Project Part 1 -482

Introduction

For the following movies dataset, we will be analyzing which variables affect the revenue of a movie? We feel revenue is the most important variable as it truly shows if the movie was successful or not. We believe this analysis as this information could be useful and valuable to movie production houses and director. Additionally, as a group we are really interested in movies and would like to learn more about the trend in the industry. We are using tidyverse tools to help us in our analysis. The dataset is originally an excel file that we have imported.

```
## # 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 2.37e8 Actio~ http://~  19995 culture~ en               Avatar
##  2     1 3.00e8 Adven~ http://~    285 ocean d~ en               Pirates o
f th~
##  3     2 2.45e8 Actio~ http://~ 206647 spy bas~ en               Spectre
##  4     3 2.50e8 Actio~ http://~  49026 dc comi~ en               The Dark
Knig~
##  5     4 2.60e8 Actio~ http://~  49529 based o~ en               John Cart
er
##  6     5 2.58e8 Fanta~ http://~    559 dual id~ en               Spider-Ma
n 3
##  7     6 2.60e8 Anima~ http://~  38757 hostage~ en               Tangled
##  8     7 2.80e8 Actio~ http://~  99861 marvel ~ en               Avengers:
Age~
##  9     8 2.50e8 Adven~ http://~    767 witch m~ en               Harry Pot
ter ~
## 10     9 2.50e8 Actio~ http://~ 209112 dc comi~ en               Batman v
Supe~
## # ... with 4,794 more rows, and 14 more variables: overview <chr>,
## #   popularity <dbl>, release_date <dttm>, 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 data:

As a group, we looked at the original movies dataset and decided to keep only the column we wanted to research. Columns that were not going to be used, had no meaningful information or were redundant, such as overview, original title, website, keywords and tagline, were removed. Afterwards we tidied the dataset. We met rule 1 as each variable had its own column. We met rule 2 as observation had its own row. We met rule 3 by

ensuring each value had its own cell. To do so, we separate values with mutiple cell into its own row. We got rid additional value we did not need in genre and company to have a more clear,concise and understandable data set.

We also filtered the dataset to only include movies in and after 2003 to 2017 (last 15 years) to find current impact on Revenue variable. The data set stop at the year 2017 except for a few outliers showing the date such as 2027 which we removed as they were likely errors. Comparing revenue trends from 1930 to 2020's would not be correct as the industry is very different now. Plus it might provide false result as the data does not take inflation into consideration. We filtered to only have English movies as that is the most popular and common language used around the world. We filtered out movies that were stuck in post production or were rumored as those movies would not actually have revenue and would falsely skew our dependent variable. Lastly, we filtered out any movies with a revenue of $0 as it seem that that is caused by a lack of information about the revenue generated. shows how you actually met rules.

```
movies_s <- movies %>%
  select(1:3,7,10:14,16,17,20:22) %>%
  separate(genres,into = c("genre")) %>% #tidying genre to be only 1 genre (i
.e. Action Mystery > Action)
  separate(director,into = c("director_first","director_last")) %>%
  separate(company_production,into = c("company")) %>%  #shortens company nam
e to 1 word
  filter( release_date >= "2003-01-01" & release_date <= "2017-12-31") %>% #o
nly movies in from 2007 to 2017
  filter(original_language == "en") %>% #only en movies
  filter(status == "Released") %>%
  filter(revenue > 0)
```

```
## # A tibble: 1,926 x 15
##    index budget genre original_langua~ popularity release_date        reve
nue
##    <dbl>  <dbl> <chr> <chr>                 <dbl> <dttm>                <d
bl>
##  1     0 2.37e8 Acti~ en                    150.  2009-12-10 00:00:00  2.7
9e9
##  2     1 3.00e8 Adve~ en                    139.  2007-05-19 00:00:00  9.6
1e8
##  3     2 2.45e8 Acti~ en                    107.  2015-10-26 00:00:00  8.8
1e8
##  4     3 2.50e8 Acti~ en                    112.  2012-07-16 00:00:00  1.0
8e9
##  5     4 2.60e8 Acti~ en                     43.9 2012-03-07 00:00:00  2.8
4e8
##  6     5 2.58e8 Fant~ en                    116.  2007-05-01 00:00:00  8.9
1e8
##  7     6 2.60e8 Anim~ en                     48.7 2010-11-24 00:00:00  5.9
2e8
##  8     7 2.80e8 Acti~ en                    134.  2015-04-22 00:00:00  1.4
```

```
1e9
##  9     8 2.50e8 Adve~ en                       98.9 2009-07-07 00:00:00  9.3
4e8
## 10     9 2.50e8 Acti~ en                      156.  2016-03-23 00:00:00  8.7
3e8
## # ... with 1,916 more rows, and 8 more variables: runtime <dbl>, status <c
hr>,
## #   title <chr>, vote_average <dbl>, director_first <chr>, director_last <
chr>,
## #   country_production <chr>, company <chr>
```

Next we decided to learn more about our dependent variable Revenue by finding the descriptive stats. Our tidied and filtered dataset has 1926 movies, with a range of zero to billion in revenue. We will try to find out some of the variables has a impact on revenue.

