Data Analysis Project

James Tran

May 8, 2018

The area of focus on this project is to figure out how many people watch movies based on their ratings.So many great movies are missed due to poor reviews based off of personal preference. I want to figure out how many people believe these reviews and base their movie selections off of ratings. I will be using an IMDB movie data set that I found on Kaggle. This dataset contains roughly 5000 movies.

## Load the required libraries

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.4

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(psych)

## Warning: package 'psych' was built under R version 3.4.4

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(readxl)  
movie <- read\_excel("C:/Users/MSSA/Desktop/movie.xlsx")

## View current data

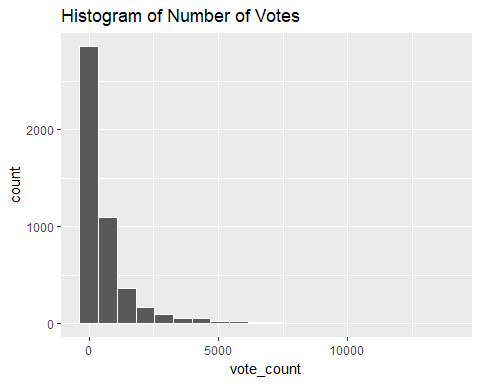
summary(movie)

## budget genres homepage   
## Min. : 0 Length:4803 Length:4803   
## 1st Qu.: 790000 Class :character Class :character   
## Median : 15000000 Mode :character Mode :character   
## Mean : 29045040   
## 3rd Qu.: 40000000   
## Max. :380000000   
##   
## id keywords original\_language original\_title   
## Min. : 5 Length:4803 Length:4803 Length:4803   
## 1st Qu.: 9014 Class :character Class :character Class :character   
## Median : 14629 Mode :character Mode :character Mode :character   
## Mean : 57166   
## 3rd Qu.: 58611   
## Max. :459488   
##   
## overview popularity production\_companies  
## Length:4803 Min. : 0.000 Length:4803   
## Class :character 1st Qu.: 4.668 Class :character   
## Mode :character Median : 12.922 Mode :character   
## Mean : 21.492   
## 3rd Qu.: 28.314   
## Max. :875.581   
##   
## production\_countries release\_date revenue   
## Length:4803 Min. :1916-09-04 00:00:00 Min. :0.000e+00   
## Class :character 1st Qu.:1999-07-14 00:00:00 1st Qu.:0.000e+00   
## Mode :character Median :2005-10-03 00:00:00 Median :1.917e+07   
## Mean :2002-12-27 23:45:54 Mean :8.226e+07   
## 3rd Qu.:2011-02-16 00:00:00 3rd Qu.:9.292e+07   
## Max. :2017-02-03 00:00:00 Max. :2.788e+09   
## NA's :1   
## runtime spoken\_languages status tagline   
## Min. : 0.0 Length:4803 Length:4803 Length:4803   
## 1st Qu.: 94.0 Class :character Class :character Class :character   
## Median :103.0 Mode :character Mode :character Mode :character   
## Mean :106.9   
## 3rd Qu.:118.0   
## Max. :338.0   
## NA's :2   
## title vote\_average vote\_count   
## Length:4803 Min. : 0.000 Min. : 0.0   
## Class :character 1st Qu.: 5.600 1st Qu.: 54.0   
## Mode :character Median : 6.200 Median : 235.0   
## Mean : 6.092 Mean : 690.2   
## 3rd Qu.: 6.800 3rd Qu.: 737.0   
## Max. :10.000 Max. :13752.0   
##

## Displaying data in order to see how many votes were counted.

On average you can see that most movies contain betwen 1,000 - 3,000 votes. With some movies rarely exceeding the 3,000 vote mark.

ggplot(aes(x = vote\_count), data = movie) + geom\_histogram(bins = 20, color = 'white') + ggtitle('Histogram of Number of Votes')



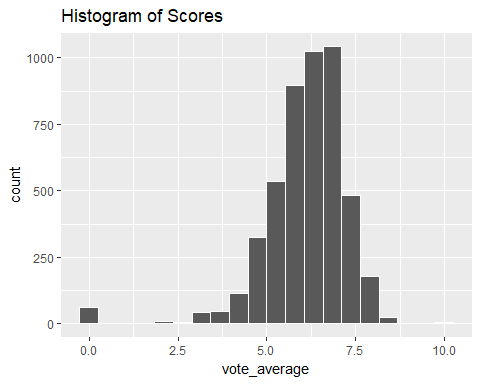
summary(movie$vote\_count)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 54.0 235.0 690.2 737.0 13752.0

## Display the data in order to see the average ratings each movie received.

On a scale of 0-10 you can see most movies received an average rating of 5-7. With the low being 0.0 and high being 8.

ggplot(aes(x = vote\_average), data = movie) + geom\_histogram(bins = 20, color = 'white') + ggtitle('Histogram of Scores')



summary(movie$vote\_average)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 5.600 6.200 6.092 6.800 10.000

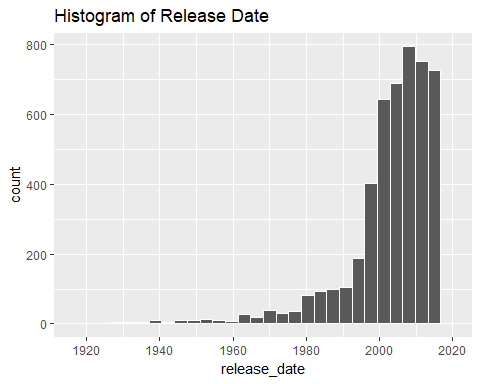
## Displaying the average movie release date.

As you can see most movies were released after the year 2000.

ggplot(aes(x = release\_date), data = movie) + geom\_histogram(color='white') +  
 ggtitle('Histogram of Release Date')

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

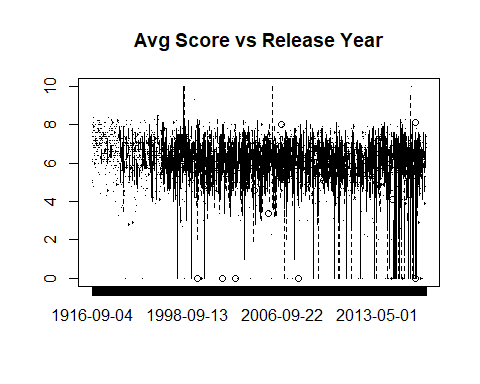
## Warning: Removed 1 rows containing non-finite values (stat\_bin).



## Displaying the highest scoring movie compared to the year

The movies with the highest average scores were produced in the early 2000’s with the recent movies declining in ratings.

boxplot(vote\_average ~ release\_date, data=movie, col='indianred')  
title("Avg Score vs Release Year")



Conclusion The most important factor that will affect a movies score is the duration. The longer the movie the higher the score. The number of critic reviews is important. The more reviews a movie gets the higher the score will be. The year in which a movie is released does not inherently affect a movies score.