Programming for Data Analytic

November 2019

1 Second Assessment. First Project

1.1 Dataset Overview

The adult dataset contains data extracted from the (U.S.) census bureau. It contains approx. 49,000 records of census information taken in 1994 from many diverse demographics. The dataset for this project is made up of the following fields.

1. Age: continuous
2. Workclass: 8 values [Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.]
3. Fnlwgt: continuous. The # of people the census takers believe that observation represents.
4. Education: 16 values The highest level of education achieved for that individual [Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.]
5. Education-num: continuous. Highest level of education in numerical form.
6. Marital-status: 7 values [Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.]
7. Occupation: 14 values [Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces]
8. Relationship: 6 values. Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. We will be ignoring this attribute.
9. Sex: Male, Female
10. Capital-gain: continuous.
11. Capital-loss: continuous.
12. Hours-per-week: continuous. Hours worked per week.

13. Native-country: (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.)

14. Income: Yes, No. Whether or not the person makes more than $50,000 per annum income.

1.2 Project Specification

The objective of this project is to provide an insight into some of the relationships and trends that exist within the adult dataset. Please note you should use Pandas, Numpy and Matplotlib as means of analysing data within this dataset.

1.2.1 Tasks to do

1. Analysis the distribution of all work classes for different genders using an appropriate visualization technique. Use appropriate features for your visualization, See Figure 4. Note that you might need to add more features to this visualization e.g., labels, legend etc...

2. Investigate the relationship between the level of education and their hours of work per week for females only. Which education level has a larger variety of working hours? Use a visualization technique with appropriate features to show the result.

3. Use a visualization technique and depict which country has the maximum entry in the dataset.

4. Repeat the previous analysis but this time exclude the country which has the maximum entry and depict which two countries have the maximum of entries.

5. Investigate the relationship between age and working hours. Use a visualization technique with appropriate features to show the result. Which age interval has higher working hours. Use annotation and show a point that has relatively the maximum working hours.

6. Analysis education levels by using a visualization technique and find out the outliers for this feature of the dataset.

7. The information about *income* tells whether or not the person makes more than $50,000 per annum. Analysis this information using an appropriate visualization technique to depict how the income for males and females are different for each of the income class (i.e., $> 50k$ and $< 50k$). See Figure 2 Note that you might need to add more features to this visualization e.g., labels, legend etc..

8. Analysis the relationship between education and martial status by using an appropriate visualization technique. Depict which martial status has less different types of education.

9. Analysis the relationship between occupation and martial status by using an appropriate visualization technique. Depict which occupation has less different types of martial status. See Figure 5 Note that you might need to add more features to this visualization e.g., labels, legend etc...

10. Investigate and compare the relationship between people with bachelor degree and people with master degree with respect to their occupation. See Figure 1 Note that you might need to add more features to this visualization e.g., labels, legend etc..

11. Relationship feature in this dataset has a number of different values such as *wife*, *Unmarried* and etc. Analysis this feature with respect to the private workclass. Use a visualization technique that can show which relationship has the maximum number of *private* workclass, which one has the second maximum and so on. See Figure 3 to get an idea on what needs to be done. Note that you might need to add more features to this visualization e.g., labels, legend etc...

12. Analysis marital-status feature by using a visualization technique and find out the outliers for this feature in the dataset.

