Data Mining: Assignment 1

Census-Income

Tasks: Data Cleaning, Exploration, and Visualization

Assigned: August 14th,2018

Due: September 4th,2018

Points: 60 points

This data set contains census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables.

The data and data description can be obtained in Moodle.

The dataset contains 199,523 objects and 41 attributes.

Questions

Business understanding

1. What is the purpose of the census income survey? [5 points]

2. What are the uses of the census income survey? [5 points]

Data understanding

3. Identify type for each attribute (nominal, ordinal, interval, ratio) in the data file. [5 points]

4. Verify data quality: [30 Points]

    - Implement data cleaning approaches to census-income data in R.

    - Explain what you do to clean the data.

    - Provide the R codes (R markdown).

5. Give simple appropriate statistics (range, mode, mean, median, variance, counts, etc.) for 10 most important attributes and describe what they mean or if you found something interesting. [5 points]

6. Visualize 10 most important attributes appropriately (histogram, bar chart, etc.). Provide an interpretation for each chart. [5 points]

7. Explore relationships between attributes for 5 relationships. Look at the attributes and then scatter plots, correlation, etc. as appropriate. Explain the results. [5 points]