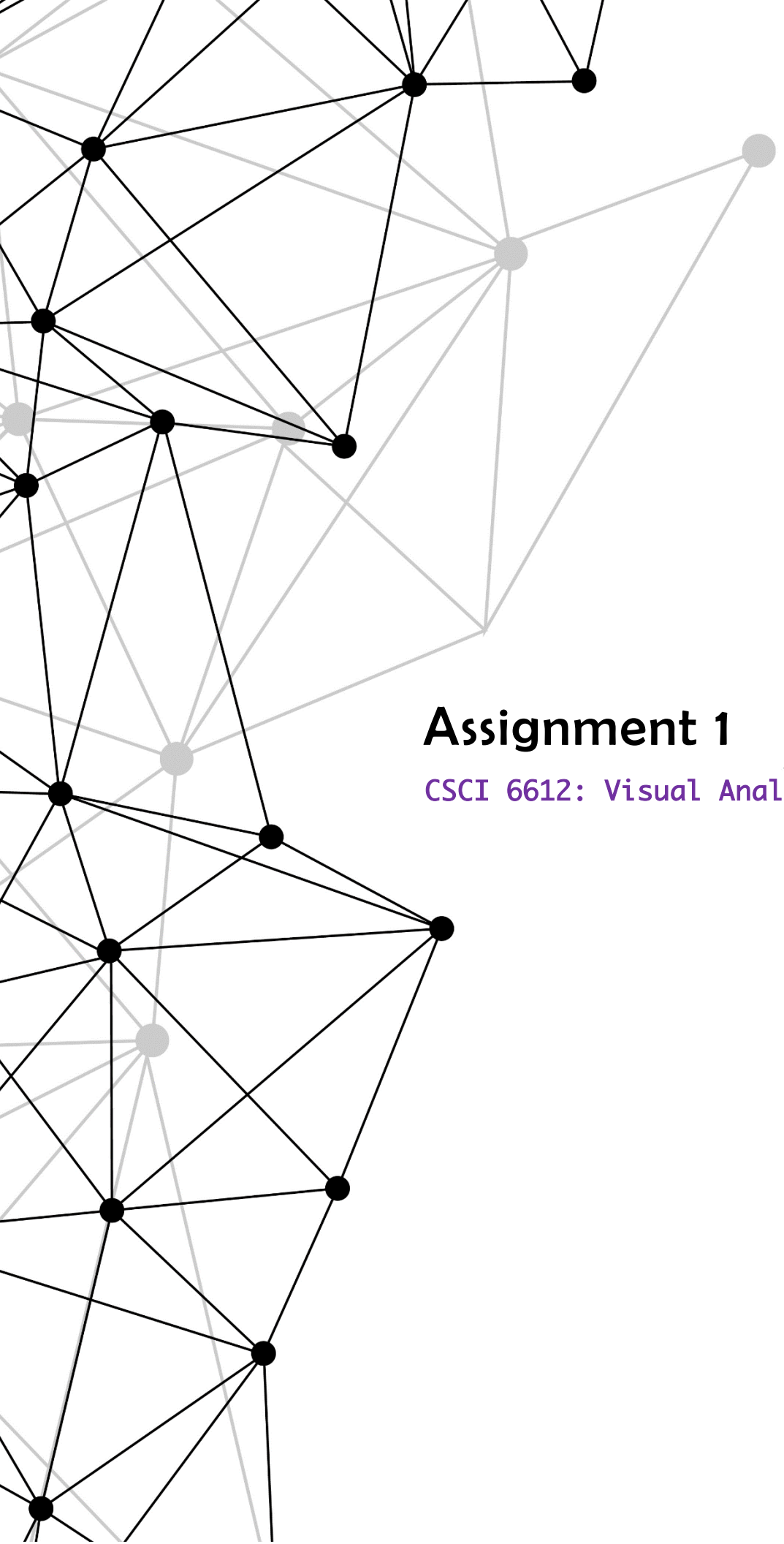


Assignment 1

CSCI 6612: Visual Analytics

Gaganpreet Singh
B00819217

Introduction

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

   The provided dataset has the following problems:

   a. The categorical values ('Workclass', 'occupation') in the given dataset has data in both Upper case and lower Case.

   b. The 1st column 'Age' has negative values. Also, many rows have the value as 0. For the given dataset, I have assumed that ideal value of Age is greater than 0.

   c. The columns 'Workclass' has both missing and corrupted values. The missing values are present as '?' and the corrupted values generally have spelling mistakes. On analyzing, I found that there were 75 unique values for this column in the given dataset. However, as per the given information, they should be only 7 types of 'Workclass'.

   d. The columns 'Occupation has both missing and corrupted values. The missing values are present as '?' and the corrupted values generally have spelling mistakes. On analyzing, I found that there were 474 unique values for this column in the given dataset. However, as per the given information, they should be only 15 types of 'Occupation.

   e. The column 'Education-num' have negative values. On analysis, I found that they have a one to one mapping with 'Education' column.

   f. The column 'Salary' has missing values. There are 827 rows having salary value as '?'.

   g. The value for the column 'Capital-gain' is 0 for the entire dataset. This column does not give any meaning full information.

