HOUSE PRICE PREDICTION

Group 4

Oluwatosin Oyediran, Mark Gopani, Zachary Dareshori
IST 718 - Big Data Analaysis

Table of Contents

1. Abstract

The objective of the project is to predict median house prices in different regions of California based on 1990 Census data. The dataset is fetched from Kaggle and can be found using this link(https://www.kaggle.com/datasets/fatmakursun/hausing-data). Since the data is actual census data therefore the observations of the data are of specific locations with available geographical coordinates. Along with the coordinates of the locations, other variables specify the demographics and provide information about the house layout plans and median age of the house at those locations. Taking into account these different components for a location, we are trying the predict the median house price of the regions and compare factors affecting prices in different regions.

The predictions that were attempted can be framed as : Given the other variables present in the dataset, can we predict the price of an unseen home? Can we generate a relatively reliable prediction of home value given only longitude, latitude, total rooms, and total bedrooms? Can we use median age to help predict housing prices on its own?

Other inferences that we try to answer are: Which variables are most influential on median home price? i.e. how much does median home price change with a change in each variable? Does the impact of median income on median home value remain relatively constant throughout the dataset, or are their outliers? What might those outliers tell us?

What is the influence on median home value of the interaction between population and total rooms ? Do home values tend to be higher where populations are more concentrated?

Is there a positive correlation between the number of non-bedroom rooms and median home value?If we partition the dataset between urban and rural locations, do our relationships hold true? Is this a potential confounding variable, or is its influence already captured by the existing variables?

2. Data Cleaning and Exploration

The data contains one observation per census block group from certain districts in California. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people).

2.1 Dataset variables defined:

Longitude: A measure of how far west a house is; a higher value is farther west

Latitude: A measure of how far north a house is; a higher value is farther north

HousingMedianAge: Median age of a house within a block; a lower number is a newer building

TotalRooms: Total number of rooms within a block

TotalBedrooms: Total number of bedrooms within a block

Population: Total number of people residing within a block

Households: Total number of households, a group of people residing within a home unit, for a block

MedianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

MedianHouseValue: Median house value for households within a block (measured in US Dollars)

