Problem Set 1

MGSC 310, Fall 2019, Professor Hersh (BEST PROFESSOR EVER!!!)

Geoffrey Hughes

9/5/2019

```
library("tidyverse")
```

```
## -- Attaching packages -----------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ---------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

a. Getting & Setting the working directory

```
getwd()
```

```
## [1] "/Users/geoffreyhughes/Documents/MGSC_310/MGSC310"
```

```
setwd("/Users/geoffreyhughes/Documents/MGSC_310/MGSC310")
```

b. Importing the downloaded dataset (movie_metadata.csv)

```
movies <- read.csv("Datasets/movie_metadata.csv")
```

c. Dimensions of the dataset

5043 observations with 28 variables

```
dim(movies)
```

```
## [1] 5043   28
```

d. Variable Names

```r
names(movies)
```

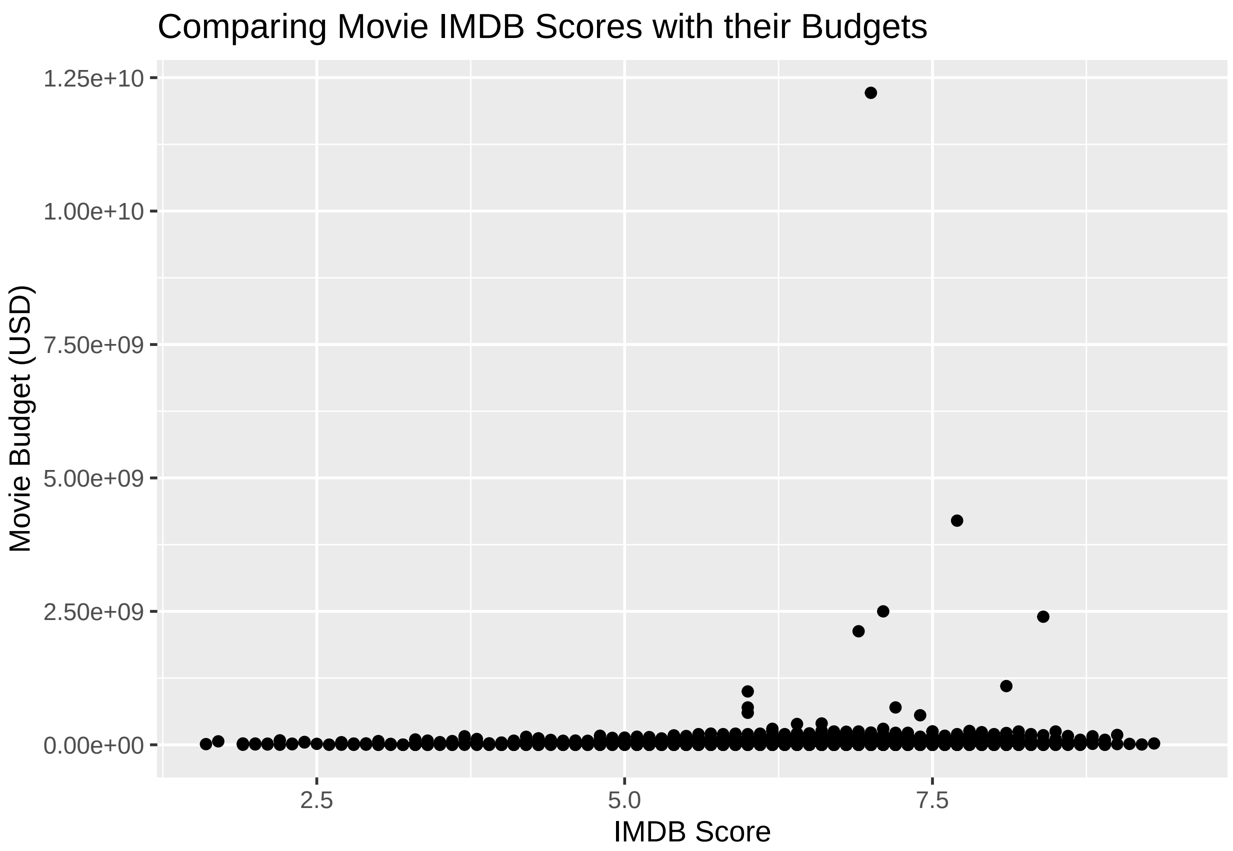
```
##  [1] "color"                     "director_name"            
##  [3] "num_critic_for_reviews"    "duration"                 
##  [5] "director_facebook_likes"   "actor_3_facebook_likes"   
##  [7] "actor_2_name"              "actor_1_facebook_likes"   
##  [9] "gross"                     "genres"                   
## [11] "actor_1_name"              "movie_title"              
## [13] "num_voted_users"           "cast_total_facebook_likes"
## [15] "actor_3_name"              "facenumber_in_poster"     
## [17] "plot_keywords"             "movie_imdb_link"          
## [19] "num_user_for_reviews"      "language"                 
## [21] "country"                   "content_rating"           
## [23] "budget"                    "title_year"               
## [25] "actor_2_facebook_likes"    "imdb_score"               
## [27] "aspect_ratio"              "movie_facebook_likes"
```

[1] “color” “director_name” “num_critic_for_reviews” “duration”
[5] “director_facebook_likes” “actor_3_facebook_likes” “actor_2_name” “actor_1_facebook_likes”
[9] “gross” “genres” “actor_1_name” “movie_title”
[13] “num_voted_users” “cast_total_facebook_likes” “actor_3_name” “facenumber_in_poster”
[17] “plot_keywords” “movie_imdb_link” “num_user_for_reviews” “language”
[21] “country” “content_rating” “budget” “title_year”
[25] “actor_2_facebook_likes” “imdb_score” “aspect_ratio” “movie_facebook_likes”

e. Scatterplot of IMDB on the x-axis and movie budgets on the y-axis.