```
movies_s %>%
  filter(!is.na(revenue)) %>%
  summarise(count=n(),avg.rev=mean(revenue),median.rev=median(revenue),max_re
v=max(revenue),min.rev=min(revenue),sd.rev=sd(revenue)) %>%
  arrange(desc(avg.rev))
```

```
## # A tibble: 1 x 6
##   count    avg.rev median.rev    max_rev min.rev     sd.rev
##   <int>      <dbl>      <dbl>      <dbl>   <dbl>      <dbl>
## 1  1926 136860734.  63271536. 2787965087       7 210238351.
```

Next we decided to learn more about the impact of Genre on Revenue. We picked genre first as the we it is the most broad-scoped variable.We have all the descriptive stats for genre. We filtered to only show genres with over 50, which is approximately 10% of the max number of movies in a particular genre. Generally an ideal sample consists has at least 10% of the population. Our sample is bigger the base 10% would have confined our results to just a few genres. We are following that standard in an effort to get accurate results and analysis. A smaller sample size could be easily skewed by a couple high revenue generating movies. Animation movies have the highest average revenue, which is surprising as generally those movies are aimed at a children. It is possible that the avg revenue is higher because parents have to accompany kids for the movies even if they personally might not be interested in the movie.

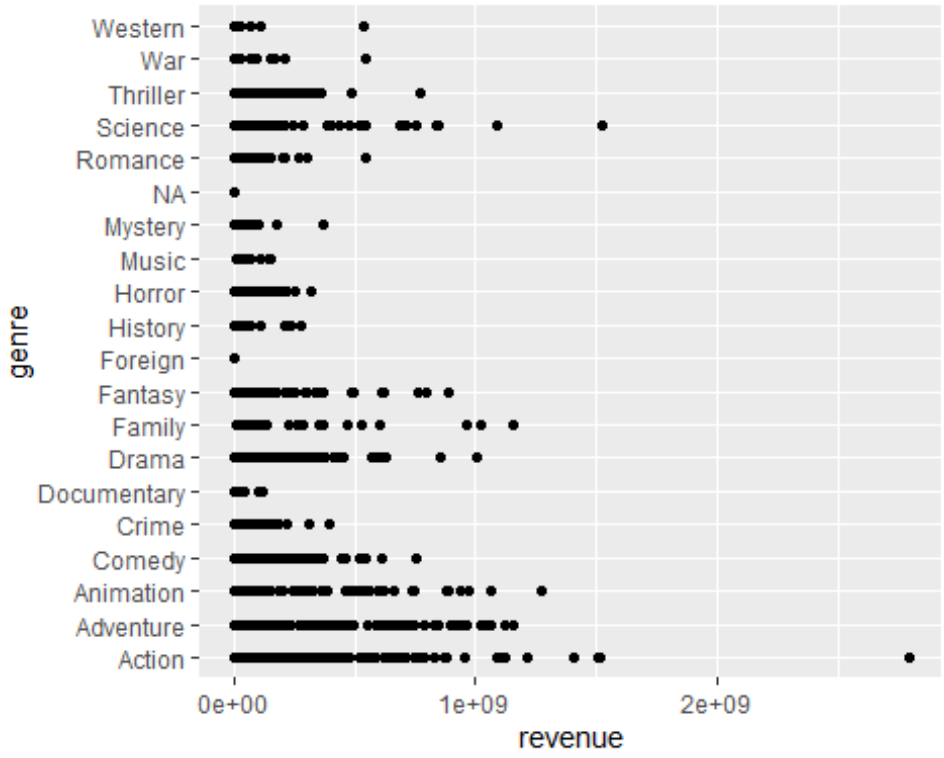
```
movies_s %>%
  group_by(genre) %>%
  summarise(count=n(),avg.rev=mean(revenue),median.rev=median(revenue),max_re
v=max(revenue),min.rev=min(revenue),sd.rev=sd(revenue)) %>%
  arrange(desc(avg.rev)) %>%
  filter(count>50)
```

