Fundamentals of Machine Learning
Homework 1
Spring 2024
Yan Konichshev

BEGINNING OF THE REPORT.

Question 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?*

Variables 2 and 3 = total number of rooms in a given block and number of bedrooms in a given block respectively

Variables 4 and 5 = population in the block, number of households in a block (note # of households and # of houses are not the same thing)

1) I compared the correlation coefficients of the variables 2, 3, 4, and 5 with respect to the median house price. In addition to that, I have normalized both 2 and 3 by predictors 4 and then 5 and compared the results. Additionally, I have plotted the regression of population to median price and number of households with respect to the median price.

2) Correlation coefficient is a great way of statistically measuring the relationship between two variables. Intuition was to compare how high/low coefficients will go once I "normalize" them by either population or number of households. To give more meaning to the number of bedrooms and rooms in the block I needed to provide them with some additional feature, which was "per capita," or "per household".

3) Let me address why predictors 4 and 5 are not very useful by themselves to predict the median house value. As it could be seen from the graphs below (Fig 1 and 2), there is no relationship between the population number / number of households with the median household price. This is also shown by the incredibly small R^2 evaluation metric that approaches 0. Secondly, it probably is a good idea to normalize 2 and 3 by 4, or 5 not simply because it is providing more meaning, but also because the correlation coefficients slightly improve (please see table 1).

Variable	Correlation coefficient
Population	-0.0246497
Number of households	0.0658427
Number of rooms	0.134153
Number of bedrooms	0.134154
Rooms normalized by population	0.209482
Bedrooms normalized by population	0.113095
Rooms normalized by households	0.151948
Bedrooms normalized by households	0.0582604

Table 1

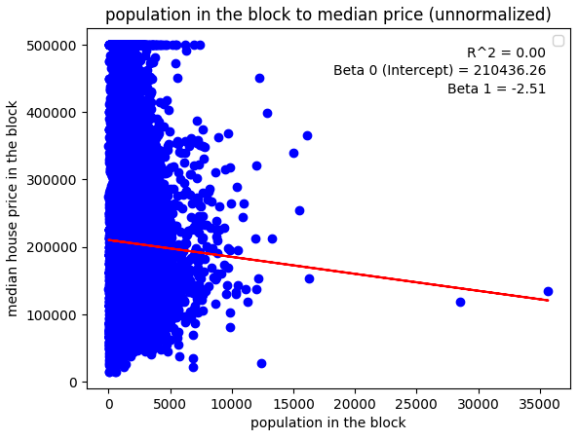

Fig 1

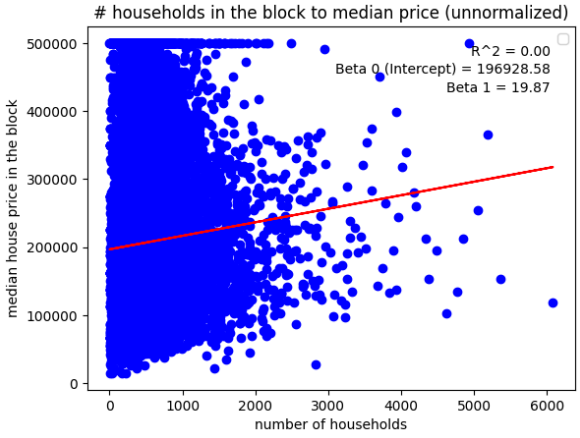

Fig. 2

4) I think my findings point that there is a potential exploring the normalized predictors 2 and 3 to further look for dependencies and correlations between different variables. It was a good idea to standardize/normalize the predictor variables 2 and 3 by population and/or because although the number of bedrooms and/or number of total rooms affect the housing price there is MUCH more we need to take into account when valuing a house. This could be shown by small numbers we got even after the “normalization". Additionally, I think that the number of people living in the block, as well as number of households **are not reflective** of the price of the housing (e.g. rich neighborhoods have very few residents / households, but insanely high prices, as well densely populated areas also do have higher prices and etc.)