```r
ggplot(movies, aes(x = imdb_score, y = budget)) +
  geom_point() +
  labs(x = "IMDB Score",
       y = "Movie Budget (USD)",
       title = "Comparing Movie IMDB Scores with their Budgets")
```

```
## Warning: Removed 492 rows containing missing values (geom_point).
```



f. Remove movies with budgets > $400 million

```
dim(movies)
```

```
## [1] 5043   28
```

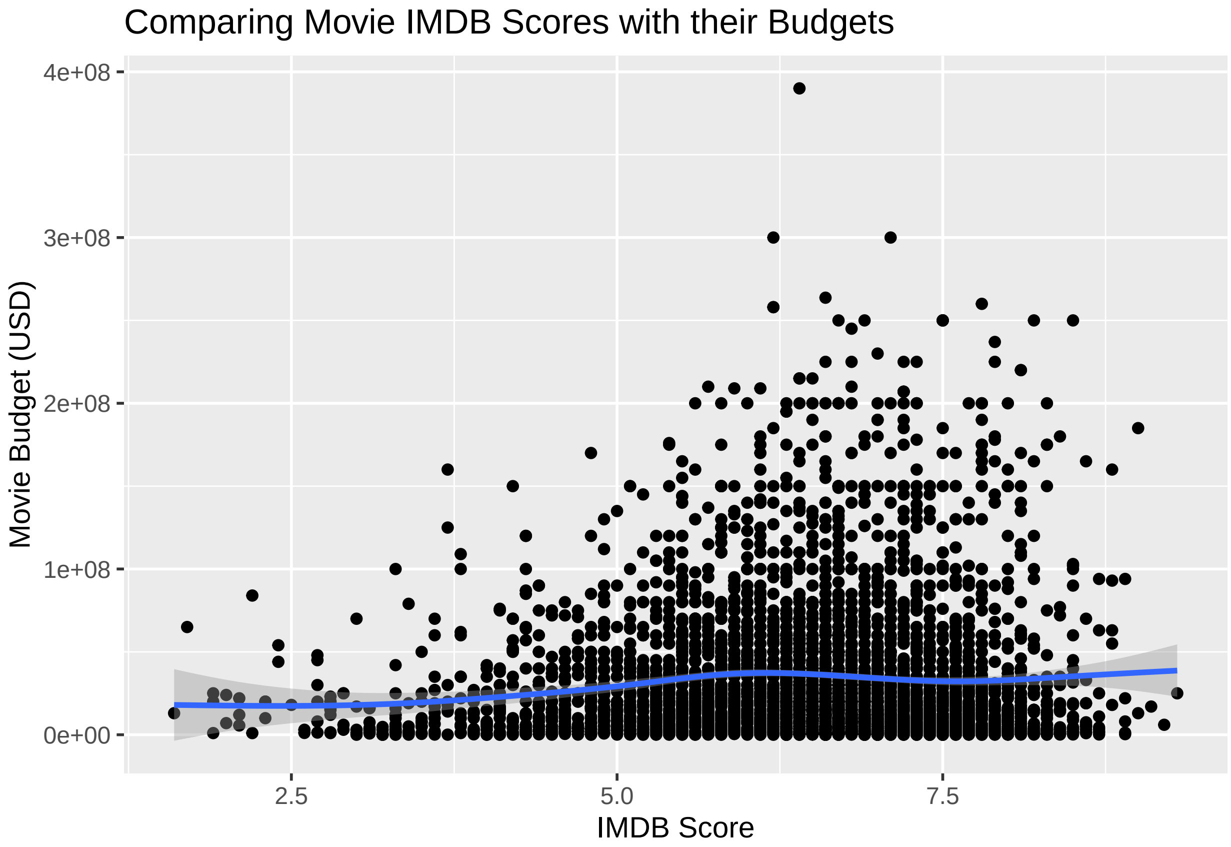
```
movies <- movies %>% filter(budget < 400000000)
```

Went from 5043 movies to 4539 movies!

g. Create a trendline in the ggplot

