Qingyu Zhang / Andy

N-number: N19903322

Prof. Pascal Wallisch

Fundamentals of Machine Learning

HW1 Report: Houses

Question 1:

1) What was done:

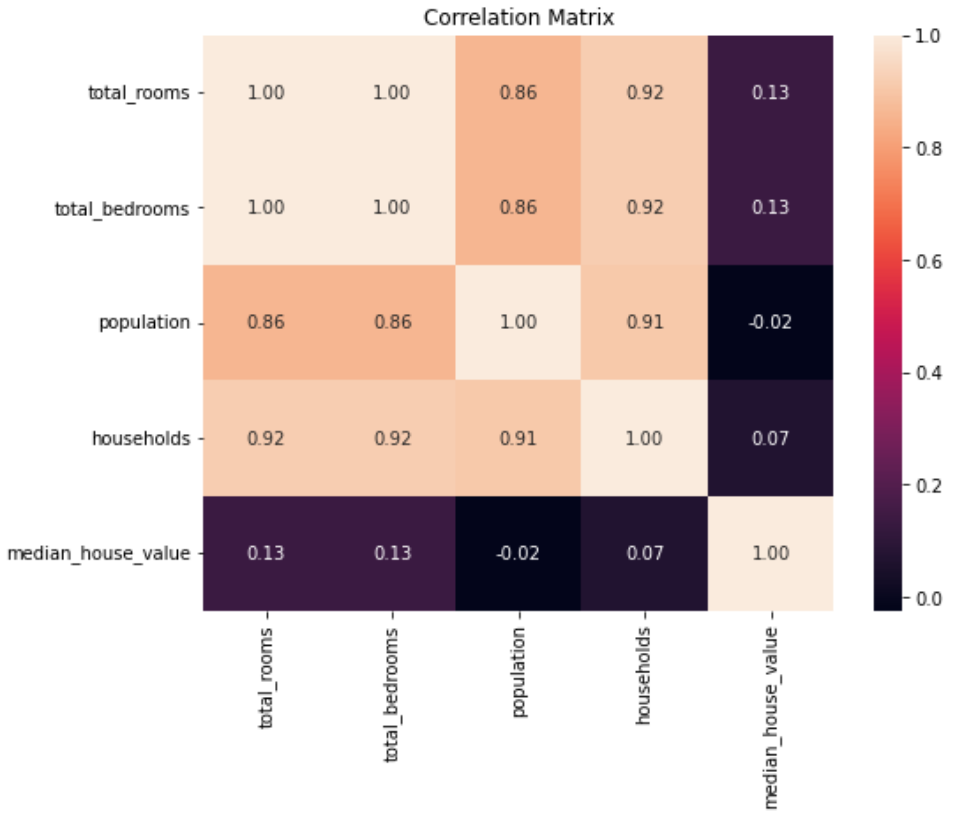
I extracted the following features and used them to make a correlation heatmap to see the correlation strength between each other: 'total_rooms', 'total_bedrooms', 'population', 'households', 'median_house_value'

2) Why this was done:

To answer the 2 problems in Question 1, I don't think it's enough to just rely on logic and theoretical analysis. Hints mention that the raw metrics such as the number of rooms in a block are misleading because each block is not the same size. Calculating and checking the correlation between the four features mentioned above and the price of the house can provide data evidence for our analysis. The strength of correlation visualizes whether a feature is a good / useful linear predictor of house values.

3) Findings:

The correlation coefficients of “number of rooms”, “number of bedrooms”, “population”, and “number of households” with “median house value” are 0.13, 0.13, -0.02 and 0.07 respectively. Correlation heatmap is shown below:



4) Interpretation:

As mentioned in Hints, blocks are of different size. The intent of using the number of rooms and number of bedrooms data may be to determine the representative house type (how many rooms are in a house) for the block. If the house types in the block are generally larger, then it stands to reason that home prices will be more expensive. But the raw data doesn't reflect this information because the size of the block is inconsistent (a high number of rooms in a block could be because this block is bigger). So, it makes sense to standardize variables 2 and 3.

