IST 3420 – Introduction to Data Science and Management, Fall 2017, Chen

**Project Evaluation Form – Milestone 2: Data Analysis I**

**Submission Due Oct 6, 11:59 PM**

**Instruction:**

1. Collect and cleanse data. The project report should include:

* Introduction (refined from M1)
* Data Source and Collection (refined from M1)
* Data manipulation (newly developed)
* Data summarization and visualization (newly developed)

Read the evaluation criteria carefully on the next page for the detail.

1. Use R Markdown to write your project report. You need to use proper Markdown syntax to format your report. Do not use MS Word or other format.
2. Please submit the following documents into Canvas:

* The project report written in R Markdown;
* The Word/HTML/PDF report that is directly generated from your R Markdown file;
* The Evaluation form with full project team information (see below table).

**Project Team Information (filled in by students)**

|  |  |  |
| --- | --- | --- |
| **Member name** | **Percent contribution** | **Activities completed by the member** |
| Adam Forestier | 25% | Coded almost all of this part of the project, Assisted with writing, attended group meeting, came up with analysis and correlation ideas |
| Brandon Jones | 25% | Attended group meeting, verified duplicates needed to be removed, reviewed group work |
| Nathaniel Williams | 25% | Did majority of writing interpretations, provided analysis concepts, assisted with coding, attended group meeting |
| Parika Gupta | 25% | Attended group meeting, found data that needed to be cleansed |

**Evaluation Summary – M2 (filled in by instructor)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **Target %** | **Comments** | **Evaluation** |
| * Convert project proposal into R Markdown format. * Extract and transform potential variables from the data source(s). Cleanse your dataset(s). | 20 | The columns need a brief explanation. Some variables such as “gross” and “color num\_critic\_for\_reviews  ” are hard for the audience to understand. | 19 |
| * Manipulate your dataset properly. * Properly detect outliers and deal with missing values in your dataset. | 30 | Using movie title to check duplicated records may be problematic. Two movies by different directors or in different years may have the same name.  No procedures to deal with missing values and outliers. ???  As shown in summary statistics, there are many missing values. | 20 |
| * Summarize and visualize data by using appropriate methods. * Professionally interpret your data summary and visualization. * Provide summary statistics, correlation table/plot, and at least 4 professional graphs with detailed and proper interpretations. | 30 | Need to have detailed and proper interpretations for summary statistics and correlation tables.  The interpretation of linear regression lines may be incorrect: the variables are skewed and the linear regression lines are highly influenced by the outliers. You may need to log transform some skewed variables. Some relationships may not be linear. Also these scatter plots with linear regression lines are still plots. They are NOT regression analyses and thus should be included in the visualization section.  You can plot the number of movie overall the years, as a companion to the “year and score” plot.  Is there a way to sort the content rating (x-axis items) in the “Content Scores” plot?  You can also explore whether other factors (such as country and language) have influence on movie scores. There are many ways to better understand your data and be prepared for regression analysis and predictive analytics in milestone 3. | 28 |
| * Format your project report in a professional way. * Write your project report by using appropriate Markdown syntax. | 20 |  | 20 |
| The report satisfies all of the following criteria:   * It tells a very interesting story; * The data manipulation methods are professionally applied; * The whole document is well written without many grammar or writing issues. | 10 bonus |  | 5 |
| **Total** | **100** |  | **92** |