# R Notebook - Movie Industries Project

Maya Reese Farmer

October 15, 2021

**Introduction**

I will be demonstrating some introductory statistics and data visualization techniques using the Movie Industry dataset from kaggle.com
*this data was scraped from IMDb

The 'movies' dataset contains 6820 movies (220 movies per year, 1986-2016) and has 15 columns. Each movie has the following attributes:

- budget

- company
- county
- director
- genre
- gross (revenue)
- name
- rating (R, PG, PG-13, etc.)
- released (date)
- runtime (minutes)
- score (IMDb user rating)
- votes
- star
- writer
- year

```
## importing dataset into r
movies <- read.csv("~/Downloads/movies.csv")
attach(movies)

## view first 6 rows of data
head(movies)
```

```
##                                         name rating     genre year
## 1                                The Shining      R     Drama 1980
## 2                            The Blue Lagoon      R Adventure 1980
## 3 Star Wars: Episode V - The Empire Strikes Back     PG    Action 1980
## 4                                  Airplane!     PG    Comedy 1980
## 5                                Caddyshack      R    Comedy 1980
## 6                            Friday the 13th      R    Horror 1980
##                       released score   votes           director
```

```
## 1 June 13, 1980 (United States)   8.4  927000    Stanley Kubrick
## 2  July 2, 1980 (United States)   5.8   65000     Randal Kleiser
## 3 June 20, 1980 (United States)   8.7 1200000     Irvin Kershner
## 4  July 2, 1980 (United States)   7.7  221000       Jim Abrahams
## 5 July 25, 1980 (United States)   7.3  108000       Harold Ramis
## 6   May 9, 1980 (United States)   6.4  123000 Sean S. Cunningham
##                    writer           star        country  budget     gross
## 1           Stephen King Jack Nicholson United Kingdom 1.9e+07  46998772
## 2 Henry De Vere Stacpoole Brooke Shields  United States 4.5e+06  58853106
## 3          Leigh Brackett    Mark Hamill  United States 1.8e+07 538375067
## 4            Jim Abrahams    Robert Hays  United States 3.5e+06  83453539
## 5      Brian Doyle-Murray    Chevy Chase  United States 6.0e+06  39846344
## 6           Victor Miller   Betsy Palmer  United States 5.5e+05  39754601
##              company runtime
## 1       Warner Bros.     146
## 2  Columbia Pictures     104
## 3          Lucasfilm     124
## 4 Paramount Pictures      88
## 5     Orion Pictures      98
## 6 Paramount Pictures      95
```

### Queries and Data Visualization

I'll be querying the data to answer a series of questions and will present my findings using either tables or ggplot2 for data visualization.

**Genre, Budget, and Revenue** The first thing that would be interesting to look at is the relationship between movie genre and budget: is there a significant difference in budget between movies of different genres?
Because genre is a categorical variable, I will perform an analysis of variance (ANOVA) of budget and movie genre.

```
## create one-way anova model
genre.aov = aov(budget~genre, data=movies)
summary(genre.aov)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## genre         15 1.870e+18 1.247e+17    90.2 <2e-16 ***
## Residuals   5481 7.576e+18 1.382e+15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2171 observations deleted due to missingness
```

The output shows that different genres have significantly different budgets becuase the p value is significantly less that 0.05.

I think the best way to visualize this data would be to use a crossbar plot in ggplot2. These plots will show the mean and standard deviation (sd) in budget for each genre. To calculate mean and sd more efficiently, I can use a summary function.

```
library(plyr)
## first determine whether there are NAs in the budget
```

```
## column
sum(is.na(budget))
```

```
## [1] 2171
```

