OMIS 482 Project Part 1

Francisco Lozano, Amelasky Mendez, Rachel Worden, Tyler Williams and Gary Dhami

9/26/2021

## Introduction

Hi, we are a group of movie reviewers! We pick a movie every month, watch it, and get together to discuss it. We love the movies so much that we wanted to ask what makes a certain movie successful?

Using the movies dataset below, our group plans to analyze what factors affect the success of a movie. We believe a movie is successful by how much revenue a movie makes. But that is not all. By analyzing what determines success we can get a better idea of what other variables make a movie a hit. Determining this would allow movie producers to know what they can do to maximize success for their movies. We also plan to look at general trends in the movie industry to see what other patterns exist in the data. The analysis is done using data from an excel file and will be manipulated and visualized using tidyverse.

## Importing the data:

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

Movies\_Data <- read\_csv("movies\_project.csv")

## Rows: 4804 Columns: 22

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (14): genres, homepage, keywords, original\_language, original\_title, ove...  
## dbl (8): index, budget, id, popularity, revenue, runtime, vote\_average, vot...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

print(Movies\_Data)

## # A tibble: 4,804 x 22  
## index budget genres homepage id keywords original\_langua~ original\_title  
## <dbl> <dbl> <chr> <chr> <dbl> <chr> <chr> <chr>   
## 1 0 237000000 Actio~ http://~ 19995 culture~ en Avatar   
## 2 1 300000000 Adven~ http://~ 285 ocean d~ en Pirates of th~  
## 3 2 245000000 Actio~ http://~ 206647 spy bas~ en Spectre   
## 4 3 250000000 Actio~ http://~ 49026 dc comi~ en The Dark Knig~  
## 5 4 260000000 Actio~ http://~ 49529 based o~ en John Carter   
## 6 5 258000000 Fanta~ http://~ 559 dual id~ en Spider-Man 3   
## 7 6 260000000 Anima~ http://~ 38757 hostage~ en Tangled   
## 8 7 280000000 Actio~ http://~ 99861 marvel ~ en Avengers: Age~  
## 9 8 250000000 Adven~ http://~ 767 witch m~ en Harry Potter ~  
## 10 9 250000000 Actio~ http://~ 209112 dc comi~ en Batman v Supe~  
## # ... with 4,794 more rows, and 14 more variables: overview <chr>,  
## # popularity <dbl>, release\_date <chr>, revenue <dbl>, runtime <dbl>,  
## # status <chr>, tagline <chr>, title <chr>, vote\_average <dbl>,  
## # vote\_count <dbl>, cast <chr>, director <chr>, country\_production <chr>,  
## # company\_production <chr>

## Tidying the data:

Before we can manipulate the dataset and start to analyze the data, we need to tidy it up! We noticed that the columns “keywords”, “index”, “overview”, “homepage”, “ID”, “Tagline”,“runtime” and “Cast” were not relevant, so we decided to omit these columns because they add no value to our goal of trying to figure out what impacts a movie’s success. As of right now our knowledge in Rstudio doesn’t allow us to perform analytics on text columns and these categories are mostly character/text columns. In order to find the answers to our questions we plan to use mostly integer values. We want to make sure we abide by “rule of 3” ensuring each variable has its own column, row, cell. In separating release\_date we left the resulting columns to data type chr because month, day, and year aren’t columns that can be summarized with mean,max,sum, and so forth. Although they are columns that can be used to group by. Next, since our data set is not in the tidyverse package, we had to assign our data frame to the name “movies.” Because our dataset goes all the way back to 1930 and even predicts films to 2029, we found that using our data set from 1980 to 2017 would give us a more accurate look at how film has evolved. We also found that in order to measure a movies success it has to be released, so we filtered out the movies that were never released. Finally, we removed all the outliers in the data set. Most of the outliers were because the movies were really popular and broke records. For example, Avatar was a popular movie that revolutionized the movie industry since it used a bunch of new technology such as new motion capture animation technologies.

movies <-select(Movies\_Data,2,3,7:8,10:14,16:18,20:22) %>%  
 filter(!is.na(revenue),!is.na(release\_date)) %>%  
 separate(director,into = c("director\_first\_name","director\_last\_name"), sep = " ",convert = "TRUE") %>%  
 separate(release\_date, into = c("month", "day", "year"), sep = "/") %>% filter(year>= 1980 & year <= 2017,status=="Released",original\_title != "Avatar",original\_title != "Titanic")

