**Data Set Title**

**Exploratory Analysis**

Jacob Redmon, [jredmon@bellarmine.edu](mailto:jredmon@bellarmine.edu)

1. **INTRODUCTION**

This data set contains some basic data from movies in the Internet Movie Database. It contains information about the number of Facebook likes the director and starring actors have, which will provide some interesting analysis. The data set can be found at <https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset>.

1. **DATA SET DESCRIPTION**

This data set contains 5043 samples with 28 columns with various data types. A complete listing is shown in **Table 1**.

For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| movie\_title | /object | 0% |
| title\_year | /float64 | 2% |
| country | /object | 0.09% |
| language | /object | 0.2% |
| budget | /float64 | 9.7% |
| gross | /float64 | 17.5% |
| imdb\_score | /float64 | 0% |
| duration | /float64 | 0.3% |
| genres | /object | 0% |
| plot\_keywords | /object | 3% |
| color | /object | 0.3% |
| content\_rating | /object | 6% |
| aspect\_ratio | /float64 | 6.5% |
| facenumber\_in\_poster | /float64 | 2.5% |
| num\_critic\_for\_reviews | /float64 | 0.9% |
| num\_user\_for\_reviews | /float64 | 0.4% |
| num\_voted\_users | /int64 | 0% |
| director\_name | /object | 2% |
| actor\_1\_name | /object | 1.4% |
| actor\_2\_name | /object | 0.2% |
| actor\_3\_name | /object | 4.5% |
| director\_facebook\_likes | /float64 | 2% |
| actor\_1\_facebook\_likes | /float64 | 0.1% |
| actor\_2\_facebook\_likes | /float64 | 0.2% |
| actor\_3\_facebook\_likes | /float64 | 0.4% |
| cast\_total\_facebook\_likes | /int64 | 0% |
| movie\_facebook\_likes | /int64 | 0% |
| movie\_imdb\_link | /object | 0% |

1. **Data Set Summary Statistics**

Narrative introduction to the section.

**Table 2: Summary Statistics for XXX (name of dataset)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
|  |  |  |  |  |  |  |  |  |

There should be a table for **EACH** categorical variable.

Table 3: Proportions for XXX (n=yyy)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
|  |  |  |

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

Table 4: Correlation Table/Tables

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

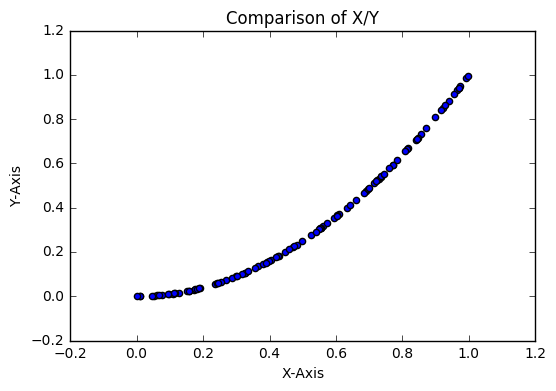
After the table with the raw data, include a heatmap of the correlation matrix as a figure.

1. **DATA SET GRAPHICAL EXPLORATION**

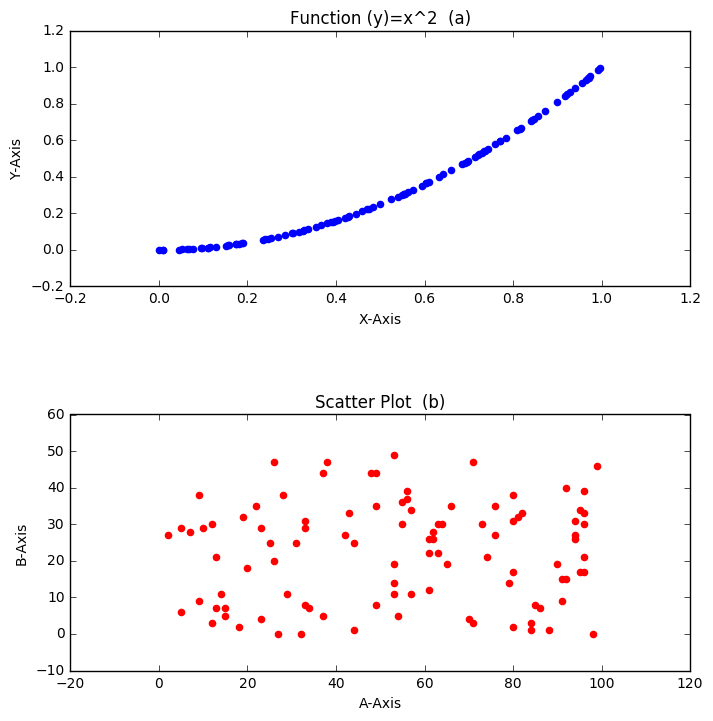
Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

* 1. *Distributions*
  2. *ScatterPlots / Pairwise Plots (continuous variables)*
  3. *Barcharts (categorical variables)*
  4. *Other Plots - don’t skimp – there are likely other plots that would be useful that I haven’t already specified. Include those in this section.*

All figures should be cited formatted like this and mentioned in the text.



**Figure 1: Comparison of X/Y from dataset (single plot) (8 pt)**



**Figure 2: (a) Function Output (b) A against B (multiple plots) (8 pt)**

1. **SUMMARY OF FINDINGS**

Finish up with a paragraph or two of summarizing your findings about this data set.