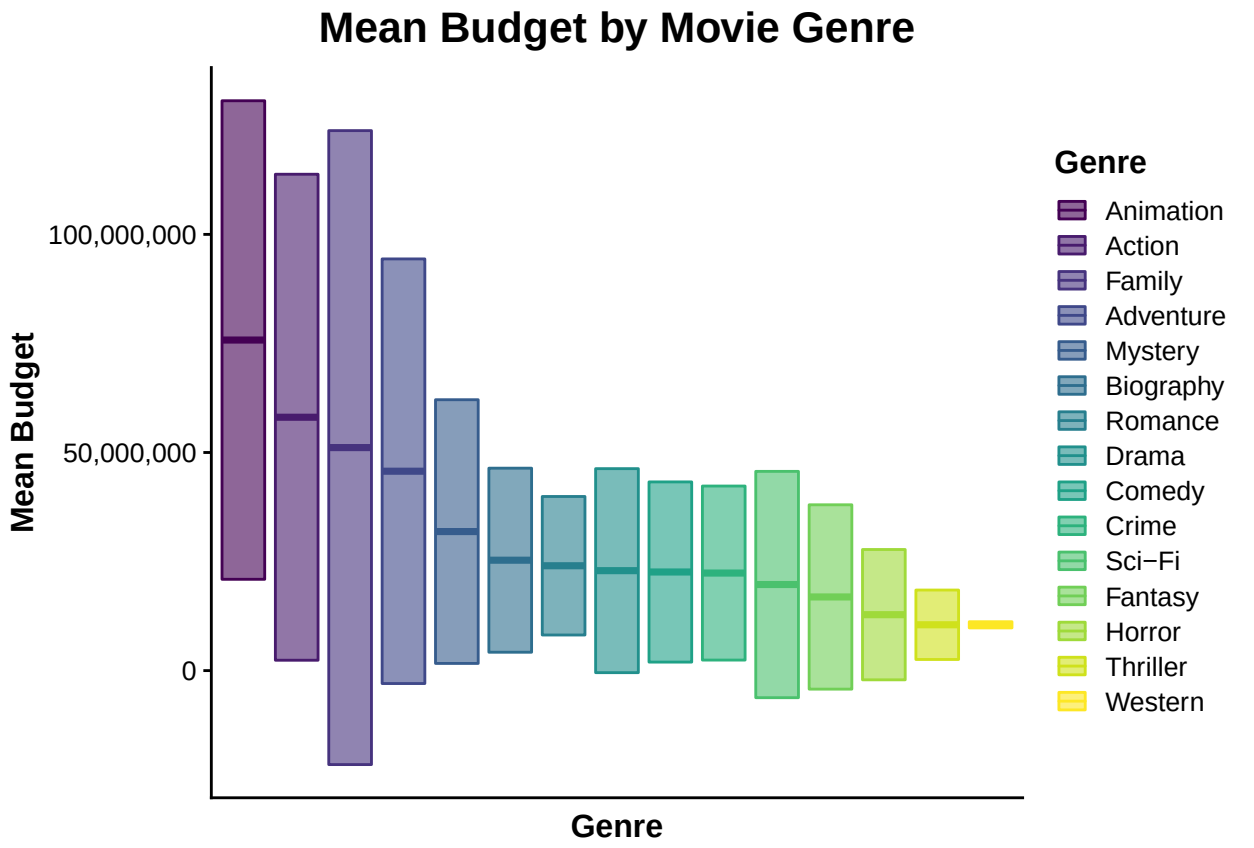
```
## summary function to calculate mean, sd, and standard
## error removing NAs
summary.genre <- ddply(movies, c("genre"), summarise, N = length(budget),
    mean = mean(budget, na.rm = TRUE), sd = sd(budget, na.rm = TRUE),
    se = sd/sqrt(N))
summary.genre
```

```
##        genre    N     mean         sd         se
## 1     Action 1705 58084599 55701126.6  1348968.4
## 2  Adventure  427 45708389 48667393.8  2355180.3
## 3  Animation  338 75785197 54850883.3  2983494.7
## 4  Biography  443 25312317 21097934.8  1002393.1
## 5     Comedy 2245 22607802 20653227.6   435892.9
## 6      Crime  551 22363566 19946540.1   849751.2
## 7      Drama 1518 22914609 23380692.2   600097.0
## 8     Family   11 51125000 72666332.2 21909723.5
## 9    Fantasy   44 16885714 21128810.2  3185288.0
## 10   History    1   323562         NA         NA
## 11    Horror  322 12825159 14935014.6   832295.8
## 12     Music    1      NaN         NA         NA
## 13   Musical    2      NaN         NA         NA
## 14   Mystery   20 31876471 30227006.3  6758964.1
## 15   Romance   10 24040000 15877909.2  5021035.7
## 16    Sci-Fi   10 19733750 25942298.0  8203674.9
## 17     Sport    1      NaN         NA         NA
## 18  Thriller   16 10511111  7957299.2  1989324.8
## 19   Western    3 10500000   707106.8   408248.3
```

