**ITP 449, FALL 2020**

**HOMEWORK 4**
**30 POINTS**

For each one of the following questions, write Python code in PyCharm. For each question, create a new Python file. Name it *HW4_Q1_lastname_firstname.py* Create a header in each file using *comments* to display your name and HW information. After that write your Python code.

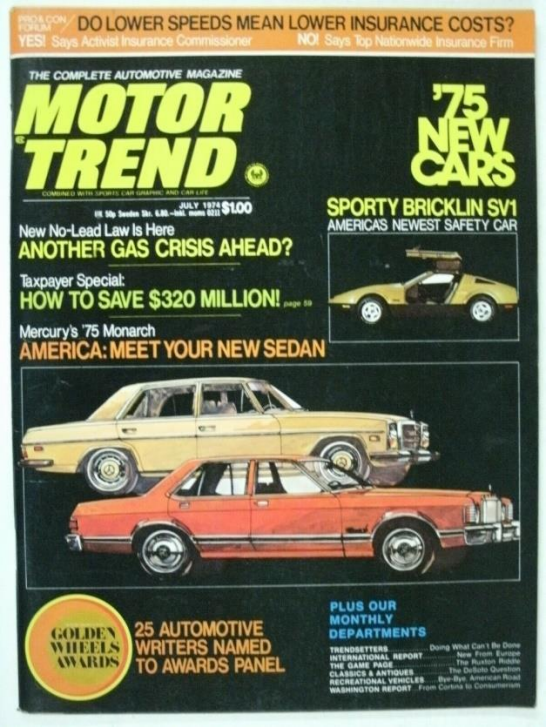
*# Tommy Trojan*
*# ITP 449 Fall 2020*
*# HW2*
*# Question 1*

- Apart from the above comments, include single line comments describing the core logic of your algorithm / code.
  As an example,

  *#Obtaining Tuples of Car Attributes (Car Name, mpg, cyl, disp, hp, gear) and generating pandas series.*
  *#Creating a DataFrame using the mtcars dataset.*

