COSC 311 - Lab 3

Dr. Joe Anderson

Due: 1 October

1 Objectives

1. Practice efficiently manipulating data with Python
2. Use the `matplotlib`, `pandas` libraries
3. Gain familiarity with data import and plotting

2 Tasks

1. Download the "Adult" data set from the UCI Machine Learning data repository: `https://archive.ics.uci.edu/ml/datasets/Adult`. This dataset is record of adults, along with various occupational and lifestyle attributes. Each adult is "labeled" as to whether or not they make more or less than $50k per year. Using this as a driving label, one would typically want to design a process to determine what combinations of factors enable a person to make more than $50k per year.
2. Read the data into a `pandas` DataFrame object.
3. Pivot the data so that you can:
   (a) Rank each occupation in terms of most likely to earn more than $50k.
   (b) Calculate the median age of people who make more and less than $50k. Calculate the medians for each again, but now split apart by the sex of the adults.
   (c) Calculate the mean and median number of years of education held by adults who make more and less than $50k.
   (d) Plot a histogram to see the distribution of years of school held by adults who make more and less than $50k.
   (e) Plot a histogram to see the distribution of ages of adults who make more and less than $50k.
4. Practice some other types of visualization with your choice of variables:
   (a) Plot at least one *bar plot* that shows a trend within a variable that does not have inter-relationship.
   (b) Plot at least one *line plot* that shows a trend, where the x-axis variable manifests fully across the plot domain.
   (c) Plot at least one *scatter plot* of the data that suggests a relationship between two discrete variables, where the x-axis has an inter-related progression (amount, cost, process, etc.) but is not fully/evenly observed across the plot domain.
   (d) Plot a histogram of a quantity that varies across the x-axis domain, but where we would like to consider a general distribution, rather than individual observations.

5. Using pivoting, plotting, sorting, etc., attempt to answer the following questions with the data. Record your responses, with visual examples, in your notebook.
   (a) What is the relationship between between gender and whether a person makes more than $50k?
   (b) When a person makes more than $50, what is the relationship between gender and occupation? What about for less?
   (c) When are the "richest" professions in each possible native country?
   (d) What is the relationship between race and level of education? Does it further seem to relate to whether a person makes more than $50k?

3 Submission

Zip your source files and upload them to the assignment page on MyClasses. Be sure to include all source files, properly documented, a `README` file to describe the program and how it works, along with answers to any above discussion questions.