DSCT Assignment 4 Report

Done by: Zheng Yi Tao, Zheng Min, Vincent Tan and Juin Aing

Data Exploratory (Q1)

1) Checking for missing records using *missmap* function

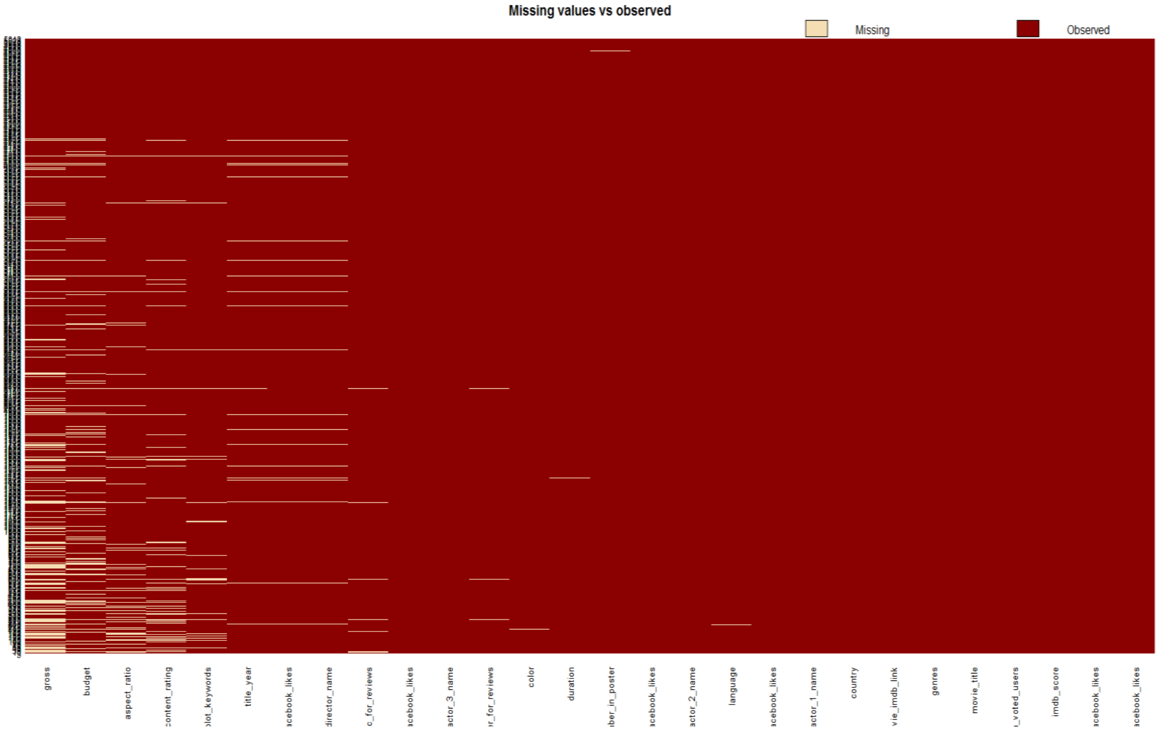
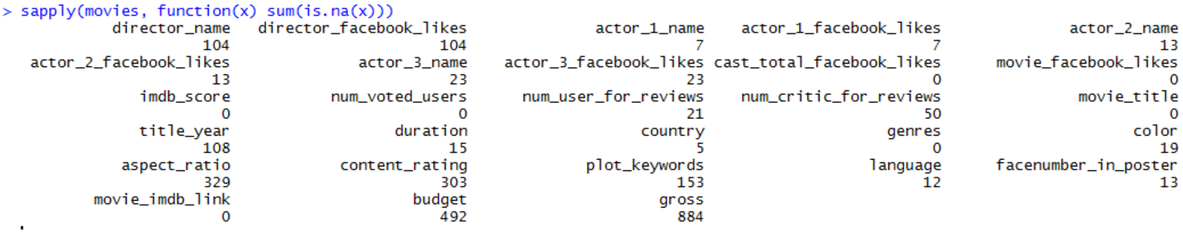


Figure 1: There are many missing data especially gross and budget.

