**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.

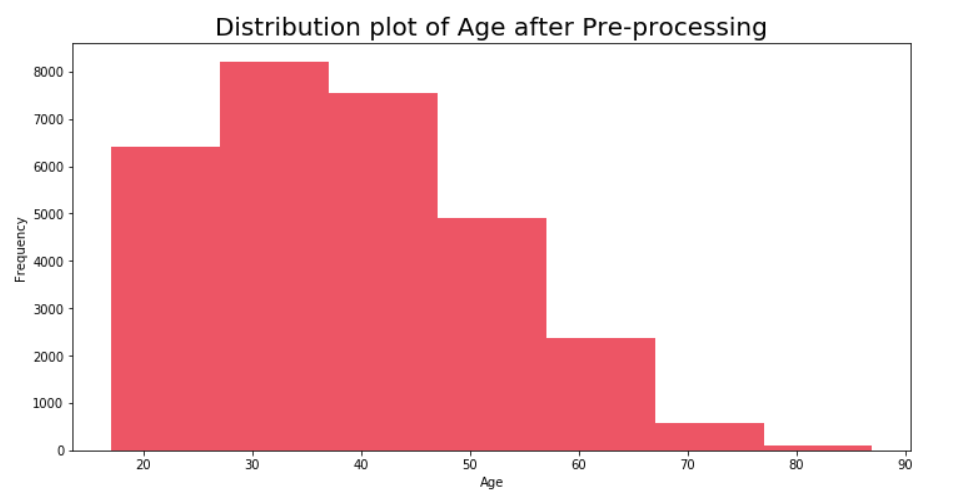
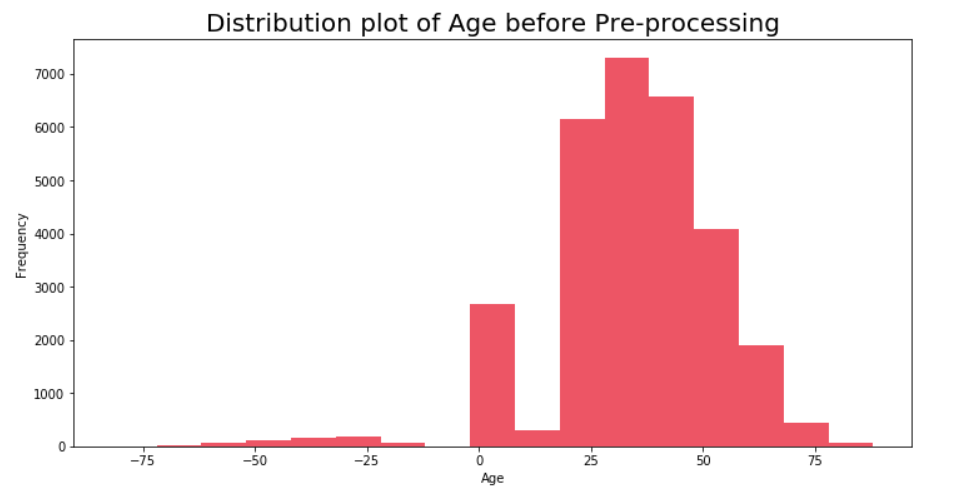
A2. For every column that you had to work on, explain how you fixed the data and justify your decision. If you used any libraries, briefly describe how you used them.

