Movie_Rating

Ashish Pal

October 7, 2017

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
library('caret')
```

```
## Loading required package: lattice
```

Frame

Data Visualization (EDA) for Movie Rating based on given Factors.

Acquire

```
setwd("C:/Users/Ashish/Desktop/GL/Day2/HomeWork/Movie_Rating Assg")

Movie_data = read.csv("Movie-ratings.csv", stringsAsFactors = FALSE)

Movie_data1 = Movie_data
```

Refine & Transform

```
str(Movie_data1)
```

```
## 'data.frame':    562 obs. of  6 variables:
##  $ Film                    : chr  "(500) Days of Summer " "10,000 B.C." "12 Rounds " "127 Ho
urs" ...
##  $ Genre                   : chr  "Comedy" "Adventure" "Action" "Adventure" ...
##  $ Rotten.Tomatoes.Ratings..: int  87 9 30 93 55 39 40 50 43 93 ...
##  $ Audience.Ratings..      : int  81 44 52 84 70 63 71 57 48 93 ...
##  $ Budget..million...      : int  8 105 20 18 20 200 30 32 28 8 ...
##  $ Year.of.release         : int  2009 2008 2009 2010 2009 2009 2008 2007 2011 2011 ...
```

```
#Check for missing value or NA value
colSums(is.na(Movie_data1))
```

```
##                      Film                     Genre
##                         0                         0
## Rotten.Tomatoes.Ratings..       Audience.Ratings..
##                         0                         0
##       Budget..million...          Year.of.release
##                         0                         0
```

```
colSums(Movie_data1 == '')
```

```
##                      Film                     Genre
##                         0                         0
## Rotten.Tomatoes.Ratings..       Audience.Ratings..
##                         0                         0
##       Budget..million...          Year.of.release
##                         0                         0
```

```
colSums(Movie_data1 == 0)
```

```
##                      Film                     Genre
##                         0                         0
## Rotten.Tomatoes.Ratings..       Audience.Ratings..
##                         3                         2
##       Budget..million...          Year.of.release
##                         8                         0
```

```
#This clearly shows there is no NA or missing value in our dataset
# But there are 8 Movies who's Budget is Zero which is practically impossible.
# Let's replace it with mean of Budget.


s = mean(Movie_data1$Budget..million...)
Movie_data1$Budget..million...[Movie_data1$Budget..million... == 0] = s


apply(Movie_data1,2, function(x) length(unique(x)))
```

```
##                      Film                     Genre
##                       562                         7
## Rotten.Tomatoes.Ratings..        Audience.Ratings..
##                        98                        74
##        Budget..million...           Year.of.release
##                        99                         5
```

```
##We can also see that Genre & Year.of.release is a Factor.Let's convert it into Factor

Movie_data1$Genre = as.factor(Movie_data1$Genre)
Movie_data1$Year.of.release = as.factor(Movie_data1$Year.of.release)

col_name = c ("Film","Genre","Critics.Ratings","Audience.Ratings","Budget.Million","Year")
colnames(Movie_data1) = col_name
str(Movie_data1)
```

```
## 'data.frame':    562 obs. of  6 variables:
##  $ Film            : chr  "(500) Days of Summer " "10,000 B.C." "12 Rounds " "127 Hours" ...
##  $ Genre           : Factor w/ 7 levels "Action","Adventure",..: 3 2 1 2 3 1 3 5 3 3 ...
##  $ Critics.Ratings : int  87 9 30 93 55 39 40 50 43 93 ...
##  $ Audience.Ratings: int  81 44 52 84 70 63 71 57 48 93 ...
##  $ Budget.Million  : num  8 105 20 18 20 200 30 32 28 8 ...
##  $ Year            : Factor w/ 5 levels "2007","2008",..: 3 2 3 4 3 3 2 1 5 5 ...
```

```
# Number of movies in each year in our dataset.
#2007
Movie_data1 %>% filter(Year == 2007) %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    79
```

```
#2008
Movie_data1 %>% filter(Year == 2008) %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   125
```

```
#2009
Movie_data1 %>% filter(Year == 2009) %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   116
```

```
#2010
Movie_data1 %>% filter(Year == 2010) %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   119
```

```
#2011
Movie_data1 %>% filter(Year == 2011) %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   123
```

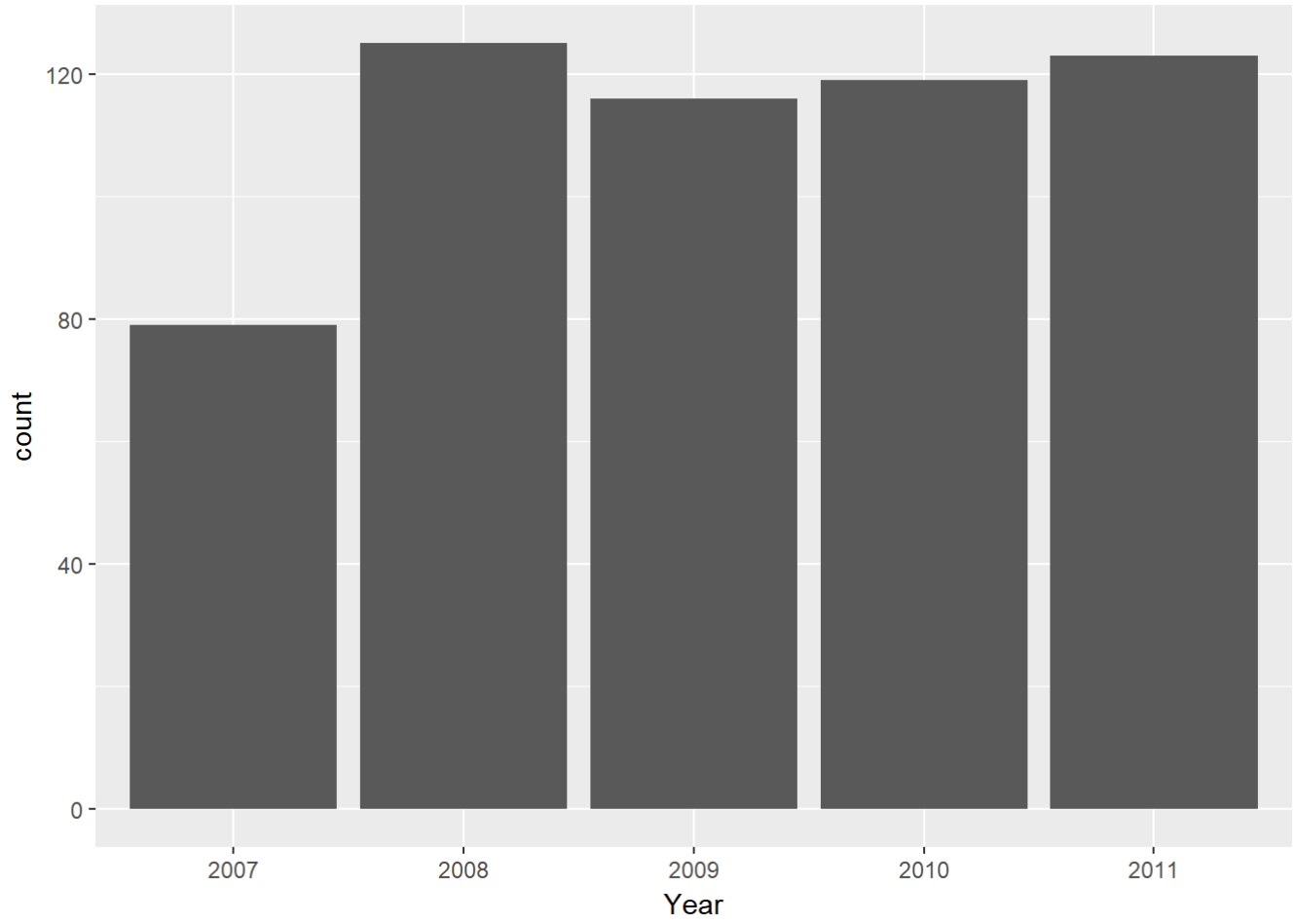
```
# Let's gather Ratings into one variable -
Movie_data2 = Movie_data1 %>% gather("Rating.Source","Rating",3:4) %>% arrange(Year)
str(Movie_data2)
```