Question 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)?*

1) I must notice that this data is extremely noisy and hard to work with because even when I added "normalized" the data, I *have not achieved* extreme results of improving the quality of the prediction, or improving the model estimators. Thus, I have plotted all the potential variations of the relations, such as # rooms vs. median housing price (not normalized), # bedrooms vs. median housing price (not normalized), # rooms normalized with respect to the population vs. median housing price, # bedrooms normalized with respect to the population vs. median housing price, as well as same variables with respect to the number of households.

2) I was looking for any improvement of R^2, and honestly just for a better fit of the regression with respect to the datapoints. I wanted to see whether there will be some sort of dependency in between one of the aforementioned predictors and the outcome we were trying to predict.

3) For all the observations, please refer to the figures 3-8 above. Although, as I have mentioned in the first section of this answer, I was bot able to find a significant relation between these variables, I was able to start exploring more about this dataset. I would stick further with the rooms and bedrooms **per capita,** since by looking comprehensively at the R^2 values of these normalized values I was able to make out the meaning out of these variables. Please see my decision making process with exact digits below:

Correlation coff between 'number of rooms' and 'house_value': 0.1341531138065631 ===> original unnormalized

Correlation coff between 'number of bedrooms' and 'house_value': 0.1341536985700889 ===> original unnormalized

Correlation coff between 'number of rooms normalized by households' and 'house_value': 0.15194828974145796 ==> improved slightly after normalization

Correlation coff between 'number of bedrooms normalized by households' and 'house_value': 0.0582604339126752 ==> decreased by 2.6 after normalizing

Correlation coff between 'number of rooms normalized by population' and 'house_value': 0.20948196900668967 ==> improved by 1.53 after normalizing by population

Correlation coff between 'number of bedrooms normalized by population' and 'house_value': 0.11309509846221796 ==> slightly decreased by 1.18 after normalizing by population

4) I have noticed that the correlation coefficient, and thus the R^2 values went down when I used to normalize by the number of households significantly, which was interesting, since the intuition was that it would be more meaningful to use #rooms/bedrooms with respect to the