Now the summary dataset is ready for ggplot.

```
library(ggplot2)
library(cowplot)
library(scales)
library(viridisLite)

ggplot(na.omit(summary.genre), aes(x = (reorder(genre, -mean)),
    mean, color = reorder(genre, -mean), fill = reorder(genre,
        -mean))) + geom_crossbar(aes(ymin = mean - sd, ymax = mean +
    sd), alpha = 0.6, width = 0.8) + scale_y_continuous(labels = comma) +
    theme_cowplot(12) + scale_fill_viridis_d(name = "Genre") +
    scale_colour_viridis_d(name = "Genre") + theme(axis.text.x = element_blank(),
    axis.ticks.x = element_blank(), legend.title = element_text(face = "bold"),
    axis.title.x = element_text(face = "bold"), axis.title.y = element_text(face = "bold"),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5)) +
    labs(x = "Genre", y = "Mean Budget") + ggtitle("Mean Budget by Movie Genre")
```




The figure shows mean budget broken down by movie genre. Each box and color represents a different genre (as shown in legend). Each box represents the mean (center line) and sd in budget for each genre. Based on this graph, we can see that Animation, Action, and Family movies tend to have the highest budgets, but there's also a lot of variation in budget as well. On the other end of the spectrum, Westerns, Thrillers, and Horror films tend to have the lowest budgets, and there's much less variation around the mean.

