CSCI 6612

Assignment 1

VISUAL ANALYTICS

MEGHNA RAMACHANDRA HOLLA | B00812604

A1. What problems do you see in the data? In which columns?

- The dataset has no empty values or NaN values, but columns like ‘workclass’, ‘occupation’ and ‘salary’ have ‘?’ character that cannot be identified with a straight-forward ‘.isnan’ or ‘.isnull’ functions.
- There are typographical errors in ‘workclass’ and ‘occupation’ columns
- Column ‘age’ has negative and zero values. This is a problem as age cannot be 0 or negative.
- Column ‘education-num’ have negative values. Assuming that these are identifiers of column ‘education’, they cannot have negative values.
- Column ‘capital-gain’ has value 0. This does not give any useful information.
- Column ‘fnlwgt’ has numeric values, but the column name is not understandable to make interpretations.

A2. For every column that you had to work on, explain how you fixed the data and justify your decision. If you used any libraries, briefly describe how you used them.

1. **Age**
    a. Replaced all ‘?’ values with 0 in order make the corrupt data consistent
    b. Implemented Forward Filling of Age with values equal to 0. This was done to maintain the distribution of data. By filling all the 0s with either mode or mean would alter the distribution. Therefore, chose forward filling method.
    c. I observed that the range of Ages (-17 to -82) with negative values was similar to the range of positive values (17 to 90). Therefore, assuming that the minus sign (-) was a typographical error, I replaced negative values with their positive values.
    d. Libraries used: Abs() function for converting negative values into positive values. Pandas library to replace 0s with ‘ffill’
2. **Workclass**
    a. Rectified typographical errors by creating a robust dictionary of regex patterns. Iterated through this dictionary and replaced matching words with correct values.
    b. Replaced all ‘?’ values with values obtained by Forward Filling method. The values did not have any noticeable pattern with respect to other columns, therefore used ‘ffill’ method to maintain the distribution.
3. **FnlWgt**
    a. Although the purpose of ‘fnlwgt’ column was not understood, it was not dropped because it might have some useful information that could be used in the future.
    b. I checked if these values are unique to understand if they are identifiers of some kind, but they aren’t unique.
    c. Libraries used: Pandas library to check if there are unique values using ‘is_unique’

4. Education-Num

a. No noticeable correlation was observed with respect to other columns except with ‘education’ column.

b. Assuming that ‘education-num’ is the identifier of ‘education’ column, I replaced all the values (along with negative) to the mode of education’s identifier. For example, if ‘education’ value is ‘HS-grad’, I replaced all ‘education-num’ with the most frequent values of ‘education-num’ for ‘HS-grad’ ie.,9

c. Libraries used: Pandas library to access a set of rows using ‘loc’ and to find the most frequent value in a column using ‘mode’

5. Occupation

a. Rectified typographical errors by creating a robust dictionary of regex patterns. Iterated through this dictionary and replaced matching words with correct values.

b. Replaced all ‘?’ values with values obtained by Forward Filling method. The values did not have any noticeable pattern with respect to other columns, therefore used ‘ffill’ method to maintain the distribution.

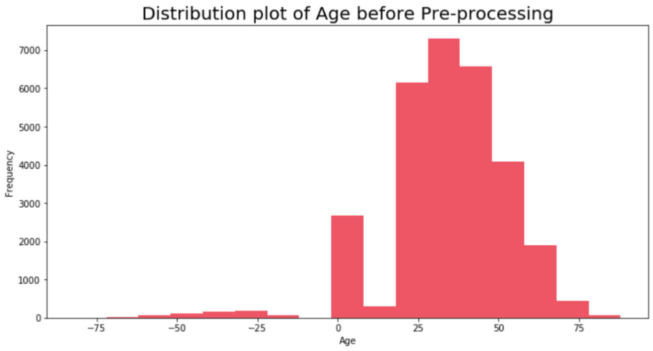
6. Salary

a. Replaced all ‘?’ values with values obtained by Forward Filling method. The values did not have any noticeable pattern with respect to other columns, therefore used ‘ffill’ method to maintain the distribution.

