California Housing Data Project

ADS1001 Semester 1 2023
Final project report

		Name, Surname	Monash ID	Contribution (%)	Contribution (description)
1.	1.	Emilia Zhang	33154309	20	Analysis on median income and median house value on different no. of people in households and organisation of report. Composing Part 5 Conclusion section of report.
2.	2.	Aden Siau	33890277	20	Composing Part 4 Modelling section of report. Analysis on relationship of block population and number of rooms on house price.
3.	3.	Hieu Nguyen	33936447	20	Analysis of house values depending on ocean proximity and respective household incomes, aiding in analysis of median income and median house value on various household densities.
4.	4.	Jaehong Kang	33890439	20	Composing Part 1 Introduction section of report. Analysis on relationship between household income, household density, and no. of rooms on house price.
5.	5.	Joshua	33890374	20	Analysis on relationship between median income and house age on house value. Composing Part 2 of report.

Part 1

Description of the project:

This project was an investigation into the California Housing Data that contains information from the 1990 California census. The data described houses found in a given California block, which generally summarised *location (longitude, latitude, ocean proximity)*, the general details of the houses in the block (*total rooms, total bedrooms, number of households, median house value, median house age*), and information relating to the people who live in the block (*population, median income*). The data was almost entirely numerical, except for one categorical variable which was ‘Ocean Proximity’. The overarching question the group posed was:

Which variables had the largest impact on the median house value of a block?

To achieve this analysis, the group undertook some exploratory data analysis to answer some sub-questions which aimed to investigate the correlation between different variables in the data. In order to analyse the California Housing, the data also had to be cleaned to remove any outliers and null values which would impact the results of our analysis. The dataset also had some limitations such as capped figures which required the group to use their knowledge of Python to overcome and interpret the data as best as possible.

Part 2

Details of cleaning, preprocessing and manipulation of data in Python

The first part of the investigation involved data manipulation and preprocessing of the California Housing Data. The data needs to be altered and cleaned so that proper analysis can be carried out. To begin with, an isna() function was run on the dataset to see if there would be any NaN values. Fig 1.1 presents that there were 207 entries in the 'total bedrooms' column of the dataset which had NaN values. The group cleaned the data by removing the rows which contained these 'NaN' values by using the dropna() function. This was done as the number of blocks with NaN entries was a small percentage of the total dataset, and calculations could not be completed without removing them.

In the preprocessing stage, the group calculated the average household population for each block. This was done by dividing the block population by the number of households. This formed a new variable 'household_average' which was then added back to the data frame. This new column then had its entries rounded to the nearest integer, as it would be more logical to have the number of people per household be a whole number.

Fig 1.2 shows that there were households with ridiculous numbers of people living in them. Therefore, the data was further cleaned by removing houses with a *household average of more than 10.* This was because it was concluded that regular households would most likely not have this many people living under one roof, and therefore these entries would likely have been apartments. There was also a general trend that these entries also had much lower median income, which skewed the data. As a result, by choosing to remove entries that fulfilled the above criteria, we could have a more accurate analysis of the housing data by more likely being able to focus on analysing landed properties/ low-density housing, such as suburban homes or smaller townhouses.

The group also added dummy variables for Ocean Proximity, as it is the only variable which is categorical. To do this, the 'get dummies' function was used, which added a 1 to the corresponding ocean proximity category for each block, and a 0 for the rest. This let us to work with all numerical data, which allowed for calculations in the analysis. We then merged the dummy variables into the original set by using the pandas 'merge' function, which was easily done as the variables were all correctly assigned to their specified ocean proximity.

We then decided at this point to not use the ocean proximity column for future analysis, as it being categorical would be difficult to be used in analysis, and it had essentially been replaced by the dummy variables, which can be more easily filtered out and sorted for our explanatory data analysis. However, we chose not to remove it although it would be unused for analysis, as it would be easier to visualise graphs that featured ocean proximity, and is actually easier to understand when not doing calculations.

Additionally, new columns featuring standardised values of the median incomes, house age, and house value variables were calculated and added to the data frame. This was achieved through using the zscore() function from the scipy library, which deducted the mean of the variable from each data point, and divided the result by the standard deviation of the variable. This calculation was also done by using lambda functions, but the former method would be less susceptible to human error. The standardised values (z-values) let us compare the variables better, as the variables had large variation between their figures, and as these variables would now be equally scaled, any future visualisation of the data would be neater.

Part 3

Summary of exploratory data analysis and any significant conclusions

Block population and median house price

The correlation between population and median house value was -0.0246. As correlation is between -1 and 1, this correlation is very weak, showing that there is a weak relationship between block population and median house price. The negative aspect of the correlation suggests that an increase in block population will lead to a decrease in median house price for a block. (Figure 3.1)

Further, from the investigation into the correlation between ocean proximity and median house value, the highest valued category for ocean proximity was ‘Near Bay’, followed by ‘Near Ocean’, ‘<1H Ocean’ and then ‘Inland’. On the other hand, the average population for each ocean proximity category was calculated. The results were almost the opposite, with the highest populated ocean proximity category being ‘<1H Ocean’, then ‘Inland’, ‘Near Ocean’, and then ‘Near Bay’. (Fig 3.2) This shows that the ocean proximities with lower population densities are worth more than the ones with higher ocean proximity, further supporting that population and median house prices are negatively correlated

Average house value and income for each region in proximity to the ocean

The highest average house value is $259097.08 in ‘Near Bay’ although the average income in ‘Near Bay’ is in second place, $41,700. ‘Inland’ is identified as a location having the lowest average house value and income which are $ 124863.96 and $32,100 respectively. While the average house value of ‘<1H Ocean’ nearly being the lowest one with a value of $240234.94, its average income is the highest at $42,300. ‘Near Ocean’ has an average house value valued at $249288.90 and an average income valued at $40,100. (Figure 3.3)

Highest median house value compared to average house value of each region

The highest median house value is $500,001 (as the data was capped at $500,000) which is almost double the average house value in ‘Near Bay’, ‘<1H Ocean’, and ‘Near Ocean’. It has a huge difference from the lowest average house income, ‘Inland’, which is around four times. The highest number of households in every income level all choose to live in ‘<1H Ocean’ in medium house value, it might be due to the high income. The second-highest number of households is ‘Inland’ with most of the households having low income and living in low house value. (Figure 3.4)

What is the average median household income for each no. of people in a household?

As shown in *Figure 3.5,* the average median income per person in households was found by creating separate data frames and isolating each numerical value of ‘household_average’ from the cleaned and pre-processed California housing data. With each data frame containing data based on an average number of people in each household, the average median income could be found (rounded to the nearest dollar):

1 people households: $27,058

2 people households: $36,543

3 people households: $42,449

4 people households: $32,276

5 people households: $27,989

6 people households: $31,731

Thel boxplot in Figure 3.7, contains all the median incomes from every household with up to 8 people, household of 3 are predominately contain the wealthiest households.

Furthermore, the average income in California is $33,719, thus, the following conclusions were made.

Households with an average of 2-3 people have the highest average median income and is over the California average income. This data most likely reflects stable families/couples that are highly educated and have high income jobs. 2-3 people households are the most frequent population per household in the entire data, further depicting that most of the people in California are well off.

1 people household have the lowest average median income, portraying that out of those that live by themselves, most are low-income earners, such as students.

4-6 people households all have an average median income below the California income. It is likely that this data is impacted by families where the parents are not highly educated. Poor education leads to less opportunities for high income jobs, which correlates to the below average median income for these families. Furthermore, poor education leads to less knowledge about reproduction, which would result in unwanted pregnancies, which would represent why families of larger proportions have low average median income.

Ultimately, it is concluded that there are other external factors that are not outlined in the overall California Housing Data that affect both variables, therefore, the 2 variables are not reliable to predict each other's. This is further proven by modelling of median income against household average without the rounding, and by binning the household averages.

What is the median house value for each no. of people in household?

1 people households: $187500

2 people households: $218450

3 people households: $186100

4 people households: $149200

5 people households: $137500

6 people households: $157500

Median house value for each no. of people per household was found by grouping median house value values (seen in figure 3.8) into their household averages and calculating their median. The median was used instead of the mean as a household of 6 people had a lower bound outlier which disproportionately negatively skewed as it pulls the mean down, making its average value lower than its actual centre (seen in figure 3.9).

Looking at Figure 3.8, households of 2 people have the highest median house value followed by 1 and 3-people households. The lowest being 5 people households followed by 4 people households then 6. We can compare this to the average income of the respective households to understand how income in different households can affect the house value.