```
## 'data.frame':    1124 obs. of  6 variables:
##  $ Film          : chr  "30 Days of Night" "88 Minutes" "Across the Universe" "Alien vs. Pred
ator -- Requiem" ...
##  $ Genre         : Factor w/ 7 levels "Action","Adventure",..: 5 4 6 5 7 1 3 3 3 7 ...
##  $ Budget.Million: num  32 30 45 40 100 150 61 6 21 20 ...
##  $ Year          : Factor w/ 5 levels "2007","2008",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Rating.Source : chr  "Critics.Ratings" "Critics.Ratings" "Critics.Ratings" "Critics.Rating
s" ...
##  $ Rating        : int  50 5 54 14 79 71 69 1 40 67 ...
```

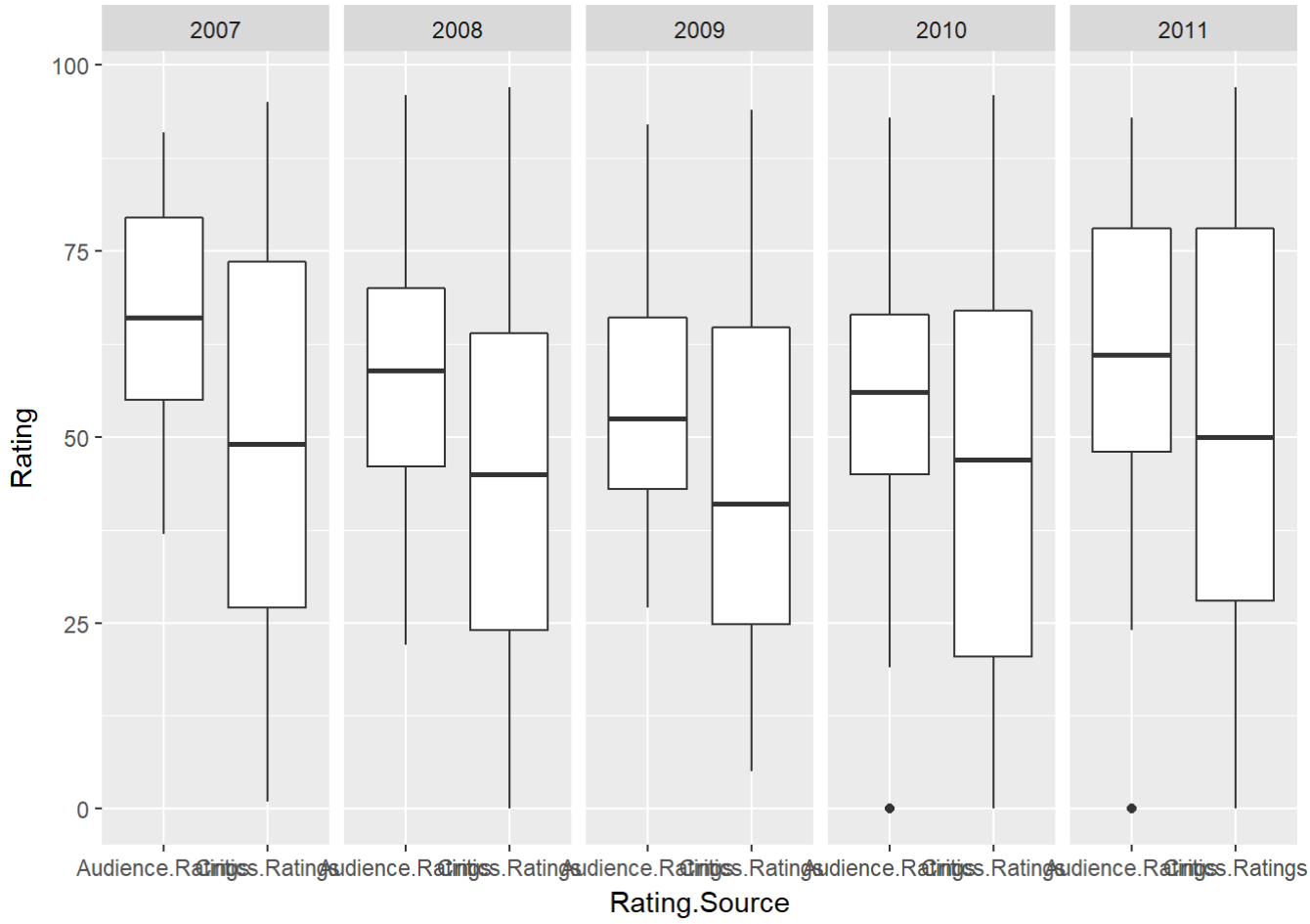
Explore

1. Movies Data analysis and checking outliers -

