Group 2 Poster

Analysing California Housing Data

1) Overview

About: California Housing Data that contains information from the 1990 California census involving:

- Location such as longitude, latitude and ocean proximity
- Details of the houses in a block like housing median age, total no. of rooms within a block, total no. of bedrooms within a block, population per block, total no. of households for a block
- Information about the worth of house. E.g. median house value and median income measured in ten thousands of dollars

Data was almost entirely numerical, but contained one categorical variable: 'Ocean Proximity'.

Main question: What variables had the most significant impact on median house value in each block?

Before exploratory analysis: cleaning data by removing outliers and null values which can negatively affect the results of our analysis. Pre-processing some variables into data that can be more easily analysed such as taking population per household to find average number of people in each household.

Exploratory analysis: Finding answers to sub-questions aimed in investigating correlations and relationships between different variables from the dataset in relation to median house value. Final result will be analysis of what variables has the most significant influence over the median house value in California.

Topics we created to focus our interpretation and analysis on:

- What is the **dominant type of household** in each block and how does this impact house value and income?
- What is the relationship between household income and house price?
- What is the **median household income** for each **no. of people in household**?
- What is the relationship between **no. of rooms** and bedrooms (house size) and the **house value**?
- What are the **average house values** and **income** for **each region** in proximity to the ocean?
- Compare the **highest** median house value and the **average** house value of each **region in proximity to the ocean**. → What type of income levels or families want to live in different ocean proximities?
- Do **smaller** or **larger** block populations attract **higher** or **lower** house prices
- Rooms and bedrooms relationship with location and hence house price
- Is there a relationship with **median household value** and the **age of the house?**
- What relationship is there between house income and age of the house?

2) Preprocessing and Manipulation of Data

1 Removing all blocks with 'NaN' values

- Before analysing the dataset, we checked for NaN values in the data that needed removing.
- This was done using the isna() function.
- Found 207 NaN values in total bedrooms column.
- Used drop function to remove their respective entries

```
1 cali.isna().sum()

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
```

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

2 Creating dummy variables for ocean proximity column

- Made ocean category column into separate dummy variable for each unique value in column
- This is to aid in future analysis where ocean proximity can be used as a numerical variable

```
1 dummies = pd.get_dummies(cali_cleaned1['ocean_proximity'])
2 cali_cleaned2 = pd.merge(cali_cleaned1, dummies, left_index=True,
3          right_index=True)
4 cali_cleaned2
```

ocean_proximity	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN
NEAR BAY	0	0	0	1	0
NEAR BAY	0	0	0	1	0

3 Standardising certain columns

- We observed that housing age, value, and median income had different scaling, which would make visualisation very difficult.
- Thus, we decided to standardise their respective values by calculating their z-scores using lambda functions

```
1 cali_cleaned = cali_cleaned.assign(zmedianincome = lambda x :
2     ((x['median_income']-x['median_income'].mean())/x['median_income'].std())) # z-score median income
3
4 cali_cleaned = cali_cleaned.assign(zhouseage = lambda x :
5     ((x['housing_median_age']-x['housing_median_age'].mean())/x['housing_median_age'].std())) # z-score house age
6
7 cali_cleaned = cali_cleaned = cali_cleaned.assign(zhousevalue = lambda x :
8     ((x['median_house_value']-x['median_house_value'].mean())/x['median_house_value'].std())) #z-score house value
```

housing_median_age
41.0
21.0
52.0

median_income	median_house_value
8.3252	452600.0
8.3014	358500.0
7.2574	352100.0

zmedianincome	zhouseage	zhousevalue
2.345106	0.982139	2.128767
2.332575	-0.606195	1.313594
1.782896	1.855723	1.258152

4 New derived variable 'household density'

- We decided to calculate the number of people living in each household, to determine what kind of housing the people lived in
- This was done by dividing population variable by household variable

```
1 cali_cleaned['household_average'] = cali_cleaned['population']/cali_cleaned['households']
2 cali_cleaned['household_average'] = round(cali_cleaned['population']/cali_cleaned['households'], 0)
```

population	households	household_average
322.0	126.0	3.0
2401.0	1138.0	2.0
496.0	177.0	3.0

3) Data Analysis

- **Block population and median house price**
  - The correlation between population and median house value was -0.0246.

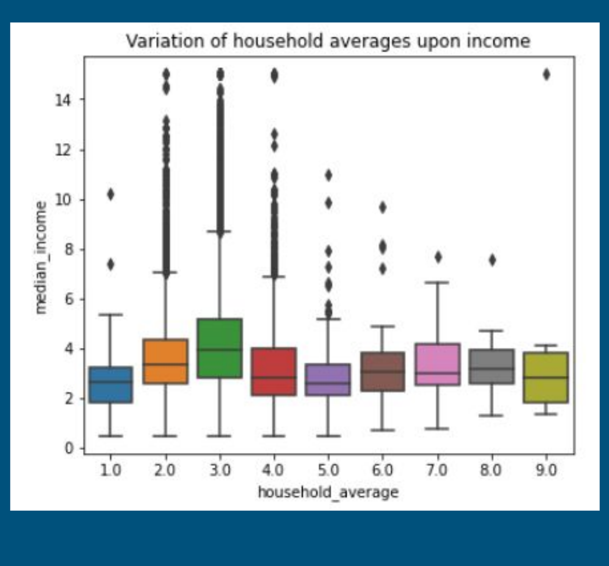
```
In [51]: population_value_correlation = cali_cleaned['population'].corr(cali_cleaned['median_house_value']).round(4)

         print('correlation between population and median house value:', population_value_correlation)

         correlation between population and median house value: -0.0246
```