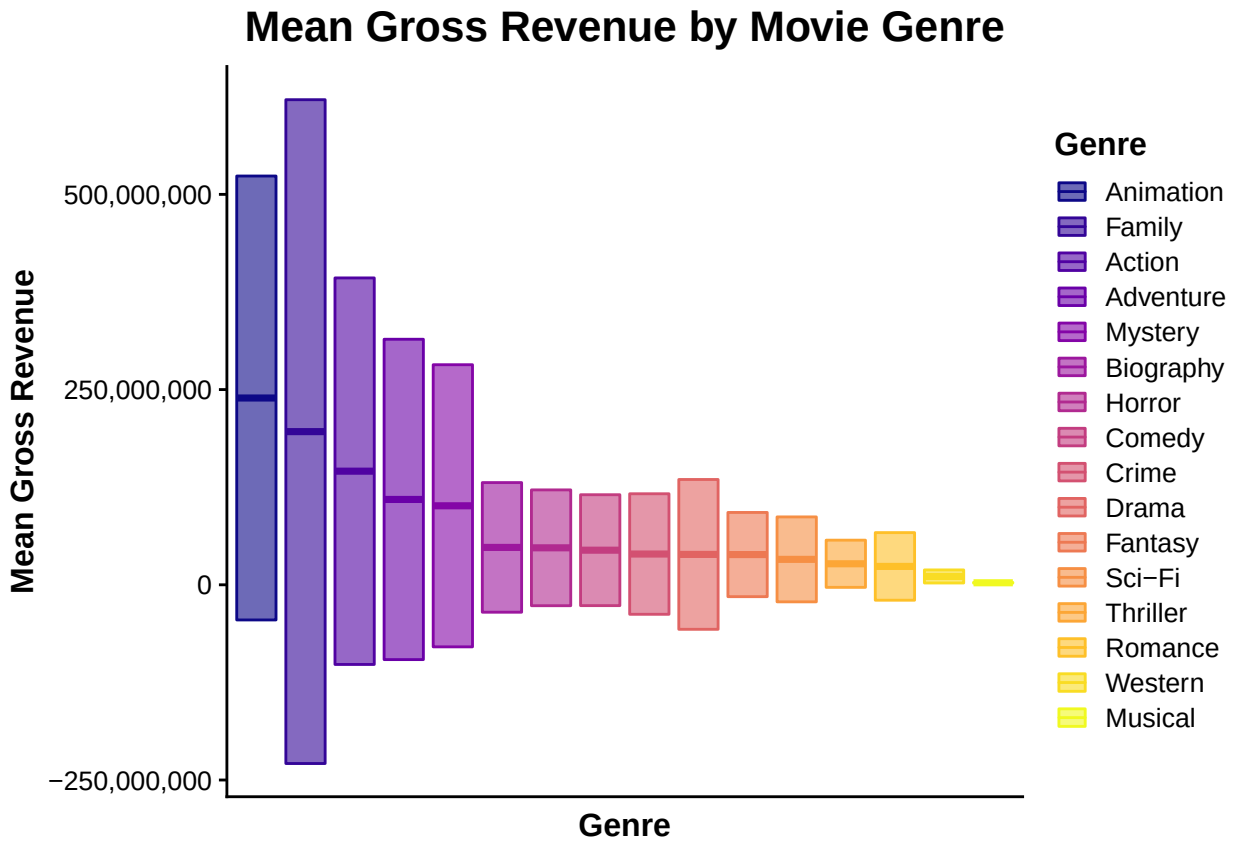
Next, let's see whether this pattern holds when we compare genre to gross revenue: do Animation, Action, and Family movies also rake in the highest gross revenue?

```
## create summary data comparing genre and gross revenue
summary.gross <- ddply(movies, c("genre"), summarise,
               N    = length(gross),
               mean = mean(gross, na.rm = TRUE),
               sd   = sd(gross, na.rm = TRUE),
               se   = sd / sqrt(N)
)
summary.gross
```

```
##        genre    N      mean          sd        se
## 1     Action 1705 145508581 247515833.1   5994332
## 2  Adventure  427 109325230 205149463.3   9927878
## 3  Animation  338 239229987 284266475.7  15462058
## 4  Biography  443  47874323  83004471.8   3943661
## 5     Comedy 2245  44331874  71029066.8   1499091
## 6      Crime  551  39401196  77171122.0   3287600
## 7      Drama 1518  38930959  95928404.6   2462132
## 8     Family   11 196172492 425078978.8 128166134
## 9    Fantasy   44  38709329  53933631.2   8130801
```

```
## 10   History    1       NaN          NA        NA
## 11    Horror  322  47372409  74162666.7   4132924
## 12     Music    1    110014          NA        NA
## 13   Musical    2   2595346    534701.4    378091
## 14   Mystery   20 101183528 180676747.7  40400549
## 15   Romance   10  23549375  43285526.2  13688085
## 16    Sci-Fi   10  32561233  54386726.5  17198593
## 17     Sport    1   1067629          NA        NA
## 18  Thriller   16  26935259  30215705.6   7553926
## 19   Western    3  10675295   8355948.8   4824309
```

```
ggplot(na.omit(summary.gross), aes(x = (reorder(genre, -mean)),
    y = mean, color = reorder(genre, -mean), fill = reorder(genre,
        -mean))) + geom_crossbar(aes(ymin = mean - sd, ymax = mean +
    sd), alpha = 0.6, width = 0.8) + scale_y_continuous(labels = comma) +
    theme_cowplot(12) + scale_fill_viridis_d(name = "Genre",
    option = "plasma") + scale_colour_viridis_d(name = "Genre",
    option = "plasma") + theme(axis.text.x = element_blank(),
    axis.ticks.x = element_blank(), legend.title = element_text(face = "bold"),
    axis.title.x = element_text(face = "bold"), axis.title.y = element_text(face = "bold"),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5)) +
    labs(x = "Genre", y = "Mean Gross Revenue") + ggtitle("Mean Gross Revenue by Movie Genre")
```




The graph shows that the top 3 genres from the budget graph (Animation, Action, and Family) are also the highest revenue genres *although action and family switch. I think it's interesting to note that Family movies have a lot of variation around the mean, which indicates that they can be highly lucrative or they cab be extremely costly and lose money. So making Family movies might be more risky (there are probably

several factors that influence how much money Family movies make; further analysis could elucidate these factors).

