Suvir Wadhwa

Fundamentals of Machine Learning: Homework 1

Question 1

To show why it is a good idea to standardize the predictor variables 2 and 3, the variance of the total rooms and bedrooms was calculated before and after standardization. Similarly, for the second part of the question, to show why the population and households are not very useful predictors, we plotted a scatter plot for each of them and calculated their correlation coefficients.

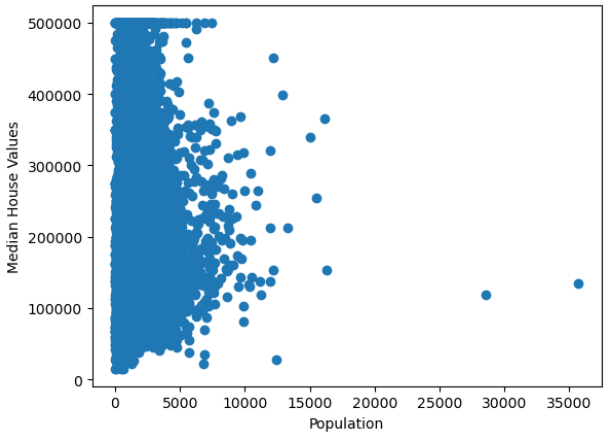
The first part was done in the aforementioned way to show the impact standardization would have on the variance. Standardizing the predictor variables 2 and 3 helped to ensure that all the variables are on the same scale and have a mean of 0 and a standard deviation of 1. This is important because linear regression models assume that the variables are on the same scale and have similar distributions. In the second part, finding the correlation coefficient and visually representing it helped us visualize and see the relationship between each of the variables 4 and 5 and the output.

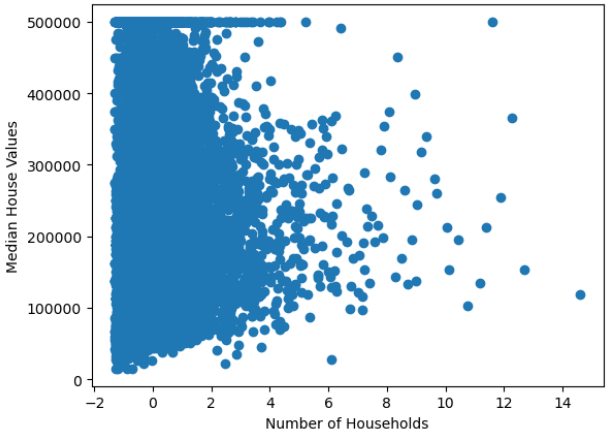
For the first part, the variance before and after is as follows:

Variance before normalization:
total_rooms = 4.759215e+06
total_bedrooms = 1.531810e+05
Variance after normalization:
total_rooms = 1.0
total_bedrooms = 1.0

Suvir Wadhwa

For the second part, the correlation matrix and scatter plots are:





	population	households	median_house_value
population	1.000000	0.907222	−0.024650
households	0.907222	1.000000	0.065843
median_house_value	−0.024650	0.065843	1.000000

Looking at the findings for part one, you can see that standardization made the variance equal one, this ensures that one variable won't dominate the other in terms of the outcome variable. Similarly, seeing the correlation coefficients to house value for part 2, it can be seen that the values are -0.02 and 0.06 for population and households, respectively. These are weak coefficient values, showing that there isn't much use.

Question 2

To find out whether we should standardize columns 2 and 3 with either columns 4 or 5, we found the correlation coefficients with respect to both: population and households. We then find the greater absolute correlation and that will be the better column. We then standardize the columns 2 and 3 with the resulting column we found. Finally, we can create a plot to show our findings.

The code gave us a result saying that we should normalize with households as we got a coefficient value of 0.84 which was higher.

The findings show the relationship in terms of coefficients between bedrooms/rooms and households. The better column to normalize by is 5 i.e. Households.