Whilst 2-3 people households have the highest income, it is interesting to see here that 1 person households have the second highest median house value whilst having the lowest average income (seen in Figure 3.6). Looking at the boxplots for median house value for household averages, we see that a household average of 1 person does not have any outliers meaning that the average is a good approximation of the average median house value. One possibility to consider is that the 1-person households may be renters, therefore afford to live in a house on lease with a 1-person's income. However, there isn’t enough data to confirm this theory.

Compared to the median house value average of the whole dataset of $206886 USD, we can see that only the 2 people's household median house value exceeds the population mean. The other groups all fall under the population mean with the highest being 1 person household with a median house value of $187500 USD.

Relationship between median income and house value.

By creating a regression linear plot of the two variables, we can clearly see that median income and median house value have some sort of linear relationship. This can be explained by the correlation coefficient of approximately 0.6895984… The correlation coefficient displays a moderate positive linear relationship between median income and median house value where median income increases by $1 USD, then median house value generally increases by $0.69. Therefore, the exploratory variable median income is a significant variable to consider that influences median house value.

Relationship between no. of rooms and house value.

To find the correlation/relationship between the explanatory variable ‘total rooms’ and response variable ‘median house value’, a linear model needed to be created: with a coefficient of 22619.115, as seen in figure 4.7, depicting that as the number of rooms increase, the house value increases by $22,619.12. However, figure 3.# shows that the correlation score was 0.1334, this presents that the 2 variables have a very weak relationship.

To ensure this finding was viable, another variable was created: average rooms per household, which was found by dividing total rooms in each region by the number of households. This new column showed that houses with 4-7 rooms were the most frequent. When finding the correlation between average rooms per household and median house value, the correlation score was 0.1515, which is close to the relation between total rooms and median house value.

Moreover, it was interesting to find that there were regions with average rooms of over 100 in a household, as shown in figure 3.13. This showed that there were some apartment building that we missed, thus, I only focused on regular houses. This was to drop average houses with more than 8 rooms. After finding the relation between the pre-processed average room per household data and median house value, the correlation was 0.2287.

The data conclusively showed that the number of rooms in the region or per household in the region does not highly impact median house value, as there are other external factors that impact.

Relationship between median household value and age of house.

Before analysing the relationship between these two variables, the other variables in the dataset were removed to control the analysis and prevent other variables from influencing the relationship between household value and house age. It was then discovered that these two variables had non-symmetrical distributions, with the former having a positively skewed distribution and the latter being irregular (Figure 3.14). These variables were then modified by standardising them, as well as by squaring and cubing their values (Figure 3.15). The cubing transformation made the household value and age of house variables symmetric, and thus this transformation would be used to analyse their relationship. However, this did not change the relationship between them, as the correlation coefficient stayed at 0.11 (Figure 3.16), which was the same as their unmodified versions. Thus, it can be concluded that there is a very weak positive linear relationship between household value and house age.

Relationship between household income and house age.

To minimise external factors influencing the two variables, the dataset was again modified by removing the other variables not involved in the analysis. There was a very weak relationship between the two variables, and doing the same procedures as the ones done to find the relationship between house value and house age (ie. transformation of variables), their relationship failed to change significantly, and thus these two variables on their own have a very weak correlation with each other (Figure 3.17). It was discovered however, that household income had a very strong relationship to house value, both before and after transforming the variables given, with their correlation coefficient remaining at around 0.7. Consequently, the dataset was split between homes that were below the maximum age of a given house in the dataset, 52 years old, and those only aged 52 years old, to see if they shared any difference in their relationships with the other variables, including house price. It was then found that these 2 datasets were actually very similar, with both having similar distributions of household income (Figure 3.18), as well as similar mean/median values for both age groups (Figure 3.19). It can be concluded that house age is not a significant factor in influencing house price, and reinforces the relationship between household income and house value due to them not being affected by age.

Part 4

Description of modelling

Predicting real Median House Value of capped blocks

Linear Model:

Predicted Median House Value = -24432(longitude) + -2.2571(latitude)+ 931(housing median age) + -7 (total rooms) + 87 (total bedrooms) + -33 (population) + 54 (households) + 38343 (median income) + -24504 (<1H Ocean) + -63484 (Inland) +140598 (Island) + -3155 (Near Bay) + -21059 (Near Ocean) -2062340

The data was capped at $500000 for blocks which had a median house value that was more than $500000. The group decided that this would inaccurately represent the housing value in California so a multiple linear regression was created to predict the real median house value for blocks at $500000 in the original data. The model used all variables to predict the median house value (Fig 4.1). To do this, the model was trained on a data frame that contained all blocks with a median house value that was not equal to $500,000 (Fig 4.1). Then it was applied to a data frame that contained all blocks with a median house value of $500,000 or more (Fig 4.2). The model predicted the real uncapped median house value of the blocks and these values were added to the corresponding block (Fig 4.2, 4.3). However, many of the predicted median house values were less than $500,000 (75% was below $454766) (Fig 4.4), which would be incorrect as blocks with a median house value at that price in the original data would have been more than or equal to $500,000. Therefore we decided to rule out the modelling as it would have made the data more inaccurate. The attempts to fix this issue included I using only the X variables which had the strongest correlation with Median House Value as well as training the model using the entire original data set instead of just the non-capped blocks. However, both of these changes made marginal improvements, none of which were significant enough to apply.

—------

Linear model: y = -0.133x + 4.259

Initially, the approach to find if household average (explanatory variable) affects median household income (response variable) was taken to observe whether or not, the 2 variables can be used to predict the other. However, upon observation of the data, there is a very weak relationship between the two, with a correlation of -0.057, as shown in Figure 4.5 For every people per household, the spread of median income per region was greatly spread, thus a linear regression model was not suitable. This is further supported by the testing score of 0.004, as shown in Figure 4.6.

Linear model: y = 22619.115x + 84079.691

Although it is seen in Fig 4.7, that households with more rooms tend to have greater value, the data is too spread, meaning that the predictability of the linear model is very weak. This is further supported by the testing score of 0.059, presented in Fig 4.8.

Part 5

Conclusion to analysis

In relation to what variables have the most significant influence over median house value, it appears that lower population density regions are worth more than regions with higher population density. We can also expect a positive correlation between median income and median house value as generally in inland regions we have a lower house value and lower median income. In contrast, regions near the bay and ocean generally have higher house values and income.

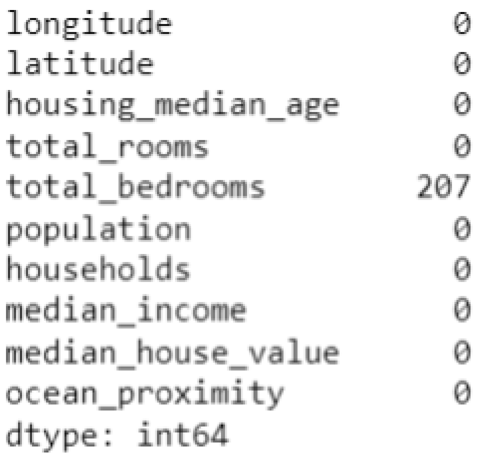
It is inferred that households of 2-3 people generally earn more and have higher average incomes as well as high-valued houses which reflects a stable and comfortable lifestyle for Californian families/couples. However, it is interesting to note that despite having the lowest average income, 1-person households also have the second highest median house value. This contradicts the idea that a higher median income leads to a higher median house value. However, there is a possibility that some 1 person households are tenants living on lease which can explain how they may manage to live in a property with a high house value with low income.

In response to the main problem, the variables that had the most significant impact on median house value were ‘Ocean Proximity’, ‘Median income’ and ‘Population Density’. These exploratory variables indicated clear patterns in impacting median house value that was also minimally influenced by external variables that aren’t addressed in the database.

For future notes, more cleaning involving filtering the data would assist in an accurate analysis of what variables affect the median house value. Also, creating and comparing different exploratory variables that are unaffected by the effects of other variables and significantly impact the median house value would give a better visualisation of the effect of specific exploratory variables on the response variable. Finally, something to consider for future projects would be developing a deeper perception of the data’s background which is essential for producing in-depth answers to relevant questions that clients may hold.