Another interesting factor to point out is that some genres that had lower budgets actually tend to make more money, and vice versa. For example, Horror films had some of the lowest budgets, but they actually rake in a decent amount of revenue. On the other hand, Romance movies had a higher budget, but fall in the bottom 3 when you consider gross revenue.

---

Finally, let's consider the relationship between budget and revenue: do higher budget films also generate more revenue?

First, we can perform a Pearson correlation to determine whether there's a significant correlation between budget and gross revenue.

```
cor.test(budget,gross, method = c("pearson"))
```

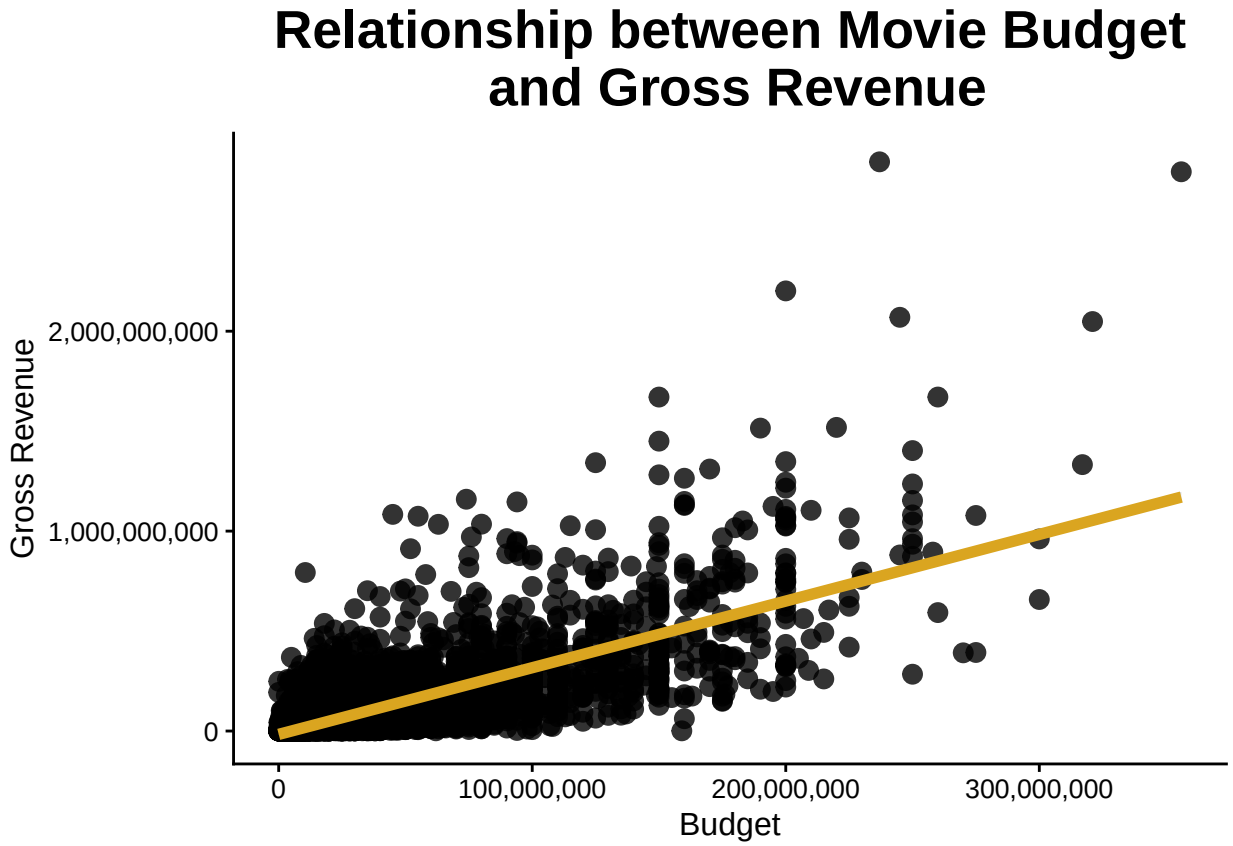
```
##
##  Pearson's product-moment correlation
##
## data:  budget and gross
## t = 81.198, df = 5434, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7281425 0.7521743
## sample estimates:
##       cor
## 0.7403949
```

Based on the output, we can see that there is a significant, positive relationship between budget and gross revenue ($r=0.74$, $p<0.001$), indicating that movies with higher budgets tend to generate more revenue.