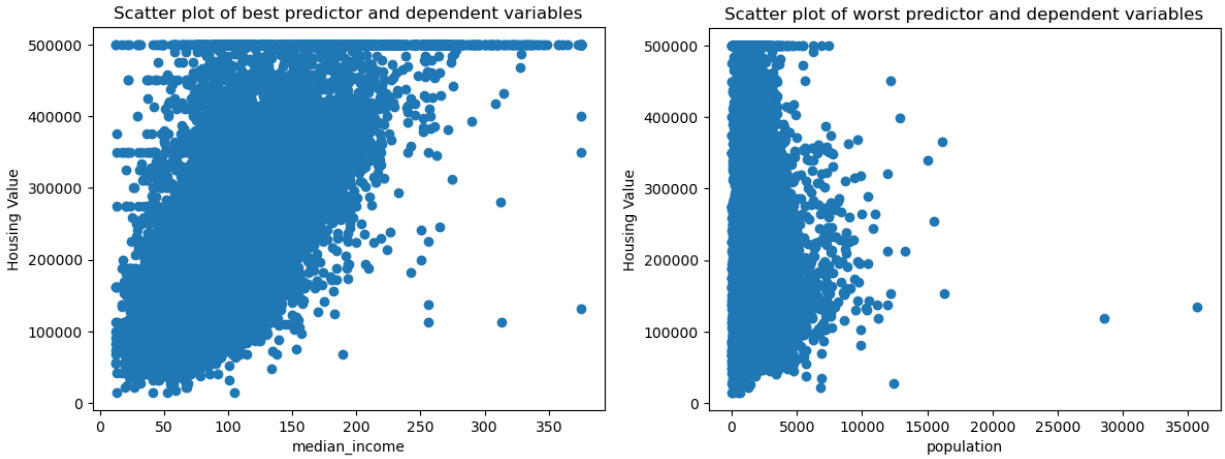
Question 3

The code for this question takes a dataset and performs linear regression on each predictor variable to determine which one has the highest R-squared value. It then plots a scatter plot of the best predictor against the output variable and identifies the predictor variable with the lowest R-squared value and plots a scatter plot for it. We then find the columns with the highest and lowest R^2 values and plots scatter plots for them.

The above approach was used to identify the best predictor variable by fitting a linear regression model for each predictor variable and calculating its R-squared value. The R-squared value measures the proportion of variation in the dependent variable (in this case, median house value)

that is explained by the independent variable (each predictor variable). The higher the R-squared value, the better the predictor variable.

From the code, the best predictor was the median income, and the worst predictor was the population. The scatter plots for both are below:



The findings show that the best predictor is the median income and the worst is the population. Inspecting the scatter plot, it can be seen that there is a lot of data at the top. This could be data for the housing value that is above the threshold of our data range. Hence, restricting our findings partially. If we could extend the range or get credible values for this data, then we would be able to predict better.

Question 4

The code for this question performs linear regression modeling to predict house values using several predictors. The predictors are split into training and test sets using the train_test_split() function. The code then fits a multiple linear regression model using all the predictors and calculates the mean squared error and R-squared value of the model on the test data. The code then fits a simple linear regression model using the single best predictor and calculates the mean

squared error and R-squared value of the model on the test data. Finally, the code compares the mean squared errors of the two models and outputs a message indicating which model has a better fit to the data. The multiple regression model is preferred if it has a lower mean squared error, indicating a better fit.

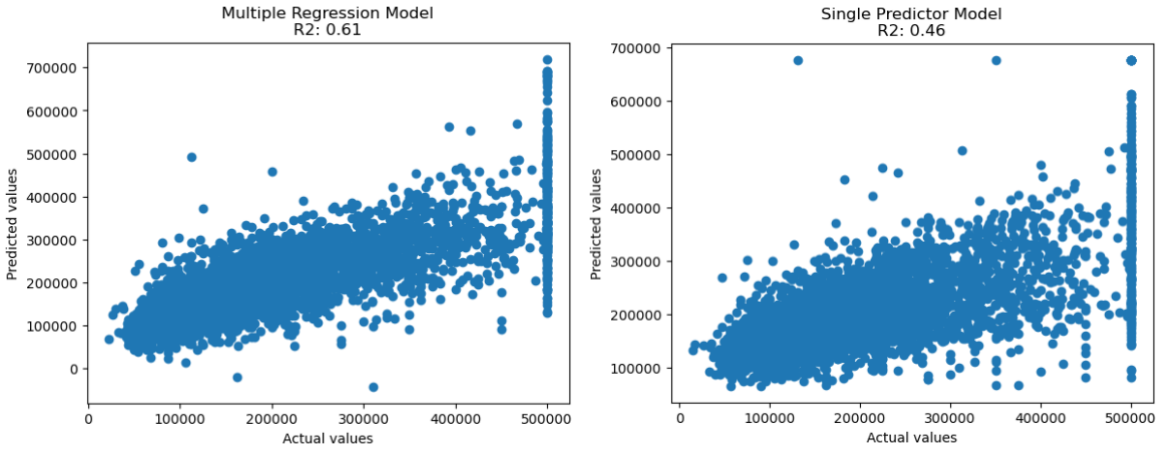
This approach was used because it allows us to compare the predictive performance of the full model (multiple regression) against that of the best single predictor model. By fitting both models and comparing their respective MSE and R2 values, we can determine which model provides a better fit to the data. The multiple regression model may be better able to capture the combined effects of all the predictors on the outcome variable, while the single predictor model may be better able to capture the effect of a single predictor on the outcome variable.

