Phase 1- Data Science and Machine Learning

Siddharth Tripathi

July 2020

1 Executive Summary

The Data set is the property data set of Auckland city New Zealand which was given already as a part of assignment of Phase 1. It contains different features related to property such as land area, rooms, suburb etc. The data describe various attributes of a property.

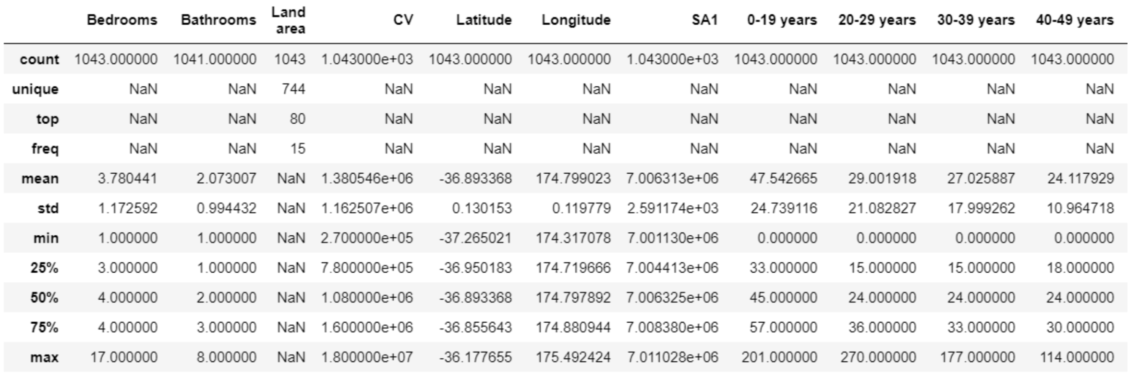
In total, there are 1051 records of different property which contain about 13 different features which describe various aspect of data. In our data set, CV column describes the price of each attribute (property). This is also our target variable in house price prediction. The other attributes are independent and help in target prediction. These are the explanatory variables about the property such as land area, rooms, suburb etc.

In the first part of assignment, we perform Data Collection and API calling. Two new columns are added in the original data set which are of census population 2018 and deprivation index 2018. The data set contains null values and duplicate records which will be handled in data cleaning and pre-processing. In the second part, data analysis and pre-processing is done on the data set obtained from part 1. Firstly, data is cleaned and pre-processed which includes handling null values, dropping columns and encoding of categorical columns. A correlation analysis is then performed on the final data set to find out which features affect the price most.

This report highlights the important observations and patterns from the data set. The Jupyter Notebook which contains the code for this assignment is more detailed. All actions are explained in details in our notebook via markdown cells.

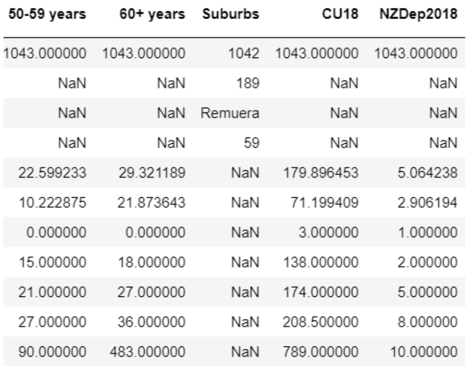
2 Initial Data Exploration

The data exploration begins with the statistical summary for all the records in a table (Fig 1).



	Bedrooms	Bathrooms	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years
count	1043.000000	1041.000000	1043	1.043000e+03	1043.000000	1043.000000	1.043000e+03	1043.000000	1043.000000	1043.000000	1043.000000
unique	NaN	NaN	744	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	80	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	15	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	3.780441	2.073007	NaN	1.380546e+06	-36.893368	174.799023	7.006313e+06	47.542665	29.001918	27.025887	24.117929
std	1.172592	0.994432	NaN	1.162507e+06	0.130153	0.119779	2.591174e+03	24.739116	21.082827	17.999262	10.964718
min	1.000000	1.000000	NaN	2.700000e+05	-37.265021	174.317078	7.001130e+06	0.000000	0.000000	0.000000	0.000000
25%	3.000000	1.000000	NaN	7.800000e+05	-36.950183	174.719666	7.004413e+06	33.000000	15.000000	15.000000	18.000000
50%	4.000000	2.000000	NaN	1.080000e+06	-36.893368	174.797892	7.006325e+06	45.000000	24.000000	24.000000	24.000000
75%	4.000000	3.000000	NaN	1.600000e+06	-36.855643	174.880944	7.008380e+06	57.000000	36.000000	33.000000	30.000000
max	17.000000	8.000000	NaN	1.800000e+07	-36.177655	175.492424	7.011028e+06	201.000000	270.000000	177.000000	114.000000

Figure 1: Statistical Summary - Properties data set



50-59 years	60+ years	Suburbs	CU18	NZDep2018
1043.000000	1043.000000	1042	1043.000000	1043.000000
NaN	NaN	189	NaN	NaN
NaN	NaN	Remuera	NaN	NaN
NaN	NaN	59	NaN	NaN
22.599233	29.321189	NaN	179.896453	5.064238
10.222875	21.873643	NaN	71.199409	2.906194
0.000000	0.000000	NaN	3.000000	1.000000
15.000000	18.000000	NaN	138.000000	2.000000
21.000000	27.000000	NaN	174.000000	5.000000
27.000000	36.000000	NaN	208.500000	8.000000
90.000000	483.000000	NaN	789.000000	10.000000

Figure 2: Statistical Summary - Properties Data Set

3 Correlation and Relationship

A correlation and relationship analysis was done using Seaborn library. A heatmap was produced depicting the correlation and relationship of all the columns with each other (Fig 3). The color code is explained by the scale on the right side of table. Dark spots mean negative high co-relation while light spots mean positive co-relation.