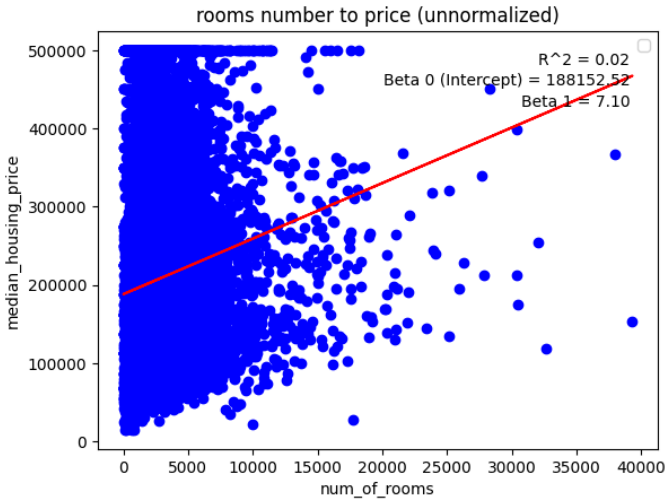

Fig. 3

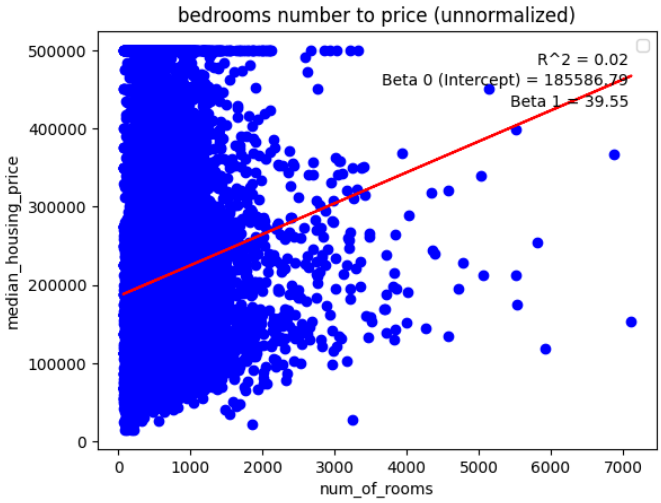

Fig. 4

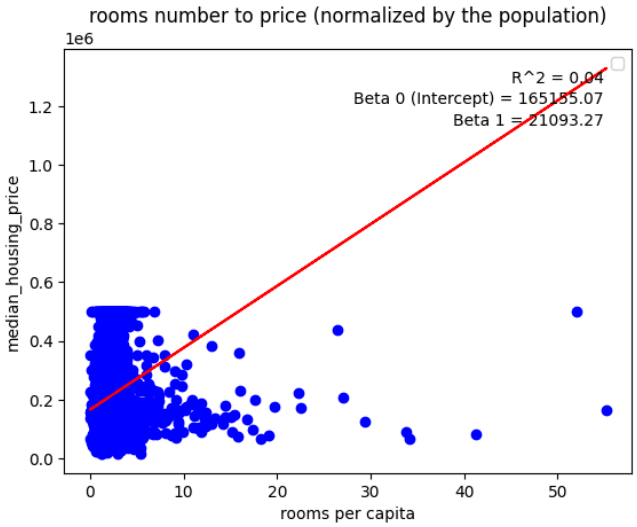

Fig. 5

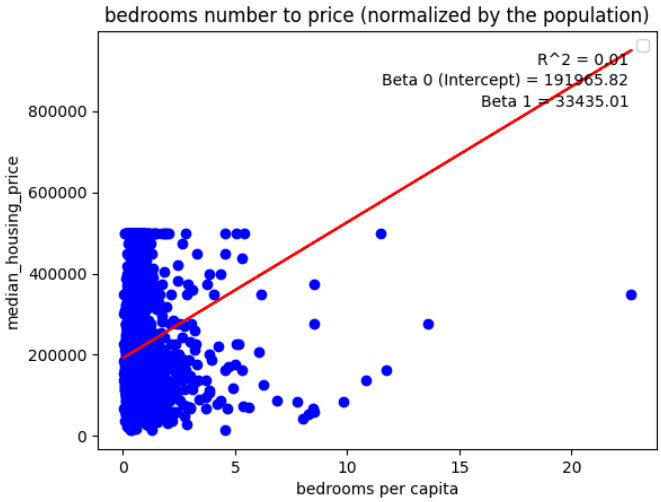

Fig. 6

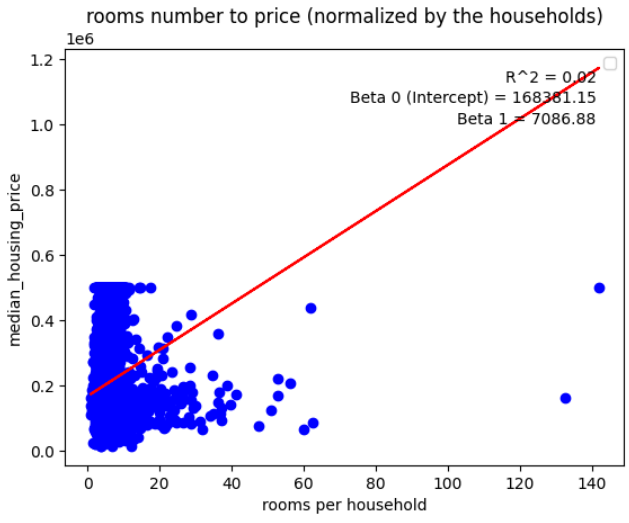

Fig. 7

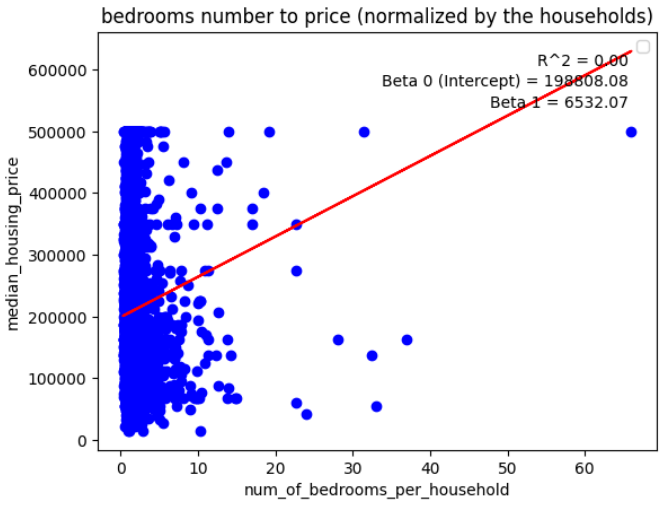

Fig. 8

number of households, than by population. Interestingly enough, even though I “gave” meaning to previously useless predictors such as number of rooms and bedrooms it still didn’t help me much, as I was not able to see any significant improvements in predictions, but decided to move on with the normalized predictors 2 and 3 with respect to the **population**, as it showed to have better results.

Question 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?]*

1) I have pragmatically approached this task and simply created scattered plots and linear regressions on top of it for all the raw and normalized predictors. After that, I have compared both correlation coefficients and the R^2 values and started to think which ones would be most and least predictive.

2) I decided to go with the linear regressions and scatter plots approach because linear regression could be great way of seeing whether there is, or there isn't dependency between a predictor and outcome. In addition to that, there is a way to measure how "good" the fit of the regression is (by examining the evaluation metrics such as RMSE, R^2 and etc.)

