Data Overview: Movies!

Maryann Zhao and Leyla Akay

## Data Description

Source: <https://www.kaggle.com/tmdb/tmdb-movie-metadata>

The observational units of our dataset are individual movies. While the dataset includes 20 variables, there are ~10 that are of interest to us. Quantitative variables include budgets for the movies, popularity, revenue produced, runtime, and average votes for the movie. Categorical variables include genres, keywords that describe the movie, original language the movie was released in, production company, and production country.

### Summary Statistics

From a skim of the summary statistics, we see that the dataset has 4803 observational units and includes several variable types that can be used for categorical variables (character) and continuous variables (integer, numeric). Looking at a few of the variable that we're interested in, we see that the budget, revenue and popularity seem to be right skewed with a mean higher than the median. The production companies have a significant number of unique n but we may be interested in investigating the characteristics of the movies produced from large companies compared to smaller ones. Many of the movies fall under several genres, so we could investigate the association of any movie that falls in certain genres with the other variables such as revenue.

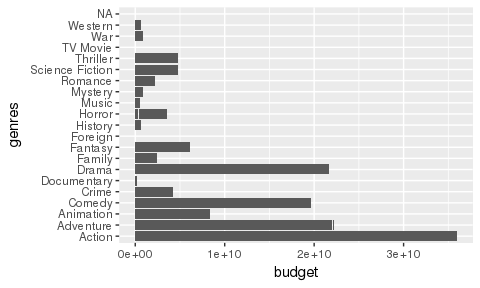
tmdb\_5000\_movies <- read\_csv("~/LM HW/SLM\_Project/tmdb\_5000\_movies-OG.csv")  
tmdb2 <- read\_csv("tmdb\_2.csv")  
  
library(readr)  
library(purrr)  
df <- readr::read\_csv("tmdb\_5000\_movies-OG.csv")  
genres.table <- purrr::map(df$genres, jsonlite::fromJSON)  
genres.table <- purrr::map\_df(genres.table, ~data.frame(x=.x), .id="movieID")  
  
skim(tmdb\_5000\_movies)

## Skim summary statistics  
## n obs: 4803   
## n variables: 20   
##   
## Variable type: character   
## variable missing complete n min max empty n\_unique  
## genres 0 4803 4803 2 230 0 1175  
## homepage 3091 1712 4803 17 138 0 1691  
## keywords 0 4803 4803 2 3783 0 4222  
## original\_language 0 4803 4803 2 2 0 37  
## original\_title 0 4803 4803 1 86 0 4801  
## overview 4 4799 4803 23 1000 0 4799  
## production\_companies 0 4803 4803 2 1155 0 3697  
## production\_countries 0 4803 4803 2 517 0 469  
## spoken\_languages 0 4803 4803 2 474 0 544  
## status 0 4803 4803 7 15 0 3  
## tagline 844 3959 4803 3 252 0 3944  
## title 0 4803 4803 1 86 0 4800  
##   
## Variable type: Date   
## variable missing complete n min max median  
## release\_date 1 4802 4803 1916-09-04 2017-02-03 2005-10-03  
## n\_unique  
## 3280  
##   
## Variable type: integer   
## variable missing complete n mean sd p0 p25  
## budget 0 4803 4803 2.9e+07 4.1e+07 0 790000   
## id 0 4803 4803 57165.48 88694.61 5 9014.5  
## runtime 80 4723 4803 106.96 22.72 0 94   
## vote\_count 0 4803 4803 690.22 1234.59 0 54   
## median p75 p100 hist  
## 15000000 4e+07 3.8e+08 ▇▂▁▁▁▁▁▁  
## 14629 58610.5 459488 ▇▁▁▁▁▁▁▁  
## 104 118 338 ▁▁▇▁▁▁▁▁  
## 235 737 13752 ▇▁▁▁▁▁▁▁  
##   
## Variable type: numeric   
## variable missing complete n mean sd p0 p25 median  
## popularity 0 4803 4803 21.49 31.82 0 4.67 12.92   
## revenue 0 4803 4803 8.2e+07 1.6e+08 0 0 1.9e+07  
## vote\_average 0 4803 4803 6.09 1.19 0 5.6 6.2   
## p75 p100 hist  
## 28.31 875.58 ▇▁▁▁▁▁▁▁  
## 9.3e+07 2.8e+09 ▇▁▁▁▁▁▁▁  
## 6.8 10 ▁▁▁▂▇▇▁▁