```
ggplot(Movie_data1) + aes(Year)+geom_bar()
```




```
ggplot(Movie_data2)+aes(Rating.Source,Rating) +geom_boxplot()+ facet_grid(~Year)
```



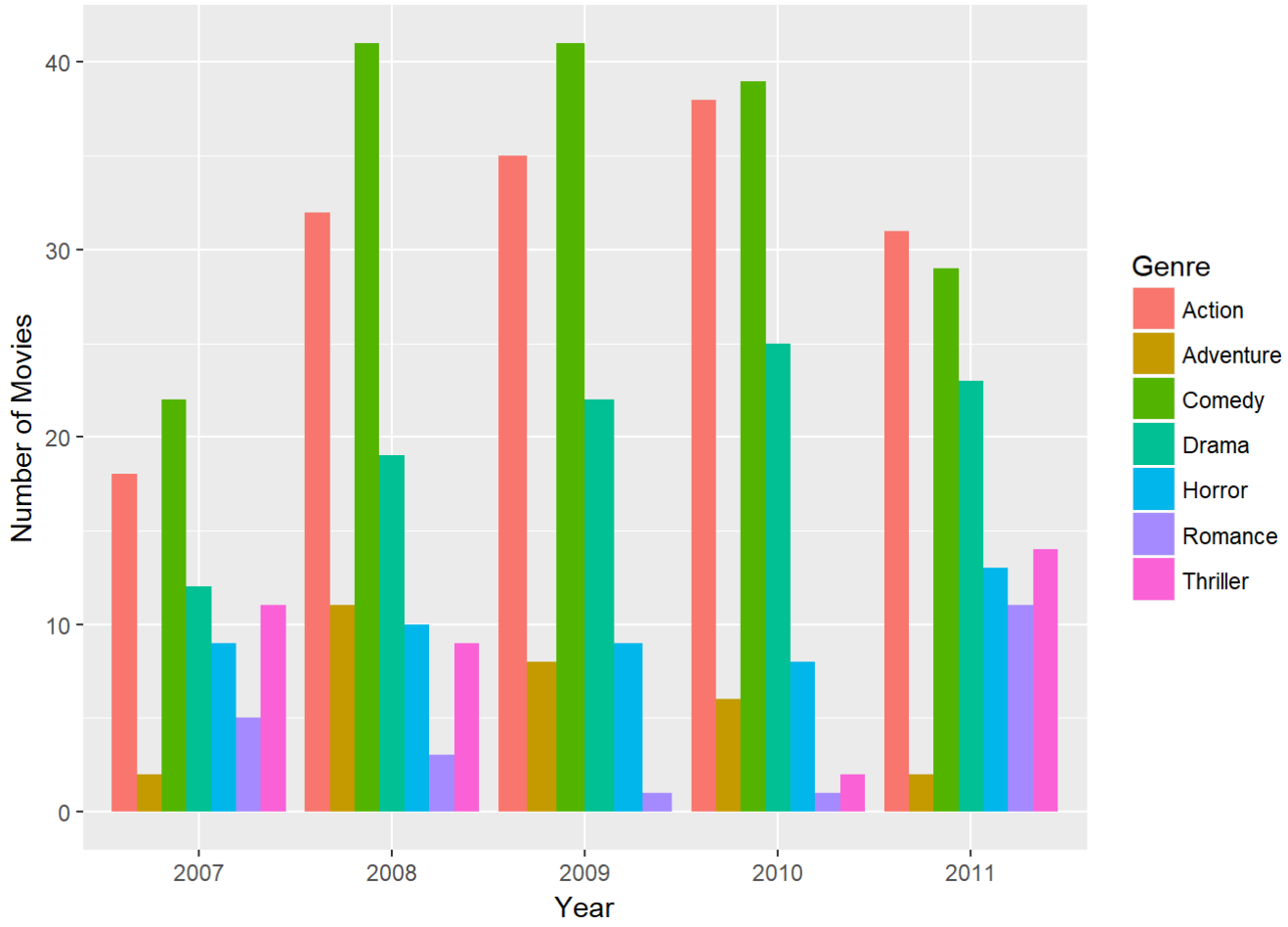
Infer 1-

This shows Year 2007 is having comparatively less movies than oher years in our Dataset.

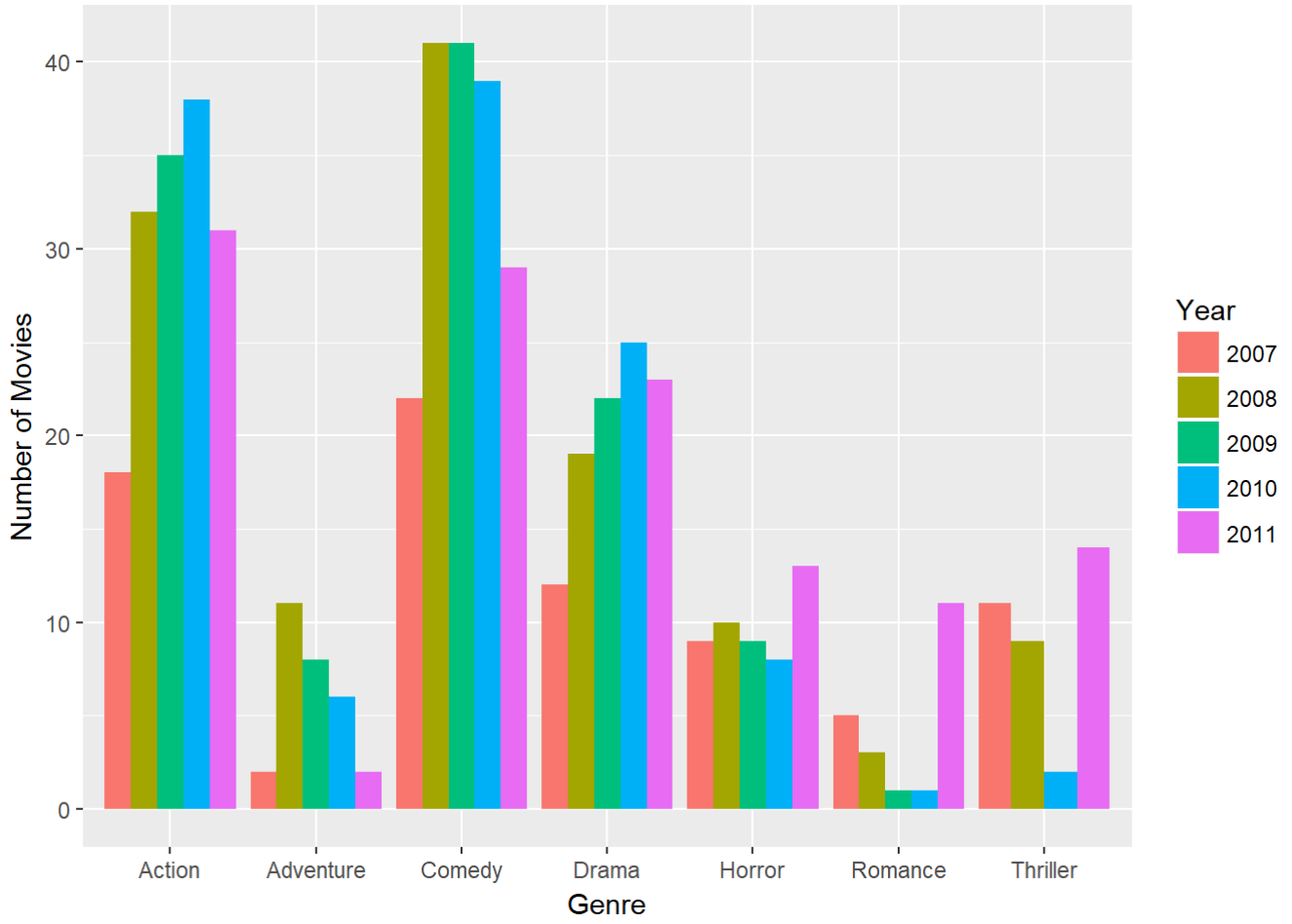
Another Important Observation is Audience Rating average is greater than Critics Rating. Also Critics Rating is spread over wide range compare to Audience Rating.

2 . Exploratory Analysis on Genre & Year