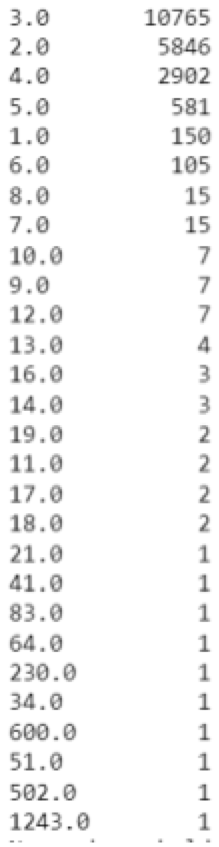
Figures

Part 2 - Details of cleaning, preprocessing and manipulation of data in Python



```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        207
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

Fig 1.1: part2.Details of cleaning, preprocessing and manipulation of data in Python



```
3.0       10765
2.0        5846
4.0        2902
5.0         581
1.0         150
6.0         105
8.0          15
7.0          15
10.0          7
9.0           7
12.0          7
13.0          4
16.0          3
14.0          3
19.0          2
11.0          2
17.0          2
18.0          2
21.0          1
41.0          1
83.0          1
64.0          1
230.0         1
34.0          1
600.0         1
51.0          1
502.0         1
1243.0        1
```

Fig 1.2 - Part 2.Details of cleaning, preprocessing and manipulation of data in Python

Part 3 - Summary of exploratory data analysis and any significant conclusions

Figures:

Block population and median house price

Figure 3.1

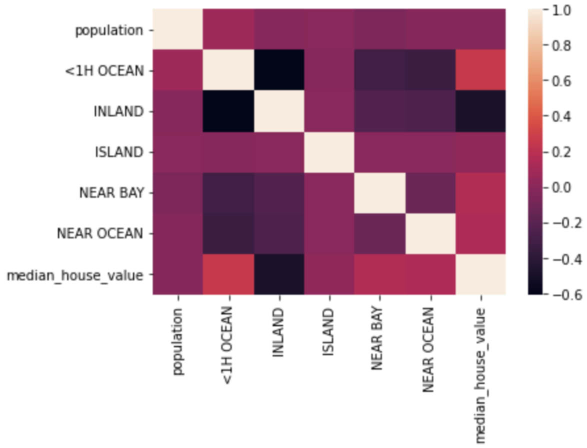
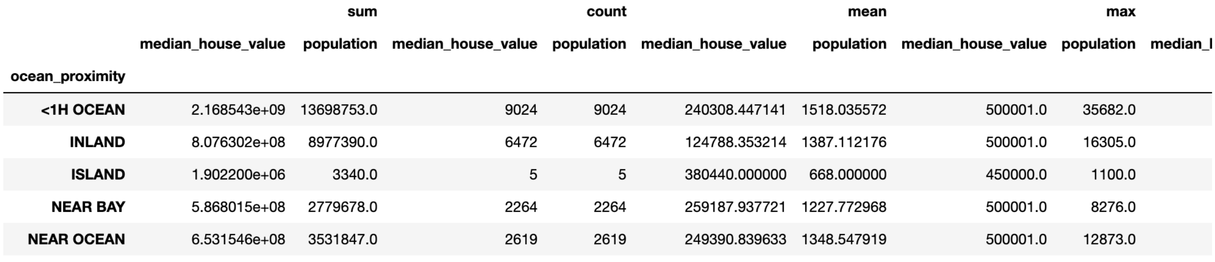
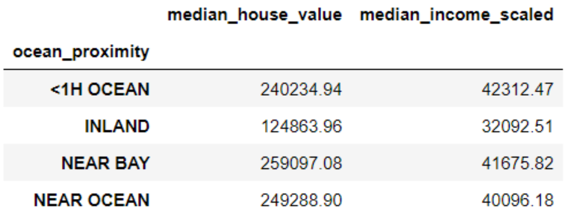


Figure 3.2



	sum		count		mean		max		
	median_house_value	population	median_house_value	population	median_house_value	population	median_house_value	population	median_l
ocean_proximity									
<1H OCEAN	2.168543e+09	13698753.0	9024	9024	240308.447141	1518.035572	500001.0	35682.0	
INLAND	8.076302e+08	8977390.0	6472	6472	124788.353214	1387.112176	500001.0	16305.0	
ISLAND	1.902200e+06	3340.0	5	5	380440.000000	668.000000	450000.0	1100.0	
NEAR BAY	5.868015e+08	2779678.0	2264	2264	259187.937721	1227.772968	500001.0	8276.0	
NEAR OCEAN	6.531546e+08	3531847.0	2619	2619	249390.839633	1348.547919	500001.0	12873.0	

Average house value and income for each region in proximity to ocean



	median_house_value	median_income_scaled
ocean_proximity		
<1H OCEAN	240234.94	42312.47
INLAND	124863.96	32092.51
NEAR BAY	259097.08	41675.82
NEAR OCEAN	249288.90	40096.18

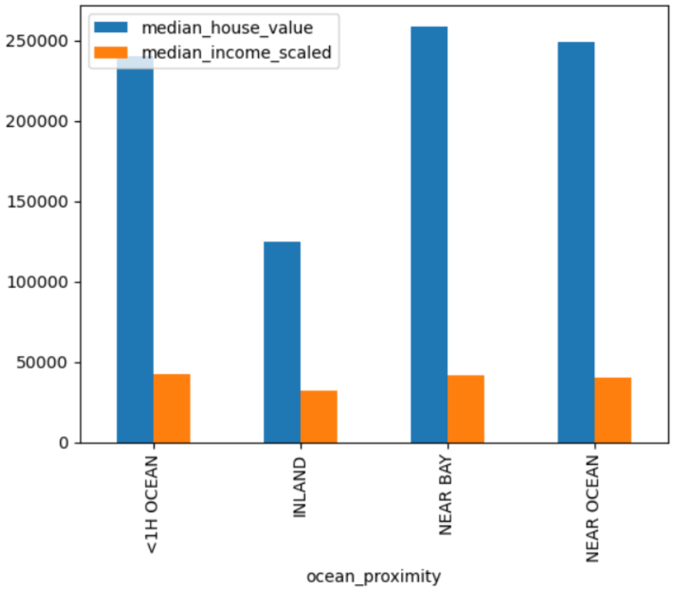
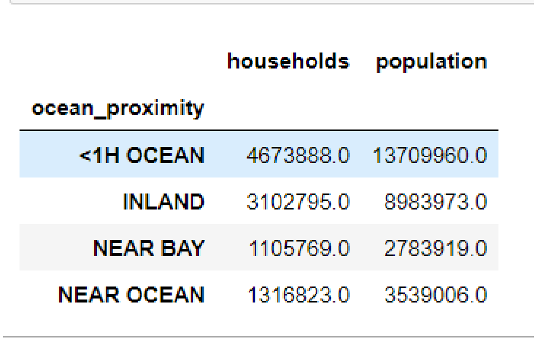


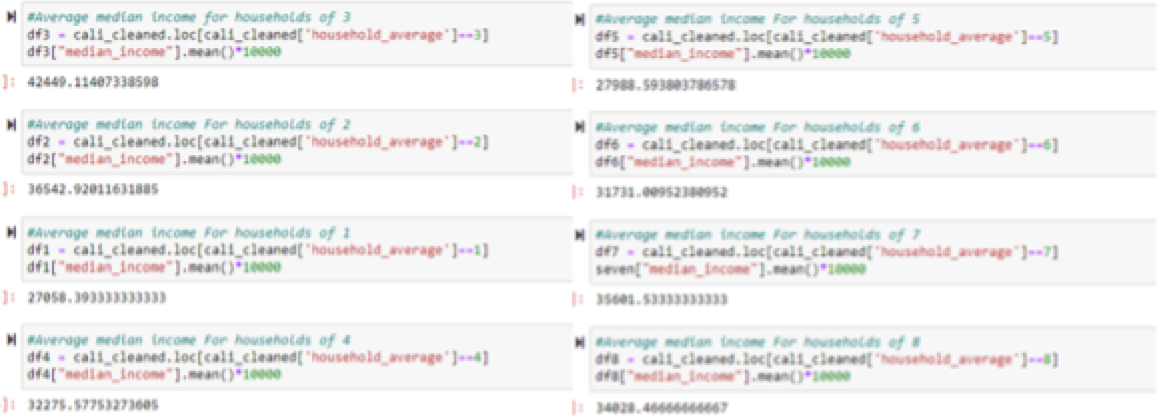
Figure 3.3

Highest median house value compared to average house value of each region - What type of income levels or families want to live in different ocean proximities?



	households	population
ocean_proximity		
<1H OCEAN	4673888.0	13709960.0
INLAND	3102795.0	8983973.0
NEAR BAY	1105769.0	2783919.0
NEAR OCEAN	1316823.0	3539006.0

Figure 3.4



```
#Average median income for households of 3
df3 = cali_cleaned.loc[cali_cleaned['household_average']==3]
df3["median_income"].mean()*10000
```

