Individual Project 3
DS160-01
Introduction to Data Science
Spring 2020

Working with Pandas/Matplotlib/Seaborn (100 points)

Goal: The goal of this project is to use some of the features/functions provided by the Python Pandas library to clean up a data set and then visualize it with Pandas/Matplotlib/Seaborn.

Instructions: Create a new notebook titled **IP3_XXX**, where **XXX** are your initials. Also create a GitHub repository titled **IP3_XXX** to which you can push your code. Then complete the following:

Part 1

1. You have been given four .csv files. Each file represents a different German power company's wind turbine power generation data. Your first goal is to merge these datasets together. To do this, write a function that takes a filename (of and returns a dataframe. Your function should do the following:
    a. Convert the Date column to a date (notice, the data are in d/m/y format – take that into consider when you convert.
    b. Since none of the files have a column for company, you need to add one called Company that contains the Company name (you can extract this from .csv filename. Hint: dataframes have an insert function.
2. Execute your function four times, passing each file to it. Once complete, you will have four dataframes that you can merge.
3. Stack the dataframes on top of each to produce one large dataframe that has 98 columns, one for company, one for date and one for each timepoint.
4. You are going to visualize aggregate data first. To start, you need to create a new dataframe that is grouped by Company and the year (e.g. 50Hertz 2019, 50Hertz 2020, etc.). You can group by multiple columns by putting the column list in []. Group by the mean() function. The first few columns of your new dataframe should look like this:

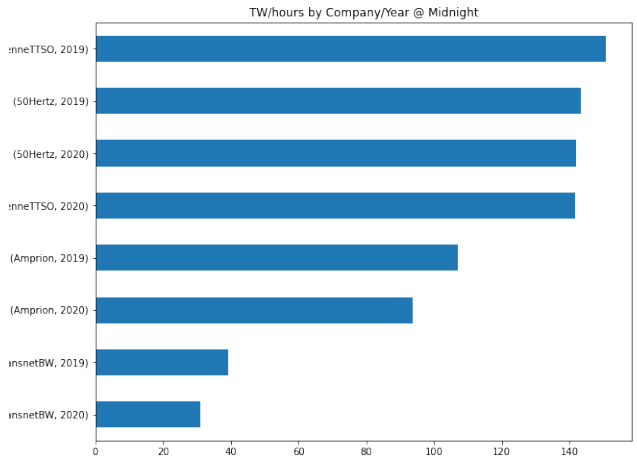
		00:15:00	00:30:00	00:45:00	01:00:00	01:15:00
Company	Date					
50Hertz	2019	142.351145	142.366412	142.000000	141.900763	141.381679
	2020	141.124060	140.936090	140.838346	140.281955	140.037594
Amprion	2019	106.679389	106.091603	105.549618	108.083969	108.282443
	2020	94.312030	94.424812	94.033835	93.657895	93.563910
TenneTTSO	2019	150.450992	150.152901	149.611679	148.946412	148.629847
	2020	141.457105	141.203120	141.206278	140.931617	140.642105
TransnetBW	2019	39.462290	39.165191	39.138321	39.369084	39.028779
	2020	30.604737	30.354023	29.871128	29.731955	29.642481

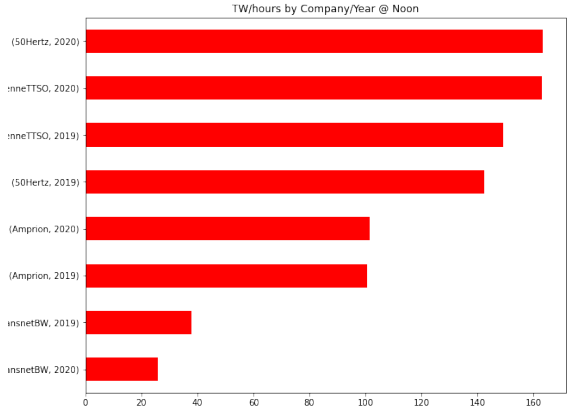
5. The first plot you want to make is horizontal bar chart that shows the average amount of power generated by each company, in each year for midnight and another bar graph that shows the same for noon. To this:
    a. Write the code to extract a dataframe for each time point (00:00:00 and 12:00:00) and put them in their own objects name midnight and noon. When you have done this you will have two Pandas Series that look like below.
    b. Write the code to create a horizontal bar chart for each dataset that look the those below. **Note: These are two separate charts, they are not subplots.**
    c. Save each chart separately.

