### COMP8042 Analytical & Scientific Programming

### Project 2

Student: Kevin O’Mahony

ID: R00105946

Date: 06/12/2016

### Introduction

This document describes the output of my Python application for exploring some features of the Kaggle, IMDB movie dataset as part of the second item of project work for the Application & Scientific Programming module.

The first issue to address is the percentage of null values in our data set. Below is the null value counts per attribute in our data set:

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Name** | **Null Count** | **Attribute Name** | **Null Count** |
| gross budget  content\_rating  aspect\_ratio  title\_year  director\_name  director\_facebook\_likes  plot\_keywords  num\_critic\_for\_reviews  color  num\_user\_for\_reviews  actor\_3\_name  actor\_3\_facebook\_likes  language | 464  312  125  121  89  84  84  62  14  11  10  8  8  7 | facenumber\_in\_poster  duration  country  actor\_2\_facebook\_likes  actor\_2\_name  movie\_facebook\_likes  imdb\_score  actor\_1\_facebook\_likes  genres  actor\_1\_name  movie\_title  num\_voted\_users  movie\_imdb\_link  cast\_total\_facebook\_likes | 7  7  6  5  4  1  1  0  0  0  0  0  0  0 |

The attributes we will be looking at that have non-zero null counts:

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Name** | **Null Count** | **Attribute Name** | **Null Count** |
| gross  budget | 464  312 | director\_name  title\_year | 84  89 |

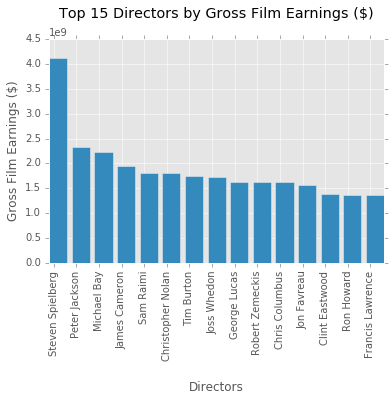
If we delete any row that has a null value for one of the above attribute’s, we reduce the original data set size from 4067 to 3427 rows. This represents a loss of 640 rows or 15.7% of the original data. In order to preserve as much of the data as possible we will delete null values on a per query basis i.e. if our query looks at the gross and budget attributes we will only delete those rows with null gross and budget entries. Below are the counts of rows removed for each of our queries.

|  |  |
| --- | --- |
| **Query** | **Rows deleted** |
| Most successful Director | 467 |
| Most successful Actor | 464 |
| Analysis of the distribution of gross earnings | 468 |
| Gross earnings versus IMDB movie scores | 465 |
| Genre Analysis | 1 |
| Gross Earnings / Budget ratio | 640 |

Loading the file into a data frame and removing null values each time we run a query is not very efficient. However, the IMDB movie data set is relatively small and for this application it was deemed practical.

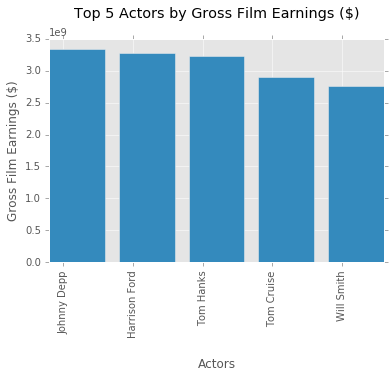
**1. Most Successful director or actors**

We are asked to produce a bar graph showing either the top directors or top actors based on the gross movie earnings, as selected by the user. The user should also be able to specify the number of directors/actors to display. Below is the chart of the top 15 directors:

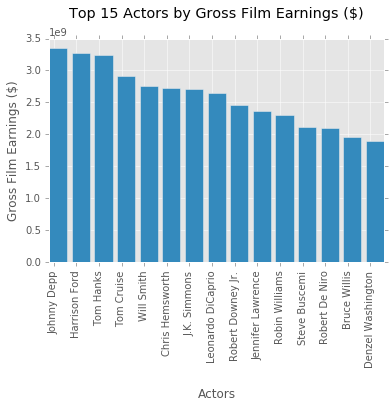


We can see Steven Spielberg leads the Directors list, with almost twice the gross earnings of any other director in the list.