```
]: 42449.11407338598
```

```
#Average median income For households of 2
df2 = cali_cleaned.loc[cali_cleaned['household_average']==2]
df2["median_income"].mean()*10000
```

```
]: 36542.92011631885
```

```
#Average median income For households of 1
df1 = cali_cleaned.loc[cali_cleaned['household_average']==1]
df1["median_income"].mean()*10000
```

```
]: 27058.393333333333
```

```
#Average median income For households of 4
df4 = cali_cleaned.loc[cali_cleaned['household_average']==4]
df4["median_income"].mean()*10000
```

```
]: 32275.57753273605
```

```
#Average median income For households of 5
df5 = cali_cleaned.loc[cali_cleaned['household_average']==5]
df5["median_income"].mean()*10000
```

```
]: 27988.593803786578
```

```
#Average median income For households of 6
df6 = cali_cleaned.loc[cali_cleaned['household_average']==6]
df6["median_income"].mean()*10000
```

```
]: 31731.00952380952
```

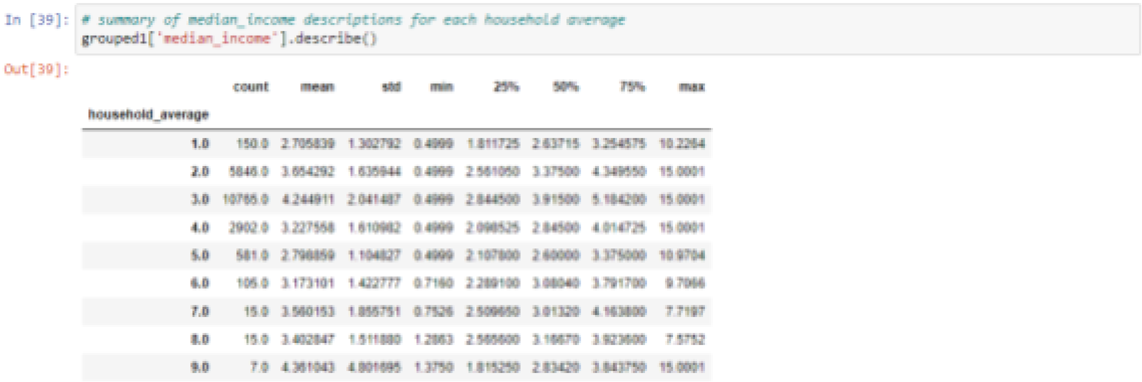
```
#Average median income For households of 7
df7 = cali_cleaned.loc[cali_cleaned['household_average']==7]
seven["median_income"].mean()*10000
```

```
]: 35601.53333333333
```

```
#Average median income For households of 8
df8 = cali_cleaned.loc[cali_cleaned['household_average']==8]
df8["median_income"].mean()*10000
```

```
]: 34028.46666666667
```

Fig 3.5 - Average median income for each household

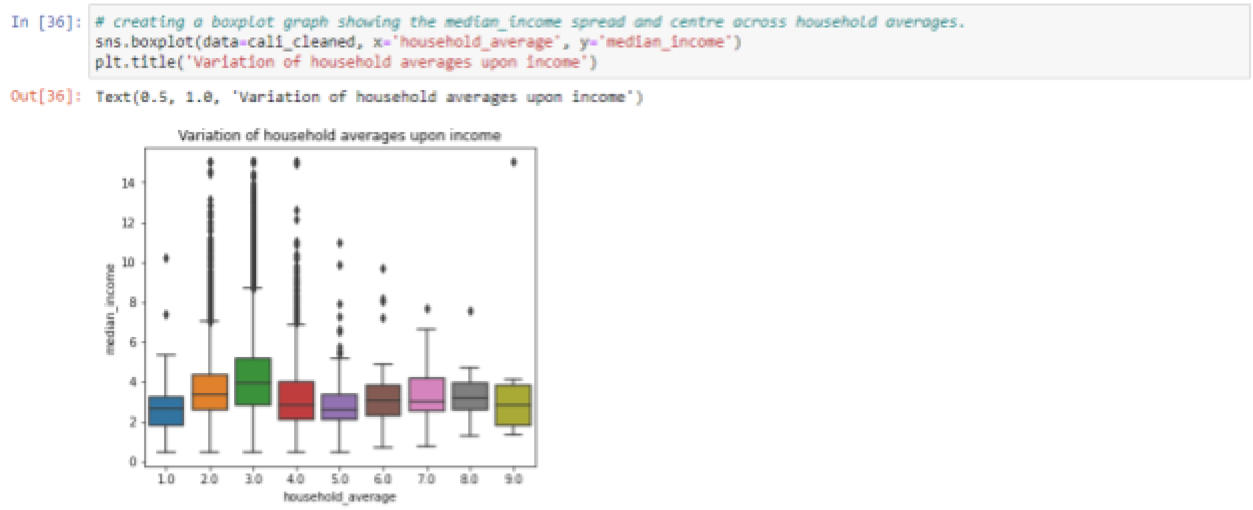


```
In [39]: # summary of median_income descriptions for each household average
         grouped1['median_income'].describe()
```

Out[39]:

household_average	count	mean	std	min	25%	50%	75%	max
1.0	150.0	2.705839	1.302792	0.4999	1.811725	2.63715	3.254575	10.2264
2.0	5846.0	3.654292	1.635944	0.4999	2.561050	3.37500	4.349550	15.0001
3.0	10765.0	4.244911	2.041487	0.4999	2.844500	3.91500	5.184200	15.0001
4.0	2902.0	3.227558	1.610982	0.4999	2.098525	2.84500	4.014725	15.0001
5.0	581.0	2.798859	1.104827	0.4999	2.107800	2.60000	3.375000	10.9704
6.0	105.0	3.173101	1.422777	0.7160	2.289100	3.08040	3.791700	9.7066
7.0	15.0	3.560153	1.855751	0.7526	2.509650	3.01320	4.163800	7.7197
8.0	15.0	3.402847	1.511880	1.2863	2.565600	3.16670	3.923600	7.5752
9.0	7.0	4.361043	4.801695	1.3750	1.815250	2.83420	3.843750	15.0001

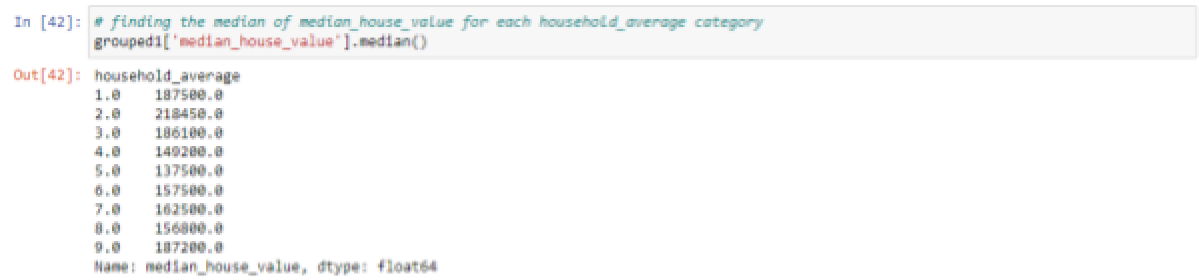
Figure 3.6 - Descriptive statistics of median income per no. of people per household



```
In [36]: # creating a boxplot graph showing the median_income spread and centre across household averages.
         sns.boxplot(data=cali_cleaned, x='household_average', y='median_income')
         plt.title('Variation of household averages upon income')

Out[36]: Text(0.5, 1.0, 'Variation of household averages upon income')
```

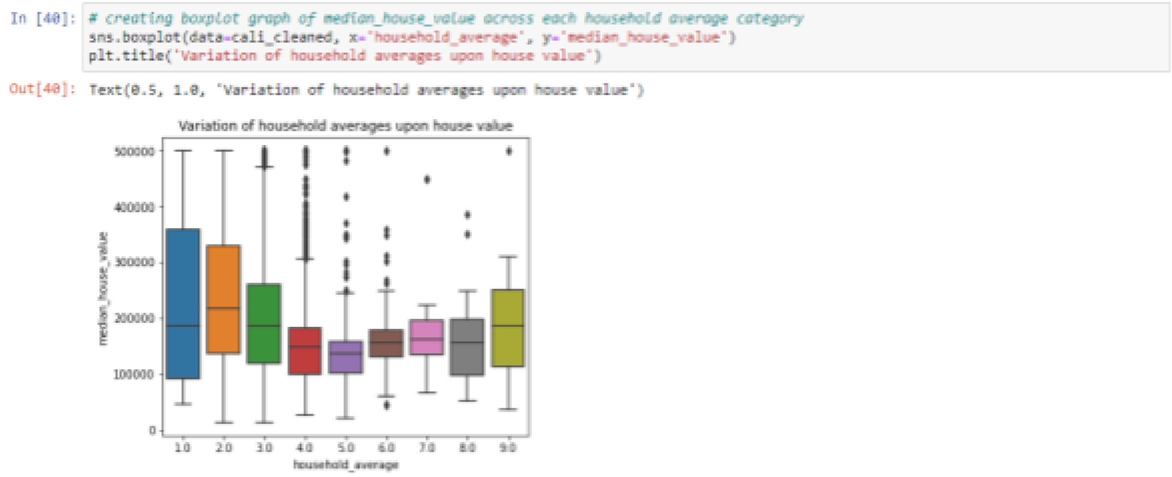
Figure 3.7 Boxplots of median household income against no. of people in a household

<u>Median house value for each no. of people in household</u>



