MAT 128 Lab 6: Selecting dataframe rows based on conditions

Motivating question: Do taxis with more passengers take longer trips?

We will make a hypothesis, which is a tentative, testable answer to a question. Since the cost of the taxi trip with more passenger would be split between more people, it seems likely that taxi trips with more passengers are longer. For example, maybe someone who would usually take the subway by themselves would decide to split the cost of a taxi if they are with a friend.

Hypothesis: Taxi trips with more than 1 passenger are longer on average than taxi trips with 1 passenger.

Get the Data: We will use the 2016 Green Taxi Trip Dataset from Labs 3 and 5. Follow the instructions in Lab 3 if you do not already have the dataset. You only need to download the trip data for a single day.

Test your Hypothesis: To test our hypothesis, we will compute the average length of a trip taken by single passengers and by groups of 2 or more passengers.

1) Write code to import the pandas library and read your CSV file into the dataframe taxi.

2) Compute the average length of a trip taken by solo passengers as follows:

a) Create a new variable that holds the condition we want to look for:

```
solo = taxi['Passenger_count'] == 1
```

In this case, the condition is that the Passenger_count is 1.

b) Use the condition to make a new dataframe of only the rows where this condition is true:

```
solo_trips = taxi[solo]
```

c) Compute the mean of the Trip_distance column in this new dataframe:

```
solo_mean  = solo_trips['Trip_distance'].mean()
```

Note that we could do steps (b) and (c) in one line:

```
solo_mean = taxi[solo]['Trip_distance'].mean()
```

d) Print out the mean trip distance for all trips taken by only 1 passenger:

```
print("Mean trip distance for solo passengers is",solo_mean)
```

3) Compute the average length of a trip taken by a group (2 or more) of passengers:

a) Create a new variable that holds the condition we are looking for:

```
group = taxi['Passenger_count'] >= 2
```

In this case, the condition is that the Passenger_count is greater than or equal to 2.

b-d) Repeat steps b-d from step 2 using the `group` condition variable instead of the `solo` condition variable. Change the names of the new variables and the print statement to be appropriate.

When you are done, your code should look something like:

```
group_trips = taxi[group]
group_mean  = group_trips['Trip_distance'].mean()
print("Mean trip distance for groups of passengers is",group_mean)
```

3) Compare the two means. What is the answer to our hypothesis? How sure are you about your answer? Later in the course, we will see more precise ways to check ("test") our hypothesis.

Possible conditions

We can make all kinds of conditions. Here are some examples:

Person tips by credit card (cash is not recorded):

```
tippers = taxi['Tip_amount'] > 0
```

Pick up at JFK airport:

```
jfk = taxi['RateCodeID'] == 2
```

(according to the description of the dataset columns a RateCodeID of 2 is used for the special JKF airport rate)

To use two conditions to select rows:

`taxi[tippers & jfk]` gives all rows where the person tipped AND was picked up at JFK

`taxi[tippers | jfk]` gives all rows where the person tipped OR was picked up at JFK OR both

To get all taxi trips with 2-4 passengers:

```
pass_2_or_more = taxi['Passenger_count'] >=2
pass_4_or_less = taxi['Passenger_count'] <=4
taxi[pass_2_or_more & pass_4_or_less]
```