```
ggplot(movies, aes(x = imdb_score, y = budget)) +
  geom_point() +
  stat_smooth() +
  labs(x = "IMDB Score",
       y = "Movie Budget (USD)",
       title = "Comparing Movie IMDB Scores with their Budgets")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



There is a *very* slight positive relationship between higher budgets = higher IMDB scores, but in some places there is a negative relationship. I would say there is NOT a significant relationship.

h. Sub-plots by content_rating in ggplot

```
movies$rating_factor <- factor(movies$content_rating)

ggplot(movies, aes(x = imdb_score, y = budget)) +
  geom_point() +
  stat_smooth() +
  facet_wrap(~rating_factor, scales = "free") +
  labs(x = "IMDB Score",
       y = "Movie Budget (USD)",
       title = "Comparing Movie IMDB Scores with their Budgets")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
```

```
## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
```

```
## Warning: Computation failed in `stat_smooth()`:
```

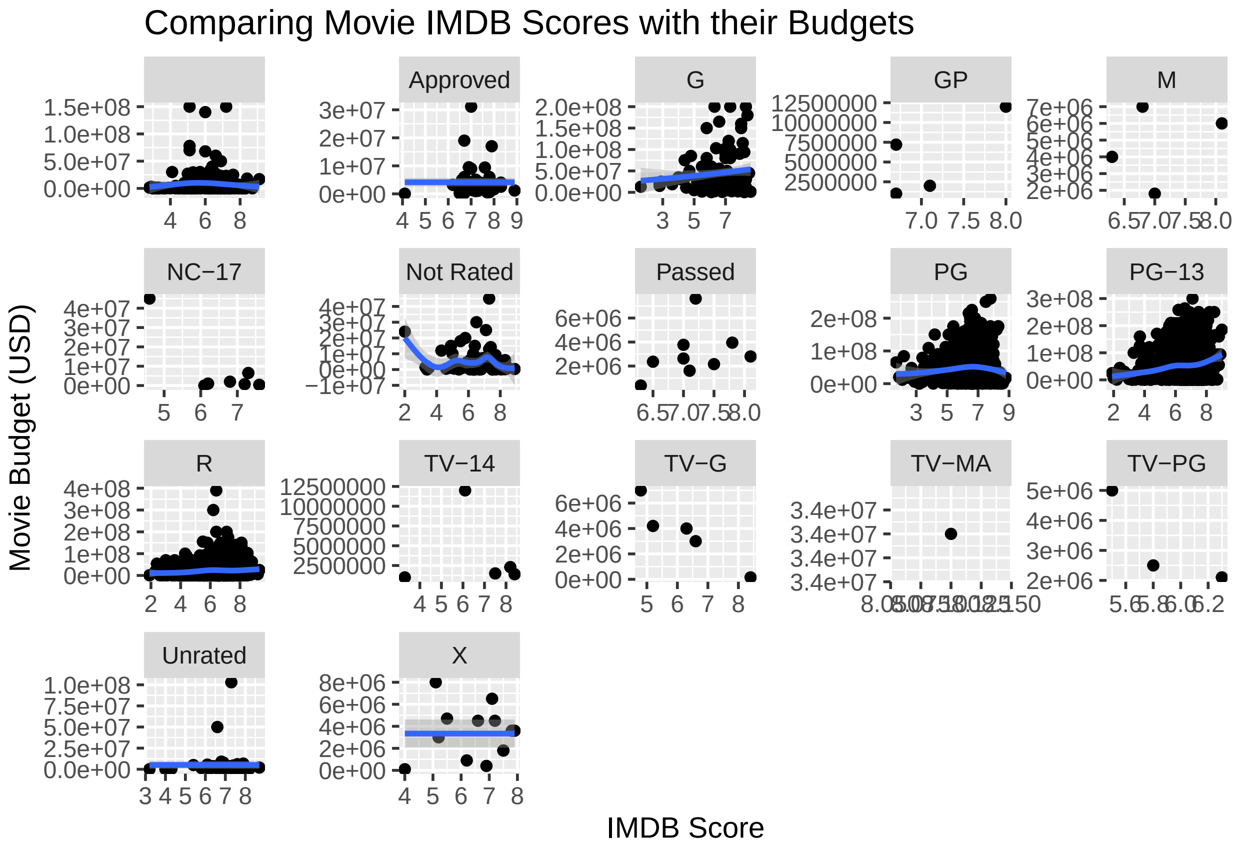
```
## x has insufficient unique values to support 10 knots: reduce k.

## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.

## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.

## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.

## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
```