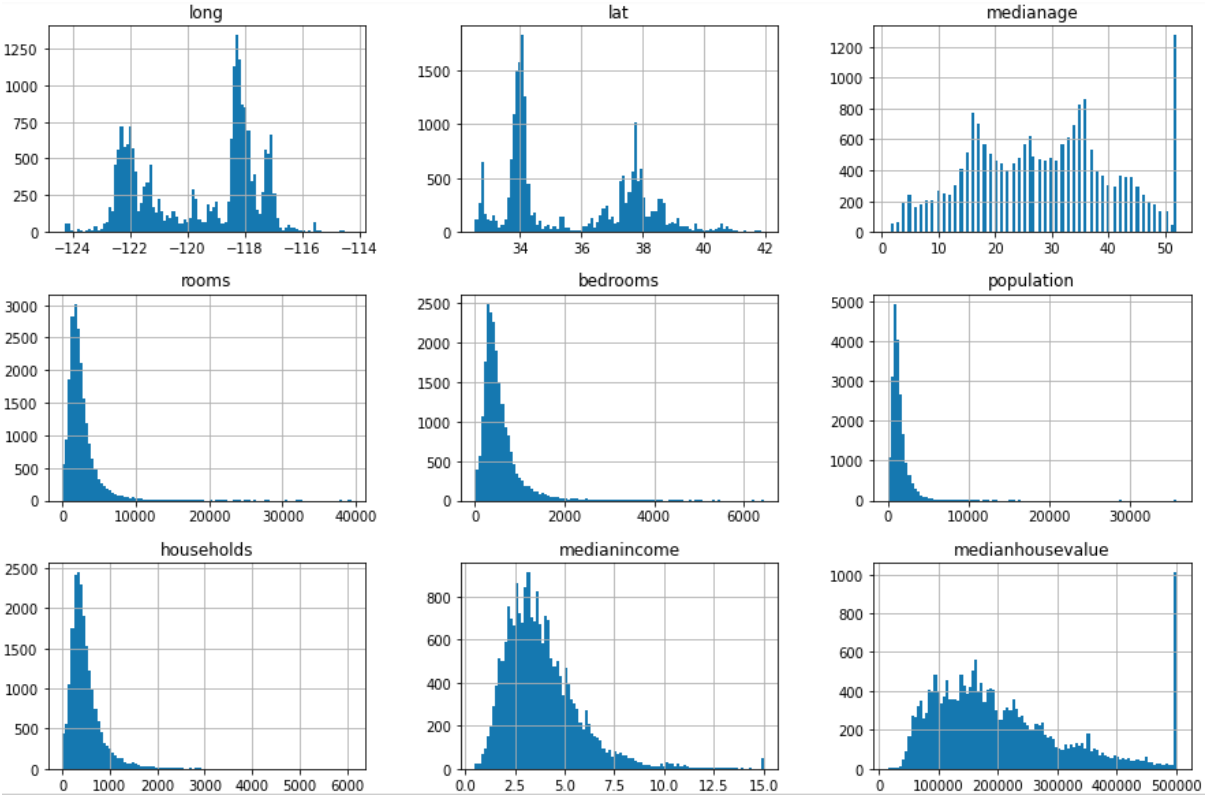
2.1 Data Exploration

The observations with null or zero values have been already excluded in the dataset

```
from pyspark.sql.functions import isnan, when, count, col
df.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in df.columns]).show()
```

```
+----+---+---------+-----+--------+----------+----------+------------+----------------+
|long|lat|medianage|rooms|bedrooms|population|households|medianincome|medianhousevalue|
+----+---+---------+-----+--------+----------+----------+------------+----------------+
|   0|  0|        0|    0|       0|         0|         0|           0|               0|
+----+---+---------+-----+--------+----------+----------+------------+----------------+
```

Distributions of each variable.



3. Data Exploration Insights

In this dataset we observed block group mean population of 1427 people.
We observed that over 1200 block groups have a high median age of around 50 years , and there are bout 1000 block groups with high median house value around $500,000. Many variables show right skewed distribution: most block groups have number of rooms ranging from 1000-5000, most number of bedrooms ranging from 400-800 and number of households ranging from 250- 750. The median income variable shows most block groups have income between $12000-$40000 with few block groups showing median income more than $150000. These block groups are probably the urban areas of California.
We see around 70 blocks with median value higher than rest of the distribution. The skewed distributions with outliers might affect the house price prediction of the block.
In our analysis below ,we see if median age can be used as the only predictor as well as divide the regions based on whether they are urban or rural to get a comparative insight of the house pricing.

4. Methodology

4.1 Which variables are most influential on median home price?

We used a Vector assembler to put the original features into a feature vector column and then used a standard scaler to standardise all the columns with all features are centred around 0 and having variance in the same order. The input column was the feature vector predicting the median house value. We then fitted the data onto a linear regression model and then extracted the coefficients of each original variable to determine its importance in predicting the outcome variable.

4.2 Given the other variables in the dataset, can we predict the price of an unseen home?

We scaled the variable 'medianhousevalue' to express the house value in units of 100,000 to match the other data scales. A target such as 453700 became 4.537. we then decided to add a few variables like:

- Rooms per household: The number of rooms in households per block
- Population per household: How many people live in the households per block
- Bedrooms per block: How many rooms are bedrooms per block

We also dropped some columns that we believed were not useful, like longitude, latitude, medianage and total rooms. A vector assembler was then used to put features into a feature vector column, and a standard scaler was used to scale the data. The input column was the features, and the output column was populated with rescaled features. We split the data into training and test sets using the random split method, ran a linear regression and then generated predictions for our test data.

4.3 Can we generate a relatively reliable prediction of house value given only longitude, latitude, total rooms, and total bedrooms?

With the already performed pre-processing, A vector assembler was used to turn the predictor variables into a "features" column of the above listed variables for processing by spark. A standard scaler was then applied to the data. Then, a linear regression was done. We then computed the RMSE and R-squared scores.

4.4 Can we use median age to help predict housing prices on its own?

We defined medianage as our label/predictor variable, split the data into training and testing loaded a vector assembler to assembles the features in a vector, selected the features and label column, because we had already vectorized our features. We then fit the testing and training set to a liner regression model and lastly computed the r-square score.

4.5 Can rooms per person function as a predictor variable? Do home values tend to be higher where populations are more concentrated?

A derived column was created by dividing the values in the 'rooms' column by the values in the 'population' column and saving the results to a new column called 'rooms_per_person'. A vector assembler was then used to turn the predictor variables into a "features" column for processing by spark. A standard scaler was also applied to the data. Then, a linear regression was done.

4.6 Is there a positive correlation between the number of non-bedroom rooms and median home value?

A derived column was created by dividing the values in the 'bedrooms' column by the values in the 'rooms' column. The results were saved to a new column called 'pct_bedrooms'. This column showed the percentage of rooms in a given area that were bedrooms.

A vector assembler was used to turn the predictor variables into a "features" column for processing by spark. A standard scaler was then applied to the data. Then, a linear regression was done.

4.7 If we partition the dataset between urban and suburban/rural locations, do our relationships between predictor and target variables hold true?