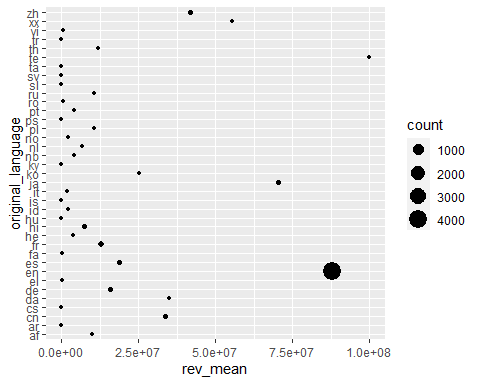
## Warning: Expected 2 pieces. Additional pieces discarded in 391 rows [52, 138,  
## 139, 153, 167, 177, 194, 218, 227, 260, 296, 324, 325, 339, 370, 371, 374, 378,  
## 384, 390, ...].

## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 11 rows [44, 220,  
## 304, 351, 654, 664, 989, 1728, 2603, 2712, 3207].

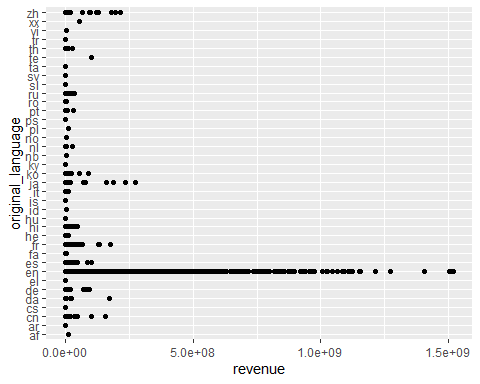
## Analysis 1: Does Language effect Revenue? - Francisco Lozano

We wanted to see if language effects revenue, so we grouped the data set by the column original\_language. After the data set was grouped we ran descriptive statistics such as mean, median, max, min, and standard deviation. Finally we plotted the average of the revenue on a scatter plot. From the graph we saw that English and Tegulu had the highest averaged revenue. We first hypothesized that English would be the first, so we were surprised Tegulu had a high average revenue. Although we then noticed that the size of the point was small meaning that not a lot movies were made in that language. We then made another graph that plotted the movies revenue based on original language to dive deeper into the data. From there we noticed the data was skewed because Tegulo only had one movie made in that language resulting in the average to be high. While English had the most movies made in that language. We also saw that English did have a higher influence in revenue since all of the top movies based on revenue were originality made in English. Confirming that language does effect revenue and the success of a movie.

movies %>%  
group\_by(original\_language) %>%  
summarise(count=n(),rev\_mean = mean(revenue,na.rm = TRUE),rev\_med=median(revenue),rev\_max=max(revenue),  
 rev\_min=min(revenue),rev\_sd=sd(revenue)) %>%  
ggplot() +  
geom\_point(mapping = aes(y = original\_language, x=rev\_mean,size=count))



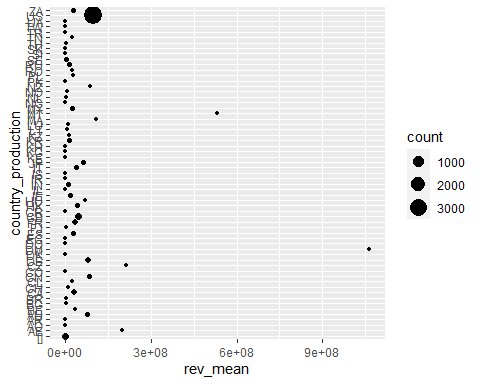
movies %>%  
ggplot() +  
geom\_point(mapping = aes(y = original\_language, x=revenue))



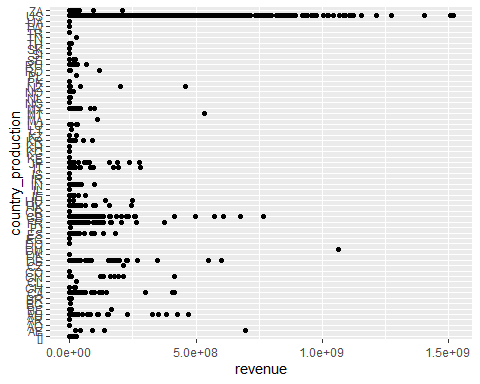
## Analysis 2: Does Country of production have an effect on revenue?- Amelasky Mendez

