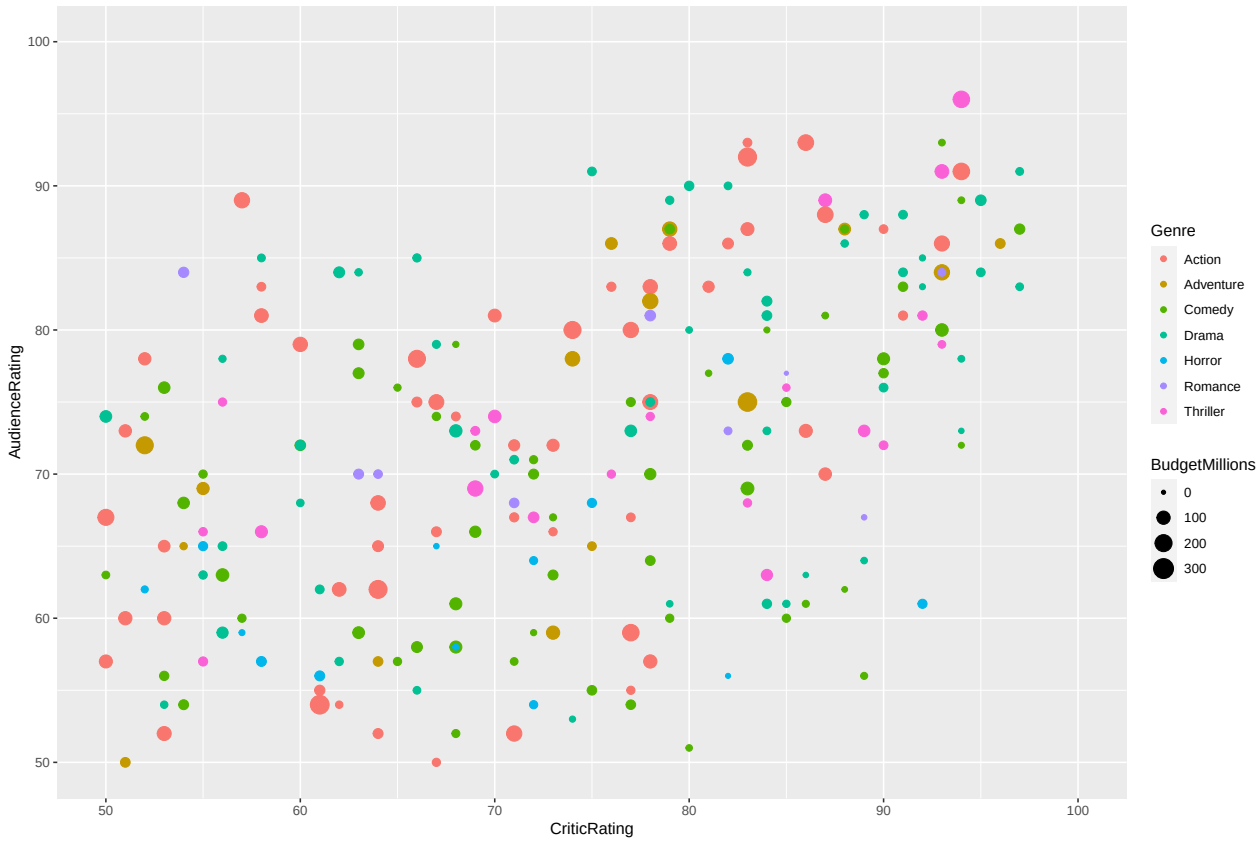
#Coordinates

```
m <- ggplot(data=movie.ratings, aes(x=CriticRating, y=AudienceRating, size=BudgetMillions, color=Genre))
m + geom_point()
```



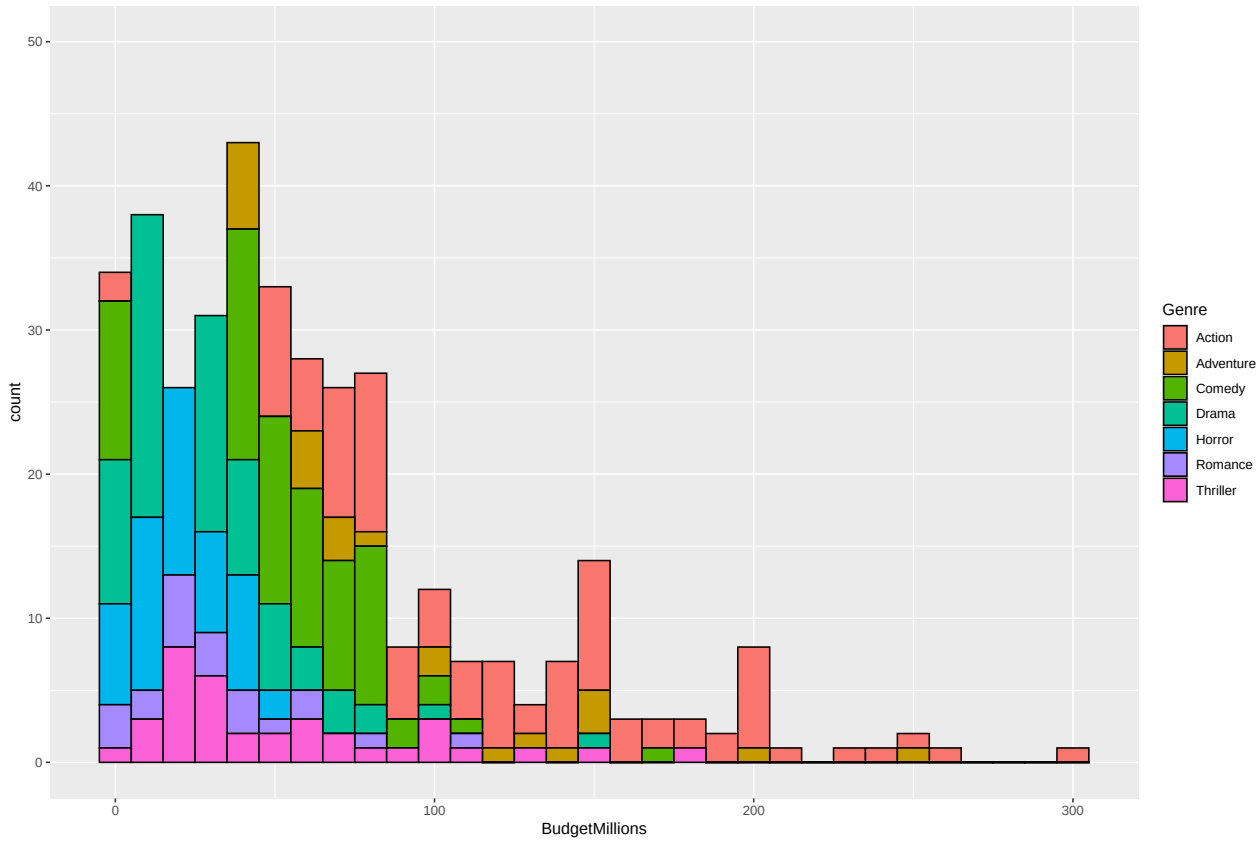

```
m + geom_point() + xlim(50, 100) + ylim(50,100)
```

```
## Warning: Removed 335 rows containing missing values (geom_point).
```