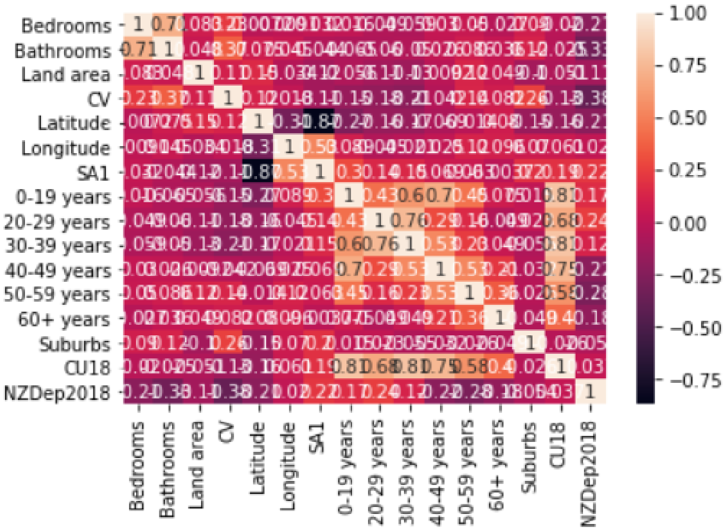


Figure 3: Heatmap - Correlation and Relationship

4 Analysis

4.1 Patterns in Data

From the above co-relation heat-map if we see co-relation of different features with our target variable CV, we can estimate that which features are the important features which influence the price of property. The negative co relation represent a inverse relation meaning that means if value for feature increases, the CV or price will increase and vice-versa.

4.2 Inferences And Analysis from Data

From our correlation table, different inferences about our data and pattern in it can be made. Some of these inferences which I observed are as follows -

1. **Bedrooms/CV**:The co-relation between Bedroom and CV is 0.227730. The positive correlation means that more the number of bedrooms, more the price of property which is an obvious thing.
2. **Bathrooms/CV**: The co-relation between Bathrooms and CV is 0.371953. Again, positive correlation represents a directly proportional relationship. More bathrooms more the price. Two bathrooms between 4 rooms is more convenient compared to one bathroom between 4 rooms.

3. **Land Area/CV**: The co-relation between Land Area and CV is 0.107736. This means that both are directly proportional and more the land area, more is the price.
4. **Latitude/Longitude/CV**: The co-relation between them and CV is positive and negative but these factors do not have a influence on the prices of any property. Hence while fitting regression models on this data set, these columns can be dropped.
5. **0-19-..60years/ CV**:For all the age groups also, there are different values of co-relation(mostly negative). Age groups have no significant influence on prices of house. However, one observation from this data set is that if an area has high population of people below 40 years, the prices of properties are low. This is explained by the negative correlation (except for 50-59 yrs). The possible explanation for this is that young people have lower salaries and lesser savings compared to those in older age. Therefore, their population would be higher in areas which have low property prices.
6. **Suburbs/CV**: The co-relation between Suburbs and CV is 0.257042 which means that the suburb in which property is located effects the prices of it. And this is true observation in real world as well. Some properties are costly because they are located in a posh suburb
7. **CU18/CV**: The census population shows a negative correlation of -0.125246. This indicates that rise in population will decrease the price of property.
8. **NZDep18/CV**: The deprivation index is another important feature which influences the price of properties. The NZDep18 has a negative correlation of -0.378413. Higher the deprivation index of a property, lower will be the price of the property.

Above are some observations about data pattern and information from the correlation and relationship of attributes with the target variable CV - Price of property.

The Machine learning mode which is implemented on our data set is Linear Regression since the target variable is not categorical but continuous data. The data set is first divided into training and testing data in the ratio of 80% data training and 20% data testing.

Based upon our correlation heat map and matrix, all the attributes having correlation lower than 0.20 are dropped. The features left are - Bedrooms, Bathrooms, Suburbs, NZDep18 and 30-39 years. Upon fitting the model prediction score for the model is calculated in order to analyze the performance of our model. This score comes out approximately 28% which is very low. Even after multiple iterations and pre-processing, a maximum of 45% score could be achived. This indicates that either the Machine Learning Model we chose for our data set is not appropriate, or data needs to be cleaned/pre-processed even more to remove discrepancies.

5 Conclusion

In conclusion, it can be said that correlation matrix and heat maps can help us analyze various patterns in our data set. They can also help to identify key features influencing our target variable.