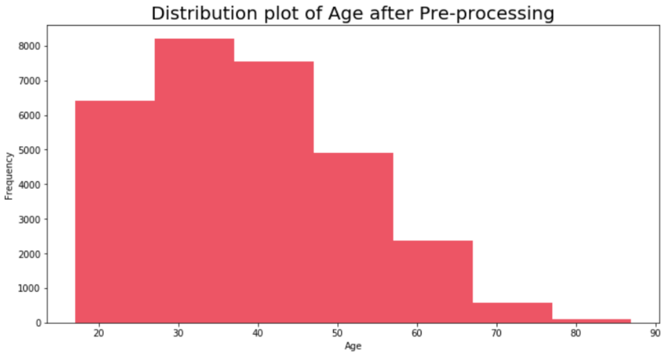
Other Libraries used:

- Numpy library to check if the data has NaN or null values, and to set a range of values for histogram with numeric type.
- Matplotlib to display categorical and numeric histograms and to set labels.

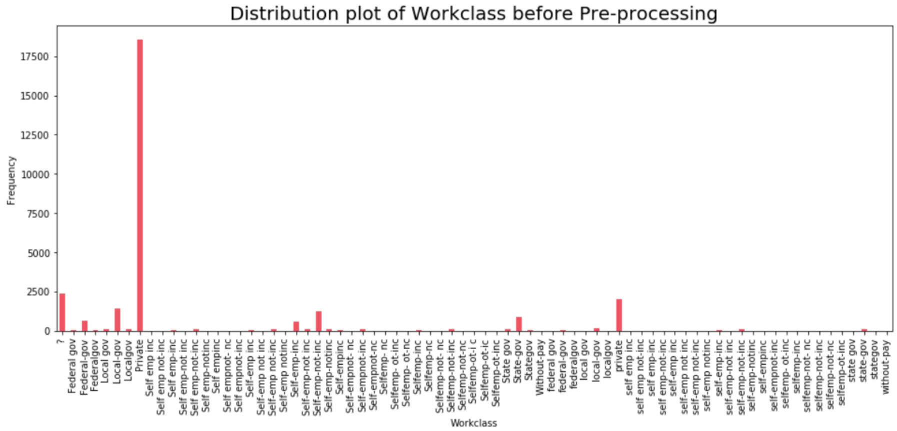
A3. Show the histograms for the columns that you fixed. (before and after)

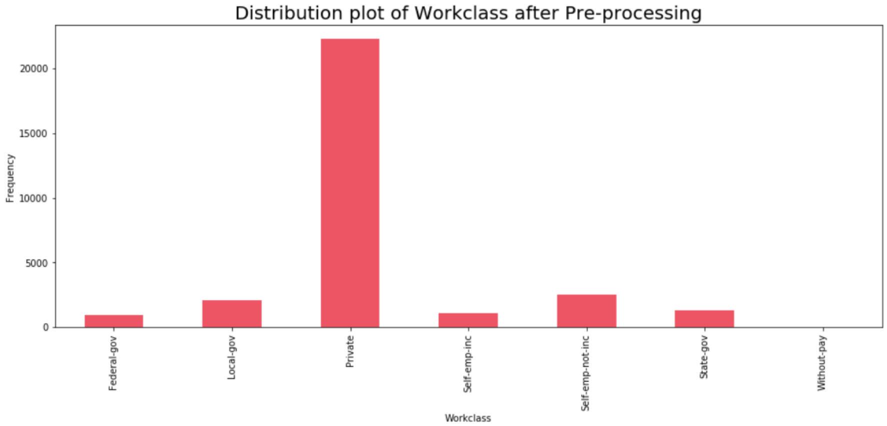


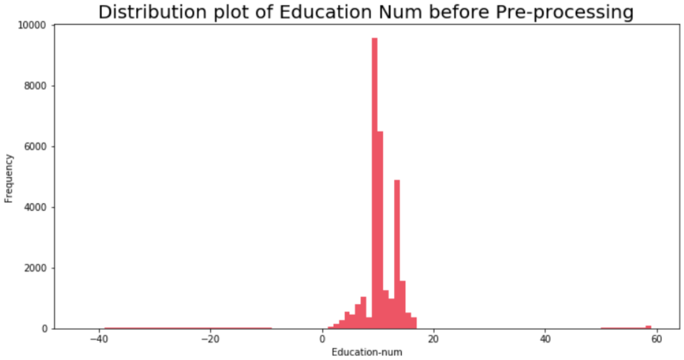
Distribution plot of Age before Pre-processing