```
## # A tibble: 9 x 7
##   genre     count    avg.rev median.rev    max_rev min.rev     sd.rev
##   <chr>     <int>      <dbl>      <dbl>      <dbl>   <dbl>      <dbl>
## 1 Animation    72 339690689. 264561814  1274219009   73706 301735288.
## 2 Adventure   154 297027143. 186951741  1153304495  480314 291644835.
```

```
## 3 Fantasy      52 214087987. 148808887   890871626    6399 216002555.
## 4 Action      328 199775506.  92042250. 2787965087     126 292983815.
## 5 Thriller     87  96542742.  60222298   767820459   56825 122880212.
## 6 Comedy      413  84329592.  49830607   752600867       7  98162263.
## 7 Horror      114  72513356.  63645516.  320170008    1632  64701628.
## 8 Drama       439  71849859.  34234008  1004558444      92 110973075.
## 9 Crime        70  66060795.  40289200   392000694   10018  74580029.
```

Next we decided to graph the show the overall trend in revenue between the various genres and to find any outliers. We noticed that one particular movie in action genre is big outlier.

```
ggplot(movies_s)+
  geom_point(mapping = aes(x=revenue,y=genre))
```



Next, we decided to find out which movie was the outlier. We found out that Avatar is the outlier and it makes sense as that movie was really popular and set multiple revenue records.

```
movies_s %>%
  group_by(title) %>%
  summarise(max.rev=max(revenue)) %>%
  arrange(desc(max.rev))
```

```
## # A tibble: 1,926 x 2
##    title                                max.rev
##    <chr>                                  <dbl>
```

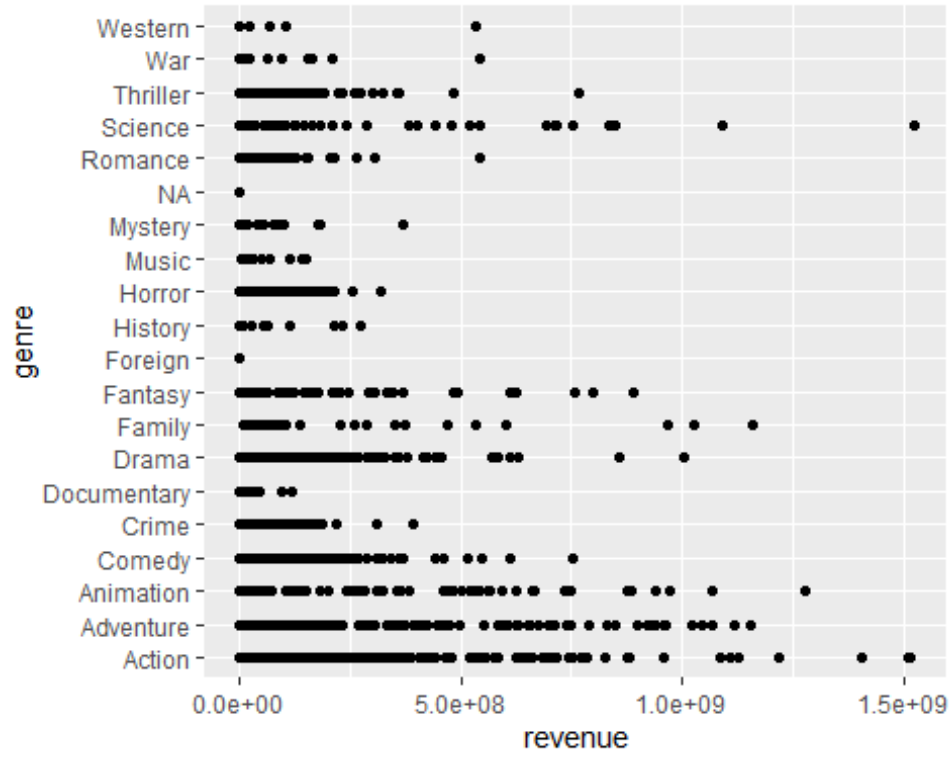
```
##  1 Avatar                                2787965087
##  2 The Avengers                          1519557910
##  3 Jurassic World                        1513528810
##  4 Furious 7                             1506249360
##  5 Avengers: Age of Ultron               1405403694
##  6 Frozen                                1274219009
##  7 Iron Man 3                            1215439994
##  8 Minions                               1156730962
##  9 Captain America: Civil War            1153304495
## 10 Transformers: Dark of the Moon        1123746996
## # ... with 1,916 more rows
```

Upon finding out which movie is the outlier, We decided to filter it out to un-skew our results and make it more visually appealing.

```
movies_1 <- movies_s %>%
 filter(!(title=="Avatar")) %>%
 filter(!(genre=="NA" & genre=="Foreign"))
```

Finally, we have our graph without any massive outliers in either direction. The results show that Action, Adventure, and Animation movies are some of the top revenue grossing movies.

