Laboratory work 1

Please save you lab wors as follows befor uploading on ELSE: Name_Surname_AED_Lab1

The dataset for this lab work contains data on prices of homes in Boston and possible predictor variables. Please see the description of variables below. You can find the tasks for the lab work under the variable descriptions.

crim
per capita crime rate by town.
chas
Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
rm
average number of rooms per dwelling.
age
proportion of owner-occupied units built prior to 1940.
dis
weighted mean of distances to five Boston employment centres.
tax
full-value property-tax rate per \$10,000.
lstat
lower status of the population (percent).
medv
median value of owner-occupied homes in \$1000s.

1. Load and print the first 20 observations of the dataset. Report if you see any unusual values.
2. Discuss what effects you would expect to see on the med home values (medv) for each variable.
3. Compile the table with summary statistics (min, max, med, etc). Add the measure of variability (var, skew) to this table. Comment on the table and report briefly if you see anything unusual in the statistics of your variables.
4. Check the types of your data. Change the types as appropriate (if any categorical variable present change its type to category).
5. Substitute the NaN values with appropriate measures of central tendency (mean, median, or mode – for the categorical variable – if you can't change to mode, then check what is the most frequent value of that variable (you can use value_counts) and change it to the most frequent value). You may want to do this procedure for all variables to make sure that you did not miss a variable because you were not able to see that the variable contains NaNs while inspecting the table.
6. Produce the histograms of all variables (except Chas) and comment on their distributions (for each variable separately). Notice any outliers, or fat tails (like in the case of tax). Put this into context knowing what your variables mean.
   So don't just say that tax has a fat right tail, but something like "it appears that our dataset has the majority of houses with relatively low tax rates, and a set of properties that are highly taxed."
7. Create box plots for all variables where you split by the Chas variable (make sure to adjust the number of axes). Comment shortly on each box plot separately noticing if the distributions are located higher for properties on the river versus those not on the river. What does it mean when you put it into context with what your variables mean? Can you make a guess if these houses are preferred by Bostoners? Are these high end residences? How do you explain that they seem to be valued higher when it

comes to price, but are in the same time on the older side when it comes to building's age?

8. Create the scatter plots for each pair of variables. Comment on how the variables correlate with the **medv** variable. Do these correlations make sense? Explain why? Do they confirm what you wrote in the beginning of this work where you hypothesized how these variables will affect the houses' values? Notice any other correlations between the pairs or variables if obvious from the scatter plots.
9. Create the heatmap and add the correlation coefficients to it. What are the 5 strongest correlations that you see? Comment on their sign (only if not done so previously).
10. Based on your hypotheses in (2) choose the variables that you believe could have an impact on the median value of a home. Run a regression with these variables as predictors and medv as the dependent variable. Discuss your results: interpret the coefficients, discuss if the signs of their effects are as you expected, discuss their p-values and the R-squared of the regression.