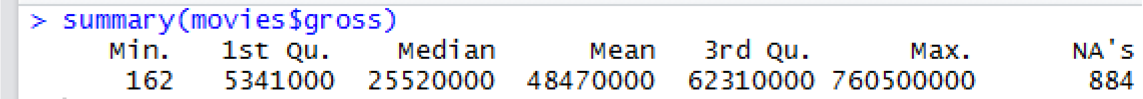


```
> sapply(movies, function(x) sum(is.na(x)))
            director_name   director_facebook_likes              actor_1_name    actor_1_facebook_likes              actor_2_name 
                      104                       104                         7                         7                        13 
   actor_2_facebook_likes              actor_3_name    actor_3_facebook_likes cast_total_facebook_likes      movie_facebook_likes 
                       13                        23                        23                         0                         0 
               imdb_score           num_voted_users      num_user_for_reviews    num_critic_for_reviews               movie_title 
                        0                         0                        21                        50                         0 
               title_year                  duration                   country                    genres                     color 
                      108                        15                         5                         0                        19 
             aspect_ratio            content_rating             plot_keywords                  language     facenumber_in_poster 
                      329                       303                       153                        12                        13 
          movie_imdb_link                    budget                     gross 
                        0                       492                       884 
```

Figure 2: Sum of missing values for each variables

More specifically, gross has the highest of 884 missing data and budget has 492.

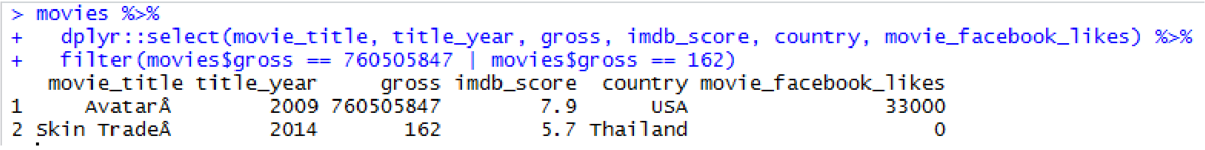
2) Summary of gross, the target variable



```
> summary(movies$gross)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
      162   5341000  25520000  48470000  62310000 760500000       884 
```

Figure 3: Summary of Gross values

The summary show that gross has a minimum of 162 and max of 760500000 which is a huge range, thus having huge differences between median and mean as mean is more susceptible to huge range.



```
> movies %>%
+   dplyr::select(movie_title, title_year, gross, imdb_score, country, movie_facebook_likes) %>%
+   filter(movies$gross == 760505847 | movies$gross == 162)
   movie_title title_year     gross imdb_score  country movie_facebook_likes
1      AvatarÂ        2009 760505847        7.9      USA                33000
2 Skin TradeÂ        2014       162        5.7 Thailand                    0
```

Figure 4: Highest and lowest grossing movies

Figure 4 shows that Avatar is the highest grossing movie and Skin Trade is the lowest.