We can visualize this relationship using a scatterplot.

```
ggplot(na.omit(movies), aes(budget, gross)) + geom_point(size = 3,
    alpha = 0.8) + geom_smooth(method = lm, se = TRUE, color = "goldenrod",
    size = 2) + labs(x = "Budget", y = "Gross Revenue") + theme_cowplot(12) +
    ggtitle("Relationship between Movie Budget\n and Gross Revenue") +
    theme(plot.title = element_text(size = 20, face = "bold",
        hjust = 0.5)) + scale_y_continuous(labels = comma) +
    scale_x_continuous(labels = comma)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




---

**Stars and Gross Revenue** Next, I want to look at the top 10 actors in the dataset based on the number of movies they're starred in.

```
library(dplyr)
```

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

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

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

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

```
library(magrittr)
## tally the number of movies by star, sorted from highest to lowest
movies %>% group_by(star) %>% tally(sort = TRUE)
```

```
## # A tibble: 2,815 x 2
##    star                  n
##    <chr>             <int>
##  1 Nicolas Cage         43
##  2 Robert De Niro       41
##  3 Tom Hanks            41
##  4 Denzel Washington    37
##  5 Bruce Willis         34
##  6 Tom Cruise           34
##  7 Johnny Depp          33
##  8 Sylvester Stallone   32
##  9 John Travolta        31
## 10 Kevin Costner        29
## # ... with 2,805 more rows
```

I can now use the output to create a new dataset that only includes data from these 10 actors.

```
## subset of movies including top 10 stars
top10star = subset(movies, star == "Nicolas Cage" | star == "Robert De Niro" |
    star == "Tom Hanks" | star == "Denzel Washington" | star ==
    "Bruce Willis" | star == "Tom Cruise" | star == "Johnny Depp" |
    star == "Sylvester Stallone" | star == "John Travolta" |
    star == "Kevin Costner")
head(top10star)
```

```
##                 name rating     genre year                         released
## 8        Raging Bull      R Biography 1980  December 19, 1980 (United States)
## 25      Urban Cowboy     PG     Drama 1980       June 6, 1980 (United States)
## 99          Blow Out      R     Crime 1981      July 24, 1981 (United States)
## 150 True Confessions      R     Crime 1981 September 25, 1981 (United States)
## 152       Nighthawks      R    Action 1981     April 10, 1981 (United States)
## 211      First Blood      R    Action 1982   October 22, 1982 (United States)
##     score  votes        director             writer               star
## 8     8.2 330000 Martin Scorsese       Jake LaMotta     Robert De Niro
## 25    6.4  14000   James Bridges       Aaron Latham      John Travolta
## 99    7.4  47000  Brian De Palma     Brian De Palma      John Travolta
## 150   6.3   7500    Ulu Grosbard John Gregory Dunne     Robert De Niro
## 152   6.4  18000   Bruce Malmuth       David Shaber Sylvester Stallone
## 211   7.7 234000    Ted Kotcheff      David Morrell Sylvester Stallone
##           country  budget     gross                        company runtime
## 8   United States 1.8e+07  23402427 Chartoff-Winkler Productions     129
## 25  United States      NA  46918287           Paramount Pictures     132
## 99  United States 1.8e+07  12000000            Filmways Pictures     108
## 150 United States 1.0e+07  12850276 Chartoff-Winkler Productions     108
## 152 United States 5.0e+06  19905359           Universal Pictures      99
## 211 United States 1.5e+07 125212904                Anabasis N.V.      93
```

Now I want to evaluate the relationship between actor and gross revenue to determine whether actors, and the movies they star in, differ in the amount of revenue they generate.
To formally test this relationship, I'll perform an ANOVA using star and gross revenue.