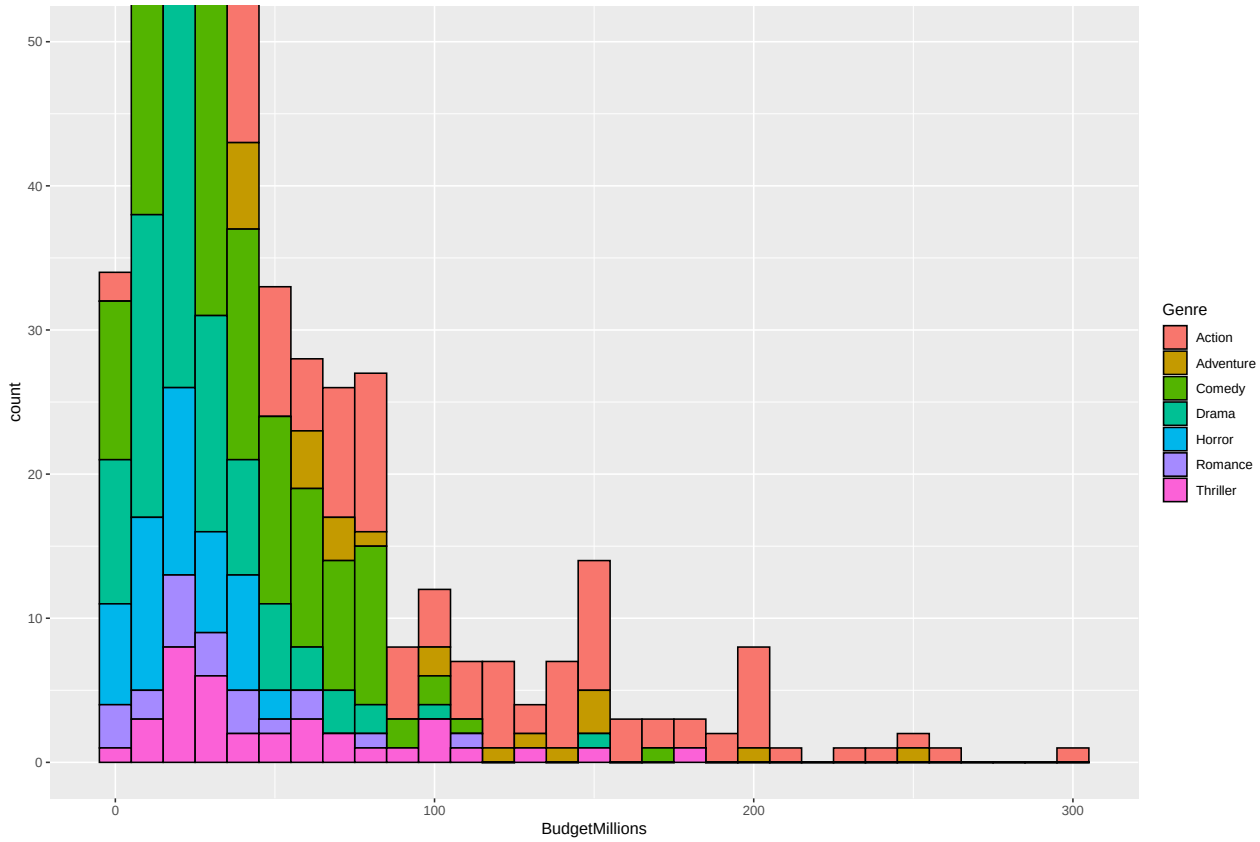

```
n <- ggplot(data=movie.ratings, aes(x=BudgetMillions))
n + geom_histogram(binwidth = 10, aes(fill=Genre), color = "black") + ylim(0, 50)
```

```
## Warning: Removed 11 rows containing missing values (geom_bar).
```