On the other hand, variables 4 and 5 (population and number of households) are quite strongly influenced by block size. And block size does not seem to have much to do with median home values. So, variables 4 and 5 are not useful predictors. Small correlation coefficients like -0.02 and 0.07 provide strong support for this speculation.

Question 2:

1) What was done:

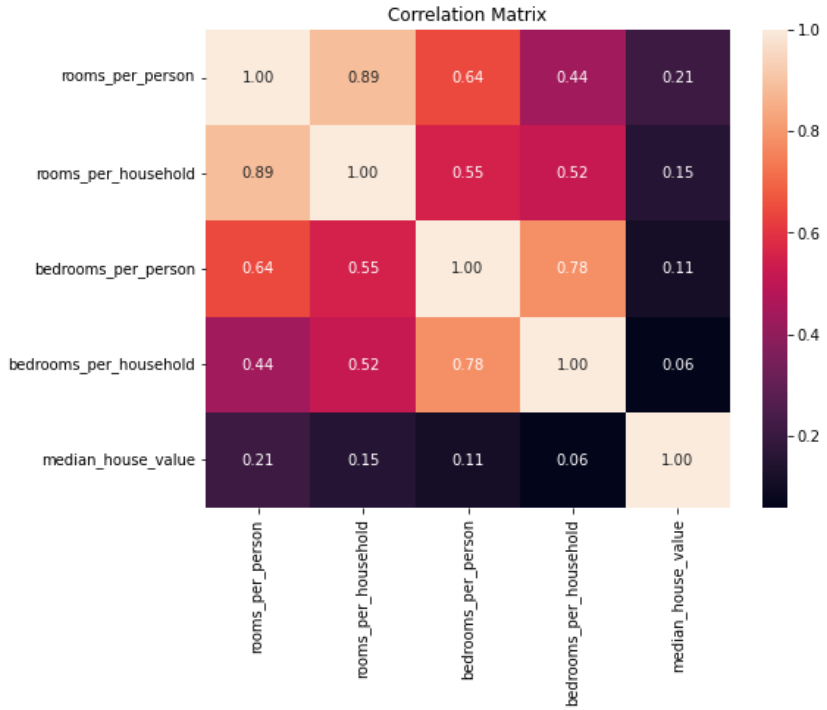
   Computed room/bedroom counts per population and per household (4 standardized features). Draw correlation heatmap with median house value to determine the better standardization approach.

2) Why it was done:

   The purpose of normalization is to find a way to reasonably compare room and bedroom numbers across blocks of different sizes. Correlation coefficients can show which normalized feature has better prediction effect.

3) Findings:

   The correlation coefficients of "rooms per person", "rooms per household", "bedrooms per person", and "bedrooms per household" with "median house value" are 0.21, 0.15, 0.11, 0.06 respectively. Features normalized by population have higher coefficients.



4) Interpretation:

Clearly the graph shows that it's better to normalize by population. Higher correlation coefficients mean that features normalized by population are more predictive measures for median house value.

Question 3:

1) What was done:

Conducted simple linear regressions for each predictor. Identified which predictor account for the most and least variance in median house values (Recorded R-Squares). Generated scatter plots for median income (most predictive) and population (least predictive) against median house value to visualize trends.

2) Why it was done:

R-Square value is critical in evaluating the effect of a simple linear regression model. The predictor has highest R-Square value explains most variances in the outcome variable (median house values), therefore is the most predictive predictor. Scatter plot visualizes distribution of datapoints, therefore is useful to see whether there's limitation / problem in dataset.

