Data Analysis of housing prices

Gilberto Juárez Rangel
(Dated: 5 de mayo de 2024)

I. DATASET

The California housing prices dataset based on the 1990 census data consists of 9 numerical features (8 of them numerical and 1 categorical) and 1 target (the housing price) with a total of 20640 rows. The features included in the dataset are the following:

- longitude (°)
- latitude (°)
- hosing median age (years)
- total_rooms (#)
- total_bedrooms (#)
- population (#)
- households (#)
- median_income ($)
- ocean_proximity (Categorical)

II. DATA EXPLORATION PLAN

To get familiarity with the dataset, we will use a pairplot to see the histograms of each feature and if there exist any relationship between them.

We will also search if there exists nan, null or non-sense data (e.g. of non-sense data, height: -1,500 m). For this we need to know what kind of non-sense data we can encounter for each feature, the possible values for each feature and target are:

- longitude: $-180 < x < 180$
- latitude: $-90 < x < 90$
- housing median age: $0 < x < 172$ (I searched for the oldest registered house in California).
- total_rooms: $x > 0$
- total_bedrooms: $0 < x < \text{total\_rooms}_x$
- population: $0 < x$
- households: $0 < x$
- median_house_value: $0 < x$
- median_income: $0 < x$

If some data gets out of range its considered non-sense and we must clean it.

III. CLEANING AND FEATURE ENGINEERING

After exploring the dataset I figured out there exist 207 nan values. Since this is a very small amount compared with the number of rows available in our dataset I decided to remove the rows.
We also found that there is only 5 rows with the category 'Island' of the ocean_proximity feature. Since this is a small amount of data we decided to remove it.

When I plotted the histograms of the raw data I found some skewness in the numerical values except for the latitude, longitude and housing median age features, this can be seen in figure 1.

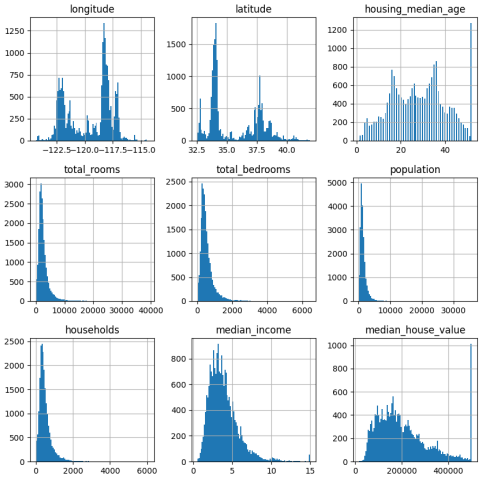

Figura 1: Histogram of the 8 features and the target data with 100 bins.

To reduce the skewness we tried 3 different transformations (square root, logarithm and boxcox). I used the normaltest to get the best transformation for each feature. In this case we used the boxcox transformation for the total_rooms, total_bedrooms, population, households and median_income features and the logarithm transformation for the median_house_value target. After this transformation I plotted the histograms and we got a more normalized distributions, this can be seen in figure 2.

After transforming our data we removed the outliers by using a boxplot.

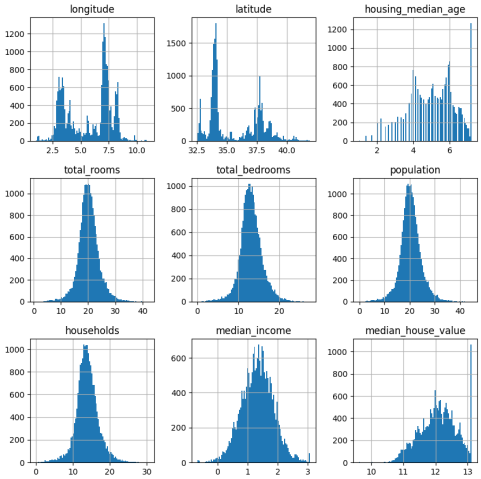

Figura 2: Histograms of the 8 features and target after being transformed by boxcox and logarithm function.

IV. FINDINGS AND INSIGHTS

Now that the data has been cleaned and transformed, I did a pair plot to see if any information can be observed between feature-feature and feature-target, this is shown in figure 3.

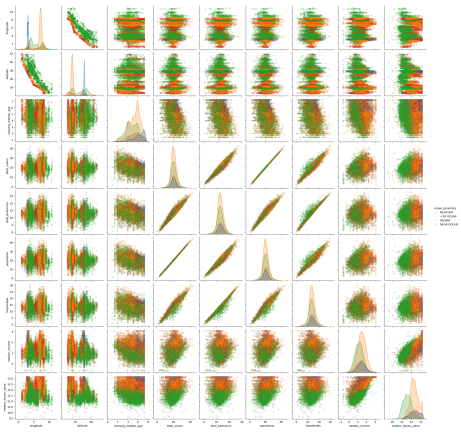

Figura 3: Pair plot between feature-feature and feature target for each ocean_proximity categorical data (the order from top to bottom and from left to right is the same as is shown in II).

From figure 3 we can observe that a linear relationship between the longitude and latitude data exist. We also see a linear relationship between total_rooms, total_bedrooms, population and households, but the one with strongest linear relationship is between population and total_rooms. This makes sense since usually the increment of population requires an increment of rooms. Since the linear relationship between the population and total_rooms is almost perfect, we decided to remove one of them to avoid multi-correlation with the target.

From the feature-target plots (bottom row or right column), no relation is found, the closest relation we can observe is a linear relationship between median_income and the median_house_value. We can also see that the inland category is usually cheaper than a house near to the ocean (Near ocean, ¡1H ocean and Near bay).

V. HYPOTHESIS

1. Hypothesis 1: There is no correlation between the spatial features (Longitude and Latitude) and the housing price.
2. Hypothesis 2: There exists a linear correlation between the median_income and the median_house_value.
3. Hypothesis 3: In land category has better correlation than the other 3.

By using the Pearson correlation between the median_income and the median_house_value we obtain a value of 0.677 which is close to what is usually considered a strong linear relationship $0.7 \leq |r| \leq 1$. In this case we can say that there exists a moderate linear correlation between the median_income and the median_house_value accepting our null hypothesis. We must consider that this relationship is after the transformations done previously and the real relation is an inverse boxcox.

VI. SUGGESTIONS

To further analyze this dataset we must search for polynomial relationships between the features and target and whether is necessary to keep all the linearly related features (i.e. total_rooms, total_bedrooms, population, households). We shall also explore how the median_house_value is related to the spatial coordinates, this could be done by doing a kind of heatmap. Lastly we should consider an analysis for each type of category of the ocean_proximity feature since this could be an important feature to consider when analyzing the median_house_value.

VII. QUALITY OF THE DATASET

This dataset has a small amount of nan or null data, this was helpful to avoid removing a considerable amount of rows just by missing values. When using a boxplot to remove outliers after transformation we lost around 2 thousand rows. This means we kept around 90 % of the data available in the dataset for analysis.