```
ggplot(movies_1)+
  geom_point(mapping = aes(x=revenue,y=genre))
```



Next, we decided to see the impact of budget on revenue. We noticed that it is difficult to analyze the result with out a graph. As such we decided to create a graph.

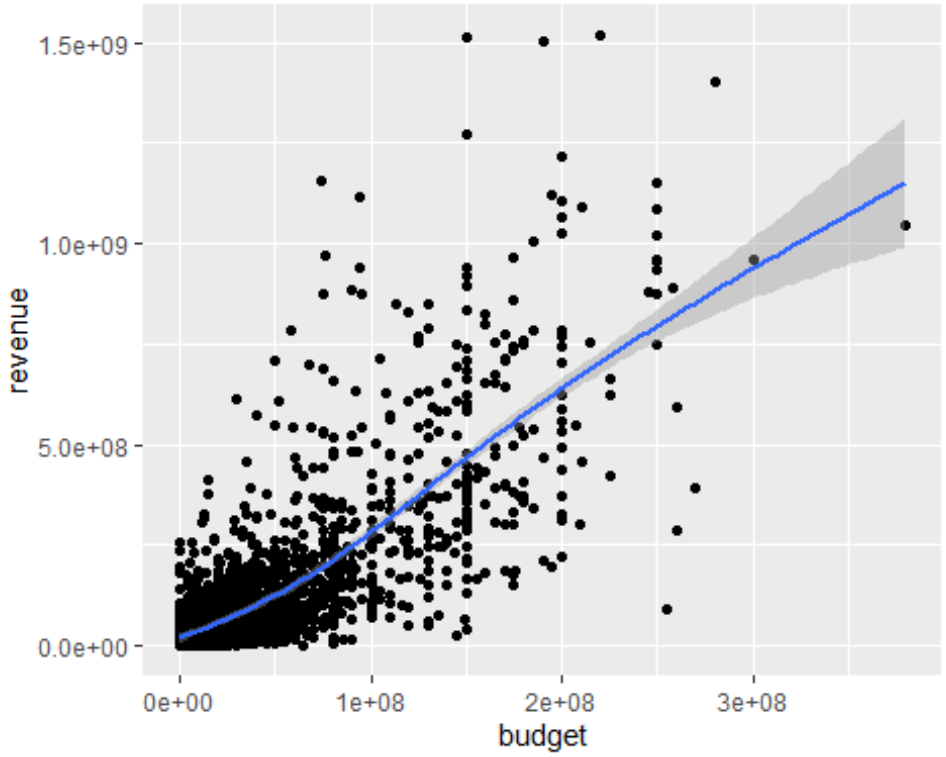
```
movies_1 %>%
  filter(!is.na(revenue)) %>%
  group_by(budget) %>%
  summarise(avg.rev=mean(revenue)) %>%
  arrange(desc(avg.rev))
```

```
## # A tibble: 246 x 2
##       budget     avg.rev
##        <dbl>       <dbl>
##  1 220000000 1519557910
##  2 280000000 1405403694
##  3 380000000 1045713802
##  4  94000000 1029612258.
##  5 250000000  966106140.
##  6 300000000  961000000
##  7 258000000  890871626
##  8 245000000  880674609
##  9 113000000  850000000
## 10 210000000  775382326
## # ... with 236 more rows
```

We created a scatter plot as both revenue and budget are numerical variables.We saw that generally as budget went up the revenue of the movie went up. There were a few exception to these though.