2. **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.**

   The libraries used for pre-processing of the given dataset are:

   - **pandas & NumPy:** used for data manipulation and analysis of the dataset
   - **seaborn & matplotlib:** for plotting distribution charts for all the columns in the given dataset.
   - **re:** to perform search and replace operations for the corrupted values in columns having categorical data.
   - **csv:** to read csv files.

Preprocessing Steps:

a. **Case Sensitive data:**
All the rows in categorical columns were converted to lower case characters.

b. **For column 'age':**
All the negative values and 0's are replaced by the mean values of this column.

The reason for replacing the missing value with mean is that it is an efficient way that is simple and quick.

c. **For column 'Workclass' and 'occupation':**
i) The corrupted values were corrected by performing a search and replace operation using regular expressions.
ii) The missing values have been replaced by the most frequent value (i.e. *mode*) in the column.

The reason for using mode is that it works well with categorical data. Also, there is no perfect way to make up for missing data set values. T

d. **For column 'education-num':**
A new dictionary was created to map the value of the column 'education' to that of 'education-num'. Then all the existing values of this column were replaced by using the map.

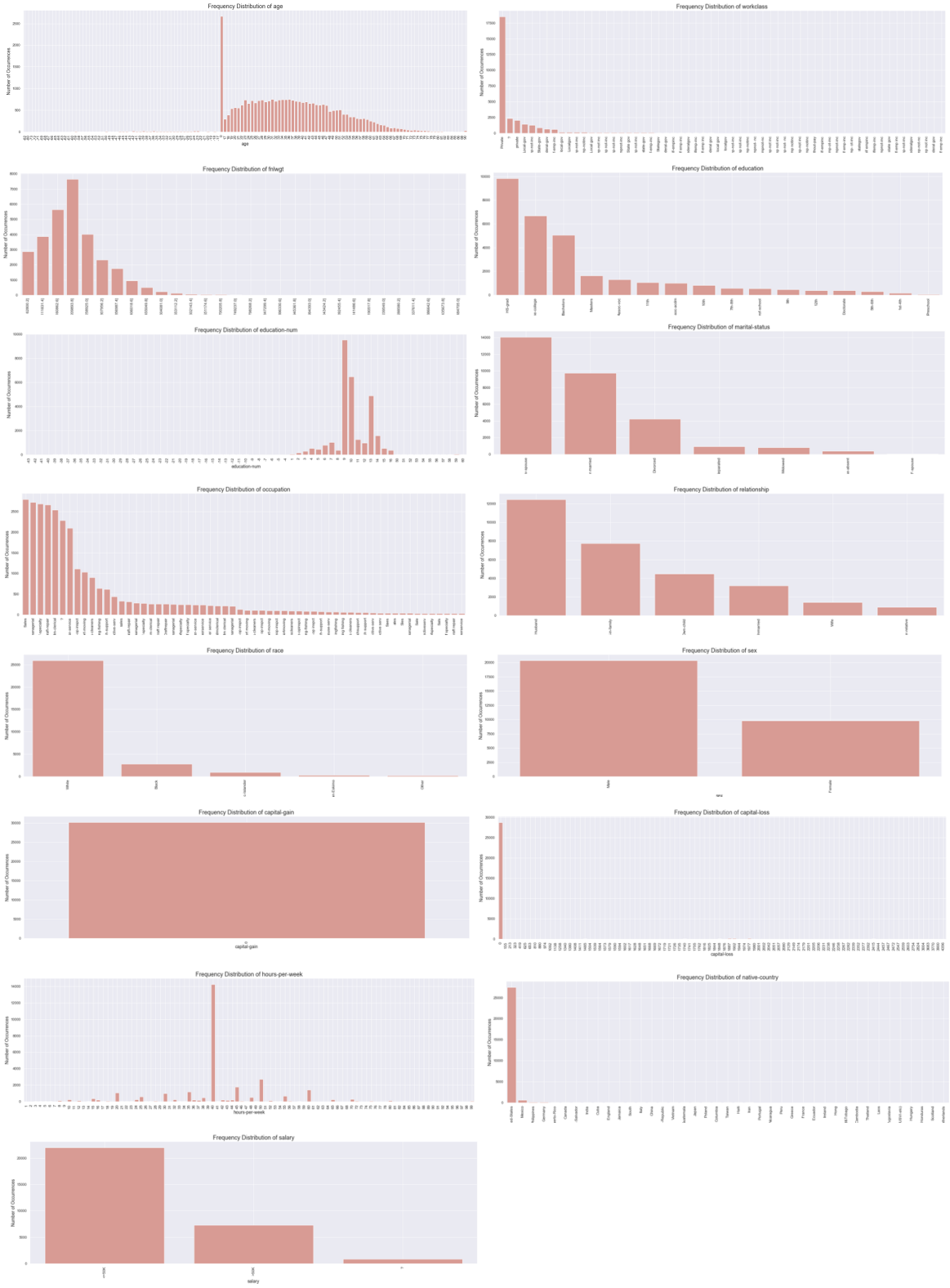
e. **For column 'salary':**
The missing values (originally present as '?') were replaced by the most common occurrence of that column i.e. the *mode* of that column.

f. **For column 'capital-gain':**
Since all the rows have a value of '0' for this column, it does not convey any meaning full information. Hence, this column is dropped.