3) Looking at gross earnings over the years

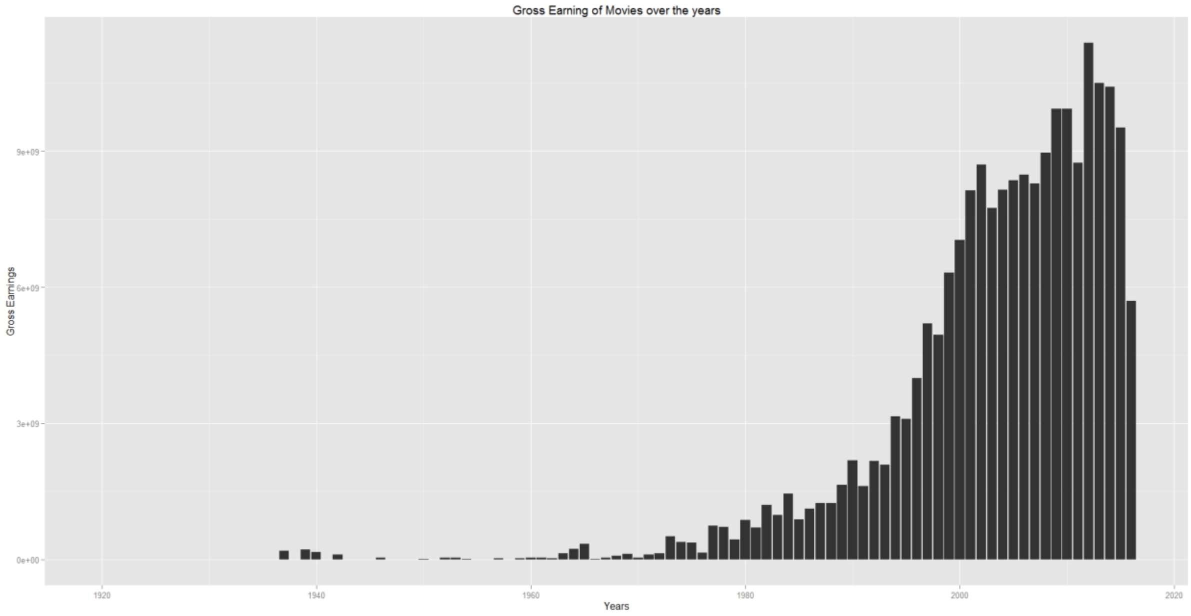


Figure 5: Gross earnings of movies over the years

Figure 5 showed that movies started generating more revenues only from the 1980s onwards and has become significantly high from the year 2000 onwards.

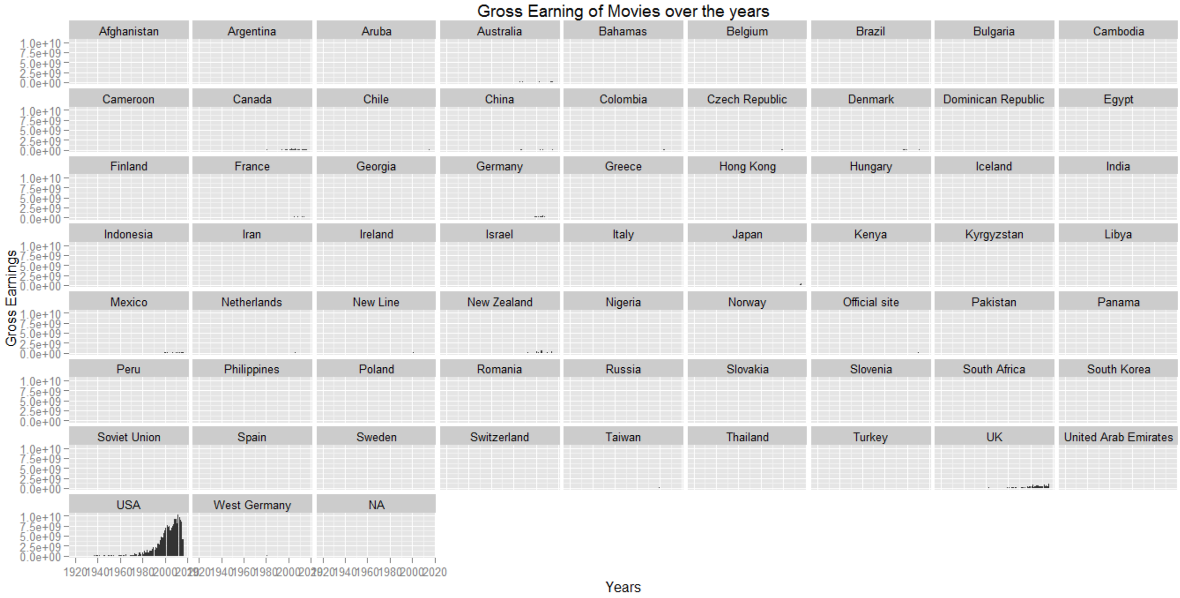


Figure 6 Gross over for each country

USA dominated the movies industry as shown in the graph, with UK and Canada following next with a super big gap. The gross sales are so high for USA could be that, the Hollywood produce thousands of movies yearly and quite a good portion of them cast globally.

Data cleaning and Transformations (Q2,3)

Initial steps taken:

1) Duplicates removed.
2) All the movie titles has its noisy character removed.

Content Ratings:

1) Removed all the content rating that has "TV" in them, because the business need is to find out movies' impact on gross and not TV shows.
2) Found out that the rating formats are taken from USA and Australia.
3) Reduced the number of categories to 5:
    a. "GP" and "G" to just "**G**"
    b. "PG-13", "Approved" and "Passed" to just "**PG**"
    c. "X" and "R" to just "**R**"
    d. "NC-17" and "M" to just "**NC**"
    e. "Not Rated" and "Unrated" to just "**Unrated**"

Countries:

1) 2 erroneous data from countries with values "Official site" and "New Line" are changed to "USA" based on the movie title.
2) Selecting only movies that are from "USA"
    a. Reason is that Gross and Budget are use based on the country's currency. Thus, to normalise the result, only movies made in "USA" are selected.
    b. This has reduced 1189 data out of 4844.