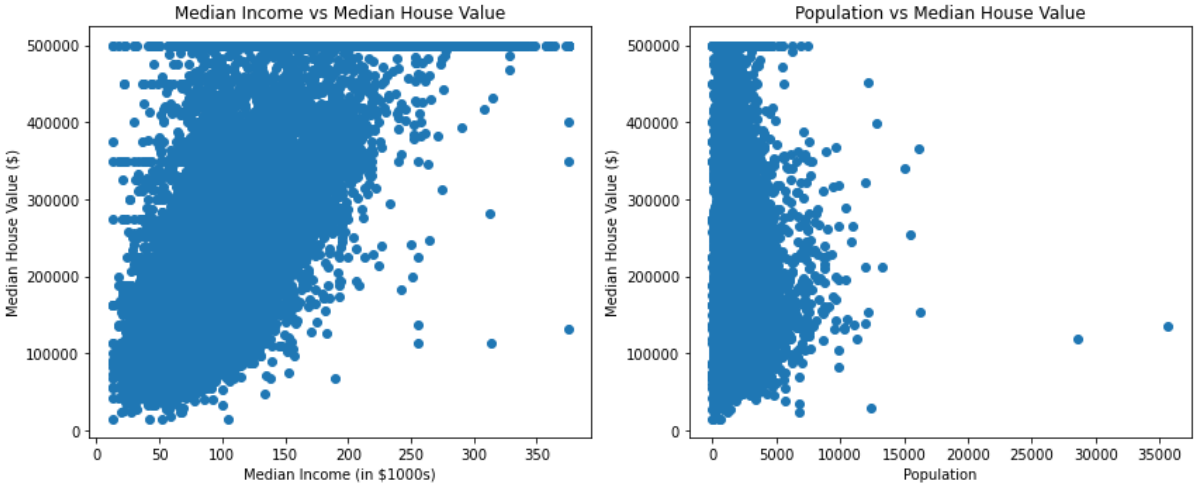
3) Findings:

"Median income" has highest R-Square value of about 0.473 while "population" has lowest R-Square of about 0.0006.

```
R-squared values are: {'housing_median_age': 0.011156305266710853, 'rooms_per_person': 0.04388269533891975,
'bedrooms_per_person': 0.012790501296178869, 'population': 0.0006076066693256887, 'households': 0.0043352546340906795,
'median_income': 0.47344749180719903, 'ocean_proximity': 0.15780848616855125}

The most predictive predictor is "median_income", the least predictive predictor is "population".
```

The scatter plot for median income vs. median house value shows a clear positive trend, but house prices appear to be capped at $500,000, which is a limitation of the dataset. The scatter plot for population vs. median house value shows no clear trend, shows again population's weak predictive ability.



4) Interpretation:

Based on the R-Square values, clearly the most predictive predictor is "median income", the least predictive predictor is "population". Population size is almost irrelevant in predicting house value. Additionally, we are able to note from the scatterplot that the dataset does not contain data for blocks where the median house value is above $500,000. Or perhaps blocks with median home values above $500,000 have had their median home values forced down to $500,000. The imposed house price cap may limit predictive potential of "median income". It might be more predictive if without the price cap.

Question 4:

1) What was done:

   Built a multiple regression model using all predictors. Compared it to the simple regression model using only median income. Recorded R-Square values of the 2 models. Used train-test split to build and test the models (test size = 0.2).

2) Why it was done:

   Multiple linear regression will definitely increase the value of the R-squared relative to simple linear regression with a single predictor. Each additional predictor only increases the R-squared value, but the question is by how much. If it only increases a little, then we can reasonably say that the other predictors are redundant. On the other hand, collinearity can be a potential problem when we use complex models with many predictors, because the model may fit the noise in the data. That's why we want to use a train-test split.

3) Findings:

   Full model (all predictors) has R-Square of about 0.591, meaning the combined predictors account for 59.1% of the variance in median house value. Single predictor model (only median income) has R-Square of about 0.459, meaning median income alone explains 45.9% of the variance.

```
R-Squre of the multiple regression model is: 0.5914166923864895; R-Squre of the simple best predictor model is
0.45885918903846656
```

4) Interpretation:

   The multiple linear regression model explains about 59.1% of variance in median house value, which seems good if we didn't compare. While adding predictors improves the

model's explanatory power, a large portion of the variance is still explained by median income alone, showing its strong predictive power. Therefore, I would say that the other predictors are basically redundant. Considering the overfitting problems associated with adding six additional predictors, I think the simple linear regression model with only the best predictor is better.

