**California Housing Data**

ADS1001 Semester 1,2023 Project Report – Group 3 Monday

|  |  |  |
| --- | --- | --- |
|  | Name | Monash ID |
| 1. | Tan Zhi Xian | 32795718 |
| 2. | Nithik Vijayanand | 33890285 |
| 3. | Nenja Ivanovic | 33890250 |
| 4. | Sanugi Fernando | 33767130 |
| 5. | Stefano Nicholas Rusli | 33511284 |
| 6. | Chung Ling Lai | 33067856 |

**Part 1 Description of the project:**

The California housing dataset consists of information regarding 20639 samples based on longitudes and latitudes from the 1990 Census blocks. Based on these identifiers, these samples are linked to a range of residential characteristics in California. With regard to our main objective in our project analysis, we were able to go through a variety of project aims due to the large spectrum of questions we could answer from this dataset containing highly relevant variables. We ultimately decided our main question to be 'Does attributes or location have the most impact when determining the median house value? Therefore, our hypothesis for this analysis is that location holds the most influence when determining the median house value. The first step in this analysis is to clean the data by removing outliers and missing values during the preprocessing. In order to better visualize our results, we will examine the different relationships between the variables using heatmaps and various statistical graphs to illustrate their correlations. Thereafter, we will be doing a comparison of the relationship between attributes and the location variables to ultimately decide which has the higher correlation in relation to Median House value. Formulating a direct response for our project aim will allow us to specifically find out whether locations play a major role in determining the house value, which is quite beneficial since such an assumption is quite common in the real estate industry. As a result, many people may be able to gain a better understanding of the key factors needed for their ideal home and whether the house is worthy of purchasing.

­

**Part 2 Details of preprocessing and manipulation of data in Python**

**READING:**

All the information provided on California housing was compacted into a single file, resulting in a huge data frame of 10 columns and more than 20,000 rows. It was immediately evident in the initial inspection of this data that thorough rudimentary pre-processing would be necessary prior to any analysis or modelling. The dataset had to be cleaned holistically because the broad nature of our research question made it difficult to predict which parts of it would be relevant to the investigation in the future.

**MISSING VALUES:**

The pd.info() function was used to evaluate the distribution of missing values within the raw dataset. It revealed that only the ‘total\_bedrooms’ column contained nulls [Fig. 1], having just 20,433 non-null values where the other columns had 20,640. Originally, it was believed that separating this incomplete column from the rest of the data set was the best option because it would preserve the most rows of data. The idea was to use the column-free set for most of the analysis when the ‘total\_bedrooms’ variable would not be very important, and the raw set with removed rows when it was.

However, it was realized that the effort required to navigate two parallel data frames for the rest of the project was not worth a few rows in the context of tens of thousands. In the end, 207 rows were simply dropped using pd.dropna(), indices were reset, and this data was named ‘raw\_data\_without\_nulls’.

**OUTLIERS:**

With the deficient entries removed, it was possible to begin dealing with outliers. With the intention of getting a preliminary sense of the distributions of all the variables as well as the general location and quantity of each of their outliers, a rough comparative boxplot was created. This plot had many limitations, particularly that all of the variables were plotted on the same vertical axis although they were measuring different things [Fig. 2]. Primarily, the values in ‘median\_house\_value’ were generally in the hundreds of thousands range, so it was difficult to properly see the boxplots of columns like ‘housing\_median\_age’ which only contained floats less than 100.

Nevertheless, it was clear that certain variables had lots of outliers, and that all of them occurred above the upper fence. These were more suitably replotted, and all the rows in ‘raw\_data\_without\_nulls’ with one or more data points above the corresponding upper fence were removed such that the data remaining had no outliers in any column, and no new missing values were created [Fig. 3]. To achieve this, 6,363 rows needed to be eliminated. Although this was a substantial loss, it was a vital step in ensuring that all conclusions drawn from the data were consistent with each other and as accurate as possible. The set of 14,070 outlier-free rows was named, ‘raw\_data\_without\_nulls\_no\_outliers’.

**FINAL TOUCHES:**

A few additional measures were taken to make analysis and modelling more convenient. All the rows containing outliers were recorded in an independent data frame for reference (‘data outliers’), the categorical ‘ocean\_proximity’ data was encoded as float values [Fig. 4], and appropriate columns were standardized. Each variation of data was split into standard testing and training groups [Fig. 5], but more elaborate divisions were avoided at this early stage in the project.