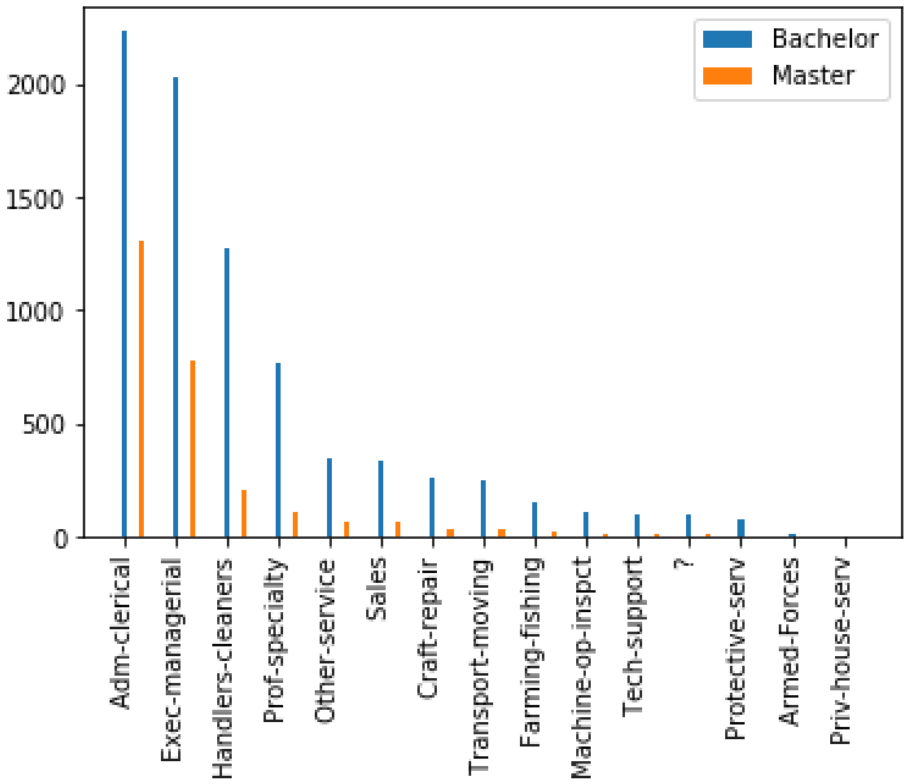

Figure 1: Example

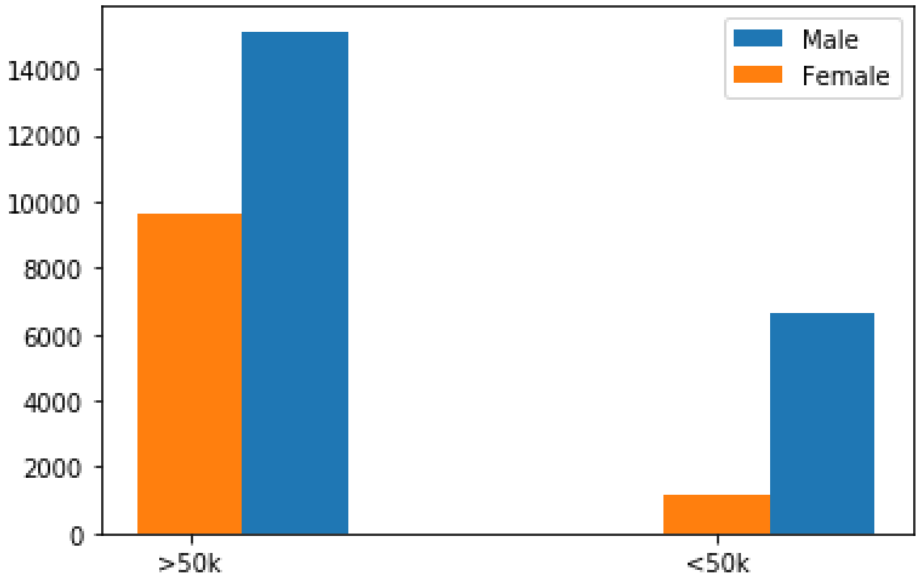

Figure 2: Example

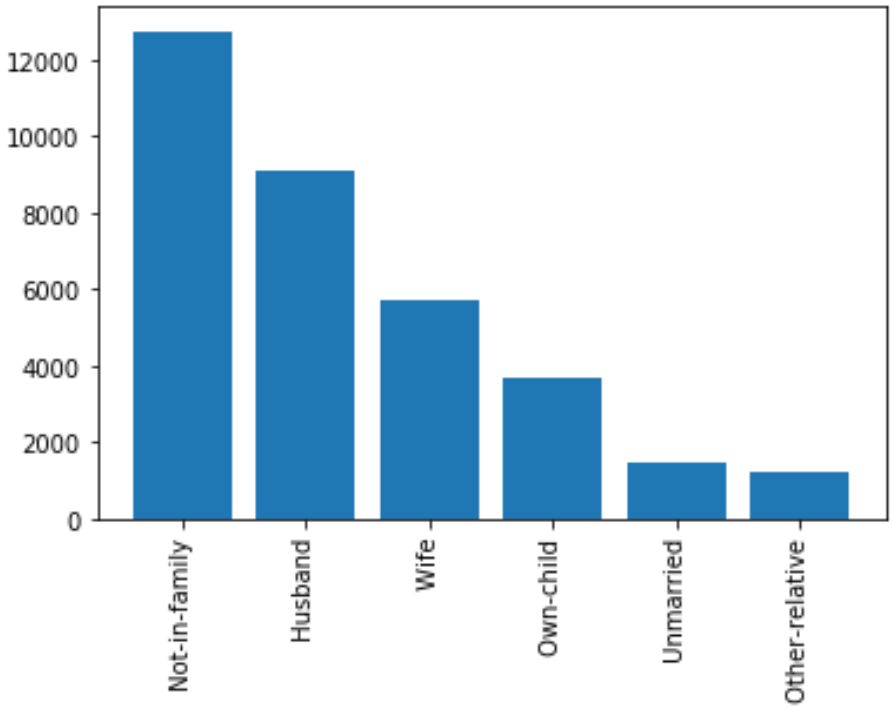

Figure 3: Example

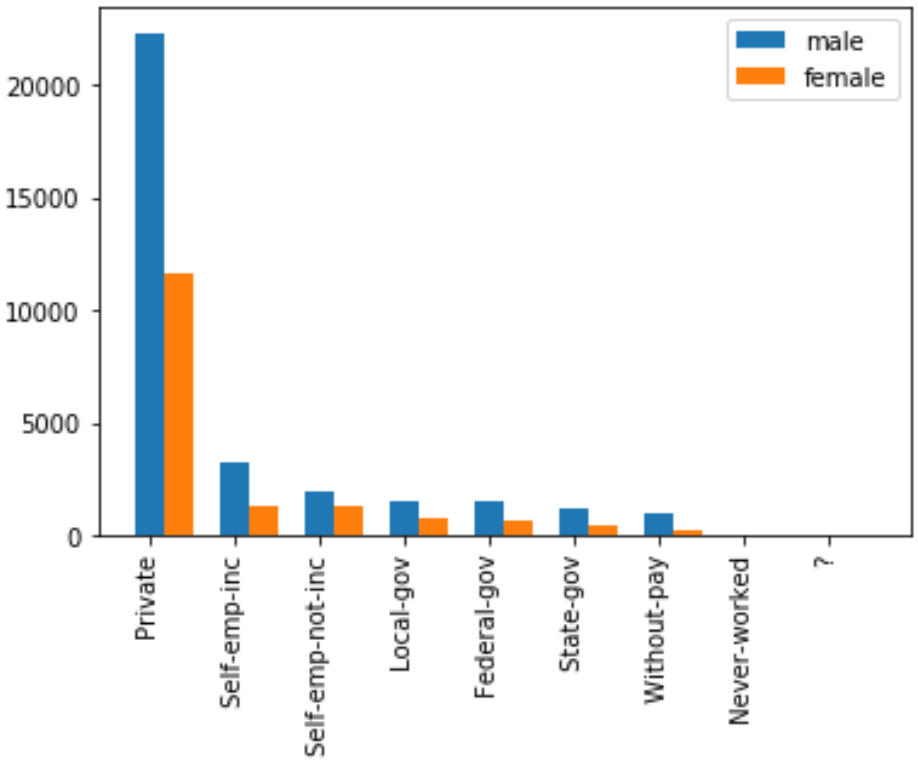

Figure 4: Example

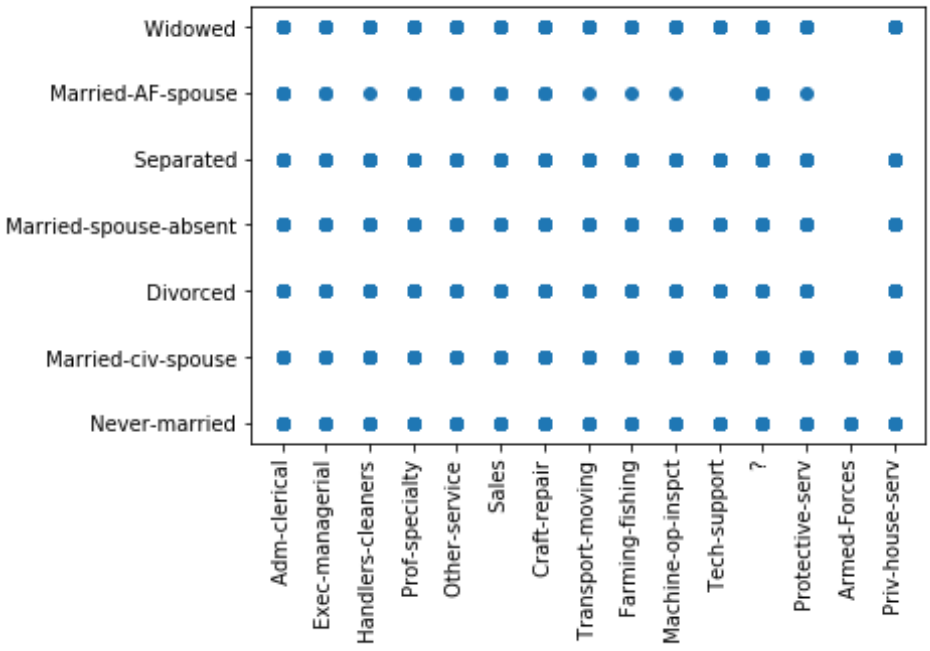

Figure 5: Example

1.3 Specification

Each visualization should have appropriate features such as a clear title, x axis title, y axis title and clear xticks and yticks if possible.

Data needs to be cleaned, e.g., some column names might have an space at the beginning or at the end of the text e.g., ' workclass' instead of 'workclass'. Columns should be check for *nan* values.

1.4 Submission and Deadline

Please use the python file template that is provided for you and complete your project in that file. Each task needs to be implemented as a separated function with one line interpretation as a comment below the function.

Please write your name and student ID as a comment in the designated area in the provided python file.

The template file should be re-named at the end using your student ID followed by letter s, for example if your student ID is: 1234567 the the python file should be named: s1234567.py

The deadline for this project is 27th of November at 23:59. Late submission is accepted with %10 penalty and the deadline for the late submission is 4th of December at 23:59.

Please submit your project via Canvas.

1.5 Rubric

This rubric is subject to change.

1. Correct task implementation (visualization) with meaningful labels, annotation (if needed) legend and etc. (100%)

2. Correct task implementation (visualization) with less/minimum meaningful labels, annotation etc. (80%)

3. Partly correct task implementation (visualization) with partly meaningful labels, annotation (if needed) legend and etc. (50%)

4. Wrong task implementation (visualization). (0%)