#Zoom

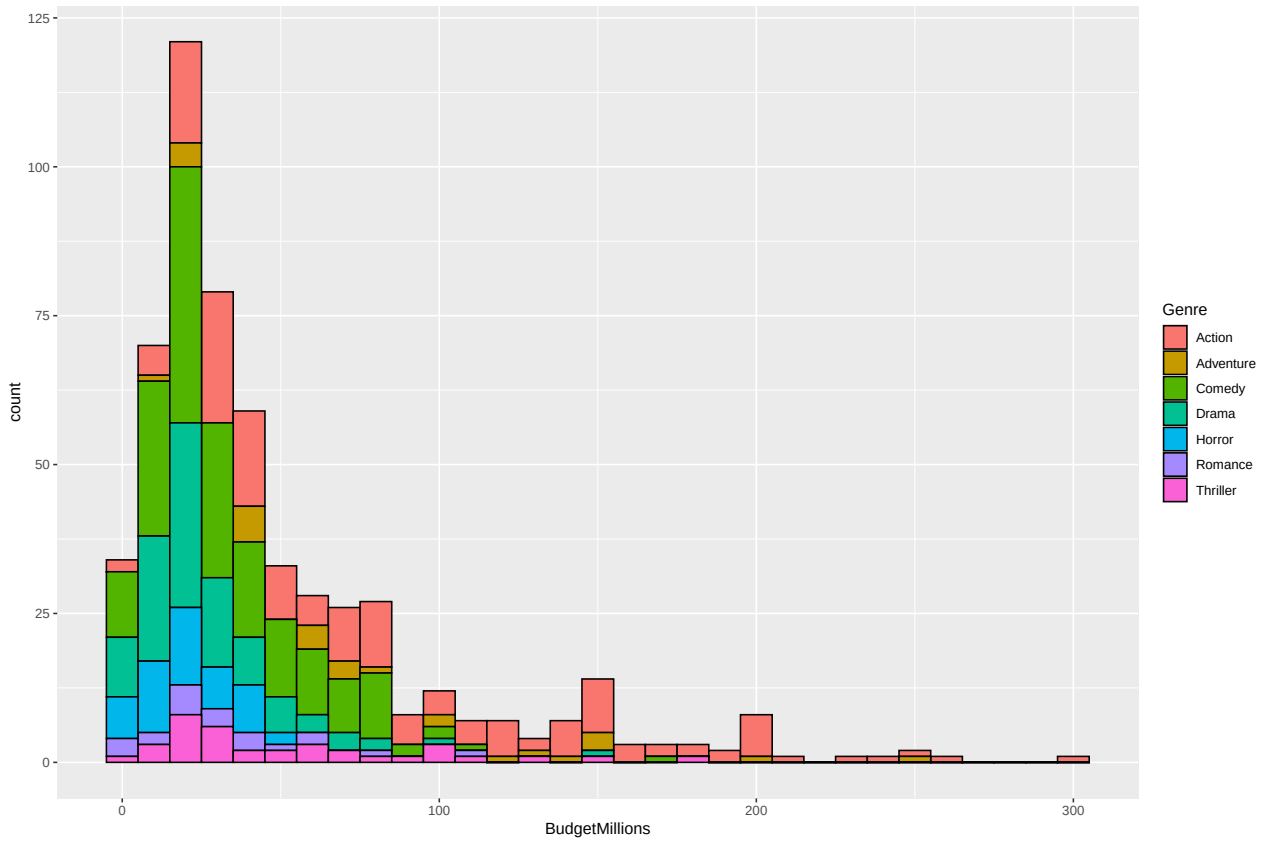
```
n + geom_histogram(binwidth = 10, aes(fill=Genre), color = "black") + coord_cartesian(ylim=c(0,50))
```



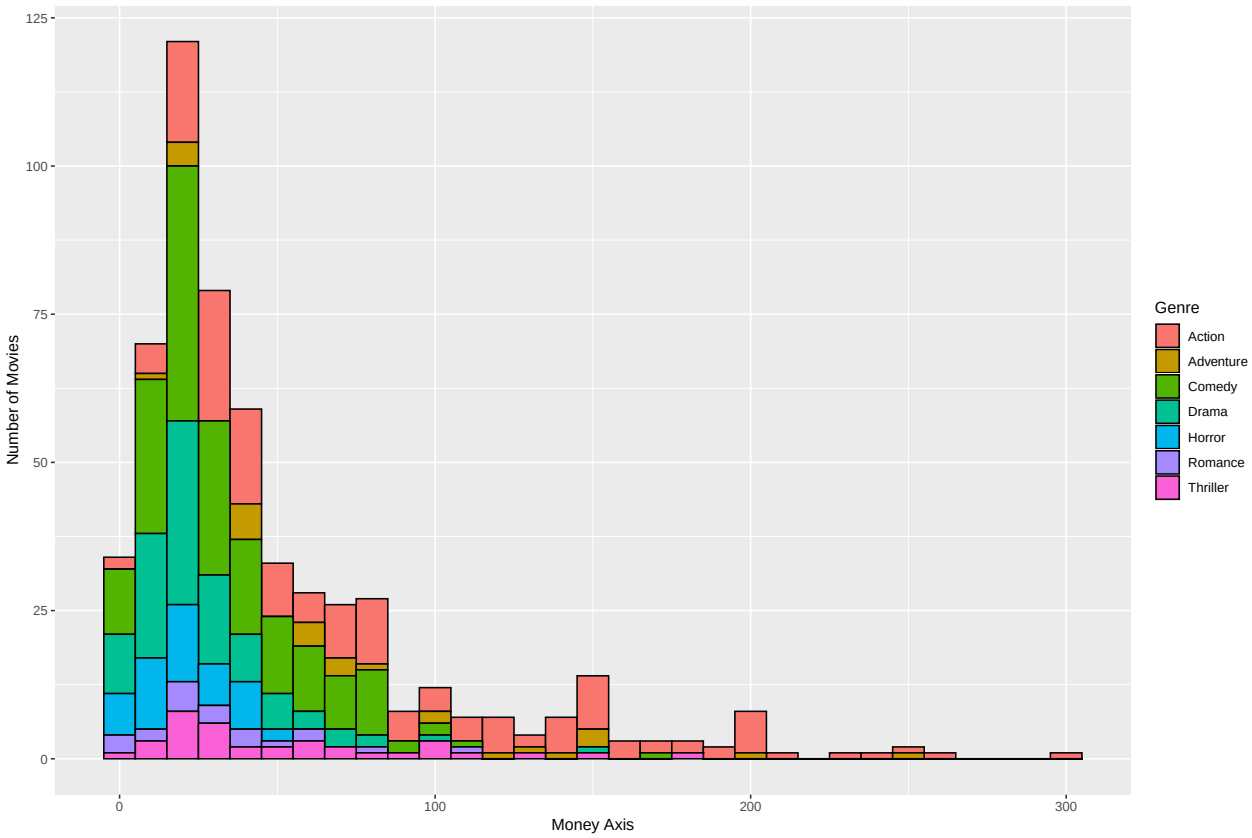
#Improve #1:

