Lab 1: Exploratory Data Analysis

This lab is due at 11:59 PM ET on January 30, 2018.

You may work individually or in groups of two. If working in a group, only one student needs to submit the Jupyter notebook. This lab is worth 100 points.

I recommend that you use the latest Anaconda Python 3.6 distribution for this lab.

1. Get the data

For this lab, you will be using the "Adult" dataset, found at the UCI Machine Learning Repository. Download the datafile (adult.data) and the dataset description to your local machine.

2. Understand the data in context (10 points)

One of the most important aspects of any data mining project is to gain an understanding of the business or experimental context in which the data exists. Since we did not collect this data ourselves, it is important that we understand where the data came from and what it is for. Using the dataset description as a source, write a paragraph describing the purpose of the dataset and of the data mining task(s) you might use to analyze it. Consider the following questions:

- Why was this data collected?
- What does each record represent?
- Where did the data originally come from?
- What is the principal question that our data mining task seeks to answer?
- Are there other questions that we might be able to answer with this data?
- How will you know if you have mined useful data from it?
- How would you measure the effectiveness of a good analysis?

I don't expect you to be (or become) an expert on the US Census, but I do expect you to have some understanding of the context for this dataset. You don't need to do research for this lab; if you can't get satisfactory answers from the dataset description, do your best to imagine the context for this data and work with that. Feel free to make assumptions, but be sure to state those assumptions in your writeup. The point here is to think about the *data in context*.

3. Understand the data (75 points)

Next, you will perform an exploratory analysis of the dataset:

- Describe the meaning and type (e.g. categorical/numeric, binary/discrete/continuous) of the data for each attribute (10 points)
- Verify data quality: explain any missing values, duplicate data, or outliers. What, if anything, do you need to do about these? (10 points)
- Provide appropriate basic statistics (e.g. range, mode, mean, median, variance, counts) for the most important/interesting attributes. Describe what these mean and why they are important or interesting. (10 points)
- Visualize the most important or interesting attributes (at least 5) using appropriate techniques. For each visualization, provide an interpretation explaining why it is appropriate or interesting. What does each visualization tell us? (15 points)

- Explore the relationships among the attributes, excluding the class attribute. Use scatter plots, correlation matrices, cross-tabulations, group-wise averages, or other appropriate techniques. Explain and interpret any interesting relationships. (15 points)
- Identify and explain any interesting relationships between the class attribute and the other attributes. You may refer to earlier visualizations or create new ones. (10 points)
- What attributes could you add to this dataset, either by altering the data collection process or by creating new attributes from existing ones? Explain in detail. (5 points)

4. Exceptional Work (10 points)

For this section, you are free to provide whatever additional analysis (related to data quality and data preprocessing) you wish. Some ideas for this part to consider include:

- Implement dimensionality reduction using PCA or some other technique, then visualize and interpret the results.
- Implement the feature creation, perhaps using one of the approaches you described in part 3.
- Implement the data cleaning steps you identified in part 3. Be sure to show the effects of cleaning through appropriate statistical analyses and visualizations.
- Select two small (100 records each) data subsets using both simple random sampling and some form of stratified sampling, then repeat some of the basic analyses from part 3. Do you get similar results? Why or why not?

You may choose any one of the items above or pursue something of your own device. To get full credit for this part, your "exceptional work" need not be lengthy or deeply involved, but it must be non-trivial and provide insight not already obtained anywhere else in the report.

5. Submit your work (5 points)

Create a single notebook containing all code, visualizations, and written descriptions for this lab. Make sure that:

- All your work is in one notebook. (1 point)
- All the code cells have been run and outputs are shown. (1 point)
- Outputs don't contain any errors. If there are persistent errors, provide an explanation. (1 point)
- Your notebook does not contain any extraneous cells (e.g. tests, examples) that are not part of your analysis. (1 point)
- You have included a markdown cell near the top of your notebook containing your name(s), the course, the lab (e.g. "Lab 1, Exploratory Data Analysis"), and the date. (1 points)

Save your notebook using a descriptive filename, such as **`lab1_burkhardt_5.ipynb`**, then submit the notebook (.ipynb file) via Carmen, under Lab 1.

There's no need to submit your data file(s).