California Housing: A Data-Driven Analysis
Group 3

ADS1001 Semester 1 2023
Final project report

	Name, Surname	Monash ID	Contribution (%)	Contribution (description)
1.	Byron, Shim	33903638	16%	- Households and Population Analysis - Presentation Slides - Report
2.	Benjamin, Hall	33890331	20%	- Cleaning and removing outliers - City Proximity Analysis - Bedlessness Analysis - Presentation Slides - Report - Poster
3.	Deslyn, Mulio	33368066	16%	- Presentation Slides - Poster - Island blocks' Low income and High House Value analysis
4.	Batheendra, Nanayakkara	33890471	16%	- Merging Python Notebooks - Median Income and Median House Value Analysis - Report - Presentation Slides
5.	Simin, Liang	33499179	16%	- Median Income and Median House Value Analysis - Presentation Slides - the relation between population density and median house value -
6.	Thi, Do	33949891	16%	- Presentation Slides - Poster - Report - Island blocks' Low income and High House Value analysis

Part 1 Description of the project:

The project explores the California housing data, which contains information from California housing blocks as of the Californian census in 1990. There is a variety of both numerical and categorical data present. For each block, the numerical data includes 'longitude' and 'latitude', the 'median house age', the 'total rooms', 'total bedrooms', the 'population' and number of 'households' as well as both the 'median income' and 'median house value'. The categorical data provides an indication of the proximity of the block to the ocean, with each block being labelled as either 'inland', 'near ocean', 'near bay' or 'island'. Some of the objectives of the group project include the cleaning of data and the removal of outliers, while taking on some exploratory data analysis. The correlations between variables will also be investigated, such as whether or not proximity to the ocean or cities tends to affect house price and number of homeless people. Furthermore, graphics were created which illustrate the geographical distribution of house prices and house age around the state.

Part 2 Details of preprocessing and manipulation of data in Python

The Californian housing data contained 20640 rows by 10 columns. To prepare the data for analysis, it was important to remove outliers and filter the data to only include usable data values. The data was initially checked for any NaN values, which are any missing values in the dataset. It was found that there were 207 missing values under the total bedrooms variable, and that every other variable had zero missing values and was usable. The rows which contained these missing values of data were then removed from the dataset so the data could be analysed. The categorical data, 'ocean proximity' was then analysed for any anomalies. This was done by using the code housing ['ocean_proximity']. value_counts(). This showed how many blocks are located 'inland', 'near ocean', 'near bay' and 'island'. All values appeared to be fairly normal apart from the clear outlier of island blocks, which were only 5 of out of the 20433 remaining blocks. This number was dwarfed by the other ocean proximity values, which were 2270 near bay, 2628 near ocean, 6496 inland and 9034 which were less than an hour from the ocean. However, only having a small amount of island blocks was expected. Given the magnitude of the dataset, it was expected that there would be a large number of outliers among the range of variables. To identify outliers, we first identified if there were any unique values for the whole data frame. Once we found that there were no unique values we created a unique value for each row to enable merging of data sets. We then created two new variables to help us identify outliers. These new variables were "population density" which was created by dividing "population" by "households", and "bedless", which was created by dividing "population" by "total bedrooms" . We then calculated the upper and lower fences and plotted the variables using box plots. Any rows containing values lying above the upper fence or below the lower fence were removed from the dataset. However, at this point it was observed that over 5% of the original data had been removed due to the presence of outliers. This created a difficult situation where not every row with an outlier could be removed as too much data would be lost. Therefore, we only removed outliers that were 1.5 times above the upper and below the lower fences from variables that were used in our analysis. Furthermore, distance to the closest city was also calculated as it was an important part of our analysis. First a new dataset was created featuring the longitude and latitude of major cities in california, the longitude and latitude were found through google maps. Then, the euclidean distance was calculated between each block and each city and the minimum distance was appended to the housing dataset. In hindsight, it would have been useful to append the name of the closest city to the dataset as well.

Part 3 Summary of exploratory data analysis and any significant conclusions

Out of the 20640 rows of data, 207 rows contained NaN values. Once these were removed, we then removed further outliers leaving 19543 data points left. Overall, we removed only 5.4% of our raw data. To begin with our exploratory analysis, we first looked at the relationship between “median income” and “median house value”. To present the data in a way which showed geographical location, the latitude and longitude were plotted on a scatter plot, with longitude on the x-axis and latitude on the y-axis. This presented a scatter plot with thousands of blue dots of the same size spreading across the plot with the majority lying along a band from the top left corner to the bottom right, with each dot representing a different block. From this graph it was clear where the California coastline was, but just seeing the locations of the blocks gave a very limited insight into the relationships between other variables and the locations of the blocks. The next step was to differentiate between high density block locations and low density block locations. To do this, each point was made more transparent by adding alpha=0.1 into the scatterplot code. This meant that regions of the state containing a very few number of blocks were quite a faint bluish colour and high density areas were a more solid, darker blue colour. The next step was to integrate other variables into the plot. By coding the size of the circles to represent the 'median income' and the colour of the circles to represent the 'median house value', we could visually see the distribution of the median house values. We then used this same approach for the “population” and “median house value”. For these variables we coded the size of the circle to represent the “population” and the colour to represent the “median house value”. From this we saw that the more populated areas generally have a higher “median house value”.

