### Required R libraries

## Loading required package: plyr

## Loading required package: ggplot2

### Objective

To analyse a dataset of movie rating

### Starting Point

The dataset if obtained from the below site: <http://vincentarelbundock.github.io/> Dataset reference: [Movie information and user ratings from IMDB.com.](http://vincentarelbundock.github.io/Rdatasets/csv/ggplot2/movies.csv)

### Obtaining the data

1. Download the data in a csv sheet

theURL <- "http://vincentarelbundock.github.io/Rdatasets/csv/ggplot2/movies.csv";

movie\_data <- read.table(file = theURL, header = TRUE, sep = ",");

head(movie\_data);

#State column is named as 'X". Changing the column name to US\_State

colnames(movie\_data)[1] <- "RefNo";

head(movie\_data);

#Write the data to a CSV

write.table(movie\_data, file = "movie\_data.csv", sep = ",", row.names = FALSE);

#The saved file is uploaded to githib and below is the URL which will be used hereafter.

#https://raw.githubusercontent.com/arunk13/MSDA-Assignments/master/BridgeCourse/Week5/movie\_data.csv

### Raw Data

## RefNo title year length budget rating votes r1 r2

## 1 1 $ 1971 121 NA 6.4 348 4.5 4.5

## 2 2 $1000 a Touchdown 1939 71 NA 6.0 20 0.0 14.5

## 3 3 $21 a Day Once a Month 1941 7 NA 8.2 5 0.0 0.0

## 4 4 $40,000 1996 70 NA 8.2 6 14.5 0.0

## 5 5 $50,000 Climax Show, The 1975 71 NA 3.4 17 24.5 4.5

## 6 6 $pent 2000 91 NA 4.3 45 4.5 4.5

## r3 r4 r5 r6 r7 r8 r9 r10 mpaa Action Animation Comedy

## 1 4.5 4.5 14.5 24.5 24.5 14.5 4.5 4.5 0 0 1

## 2 4.5 24.5 14.5 14.5 14.5 4.5 4.5 14.5 0 0 1

## 3 0.0 0.0 0.0 24.5 0.0 44.5 24.5 24.5 0 1 0

## 4 0.0 0.0 0.0 0.0 0.0 0.0 34.5 45.5 0 0 1

## 5 0.0 14.5 14.5 4.5 0.0 0.0 0.0 24.5 0 0 0

## 6 4.5 14.5 14.5 14.5 4.5 4.5 14.5 14.5 0 0 0

## Drama Documentary Romance Short

## 1 1 0 0 0

## 2 0 0 0 0

## 3 0 0 0 1

## 4 0 0 0 0

## 5 0 0 0 0

## 6 1 0 0 0

### Cleaning and subsetting

#First column can be removed.

movie\_data <- movie\_data\_raw[-1];

+Summary

summary(movie\_data);