The code produced the following results:

Multiple Regression Model
MSE: 5354465181.29
R2: 0.61

Single Predictor Model
MSE: 7109257106.15
R2: 0.46
The multiple regression model has a lower MSE and is a better fit for the data.

Suvir Wadhwa



The results above show us that the multiple regression gives us a R^2 value of 0.61. It indicates 61% of the variance in the data. In comparison to the single predictor model which has an R^2 value of 0.46, the multiple regression is a better predictor of housing value. However, it is important to always consider collinearity.

Question 5

The code first calculates the Pearson correlation coefficient between standardized variables 2 and 3, and between standardized variables 4 and 5, which helps determine if there is potentially a concern regarding collinearity between variables in the dataset. It then creates scatterplots of two pairs of variables, total_rooms and total_bedrooms, and population and households, to visually examine their relationship. The code again calculates the Pearson correlation coefficient between the variables in these two pairs.

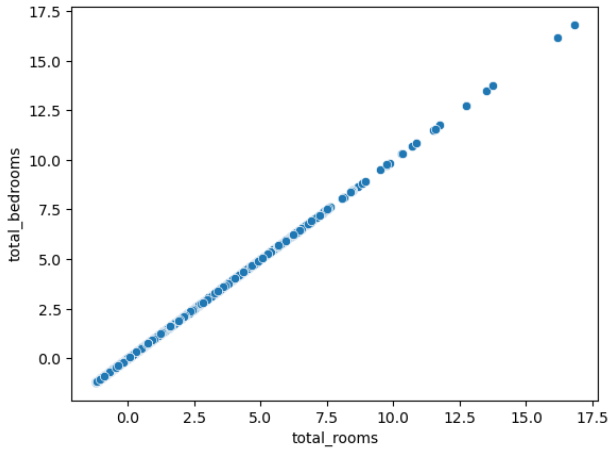
The Pearson correlation coefficient is a measure of the strength and direction of the linear relationship between two variables. By calculating the Pearson correlation coefficient between variables 2 and 3, as well as variables 4 and 5, we can determine if there is a strong linear relationship between these variables, which may indicate the presence of collinearity.

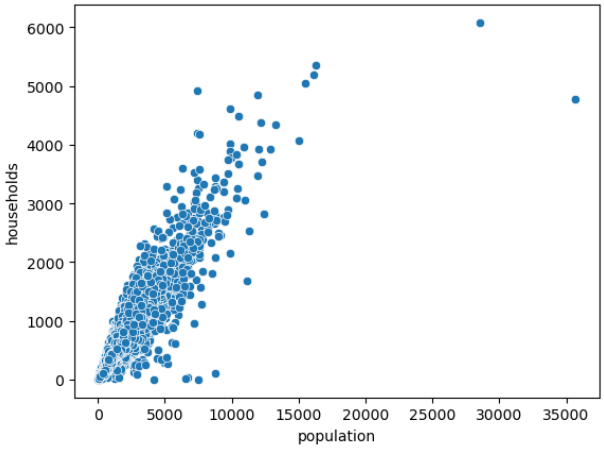
Additionally, by examining scatterplots of the variables, we can visually assess the presence of any non-linear relationships that may also indicate collinearity. This approach helps to identify any potential issues with collinearity in the data, which is important for building accurate regression models.

