Data Analytics

Course: 18-787

Recitation 1

Spring 2023

Carnegie Mellon University

Assignment 1 overview

This assignment will help you to get familiar with:

- Cleaning raw data: removing unnecessary rows/columns, linear interpolation, combining datasets, etc.
- Feature selection techniques
- Different data analytics techniques such as curve fitting, etc.

Question 1

Procedures:

- Download the CSV file posted on canvas which contains the historical daily weather data for France in 2017
- Save it as CSV file and load it into MATLAB/Jupyter Notebook/Colab
- Fill any missing gaps using linear interpolation

Question 1 (cont'd)

- Missing values Hint: use dataframe.info() to get the summary of your dataframe

```
Data columns (total 21 columns):
Date                                365 non-null object
high Temp. (°C)                     365 non-null int64
avg Temp. (°C)                      365 non-null int64
low Temp. (°C)                      365 non-null int64
high Dew Point (°C)                 365 non-null int64
avg Dew Point (°C)                  365 non-null int64
low Dew Point (°C)                  365 non-null int64
high Humidity (%)                   365 non-null int64
avg Humidity (%)                    365 non-null int64
low Humidity (%)                    365 non-null int64
high Sea Level Press. (hPa)         365 non-null int64
avg Sea Level Press. (hPa)          365 non-null int64
low Sea Level Press. (hPa)          365 non-null int64
high Visibility (km)                365 non-null object
avg Visibility (km)                 365 non-null object
low Visibility (km)                 365 non-null object
high Wind (km/h)                    365 non-null int64
avg Wind (km/h)                     365 non-null int64
high Gust Wind (km/h)               365 non-null object
sum Precip. (mm)                    365 non-null float64
Events                              226 non-null object
dtypes: float64(1), int64(14), object(6)
```

Question 1 (cont'd)

- Be careful of the (-, ?)in your dataset.

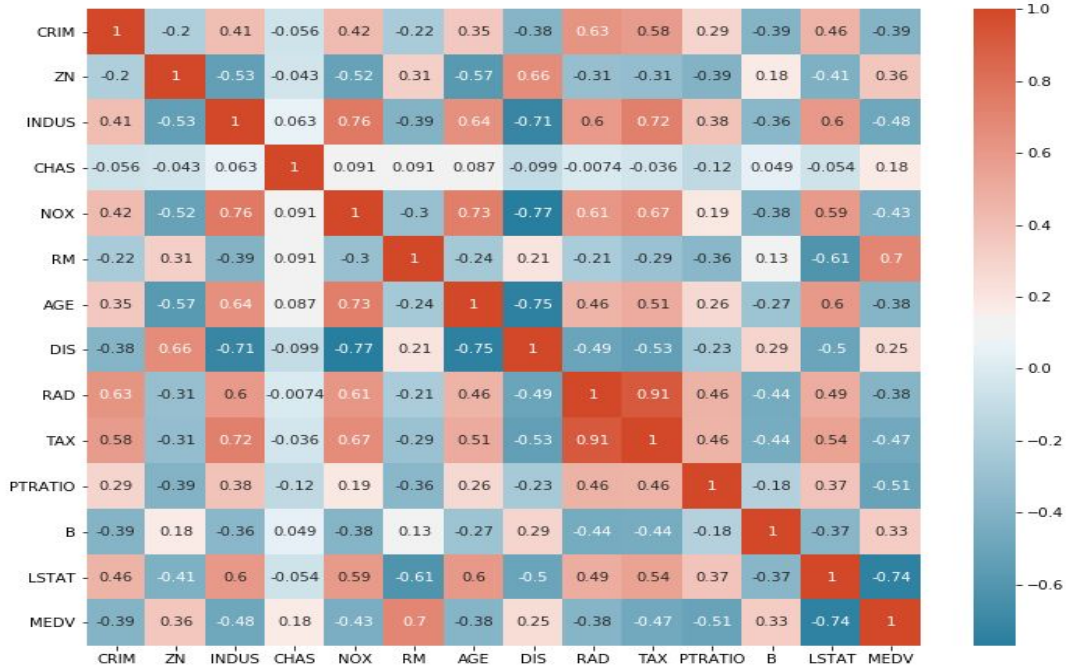
```
Data columns (total 21 columns):
Date                                365 non-null object
high Temp. (°C)                     365 non-null int64
avg Temp. (°C)                      365 non-null int64
low Temp. (°C)                      365 non-null int64
high Dew Point (°C)                 365 non-null int64
avg Dew Point (°C)                  365 non-null int64
low Dew Point (°C)                  365 non-null int64
high Humidity (%)                   365 non-null int64
avg Humidity (%)                    365 non-null int64
low Humidity (%)                    365 non-null int64
high Sea Level Press. (hPa)         365 non-null int64
avg Sea Level Press. (hPa)          365 non-null int64
low Sea Level Press. (hPa)          365 non-null int64
high Visibility (km)                363 non-null object
avg Visibility (km)                 363 non-null object
low Visibility (km)                 363 non-null object
high Wind (km/h)                    365 non-null int64
avg Wind (km/h)                     365 non-null int64
high Gust Wind (km/h)               69 non-null object
sum Precip. (mm)                    365 non-null float64
Events                              226 non-null object
```

Question 2

Calculating and plotting the correlation matrix.

A correlation matrix is a table showing correlation coefficients between variables. It is used to investigate the dependence between multiple variables at the same time.

Example of correlation matrix as a heat-map.



Source:
https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/

Question 3

Refer to Question 1

Question 4

Procedures:

- Extract the average/mean temperature data by Indexing.
- Synchronize Weather and Energy consumption Timestamps.
- Extract weather and energy consumption dates.

Create time series for both weather and energy - using the time series function (combines dates and data)

Example: You can use synchronize function for MATLAB and join, merge, etc.,… for python.

- Make a scatter plot of energy consumption against mean temperature.

Question 5

Fitting a quadratic model to the energy versus temperature.
A quadratic model is of the form **y = a2x^2+a1x+a0** where a2, a1 and a0 are coefficients that minimize the square error.

Sample code:

Python

```
import numpy as np
np.polyfit(x, y, 2)
```

MATLAB

```
model = polyfit(x,y,2);
```

Question 6

- The optimal minimum temperature corresponds to the lowest energy consumption.

Question 7

- Use multivariate linear regression
- Instead of passing one variable, you use multiple variables.

- MATLAB (Hint: you can use stepwiselm,stepwisefit,…)
- Rsquared

In python

- https://pypi.org/project/stepwise-regression
- https://github.com/AakkashVijayakumar/stepwise-regression/blob/master/stepwise_regression/step_reg.py

Question 8

- Feature variables: X, X^2

Same steps as Question 7

Which variables are selected?

Compare the old and new R^2 values

Question 9

- Feature variables: X, X^2, dummy variables of weekdays

Same steps as Question 7

Hint:
pandas: pd.get_dummies()
Matlab:

- dummyvar function or
- you can deal with it logically

Question 10

- Freestyle

What to submit?

Submission Files

Submission Instructions

- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb]**
- Assignment report(.pdf) - remember to name the file as instructed
    - Indicate the libraries you have used in your code at the beginning of the report (After the title page)
- Data files (as given)

Submission process:

1. Put source code file and data files in a single folder
2. Name of the folder should be the same as your andrewID
3. Zip this folder and attach the zipped file on the assignment submission page (CANVAS)
4. After attaching the zipped file, click on "Add Another File" from the assignment submission page and attach your report
5. Submit your assignment

Make sure your code run well before submitting!!!

Thank you!