#Theme

```
o <- ggplot(data=movie.ratings, aes(x=BudgetMillions))
h <- o + geom_histogram(binwidth = 10, aes(fill = Genre), color = "Black")
h
```

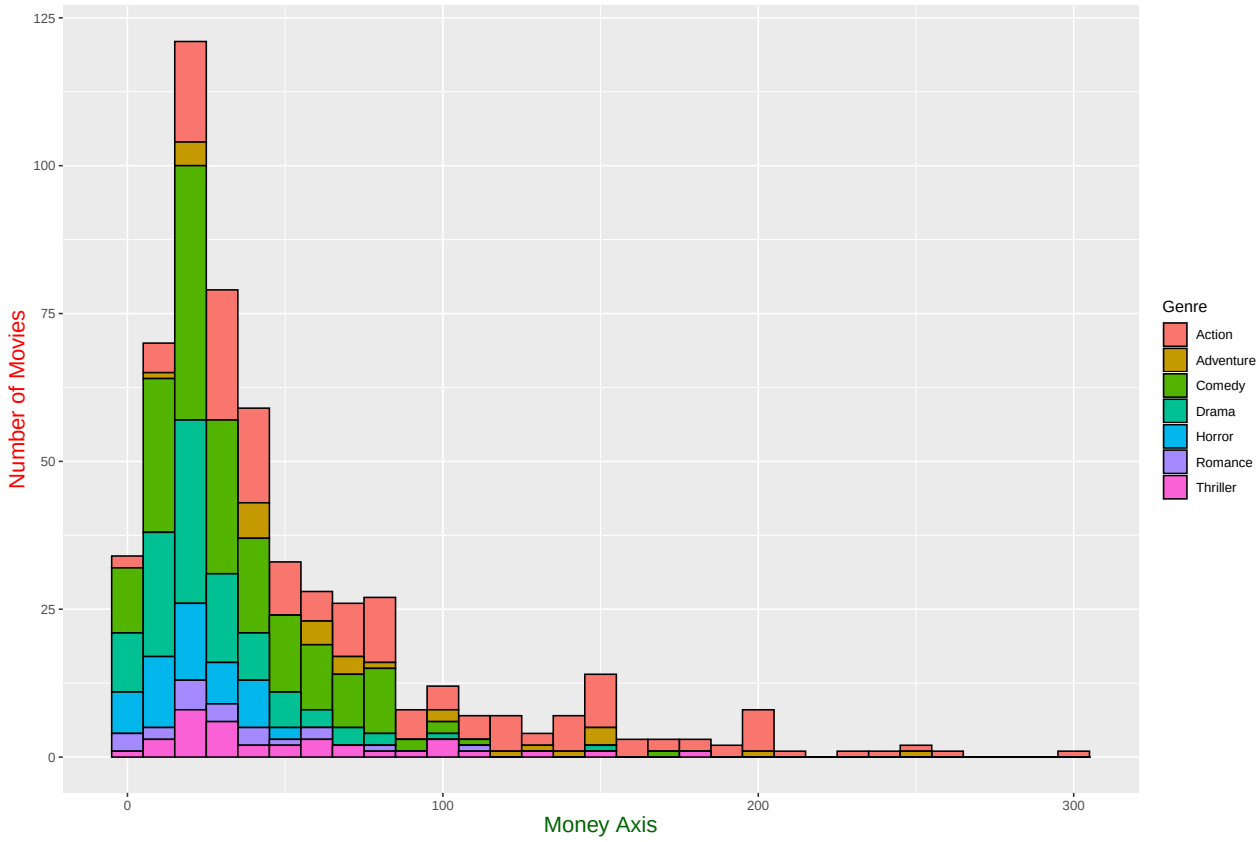



```
h + xlab("Money Axis") + ylab("Number of Movies")
```