Fig. 9

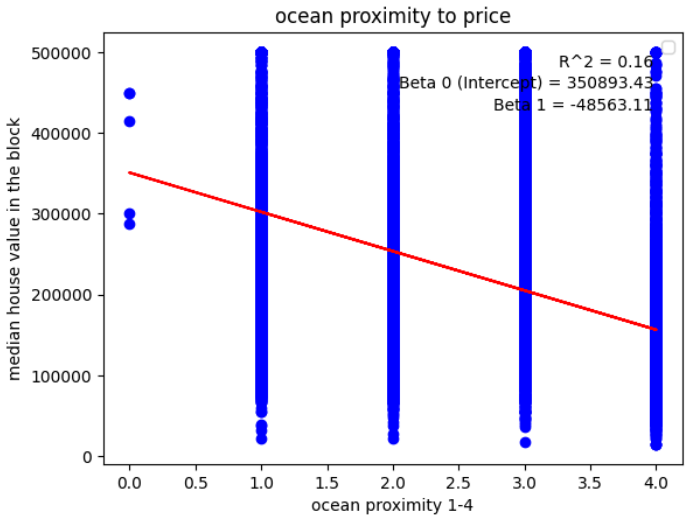

Fig. 10

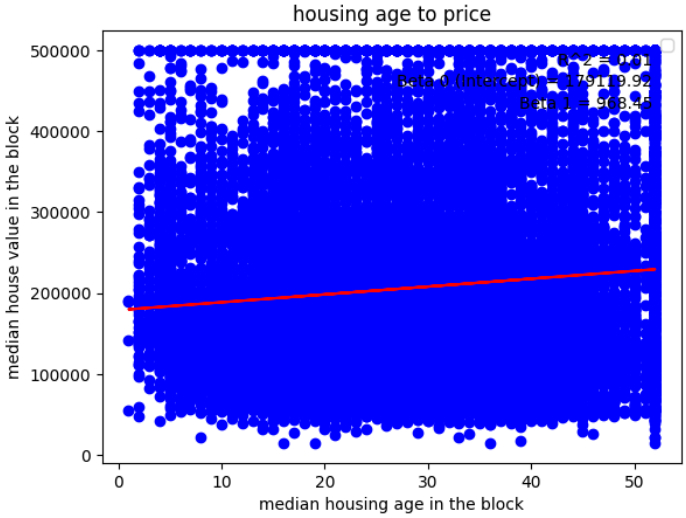

Fig. 11

3) For the regressions of the predictors 4 and 5, please refer to the figs 1-2 above. For the regressions of the normalized predictors 2 and 3 (number of bedrooms and number of rooms per capita), please refer to the figs 5-6. I have plotted and constructed new regressions with the remaining 4 factors, please see figs 9-11 for that. The least predictive variable turned out to

be **median housing age** in the given block. As it can be seen from the fig. 11, there is no relation whatsoever, with R^2 being almost 0, datapoints are scattered all over the place. The most predictive feature of the median housing price, on the other hand, would be the median income in the given housing block. As it can be seen, R^2 reached a record-breaking 0.47, which was unusual for this dataset up to this point. Additionally, it is worth mentioning yet another *big limitation* of this dataset is the **"ceiling" of housing prices** which is set at 500,000, as there are no points in the dataset beyond 500,000.

4) To reiterate, the best predictor of this dataset is the median income, and the least predictive feature would be the median housing age in the block. Speaking in other words, it does really matter how rich people are in the housing block as it hugely influences the median housing price. On the other hand, it completely doesn't matter whether the median age of the buildings in the block is high or low, as it **does not** influence the median price of the housing the block. Finally, we could have had aimed at better results if we had a complete picture of the housing market by not having an artificial "ceiling" of median prices.

Question 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?*

1) I took all the predictors we need (columns with values for these features) and fitted them into the array reshaping the data as necessary. After that, I have trained the model with 7 predictors and one outcome. Just for the curiosity, I did it with both normalized predictors 2 and 3, and with raw predictors 2 and 3. Also, I have tested my model using testing dataset, which I separated from the training dataset by the principle of 20/80, where 20% of data was isolated for testing purposes and the rest 80% for training.
2) The idea behind fitting all these predictors into a single model is to provide the model with a "bigger picture" of what was actually happening, because as we have seen from the previous questions, single predictors no matter normalized, or not normalized didn't provide us with desired outcomes and better values of estimators. By incorporating all the predictors into the model, I have allowed it to learn more about the hidden features, dependencies which might not be as obvious as they are from the beginning.
3) Comprehensive picture with having more predictors of course performed way better than **any** of the previous predictors did. I will incorporate multiple tables with results of different models and approaches I used to make the model "learn" more about the median prices.

Using all the raw predictors:
I am using RAW predictors 2 and 3 (not normalized)

Intercept: 61420.40794291612

Beta 1: 1208.5653100956806
Beta 2: -47.693666214856954
Beta 3: 217.25066034241667
Beta 4: -38.88095305142316
Beta 5: 178.33433850081292
Beta 6: 1694.0724863453358
Beta 7: -26217.362887313957

R^2 Score: 0.5878967849154422

RMSE: 74432.94343469487

Mean arithmetic error: 54184.15148779294

As it could be seen, even though I am using raw predictors without any normalization, the comprehensive model performs better than the best single predictor (median income) with R^2 being 0.59 with respect to the best estimator's R^2 being 0.47.

Now, let's take the model with normalized predictor:

Using normalized predictors:

Intercept: 206713.46705426337

Beta 1: 16423.532840938216
Beta 2: 75777.05352651316
Beta 3: -26089.565802144738
Beta 4: 2651.170666096769
Beta 5: 2375.0109541437737
Beta 6: 50453.610840641646
Beta 7: -42354.4432774165

R^2 Score: 0.5858832464347383

RMSE: 74614.56173497147

It performs more or less same to the raw estimators model at around 0.59.