## title year length

## Alice in Wonderland : 7 Min. :1893 Min. : 1.00

## Three Musketeers, The : 7 1st Qu.:1958 1st Qu.: 74.00

## Midsummer Night's Dream, A: 6 Median :1983 Median : 90.00

## Skin Deep : 6 Mean :1976 Mean : 82.34

## Underground : 6 3rd Qu.:1997 3rd Qu.: 100.00

## Anna Karenina : 5 Max. :2005 Max. :5220.00

## (Other) :58751

## budget rating votes r1

## Min. : 0 Min. : 1.000 Min. : 5.0 Min. : 0.000

## 1st Qu.: 250000 1st Qu.: 5.000 1st Qu.: 11.0 1st Qu.: 0.000

## Median : 3000000 Median : 6.100 Median : 30.0 Median : 4.500

## Mean : 13412513 Mean : 5.933 Mean : 632.1 Mean : 7.014

## 3rd Qu.: 15000000 3rd Qu.: 7.000 3rd Qu.: 112.0 3rd Qu.: 4.500

## Max. :200000000 Max. :10.000 Max. :157608.0 Max. :100.000

## NA's :53573

## r2 r3 r4 r5

## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.000

## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 4.500

## Median : 4.500 Median : 4.500 Median : 4.500 Median : 4.500

## Mean : 4.022 Mean : 4.721 Mean : 6.375 Mean : 9.797

## 3rd Qu.: 4.500 3rd Qu.: 4.500 3rd Qu.: 4.500 3rd Qu.: 14.500

## Max. :84.500 Max. :84.500 Max. :100.000 Max. :100.000

##

## r6 r7 r8 r9

## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.000

## 1st Qu.: 4.50 1st Qu.: 4.50 1st Qu.: 4.50 1st Qu.: 4.500

## Median :14.50 Median : 14.50 Median : 14.50 Median : 4.500

## Mean :13.04 Mean : 15.55 Mean : 13.88 Mean : 8.954

## 3rd Qu.:14.50 3rd Qu.: 24.50 3rd Qu.: 24.50 3rd Qu.: 14.500

## Max. :84.50 Max. :100.00 Max. :100.00 Max. :100.000

##

## r10 mpaa Action Animation

## Min. : 0.00 :53864 Min. :0.00000 Min. :0.00000

## 1st Qu.: 4.50 NC-17: 16 1st Qu.:0.00000 1st Qu.:0.00000

## Median : 14.50 PG : 528 Median :0.00000 Median :0.00000

## Mean : 16.85 PG-13: 1003 Mean :0.07974 Mean :0.06277

## 3rd Qu.: 24.50 R : 3377 3rd Qu.:0.00000 3rd Qu.:0.00000

## Max. :100.00 Max. :1.00000 Max. :1.00000

##

## Comedy Drama Documentary Romance

## Min. :0.0000 Min. :0.000 Min. :0.00000 Min. :0.0000

## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.00000 1st Qu.:0.0000

## Median :0.0000 Median :0.000 Median :0.00000 Median :0.0000

## Mean :0.2938 Mean :0.371 Mean :0.05906 Mean :0.0807

## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:0.00000 3rd Qu.:0.0000

## Max. :1.0000 Max. :1.000 Max. :1.00000 Max. :1.0000

##

## Short

## Min. :0.0000

## 1st Qu.:0.0000

## Median :0.0000

## Mean :0.1609

## 3rd Qu.:0.0000

## Max. :1.0000

##

* Out of 58788 records, 53573 record does not have information about the budget. I will remove the budget column from my dataset, as i feel that as the most of the observation is missing this variable data, hence it will not help me infer any understanding.

movie\_data <- movie\_data[-4];

Similarly 53864 movies are missing mpaa ratings. Lets find out the types of movies which are missing the mpaa ratings.