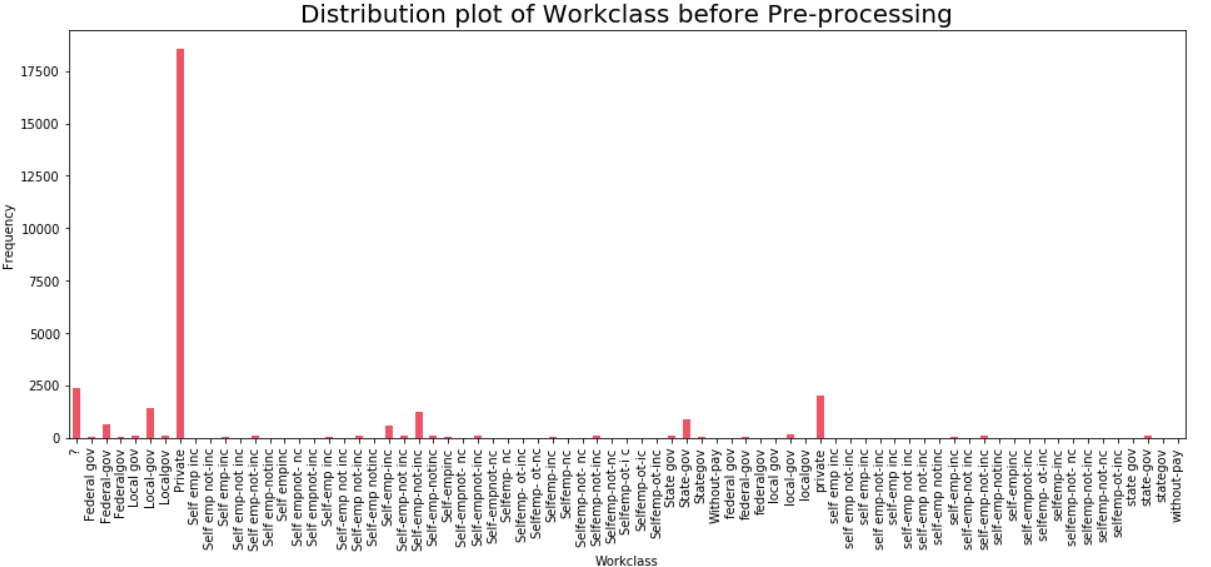
1. **Age**
   1. Replaced all ‘?’ values with 0 in order make the corrupt data consistent
   2. Implemented Forward Filling of Age with values equal to 0. This was done to maintain the distribution of data. By filling all the 0s with either mode or mean would alter the distribution. Therefore, chose forward filling method.
   3. 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.
   4. Libraries used: Abs() function for converting negative values into positive values. Pandas library to replace 0s with ‘ffill’
2. **Workclass**
   1. Rectified typographical errors by creating a robust dictionary of regex patterns. Iterated through this dictionary and replaced matching words with correct values.
   2. 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**
   1. 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.
   2. I checked if these values are unique to understand if they are identifiers of some kind, but they aren’t unique.
   3. Libraries used: Pandas library to check if there are unique values using ‘is\_unique’
4. **Education-Num**
   1. No noticeable correlation was observed with respect to other columns except with ‘education’ column.
   2. 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
   3. 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**
   1. Rectified typographical errors by creating a robust dictionary of regex patterns. Iterated through this dictionary and replaced matching words with correct values.
   2. 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**
   1. 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.