4) As discussed and shown above, this model performs way better than any single-variate model we explored in the question 3. This is because with having an access to all the collected information about housing blocks, our model is able to plot all these dots in the 8 dimensional space, thus having more contextual data, which leads to better predictions. Please note that I have run the code for the R^2 generation with different random seeds and they all seem to produce results closely approaching or fluctuating around 0.59-0.60

Question 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?*

1) To figure this out, I had to create the scatter plots with normalized rooms and bedrooms and then build a linear regression to figure out whether there is a linear relationship. I have to mention that out of curiosity, I did it with both of the normalized approaches and the unnormalized ones and found pretty interesting results.

2) I believe that building a linear regression model was a suitable approach for this, cause it could show the general tendency of the dataset, and if the predictors happen to be collinear with each other, the line fit would be perfect (read R^2 would be approaching 1). Collinearity carries an important insight about the data, which is

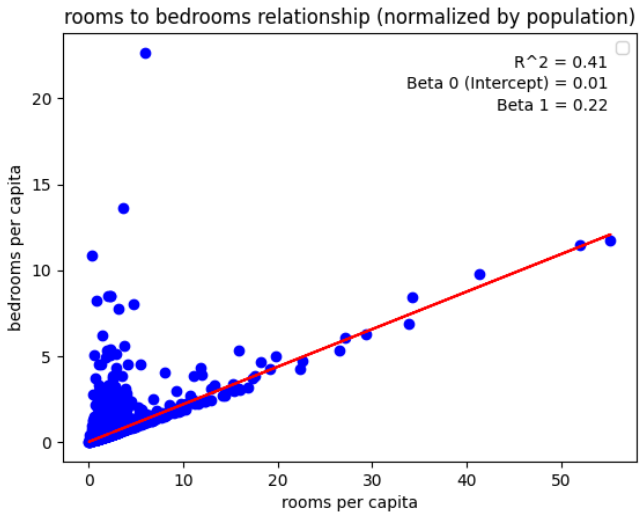

Fig. 12

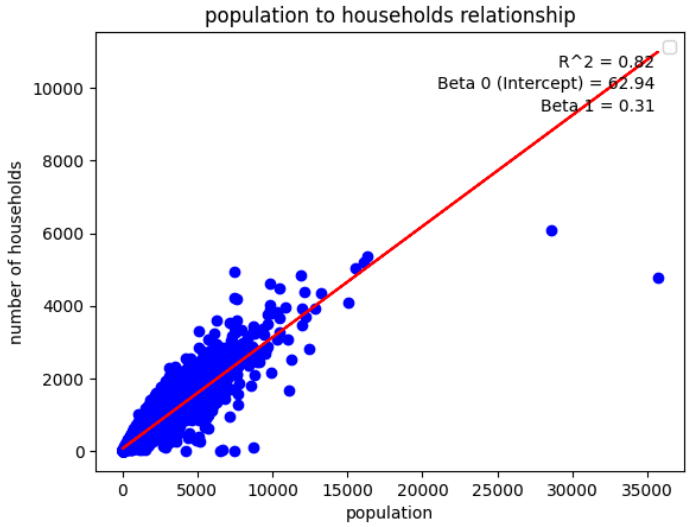

Fig. 13

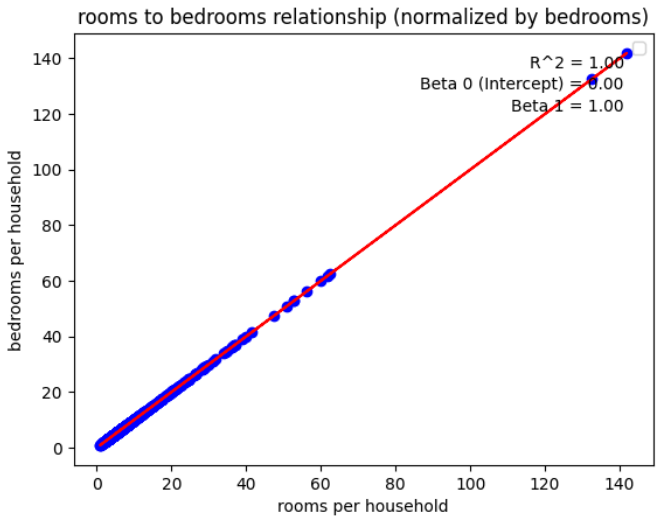

Fig. 14

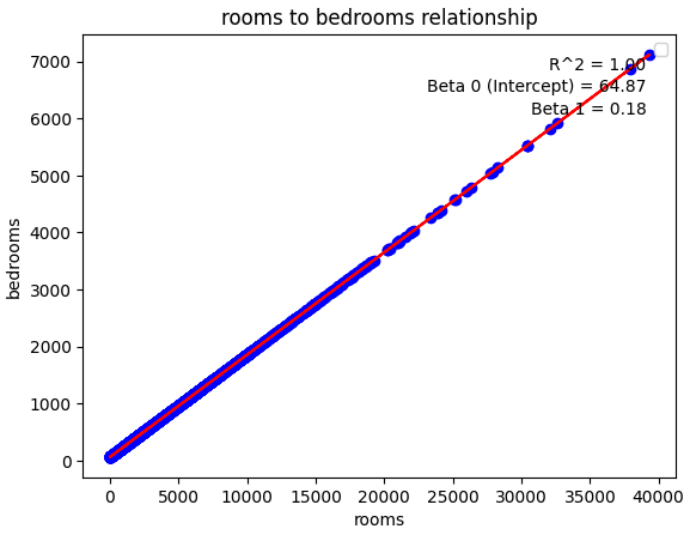

Fig. 15

basically implying that the variables are completely dependent on each other (and redundant). Thus, it is important to examine whether this is actually the case or not.

