Problem set 1 (Pandas bootcamp)
Math 446, Spring 2025
Due date: Sunday, February 2 at 23:59 on Gradescope

Instructions:

1. In this problem set/assignment, you will be using Pandas to do basic data analysis on given datasets (download them from Brightspace).

    - This problem set has less mathematical content due to the structure of the first two/three weeks.
    - You are expected to use online resources as an aid for coding (i.e. Stack Overflow, Google, Pandas documentation, so on).
    - There are some things that may be a bit ambiguous - use your best judgment and write appropriate documentation (this ambiguity happens all the time in real life) or ask if you need guidance.
    - Depending on your programming background, the estimated completion time should be approximately 3 - 12 hours. This is a large range. Because you have approximately 2 weeks to do this, a suggestion would be to start early to see which end you fall under/are closer to.

2. **Submission Guidelines:** Please complete this assignment using a Jupyter notebook. You should submit an `.ipynb` file on Gradescope.

    - The submission file should have your code and the output of the code visible.
    - It should be written so that either Farhad or Alvin can download the file and run the code, assuming the dataset is in the correct location. One way to check this is to restart your kernel and run the whole notebook and see if this results in what you expect.
    - Please write comments as necessary, and make sure your code is readable (e.g. use # or markdown or something else).

Remark. Usually, there will be a question at the end that asks for feedback regarding the course/problem set.

Questions

1. You will be analyzing the `Adult` dataset. It can be found on Brightspace or here:
`https://archive.ics.uci.edu/dataset/2/adult`. You can download the .data file from Brightspace or the uci link (you already can read it as a csv).

   a) Work with the file `Adult.data`. Create a dataframe from the `Adult` dataset using

   `pd.read_csv`

   How many data points are in the data set?

   - Something may seem off about the data. Describe something that seems suspicious to you. Part (b) will confirm or refute these suspicions.
   - Do some basic exploration of the feature 'Male' to confirm your suspicion (e.g.

     `value_counts, describe`

     or otherwise.) Hint: pay attention to the exact string involved.

   b) Now work with the file `Adult.csv`. Create a dataframe from the `Adult` using

   `pd.read_csv`

   Common names for such a dataframe would be among

   `df, data, adult`

   but you are free to choose whatever name you want. How many adults are in this dataset, and how many features/variables are there? This should confirm your suspicions in (a).

   c) How many men and how many women are represented in this dataset?

   d) What is the average age of the women?

e) What is the proportion of German citizens?

f) What are mean value and standard deviation of the age of those who receive more than 50K per year and those who receive less than 50K per year?

g) Is it true that people who earn more than 50K have at least high school education?

h) Display age statistics for each race and each gender. Use `groupby()` and `describe()`. Find the maximum age of men of `Amer-Indian-Eskimo` race. Hint: write a for loop [you may not need to use an agg function in this case].

i) Among which is the proportion of those who earn a lot (`>50K`) greater: married or single men? Consider as married those who have a marital-status starting with Married (`Married-civ-spouse`, `Married-spouse-absent` or `Married-AF-spouse`) - the rest are considered bachelors.

j) What is the maximum number of hours a person works per week? How many people work such a number of hours and what is the percentage of those who earn a lot among them (see (i) for definition of 'a lot')?

k) Count the average time of work those who earning a little and a lot for each country. Hint: write a for loop (similar to (h)).

2. You will be analyzing the `cardio` dataset (download it on Brightspace).

a) As in (1), Create a dataframe from the `cardio` dataset using

```
pd.read_csv
```

This time, if you don't input any arguments, you'll notice that if you type e.g. `df.head()`, the columns will have a semicolon. Input the relevant `sep` argument to address this. How many total people are in this dataset?

b) How many men and women are present in this dataset? Notice that unlike in (1), the labels are not explicitly present! Hint: are there any proxies for sex that you might be able to use? Height may be a good feature.

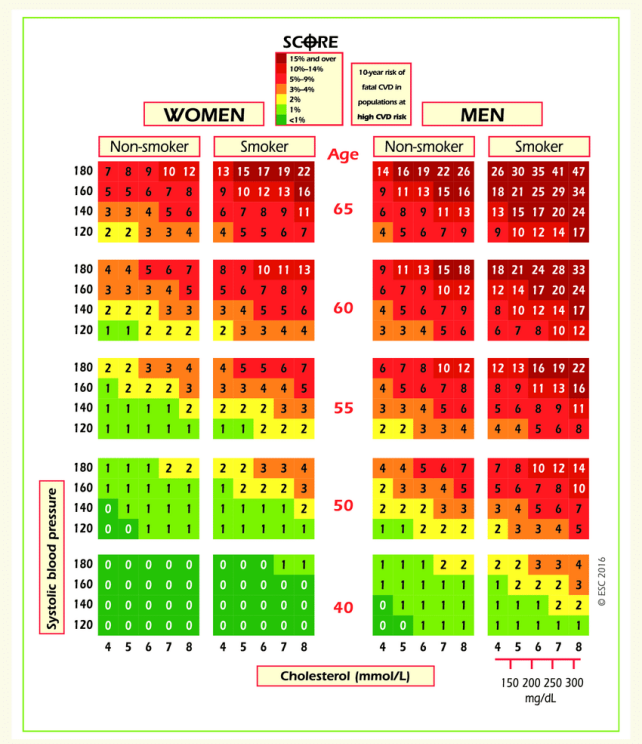
c) Who more often report consuming alcohol – men or women?

