Assignment 1

For this assignment you will experiment with Python. NumPy, and Pandas in order to perform some basic data preprocessing and analysis tasks.

You will work with a modified subset of a real data set of customer for a bank. The **data is provided in a CSV formatted file** with the first row containing the attribute names. The **description of the the different fields** in the data are provided in **this document**. You must only use Python, NumPy, Pandas, Matplotlib to perform the tasks for this assignment

1. Explore the general characteristics of the data as a whole: examine the means, standard deviations, and other statistics associated with the numerical attributes; show the distributions of values associated with categorical attributes; etc.
2. Suppose that the hypothetical bank is particularly interested in customers who buy the PEP (Personal Equity Plan) product. Compare and contrast the subsets of customers who buy and don't buy the PEP. Compute summaries (as in part 1) of the selected data with respect to all other attributes. Can you observe any significant differences between these segments of customers? Discuss your observations.
3. Use **z-score normalization** to standardize the values of the income attribute. [Do not change the original income attribute in the table.]
4. Discretize the age attribute into 3 categories (corresponding to "young", "mid-age", and "old"). [Do not change the original age attribute in the table.]
5. Use Min-Max Normalization to transform the values of all numeric attributes (income, age, children) in the original table (before the transforamtions in parts 3 and 4 above) onto the range 0.0-1.0.
6. Convert the table (after normalization in part 5) into the standard spreadsheet format. Note that this requires converting each categorical attribute into multiple binary ("dummy") attributes (one for each values of the categorical attribute) and assigning binary values corresponding to the presence or not presence of the attribute value in the original record). The numeric attributes should remain unchanged. Save this new table into a file called bank_numeric.csv and submit it along with your assignment. [Hint: you might consider using the get_dummies for Pandas data frames.]
7. Using the standardized data set (of the previous part), perform basic correlation analysis among the attributes. Discuss your results by indicating any significant positive or negative correlations among pairs of attributes. You need to construct a complete Correlation Matrix. Be sure to first remove the Customer ID column before creating the correlation matrix. [Hint:you can create the correlation matrix by using the `corr()` function in Pandas or `corrcoef` function in NumPy].
8. Using Matplotlib library and/or ploting capabilities of Pandas, create a scatter plot of the (non-normalized) Income attribute relative to Age. Be sure that your plot contains appropriate labels for the axes. Do these variables seem correlated?
9. Create histograms for (non-normalized) Income (using 9 bins) and Age (using 15 bins).
10. Using a bar graph, plot the distribution of the values of the region attribute.
11. Perform a cross-tabulation of the region attribute with the pep attribute. This requires the aggregation of the occurrences of each pep value (yes or no) separately for each value of the region attribute. Show the results as a 4 by 2 (region x pep) table with entries representing the counts. [Hint: you can either use Numpy or use aggregations functions in Pandas such as groupby() and cross-tab().] Then, either using Matplotlib directly or the plot() function in Pandas create a bar chart graph to visualize of the relationships between these sets of variables. [Hint: **This example** of creating simple bar charts using Matplotlib may be useful.]

Notes on Submission: You must submit your Jupyter Notebook (similar to examples in class) which includes your documented code, results of your interactions, and any discussions or explanations of the results. Please organize

your notebook so that it's clear what parts of the notebook correspond to which problems in the assignment. Please submit the notebook in both IPYNB and HTML formats (along with any auxiliary files). Your assignment should be submitted via D2L.

Bank Data Description

The marketing department of a financial firm keeps records on customers, including demographic information and, number of type of accounts. When launching a new product, such as a "Personal Equity Plan" (PEP), a direct mail piece or a targeted email, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product. Based on this database of prior cases, the managers decide to use data mining techniques to build customer profile models in order to predict the behavior of future customers.

The data is contained in the file bank_data.csv. Each record is a customer description where the "pep" field indicates whether or not that customer has purchased a PEP. For classification problems, this field is used as the target attribute (with "YES" and "NO") as class labels.

The data contains the following fields

id	a unique identification number (categorical, str)
age	age of customer in years (numeric, int)
income	income of customer (numeric, float)
children	number of children (numeric, int)
gender	MALE / FEMALE
region	INNER_CITY/RURAL/SUBURBAN/TOWN
married	Customer married (YES/NO)
car	Customer owns one or more cars (YES/NO)
save_acct	Customer has a savings account (YES/NO)
current_acct	Customer has a current checking account (YES/NO)
mortgage	Customer have a mortgage (YES/NO)
pep	Customer purchased a PEP, Personal Equity Plan (YES/NO)