3) As mentioned previously, I have mapped both the normalized and non-normalized features 2, 3 and features 4, 5 with each other to see if there is a concern for collinearity. As it turned out, there is a valid concert for unnormalized rooms and bedrooms predictors (see fig. 15), which could be also seen from the normalized version of it (when we normalize by number of households). However, there is some sort of multicollinearity whenever we normalize by the population number, yet still the tend seems to be worrying, as there is a clear tendency line and given the raw data and the other normalization form, one could express a valid sign for the collinearity of these two variables. On the other hand, Population to households graph (fig. 13) with an R^2 estimation of 0.82 additionally is expressing a concern for collinearity of data, however there is no additional contextual information to support this claim, although the correlation is pretty high as well.

4) It is logical that the higher the population the more households there is to expect. Thus, I would say that there is a good reason to include both the population and the number of households into the model, as there are still some hidden dependencies and outliers model might show us. On the other hand, the concern for having number of bedrooms and number of rooms to be collinear **is valid** as we have examined earlier.

Extra Credit Questions:

A) Let's map all the frequency distributions and see if there is anything that could suffice for a normal distribution. As it could be seen from the fig. 16, the only distribution that could be reasonably described as normal would be the distribution of the median income in the given block (third row, second plot) because it is more or less bell curved and centered (although not perfectly). Some might argue that median house value might look like a normal distribution, however there is an outlying prevalence of houses worth 500,000. Something like a leverage point, that doesn't let us call it a normal distribution. In addition, all the other distributions (first two rows) could not be reasonably described as normal due to the right skewness.

