Sri- Project 1

sriyoda

January 25, 2016

I performed an analysis of Box Office revenue. I obtained the data from imdb.com and boxofficemojo.com.

Loading Libraries and Tidying Data

```
library(reshape2)
library(ggplot2)
library(dplyr)
library(stringr)
library(car)
library(manipulate)
library(gridExtra)
library(cowplot)

# Loading Data
a <- read.csv('~/Datasets/movies.csv', stringsAsFactors = F)
b <- read.csv('~/Datasets/ratings.csv', stringsAsFactors = F)
c <- read.csv('~/Datasets/BoxOfficeMojo.csv', stringsAsFactors = F)
d <- read.csv('~/Datasets/movie_finance.csv', stringsAsFactors = F)
yrly_boxoffice <- read.csv('~/Datasets/yearly_boxoffice.csv', stringsAsFactors = F)

# Data Clensing
# calcuate average movie rating
b <- group_by(b, movieId) %>% summarise(mean(rating))

# extract year as separate row
a$year <- str_sub(a$title, -5,-2)
a$title <- str_sub(a$title, 1, -8)

# extract genres
a$genre <- sapply(a$genres, function(x) unlist(str_split(x, "[|]"))[1])
a$genres = NULL

# title cleaning (remove * character)
d$title <- gsub("[*]", "", d$title)
d$year <- str_sub(d$year, -5,-1)

# Converting Units
yrly_boxoffice$total_gross <- (yrly_boxoffice$total_gross * 1000000)

a <- tbl_df(a)
b <- tbl_df(b)
c <- tbl_df(c)
d <- tbl_df(d)
yrly_boxoffice <- tbl_df(yrly_boxoffice)
```

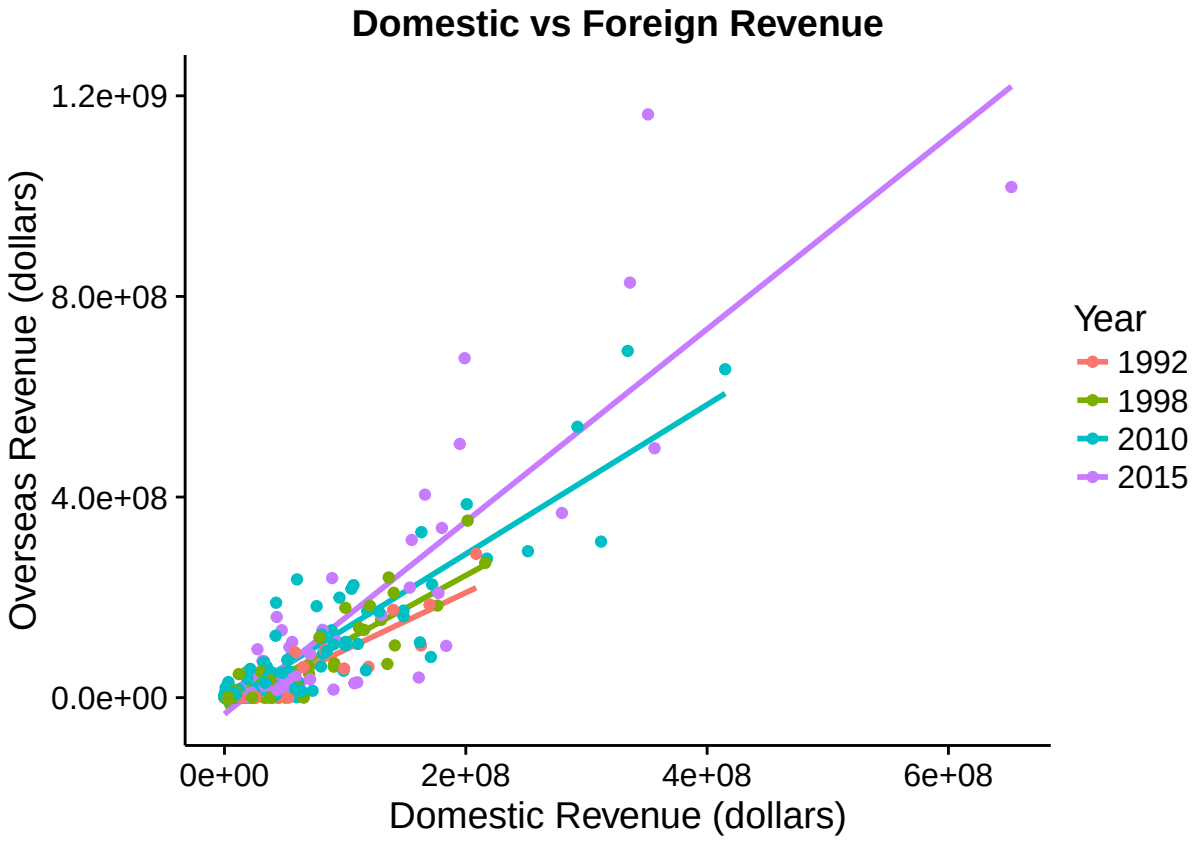
```r
# Joining Cleaned Data Sets
imdb1 <- left_join(b,a, by="movieId")
imdb1$year <- as.integer(imdb1$year)
```

```
## Warning: NAs introduced by coercion
```

```r
d$year <- as.integer(d$year)
movies <- left_join(d, imdb1, by = c('title','year'))

# rename column to rating
colnames(movies)[8] = "rating"

# removing Nas
movies <- movies[which(complete.cases(movies)),]

# calculate overseas column
movies$overseas <- (movies$worldwide - movies$domestic)

invisible(movies) # FINAL TIDY DATA FRAME

# movies subset used for some plots
movies1 <- movies[-1,]
movies1 <- filter(movies, budget > 100000000)
```