```
ggplot(movies_1)+
  geom_point(mapping = aes(x=budget,y=revenue))+
  geom_smooth(mapping = aes(x=budget,y=revenue))
```

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

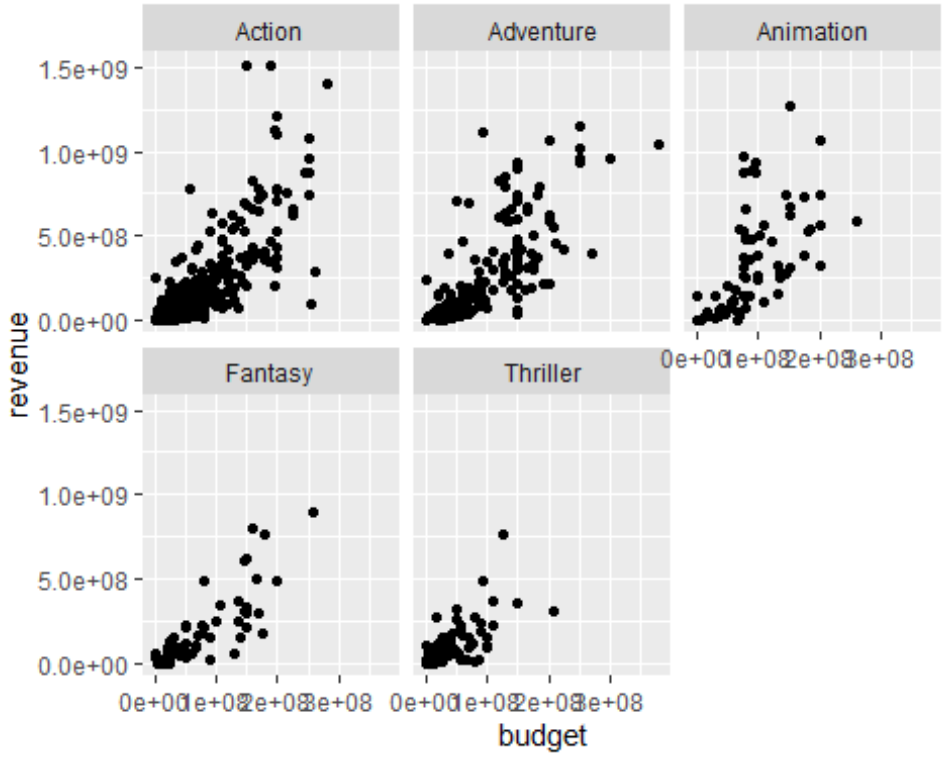


We also wanted to see the relationship of genre with revenue and budget and decided to add the top 5 genre we found earlier in to the graph.

```
movies_2 <- movies_1 %>%
filter(genre == "Animation" | genre=="Animation" | genre=="Adventure" | genre
=="Fantasy" | genre=="Action" | genre=="Thriller")
```

Upon, adding genre to our budget x revenue graph, we learned that adventure movies some of largest budget, while Action movies seems to have some of the highest revenues. Surprisingly, Animation movies often have a smaller or similar budget and comparatively seems to have great revenue.

```
ggplot(movies_2)+
  geom_point(mapping = aes(x=budget,y=revenue))+
  facet_wrap(~genre)
```



Next, we decided to see the impact of the country of production on Revenue. We started of by finding the descriptive statistics and saw that USA by far makes the most movies. This is likely due to Hollywood being in the United States and our dataset only having English language movies. We filtered the top 5 movies by count as we wanted to focus on high revenue countries but also have enough data to make accurate predictions from. We learned that for revenue, Denmark is a best in terms of average revenue, followed closely by the USA.

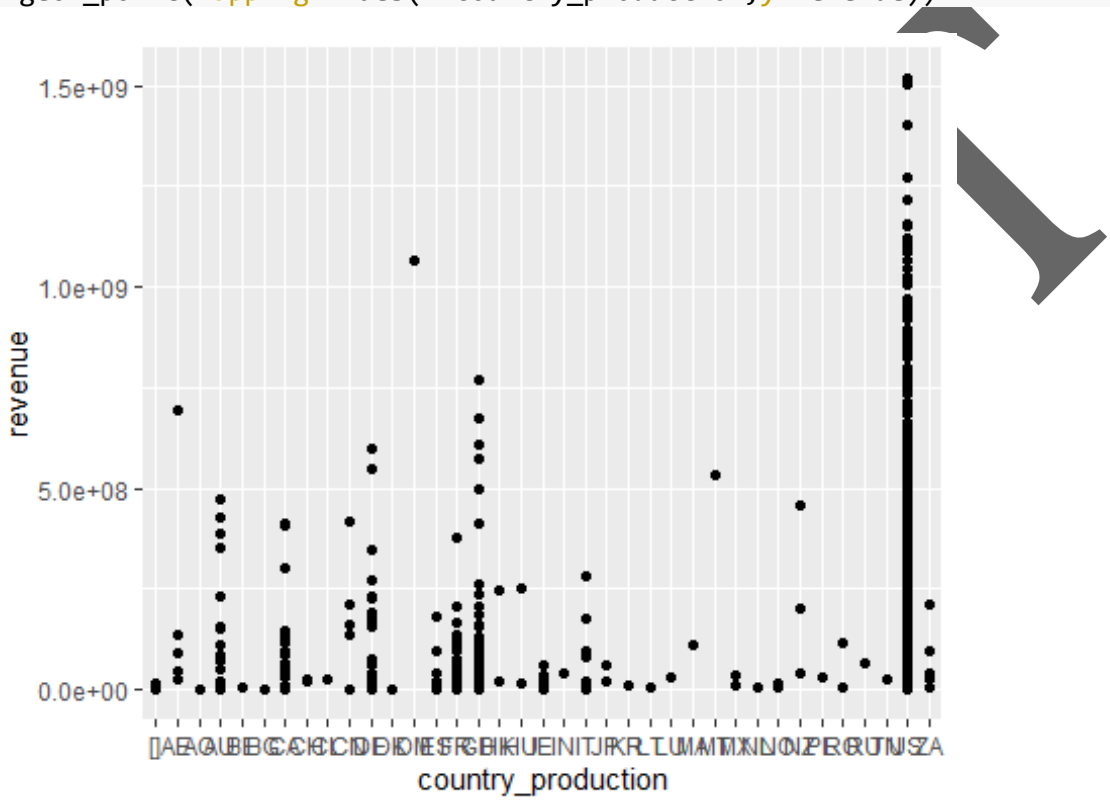
```
movies_1%>%
  group_by(country_production) %>%
  summarise(count=n(),avg.rev=mean(revenue),median.rev=median(revenue),max_re
v=max(revenue),min.rev=min(revenue),sd.rev=sd(revenue)) %>%
  arrange(desc(count)) %>%
  top_n(5,count)