```
ggplot(Movie_data1, aes(x = Year, fill = Genre)) + geom_bar(position = "dodge") + scale_x_discre
te("Year") + scale_y_continuous("Number of Movies")
```




```
ggplot(Movie_data1, aes(x = Genre, fill = Year)) + geom_bar(position = "dodge") + scale_x_discre
te("Genre") + scale_y_continuous("Number of Movies")
```



Infer 2 -

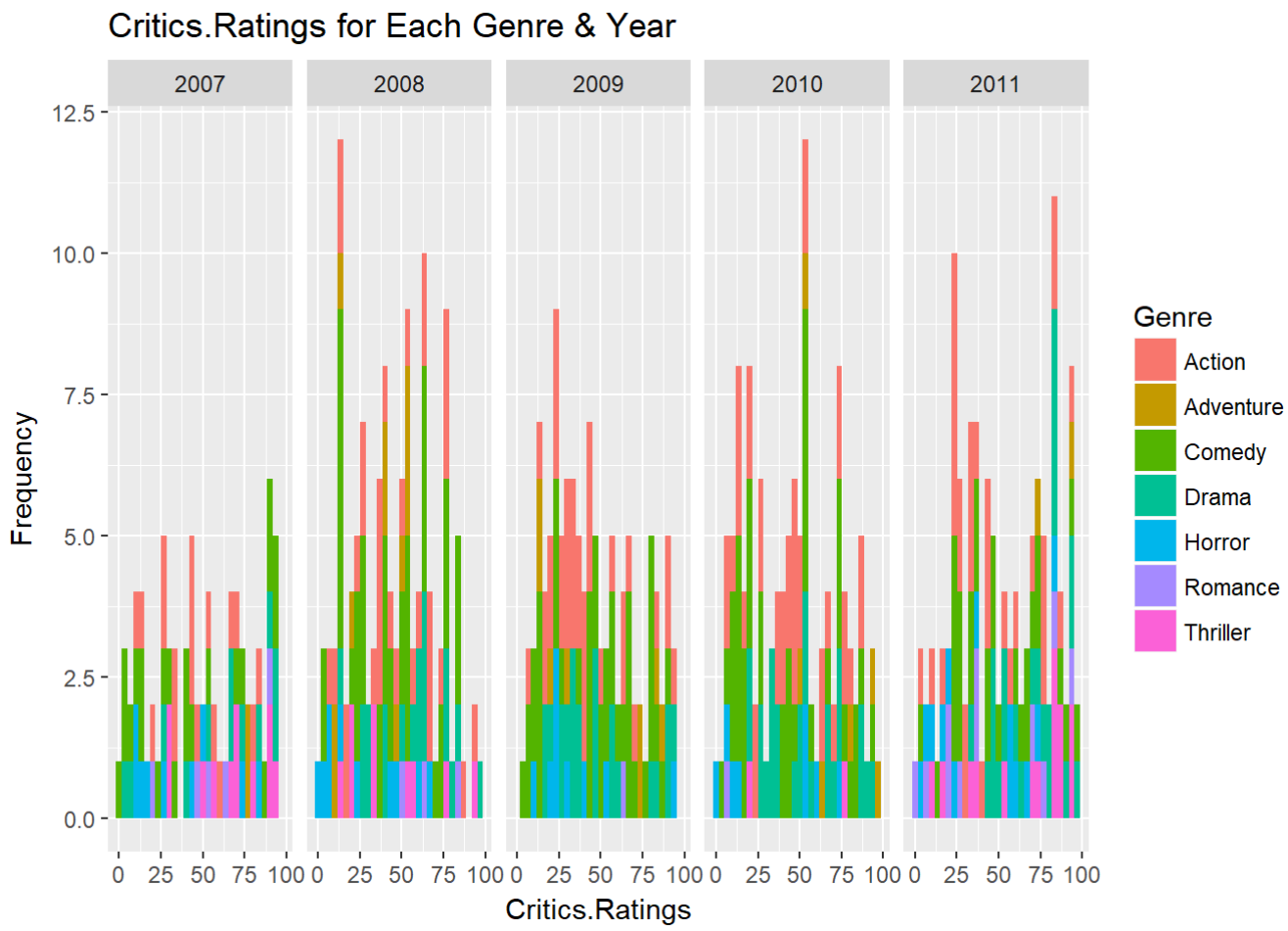
The above 2 Graphs shows that maximum Movies produced between year 2007 to 2011 are Comedy and Action followed by Drama and then Horror.

The Thriller and Romance movies number has increased in the year 2011.

3. Exploratory Analysis on Genre & Critics.Ratings / Audience.Ratings