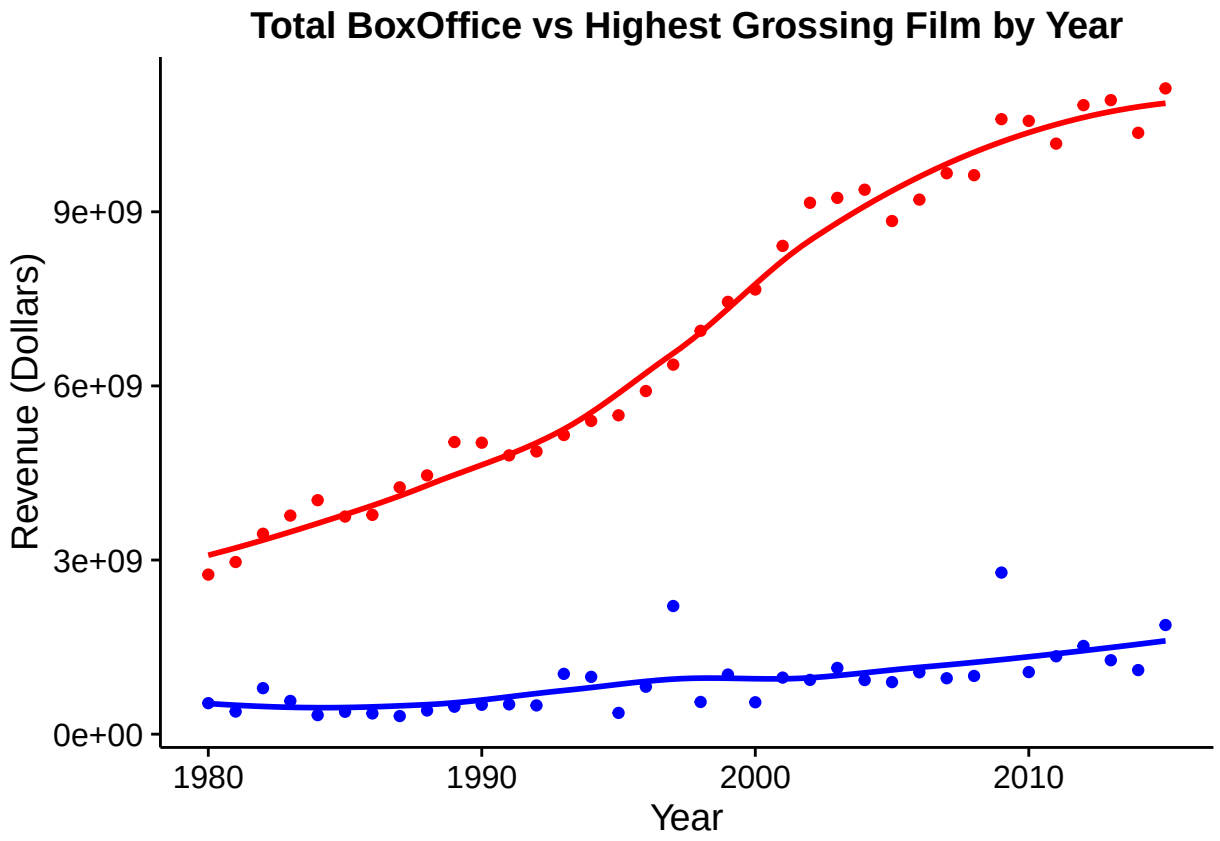
Comparing Domestic versus Foreign BoxOffice Revenues

```r
ggplot(data=movies[movies$year %in% c(1992,1998,2010,2015),]) +
    geom_smooth(aes(domestic, overseas, color=factor(year)), se=F, method = lm) +
    geom_point(aes(domestic, overseas, color=factor(year)))+scale_color_discrete(name='Year') +
    ggtitle("Domestic vs Foreign Revenue") +
    xlab("Domestic Revenue (dollars)") +
    ylab("Overseas Revenue (dollars)")
```