Datasets after extraction

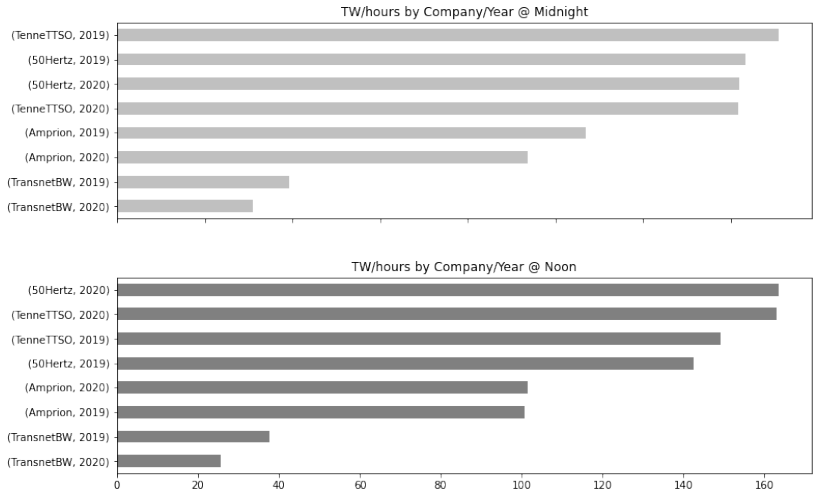
```
Company     Date
50Hertz     2019    143.320611
            2020    141.894737
Amprion     2019    106.992366
            2020     93.751880
TenneTTSO   2019    150.866183
            2020    141.603835
TransnetBW  2019     39.293740
            2020     30.899023
Name: 00:00:00, dtype: float64
```

```
Company     Date
50Hertz     2019    142.473282
            2020    163.398496
Amprion     2019    100.572519
            2020    101.357143
TenneTTSO   2019    149.229466
            2020    162.944436
TransnetBW  2019     37.772366
            2020     25.731053
Name: 12:00:00, dtype: float64
```

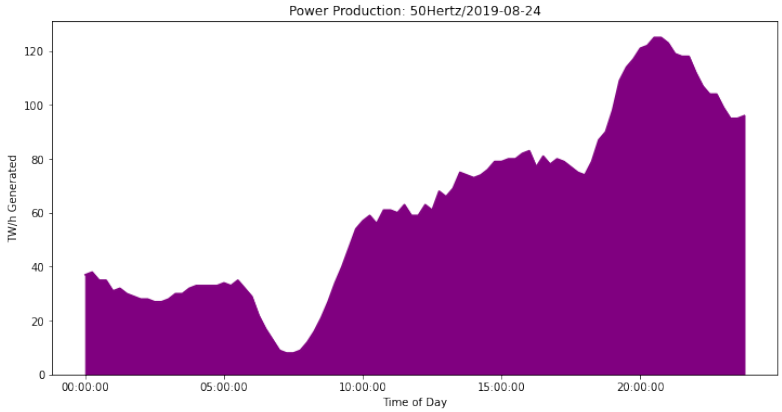




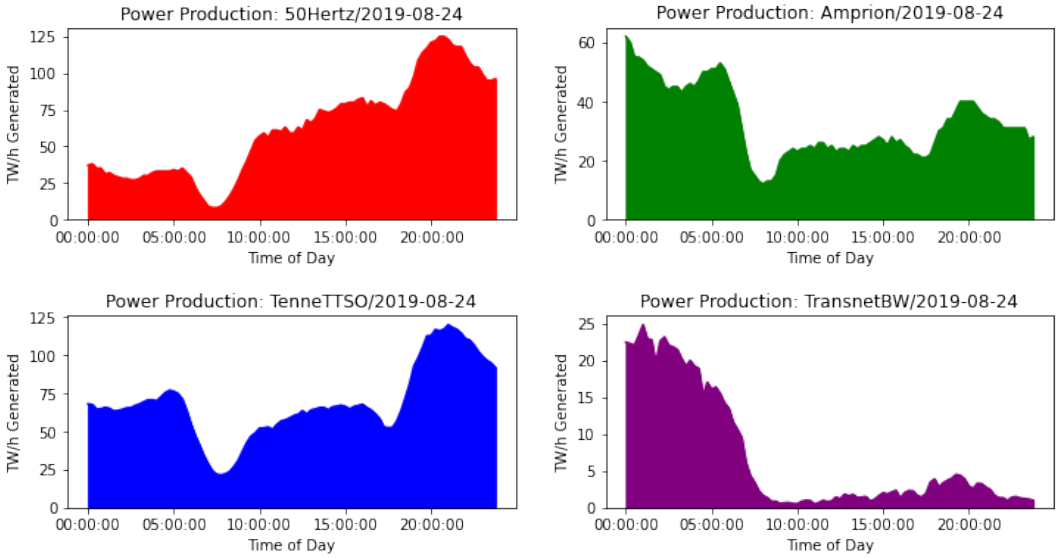
6. You have been asked to combine the two bar graphs into a single graph that shares the x-axis. And, it will be published in newspaper in grayscale, so the top graph should be silver and the bottom should be gray. Save the file as 'combined_bar_chart.png'. The chart should appear as below:



7. You have been asked to write a function that accepts the ungrouped dataframe, the company name, date, and a plotting color that generates an area graph for that particular company/day. Use this function signature: `daily_view(dataframe, company, date, color)`. Your function should produce the following graph:



BONUS (10 points): Generate a unified graph that contains four area charts, one for each company for the same day. See example below. **Note: I won't help with this since it is a bonus.**



Part 2

In class we looked at several data sets provided by Seaborn including fmri, flights, and mpg. Using one or more of the other data sets besides (*fmri, flights, mpg and penguin*) generate examples of the following plots:

1. Line Chart
2. Box/Whisker
3. Histograms and Kernel Density
4. Heatmap
5. Seaborn Pair Plot

All of your code should go in one Jupyter notebook with appropriate Markdown headers/comments as we discussed in class. Add, commit and push your code to GitHub. **You only need to push the original data files and your notebook to Github. I don't need the figure saves because I will generate them when I run your notebook.**

Project Submission: Upload a link to your GitHub repository for the project in the area provided in Moodle by the deadline specified.