## # A tibble: 5 x 7
##   country_production count    avg.rev median.rev    max_rev min.rev     sd
.rev
##   <chr>              <int>      <dbl>      <dbl>      <dbl>   <dbl>      <
dbl>
## 1 US                  1641 142786285.   68267862  1519557910      12 208716
141.
## 2 GB                    94  86853182.   30800533   767820459      46 148023
684.
## 3 FR                    34  75711383.   56896224   376141306  374743  74602
989.
```

```
## 4 CA                    27  80108453.  40547440   413106170   25000 115478
833.
## 5 DE                    26 150078769. 114479092.  599045960  871279 155869
553.
```

Next, we wanted to provide a visualization of how many movies were produced in each country and the distribution of their revenues.

```
ggplot(movies_1)+
  geom_point(mapping = aes(x=country_production,y=revenue))
```



Next. wanted to look at comparing popularity with revenue, using descriptive statistics to understand what the revenue and popularity data indicating with evidence to back up claims. Specifically, looking at whether or not the revenue increases as the movies popularity increases which would indicate a positive correlation. Also check to see what outliers exist, for example if they're any movies generate an abnormal amount of revenue compared to the rest. I wanted to organize this data to show the greatest values in descending order to make it easier to find outliers and analyze the dataset. Finally, I only want to focus on movies with a popularity score greater than 50 because these are movies that fall in to average to good spectrum. A popularity score less than 50 is generally a poor movie and as such we did not include them in this dataset.

```
movies_1%>%
  group_by(popularity) %>%
  summarise(median_rev=median(revenue),avg_rev=mean(revenue),max_rev=max(reve
nue),min_rev=min(revenue)) %>%
```

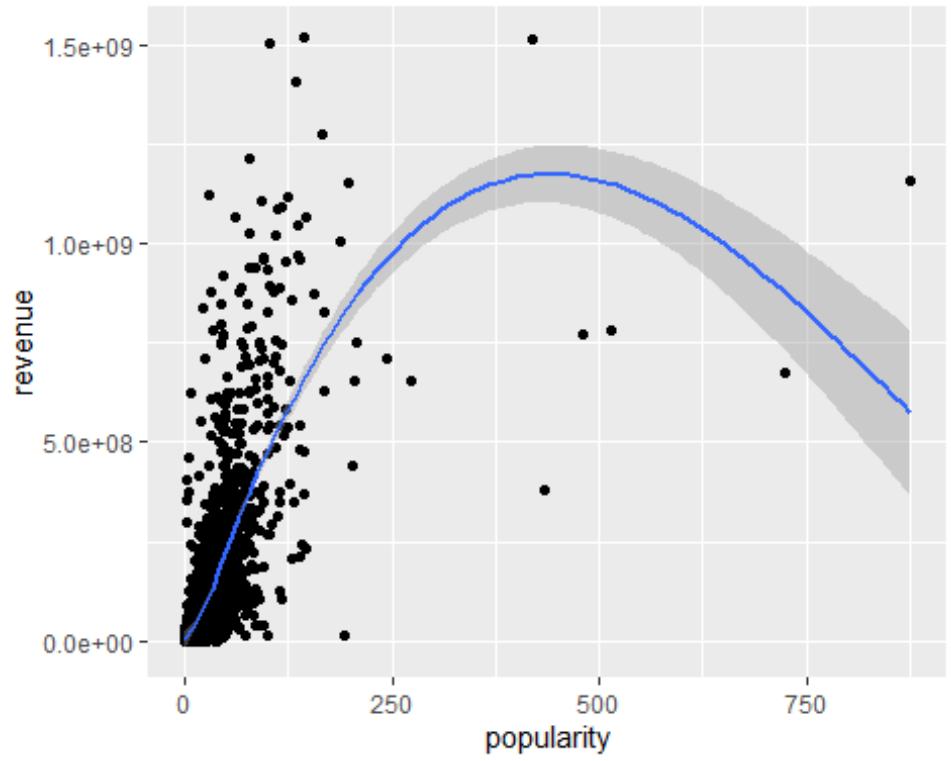
```
  arrange(desc(avg_rev)) %>%
  filter(popularity>50)
```