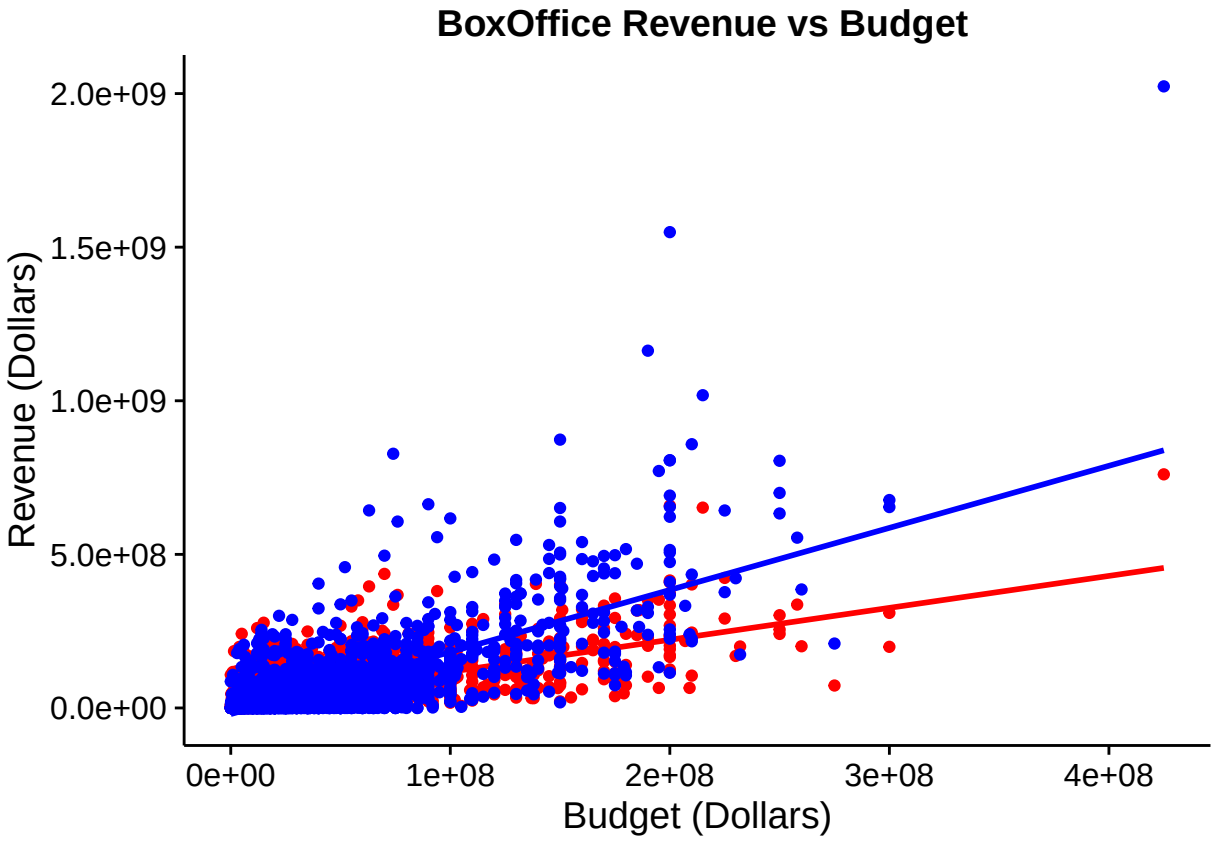
Conclusion: Foreign revenue has accounted for an increasing percentage of total revenue every year since 1992.

```
ggplot(data = yrly_boxoffice) +
    geom_smooth(aes(x=year, y=total_gross), colour='red', se=F) +
    geom_smooth(aes(x=year, y=worldwide), colour='blue', se=F) +
    geom_point(aes(x=year, y=total_gross), colour='red') +
    geom_point(aes(x=year, y=worldwide), colour='blue') +
    ggtitle("Total BoxOffice vs Highest Grossing Film by Year") +
    xlab("Year") +
    ylab("Revenue (Dollars)")
```