## Questions

*mtcars.csv* – Dataset extracted from the 1974 Motor Trend US magazine comprising fuel consumption and 10 aspects (attributes) of automobile design and performance for 32 automobiles is a famous problem dataset used in machine learning and data analysis. We will use this dataset to perform data abstraction, slicing, dicing and basic analysis using Pandas Series and DataFrame, as taught in the class.




```
> mtcars
                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

As you can observe in the cover page of the magazine, one of the main purpose of this data was to assist decision making on which car to purchase during the 1973 Oil Crisis which began in October 1973, when the members of the Organization of Arab Petroleum Exporting Countries proclaimed an oil embargo and the affected countries were Canada, Japan, the Netherlands, the United Kingdom and the United States and later extended to Portugal, Rhodesia and South Africa.

By the end of the embargo in March 1974, the oil price had risen nearly 400%, from US$3 per barrel to nearly $12 globally; US prices were significantly higher. The embargo caused an oil crisis, or "shock", with many short and long-term effects on global politics and the global economy. It was famously called the “First Oil Shock”.

1. Despite the “First Oil shock” crisis, Jack needs to buy a new car for his daily commute to work and he decides to perform data analysis using Pandas to determine the best car to buy. He decides to perform the following tasks, but requires help in coding the requirements and hence would like to approach ITP449 students for help. (5 points)

    As part of first task, help Jack perform the following:
    1. Read the csv file using Pandas. Store the output into a dataframe *frame*.
    2. Print the dataframe.
    3. You notice that the index is 0 ... 31. There is a column *Car Name*.

4. Set the index of the dataframe to the *Car Name*. In other words, make the column *Car Name* the index of *frame*.
5. Print *frame*.

| Car Name | mpg | cyl | disp | hp | drat | ... | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | ... | | | | | |
| Mazda RX4 | 21.0 | 6 | 160.0 | 110 | 3.90 | ... | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21.0 | 6 | 160.0 | 110 | 3.90 | ... | 17.02 | 0 | 1 | 4 | 4 |
| Datsun 710 | 22.8 | 4 | 108.0 | 93 | 3.85 | ... | 18.61 | 1 | 1 | 4 | 1 |
| Hornet 4 Drive | 21.4 | 6 | 258.0 | 110 | 3.08 | ... | 19.44 | 1 | 0 | 3 | 1 |
| Hornet Sportabout | 18.7 | 8 | 360.0 | 175 | 3.15 | ... | 17.02 | 0 | 0 | 3 | 2 |
| Valiant | 18.1 | 6 | 225.0 | 105 | 2.76 | ... | 20.22 | 1 | 0 | 3 | 1 |
| Duster 360 | 14.3 | 8 | 360.0 | 245 | 3.21 | ... | 15.84 | 0 | 0 | 3 | 4 |
| Merc 240D | 24.4 | 4 | 146.7 | 62 | 3.69 | ... | 20.00 | 1 | 0 | 4 | 2 |
| Merc 230 | 22.8 | 4 | 140.8 | 95 | 3.92 | ... | 22.90 | 1 | 0 | 4 | 2 |
| Merc 280 | 19.2 | 6 | 167.6 | 123 | 3.92 | ... | 18.30 | 1 | 0 | 4 | 4 |

**2.** Having obtained satisfactory results in Question1, Jack would now like to obtain the details of economic cars which are powerful and hence would like to perform following tasks: (10 points)

1. Create a DataFrame using attributes: 'Car Name', 'cyl' ,'gear', 'hp', 'mpg'. Make *Car Name* the index. Rename the columns to: Cylinders, Gear, Horsepower, Miles Per Gallon. Print the DataFrame

| Car Name | Cylinders | Gear | Horsepower | Miles per Gallon |
|---|---|---|---|---|
| Mazda RX4 | 6 | 4 | 110 | 21.0 |
| Mazda RX4 Wag | 6 | 4 | 110 | 21.0 |
| Datsun 710 | 4 | 4 | 93 | 22.8 |
| Hornet 4 Drive | 6 | 3 | 110 | 21.4 |
| Hornet Sportabout | 8 | 3 | 175 | 18.7 |
| Valiant | 6 | 3 | 105 | 18.1 |
| Duster 360 | 8 | 3 | 245 | 14.3 |
| Merc 240D | 4 | 4 | 62 | 24.4 |
| Merc 230 | 4 | 4 | 95 | 22.8 |
| Merc 280 | 6 | 4 | 123 | 19.2 |

2. Now, Jack would like to determine cars with 'Horsepower' more than 110, and add a separate column called 'Powerful' to the data frame. Print the dataframe.

| Car Name | Cylinders | Gear | Horsepower | Miles per Gallon | Powerful |
|---|---|---|---|---|---|
| Mazda RX4 | 6 | 4 | 110 | 21.0 | True |
| Mazda RX4 Wag | 6 | 4 | 110 | 21.0 | True |
| Datsun 710 | 4 | 4 | 93 | 22.8 | False |
| Hornet 4 Drive | 6 | 3 | 110 | 21.4 | True |
| Hornet Sportabout | 8 | 3 | 175 | 18.7 | True |
| Valiant | 6 | 3 | 105 | 18.1 | False |
| Duster 360 | 8 | 3 | 245 | 14.3 | True |
| Merc 240D | 4 | 4 | 62 | 24.4 | False |
| Merc 230 | 4 | 4 | 95 | 22.8 | False |
| Merc 280 | 6 | 4 | 123 | 19.2 | True |
| Merc 280C | 6 | 4 | 123 | 17.8 | True |
| Merc 450SE | 8 | 3 | 180 | 16.4 | True |

