CSCI 5015 Assignment 4
Due March 16, 11:59 PM.
50 points

Objective

Use pandas and work with multiple files

Data set

You've been given a data set from the National Highway Traffic Safety Administration (NHTSA). The data comes from their Fatality Analysis Reporting System (FARS). This particular data set represents the traffic fatalities in the United States during 2016. In each file, one line or record represents data for one accident.

For this assignment, the data is contained in a folder named **`FARS`**. Each state's data is contained in a data file named **`accident_XX.csv`**, where **`XX`** is a two-digit code for a state. The first column in the data files have the same state code/number. Use this table to match states to their code:

1	Alabama	31	Nebraska
2	Alaska	32	Nevada
4	Arizona	33	New Hampshire
5	Arkansas	34	New Jersey
6	California	35	New Mexico
8	Colorado	36	New York
9	Connecticut	37	North Carolina
10	Delaware	38	North Dakota
11	District of Columbia	39	Ohio
12	Florida	40	Oklahoma
13	Georgia	41	Oregon
15	Hawaii	42	Pennsylvania
16	Idaho	43	Puerto Rico
17	Illinois	44	Rhode Island
18	Indiana	45	South Carolina
19	Iowa	46	South Dakota
20	Kansas	47	Tennessee
21	Kentucky	48	Texas
22	Louisiana	49	Utah
23	Maine	50	Vermont
24	Maryland	52	Virgin Islands (since 2004)
25	Massachusetts	51	Virginia
26	Michigan	53	Washington
27	Minnesota	54	West Virginia
28	Mississippi	55	Wisconsin
29	Missouri	56	Wyoming
30	Montana		

The FARS folder also contains the file **`nst-est2017-alldata.csv`**, which has population estimates of the United States. This data comes from the U.S. Census Bureau. The column, **`POPESTIMATE2016`**, contains the population esitmate for each state in 2016, which is the same year as the crash data.

Coding instructions

Create a Jupyter notebook file named **assignment4.ipynb**. Unzip the file, **FARS.zip** and place the folder **FARS** in the same folder as your notebook.

Make the first cell in your notebook a Markdown cell and create a header that says "FARS Analysis" and then on the next line, have a plain paragraph that says "Written by John Nicholson" (hopefully, it is obvious that you should replace my name with your name). It should look roughly like this after you render the Markdown:

FARS Analysis

Written by John Nicholson

Use your notebook to answer the questions below. Your notebook should have both code cells and Markdown cells. The code cells should read and process the data, and the Markdown cell should explain what the associated code cell does and why you wrote the code the way you did. If you wish, add extra cells to make your main code easier to read and use.

Requirements

- No single code cell should produce all the output for more than one question. At a minimum, you will need 9 code cells, one for each question.
- If you make additional code cells, they do not have to display data. For example, you may want to write a helper function in a code cell. You still need a Markdown cell describing the function and why you wrote the function in the first place.
- All code cells need a Markdown cell explaining what the code does and why you wrote it the way you did. Explanations should be complete sentences and paragraphs, use good grammar, and be spell-checked.
- Before turning in your notebook, ensure that all code cells that produce output have their output displayed. I should not have to run your notebook in order to see the output. You should all render all Markdown cells. That is, I should not see the special Markdown characters like the **#**. I should see nicely rendered text.

Questions

For all questions, don't output codes. Output the state names, day names, etc.

1. For each state, what day of the week has the most accidents? Use the **DAY_WEEK** column. Output the day and the count. For the values output the day name, where 1 is Sunday, 2 is Monday, ... and 7 is Saturday.
2. For whole United States, i.e., all the data, what day of the week has the most accidents? Output the day and the count.
3. For each state, what hour of the day has the most accidents? Output the hour and the count.
    - A value of 99 in the HOUR means unknown.
4. For whole United States, what hour of the day has the most accidents? Output the hour and the count.
5. For each state, what is the percentage of fatal accidents involved at least one drunk driver? If the column, **DRUNK_DR**, has a 0, then no drunk drivers were involved. Any number larger than 0 indicates the number of drunk drivers involved in the accident.
6. For whole United States, what is the percentage of fatal accidents involved at least one drunk driver?
7. For whole United States, how many fatalities were caused by each type of collision? Use the **MAN_COLL** column. The values in the column are below.

0	Not Collision with Motor Vehicle in Transport
1	Front-to-Rear
2	Front-to-Front
6	Angle
7	Sideswipe – Same Direction
8	Sideswipe – Opposite Direction
9	Rear-to-Side
10	Rear-to-Rear
11	Other (End-Swipes and Others)
98	Not Reported
99	Unknown

8. For each state, what is its fatal accident rate per 10,000 people? To calculate this, count the number of accidents in a state, divide by the state's 2016 population estimate from the **nst-est2017-alldata.csv** Census data file, and then multiply by 10000. Output the states' rates in order from highest to lowest.
9. For each state, what is the rate of fatal accidents caused by drunk driving per 10,000 people? To calculate this, count the number of accidents in which a drunk driver was involved, divide by the state's 2016 population estimate from the **nst-est2017-alldata.csv** Census data file, and then multiply by 10000. Output the states' rates in order from highest to lowest.

Tips and help

Remember, we talked about files and listing directories in the previous lecture.

When you are running code, you may not see results right away. You should look to the left of the code cell for the **In [ ]:** marker. If you see a number, for example, **In [8]:**, the code in the cell is not currently running. But if you see an asterisk, for example, **In [*]:**, the code in the cell is running on the kernel, and results may not be available yet. The asterisk should change to a number when the code is done.

Remember, you can stop a running cell by choosing the "stop" button in the menu. It is the one that looks like a black square.

You can always come see me in my office or send me email. If you send me email, you should send your code as an attachment. Don't copy/paste your code into the message because that will make it harder for me to debug your code. Send your file as a **`.ipynb`** file

Submission

When your program is correct, log into D2L and locate the Dropbox for assignment 4. Upload the file

- assignment4.ipynb

into the D2L dropbox.

Contact me if you have any problems.