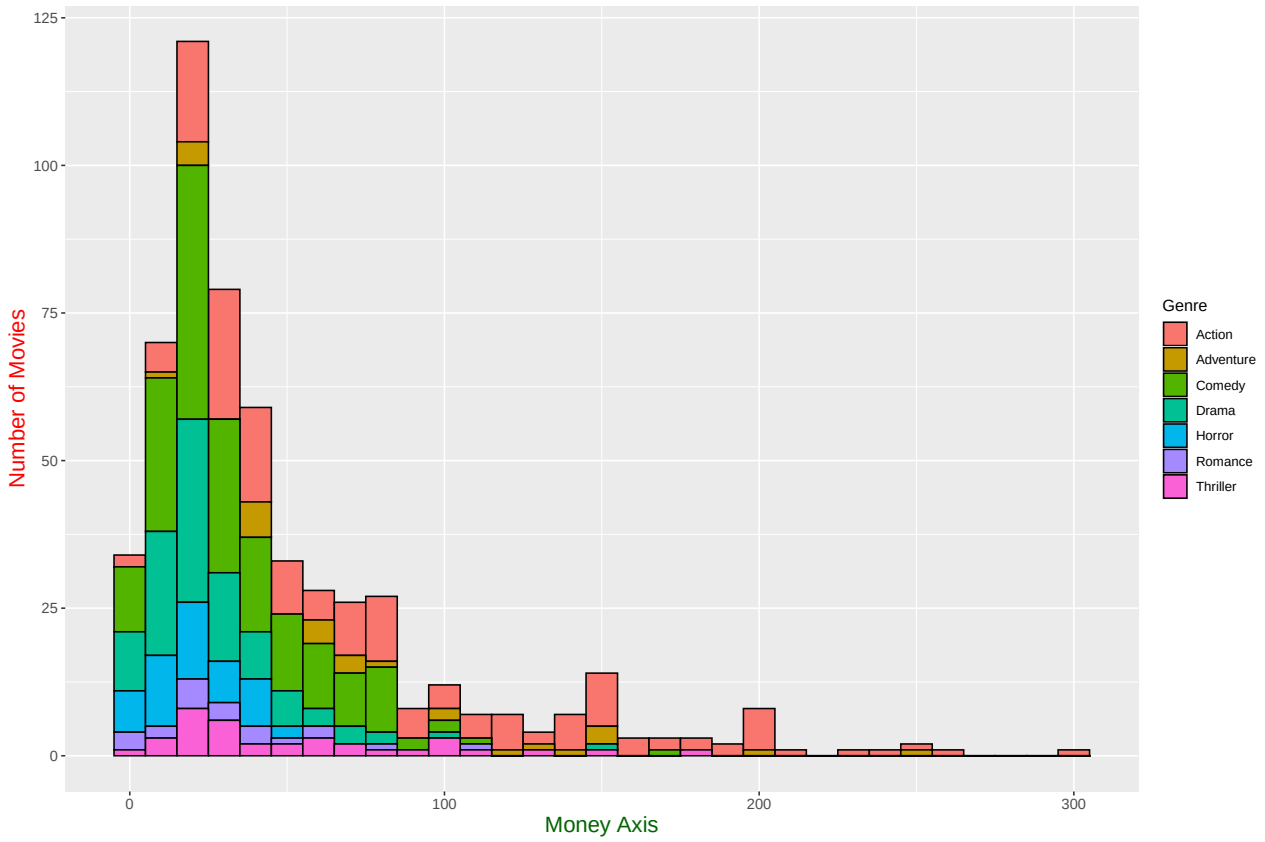
#Label Formating

```
h + xlab("Money Axis") + ylab("Number of Movies") + theme(axis.title.x = element_text(color ="DarkGreen"
```



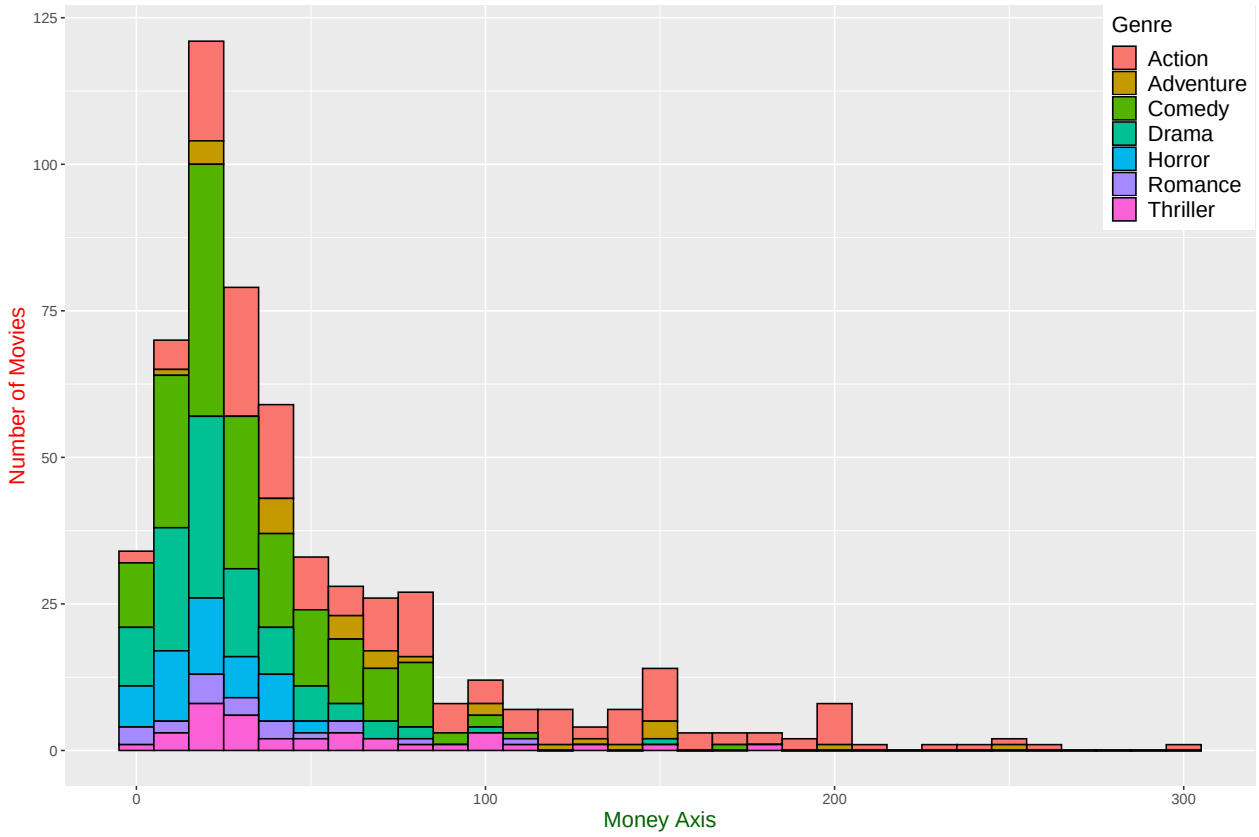
#Tick Mark Formating

```
h + xlab("Money Axis") + ylab("Number of Movies") + theme(axis.title.x = element_text(color ="DarkGreen"
```