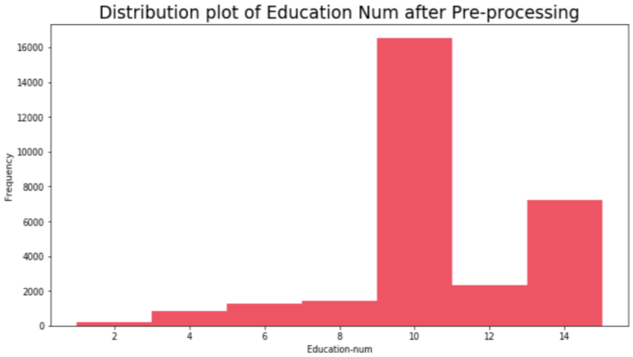


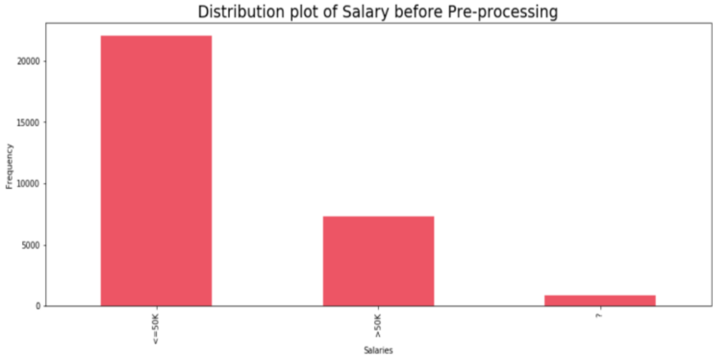
Distribution plot of Age after Pre-processing

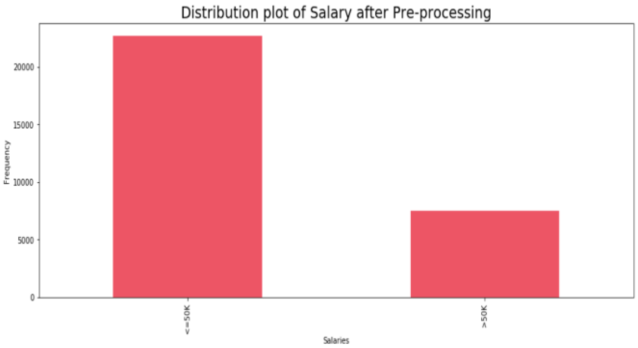

Distribution plot of Workclass before Pre-processing
Frequency
17500
15000
12500
10000
7500
5000
2500
0
?
Federal gov
Federal-gov
Federalgov
Local gov
Local-gov
Localgov
Private
Self emp inc
Self emp not-inc
Self emp-inc
Self emp-not inc
Self emp-not-inc
Self emp-notinc
Self empinc
Self empnot- nc
Self empnot-inc
Self-emp inc
Self-emp not inc
Self-emp not-inc
Self-emp notinc
Self-emp-inc
Self-emp-not inc
Self-emp-not-inc
Self-emp-notinc
Self-empinc
Self-empnot- nc
Self-empnot-inc
Self-empnot-nc
Selfemp- nc
Selfemp- ot-inc
Selfemp- ot-nc
Selfemp-inc
Selfemp-nc
Selfemp-not- nc
Selfemp-not-inc
Selfemp-not-nc
Selfemp-ot-i c
Selfemp-ot-ic
Selfemp-ot-inc
State gov
State-gov
Stategov
Without-pay
federal gov
federal-gov
federalgov
local gov
local-gov
localgov
private
self emp inc
self emp not-inc
self emp-inc
self emp-not-inc
self-emp inc
self-emp not inc
self-emp not-inc
self-emp notinc
self-emp-inc
self-emp-not inc
self-emp-not-inc
self-emp-notinc
self-empinc
self-empnot-inc
selfemp- ot-inc
selfemp-inc
selfemp-not- nc
selfemp-not-inc
selfemp-not-nc
selfemp-ot-inc
state gov
state-gov
stategov
without-pay
Workclass