```
In [42]: # finding the median of median_house_value for each household_average category
         grouped1['median_house_value'].median()

Out[42]: household_average
         1.0    187500.0
         2.0    218450.0
         3.0    186100.0
         4.0    149200.0
         5.0    137500.0
         6.0    157500.0
         7.0    162500.0
         8.0    156800.0
         9.0    187200.0
         Name: median_house_value, dtype: float64
```

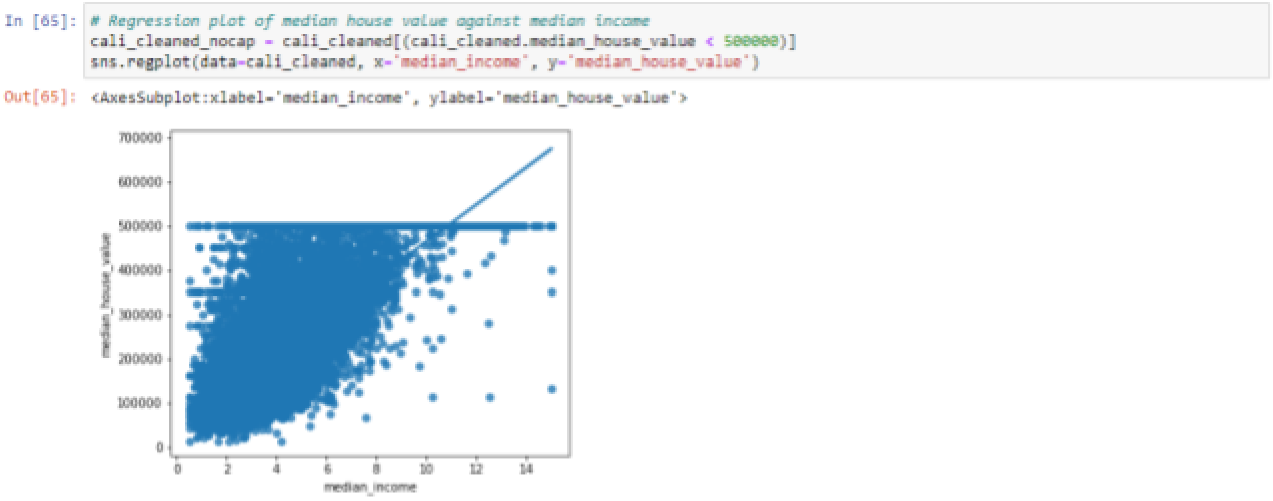
Figure 3.8



```
In [40]: # creating boxplot graph of median_house_value across each household average category
         sns.boxplot(data=cali_cleaned, x='household_average', y='median_house_value')
         plt.title('Variation of household averages upon house value')

Out[40]: Text(0.5, 1.0, 'Variation of household averages upon house value')
```

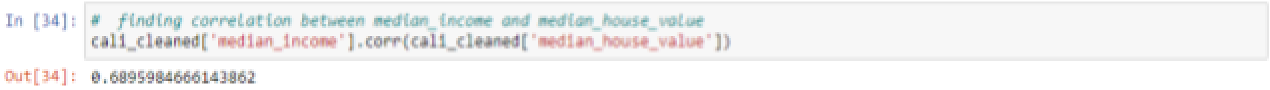
Figure 3.9 Boxplots of median house value against no. of people per household

<u>Relationship between median income and median house value.</u>



```
In [65]: # Regression plot of median house value against median income
         cali_cleaned_nocap = cali_cleaned[(cali_cleaned.median_house_value < 500000)]
         sns.regplot(data=cali_cleaned, x='median_income', y='median_house_value')

Out[65]: <AxesSubplot:xlabel='median_income', ylabel='median_house_value'>
```

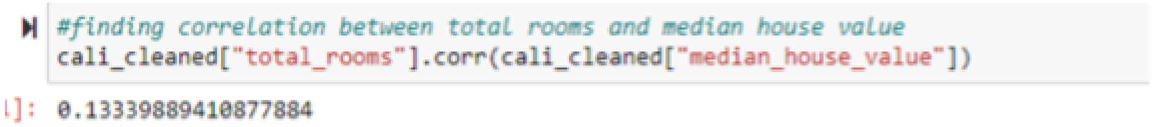
Figure 3.10 Regression plot of median house value against median income



```
In [34]: #  finding correlation between median_income and median_house_value
         cali_cleaned['median_income'].corr(cali_cleaned['median_house_value'])

Out[34]: 0.6895984666143862
```

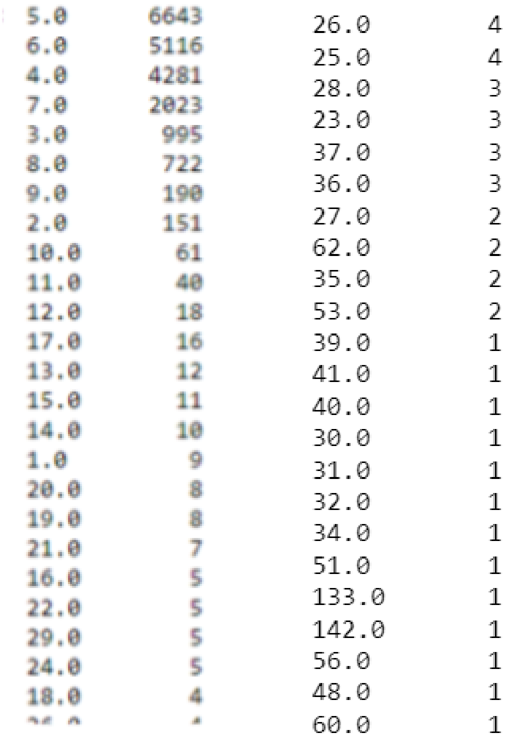
Figure 3.11 Correlation between median house value and median income

__Relationship between no. of rooms and house value.__



```
#finding correlation between total rooms and median house value
cali_cleaned["total_rooms"].corr(cali_cleaned["median_house_value"])

]: 0.13339889410877884
```

Fig 3.12 - Correlation between total rooms and median house value



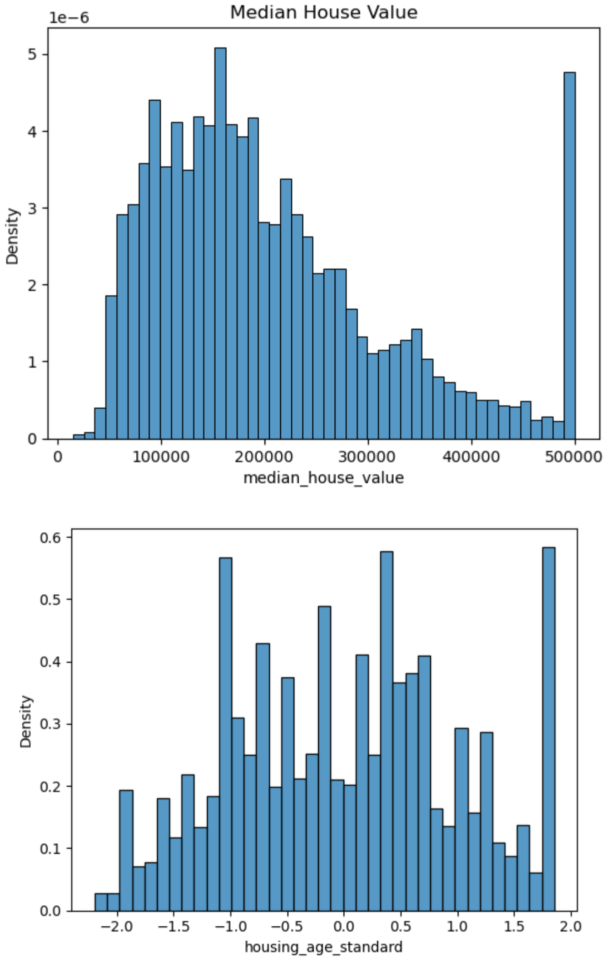
```
5.0     6643
6.0     5116
4.0     4281
7.0     2023
3.0      995
8.0      722
9.0      190
2.0      151
10.0      61
11.0      40
12.0      18
17.0      16
13.0      12
15.0      11
14.0      10
1.0        9
20.0       8
19.0       8
21.0       7
16.0       5
22.0       5
29.0       5
24.0       5
18.0       4
```