- **Average house value and income for each region in proximity to ocean**
- **Highest median house value compared to average house value of each region**

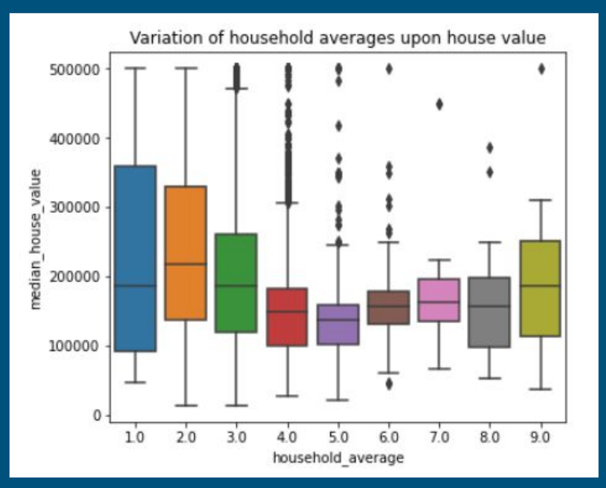
	mean	
	median_house_value	median_income
ocean_proximity		
<1H OCEAN	240234.94	4.23
INLAND	124863.96	3.21
NEAR BAY	259097.08	4.17
NEAR OCEAN	249288.90	4.01

- The highest median house value is $500001
- Double the average house price in '<1H Ocean', 'Near Bay' and 'Near Ocean'
- Four times the average house price in 'Inland'

- **Median household income for each no. of people in household**
  - 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



- **Median house value for each no. of people in household**
  - 1 people households: $187500.0
  - 2 people households: $218450.0
  - 3 people households: $186100.0
  - 4 people households: $149200.0
  - 5 people households: $137500.0
  - 6 people households: $157500.0



- **Relationship between median income and median house value.**

```
In [200]: cali_cleaned['median_income'].corr(cali_cleaned['median_house_value'])
Out[200]: 0.6895984666143862
```

- **Relationship between no. of rooms and house value.**

```
In [203]: cali_cleaned['total_rooms'].corr(cali_cleaned['median_house_value'])
Out[203]: 0.1333988941087788
```

4) Modelling

Why the model was tested

- Median house values at or above $500000 were capped at $500000
  - The data therefore did not accurately reflect the median house price
- The group decided to create a multiple linear regression model to predict the real median house value of blocks that were capped at $500000

How the model was created

- The model was trained on a dataset with blocks with a median house value less than $500000
- It used all variables as X(explanatory variable
- It used Median House Value as Y (response variable)
  - The training score was 0.6288
  - The test score was 0.6127
  - Very close, and quite high, showing it is a good model
- The model was then applied to the blocks with a median house value of $500000 as these were the blocks with a capped value

```
nocaps = cali_cleaned[(cali_cleaned.median_house_value <= 500000)]
X = nocaps[['longitude', 'latitude', 'housing_median_age', 'total_rooms', 'total_bedrooms', 'population', 'households',
           'median_income', '<1H OCEAN', 'INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN']]
Y = nocaps[['median_house_value']]

linear1 = LinearRegression(fit_intercept = True)
linear1.fit(X,Y)


X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.8, random_state = 42)

coefficients1 = np.round(linear1.coef_, 3)
intercept1 = np.round (linear1.intercept_,3)

training_score = linear1.score(X_train, Y_train)

predictions = linear1.predict(X_test)

test_score = r2_score(Y_test, predictions)

print('training score:', training_score)
print('testing score:', test_score)
print('coefficients:', coefficients1)
print('intercept:', intercept1)

training score: 0.6288125852696701
testing score: 0.6126815432933537
coefficients: [[-2.44324730e+04 -2.25714660e+04  9.31378000e+02 -6.65100000e+00
   8.70170000e+01 -3.33540000e+01  5.38430000e+01  3.83430540e+04
  -2.45039880e+04 -6.34841900e+04  1.40597689e+05 -3.15500330e+04
  -2.10594780e+04]]
intercept: [-2062340.455]
```

Results of model

- The results revealed a predicted Median House value that was lower than the $500000 for more than 75% of blocks with a capped Median house value
- This was clearly incorrect as the houses belong to the category as they were valued at or above $500000
- Therefore the group decide to not implement the predicted values, as a large proportion of them were definitely incorrect

```
caps['predicted value'].describe()

count      953.000000
mean    384073.368756
std     113774.890984
min      33989.229071
25%     301049.287940
50%     374833.281018
75%     454766.201815
max     668423.157759
Name: predicted value, dtype: float64
```

caps

median_house_value
500001.0
500001.0
500001.0
500001.0
500001.0
...
500001.0
500001.0
500001.0
500001.0
500001.0

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

5) Conclusions

Findings:

- lower population density regions are valued higher than regions with higher population density
- expect a positive correlation between median income and median house value
- i.e. inland regions have a lower house value and lower median income. In contrast, regions near the bay and ocean generally have higher house values and income.
- 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 California families/couples.
- 1 person household have the lowest average income, but the second highest median house value. Plausible reason: 1 person households are mostly tenants living on lease, justifying how they may manage to live in a property with a high house value with low income.
- In response to main problem: variables that had the most significant impact on median house value were 'Ocean Proximity', 'Median income' and 'Population Density'.
- Justification: clear pattern in impacting median house value, and are minimally impacted by external variables that aren't addressed in the dataset