CSE 5243

Instructor: Jason Van Hulse

Homework 1: Exploratory Data Analysis

Due Date: 1/21/2019 at 11:59 pm (submitted through Carmen) *[there will be a 15% penalty for all homework handed in up to 24 hours late; homework will not be accepted more than 24 hours late]*

Objective: In this lab, you will analyze the "Adult" dataset available at the UCI Machine Learning repository**.** In particular, you will concentrate on the first three phases of the CRISP-DM process model - *Business Understanding, Data Understanding,* and *Data Preparation.*

The objective of this assignment is two-fold. First, by analyzing and thinking critically about the data, you should identify interesting trends and patterns. Second, the final dataset that you create will be used to build classification models in a later homework assignment.

Approach:

1. **Business Understanding Phase**: Write a paragraph providing an overview of the data. Some questions you should consider are: Where did the data come from? What do the rows represent? Why and how was the data collected? What types of questions might you be able to analyze with this data?

   You should review the dataset description information on the webpage to get some context. Of course you will only have limited background on this topic (and I don't expect you to become an expert in the census), so do your best to imagine the context for the work, making reasonable assumptions as appropriate. At this stage, you are not analyzing individual attributes, but discussing the dataset in aggregate.

2. **Data Understanding Phase**: Perform exploratory data analysis of the dataset by looking at individual attributes and/or combinations of attributes. You should focus on identifying and describing interesting observations and insights that you might uncover from the data.
   - **A.** Describe the meaning and type of data for each attribute
   - **B.** Provide basic statistics for the attributes - e.g., counts, percentiles, mean, median, standard deviation. The statistics should be relevant for the type of attribute.
   - **C.** Visualize the most important or interesting attributes using appropriate techniques. For each visualization, provide an interpretation explaining why it is appropriate or interesting. What does each visualization tell us?

D. Verify data quality: explain any missing values, duplicate data, or outliers. What, if anything, do you need to do about these? Be specific.
E. Explore the relationships among the attributes, excluding the class attribute. Use scatter plots, correlation matrices, cross-tabulations, group-wise averages, or other appropriate techniques. Explain and interpret any interesting relationships.
F. Identify and explain any interesting relationships between the class attribute and the other attributes. You may refer to earlier visualizations or create new ones.

You should not simple provide the basic EDA information for all attributes in the data. Instead, you should focus on those that are more interesting or important, and provide some discussion of what you observe. Pay particular attention to potentially interesting bivariate (two-variable) relationships, as well as the relationship between each attribute and the class.

3. **Data Processing Phase:** Based on the insights gleaned in the *data understanding phase,* determine what type of processing that you would like to do to create a final dataset to be used for future modeling.
   A. What attributes do you decide to keep or remove? Please justify.
   B. Did you decide to implement any attribute transformations? If so, why?
   C. Did you decide to create any new features? If so, why?
   D. Implement any data cleaning steps previously identified, and show the effects of that cleaning through the use of appropriate statistics and/or visualizations.

One of the final outputs of your program should be the creation of a dataset (can be in the format of a data frame) which has all of the attributes you would like to use for the *modeling* phase of a project, as well as dealing with any outliers, noise or missing values.

Collaboration: For this assignment, you should work as an individual. You may informally discuss ideas with classmates, but your work should be your own.

What you need to turn in:

1) **Code** - please submit to Carmen any code that you used to process and analyze this data. You do not need to include the input data.
2) Written Report
   A. The report should be well-written. Please proof-read and remove spelling and grammar errors and typos.
   B. The report should discuss your analysis and observations. Present charts and graphs to support your observations. If you performed any data processing, cleaning, etc, please discuss it within the report.
   C. The written report can be in the form of a Python or R Notebook or as a Word or PDF Document.

Grading Criteria:

1. **Overall readability and organization of your report (30%)** - is it well organized and does the presentation flow in a logical manner; are there many grammar and spelling mistakes; do the charts/graphs relate to the text, etc…
2. **Business Understanding Phase (10%)** - Did you provide a reasonable level of overview information?
3. **Data Understanding Phase (40%)** - Did you find novel and/or interesting insights, or did you solely focus on simple summarizations of the data? Did you draw and present potential conclusion or observations from your analysis of the data? Did the statistics and visualizations used make sense in the context of the data?
4. **Data Processing Phase (20%)** - Did your program produce the desired output dataset, implementing the preprocessing steps that you outlined in the data processing phase? Have you justified why you eliminated certain features or examples, or included new features?

How to turn in your work on Carmen:

Please do this work in either Python or R. All the related files (code and/or report) except for the data will be tarred in a *.zip file or *.tgz file, and submitted via Carmen. Use this naming convention: “Project1_Surnames_DotNumber.zip" or “Project1_Surnames_DotNumber.tgz.” The submitted file should be less than 5MB.