Principle of Data Cleaning and Preparation
CSCI 521
Semester Project

v1.0

1 Project Description

This is a semester long project.

You will be required to clean and prepare a provided data set for data mining. The data set is part of the IMDB movie database.

The project will be broken into 4 phases. Each phase will build on the prior phase. As basic description of each phase is outlined here. More details in the remainder of the document.

- Phase 1: Select a team, and a combine the data. Due: Sep 13, 11:59 pm.
- Phase 2: Create a data dictionary of the data. Due: Oct 1, 11:59pm.
- Phase 3: Exploratory analysis of the data. Due: Nov 1, 11:59pm.
- Phase 4: Missing and invalid data. Due: Nov 22, 11:59pm

Feedback will be provided for each phase. You can select to receive feedback on a phase before its due date so you can move on to the next phase. **Once you select this option the grade becomes final**.

2 Phases

Below outlines the minimum submission requirements for each part of the project. Submissions are due by 11:59 pm on the day they are due. No late submissions will be accepted. No email submissions will be accepted.

2.1 Group selection

Self enroll in a group in myCourses. Groups may be no more than 4 people, but may be cross section. Students not enrolled in a group by the due date will automatically be assigned a group.

2.2 Phase 1

During this phase your will begin to combine and prepare your data. The following items are due:

- A report:
  - lists your group number from myCourses
  - includes your updated team name
  - the names of all team members
  - the steps you took to combine the data. I should be able to follow this process on the original data set and get the same results.

- Your combined and reduced data set; as a single TSV file.
- Any programs you wrote to combine the data. Zip this into a file called code.zip

The goal of this phase is to analyze the provide data set; they are tab separated file. Determine how the files can be combined into a single TSV file; tab separated file. A tab separated file is important due to some attributes having commas in their values.

You will then reduce this data set to only entries that have a region of the "US" in the title data set.

Information about these dataset can be found at Dataset. Using this site to gain insight into the data and their domains.

The dataset can be found at Datasets. We will only be working with a subset of these:

- **title.akas.tsv**
- **title.basics.tsv**
- **title.ratings.tsv**
- **title.name.basics.tsv**

Combining these sets together should give you a listing of actors with the movies they are know for, the information about that movie, and the ratings for that movie as a single line in the TSV file.

If an attribute, such as genre, has multiple values they must be split into multiple lines in the final TSV file.

Submit a Zip file containing the report outlined above (PDF), your combined data set, and any code used to assist you in this process. Submission of anything other than a Zip and PDFs will earn a 0 for this phase.

Zip all the required files into a zip file called **Phase1.zip**.

MyCourses will only accept one file and only save the last submission.

2.3 Phase 2

During this phase you will begin to analyze your dataset. The goal is to create a data dictionary. The following items are due:

- An updated report:
  - includes your team name
  - the names of all team members
  - all information from the prior phase
  - a data dictionary outlining every attribute in the TSV file you created in phase 1.
    - Every attribute must include the number of unique entries and total number of entries.
    - Every attribute must include a data type and weather it is a scalar, nominal, or ordinal value.
    - Numeric data must include the range of values.

A table for this is acceptable.
  - Note any inconsistencies in the data; such as lots of missing values, non-numeric mixed with numeric.
  - Explanation of the process you took to complete the tasks of this phase. Be precise and concise!

Submit a Zip file (**Phase2.zip**) containing the report outlined above (PDF), your data set, and any code used to assist you in this process. Submission of anything other than a Zip and PDFs will earn a 0 for this phase.

MyCourses will only accept one file and only save the last submission.

2.4 Phase 3

During this phase, you will preform exploratory analysis on you data set. The following items are due:

- An updated report:
  - includes your team name
  - the names of all team members
  - all information from the prior phase
  - an update data dictionary including for all numeric values:
    - Mean, median, mode, min, max
    - outliers
  - Explanation of the process you took to complete the tasks of this phase. Be precise and concise!
- Box plots, bar charts, pie charts, etc for the numeric data. Choose the best chart type for the data and explain why you chose a particular chart in your report.
- Histograms for non-numeric data that has fewer than 15 unique values.
- A few interesting bivariate charts. You do not need to provide me with all combinations of attributes; just a few.

Submit a Zip file containing the report outlined above (PDF), any code used to assist you in this process, your dataset, and the various charts (PDF), named **Phase3.zip** Submission of anything other than a Zip will earn a 0 for this phase.

MyCourses will only accept one file and only save the last submission.

2.5 Phase 4

Then final phase of the project. In this phase you will handle missing and invalid values in the data. You will also normalize numeric data. The following items are due:

- An updated report:
  - includes your team name
  - the names of all team members
  - all information from the prior phase
  - for each attribute explain how you handled any missing or invalid values. Include reasoning of why you handled them in this way.

  - for each numeric attribute explain how you normalized the data. Include reasoning of why you handled them in this way.
  - Explanation of the process you took to complete the tasks of this phase. Be precise and concise! I should be able to follow the same steps and get the same results.
- Your updated data set with missing values handled and normalization completed.

Submit a Zip file containing the report outlined above (PDF), any code used to assist you in this process, your data set, and the various charts (PDF), named **Phase4.zip** Submission of anything other than a Zip will earn a 0 for this phase.

MyCourses will only accept one file and only save the last submission.

Peer Evaluation

You must submit a peer evaluation for yourself and your team members. Failure to submit a peer evaluation will result in a 20% deduction on your project grade. A template is available on myCourses.

3 Project Constraints

This section outlines details about any project constraints or limitations.

Constraints/Limitations:

- You must use the provided dataset in its entirety.
- Your handling of missing values cannot be just to delete the row or the attribute. This is a valid option, but not for every attribute or row.
- Report should be easily readable and not just a bulleted list.

4 Grading

Your implementation will be grading according to the following:

- Phase 1: 25%
- Phase 2: 25%
- Phase 3: 25%
- Phase 4: 25%

5 Submission

Follow the submission instructions for each section of the project. Emailed and late submissions will not be accepted.