```
top10.aov = aov(gross ~ star, data = top10star)
summary(top10.aov)
```

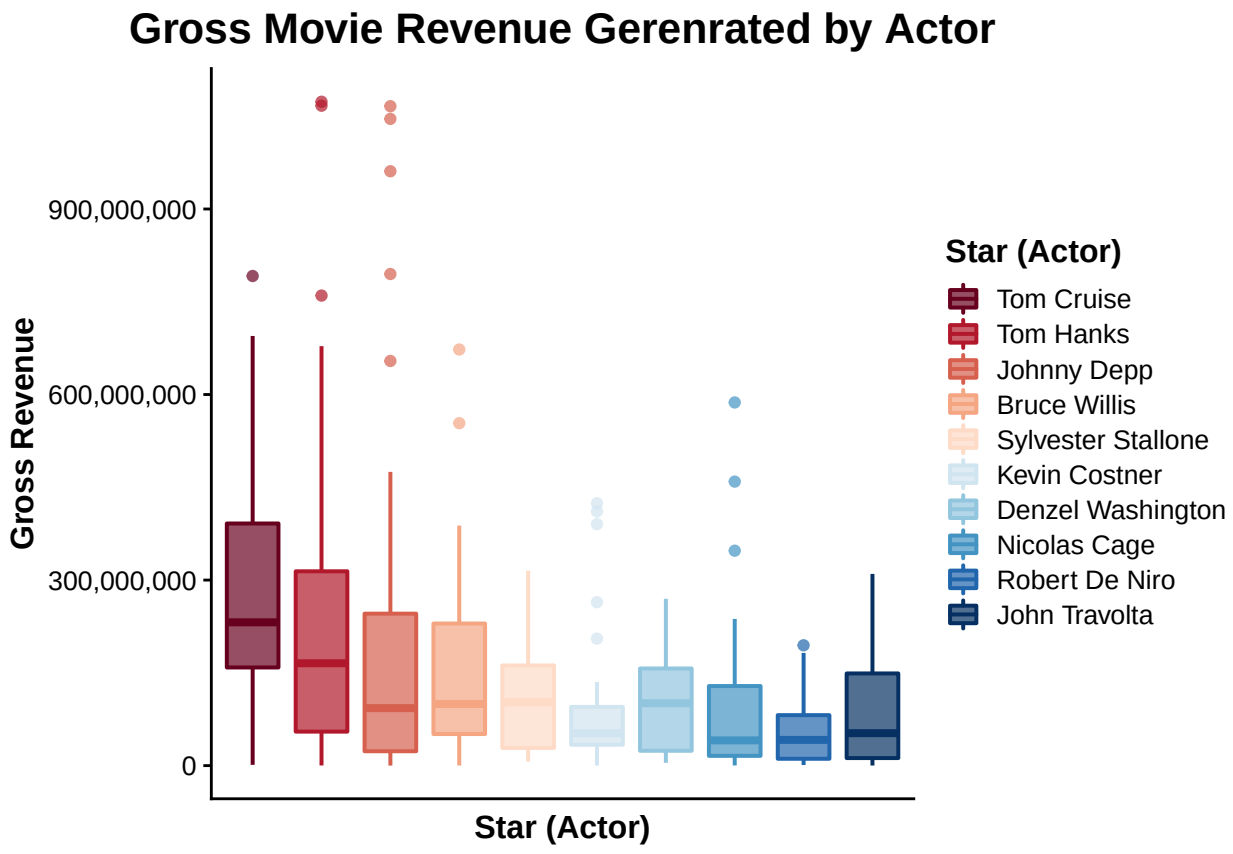
```
##              Df    Sum Sq   Mean Sq F value  Pr(>F)
## star          9 1.942e+18 2.157e+17   7.336 8.6e-10 ***
## Residuals   344 1.012e+19 2.941e+16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

Again, there are significant differences in the gross revenue generated by different actors, as the p value is significantly less than 0.