The values and metrics we got from the approach:

Pearson correlation coefficient between standardized variables 2 and 3: 0.999

Pearson correlation coefficient between standardized variables 4 and 5: 0.907



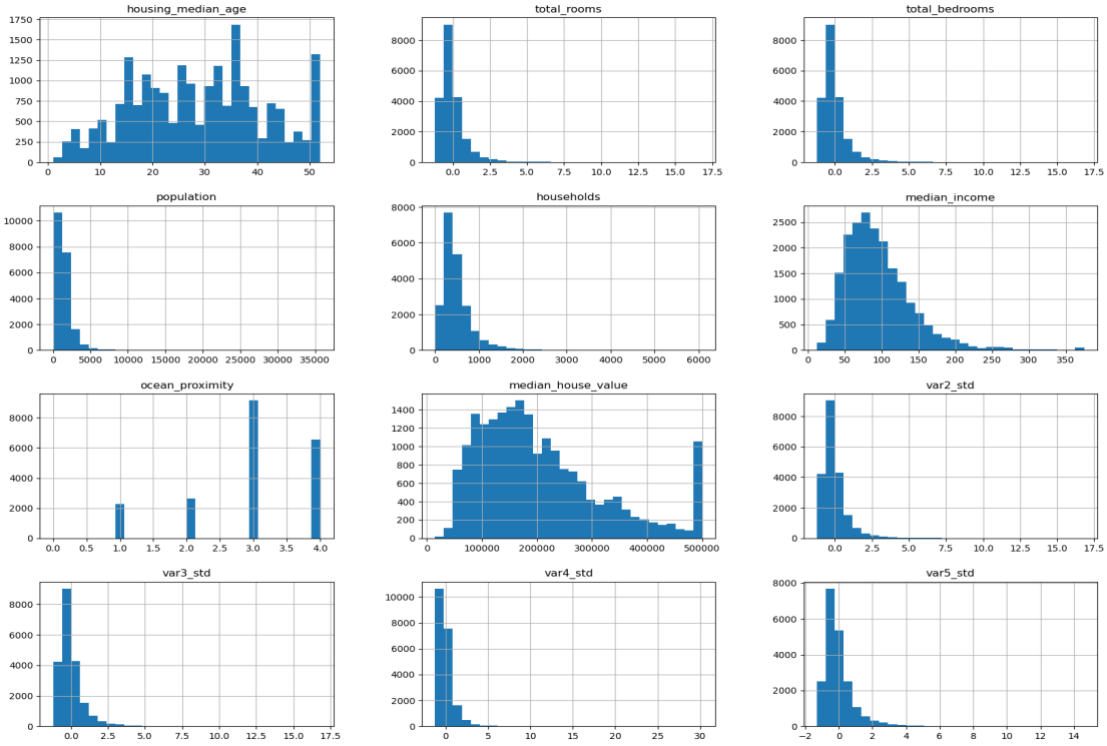


Looking at the data, there is concern about collinearity as both of the correlation coefficients are above 0.99. This is a measure of high correlation giving us a major issue of collinearity. This problem is for both columns 2 & 3, and columns 4 & 5.

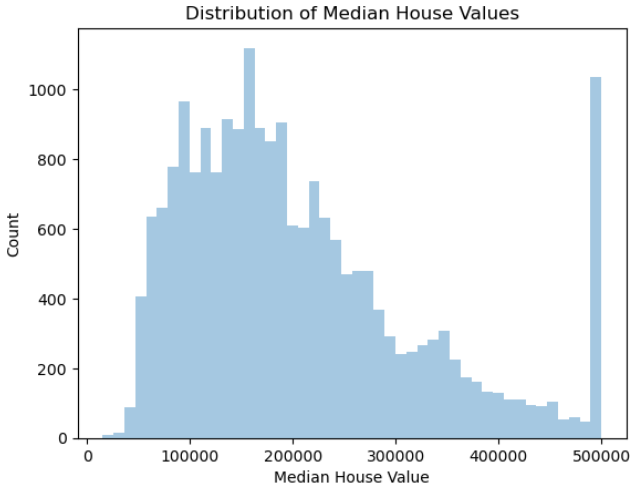
Suvir Wadhwa

Extra Credit

a) The plots for all the regression models:



None of them are normal as they are all skewed to some degree



b) Skewness of Median House Values: 0.9777632739098341

The distribution is significantly skewed, which might limit the validity of the conclusion.