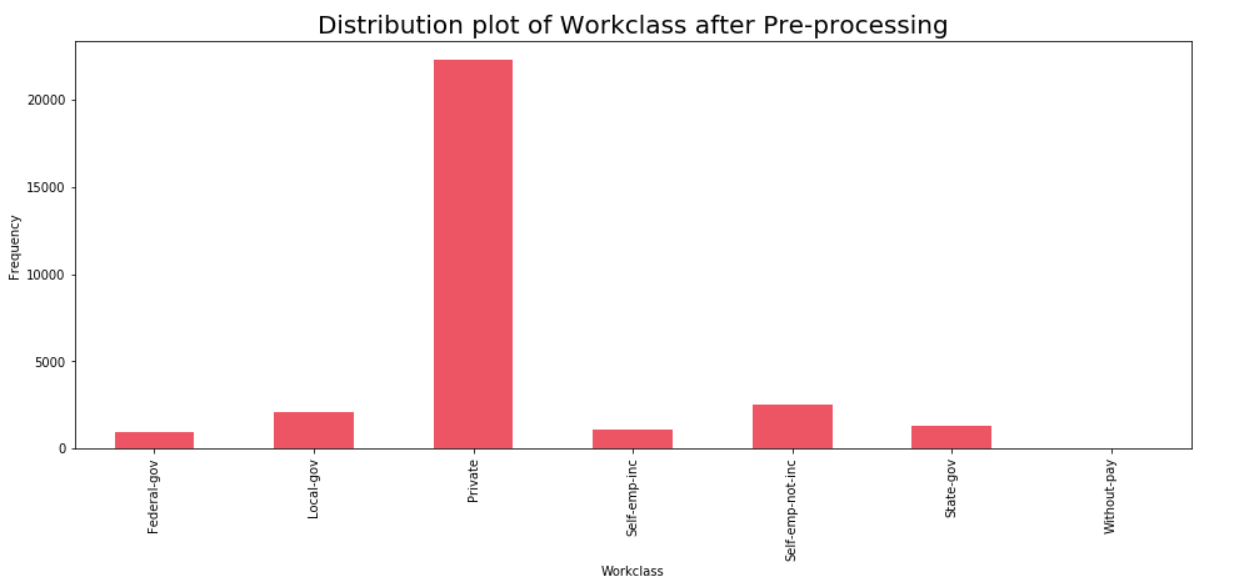
**Other Libraries used:**

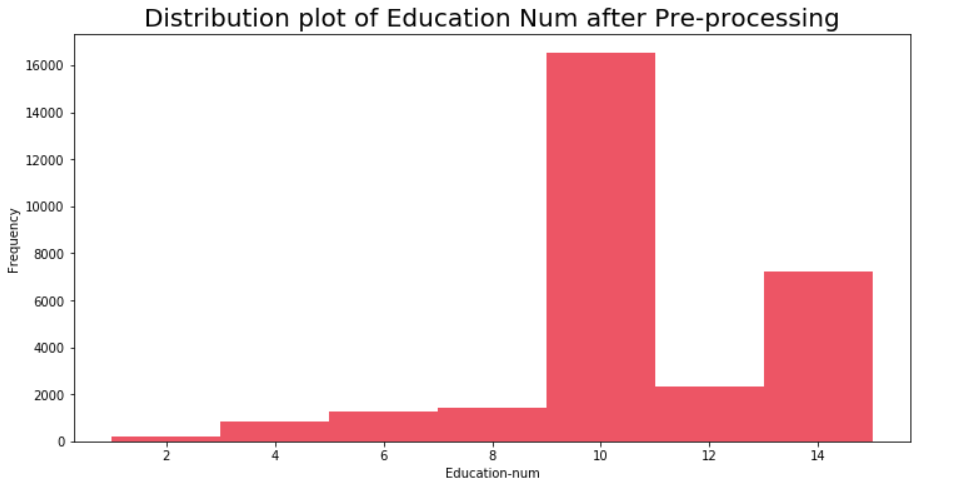
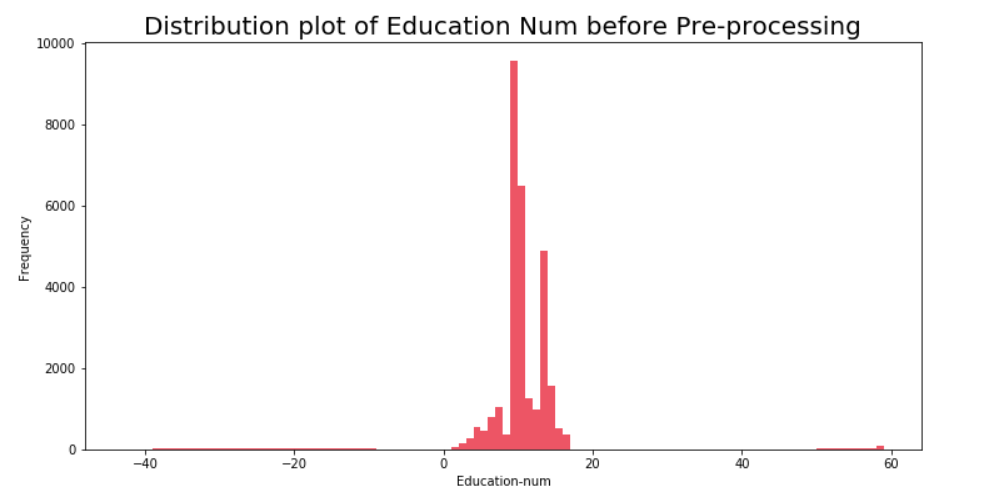
* Numpy library to check if the data has NaN or null values, and to set a range of values for histogram with numeric type.