Question 5:

1) What was done:

   Computed correlation between standardized room / bedroom counts and correlation between population / number of households.

2) Why it was done:

   Checking the collinearity between two predictors is checking the linear correlation between them. The purpose of collinearity check is to assess if certain predictors are redundant when included together.

3) Findings:

   Rooms per person vs. Bedrooms per person: Correlation = 0.641

   Population vs. Households: Correlation = 0.907

```
Question5:
Correlation btw variable 2 and 3: 0.6414637002481975; correlation btw variable 4 and 5: 0.9072222660959659
```

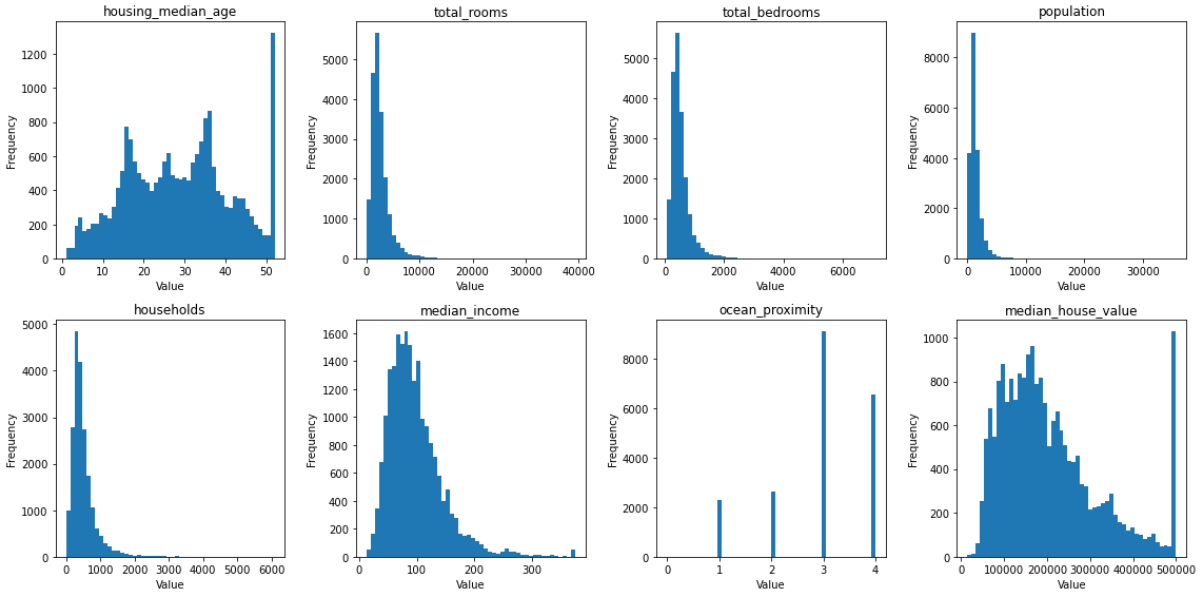
4) Interpretation:

   Correlation of about 0.641 between variable 2 (rooms per person) and variable 3 (bedrooms per person) indicates moderate collinearity. This could be a concern when both

of them are included in a model, since they are not fully independent. Correlation of about 0.907 between variable 4 (population) and variable 5 (number of households) indicates very high collinearity, meaning these variables are almost redundant. Including both of them in a model could bring issues.

Extra Credit Problems:

a) Plotted histograms of the distributions of 8 variables. Clearly it shows that none of these variables follows normal distribution.



b) The distribution of the outcome variable is the 8th histogram (median house value). Found that house values are right-skewed. And there is a price cap at $500,000, which distorts the distribution. The price cap at $500,000 likely reduces the predictive power of all the models / predictors above. It distorts the relationships in the dataset.