```
26.0      4
25.0      4
28.0      3
23.0      3
37.0      3
36.0      3
27.0      2
62.0      2
35.0      2
53.0      2
39.0      1
41.0      1
40.0      1
30.0      1
31.0      1
32.0      1
34.0      1
51.0      1
133.0     1
142.0     1
56.0      1
48.0      1
60.0      1
```

Fig 3.13 - value counts of average rooms

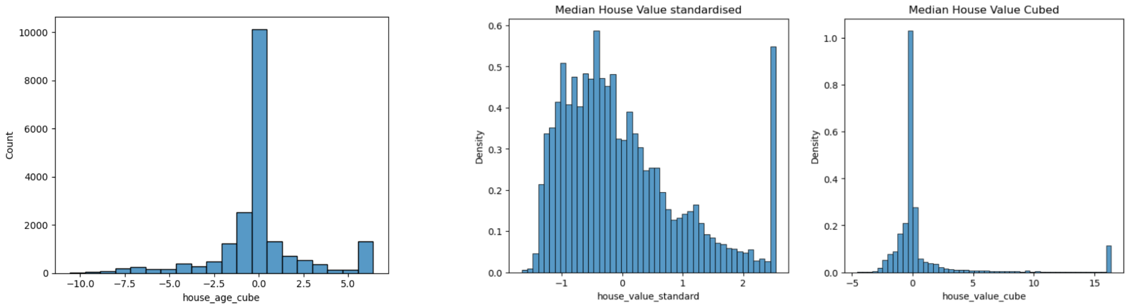
Relationship between median household value and age of house.

Figure 3.14



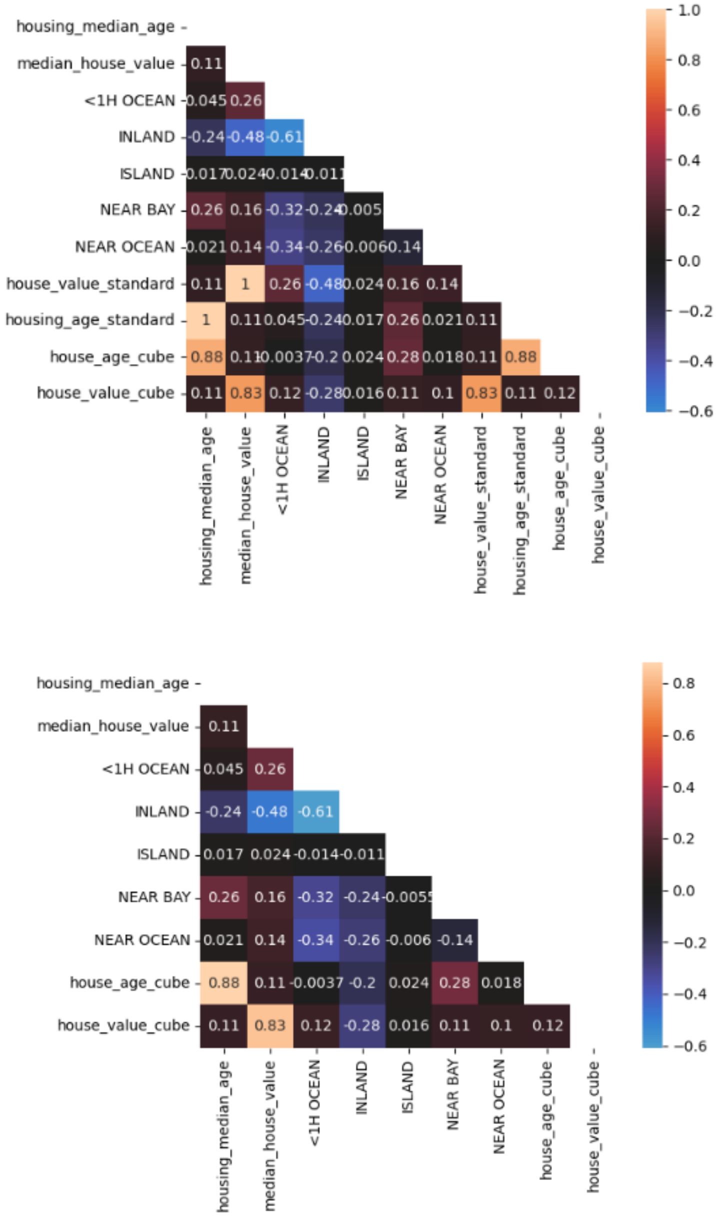
Distributions of median house value and standardised values of house age

Figure 3.15



Transformations of house age and house value respectively

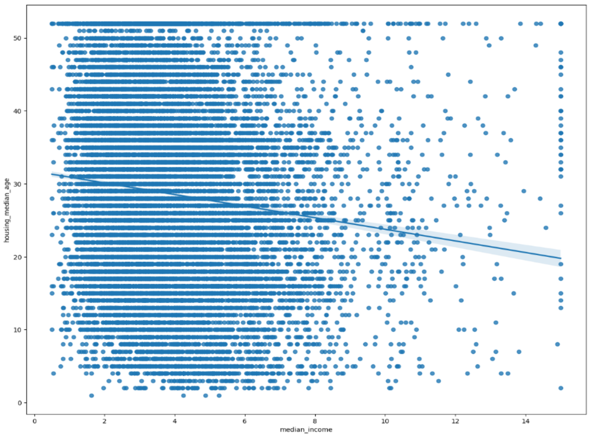
Figure 3.16



Heatmaps showing correlation coefficients between variables

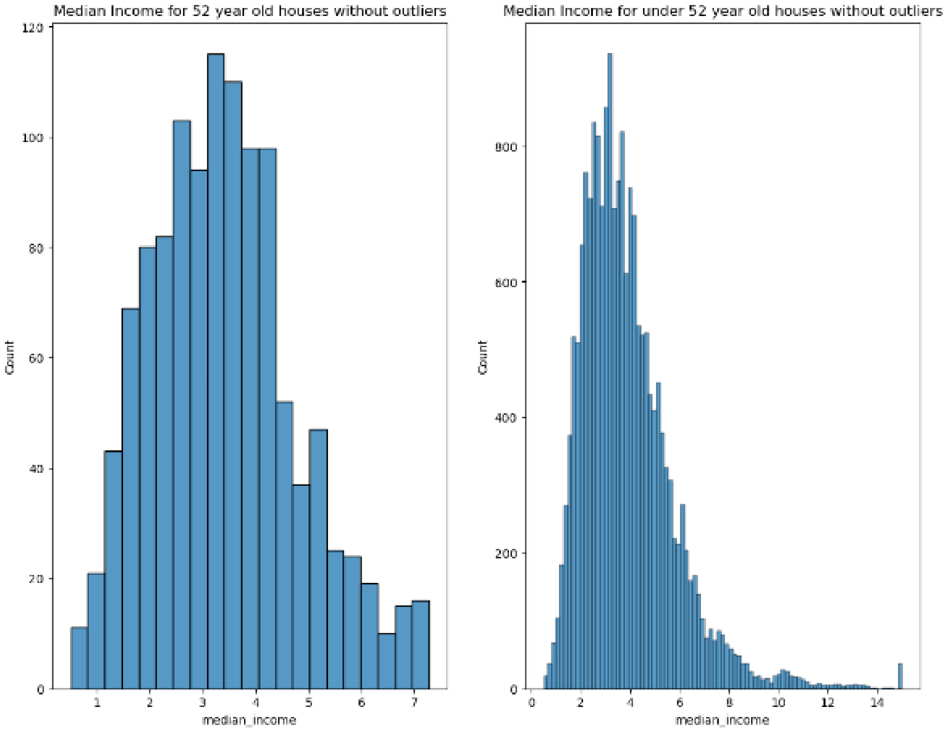
Relationship between household income and house age.

Figure 3.17



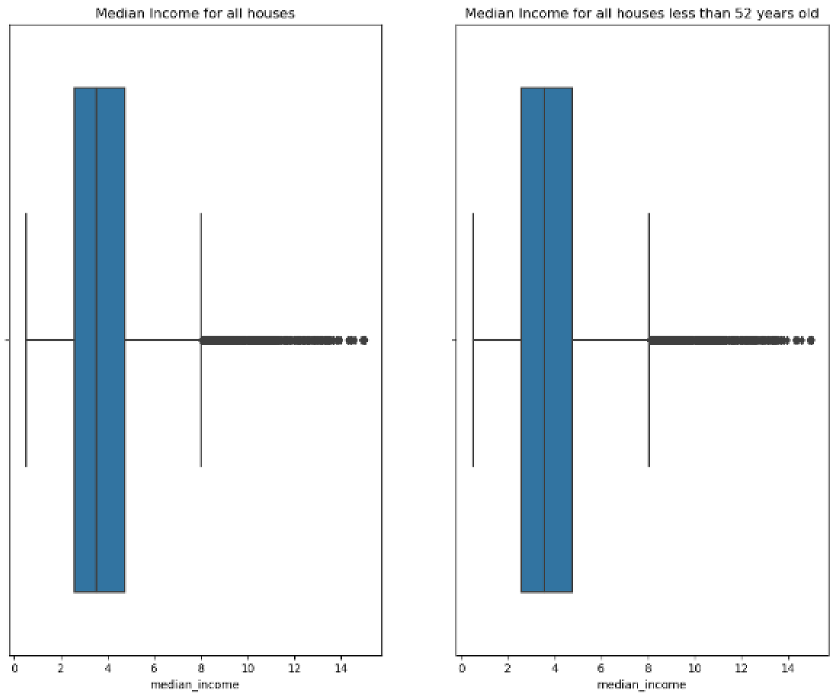
Scatterplot of median income against housing age

Figure 3.18



Distributions of median income for homes on and below 52 years old

Figure 3.19



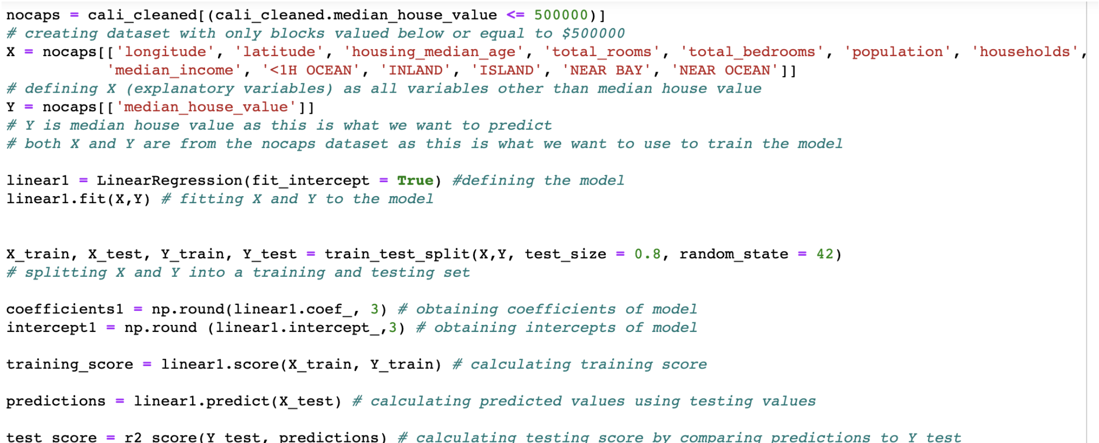
Boxplot showing range and mean median incomes for all homes and homes below 52 years old

Part 4 Summary of any undertaken modelling and any significant conclusions

Figures:

Predicting real Median House Value of capped blocks

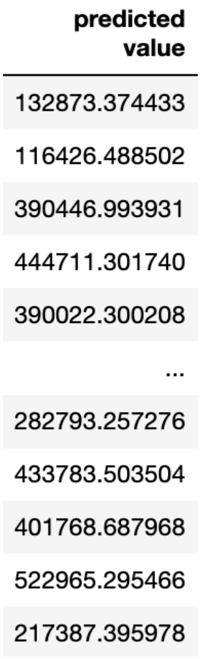
Figure 4.1: Training/testing scores, coefficients, intercept of model



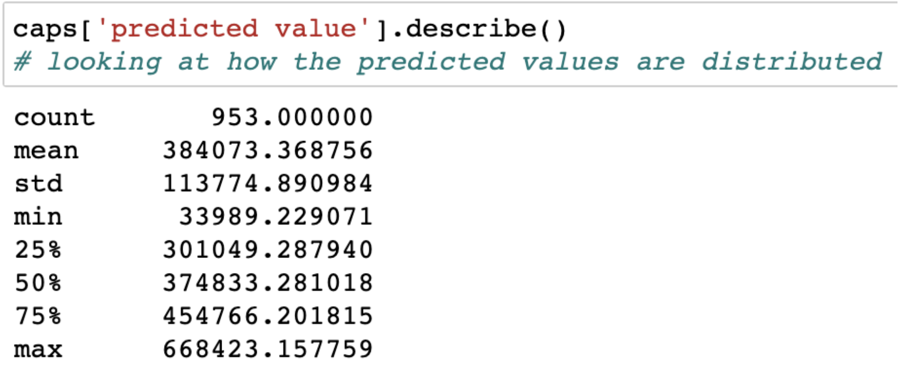
```
nocaps = cali_cleaned[(cali_cleaned.median_house_value <= 500000)]
# creating dataset with only blocks valued below or equal to $500000
X = nocaps[['longitude', 'latitude', 'housing_median_age', 'total_rooms', 'total_bedrooms', 'population', 'households',
            'median_income', '<1H OCEAN', 'INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN']]
# defining X (explanatory variables) as all variables other than median house value
Y = nocaps[['median_house_value']]
# Y is median house value as this is what we want to predict
# both X and Y are from the nocaps dataset as this is what we want to use to train the model