* Matplotlib to display categorical and numeric histograms and to set labels.

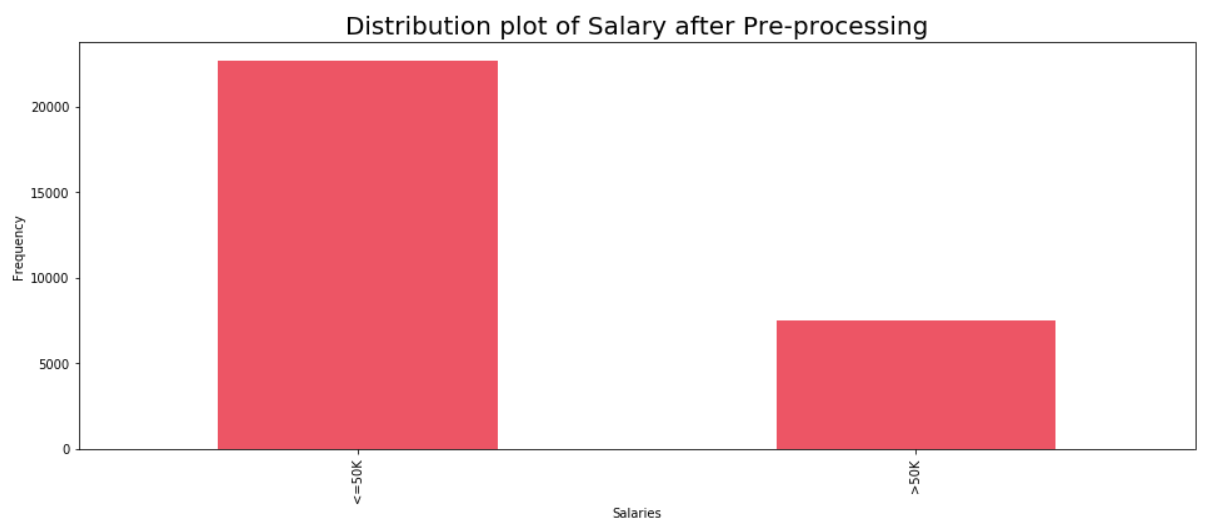
