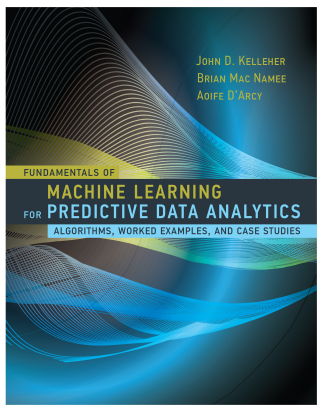
Data Quality Report Assignment

Prof. John D. Kelleher

Dublin Institute of Technology
e: john.d.kelleher@dit.ie



Assignment Description

For this assignment you need to write Python code to generate:

1. the table in the Data Quality Report containing the relevant descriptive statistics for the continuous features in the data set
2. the table in the Data Quality Report containing the relevant descriptive statistics for the categorical features in the data set
3. a histogram for each of the continuous feature with high cardinality (≥ 10)
4. a bar plot for each continuous feature with a low cardinality (< 10)
5. a bar plot for each categorical feature

Submission Deadline and Process

- Submission Deadline: 10 a.m. on Wed. 25th May 2018
- Submission Process: Email a zip file containing the required files (described later)
  1. to john.d.kelleher@dit.ie
  2. using the following subject header:

     [ENSSAT] Data Quality Assignment Team *team-name*
- Where *team-name* has been replaced by the name of your team, e.g., A, B, C, etc. I only need one email submission from each team (i.e., only one person from the team needs to submit on behalf of the team)

What files your code should take as input.

- Your code should expect the following files as **input** (note the paths to the files):
  1. Your program should expect that the dataset is in a comma separated file called 'DataSet.csv' that is stored in a directory called 'data' that is a subdirectory of the directory your program is run from (in other words the path to the dataset file should be './data/DataSet.csv')
  2. Your program should expect that the first line of the data set file contains the names of the features in the dataset.

What files your code should produce.

- Your program should output the table for the continuous features to a comma separated file using the following path

  ’./data/*team-name*-DQR-ContinuousFeatures.csv’

  and for the categorical features to a commas separated file with the following path

  ’./data/*team-name*-DQR-CategoricalFeatures.csv’

  , where *team-name* is replaced in these files names with your team-name

Example

Assuming you are team A the files should be called:

- ’./data/A-DQR-CategorialFeatures.csv’
- ’./data/A-DQR-ContinuousFeatures.csv’

and these comma separated files containing the tables should be written to the same directory that you get the input files from.

What files your code should produce.

- The format of the *team-name*-DQR-ContinuousFeatures.csv and *team-name*-DQR-CategoricalFeatures.csv files should mirror continuous and categorical feature tables in the data quality report as presented in the notes (see next slide for more info on file formats).

What files your code should produce.

- The bar charts and histograms plots can be generated using the plotly library
- The bar charts and histograms plots can be output as html files.

File Format

- The first line in each file should be a header line should be a comma separated listing the descriptive feature name for each column in the file (use the string FEATURENAME for the name of the first column).
- Each of the subsequent lines in the file should be a comma separated list with the name of the feature as the first element in the list and and then the descriptive statistics in the subsequent commas separated elements in the list in the same order as they are listed in the notes.

What you should submit

1. The Python source code you wrote for the assignment. This source code should be in a file called *team-name*-GenerateDQR.py where *team-name* has been replaced with your team-name. Also, include the names of all team members at the top of the file as a comment.
2. The Data Quality Report table for the continuous descriptive features—as identified by your code—in the dataset. This table should be in a comma separated file.
3. The Data Quality Report table for the categorical descriptive features—as identified by your code—in the dataset. This table should be in a comma separated file.
4. The html files for the bar charts and histograms for each of the features in the dataset.
5. A brief (1 page) description of the dataset that describes your analysis of the dataset in terms of missing values, outliers, feature cardinality, and your opinions as to what should be done to address these quality issues.

INSSAT
LANNION

Some useful Python modules

- The Pandas read_csv, to_csv
- Pandas dataframe operations select_dtypes, transpose, rename, insert, describe, value_counts
- The Collections module and operations, e.g. OrderedDict, Counter
- The numpy module, e.g.percentile(),mean(), median(), std()
- The str module, e.g. isnumeric
- The Plotly library is a good library to use for generating the images, for example you can use it to generate bar chart and histograms and store these figures as html files