Below is the top 5 actors chart:

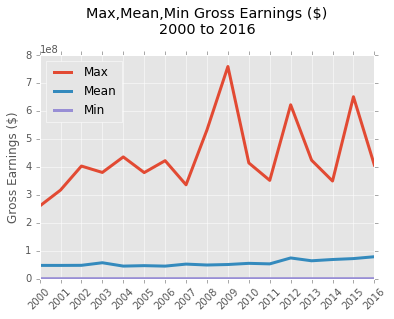


Like the top 15 director chart, the top 5 actor chart is made up of only males. Running the actors query for the top 15 brings in one female – Jennifer Lawrence at number 10.



**2. Analysis of the distribution of gross earnings**

Here we are looking at how the gross earnings is distributed over a user specified time period. A multi-line chart is used to display the maximum, mean and minimum gross earnings for each year in the range of years specified by the user.

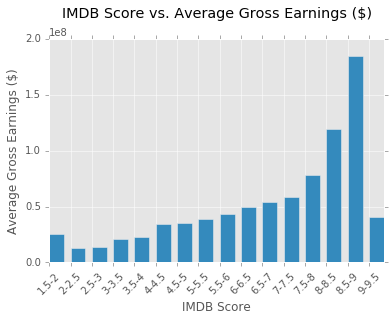


**3. A view of gross earnings versus IMDB movie scores**

This query looks at the gross earnings versus the IMDB scores. The IMDB scores in the data set are represented as a continuous variable from 0 – 10. We divide this range into categories: 0 – 0.5, 0.5 – 1, 1 – 1.5,…9.5 – 10, in order to display the relationship between gross earning and IMDB scores in bar chart format. As we can see below, in general the higher a film’s IMDB score, the higher its gross earnings. The two exceptions to this are:

* the 9 – 9.5 IMDB category is almost five times smaller than the 8.5 – 9 category’s average earnings
* the 2 – 2.5 IMDB category is nearly half the 1.5 – 2 category’s earnings figure

We also note that there was no valid data found for the IMDB ranges 0 – 1.5 and 9.5 – 10.



**4. Genre Analysis**

The 24 unique movie genres found in the data set are shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Action | Adventure | Animation | Biography |
| Comedy | Crime | Documentary | Drama |
| Family | Fantasy | Game-Show | History |
| Horror | Music | Musical | Mystery |
| News | Reality-TV | Romance | Sci-Fi |
| Short | Sport | Thriller | War |

The user selects one of the genres above and is returned the average IDMB score for that genre.

Below are some examples of the output:

There are 1 films in the News genre

Their average IMDB score is 7.4

...

There are 397 films in the Horror genre

Their average IMDB score is 5.9

...

There are 921 films in the Romance genre

Their average IMDB score is 6.4

...

There are 1548 films in the Comedy genre

Their average IMDB score is 6.2

...

There are 187 films in the History genre

Their average IMDB score is 7.1

...

There are 191 films in the War genre

Their average IMDB score is 7.0

It’s interesting to see we have only one News genre entry in the data set.

**5. Custom Query: Gross Earnings / Budget ratio**

For the custom query, I decided to look at the ratio of the gross earnings to the film budget as a way of assessing the financial success of a film. We don’t know if the budget figures are accurate, i.e. did the film come in under or over budget. Likewise, the exact makeup of the gross earnings figures are not known. Therefore, this ratio is a very crude indication of the film’s financial success, but gives us some indication for comparison purposes in this exercise. The query displays the three items: Gross/Budget ratio, Gross Earning and Budget for the 10 films with the highest ratio figure as shown over.

Multiple bar charts were used above as a convenient way to display the three very different y-axis ranges. Leading the top 10 list is ‘The Texas Chainsaw Massacre’ which had a budget of $83,532 and gross earnings of $30.85M. This film was obviously made on a tiny budget even for its release year of 1974 which is why made it to the top of the list.