linear1 = LinearRegression(fit_intercept = True) #defining the model
linear1.fit(X,Y) # fitting X and Y to the model


X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.8, random_state = 42)
# splitting X and Y into a training and testing set

coefficients1 = np.round(linear1.coef_, 3) # obtaining coefficients of model
intercept1 = np.round (linear1.intercept_,3) # obtaining intercepts of model

training_score = linear1.score(X_train, Y_train) # calculating training score

predictions = linear1.predict(X_test) # calculating predicted values using testing values

test_score = r2_score(Y_test, predictions) # calculating testing score by comparing predictions to Y_test
```

Figure 4.2: Defining capped dataset for model to be applied to/variables to be used

Fig 4.3: Example of predicted values



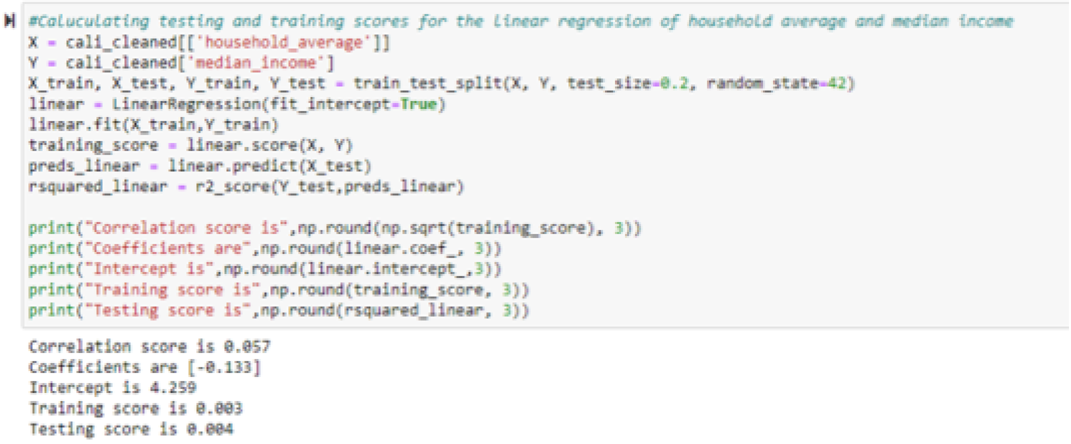
predicted value
132873.374433
116426.488502
390446.993931
444711.301740
390022.300208
...
282793.257276
433783.503504
401768.687968
522965.295466
217387.395978



```
caps['predicted value'].describe()
# looking at how the predicted values are distributed
```