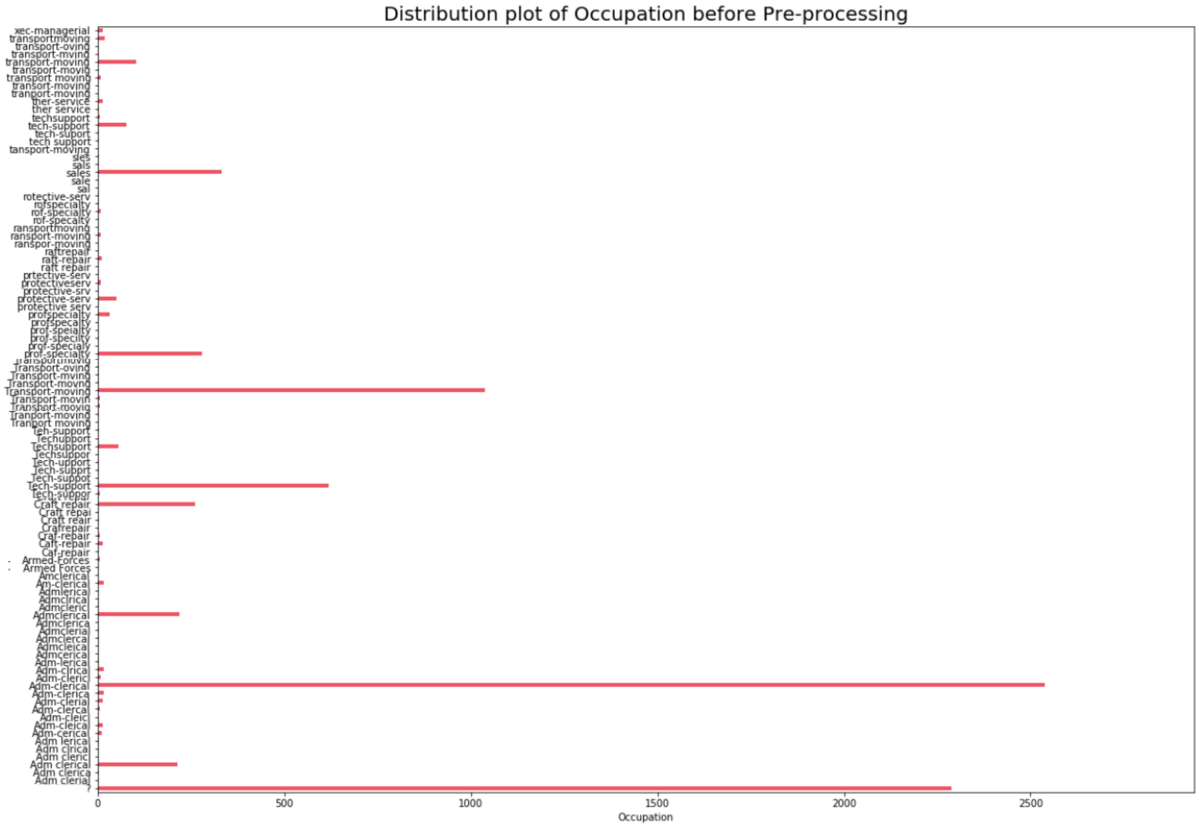

Distribution plot of Workclass after Pre-processing
Frequency
20000
15000
10000
5000
0
Federal-gov
Local-gov
Private
Self-emp-inc
Self-emp-not-inc
State-gov
Without-pay
Workclass

