Pandas Exercises

Introduction

Below are a list of tasks to be accomplished using Python. Please save your work for use later in the course.

Our goals for this lab are:

- Import Data
- Examine and Clean Data
- Merge additional data

Always remember to include appropriate headers and comments in your code.
We will use two datasets for this lab:

- ex_dm_person_demo.csv: This is a dataset with person level data, including the cluster identifier
- ex_dm_cluster.csv: This is a dataset with a cluster identifier and a trait about each cluster

Instructions

1. Create a data frame from ex_dm_person_demo.csv, and name the data frame "df_person"
    a. Inspect the text file with a text editor
    b. Import the dataset with read_table()
2. Do an initial inspection of the dataset
    a. How many rows and columns does it have?
    b. What are the dtypes for each variable in the dataset? (.dtypes attribute)
3. Correct the classes of the variables in the dataframe if needed
    a. Age, lab values, and income should be numberic of some type
    b. Clust_id, Gender, Education, and Outcome should be string
    c. Make baseline date a date value (to_Datetime())
4. Clean Data
    a. Consider the values of "999" and "" to be missing, and code them as such
    b. Check the numeric variables for outliers, code them to be missing
    c. Lab values should never be below 1. Change any lab values below 1 to be 0.

5. Create some new variables…
    a. Lab1_lab2, defined as the ratio of lab1 to lab2
    b. edu.cat, with values of "college" (grad and undergrad) and "no college"
    c. Create a categorical variable for age based on quartiles (use the cut function)
    d. rename the following variables:
        i. gender = sex
        ii. PID = ID
6. Make a data frame that only includes the following variables: P_ID, clust_id, baseline_date
7. Make a data frame which only includes the first, fourth, and fifth observations (but all variables)
8. Make a data frame which has the first, fourth, and fifth observations, and the second and fourth variables
9. Sort your original data frame (dept.data) based on gender, then education, then descending income

Merging Data:

10. Create a data frame from the file ex_dm_cluster.csv
    a. Import this with read.csv
11. Examine this dataframe
    a. How many unique values of clust_id are present? Compare these levels to the unique values of clust_id in the person level data frame.
12. Merge this information back to our original table.
    a. Perform a left join from the cluster table to the person level table. You should have all the values from the person level table, so the resulting dataframe will be the same size as the original person level dataframe. Use the merge function.

Descriptive Statistics:

13. Univariate Statistics:
    a. Describe the date range of baseline date
    b. What percent of this dataset is Male?
    c. What proportion of people have an income above $70,000?
    d. What is the mean age of the dataset?
14. Grouping Statistics:
    a. What is the mean income by gender?
    b. What is the mean income by gender and education?
    c. What is the probability of outcome == yes across gender?