We wanted to see if the country production of a movie influences the revenue it generates. We started by grouping the data by country\_production. We then ran descriptive statistics on the data that would measure the Min, Max , mean, median, and standard deviation. With these results we created a scatter plot that helped us visualize the mean amount of revenue the movies generated in each country of production. We believed that the US was going to be one of the country of productions to have high average revenue. However, it is not even in the top 5, DM had the highest average revenue followed by MT, CZ, AE, and MA. Judging by the scatter plot point size they differ from that of the US. By running another plot, we can see that the US has produced more movies than all the top 5 countries combined which created an unbalanced data in the previous plot. This shows that the place of production does have an influence on the amount of revenue a movie generates.

movies%>%  
 group\_by(country\_production)%>%  
 summarise(count=n(), rev\_min=min(revenue), rev\_max=max(revenue),rev\_mean=mean(revenue), rev\_med=median(revenue),rev\_sd=sd(revenue))%>%  
   
 ggplot()+  
 geom\_point(mapping=aes(y=country\_production, x=rev\_mean, size=count))



movies%>%  
 ggplot()+  
 geom\_point(mapping = aes(y=country\_production, x=revenue))



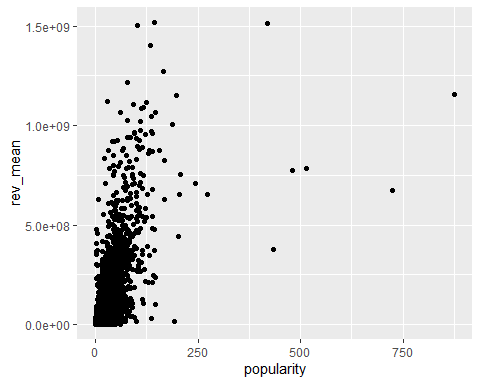
## Analysis 3: Does the movie popularity have an effect on Revenue? - Gary Dhami

We wanted to see if the popularity of a movie effects the revenue, so we grouped the data set by the column popularity. After the data set was grouped we ran descriptive statistics such as mean, median, max, and min. We can see in the tibble that the movie with the highest revenue had popularity of 144. Finally we plotted the revenue on a scatter plot. Our assumption was that the higher the popularity, the higher the revenue for the movie which does not hold true. From the graph we can see there is no real correlation between movies that have higher popularity or have higher revenues. The graph displays that the movies with the higher revenues can still have popularity less than 250 and movies with poularity higher than 250 can still have low revenues. The movie popularity has minimal effect on movie revenue.

movies %>%  
group\_by(popularity) %>%  
summarise(count=n(),rev\_mean = mean(revenue,na.rm = TRUE),rev\_med=median(revenue),rev\_max=max(revenue),  
rev\_min=min(revenue)) %>% arrange(desc(rev\_mean))

## # A tibble: 4,541 x 6  
## popularity count rev\_mean rev\_med rev\_max rev\_min  
## <dbl> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 144. 1 1519557910 1519557910 1519557910 1519557910  
## 2 419. 1 1513528810 1513528810 1513528810 1513528810  
## 3 102. 1 1506249360 1506249360 1506249360 1506249360  
## 4 134. 1 1405403694 1405403694 1405403694 1405403694  
## 5 165. 1 1274219009 1274219009 1274219009 1274219009  
## 6 77.7 1 1215439994 1215439994 1215439994 1215439994  
## 7 876. 1 1156730962 1156730962 1156730962 1156730962  
## 8 198. 1 1153304495 1153304495 1153304495 1153304495  
## 9 28.5 1 1123746996 1123746996 1123746996 1123746996  
## 10 124. 1 1118888979 1118888979 1118888979 1118888979  
## # ... with 4,531 more rows

movies %>%  
group\_by(popularity) %>%  
summarise(count=n(),rev\_mean = mean(revenue,na.rm = TRUE),rev\_med=median(revenue),rev\_max=max(revenue),   
rev\_min=min(revenue)) %>%   
arrange(desc(rev\_mean)) %>% ggplot() +   
geom\_point(mapping = aes(x = popularity, y=rev\_mean))



