MAT 128 Lab 3

Motivating Question: How long are green taxi trips?

Part A

Get the Data: We will use the 2016 Green Taxi Trip Data dataset, which stores all taxi trips taken by green taxis between January and June 2016:

https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb

Click on the "View Data" button. We are going to filter the data to only contain trips on Feb. 8, 2016. To do this:

- Click on the "Filter" button.
- On the menu that pops up, click on "Add a New Filter Condition".
- Choose "lpep_pickup_datetime" and change the "is" to "is between".
- Click on the box immediately below and a calendar will pop up. Highlight Feb. 8, 2016.
- Click on the box below that and another calendar will pop up. Select Feb. 9, 2016.
- Click the check box to the left of the first date.
- After a few seconds, the rows on the left should be filtered to only include pick-ups on Feb. 8, 2016.

Download the file as a CSV file.

Move the CSV file to the directory you save you programs in. Open with Excel to make sure it downloaded correctly.

Explore the Data

Open and read the data file 2016_Green_Taxi_Trip_Data.csv into the variable/dataframe taxi.

Make a plot with the x values being the `lpep_pickup_datetime` *(note: the first letter is a lower case L)* column, and the y values being the `Trip_distance` column.

The default plot is always a *line graph* (here there are so many points that you can't see the line), but in this case it is not very helpful for understanding the trip distances. Instead, we will make a *histogram*, which divides the trips into groups based on the time taken and plots the number of trips in each group.

To make a histogram of the values in the `Trip_distance` column with 10 groups (or bins), type:

```
taxi['Trip_distance'].plot.hist()
```

Is this histogram what you would expect? Why or why not?

We can increase the number of bins to understand the *distribution* (spread) of distances better. Type:

```
taxi['Trip_distance'].plot.hist(bins = 20)
taxi['Trip_distance'].plot.hist(bins = 40)
taxi['Trip_distance'].plot.hist(bins = 80)
```

This creates new histograms with 20, 40, and 80 bins, and plots them all on the same picture.

Is this what you expect? Why or why not?

To put each histogram on in its own plot, at the top of your program add:

```
import matplotlib.pyplot as plt
```

And after each plot command, add:

```
plt.show()
```

Challenge Questions

- Can you figure out how to add a title to the histogram?
- What does the histogram of the Total_amount column (total amount paid) look like?
- What does the histogram of the trip distances on a weekend look like? What about a holiday? Are they the same or different from a weekday one?

Part B

We can compute the mean, median, and mode of a column in the dataframe df using the following code:

```
df[‘column_name’].mean()
df[‘column_name’].median()
df[‘column_name’].mode()
```

So to compute and print out the mean, median, and mode of the `Trip_distance` column in our `taxi` dataframe, type:

```
mean =  taxi[‘Trip_distance’].mean()
print(“The mean trip distance is”, mean)
median = taxi[‘Trip_distance’].median()
print(“The median trip distance is”,median)
mode = taxi[‘Trip_distance’].mode()
print(“The trip distance taken the most is”,mode)
```

Finally, we will add two vertical lines to our histogram from Part A at the mean and median. We need to import the `matplotlib` library, which is what `pandas` uses to draw plots. To do this, at the top of you code add:

```
import matplotlib.pyplot as plt
```

To add a vertical line where the x coordinate is the mean, at the end of your code, type:

```
plt.axvline(mean)
```

This line is small, hard to see, and might blend in to some of the histogram colors. We can make it dashed, wider, and black by adding the options for these, so that the line of code becomes:

```
plt.axvline(mean, linestyle='dashed', linewidth=2, color =
"black")
```

Can you figure out which options correspond to which line properties?

To add a dashed, wider, black line at the median, type:

```
plt.axvline(median, linestyle='dashed', linewidth=2, color = "black")
```

Challenges:

- Can you make the line for the median red?
- What happens if you change the linestyle to `‘dotted’`?