```
count       953.000000
mean     384073.368756
std      113774.890984
min       33989.229071
25%      301049.287940
50%      374833.281018
75%      454766.201815
max      668423.157759
```

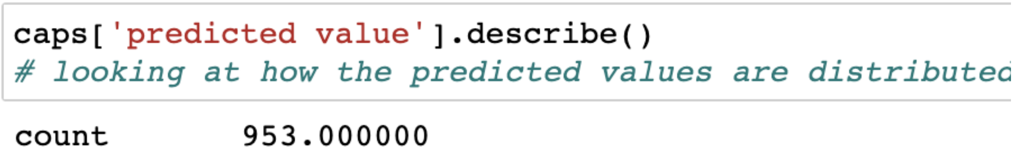
Fig 4.4: Predicted values statistics



```
#Caluculating testing and training scores for the linear regression of household average and median income
X = cali_cleaned[['household_average']]
Y = cali_cleaned['median_income']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
linear = LinearRegression(fit_intercept=True)
linear.fit(X_train,Y_train)
training_score = linear.score(X, Y)
preds_linear = linear.predict(X_test)
rsquared_linear = r2_score(Y_test,preds_linear)

print("Correlation score is",np.round(np.sqrt(training_score), 3))
print("Coefficients are",np.round(linear.coef_, 3))
print("Intercept is",np.round(linear.intercept_,3))
print("Training score is",np.round(training_score, 3))
print("Testing score is",np.round(rsquared_linear, 3))
```

```
Correlation score is 0.057
Coefficients are [-0.133]
Intercept is 4.259
Training score is 0.003
Testing score is 0.004
```

Fig 4.5 - Modelling of household average against median income



```
caps['predicted value'].describe()
# looking at how the predicted values are distributed
```

```
count       953.000000
```

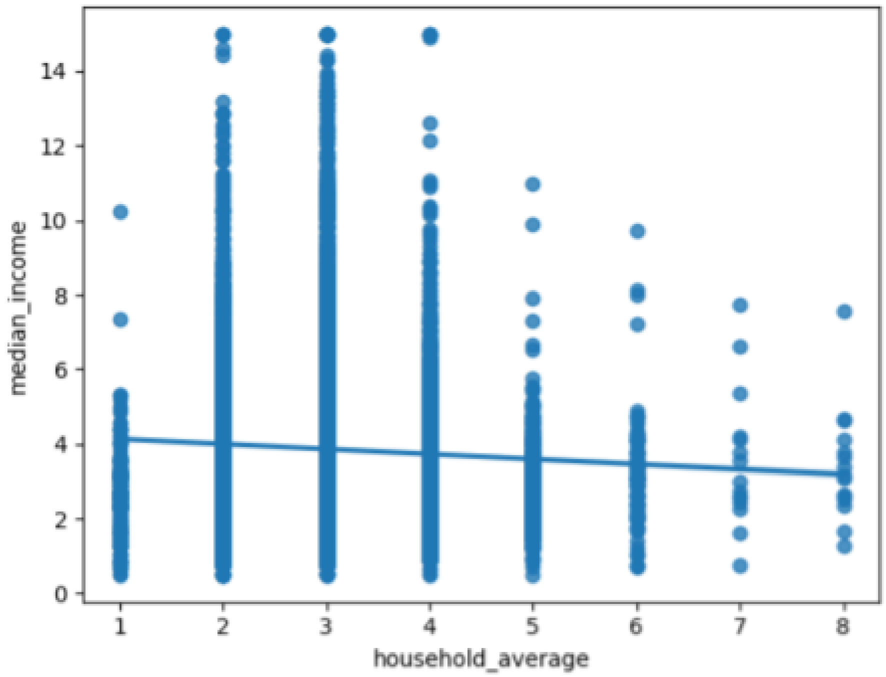
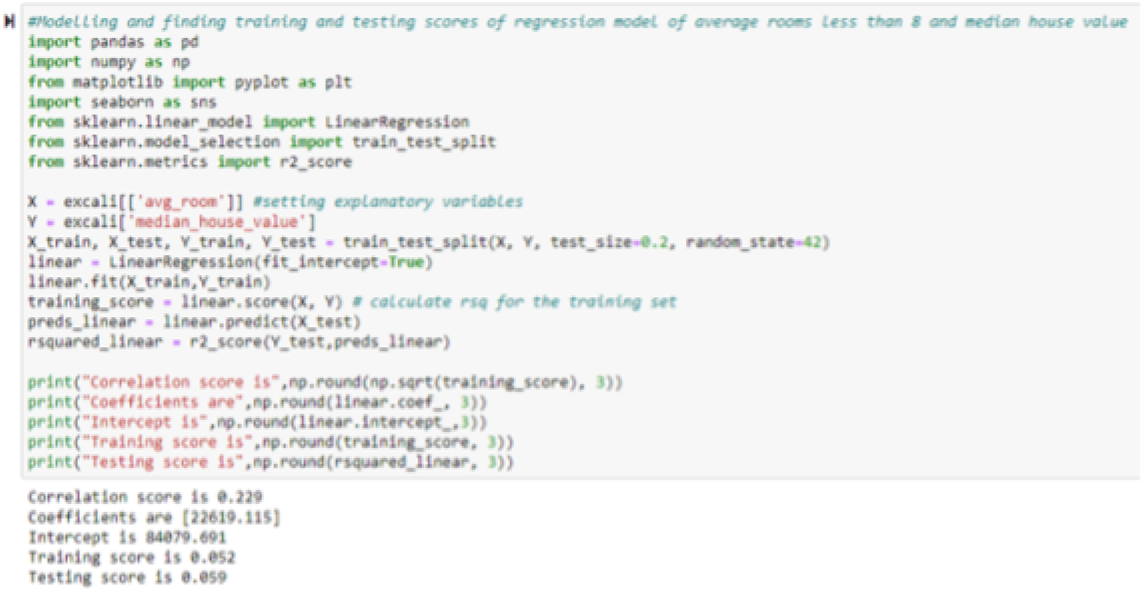


Fig 4.6 - Regression plot of household average and median income

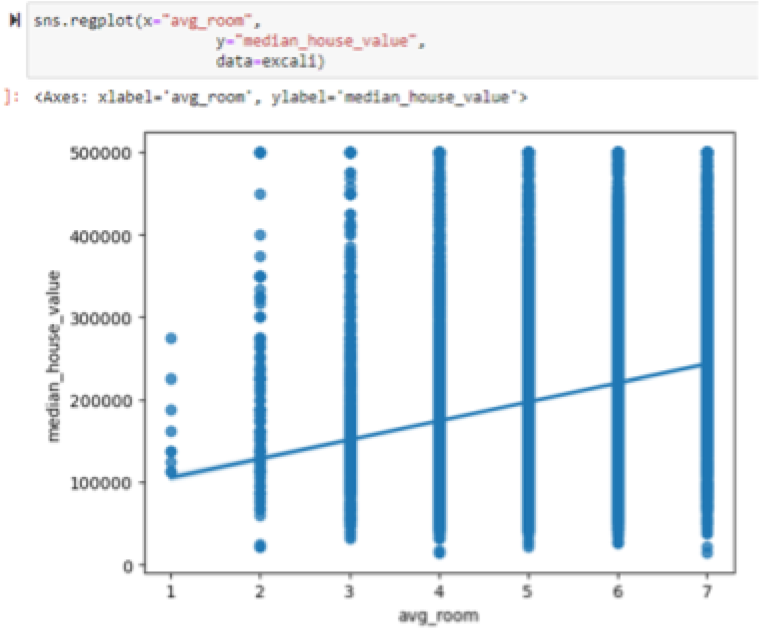
What is the relationship between no. of rooms and bedrooms (house size) and the house value?



```
#Modelling and finding training and testing scores of regression model of average rooms less than 8 and median house value
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score

X = excali[['avg_room']] #setting explanatory variables
Y = excali['median_house_value']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
linear = LinearRegression(fit_intercept=True)
linear.fit(X_train,Y_train)
training_score = linear.score(X, Y) # calculate rsq for the training set
preds_linear = linear.predict(X_test)
rsquared_linear = r2_score(Y_test,preds_linear)

print("Correlation score is",np.round(np.sqrt(training_score), 3))
print("Coefficients are",np.round(linear.coef_, 3))
print("Intercept is",np.round(linear.intercept_,3))
print("Training score is",np.round(training_score, 3))
print("Testing score is",np.round(rsquared_linear, 3))
```

```
Correlation score is 0.229
Coefficients are [22619.115]
Intercept is 84079.691
Training score is 0.052
Testing score is 0.059
```

Fig 4.7 - Modelling average room and median house value



```
sns.regplot(x="avg_room",
                    y="median_house_value",
                    data=excali)
```

```
]: <Axes: xlabel='avg_room', ylabel='median_house_value'>
```

Fig 4.8 - Regression plot of average room against median house value