Visual Analytics: CSCI 6612

CREATED BY: RAVI TULSI ZALA

B00805073

1. Data Analysis

After analyzing the dataset, some of the columns have missing data and invalid data, as explained below.

- Age column contains negative values and zeros.
- Workclass column has spelling mistakes in values like “self empnot-nc” missing values with “?” and attached words like “federalgov”.
- Columns fnlwgt, education, education-num, and marital-status do not contain invalid or missing data.
- Whereas, occupation column contains a lot of incorrect values such as “aes” and ”ad-clrical”, missing values with “?” and attached words like “Transportmoving”.
- Rest of the columns do not have any invalid data, missing values or incorrect data.

2. Data Preprocessing

For correcting negative values, I used the abs() function of Pandas. In the Age column, missing value is represented with zero. Reasons for missing data in the Age column can be a person wanted it to be a secret or data entry mistakes. Test statistics methods like Mean and Median values are useful for filling the missing values. I used Mean value of the column to fill the missing value. In order to use pandas fillna method, I replaced zeros with NaN using numpy package. Using pandas’ fillna method to fill the NaN values with the mean of the column.

I used a regular expression to remove the special characters from the list of the unique workclass or occupation column and values in the column. It proved to help tackle the issue of attached words. For correcting the typos in a given dataset, I used the SequenceMatcher class of the difflib helper. The basic idea behind using this algorithm is to find the longest sequence matches considering no junk values like spaces and special characters. Given the dataset, I set the minimum ratio to 0.87, which a best possible match between two strings. This algorithm proved to be useful in tackling the problem of spelling mistakes. Pandas package is useful in removing the missing values represented using ?. Process for the occupation follows the same algorithm for preprocessing the data.

Below are some libraries I used for data preprocessing.

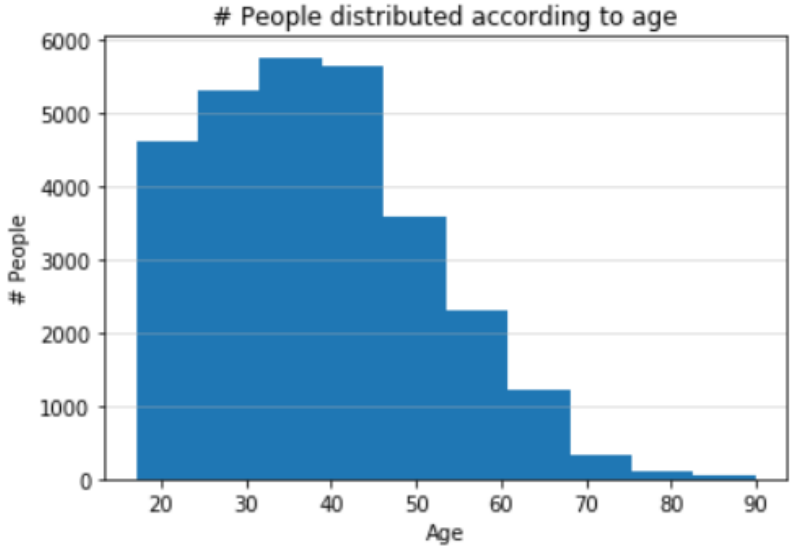
- Pandas: Pandas package provides the most powerful feature as Dataframe. It proved to be useful for dropping the missing values using dropna method, filling the values with fillna method, iterating over columns and rows, and replacing values with NaN, reading and generating CSV files. It makes Data manipulation and analysis easy [1].
- Numpy: I used np.nan while replacing the values of zeros and ? in a data with NaN. The dominant feature of the Numpy is high-performance multidimensional arrays and data manipulation using array.RE: Regular expressions are useful for finding a sequence of characters, special characters, patterns, and substrings. I used a regular expression to remove the special characters from data.
- Matplotlilb.pyplot: It is useful for visualizing data in terms of Histogram, Bar chart, Pie charts, etcetera.

- Difflib: Difflib helpers are useful in computing deltas. Classes and functions of the helpers facilitate the developers in comparing sequences. I used the SequenceMatcher class of Difflib Helpers to correct spelling mistakes according to the minimum ratio.
- Pandas.api.types : I used the "is_numeric_dtype" function of Pandas.api.types class to generalize the printing of the output [1].

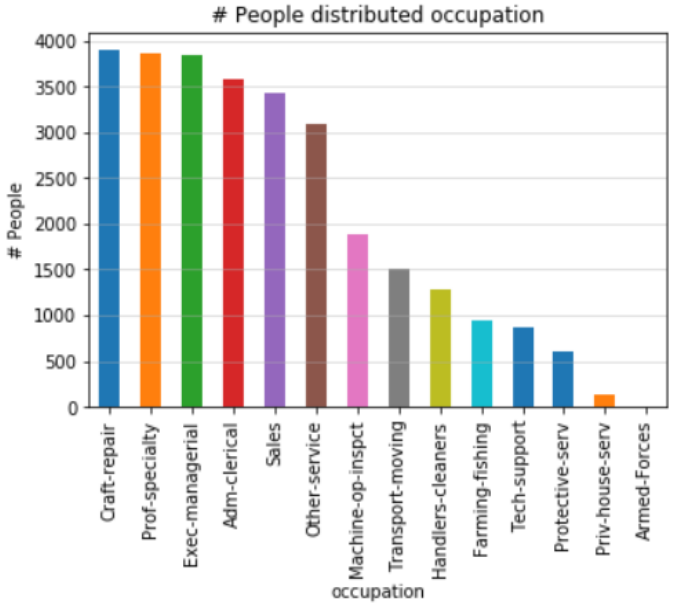
3. Plotting Data

For visualizing the data distribution, I used matplotlib.pyplot with pandas dataframe. Pandas provides xlabel(), ylabel(), and hist() functions.

- **Age Column (Numerical):**



- **Occupation Column (Categorical)**



References

[1] "Getting started — pandas 0.24.2 documentation", Pandas.pydata.org, 2019. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/getting_started/index.html. [Accessed: 25- May- 2019].

[2] "pandas.Series.astype — pandas 0.7.0 documentation", *Pandas.pydata.org*, 2019. [Online]. Available: http://pandas.pydata.org/pandas-docs/version/0.7.0/generated/pandas.Series.astype.html. [Accessed: 25- May- 2019].

[3] J. C.G., "How can I replace all the NaN values with Zero's in a column of a pandas dataframe", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/13295735/how-can-i-replace-all-the-nan-values-with-zeros-in-a-column-of-a-pandas-datafra. [Accessed: 25- May- 2019].

[4] P. S, "How to determine whether a column/variable is numeric or not in Pandas/NumPy?", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/19900202/how-to-determine-whether-a-column-variable-is-numeric-or-not-in-pandas-numpy. [Accessed: 25- May- 2019].

[5] J. Vegt, "Similarity between two words", *Data Science Stack Exchange*, 2019. [Online]. Available: https://datascience.stackexchange.com/questions/12575/similarity-between-two-words. [Accessed: 25- May- 2019].

[6] S. Boston, "how to Increase the figure size of Dataframe.hist for pandas 0.11.0", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/43392588/how-to-increase-the-figure-size-of-dataframe-hist-for-pandas-0-11-0. [Accessed: 25- May- 2019].