```
ggplot(Movie_data1)+aes(x = Critics.Ratings,fill = Genre)+geom_histogram() + facet_grid(~Year)+
xlab("Critics.Ratings") +  ylab("Frequency") + ggtitle("Critics.Ratings for Each Genre & Year")
```

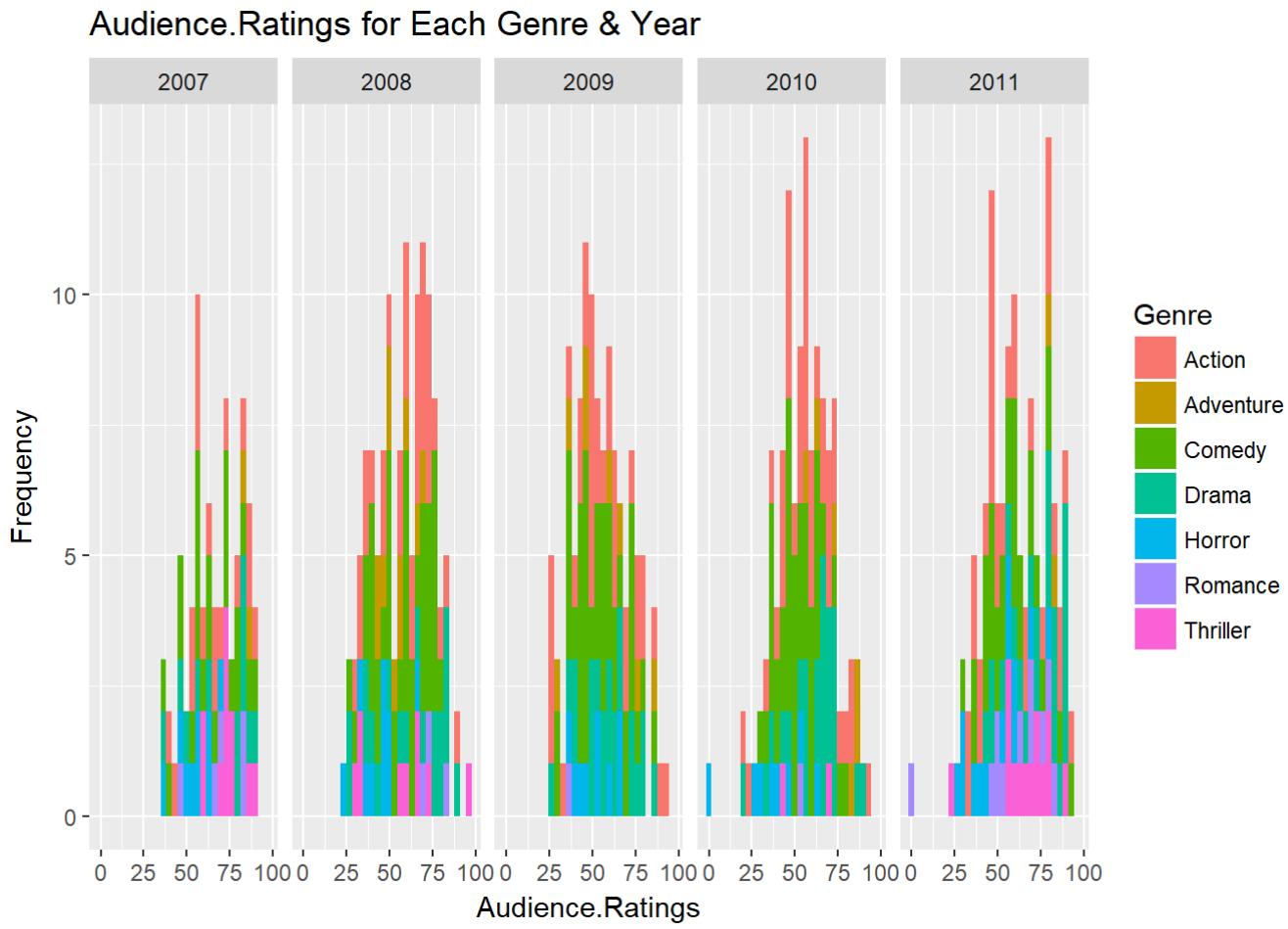
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#+scale_x_continuous(breaks=seq(20,75,10)) +
 # scale_y_continuous(breaks=seq(0,1700,250))

ggplot(Movie_data1)+aes(x = Audience.Ratings,fill = Genre)+geom_histogram() + facet_grid(~Year)+
xlab("Audience.Ratings") +  ylab("Frequency") + ggtitle("Audience.Ratings for Each Genre &
Year")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

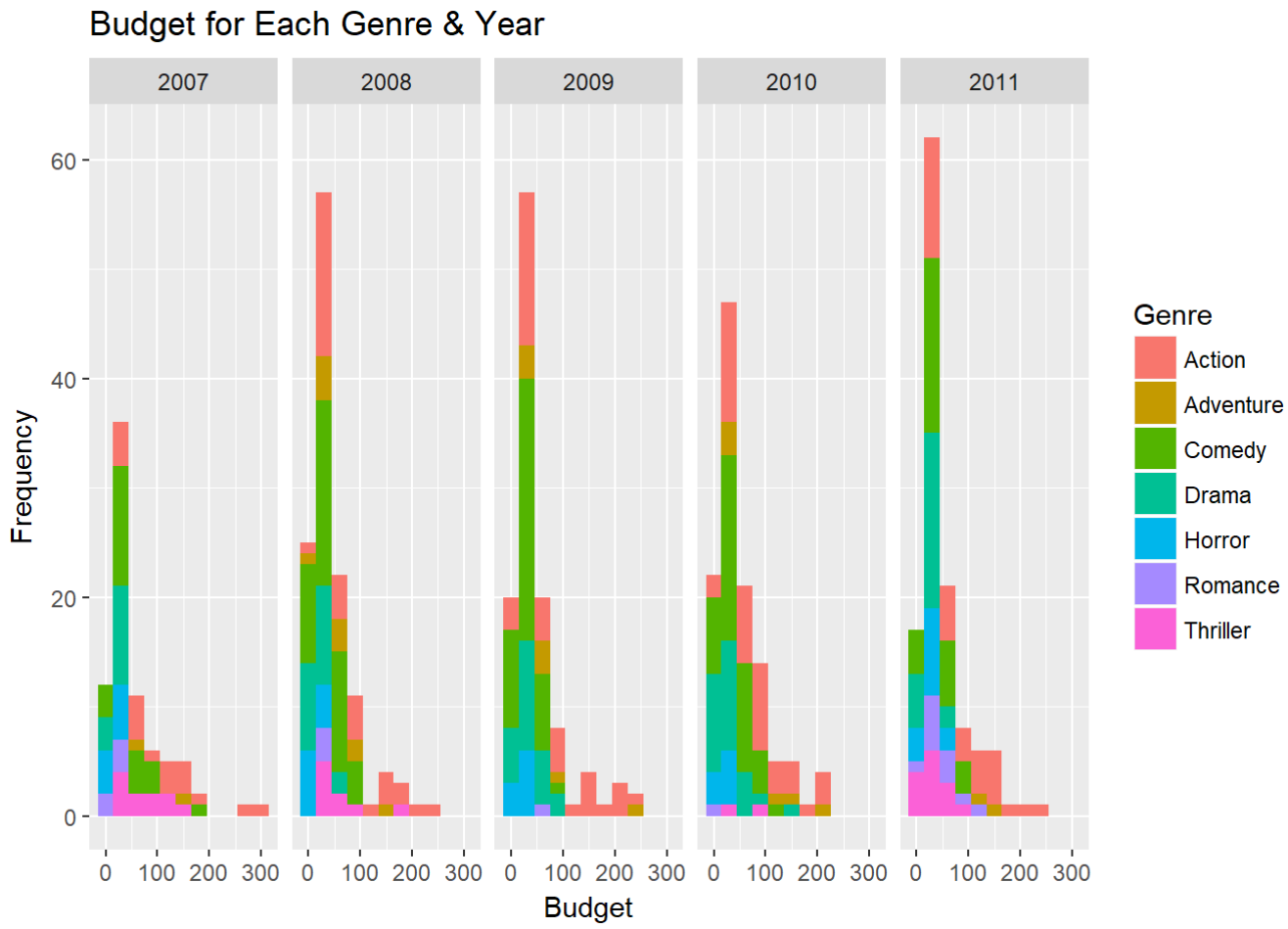


Infer 3 -

The Above 2 graphs clearly shows that Overall Movie ratings(Audience & Critic) has increased from Year 2007 to 2011.

4 .Exploratory Analysis on Genre & Budget & Year