Next, we'll visualize these differences using a boxplot.

```
ggplot(top10star, aes(reorder(star, -gross), gross, color = reorder(star,
    -gross), fill = reorder(star, -gross))) + geom_boxplot(size = 0.7,
    alpha = 0.7) + theme_cowplot(12) + scale_y_continuous(labels = comma) +
    scale_color_brewer(name = "Star (Actor)", palette = "RdBu") +
    scale_fill_brewer(name = "Star (Actor)", palette = "RdBu") +
    theme(axis.text.x = element_blank(), axis.ticks.x = element_blank(),
        axis.title = element_text(face = "bold"), plot.title = element_text(size = 16,
            face = "bold", hjust = 0.5), legend.title = element_text(face = "bold")) +
    labs(x = "Star (Actor)", y = "Gross Revenue", title = "Gross Movie Revenue Gerenrated by Actor")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



Based on my earlier output, the number of movies an actor is in doesn't necessarily seem to be correlated with how much revenue that movie generates. For example, Nicolas Cage has starred in the most movies, but he falls in the bottom 3 in terms of revenue generated. This is also true for Robert De Niro.

However, movies starring Tom Cruise, Tom Hanks, and Johnny Depp generate the most revenue among this group of actors.

---

**Top 5 Movies based on IMDb Score** Finally, I want to determine what the top 5 movies in our dataset are based on IMDb score.

```
## order movies based on score then reduce data frame to top 5 movies
movie.rank = arrange(movies, desc(score))
movie.rank2 = movie.rank[1:5,]
movie.rank2
```

```
##                                       name rating     genre year
## 1                    The Shawshank Redemption      R     Drama 1994
## 2                             The Dark Knight  PG-13    Action 2008
## 3                            Schindler's List      R Biography 1993
## 4                               Pulp Fiction      R     Crime 1994
## 5 The Lord of the Rings: The Return of the King  PG-13    Action 2003
##                          released score   votes          director
## 1  October 14, 1994 (United States)   9.3 2400000    Frank Darabont
## 2     July 18, 2008 (United States)   9.0 2400000 Christopher Nolan
## 3  February 4, 1994 (United States)   8.9 1200000  Steven Spielberg
## 4  October 14, 1994 (United States)   8.9 1900000 Quentin Tarantino
## 5 December 17, 2003 (United States)   8.9 1700000     Peter Jackson
##              writer           star       country   budget      gross
## 1      Stephen King    Tim Robbins United States 2.50e+07   28817291
## 2    Jonathan Nolan Christian Bale United States 1.85e+08 1005973645
## 3   Thomas Keneally    Liam Neeson United States 2.20e+07  322161245
## 4 Quentin Tarantino  John Travolta United States 8.00e+06  213928762
## 5    J.R.R. Tolkien    Elijah Wood   New Zealand 9.40e+07 1146030912
##                      company runtime
## 1 Castle Rock Entertainment     142
## 2              Warner Bros.     152
## 3        Universal Pictures     195
## 4                   Miramax     154
## 5           New Line Cinema     201
```

I will use a lollipop graph to visualize the differences in IMDb scores among these films.

```
ggplot(movie.rank2, aes(x = reorder(name, -score), y = score,
    color = reorder(name, -score), fill = reorder(name, -score))) +
    geom_segment(aes(x = reorder(name, -score), xend = reorder(name,
        -score), y = 0, yend = score), size = 2) + geom_point(size = 7,
    shape = 21, alpha = 0.6) + labs(x = "", y = "IMDb Score",
    title = "Top 5 Movies based on IMDb Score") + theme(panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), panel.background = element_blank(),
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
    axis.ticks.x = element_blank(), axis.text.x = element_text(size = 11),
```

```
    axis.title = element_text(face = "bold"), axis.text.y = element_text(size = 11),
    legend.position = "none") + scale_x_discrete(labels = c("The Shawshank\n Redemption",
    "The Dark Knight", "Pulp Fiction", "Schindler's List", "Lord of the Rings:\n Return of the King")) +
    scale_fill_manual(values = alpha(c("#cad2c5", "#84a98c",
        "#52796f", "#354f52", "#2f3e46"), 0.85), guide = "none") +
    scale_color_manual(values = c("#cad2c5", "#84a98c", "#52796f",
        "#354f52", "#2f3e46"), guide = "none") + ylim(0, 10)
```