In order to partition the dataset between rural and urban, we used the withColumn() function to pass the longitude and latitude for each row to a UDF that would call an API to determine the city name. The city name was then checked against a list of cities in California. If the city name was present in the list, the UDF would return the value 1. If not, the location was determined to be non-urban, and the UDF would return a 0. With this new column containing a 1 or a 0, we were able to split the dataset into two separate Spark dataframes. Each dataframe was then passed through a pipeline that ran a vector assembler, standard scaler, and linear regression on the data.

5. Model Inferences and Predictions

5.1 Which variables are most influential on median home price?

We saw that among the original variables, median income had the highest positive coefficient. Number of households and bedrooms and median age are good influential in predicting the house price, population has a negative coefficient showing that if the population of the block is lower, we can expect a higher house pricing in the block. The number of rooms has a negative coefficient which makes us question the type of room as defined by the dataset as one would expect the house price to go up with more number of rooms.

coefficient	value
medianincome	90617.207006
households	47966.126963
bedrooms	42305.405012
medianage	23672.715963
population	-40200.914108
rooms	-42852.377798

5.2 Given the other variables in the dataset, can we predict the price of an unseen home?

Upon running the regression, we got the r- squared score which was 0.55 and the RMSE which was 0.78. We then decided to add a grid search to select the best value based on the hyperparameters. We got slightly different r-squared and RMSE scores which were 0.54 and 0.77 respectively.

Based on these parameters, we see that the performance of the model is not great. We could possibly improve the model performance by pre-processing the data further by removing outliers, transforming skewed features, feature extraction by PCA etc. Therefore, we cannot confidently predict the house price of an unseen house based on the variables in the dataset.

5.3 Can we generate a relatively reliable prediction of house value given only longitude, latitude, total rooms, and total bedrooms?

We tried to find out if we can predict house prices based on fewer variables, so we ran a regression with just the variables listed above. The r-squared score was 0.30 and the RMSE

was 0.98. Based on these parameters, we concluded that those variables cannot give a reliable prediction of the price of a house.

5.4 Can we use median age to help predict housing prices on its own?

Based on the low r-squared score from the linear regression run, which was 0.012, we conclude that median age on its own is a bad predictor of house price.

5.5 Can rooms per person function as a predictor variable? Do home values tend to be higher where populations are more concentrated?

The inference goal for this question was to determine if the new derived variable "rooms per person" could be used as a predictor. To determine the answer, we ran a linear regression on the dataset after creating the derived variable. Before the linear regression was run, a vector assembler and standard scaler were applied to the dataset. The final results can be seen in the table below. 'Rooms per person' functioned as a weak positive predictor. Its coefficient was much smaller than the other variables in the dataset.

	coefficient	value
5	medianincome	89693.200235
4	households	57317.634021
2	bedrooms	32864.543950
0	medianage	23849.790742
6	rooms_per_person	6106.628196
3	population	-36346.404638
1	rooms	-46508.854283

5.6 Is there a positive correlation between the number of non-bedroom rooms and median home value?

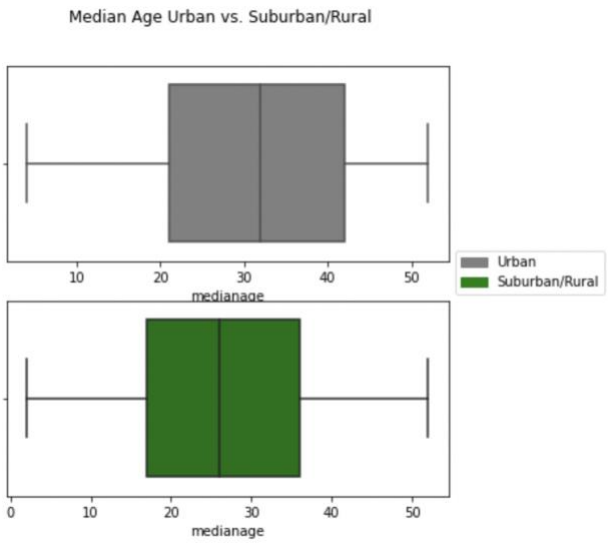
The inference goal for this question was to determine if the new derived variable "pct bedrooms" could be used as a predictor. 'pct bedrooms' is ratio of bedrooms to total rooms. In other words, it describes the proportion of rooms in each block that are bedrooms. To determine the answer to this research question, we ran a linear regression on the dataset after creating the derived variable. Before the linear regression was run, a vector assembler and standard scaler were applied to the dataset. The final results can be seen in the table below. 'pct_bedrooms' functioned as a strong positive predictor. In other words, each rise in 'pct-bedrooms' by one standard deviation was associated with a rise in median home value of $24,427.