We see the strongest relationship between mvoie IMDB score and budget in G and PG-13 movies, which are both relatively linear positive relationships.

i. Use ggridges to produce a ridgeline density plot graph by genre

```
library('ggridges')
```

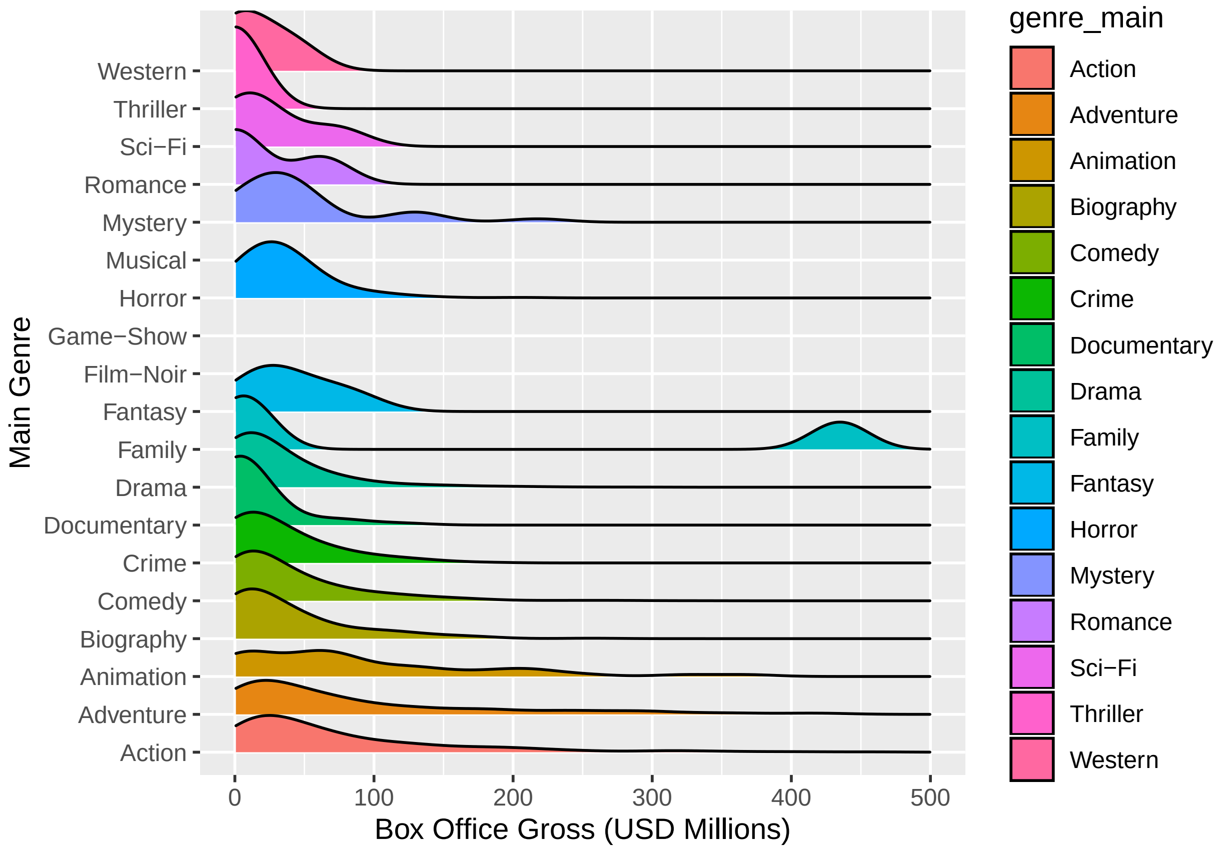
```
##
## Attaching package: 'ggridges'
## The following object is masked from 'package:ggplot2':
##
##     scale_discrete_manual
```

```
movies <- movies %>%
  mutate(genre_main = unlist(map(strsplit(as.character(movies$genres),"\\|"),1)),
         grossM = gross / 1000000,
         budgetM = budget / 1000000)

ggplot(movies, aes(x = grossM, y = genre_main, fill = genre_main)) +
  geom_density_ridges() +
  scale_x_continuous(limits = c(0, 500)) +
  labs(x = "Box Office Gross (USD Millions)",
       y = "Main Genre")
```

```
## Picking joint bandwidth of 20.9
```

```
## Warning: Removed 666 rows containing non-finite values
## (stat_density_ridges).
```