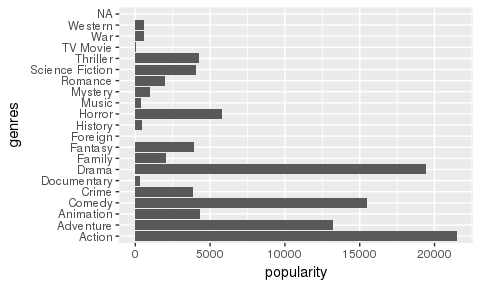
### Graphical Display

1. From the plots that evaluate the budgets and popularity for the different genres, it is interesting to see that some of the genres that have higher budgets, such as action, also seem to have high popularity. These graphs aren't necessarily useful for seeing linear associations but it gives us an idea of what the distributions between different genres look like.

genres <-tmdb2$genres1  
budget <- tmdb2$budget  
ggplot(tmdb2, aes(x=genres, y=budget))+  
 geom\_col()+  
 coord\_flip()

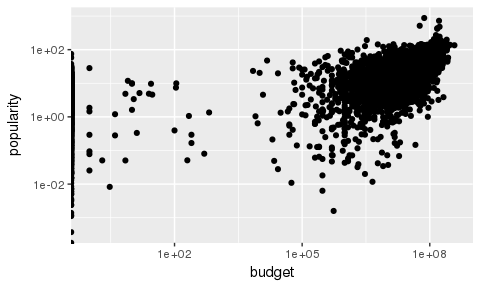


ggplot(tmdb2, aes(x=genres, y=popularity))+  
 geom\_col()+  
 coord\_flip()



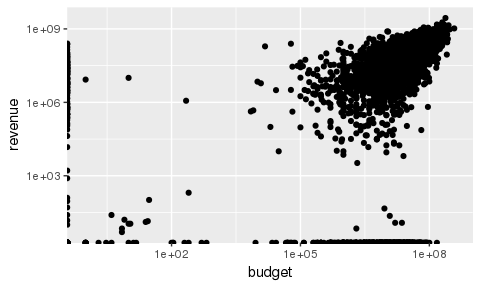
1. This scatterplot is interesting... Immediately from the dense plot on the right, it looks like there could potentially be some relationship between budget and popularity. The left side of the plot is interesting though because there are quite a few data points that have budget value of 0 but has high popularity. Not sure what to make of it yet but will look into it in the future.

ggplot(tmdb2, aes(x=budget, y=popularity)) +  
 geom\_point() +  
 scale\_x\_log10() +  
 scale\_y\_log10()



1. Similar to the plot before, these look as though there could be a potential relationship between budget and revenue, but again there are data points with zero values for revenue and budget.

ggplot(tmdb2, aes(x=budget, y=revenue)) +  
 geom\_point()+  
 scale\_x\_log10() +  
 scale\_y\_log10()



### Thoughts on data

Generally, this data set should be a representative sample of the population of movies out there, especially considering the size of the dataset. Some trends were expected, such as the association between the revenue and popularity, but some of the plots were unexpected because of the zero values. We are interested in investigating the relationships between genre and revenue or budget and popularity. There are some other variables we haven't looked at, such as production company and country, that we want to look at as well.

### Pairs

Github repository: <https://github.com/zhaozhaoo/SLM_Project> Name: SLM\_Project