```
## # A tibble: 323 x 5
##    popularity median_rev    avg_rev    max_rev    min_rev
##         <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
##  1      144.  1519557910 1519557910 1519557910 1519557910
##  2      419.  1513528810 1513528810 1513528810 1513528810
##  3      102.  1506249360 1506249360 1506249360 1506249360
##  4      134.  1405403694 1405403694 1405403694 1405403694
##  5      165.  1274219009 1274219009 1274219009 1274219009
##  6       77.7 1215439994 1215439994 1215439994 1215439994
##  7      876.  1156730962 1156730962 1156730962 1156730962
##  8      198.  1153304495 1153304495 1153304495 1153304495
##  9      124.  1118888979 1118888979 1118888979 1118888979
## 10       93.0 1108561013 1108561013 1108561013 1108561013
## # ... with 313 more rows
```

Show the relationship between revenue and popularity, based on the graph the data shows that typically movies with a greater popularity generate a higher revenue, but this is not always the case. For exampe the movie that generated the largest revenue was not considered the most popular. Based on the graph movies that didn't generate a large amount of revenue were not very popular as a result

```
ggplot(movies_1)+
  geom_point(mapping = aes(x=popularity, y=revenue))+
  geom_smooth(mapping = aes(x=popularity,y=revenue))
```

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



Next we decide to explore the impact of the vote_average, which is the average IMDB score for each movie, and revenue.The descriptive statistics shows that overall a majority of movies score between the 5.0 and 8.0 rating and the revenue can be significant for those movies. Revenue is much higher for movies that score in the upper 7's then the rest, yet most movies seems to be the in the 6.0 range, which on average score less.

```
movies_1 %>%
  group_by(vote_average) %>%
  summarise(count=n(),
            median.rev=median(revenue),
            avg.rev=mean(revenue),
            max_rev=max(revenue),
            min.rev=min(revenue)) %>%
  arrange(desc(count))
```

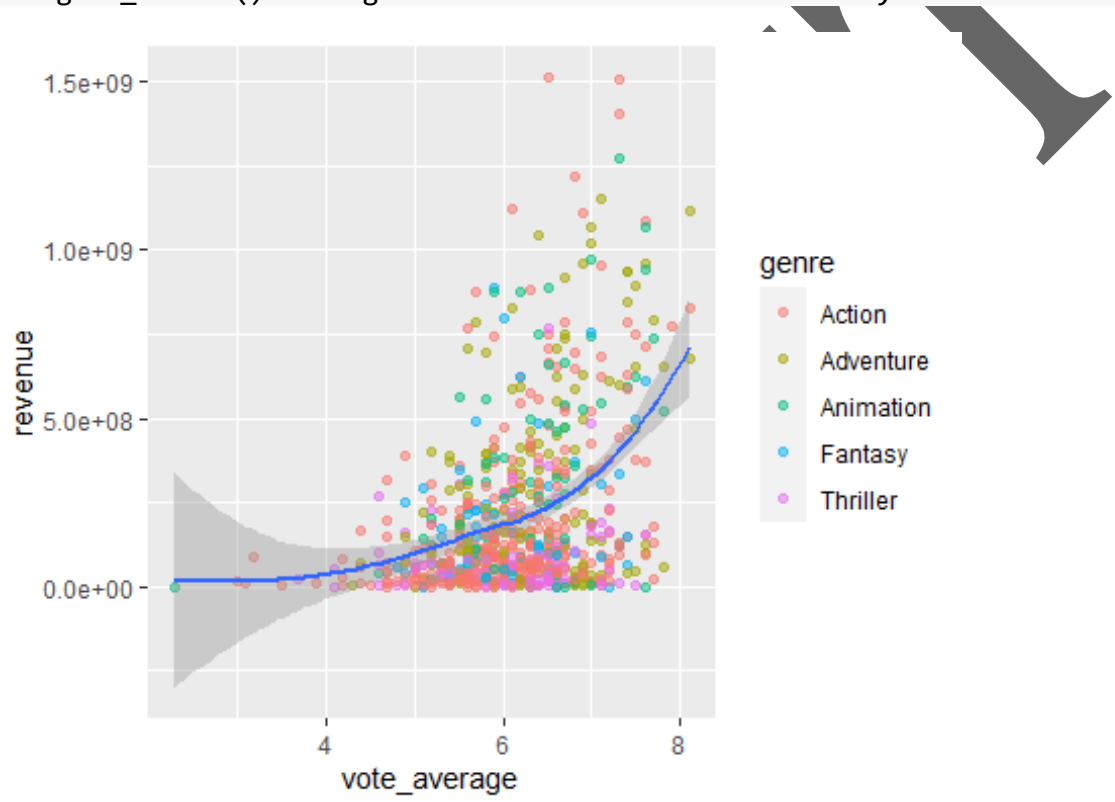
```
## # A tibble: 55 x 6
##    vote_average count median.rev    avg.rev    max_rev min.rev
##           <dbl> <int>      <dbl>      <dbl>      <dbl>   <dbl>
##  1          6     109  57231524   99965644.  836297228   31081
##  2          6.5   100  81573724. 151350556. 1513528810   44462
##  3          6.2    99  82966152  135063919.  877244782   10018
##  4          6.7    97  77000000  163717040.  966550600  113783
##  5          6.6    93  48902953  123979283.  752100229    7202
##  6          5.8    91  77920346  130592496. 1091405097   17479
##  7          6.3    91  51070807  119467763.  880674609   10000
##  8          6.1    90  61586298  125797929. 1123746996      12
##  9          5.9    86  82495204  128671490.  890871626  111731
```

