Assignment

March 22, 2023

0.1 Organizing and analyzing data in a pandas dataframe

This is an initial exercise to familiarize you with setting up a pandas dataframe to permit further data analysis in Python. Please follow these steps:

0.1.1 Step 1

- Import the pandas library and other libraries you may need for this project. You may need to gradually add more import statements as more things occur to you.
- Download the "test" and "train" data from the kaggle "titanic" problem website.
- Import the training set (the "train.csv" file) into a pandas dataframe.

[ ]:
```
## Python code here
```

Notice that, for each passenger, the data includes their class of travel, some demographic information, and whether or not they survived. Using the pandas grouping and summarizing commands, find:

- the average age of all passengers
- the average ages of the male and female passengers
- The percentage of survivors among male and among female passengers.

[ ]:
```
## Python code here
```

Build a two x two contingency table comparing gender vs survival. What are the conditional probabilities:

- P(survived | female)
- P(survived | male)

Discussion: How can you decide if this difference is statistically significant?

[ ]:
```
## Python calculations here
```

Group the data into age ranges that seem appropriate to you and compute the fraction of survivors in each age group. Does this lead to any hypotheses?

[ ]:
```
## Python calculations here
```

0.1.2 Explore the data

Try some other ways of slicing this data and discuss with your classmates any interesting hypotheses that you develop.