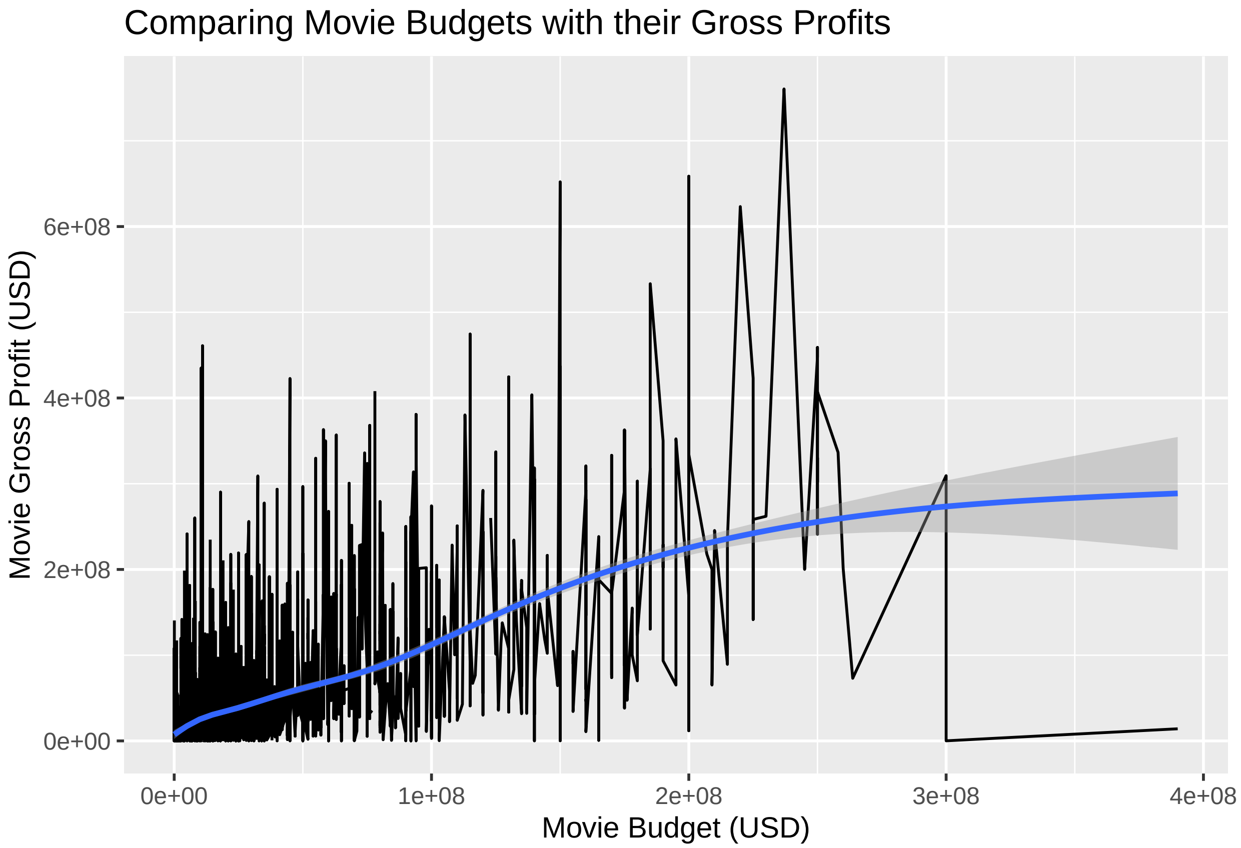


j. A few graphs showing the relationship between movie budget and gross profit

```
ggplot(movies, aes(x = budget, y = gross)) +
  geom_line() +
  stat_smooth() +
  labs(x = "Movie Budget (USD)",
       y = "Movie Gross Profit (USD)",
       title = "Comparing Movie Budgets with their Gross Profits")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

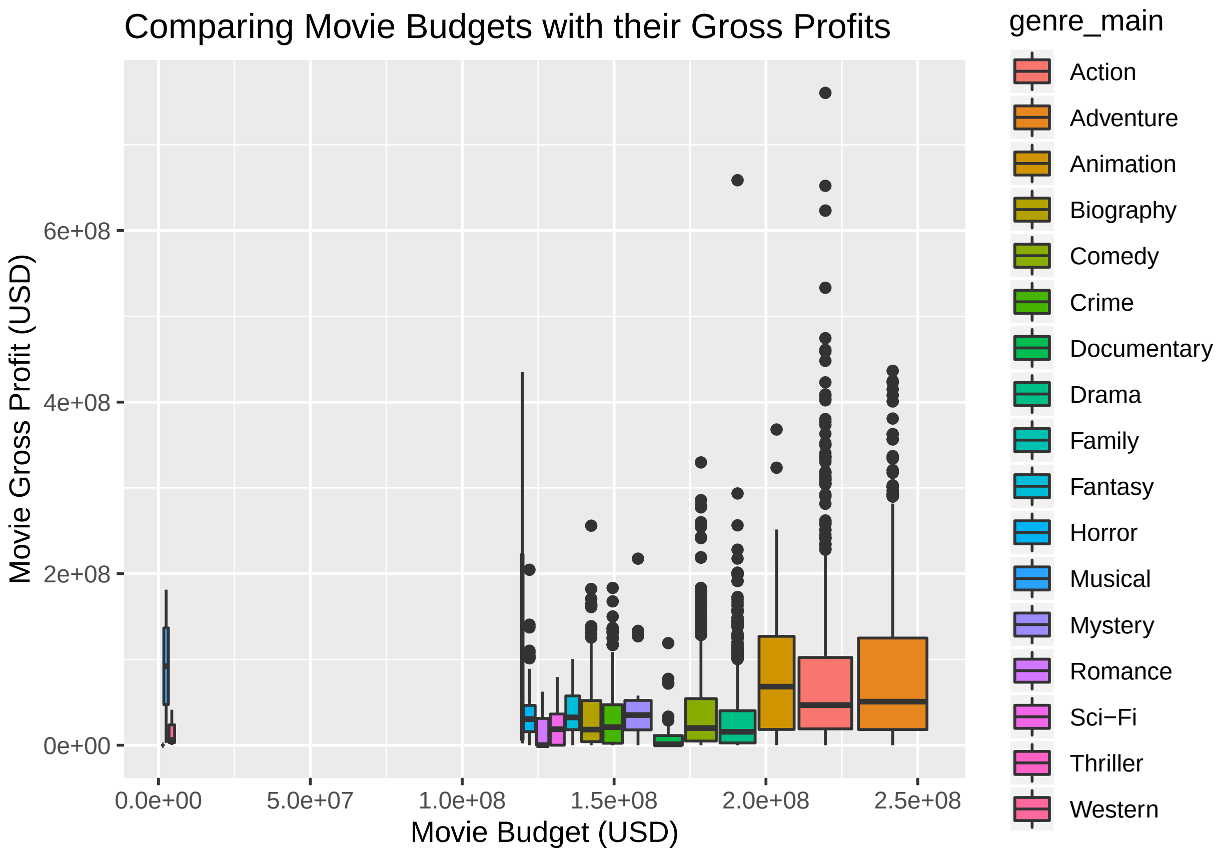
```
## Warning: Removed 660 rows containing non-finite values (stat_smooth).
```



This first graph shows an almost logarithmic relationship between budget and gross, which implies an increase in gross profit as budget increases, but with **diminishing returns**.

```
ggplot(movies, aes(x = budget, y = gross, group = genre_main, fill = genre_main)) +
  geom_boxplot() +
  labs(x = "Movie Budget (USD)",
       y = "Movie Gross Profit (USD)",
       title = "Comparing Movie Budgets with their Gross Profits")
```

```
## Warning: Removed 660 rows containing non-finite values (stat_boxplot).
```



From this second graph, I can see many high grossing outliers in the Action genre. Also the Crime genre seems to have one of the lowest average gross profit for such a high budget, as well as having so few high outliers (which themselves are some of the lowest outliers).