2.s986
ALBAN COBI

DUE: 17 Sept, 2020

PSET #2

PROBLEM 1 DATA CURATION WITH PYTHON

☑ Download Concrete Compressive Strength data set

☑ Build Jupyter Notebook addressing the following:

☑ a) Create block of code to import relevant libraries/packages to import and manage the data

pandas package – for working with data sets
matplotlib package – for plotting/visualizing data
numpy — " — – for scientific computing/numerical computing
seaborn — " — – for making statistical graphics

☑ b) Create a block of code to load dataset and query number of rows & columns

use len(data) , variable name to check rows
use len(data.columns) to check columns

☑ c) How many columns are input variables? 8
How many columns are output variables? 1

I got this from the Concrete_Readme.txt file

☑ d) Create block of code to eliminate "Fine Aggregate" (Component 7) from data set.

First name the columns by creating a new list variable
Then use the del variable[7] to delete column 7

☑ Normalize input & output variables

mean = 0

Standard deviation = 1

I used the preprocessing function from sklearn library.

Syntax: data_normalized = preprocessing.scale(data)

↑ function to normalize $\mu = 0$, $\sigma = 1$

↑ original data

PROBLEM 2 | DATA CURATION ON DATA I CHOOSE

I decided to find a dataset of my own, and I found a dataset of ~500 rows and 14 features that contains data on the house prices in Boston. The data was available from StatLib archive.

I implemented the code to curate that data in Jupyter Notebooks.

Null Variable Rows

Looking at my data, there are no "NaN" rows but the 10th column titled:

"pupil to teacher ratio"

seems to contain some rows with a backslash. I assume this is missing data.

There are 49 rows with this missing data. Based on the data I would either eliminate these rows or eliminate the entire column and see what my analysis outputs. I can't see a way to "replace" these rows since there is no information on the town where each house is located.

I tried to look up online how to detect a cell in a DataFrame structure that does not have a numerical value but I kept getting errors with the commands I tried. So I wrote down the pseudocode here:

1. scan dataframe and replace '\' with NaN
2. Drop null values with dropna function

Now the dataset should have all rows with that missing value deleted.

I think I got it to work now...

see code.