	coefficient	value
5	medianincome	97616.649336
4	households	52986.962032
6	pct_bedrooms	24427.247421
0	medianage	22717.748982
2	bedrooms	3159.010011
1	rooms	-4685.126879
3	population	-44320.908885

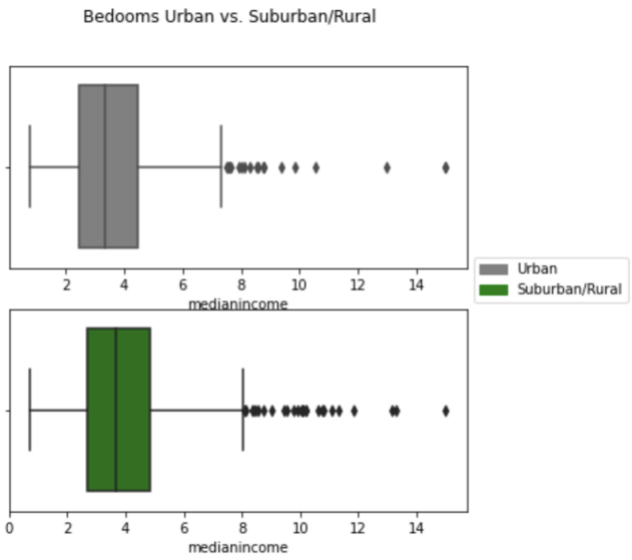
5.7 If we partition the dataset between urban and suburban/rural locations, do our relationships between predictor and target variables hold true?

The inference goal of this question was to determine if whether or not a block was located in an urban area or not functions as a confounding variable. If it was a confounding variable, it would change the relationship between the predictor variables and the independent variable (median housing value). We used a linear regression to find the coefficients in a dataset segmented into only urban data points, and another dataset of non-urban datapoints. The datasets were both standardized before the linear regression was done so that we could compare the most influential predictors among each other.

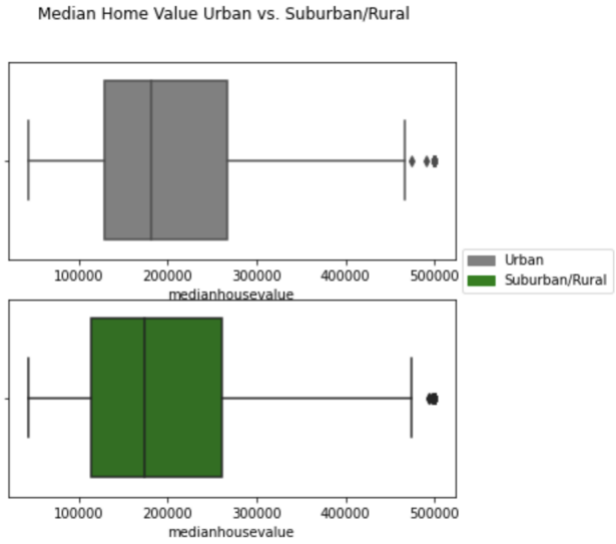
The data was segmented by calling an API on the longitude and latitude of each datapoint to determine whether it was located in a city or not. Then two separate dataframes were created. The graph below show the differences between the two dataframes:



Median age was slightly higher in the urban locations.



Median income was slightly higher in the suburban/rural locations



Median household value was slightly lower in the suburban/rural locations. However, the third quartile was also slightly higher in the suburban locations.

Next we ran the linear regression and determined the coefficients for each dataset:

	coefficient	value
2	bedrooms	139983.199310
5	medianincome	87804.787011
1	rooms	-66230.798410
3	population	-63112.865702
0	medianage	15920.135451
4	households	-12733.536026

Urban

	coefficient	value
5	medianincome	102221.458207
2	bedrooms	68110.975064
1	rooms	-50137.720001
3	population	-32982.035446
4	households	18218.784236
0	medianage	6625.422946

Suburban / Rural

The coefficients were noticeably different between the two datasets. Most notably, in the urban locations, bedrooms served as the most influential predictor. On the other hand, for the suburban/rural locations, median income served as the most influential predictor. This made sense intuitively, because in cities, bedrooms tend to significantly increase home value (and rent—speaking from personal experience). We can conclude that whether or not a datapoint was urban was indeed a confounding variable.

6. Conclusion

We learned from our analysis that there are some improvements needed to our model for better prediction. We can change the parameters that we passed to your model in various ways, change the variables in the original Data Frame, or add some more pre-processing and data cleaning steps. Overall, this project helped us apply the knowledge we gained in class, which was interesting and fun because we learned new things just by doing them.

Model	R^2	RMSE
Linear Regression	0.55	0.78
Linear Regression with parameter tuning	0.54	0.77
longitude, latitude, total rooms, and total bedrooms as input	0.30	0.98