```
ggplot(Movie_data1)+aes(x = Budget.Million,fill = Genre)+geom_histogram(binwidth = 30) + facet_g
rid(~Year)+
xlab("Budget") +  ylab("Frequency") + ggtitle("Budget for Each Genre & Year")
```



Infer 4 -

Majority of Movies are having Budget below $100 Million

High Budget movies are generally Action Movies

5. Exploratory Analysis on Genre & Critics.Ratings vs Audience Ratings

CEO's Vision

```
# Let's Analyse data statistically -
## How much Crtitics Ratings differ from Audience Ratings -

mean(Movie_data1$Critics.Ratings)
```

```
## [1] 47.40391
```

```
mean(Movie_data1$Audience.Ratings)
```

```
## [1] 58.83096
```

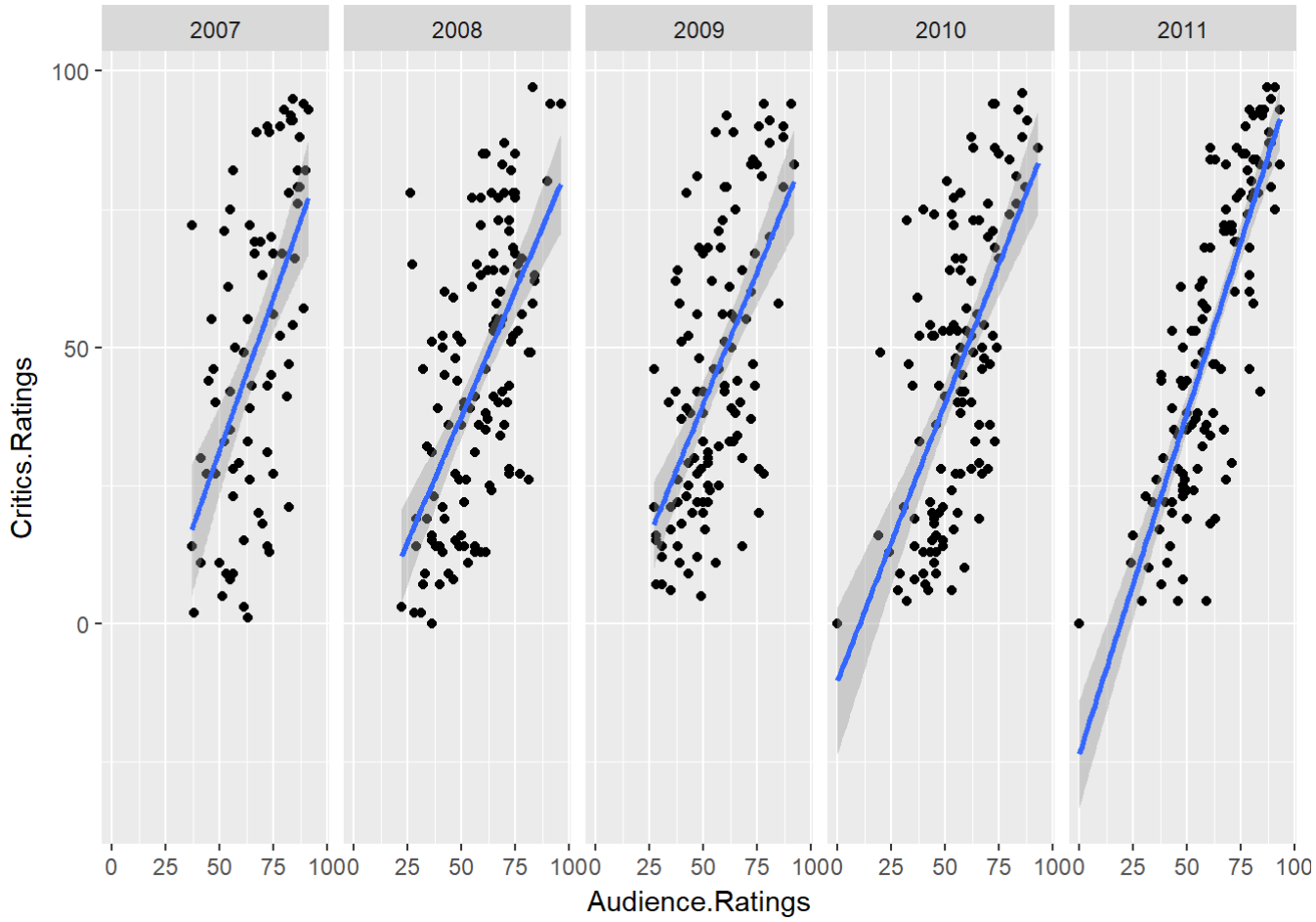
```
# As we can see overall Audience Rating's mean is greater than Crtics Rating's mean-
cov(Movie_data1$Critics.Ratings,Movie_data1$Audience.Ratings)
```

