CMPE 188 EDA homework

Instructor: Jahan Ghofraniha

1. In this assignment you will perform exploratory data analysis on the Boston dataset.
2. The dataset has been provided on Canvas.
3. Load the dataset into a Pandas dataframe.
4. Clean the data (if needed).
5. The output in this data set is Medv (median price). The rest of the columns are considered input. Separate the data into an input and output dataframes/Series. You can ignore/eliminate categorical data.
6. Perform normalization and standardization on the data. We normally normalize and standardize the input frame and keep the output intact.
7. Put the new normalized input data frame and the output into a new data frame called data_norm. Do the same for standardized data. Call the new data frame for standardized data as data_stand.
8. Perform basic EDA, i.e. descriptive stats, plot the histograms and match/verify with descriptive stats.
9. Continue with correlation analysis (calculate correlation and plot correlation heatmap) and scatter plots.
10. A preliminary Python code has been provided on Canvas in file: EDA_hw_boston.py. Use this file as a starting point and fill in the blanks with your code.
11. Identify the high correlation columns from the headmap and compare the results from those of the scatter plots. Do the results match? Explain.
12. Your homework submission includes two parts: a .py file and a pdf/word file with all the plots and explanations/comments/interpretations included. You can alternatively submit a jupyter notebook file plus the pdf version of the notebook.
13. You can collaborate on the homework assignment (2 people only/team) but you will submit individually on Canvas for grading. The assignment should include the name of the person you are collaborating with.