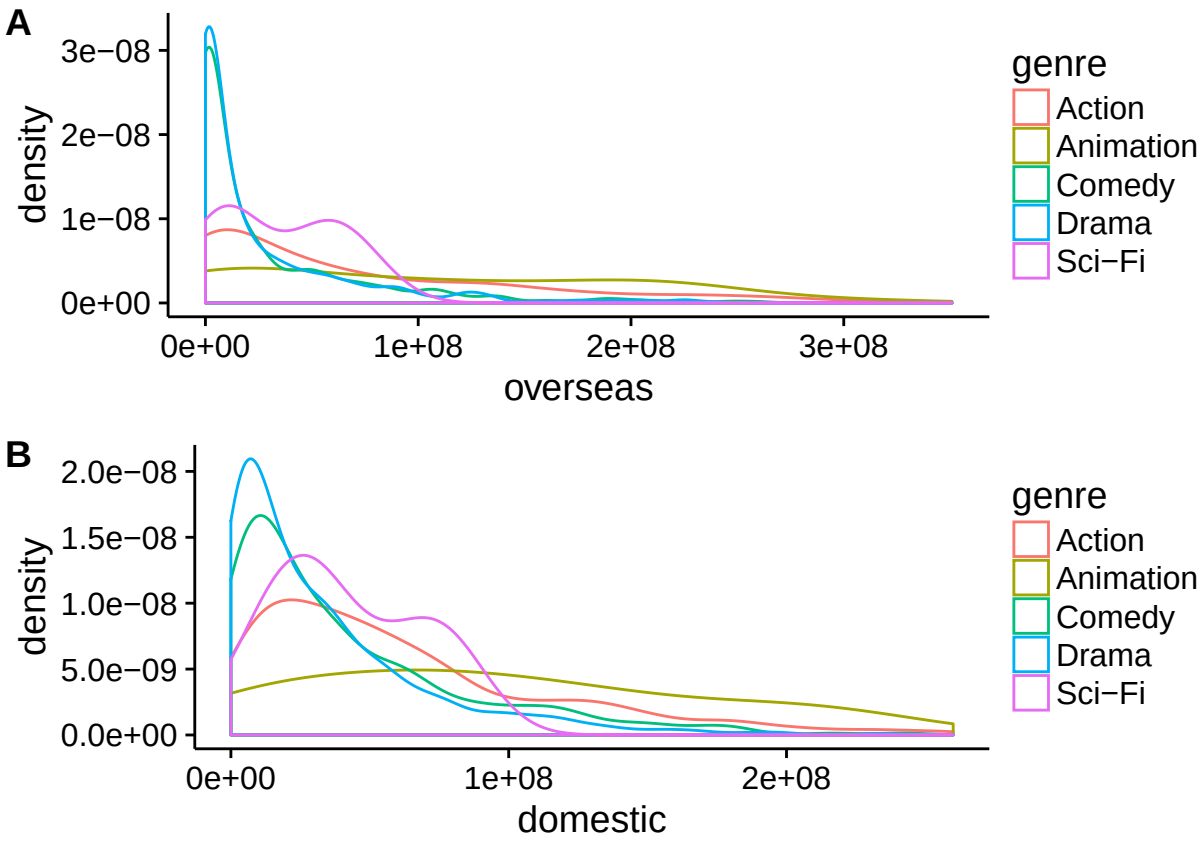
Conclusion: The Highest Grossing Movie per year accounted for a decreasing percentage of total BoxOffice Revenue

```
ggplot(data = movies) +
    geom_point(aes(x=budget, y= domestic), color='red') +
    geom_smooth(aes(x=budget, y= domestic), method='lm',formula=y~x, color='red', se=F) +
    geom_point(aes(x=budget, y= overseas), color='blue') +
    geom_smooth(aes(x=budget, y= overseas), method='lm',formula=y~x, color='blue', se=F) +
    ggtitle(" BoxOffice Revenue vs Budget") +
    xlab("Budget (Dollars)") +
    ylab("Revenue (Dollars)")
```