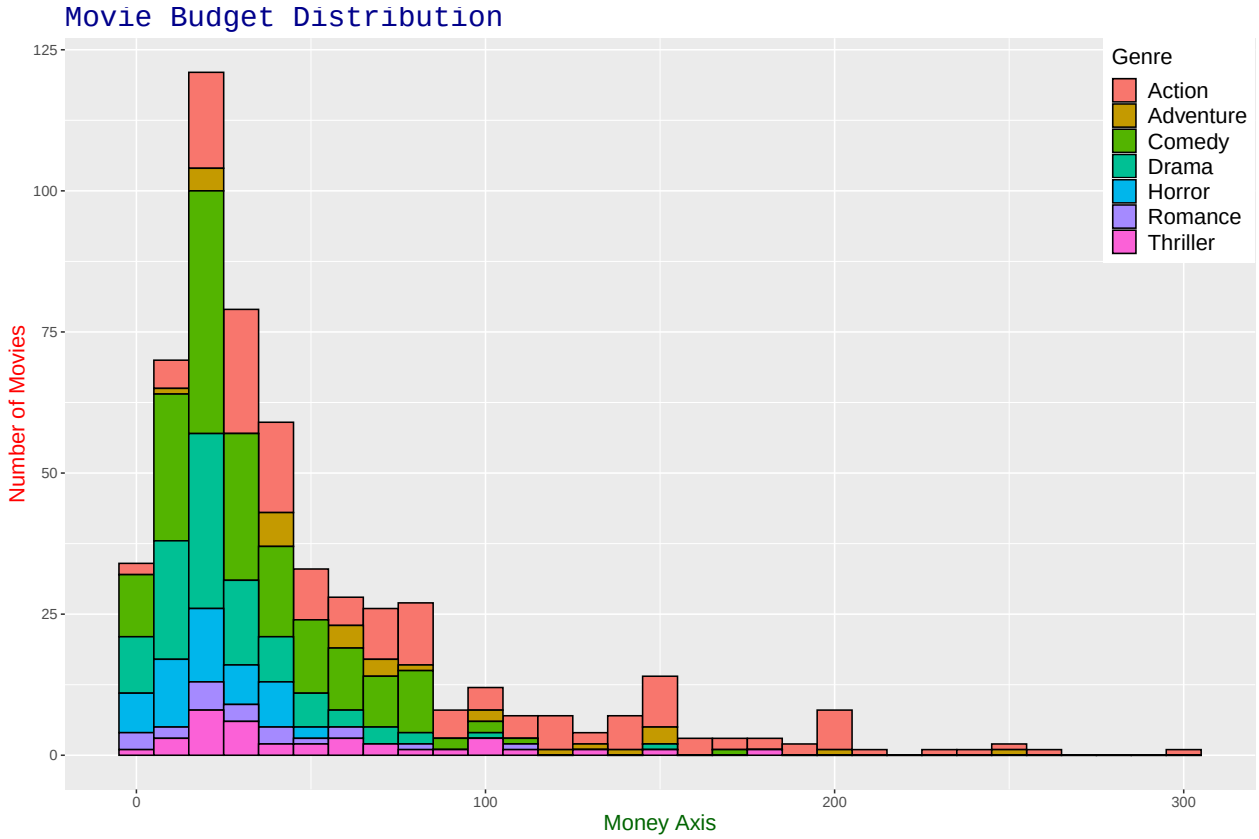
#Legend Formatting

```
h + xlab("Money Axis") + ylab("Number of Movies") +
  theme(axis.title.x = element_text(color ="DarkGreen", size=15),
        axis.title.y = element_text(color="Red", size=15),
        axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10),
        legend.title = element_text(size=15),
        legend.text = element_text(size=15),
        legend.position = c(1,1),
        legend.justification = c(1,1))
```



#Add Title

```
h + xlab("Money Axis") +
  ylab("Number of Movies") +
  ggtitle("Movie Budget Distribution")+
  theme(axis.title.x = element_text(color ="DarkGreen", size=15),
        axis.title.y = element_text(color="Red", size=15),
        axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10),
        legend.title = element_text(size=15),
        legend.text = element_text(size=15),
        legend.position = c(1,1),
        legend.justification = c(1,1),
        plot.title = element_text(color="DarkBlue", size = 20, family = "Courier"))
```



#New Challenge

```
movie.gross <- read.csv("MovieGross.csv")
head(movie.gross)
```

```
##   Day.of.Week               Director  Genre       Movie.Title Release.Date
## 1      Friday              Brad Bird action      Tomorrowland   22/05/2015
## 2      Friday            Scott Waugh action    Need for Speed   14/03/2014
## 3      Friday          Patrick Hughes action The Expendables 3   15/08/2014
## 4      Friday Phil Lord, Chris Miller comedy    21 Jump Street   16/03/2012
## 5      Friday         Roland Emmerich action  White House Down   28/06/2013
## 6      Friday              David Ayer action              Fury   17/10/2014
##                 Studio Adjusted.Gross...mill. Budget...mill. Gross...mill.
## 1 Buena Vista Studios                  202.1            170         202.1
## 2 Buena Vista Studios                  204.2             66         203.3
## 3           Lionsgate                  207.1            100         206.2
## 4                Sony                  208.8             42         201.6
## 5                Sony                  209.7            150         205.4
## 6                Sony                  212.8             80         211.8
##   IMDb.Rating MovieLens.Rating Overseas...mill. Overseas. Profit...mill.
## 1         6.7             3.26            111.9      55.4           32.1
## 2         6.6             2.97            159.7      78.6          137.3
## 3         6.1             2.93            166.9      80.9          106.2
## 4         7.2             3.62             63.1      31.3          159.6
## 5         8.0             3.65            132.3      64.4           55.4
## 6         5.8             2.85              126      59.5          131.8
##   Profit. Runtime..min. US...mill. Gross...US
## 1    18.9           130       90.2       44.6
```

```
## 2   208.0           132       43.6       21.4
## 3   106.2           126       39.3       19.1
## 4   380.0           109      138.4       68.7
## 5    36.9           131       73.1       35.6
## 6   164.8           134       85.8       40.5
```

```
dim(movie.gross)
```