## Analysis 4: Does the movie rating have an effect on Revenue? - Gary Dhami

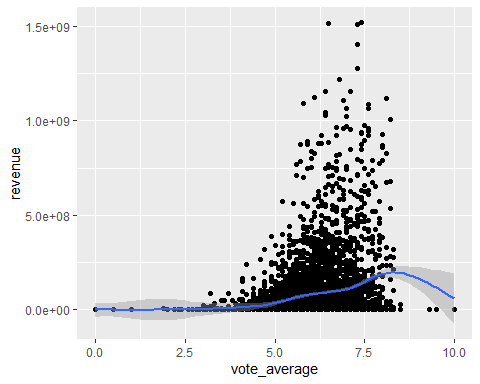
We wanted to see if the movie rating of a movie effects the revenue. So we grouped the data set by the column vote\_average. After the data set was grouped we ran descriptive statistics such as mean, median, max, and min. We can see in the tibble that the movie with the higher vote\_avergae did have the higher revenue.We plotted the revenue on a scatter plot and the graph displays a relationship between movie ratign and revenue. Our assumption was that the higher the movie rating, the higher the revenue for the movie which does hold true. we also plotted a line graph that does a great job displaying how the revenue goes up for data points with higher movie ratings. The two graphs agree with each other, leading us to believe that movie ratings has an effect on movie revenue.

movies %>%  
group\_by(vote\_average) %>% filter(revenue > 0) %>%   
summarise(count=n(),rev\_mean = mean(revenue,na.rm = TRUE),rev\_med=median(revenue),rev\_max=max(revenue), rev\_min=min(revenue)) %>% arrange(desc(rev\_mean))

## # A tibble: 60 x 6  
## vote\_average count rev\_mean rev\_med rev\_max rev\_min  
## <dbl> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 8.2 9 330485580. 234710455 1004558444 4069090  
## 2 8.1 11 329871111. 235860116 1118888979 23341568  
## 3 8 17 269141269. 109676311 926287400 7  
## 4 7.6 44 267137354. 110245152. 1084939099 246574  
## 5 7.9 25 221141141. 122126687 773328629 7103838  
## 6 7.7 35 197625353. 132511035 789804554 1270522  
## 7 7.4 74 185636219. 66225592 1519557910 15  
## 8 8.3 5 184833035. 213928762 321365567 13092000  
## 9 7.5 47 178403703. 46471023 976475550 203  
## 10 7.3 97 174952227. 85582407 1506249360 62852  
## # ... with 50 more rows

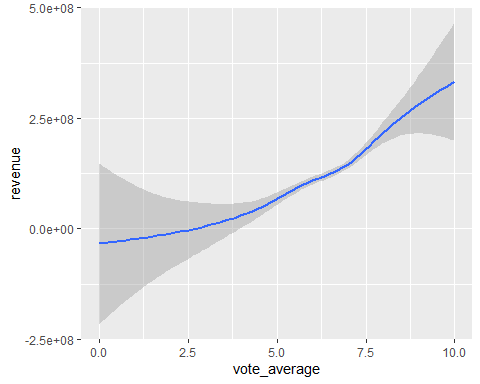
movies %>%  
group\_by(vote\_average) %>% filter(revenue > 0) %>%   
summarise(count=n(),rev\_mean = mean(revenue,na.rm = TRUE),rev\_med=median(revenue),rev\_max=max(revenue), rev\_min=min(revenue)) %>% arrange(desc(rev\_mean)) %>% ggplot(data=movies, mapping=aes(x=vote\_average, y=revenue))+geom\_point()+geom\_smooth()

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



movies %>%group\_by(vote\_average) %>% filter(revenue > 0) %>%   
ggplot() +   
geom\_smooth(mapping = aes(y = revenue, x=vote\_average))

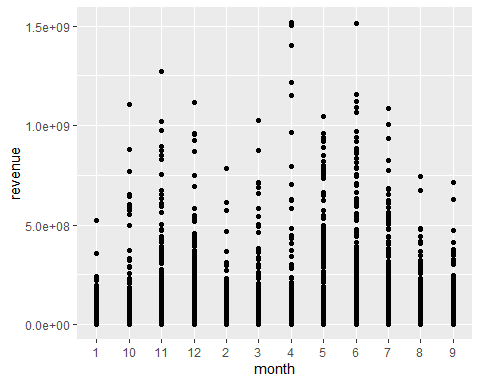
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