## year freq

## 1 2005 289

## 2 2004 1596

## 3 2003 1747

## 4 2002 1667

## 5 2001 1611

### Binning observation into decades of year

Number of observation with missing year information : 0

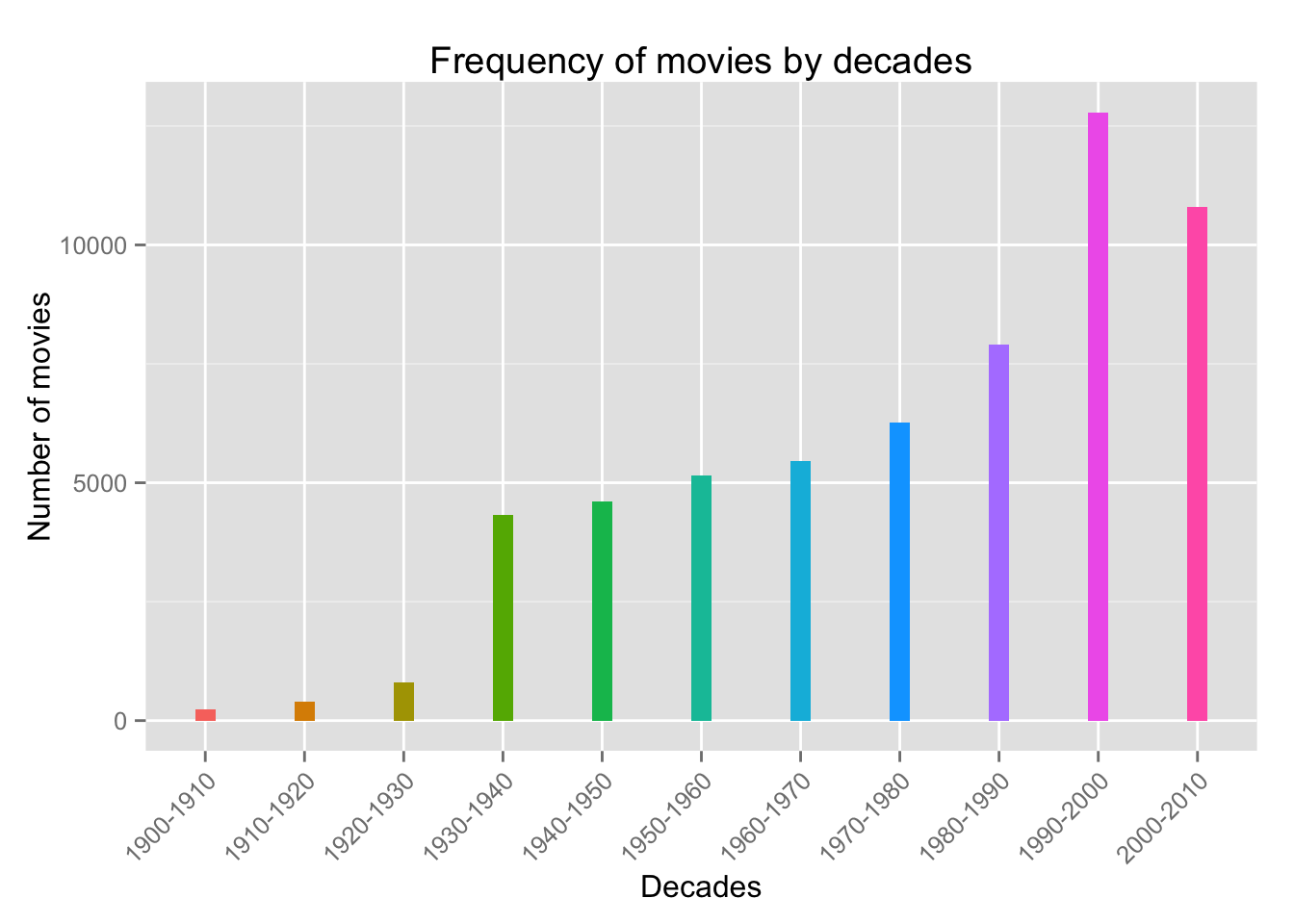
For ease of analysis, we will subset the data from 1900 - till date.

Number of movies in by decade

freq\_decade\_movie <- count(movie\_data, c("decade"));

g\_movie <- ggplot(data = freq\_decade\_movie, aes(y = freq, x = decade, fill = decade));

g\_movie + geom\_bar(stat = "identity", width = 0.2, , position = "identity") + guides(fill = FALSE) + xlab("Decades") + ylab("Number of movies") + ggtitle("Frequency of movies by decades") + theme(axis.text.x = element\_text(angle = 45, hjust = 1));



### A Study on the popularity of movies

Note: The dataset is not in a structure where we can compare the distribution of the ratings for various genres. Hence, the data frame is first melted using reshape2 package

require(reshape2);

## Loading required package: reshape2

movie\_data\_sub <- movie\_data[, c(1,2,4,5,17,18,19,20,21,22)];