d) What's the rounded difference between the percentages of smokers among men and women?

e) What's the rounded difference between median values of age (in months) for non-smokers and smokers? Note: be careful of the units in the dataset.

f) The following figure was taken from the European Society of Cardiology, where they give a SCORE scale that calculates the risk of death from a cardiovascular disease in the next 10 years.

For the upper-right rectangle that represents smoking men aged 60 to 65, we see a value of 9 in the lower-left corner and a 47 in the upper-right corner. This means that for people in this gender-age group whose systolic pressure is less than 120 the risk of a CVD is estimated to be 5 times lower than for those with the pressure in the interval [160,180).



Let's compute this ratio with our data. We do this in a couple of steps.

Create a new column called

```
age_years
```

– rounded age in years. For this task, select people aged from 60 to 64 inclusive.

Cholesterol level categories in the figure and in our data are different. In the figure, the values of cholesterol feature are as follows:

i. 4 mmol/l $\rightarrow$ 1
ii. 5-7 mmol/l $\rightarrow$ 2
iii. 8 mmol/l $\rightarrow$ 3

Calculate fractions of ill people (with cardiovascular disease) in the two groups of people described in the task. What's the ratio of these two fractions? Note that the feature

```
ap_high
```

describes "Systolic blood pressure."

g) Create a new feature – BMI (Body Mass Index). To do this, divide weight in kilograms by the square of height in meters. Normal BMI values are said to be from 18.5 to 25. Which of the following are correct? Recall that "ill" people are defined as people with cardiovascular disease.

i. The median BMI in the sample is within boundaries of normal values.
ii. Women's BMI is on average higher then men's.
iii. Healthy people have higher median BMI than ill people.
iv. For healthy and non-drinking men, BMI is closer to the norm than it is for healthy and non-drinking women.

h) The data certainly needs to be cleaned. We consider the following to be errors:

i. diastolic pressure is higher than systolic pressure. Note: diastolic pressure is given by the feature

`ap_lo`

ii. height is strictly less than 2.5%-percentile (use `pd.Series.quantile`)

iii. height is strictly more than 97.5%-percentile

iv. weight is strictly less then 2.5%-percentile

v. weight is strictly more than 97.5%-percentile

Create a new dataframe that filters out the above. What percent of the original data (rounded) did we filter out here?

3. Continue using the filtered dataset as in (2h). Your filtered dataframe should have 63259 rows. In this question, we will be focusing on visualization using the `seaborn` library. You will also want `matplotlib`.

a) To understand the features better, you can create a matrix of the correlation coefficients between the features. Plot a correlation matrix using `heatmap()`. You can create the matrix using the standard Pandas tools with the default parameters.

b) Which two features have the strongest Pearson's correlation with the gender feature?

c) From our exploratory analysis, we were able to deduce which data points represented women, and which represented men. Let's do the same thing graphically here. Create a violin plot for the height and gender using `violinplot()`. Use the parameters:

i. `hue` to split by gender,

ii. `scale` to evaluate the number of points for each gender.

In order for the plot to render correctly, you need to convert your DataFrame to long format using the `melt()` function from pandas. See for example this.

d) Create two kernel density plots (see here) of the height feature for each gender on the same chart. You will see the difference between the genders more clearly, but you will be unable to evaluate the number of points in each of them.

e) In most cases, the Pearson coefficient of linear correlation is more than enough to discover patterns in data. But let's go a little further and calculate a rank correlation. It will help us to identify such feature pairs in which the lower rank in the variational series of one feature always precedes the higher rank in the another one (and we have the opposite in the case of negative correlation). Calculate and plot a correlation matrix using the Spearman's rank correlation coefficient.

   i. Which pair of features has the largest rank correlation?

f) Create a count plot using `countplot()`, with the age on the X axis and the number of people on the Y axis. Each value of the age should have two columns corresponding to the numbers of people of this age for each cardio class.

g) What is the smallest age at which the number of people with CVD outnumbers the number of people without CVD?

4. Feedback (Optional). How long did this take you? How difficult was this assignment? Any other comments (about the course or otherwise)?