 ## Analysis 5: Does the month a movie is released effect the Revenue? Tyler Williams To see if the month of release effects the revenue, we first looked at some descriptive statistics for revenue grouped by month. By looking at the top average revenues by month, we see that June has the highest average revenue. We then created a scatter plot with revenue on the y-axis and release month on the x-axis to see the trend between month and revenue. The results show that some months are better to release a movie than others. Late sping to early summer produces movies with higher revenues often. The holiday months of Novemeber and Decemeber also see higher revenues. The winter and early fall generally are bad times to release a movie, as revenues are frequently lower than movies released in more optimal months

movies %>%   
 group\_by(month) %>%   
 summarise(count=n(),rev\_mean = mean(revenue,na.rm = TRUE),rev\_med=median(revenue),rev\_max=max(revenue),  
 rev\_min=min(revenue),rev\_sd=sd(revenue)) %>%   
 arrange(desc(rev\_mean))

## # A tibble: 12 x 7  
## month count rev\_mean rev\_med rev\_max rev\_min rev\_sd  
## <chr> <int> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 6 356 162186690. 68017922 1513528810 0 244016539.  
## 2 5 346 137186018. 21544286 1045713802 0 220806075.  
## 3 11 305 128702422. 47231070 1274219009 0 194320391.  
## 4 7 350 113472444. 46876923 1084939099 0 172944106.  
## 5 12 410 110227405. 46129256. 1118888979 0 159664598.  
## 6 4 330 77417733. 16099559 1519557910 0 199264844.  
## 7 3 356 74590083. 21148394 1025491110 0 137119141.  
## 8 2 313 57514950. 20455276 783112979 0 93971272.  
## 9 8 396 56323871. 16865092. 745000000 0 94803857.  
## 10 10 445 54934569. 11098131 1108561013 0 118516046.  
## 11 9 572 40410743. 8859292 716392705 0 78603607.  
## 12 1 362 29502412. 198426. 521170825 0 55639557.

movies %>%  
 ggplot() +  
 geom\_point(mapping = aes(x=month,y=revenue))



## Ideas:

In the column “status” filter by released only Status column look for correlation between released and months -stick with one variable ( revenue) Min, max, mean, standard deviation, of revenue language and country popularity(?)

how does it fit in the overall idea(research question)

## Questions we have:

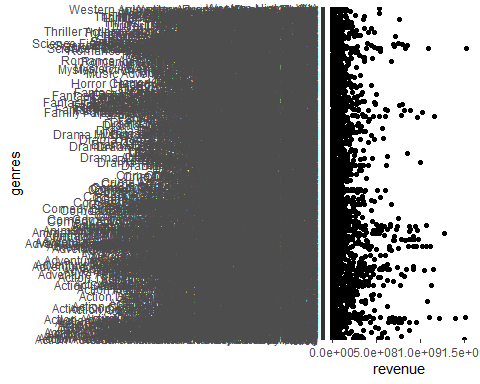
Is there a correlation between budget and the amount of revenue a movie makes? Is there a correlation between What is the most popular genre? Is there a correlation between popularity and release date Is there a correlation between how successful a movie is and their production company? Is there a correlation between movie ratings and revenue?

Movies\_Data %>%   
 group\_by(revenue) %>%   
 summarise(average = mean(revenue))

## # A tibble: 3,298 x 2  
## revenue average  
## <dbl> <dbl>  
## 1 0 0  
## 2 5 5  
## 3 7 7  
## 4 10 10  
## 5 11 11  
## 6 12 12  
## 7 13 13  
## 8 14 14  
## 9 15 15  
## 10 16 16  
## # ... with 3,288 more rows

## Mapping:

ggplot(movies)+  
 geom\_point(mapping = aes(x=revenue, y=genres))

 ## Calculations:

# #movies %>%   
# filter(!is.na(revenue)) %>%   
# summarise(count=n(),avg\_rev=mean(revenue),median\_rv=median(revenue),max\_rev=max(revenue),min\_rev=min(revenue),sd\_rev=sd(revenue))  
#   
# movies %>%   
# filter(!is.na(revenue)) %>%   
# summarise(count=n(),avg\_rev=mean(revenue))  
#   
#   
# movies %>%   
# group\_by(genres) %>%   
# summarise\_(max\_rev=max(revenue))  
#   
# movies %>%   
# filter(!is.na(revenue)) %>%   
# summarise(avg\_rev=mean(revenue))