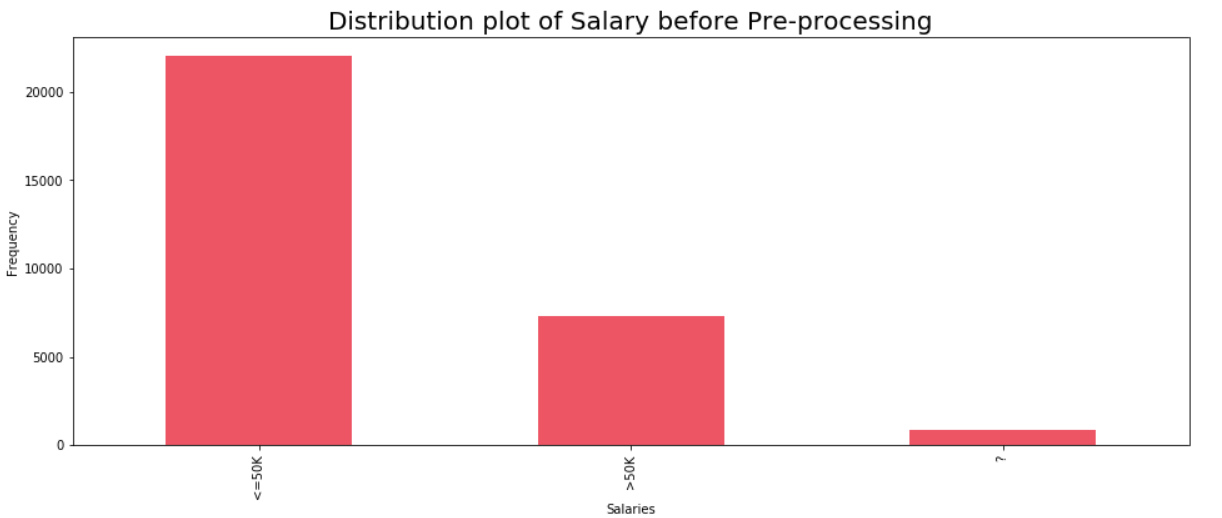
A3. Show the histograms for the columns that you fixed. (before and after)

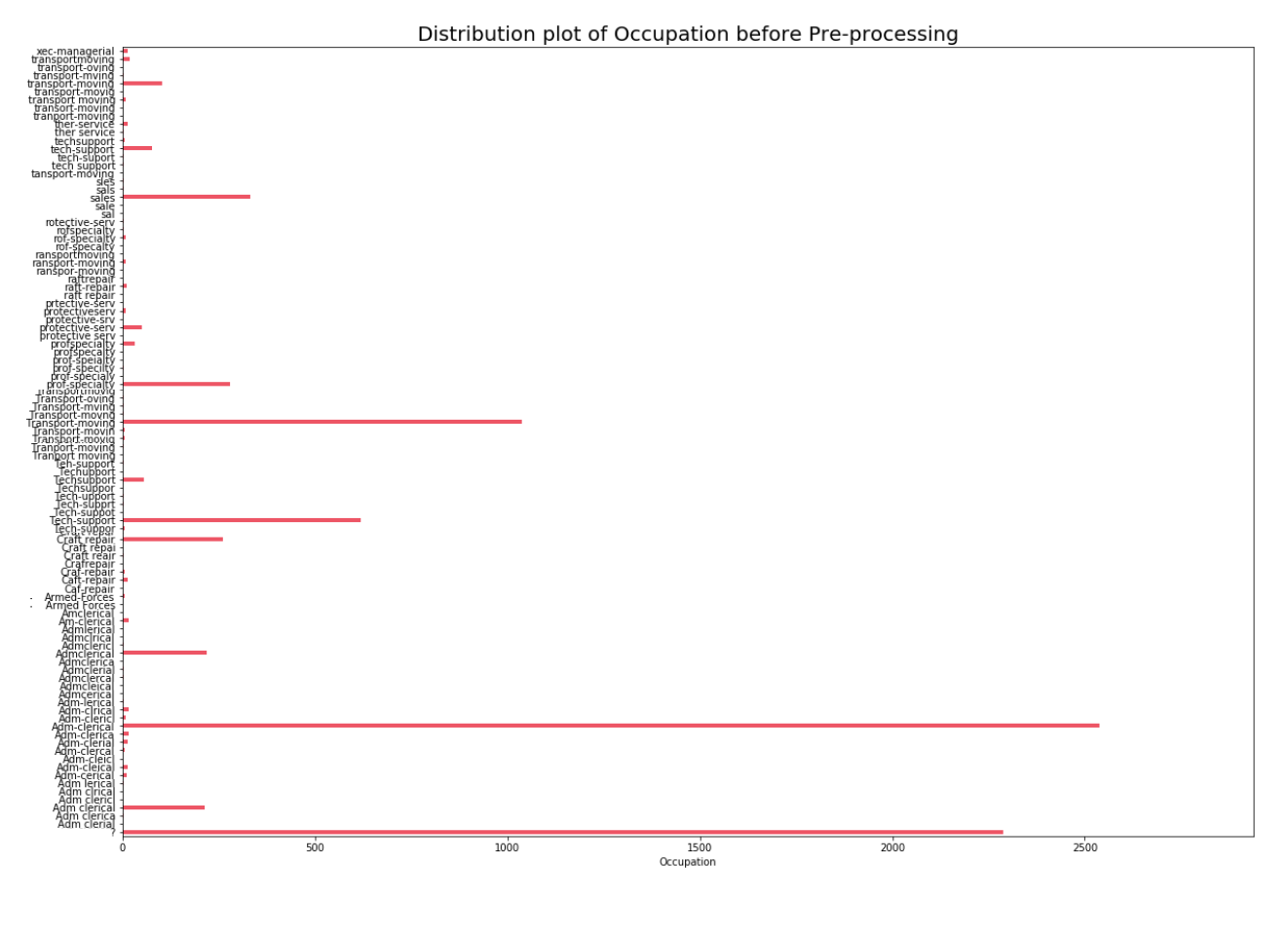