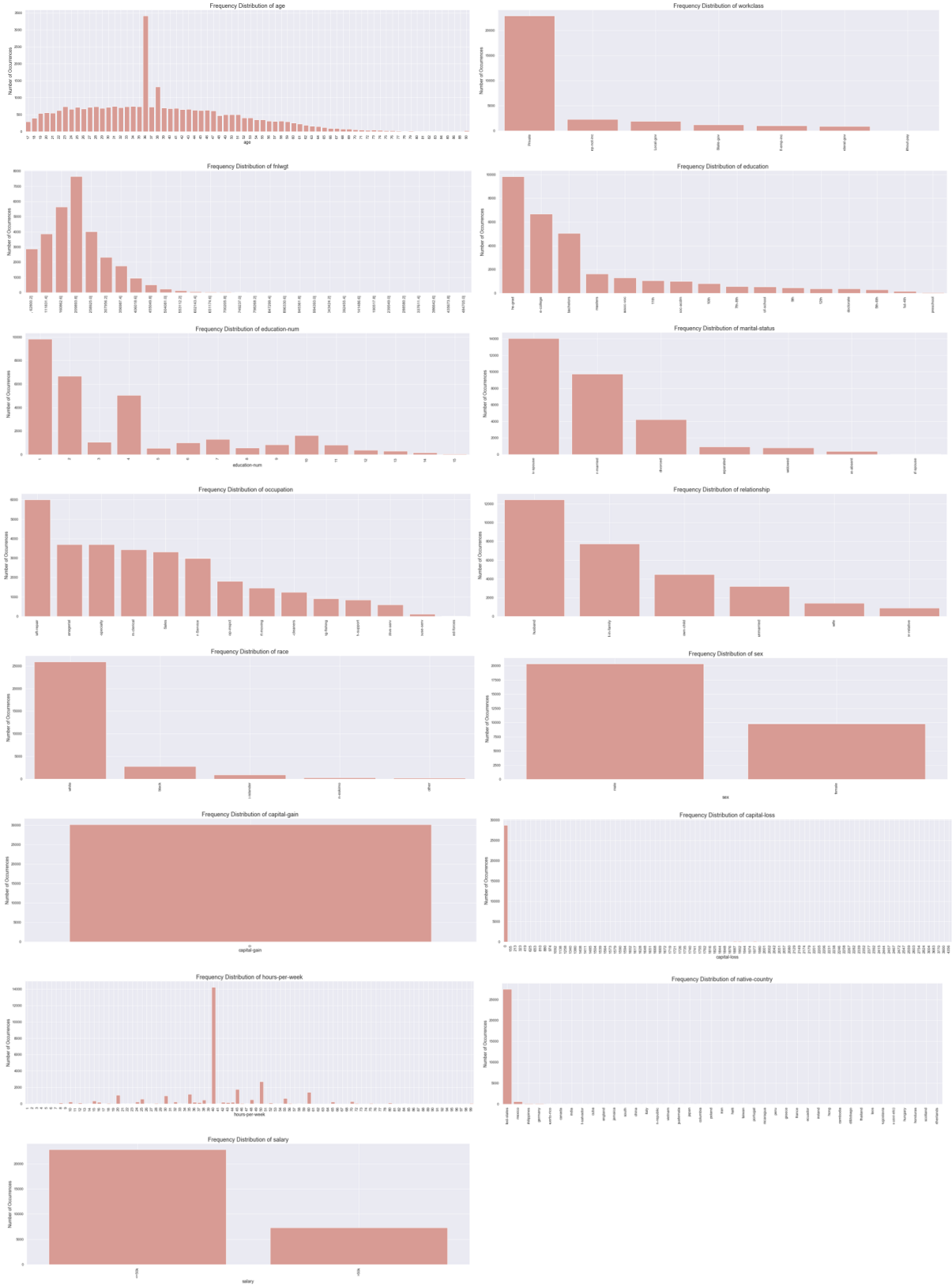
g. **For column 'fnlwgt':**
This column has numeric values. I could not interpret the actual information that it may contain. But since this column has no missing or invalid entries, I am assuming that it might carry some meaning full information. So, this column is not dropped.

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

a. Histograms before pre-processing of dataset:



b. Histograms after pre-processing of dataset:



References

[1] Handling Categorical Data in Python. (2019). Retrieved 22 September 2019, from https://www.datacamp.com/community/tutorials/categorical-data

[2] Data Mining — Handling Missing Values the Database. (2019). Retrieved 22 September 2019, from https://developerzen.com/data-mining-handling-missing-values-the-database-bd2241882e72.

[3] Python Lambda. (2019). Retrieved 22 September 2019, from https://www.w3schools.com/python/python_lambda.asp

[4] RegexOne - Learn Regular Expressions - Lesson 8: Characters optional. (2019)., from https://regexone.com/lesson/optional_characters