```
## 10          6.4    85  80916492  162990226. 1156730962   42145
## # ... with 45 more rows
```

Next, we decided to graph the results in order to provide a visual aid. The smoothing line confirms our analysis that revenue seems to increase as the average vote increases. We also learned that Adventure movies have some of the highest votes but not the highest revenue. Overall, there is mix of all genre that do well and have a great score, which shows that an approx 7.0 to 7.5 score is good indicator that a movie will do well.

```
ggplot(movies_2,mapping = aes(x=vote_average,y=revenue))+
  geom_point(mapping = aes(color = genre),alpha=0.5)+
  geom_smooth()

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Lastly we created mutations that takes 2 variables(budget & revenue) and analyzes the net profit/loss and percent gain/loss of each movie. From here we wanted to see the top 10 movies which had the most profit and arranged the table by highest percent gain. The result shows that 6 out the 10 movies are in the action genre.

```
movies_3 <- movies_1 %>%
  filter(budget > 1000) %>%
  mutate(net_profit_loss= revenue-budget) %>%
  mutate(percent_gain_loss=(net_profit_loss/budget)*100) %>%
  select("budget","revenue","title","genre","net_profit_loss","percent_gain_l
oss") %>%
```

```
  top_n(10,net_profit_loss)
  movies_3
```

```
## # A tibble: 10 x 6
##      budget  revenue title                genre  net_profit_loss percent_ga
in_lo~
##       <dbl>    <dbl> <chr>                <chr>            <dbl>
<dbl>
##  1   2.80e8   1.41e9 Avengers: Age of U~ Action      1125403694
402.
##  2   2.20e8   1.52e9 The Avengers        Scien~      1299557910
591.
##  3   1.50e8   1.51e9 Jurassic World      Action      1363528810
909.
##  4   2.00e8   1.11e9 Skyfall             Action       908561013
454.
##  5   2.00e8   1.22e9 Iron Man 3          Action      1015439994
508.
##  6   1.90e8   1.51e9 Furious 7           Action      1316249360
693.
##  7   1.95e8   1.12e9 Transformers: Dark~ Action       928746996
476.
##  8   1.50e8   1.27e9 Frozen              Anima~      1124219009
749.
##  9   9.40e7   1.12e9 The Lord of the Ri~ Adven~      1024888979
1090.
## 10   7.40e7   1.16e9 Minions             Family      1082730962
1463.
```

Conclusion

In conclusion, multiple independent variables (Genre, Budget, Country, Popularity, Rating) have an effect on our dependent variable, Revenue. We believe that to have a very successful movie in term of Revenue, the best focus should be on Action, Adventure, and Animation. A combination of any of those would work well too. The movie should have an above average budget but as that generally improves the quality of the movie, increasing its overall appeal. A lot of English movies are produced in the United States but that is not the only great location. Countries such as Denmark, Great Britain, France, and Canada are all viable locations. A popularity rating in the range of 150 is a good indicator that the movie will have a great revenue collection. Lastly, an IMBD rating of 7.1 to 7.5 and beyond is also a good indicator that the movie will be successful and generate a high revenue. Overall, we believe that a combination of these factors will lead to the successful movie in terms of revenue.