```
## [1] 290.7493
```

```
# It shows a strong Positive relation between two ratings
cor(Movie_data1$Critics.Ratings,Movie_data1$Audience.Ratings)
```

```
## [1] 0.6546554
```

```
#############################################################################
Movie_data1  %>%
  ggplot() +aes(x = Audience.Ratings, y = Critics.Ratings) + geom_point() + geom_smooth(method =
 "lm") +facet_grid(~Year)
```




```
ggplot(Movie_data1)+aes(x = Critics.Ratings,y = Audience.Ratings, col = Genre)+geom_point() + fa
cet_grid(~Year)
```


```
ggplot(Movie_data1)+aes(x = Critics.Ratings,y = Audience.Ratings, col = Genre)+geom_point() + fa
cet_grid(~Year) +
geom_smooth(method = "lm",se =FALSE) +
geom_smooth(aes(group = 1), method = "lm", se = FALSE, linetype = 2)
```


```
ggplot(Movie_data1)+aes(x = Critics.Ratings,y = Audience.Ratings, col = Genre)+geom_point() + fa
cet_grid(~Year)
```

Infer 5 -

1.This clearly shows thier is direct Positive relationship(Correlation) between Audience rating and critings Rating.

2.The graph also reflects that with increasing year the Correlation between both ratings has increased.

3.There is no clear view for Genre and Ratings

4. The Year 2011 rating is more linear than previous ratings which reflect Audience & Critics Ratings are become more similar.

6. Exploratory Analysis on Genre & Critics.Ratings vs Audience Ratings

```
# Let's make Budget into a category Small, Medium and Big Budget.
Movie_data3 = Movie_data1
range(Movie_data3$Budget.Million)
```

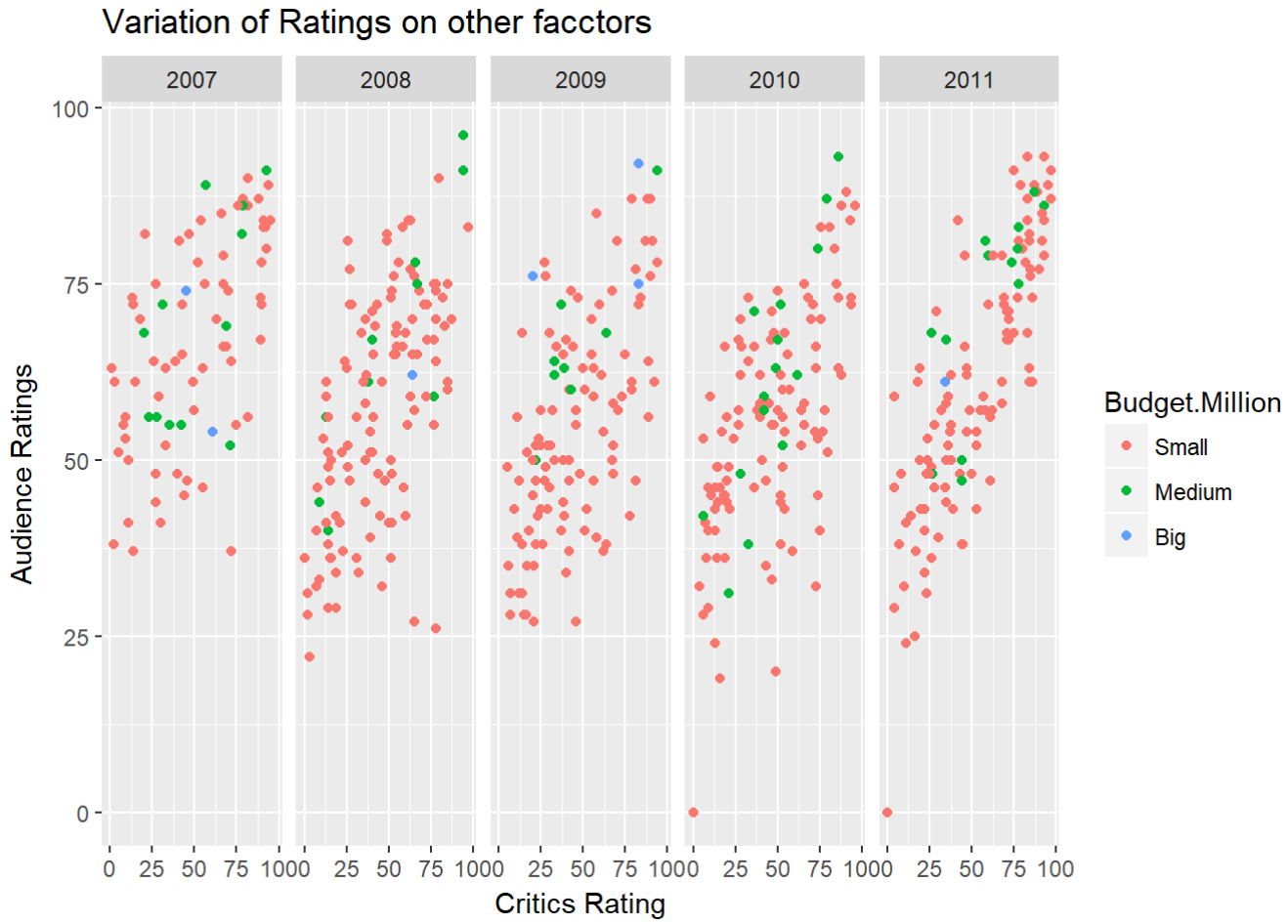
```
## [1]   1 300
```