3. Oops, Jack accidentally deleted the column 'Horsepower' from the DataFrame even though it is the master column using which the column 'Powerful' was determined but to his surprise the DataFrame still has Powerful! Print the DataFrame with the column 'Horse Power' deleted,

| Car Name | Cylinders | Gear | Miles per Gallon | Powerful |
|---|---|---|---|---|
| Mazda RX4 | 6 | 4 | 21.0 | True |
| Mazda RX4 Wag | 6 | 4 | 21.0 | True |
| Datsun 710 | 4 | 4 | 22.8 | False |
| Hornet 4 Drive | 6 | 3 | 21.4 | True |
| Hornet Sportabout | 8 | 3 | 18.7 | True |
| Valiant | 6 | 3 | 18.1 | False |
| Duster 360 | 8 | 3 | 14.3 | True |
| Merc 240D | 4 | 4 | 24.4 | False |
| Merc 230 | 4 | 4 | 22.8 | False |
| Merc 280 | 6 | 4 | 19.2 | True |
| Merc 280C | 6 | 4 | 17.8 | True |

4. Using the original DataFrame (with 'Horsepower' column), Jack would like to list cars with 'Miles Per Gallon' greater than 25.0 and sort the cars in descending order of 'Horsepower'

| Car Name | Cylinders | Gear | Horsepower | Miles per Gallon | Powerful |
|---|---|---|---|---|---|
| Lotus Europa | 4 | 5 | 113 | 30.4 | True |
| Porsche 914-2 | 4 | 5 | 91 | 26.0 | False |
| Fiat 128 | 4 | 4 | 66 | 32.4 | False |
| Fiat X1-9 | 4 | 4 | 66 | 27.3 | False |
| Toyota Corolla | 4 | 4 | 65 | 33.9 | False |
| Honda Civic | 4 | 4 | 52 | 30.4 | False |

5. Finally, Jack decides to purchase a car that is *powerful* and has the highest *Miles Per Gallon*. Help Jack filter to that car.

| Car Name | Cylinders | Gear | Horsepower | Miles per Gallon | Powerful |
|---|---|---|---|---|---|
| Lotus Europa | 4 | 5 | 113 | 30.4 | True |

3. Exploring COVID-19 data from the Johns Hopkins Center for Systems Science and Engineering data repo on github. https://github.com/CSSEGISandData/COVID-19.
There are two datasets that may help you with this analysis.

- Go to this link - https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data. Download the latest Daily Reports dataset. E.g. 06-13-2020.csv

Field description

- **FIPS**: US only. Federal Information Processing Standards code that uniquely identifies counties within the USA.
- **Admin2**: County name. US only.
- **Province_State**: Province, state or dependency name.
- **Country_Region**: Country, region or sovereignty name. The names of locations included on the Website correspond with the official designations used by the U.S. Department of State.
- **Last Update**: MM/DD/YYYY HH:mm:ss (24 hour format, in UTC).
- **Lat** and **Long_**: Dot locations on the dashboard. All points (except for Australia) shown on the map are based on geographic centroids, and are not representative of a specific address, building or any location at a spatial scale finer than a province/state. Australian dots are located at the centroid of the largest city in each state.
- **Confirmed**: Confirmed cases include presumptive positive cases and probable cases, in accordance with CDC guidelines as of April 14.
- **Deaths**: Death totals in the US include confirmed and probable, in accordance with CDC guidelines as of April 14.

- **Recovered**: Recovered cases outside China are estimates based on local media reports, and state and local reporting when available, and therefore may be substantially lower than the true number. US state-level recovered cases are from COVID Tracking Project.
- **Active:** Active cases = total confirmed - total recovered - total deaths.
- **Incidence_Rate**: Admin2 + Province_State + Country_Region.
- **Case-Fatality Ratio (%)**: = confirmed cases per 100,000 persons.
- **US Testing Rate**: = total test results per 100,000 persons. The "total test results" is equal to "Total test results (Positive + Negative)" from COVID Tracking Project.
- **US Hospitalization Rate (%)**: = Total number hospitalized / Number confirmed cases. The "Total number hospitalized" is the "Hospitalized – Cumulative" count from COVID Tracking Project. The "hospitalization rate" and "hospitalized - Cumulative" data is only presented for those states which provide cumulative hospital data.

- Go to this link for the time series data - https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

Read the dataset(s) into pandas dataframe(s).

1. What state in the US currently has the highest number of active cases?
2. What state in the US has the highest fatality rate (deaths as a ratio of infection)?
3. What is the difference in the testing rate between the state that tests the most and the state that tests the least?
4. Plot the number of daily new cases in the US for the top 5 states with the highest confirmed cases (as of today). From March 1 – today. Use Subplot 1.
5. Plot the number of daily deaths in the US for the top 5 states with the highest confirmed cases (as of today). From March 1 – today. Use Subplot 2.