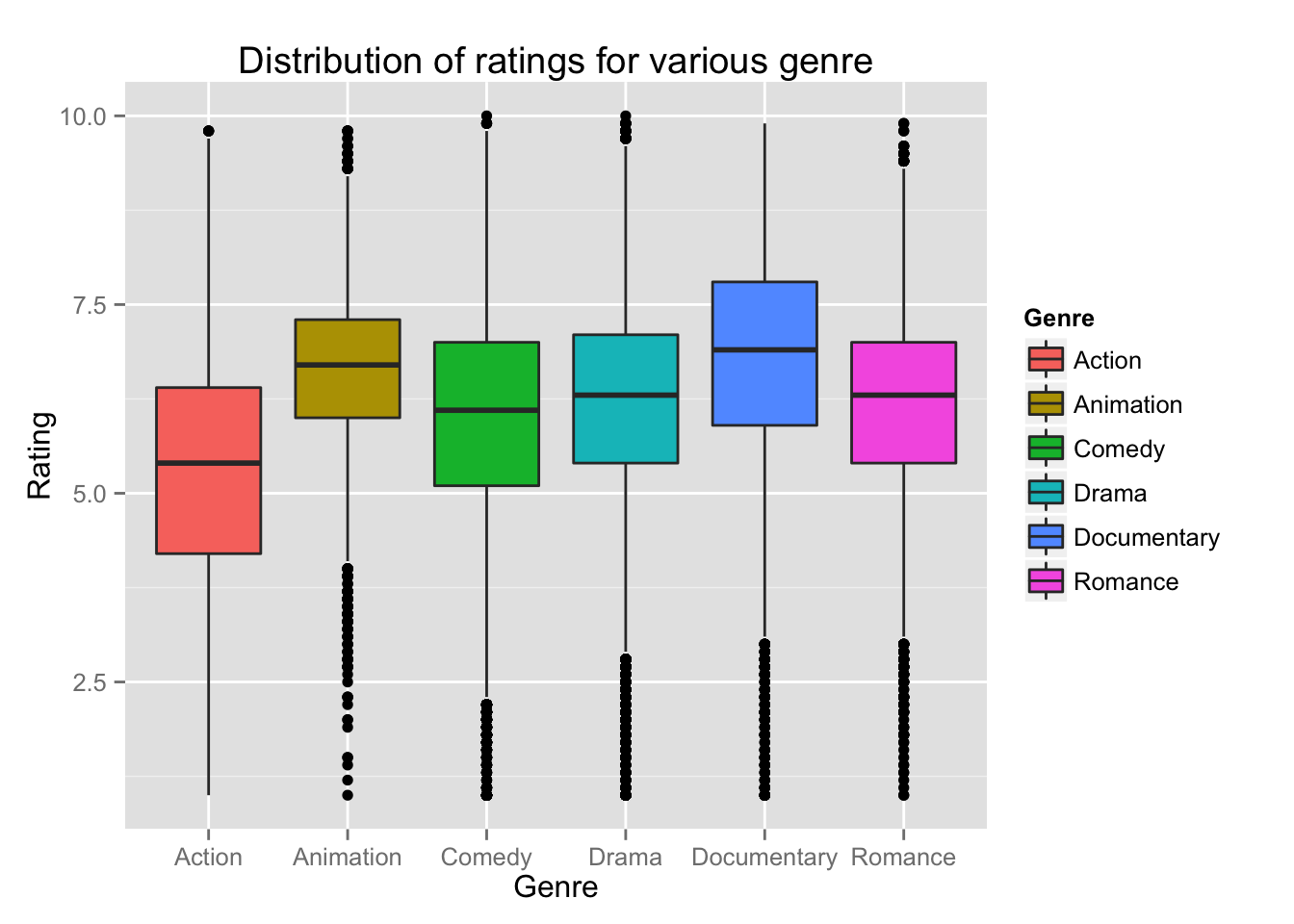
movie\_data\_sub <- melt(movie\_data\_sub, c(1,2,3,4));

names(movie\_data\_sub)[5] <- c("Genre");

