5/24/2019

Assignment 1

CSCI 6612 – Visual Analytics

Gurjot Singh
B00811724

1. Dataset

The dataset has issues in three major columns:

1) Age (Column 1) : The column has multiple negative values and many rows have age 0.
2) Work Class (Column 2) : The column has problems in multiple values. Some of the problems are as follows:
    - State-gov written as “State gov” and “stategov”.
    - Local-gov written as “local gov” and “Local gov”.
    - Self-emp-inc written as “self emp-inc” and “Self emp-inc”.
3) Occupation (Column 7) : The column also has problems in multiple values. Some of the problems are as follows:
    - Exec-managerial written as “exec-mnagerial” and “exec-managerial”.
    - Tech-support written as “Techsupport” and “tech-support”.
    - Craft-repair written as “Craft repair” and “Craft-reair”.
    - Adm-clerical written as “Adm clerical” and “Adm-cerical”.

2. Approach

The data was loaded in Jupyter Notebook using Pandas Library which converts the data into DataFrames. Matplotlib has been used to generate the histograms of the processed results. The approach used in the process is as follows:

1) Age (Column 1): The negative ages in the column were converted to positive using absolute function. Since the dataset contains Occupational information for the people, thus, the minimum age has been kept as 17 for all the data, so the age 0 has been converted to age 17 using Python.
2) Work Class (Column 2): The incorrect values in the column were converted to the correct names using string matching of the rows. An if-else function with string matching has been used in the processing of this column.
3) Occupation (Column 7): The incorrect values are processed using the if-else column as string matching to correct them to their actual occupational category.

The commons characters of all the categories have been fetched and used for matching the values with the actual category. The resultant columns are replaced with the original column and stored in a file “dataset1_processed.csv”. The results are then displayed for all rows.

3. Histogram

The histogram for the Age column(numeric) is as follows:

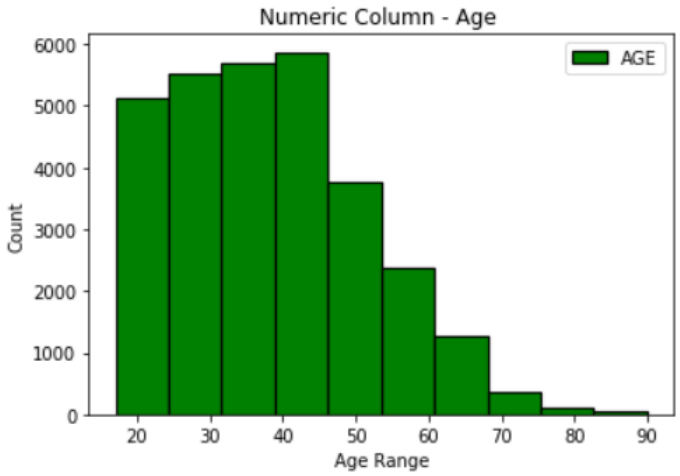

Figure 1 : Numeric Column - Age

The histogram of the Occupation column (Categorial) is as follows:

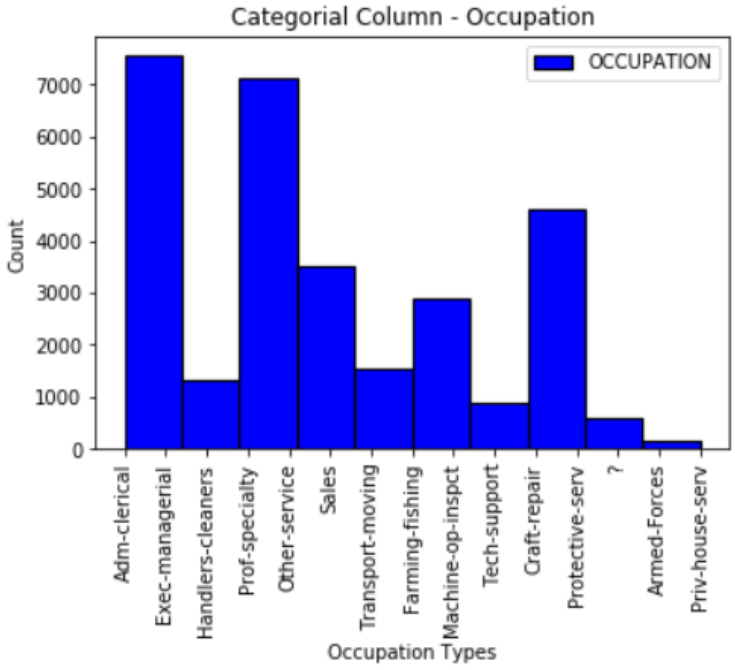

Figure 2: Categorial Column - Occupation