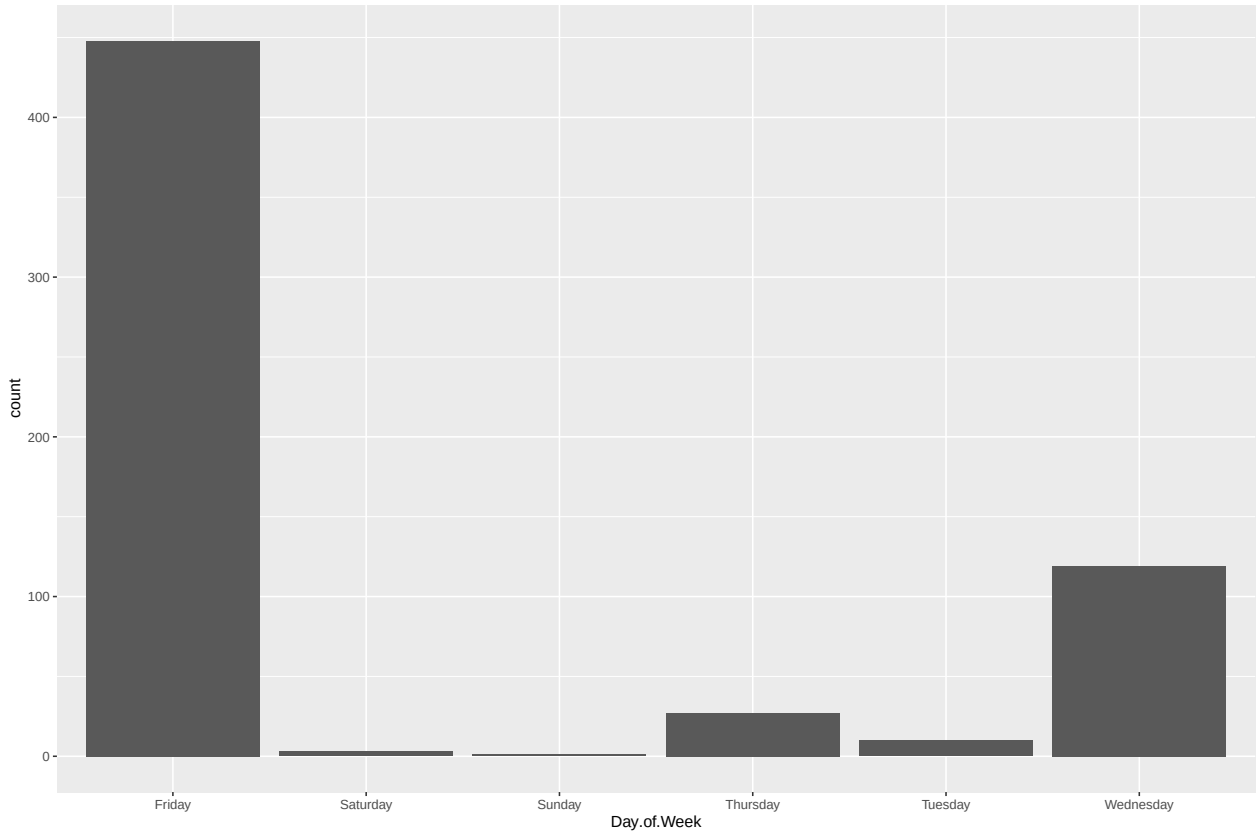
```
## [1] 608  18
```

```
colnames(movie.gross)
```

```
##  [1] "Day.of.Week"            "Director"               "Genre"
##  [4] "Movie.Title"            "Release.Date"           "Studio"
##  [7] "Adjusted.Gross...mill." "Budget...mill."         "Gross...mill."
## [10] "IMDb.Rating"            "MovieLens.Rating"       "Overseas...mill."
## [13] "Overseas."              "Profit...mill."         "Profit."
## [16] "Runtime..min."          "US...mill."             "Gross...US"
```

#Filtering the data

```
ggplot(data=movie.gross, aes(x=Day.of.Week)) + geom_bar()
```




```
filt <- (movie.gross$Genre == "action") | (movie.gross$Genre == "adventure") | (movie.gross$Genre == "a
filt2 <- (movie.gross$Studio == "Buena Vista Studios") | (movie.gross$Studio == "Fox") | (movie.gross$S

mov <- movie.gross[filt & filt2,]
head(mov)
```