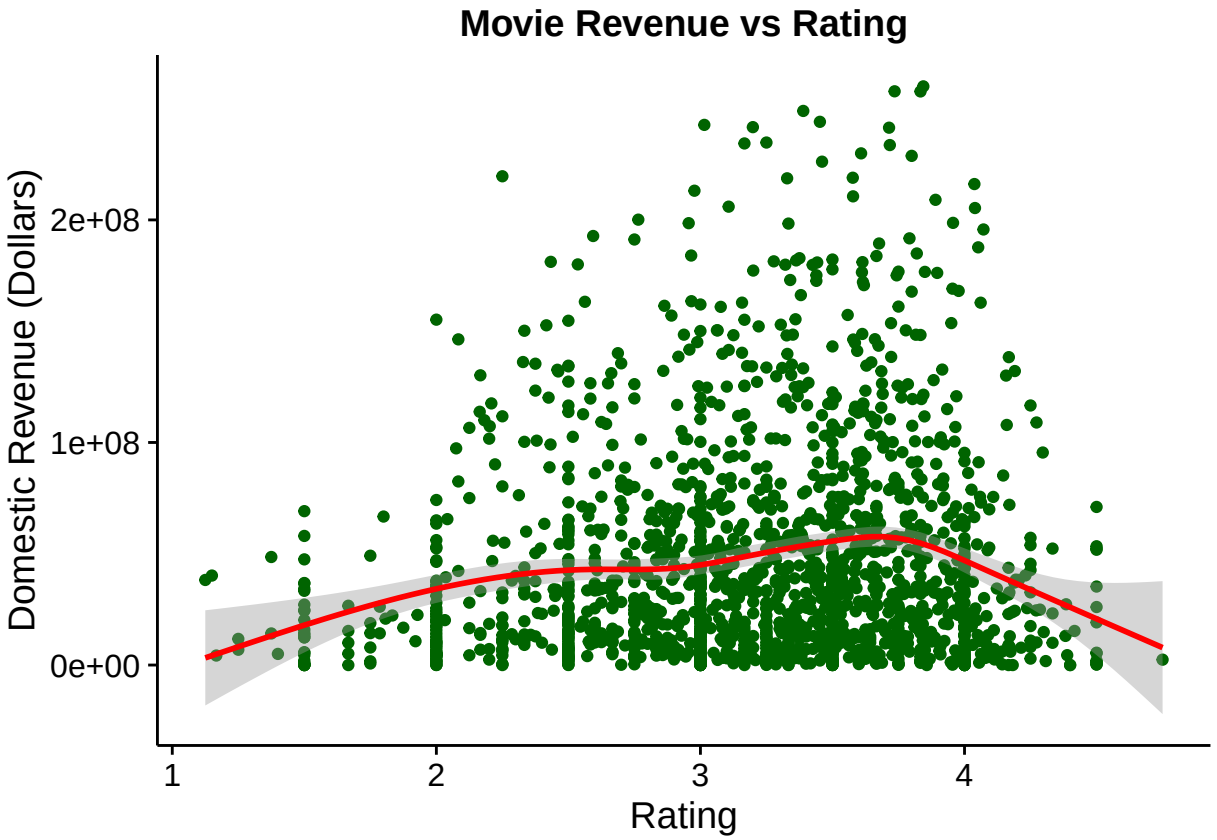
Conclusion: It seems as production budget increases, foreign to domestic revenue ratio increases. Meaning as budget increased, overseas revenue increases at a faster rate than domestic revenue. This is likely due to the fact tht big budget blockbusters are more marketable to the overseas audience. Movies like Transformers and Jurrassic Park had large budgets and made significantly more money overseas than domestically.

```
# Subsetting Genres and Rating
movies2 <- filter(movies, worldwide < 5e+08, rating > 1 & rating < 5, genre == 'Sci-Fi' | genre == 'Ani
z <- ggplot(movies2, aes(x=domestic)) + geom_density(aes(color=genre))
v <- ggplot(movies2, aes(x=overseas)) + geom_density(aes(color=genre))
plot_grid(v, z, labels=c("A", "B", "C"), ncol = 1, nrow = 2)
```



Conclusion: The density plot shows us that Drama and Comedy Genres perform worse Overseas compared to Domestically since a higher density of Drama/Comedy movies return low revenue Overseas. This could be due to fact that Drama/Comedy movies are less translatable to foreign markets since they require more cultural context. Whereas Action/Sci-Fi translate better to foreign markets because

```
ggplot(data = movies2) +
    geom_point(aes(x=rating, y= domestic), color='dark green') +
    geom_smooth(aes(x=rating, y= domestic), se=T, color = 'red') +
    ggtitle("Movie Revenue vs Rating") +
    xlab("Rating") + ylab("Domestic Revenue (Dollars)")
```



Conclusion: The rating sweet spot that generates the most revenue is between 3.5-3.8. Movies that score greater than 4 have a sharp decline in revenues. This could be due to the fact that the average movie goer more easily appreciates an average movie (cough cough** Michael Bay movies).