Missing Values:

1) Converted categorical data into Factors.
2) Removing all of values that is "NA" from variables Gross and Budget.
3) For Colours variable, replaced 1 missing field with "Color". This is because there are only 89 "Black and White" as compared to 2905 "Color".
4) Changing all movies' language to English including those that have missing language values.
    a. This is because all USA movies even though may have multiple languages, all of them does still have English.
5) Removing all rows with more than 2 missing values.
6) Removing all content_rating, aspect_ratio and facenumber_in_poster values that are missing.
7) Removed outlier from aspect_ratio and cast_total_facebook_likes.

Filtered and reordered all of the columns, leaving only those that are useful to build a predictive model.

Dichotomous Variables:

1) Genres
    a. Removed all genres that are “short”
    b. Splitting all 26 genres into dichotomous variables
    c. Combining columns of genres that are related, reducing 26 genres to just 12:
        i. Action, Adventure, War and Western becomes **Action**
        ii. Musical, Drama, Film-Noir and Music becomes **Drama**
        iii. Sci-Fi and Fantasy becomes **Fantasy**
        iv. Thriller and Mystery becomes **Thriller**
        v. Biography, Documentary and History becomes **Educational**
    d. All of the remaining genres are
        i. Action
        ii. Drama
        iii. Fantasy
        iv. Thriller
        v. Educational
        vi. Animation
        vii. Comedy
        viii. Crime
        ix. Family
        x. Horror
        xi. Romance
        xii. Sport

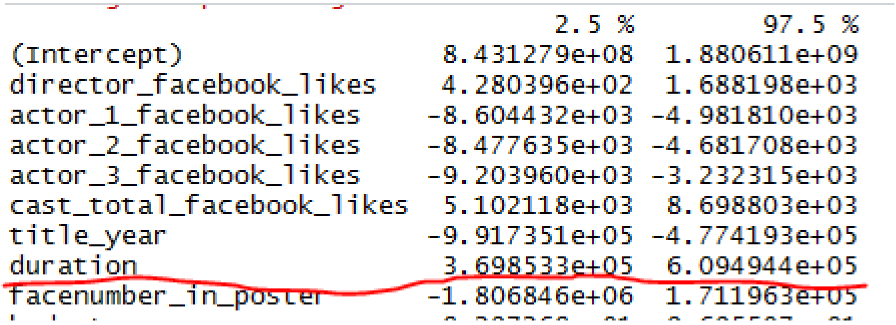
2) Remaining Variables that are converted to Dichotomous:
    a. aspect_ratio
    b. content_rating
    c. color

Prediction Model (Q4,5,6,7)

The problem stated in this scenario require us to come out with a model to identify which feature contribute to the gross earning, thus we are required to model a linear regression model to predict the gross and also to understand which attribute contributes the most.

And the model will be build and train with 80% of the total dataset, and the data are randomly selected to avoid biasness in the sequence and order of the data so to hide random data away from the model to achieve the best estimation and emulation of the real world where there are data that are not recorded in our dataset.

After the model was built, as we look at each variable at their P-value, and notice that Facebook likes for the actors and show ,Year of cast, budget, Drama-genres , aspect ratio of 2 and 2.24 have very high association with the the gross earning, and both budget and aspect ratio 2 topped the list in the p-values, which indicate the have the highest association with the gross earning.



```
                                 2.5 %        97.5 %
(Intercept)              8.431279e+08  1.880611e+09
director_facebook_likes  4.280396e+02  1.688198e+03
actor_1_facebook_likes  -8.604432e+03 -4.981810e+03
actor_2_facebook_likes  -8.477635e+03 -4.681708e+03
actor_3_facebook_likes  -9.203960e+03 -3.232315e+03
cast_total_facebook_likes  5.102118e+03  8.698803e+03
title_year              -9.917351e+05 -4.774193e+05
duration                 3.698533e+05  6.094944e+05
facenumber_in_poster    -1.806846e+06  1.711963e+05
```

Figure 7Confidence interver of duration for the built model

The 95% confidence interval of the variable “duration” lies between 3.7e+5 to 6.1e+5.

The Model is reasonably significant as the model provides a better fit than the no model, as the model was built with some variables that have high association with the gross earning.

Final Model (Q8)

There is multicollinearity exist in the model, and after removing those, the refined model is having VIF value for the rest of the variable between 1 to 1.6, which indicated that there still exists certain degree of collinearity, but it will not have any significant impact on the model as the VIF is still within the acceptable range, as VIF value >5 means high collinearity.