Next we investigated the median house value in relation to proximity to the city. From this we found that the median house price generally increases as the distance to the closest city decreases. House price generally increases as the distance to the closest city decreases. We also found that (with a correlation coefficient of 0.748) for every 1 bedroom increase there was an increase of 3 people. From our background research we learnt that California had a big homeless issue. As a result, we investigated the issue of Inadequate amount of bedrooms in California. We figured that when the ratio of people to bedrooms in a block exceeded 3:1 that there was a significantly higher chance of people being left on the streets rather than each room in each house in the block being the sleeping space of 3 people. Representing the frequency of this ratio visually meant that we could visually identify key areas of what we have dubbed “bedlessness”, Seeing that clusters of bedlessness occurred around city blocks we decided to investigate this relation. We found that as the blocks got closer and close to the city, there was a higher frequency of bedlessness and a higher maximum block ratio with a specific block in Los Angeles reaching a ratio of 18:1. Although we could see visually that a major bedlessness hotspot existed in Los Angeles, if we had appended the closest city name to the dataset we could have created a histogram that showed us the frequency of bedlessness in and around each of the cities in our dataset.

We also decided to investigate the 5 island blocks and we found some interesting insights. The mean "median_income" for all the blocks is $38723.60 while for the "Island" blocks, it is $27444.20. However, the mean "median_house_value" for all the blocks is $173,115.71 less than the mean for the "Island" blocks. So the "Island" blocks on average have a smaller "median_income" but a larger "median_house_value". Why is this? Well, we found that the mean "median_age" for the 5 "Island" blocks was 13.67 years older compared to the mean "median_age" of the entire data set. Therefore, we can make an assumption that the properties on the "Island" blocks were inherited or passed down.

Part 4 Summary of any undertaken modelling and any significant conclusions

Linear regression modelling was used to predict the median house value of a block based off median income, which was expected to have a positive correlation. It was found that on average, for each additional $10,000 made per year, house value increased by $41928.97 It was also found that on average, the median house value for someone with no income would be $44672.7, but as there was no one in the dataset with no income, this is extrapolation and thus an unreliable prediction. The testing score for the model was 0.472, suggesting that 47.2% of the variation in median house value can be explained by the variation in median income. Another regression model explored the relationship between the number of households in a block and the population of the block, which was expected to be strongly positively correlated. It was found that for each additional household on a block, the block's population was expected to increase by 2.51, suggesting the average occupancy of a Californian household is between 2 and 3 people. The testing score for the model was 0.755, suggesting that 7.55% of the variation in a block's population can be explained by the number of households in a block, which is a strong positive linear correlation.

Part 5 Conclusions in the relation to the original problem

In conclusion, this report has provided a comprehensive analysis of the California housing dataset, shedding light on various aspects of housing dynamics and related phenomena in the state. Our examination of areas with a higher population-to-room ratio has revealed concerning indications of potential homelessness. Particularly, the Los Angeles and San Francisco areas stand out as regions where a significant number of individuals either share bedrooms or lack bedrooms all together. The analysis of house prices in relation to ocean proximity has revealed that houses located on islands command significantly higher prices compared to those situated near the coast. This finding suggests that factors such as exclusivity, desirability, and limited supply contribute to the premium associated with island properties. The

examination of house age in the dataset has also shown that houses located on islands tend to have higher average ages compared to properties in other parts of California. This observation suggests that island properties may possess historical and cultural significance, contributing to their preservation and limited turnover. The higher house age on islands may also imply a higher demand for these properties due to their unique charm and limited availability. Taking into account these additional variables will allow us to gain a more complete understanding of how housing characteristics and socioeconomic factors are interconnected. Furthermore, our analysis included a regression examining the relationship between population and households, intending to predict the housing needs to accommodate future population growth. The results of the regression analysis indicate a positive and statistically significant relationship between population and households. This finding implies that as the population increases, the number of households also tends to rise proportionally. By operating this regression equation, policymakers and urban planners can forecast the housing requirements based on projected population growth. For instance, if the population is expected to increase by a certain amount, the equation can be employed to estimate the corresponding number of households needed to accommodate the growing population. This information is crucial for informed decision-making in housing planning, infrastructure development, and resource allocation.

References

List of largest cities in California by population. (2023, May 21). Wikipedia, the free encyclopedia. Retrieved April 30, 2023, from https://en.wikipedia.org/wiki/List_of_largest_cities_in_California_by_population#:~:text=List%20of%20largest%20cities%20in%20California%20by%20population,economic%2C%20polit%20...%20%2018%20more%20rows%20