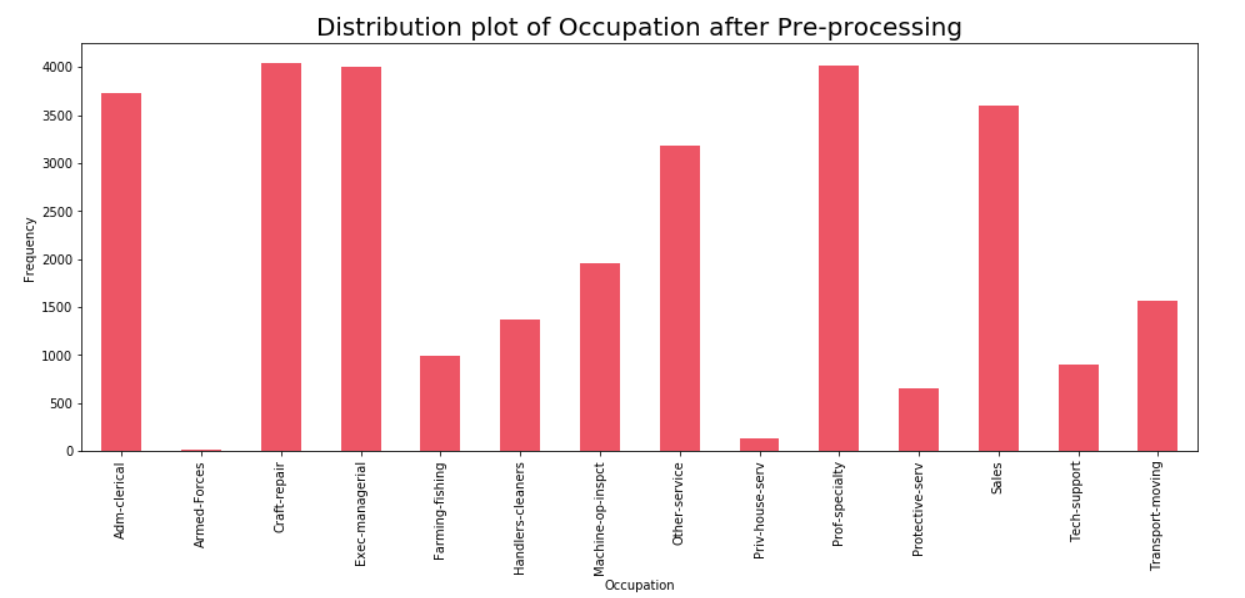












References

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/>