**Part 3 Summary of exploratory data analysis and significant conclusions**

The analysis aimed to determine whether the attributes of a house or its location played a more significant role in determining its price. We first began by examining the distribution of house age, which revealed that most houses fell within the 30 to 40-year range. Interestingly, there was a notable concentration of houses with an age capped at 52, indicating that some older houses may have been recorded as 52 years old. However, no correlation was found between house age and house value.

Turning the focus to location, we found the impact of ocean proximity on housing prices. Houses situated on islands were found to have the highest median housing value, likely due to their higher minimum value. However, the maximum housing value for island houses was lower compared to other areas. This suggests that while location influences housing value, other factors may also come into play.

Analyzing the distribution scatter plot, it was observed that as the age of existing houses increased, new houses were being built farther north and west. This shift in housing locations indicated a trend of expanding residential areas to accommodate the growing population. The analysis proceeded by categorizing houses into age groups, with each group spanning ten years. Most houses were found to fall within the 31 to 40-year range. However, no significant correlation was found between housing age and location, implying that other factors have a stronger influence on location determination.

Examining average house values, regression plots indicated that the values varied between $160,000 and $200,000 across different housing ages. Additionally, urban regions, specifically areas near the bay and the ocean, exhibited the highest average income levels compared to other regions. Consequently, urban areas also displayed higher average housing values compared to suburban areas. Analyzing region-specific graphs, no correlation was found between housing value and location within urban, suburban, and rural regions. This suggests that the relationship between location and housing value may not be straightforward and can be influenced by additional factors beyond region classification.

Following this, although areas near the ocean demonstrated relatively high median house prices, locations slightly further from the coast had significantly lower values. This finding indicates that the location of a house plays a substantial role in determining its value.

**Part 4: Summary of any undertaken modelling and any significant conclusions**

Polynomial regression modelling was applied for 5 different ocean proximity-related locations.

Initially, the variables were updated by filtering the data. After that, additional data frames were made for each category in order to be able to drop the ‘location’ and ‘latitude’ columns, ultimately removing redundancies of location datasets. Next, polynomial regression models were used in the code using the ‘polyfit' function. The models are fitted separately for the median income and median house value relationship, as well as the housing median age and median house value relationship. Lastly, the results were printed in the form of line equations best fitted for each location and each relationship. The equations offer the coefficients for the polynomial regression models, illustrating the connection between the variables.

All the given lines of best ft equations for the median income and median house value relationship have a positive correlation. According to the value of the coefficients (value multiplied by x), ‘near bay’ is the most correlated and ‘islands the least correlated. For the housing median age and median house value relationship, ‘<1H ocean’ and ‘inland were the only 2 locations which were negatively correlated, with ‘near ocean’ being the most correlated and '<1H ocean’ being the least. Overall, both ‘near bay’ and ‘near ocean’ have the greatest relative correlations between the two relationships, and ‘inland’ has the least.

**Part 5: Conclusions in the relation to the original problem**

Throughout the California housing data, there are many different variables which influence the median house value in each longitude and latitude region. However, the two primary variables which ultimately determine the price of a house are based on the attributes of a house (such as the number of rooms within the house or how old the house is) and the location of the property. Through our data modeling and visualization, we identified that the location of the house was the key factor when determining the price of a house, since there was a slightly higher level of correlation between our location variables, for example, a house located on an island or close to the ocean/bay had a significantly higher median house value than the house located in regions in land or over an hour from the shoreline. Therefore, our null hypothesis, which states the location holds a higher level of influence over the median house value, was proven to be correct.

Our analysis process was subject to a few inaccuracies as the large quantity of data failed to prove much correlation between our variables. A solution to tackle the issue for future investigations is to analyze regression based on a smaller quantity of households. By examining several random longitude and latitude locations of fewer households, there may be higher levels of correlation, which could better explain our main question.

# **Figures**

# 

# Figure 1 Raw Housing Data Information

# A graph with numbers and lines Description automatically generated with medium confidence

# Figure 2 Initial Box and Whisker Plot for Outlier Check

# A chart of a number of data Description automatically generated with medium confidence

# Figure 3 Final Box and Whisker Plot

# A table with numbers and letters Description automatically generated

# Figure 4 Binary Encoding of ‘ocean\_oroximity

# 

# Figure 5 Summary of Pre-Processing Outcom

# **Bibliography**

# [Kechit Goyal (Jul 15, 2021). Data Preprocessing in Machine Learning: 7 Easy Steps To Follow. upGrad.](https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/)