- Homework 1 Spec sheet -

Please load and use the "housingUnits.csv" data file. This dataset contains information on the housing situation in over 20,000 housing "blocks" in California, from the US Census.

Each row in the csv file represents the information from a housing block.
Columns represent (in order):

1) Median age of the houses in the block (in years)
2) Total number of rooms in a given block
3) Number of bedrooms in a given block
4) Population in the block
5) Number of households in the block
6) Median household income in the block (in thousands of dollars)
7) Proximity to the ocean (rated on a scale from 0 = closest to 4 = farthest)
8) Median house value in the block (in dollars)

We/you will want to use the first seven variables as predictors and the last variable – the house value – as the outcome variable, when answering the questions.

The granularity of the information ("block" being the unit of analysis) is not ideal, as it will introduce some noise/blur, as not all houses are representative of the average house in the block, but we'll make do. Just recall that the unit of analysis is "housing block", not "house". The good news is that this should not matter here, as we will try to predict the *median* house value in a block, trusting that the median represents the houses in a block well. Moreover, this dataset has been preprocessed and cleaned, so there is no missing data. That is also good news, as this will be something we have to deal with in later assignments.

Mission command approach: As per §4.5 of the Sittyba, we will tell you what to do ("answer these questions"), not how to do it. That is up to you. However, we want you to:

a) Do the homework yourself. Do not copy answers from someone else.
b) Restrict your methods (for now) to what was covered in the lecture/lab (in other words, linear regression methods)
c) Include the following elements in your answer (so we can grade consistently):

Each answer should contain these elements:

1) A brief statement (~paragraph) of what was done to answer the question (narratively explaining what you did in code to answer the question, at a high level).
2) A brief statement (~paragraph) as to why this was done (why the question was answered in this way, not by doing something else. Some kind of rationale as to why you did x and not y or z to answer the question – why is what you did a suitable approach?).
3) A brief statement (~paragraph) as to what was found. This should be as objective and specific as possible – just the results/facts. Do make sure to include numbers and a figure (=a graph or plot) in your statement, to substantiate and illustrate it, respectively.
4) A brief statement (~paragraph) as to what you think the findings mean. This is your interpretation of your findings and should answer the original question.

Please answer the following questions in your report:

1. Why is it a good idea to standardize/normalize the predictor variables 2 and 3 and why are predictor variables 4 and 5 probably not very useful by themselves to predict median house values in a block?

2. To meaningfully use predictor variables 2 (number of rooms) and 3 (number of bedrooms), you will need to standardize/normalize them. Using the data, is it better to normalize them by population (4) or number of households (5)?

3. Which of the seven variables is most *and* least predictive of housing value, from a simple linear regression perspective? [Hints: a) Make sure to use the standardized/normalized variables from 2. above; b) Make sure to inspect the scatter plots and comment on a potential issue – would the best predictor be even more predictive if not for an unfortunate limitation of the data?]

4. Putting all predictors together in a multiple regression model – how well do these predictors taken together predict housing value? How does this full model compare to the model that just has the single best predictor from 3.?

5. Considering the relationship between the (standardized) variables 2 and 3, is there potentially a concern regarding collinearity? Is there a similar concern regarding variables 4 and 5, if you were to include them in the model?

Extra credit:

a) Does any of the variables (predictor or outcome) follow a distribution that can reasonably be described as a normal distribution?
b) Examine the distribution of the outcome variable. Are there any characteristics of this distribution that might limit the validity of the conclusions when answering the questions above? If so, please comment on this characteristic.

Hints / Suggestions / Clarifications

*****"Normalization" usually refers to z-scoring (subtracting the mean and dividing by the standard deviation). This makes a lot of sense, as it yields unit-less data, but is not what we are asking here. We will use it later. Here is what I mean when referring to standardization in this assignment:

Logic:

1) The blocks are not of equal size. Some are big and some are small, with everything in between. We don't know which is which. This is simply a census designation. And we got the data from the census. This is all we have.
2) Therefore, the raw total number of rooms in a block is not going to be a very meaningful predictor of median house price. [I can't give away too much here as to how you can tell, because that is the answer to question 1, but think of the total number of gummy bears consumed in a state as an indicator of how much the residents like gummy bears. Unless you take the state population (by dividing the number of gummy bears eaten by the number of people) into account, the total number is not very meaningful. Think of the equation for the arithmetic mean. The sum of the values in a dataset is not independent of the number of datapoints we have, which is why we divide by the number of datapoints in the denominator, to get a more useful number].
3) Ideally, we would like to divide the number of rooms in block by the number of houses in a block, to get a sense of how big the average house is (rooms per house), to get a sense of the typical layout of a house in the block, beyond square footage.
4) However, we don't have that information. The census did not record the number of houses.
5) So we have to do the best we can. [This is starting to become a theme in this class]
6) In other words, either divide by the population in the block or by the number of households in the block. One of these is a better choice than the other. Do what makes the most sense (for question 2) and proceed from there.
7) In other words, we are trying to standardize. We do this in order to get a sense of how the average house in the block is laid out (as a predictor of price), but to do that, we have to standardize it first, as the blocks are of unequal size. [In an ideal world, the blocks would all be of the same size, or we would have information about the number of houses in the blocks. Unfortunately, and this is quite realistic – in the real world, we usually don't have that information (or not in that form), so we simply have to do the best we can, and provide a reason as to why it is a reasonable thing to do. That's usually most of what the job is, given that we are often working with pre-existing datasets. As someone might have recorded them for a different purpose (in this case the census), they might not be what we need. But they are what we have. So you have to build a bridge between what we need and what we have.]

*****You should be able to answer all of the questions in this assignment with the information from the first week only (in other words, linear regression. You should not need regularized regression, polynomial regression, nor logistic regression).

*****When asking to "make a case for something", that usually means a number. Either the number being sufficiently small (or close to zero) or large. The only numbers yielded by our algorithms we have encountered in the context of linear regression are the correlation r, the explained variance R^2 and the betas. In terms of supporting illustrations, you should be able to make do with histograms and scatter plots.