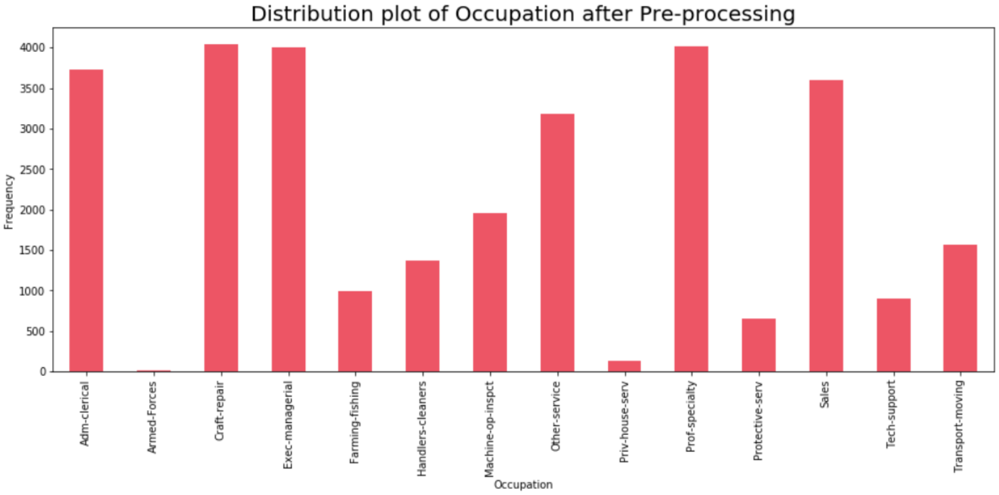

Distribution plot of Education Num before Pre-processing
Frequency
10000
8000
6000
4000
2000
0
−40
−20
0
20
40
60
Education-num


Distribution plot of Education Num after Pre-processing
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
2
4
6
8
10
12
14
Education-num


Distribution plot of Salary before Pre-processing
Frequency
20000
15000
10000
5000
0
<=50K
>50K
?
Salaries


Distribution plot of Salary after Pre-processing
Frequency
20000
15000
10000
5000
0
<=50K
>50K
Salaries


Distribution plot of Occupation before Pre-processing
xec-managerial
transportmoving
transport-oving
transport-mving
transport-moving
transport-movig
transport moving
transort-moving
tranport-moving
ther-service
ther service
techsupport
tech-support
tech-suport
tech support
tansport-moving
sles
sals
sales
sale
sal
rotective-serv
rofspecialty
rof-specialty
rof-specalty
ransportmoving
ransport-moving
ranspor-moving
raftrepair
raft-repair
raft repair
prtective-serv
protectiveserv
protective-srv
protective-serv
protective serv
profspecialty
profspecalty
prof-speialty
prof-specilty
prof-specialy
prof-specialty
Transport-oving
Transport-mving
Transport-movng
Transport-moving
Transport-movin
Tranport-moving
Tranport moving
Teh-support
Techupport
Techsupport
Techsuppor
Tech-upport
Tech-supprt
Tech-suppot
Tech-support
Tech-suppor
Craft repair
Craft repai
Craft reair
Crafrepair
Craf-repair
Caft-repair
Caf-repair
Armed-Forces
Armed Forces
Amclerical
Am-clerical
Admlerical
Admclrical
Admcleric
Admclerical
Admclerica
Admcleria
Admclercal
Admcleical
Admcerical
Adm-lerical
Adm-clrical
Adm-cleric
Adm-clerical
Adm-clerica
Adm-cleria
Adm-clercal
Adm-cleic
Adm-cleical
Adm-cerical
Adm lerical
Adm clrical
Adm cleric
Adm clerical
Adm clerica
Adm clerial
?
0
500
1000
1500
2000
2500
Occupation


Distribution plot of Occupation after Pre-processing
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
Adm-clerical
Armed-Forces
Craft-repair
Exec-managerial
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Other-service
Priv-house-serv
Prof-specialty
Protective-serv
Sales
Tech-support
Transport-moving
Occupation

References

[1] "User's Guide¶." User's Guide - Matplotlib 3.1.1 Documentation, https://matplotlib.org/users/index.html.

[2] NumPy¶. (n.d.). Retrieved from https://numpy.org/

[3] powerful Python data analysis toolkit¶. (n.d.). Retrieved from https://pandas.pydata.org/pandas-docs/stable/