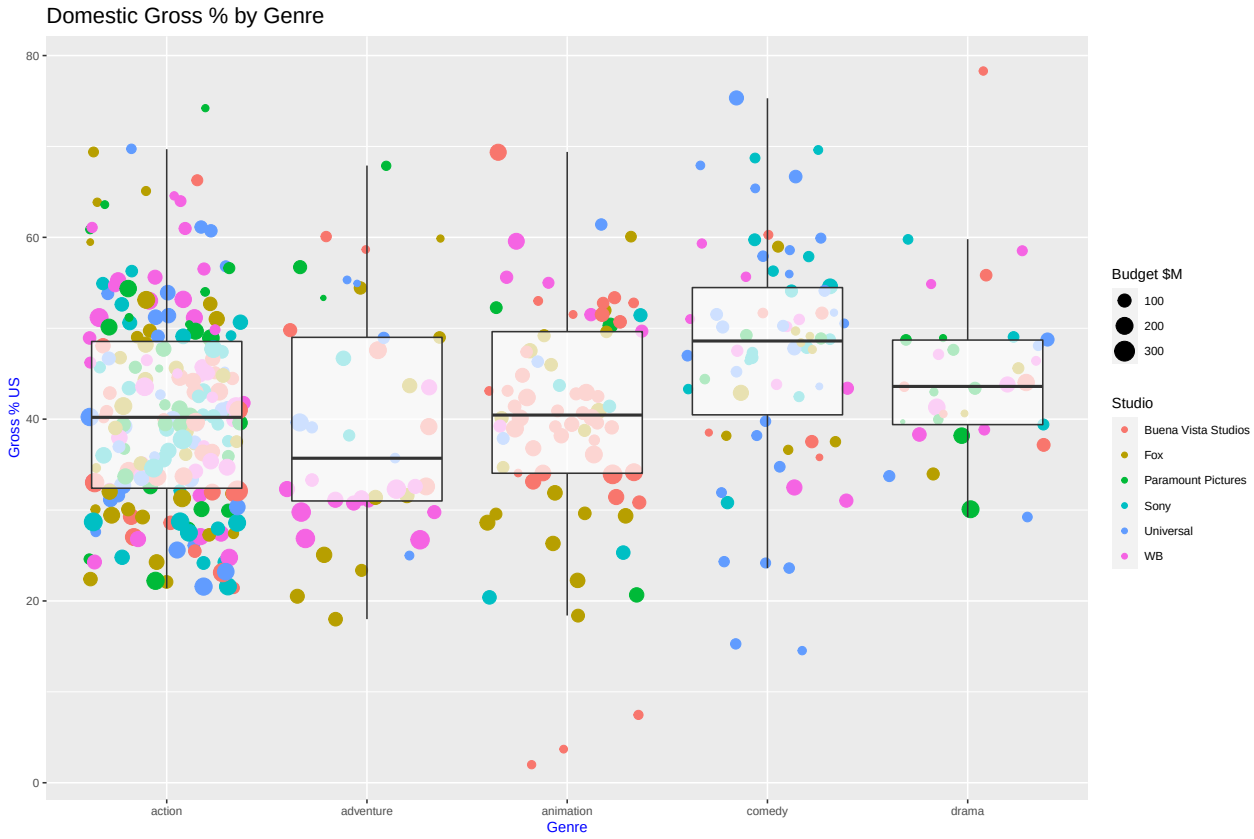
```
##   Day.of.Week               Director     Genre      Movie.Title Release.Date
```

```
## 1      Friday               Brad Bird    action     Tomorrowland   22/05/2015
## 2      Friday             Scott Waugh    action   Need for Speed   14/03/2014
## 4      Friday Phil Lord, Chris Miller    comedy   21 Jump Street   16/03/2012
## 5      Friday         Roland Emmerich    action White House Down   28/06/2013
## 6      Friday              David Ayer    action             Fury   17/10/2014
## 7    Thursday            Rob Marshall adventure   Into the Woods   25/12/2014
##                Studio Adjusted.Gross...mill. Budget...mill. Gross...mill.
## 1 Buena Vista Studios                  202.1            170         202.1
## 2 Buena Vista Studios                  204.2             66         203.3
## 4               Sony                   208.8             42         201.6
## 5               Sony                   209.7            150         205.4
## 6               Sony                   212.8             80         211.8
## 7 Buena Vista Studios                  213.9             50         212.9
##   IMDb.Rating MovieLens.Rating Overseas...mill. Overseas. Profit...mill.
## 1         6.7             3.26            111.9      55.4           32.1
## 2         6.6             2.97            159.7      78.6          137.3
## 4         7.2             3.62             63.1      31.3          159.6
## 5         8.0             3.65            132.3      64.4           55.4
## 6         5.8             2.85              126      59.5          131.8
## 7         6.0             3.16             84.9      39.9          162.9
##   Profit. Runtime..min. US...mill. Gross...US
## 1    18.9           130       90.2       44.6
## 2   208.0           132       43.6       21.4
## 4   380.0           109      138.4       68.7
## 5    36.9           131       73.1       35.6
## 6   164.8           134       85.8       40.5
## 7   325.8           125      128.0       60.1
```

```
colnames(mov) <- c("Day", "Director", "Genre", "Title", "ReleaseDate", "Studio", "AdjustedGrossMillion"
```

#Creating the chart

```
chart <- ggplot(data = mov, aes(x=Genre, y=GrossUS))
b <- chart + geom_jitter(aes(size = BudgetMillion, color = Studio)) + geom_boxplot(alpha = 0.7, outlier
b <- b +
  xlab("Genre") +
  ylab("Gross % US") +
  ggtitle("Domestic Gross % by Genre")
b <- b +
  theme(axis.title.x = element_text(size = 10, color="Blue"),
        axis.title.y = element_text(size = 10, color="Blue"),
        axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8),
        plot.title = element_text(size = 15),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),)
b$labels$size <- "Budget $M"
b
```


Domestic Gross % by Genre
Gross % US
80
60
40
20
0
action
adventure
animation
comedy
drama
Genre
Budget $M
100
200
300
Studio
Buena Vista Studios
Fox
Paramount Pictures
Sony
Universal
WB