```
Movie_data3$Budget.Million = cut(Movie_data3$Budget.Million , breaks =  c(0,100,200,301),labels
= c("Small","Medium","Big"))

str(Movie_data3)
```

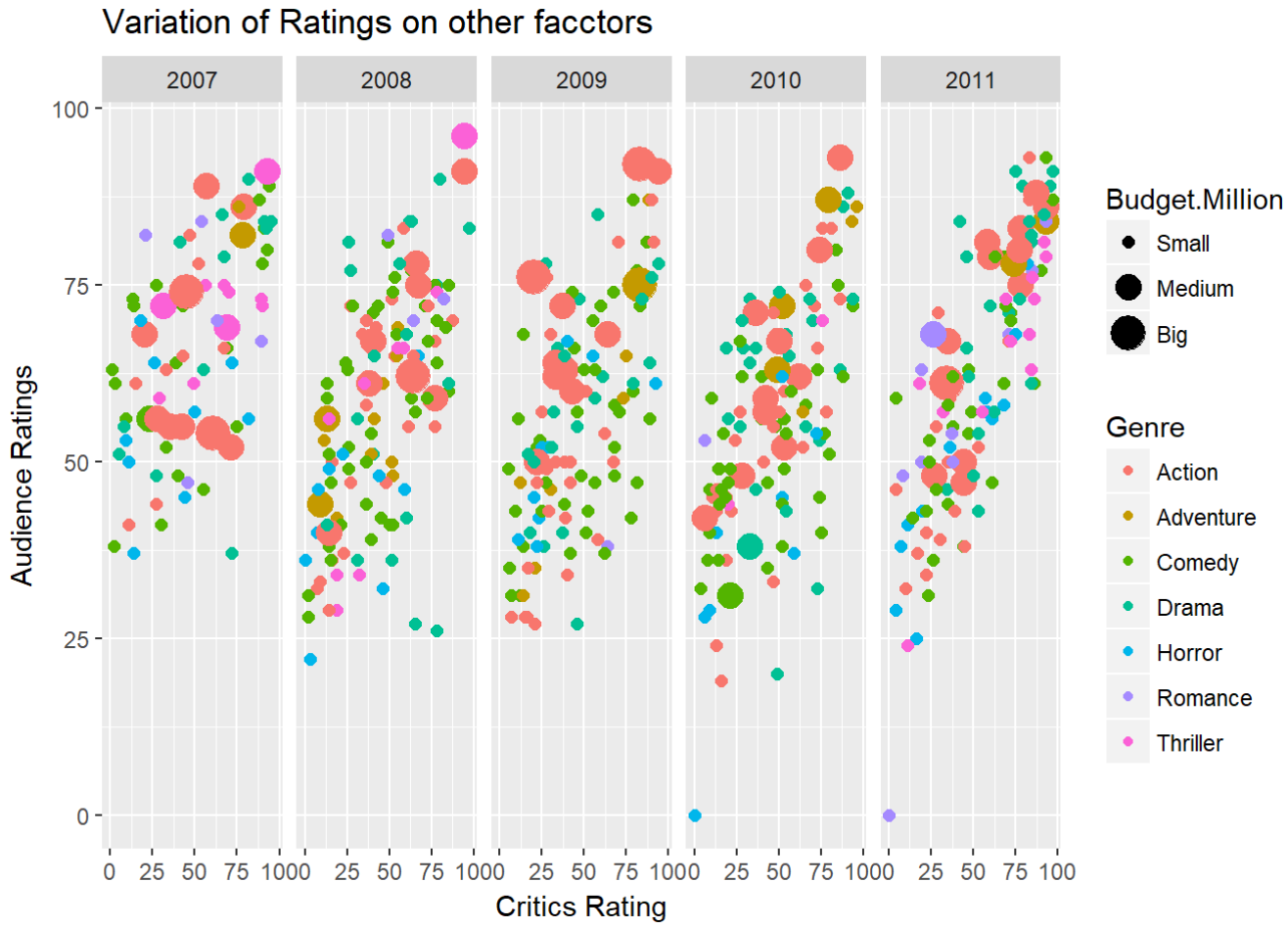
```
## 'data.frame':    562 obs. of  6 variables:
##  $ Film            : chr  "(500) Days of Summer " "10,000 B.C." "12 Rounds " "127 Hours" ...
##  $ Genre           : Factor w/ 7 levels "Action","Adventure",..: 3 2 1 2 3 1 3 5 3 3 ...
##  $ Critics.Ratings : int  87 9 30 93 55 39 40 50 43 93 ...
##  $ Audience.Ratings: int  81 44 52 84 70 63 71 57 48 93 ...
##  $ Budget.Million  : Factor w/ 3 levels "Small","Medium",..: 1 2 1 1 1 2 1 1 1 1 ...
##  $ Year            : Factor w/ 5 levels "2007","2008",..: 3 2 3 4 3 3 2 1 5 5 ...
```

```
ggplot(Movie_data3)+aes(x = Critics.Ratings,y = Audience.Ratings, col = Budget.Million )+geom_po
int() + facet_grid(~Year)+
xlab("Critics Rating") +  ylab("Audience Ratings") + ggtitle("Variation of Ratings on other facc
tors")
```



```
ggplot(Movie_data3)+aes(x = Critics.Ratings,y = Audience.Ratings, size = Budget.Million , col =
Genre)+geom_point() + facet_grid(~Year)+
xlab("Critics Rating") +  ylab("Audience Ratings") + ggtitle("Variation of Ratings on other facc
tors")
```

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



Infer 6

Apart from Infer4 this shows Budget of the Movie do not influence Crtics Rating much but have impact on Audience rating.