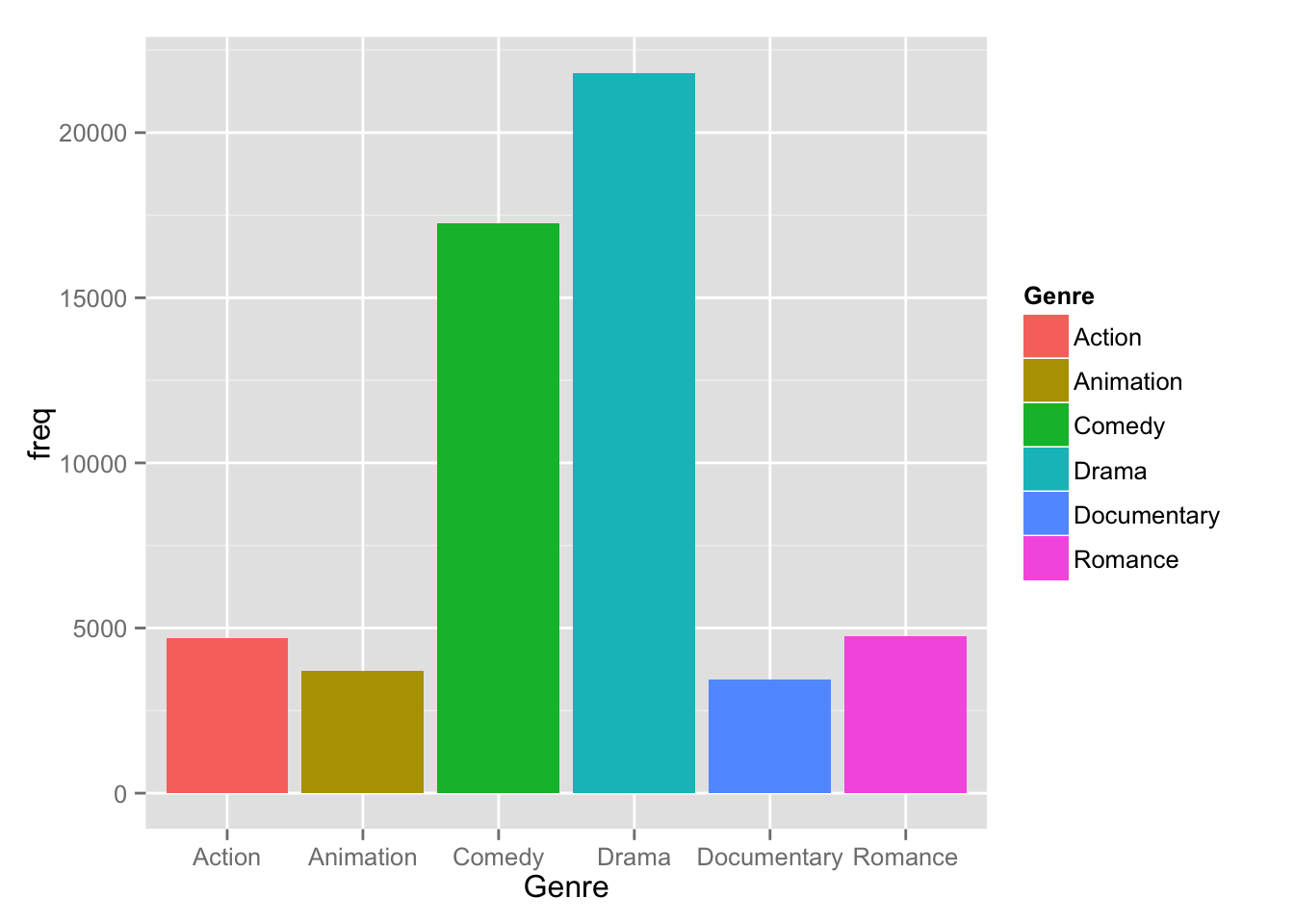
movie\_data\_sub <- subset(movie\_data\_sub, value == 1);

g\_genre <- ggplot(data = movie\_data\_sub, aes(x = Genre, y = rating, fill = Genre));

g\_genre + geom\_boxplot() + xlab("Genre") + ylab("Rating") + ggtitle("Distribution of ratings for various genre");



Before inferring popularity of each genre, we will also have to see the frequency of movies in each genre.



### Evolution of animation movie

Summary of ratings for animation movie

summary(movie\_anim$rating);

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 1.000 6.000 6.700 6.584 7.300 9.800

### Probability

Probability(animation movie) = total number of animation movie/total movie in sample space = 0.0628

Probability(movie rating > 6 and movie being animation) = 0.0454

Probability of a movie getting rating greater than 6 provided movie is animation movie =

0.7228 = 72.2764 %