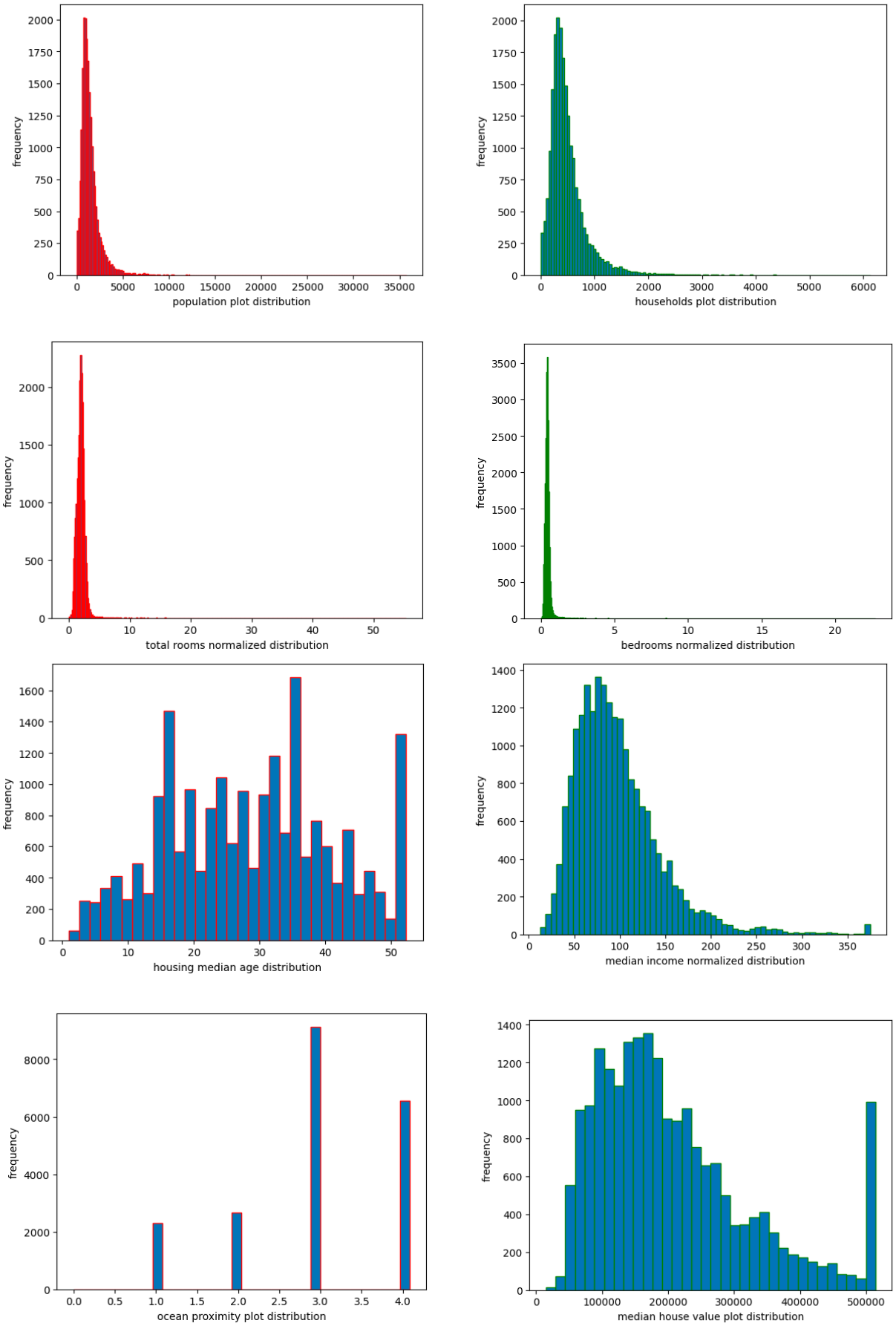
frequency
population plot distribution
households plot distribution
total rooms normalized distribution
bedrooms normalized distribution
housing median age distribution
median income normalized distribution
ocean proximity plot distribution
median house value plot distribution

Fig. 16

B) I think a huge factor that influences the validity of the conclusions is the very last bin that acts rather like a leverage point (see last row, second picture on fig. 16), which signifies that our experiment was not descriptive enough and requires more probes and records. It is also worth mentioning that based on the previous results I got that the market prices are capped at 500,000, which is rather an anomaly than an actual life data. These two factors combined contribute hugely to the quality of the regressions and the results of our predictions.

END OF THE REPORT