CALIFORNIA HOUSING

Group 2 - ADS1000

Group Members

Name	ID
Hieu Nguyen	33936447
Joshua Gonzales	33890374
Aden Siau	33890277
Jaehong Kang	33890439
Emilia Zhang	33154309

Table of Content

Project 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?

Preprocessing and Manipulation of Data

1

Removing all blocks with ‘NaN’ values

2

Creating dummy variables for ocean proximity column

3

Making z-scores for house value, age, and income

4

Derive new variable ‘household avg’

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

```
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

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

Preprocessing and Manipulation of Data

3

Standardising certain columns

```
cali_cleaned = cali_cleaned.assign(zmedianincome = lambda x :
    ((x['median_income']-x['median_income'].mean())/x['median_income'].std())) # z-score median income

cali_cleaned = cali_cleaned.assign(zhouseage = lambda x :
    ((x['housing_median_age']-x['housing_median_age'].mean())/x['housing_median_age'].std())) # z-score house age

cali_cleaned = cali_cleaned = cali_cleaned.assign(zhousevalue = lambda x :
    ((x['median_house_value']-x['median_house_value'].mean())/x['median_house_value'].std())) #z-score house value
```

- 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

housing_median_age	median_income	median_house_value	zmedianincome	zhouseage	zhousevalue
41.0	8.3252	452600.0	2.345106	0.982139	2.128767
21.0	8.3014	358500.0	2.332575	-0.606195	1.313594
52.0	7.2574	352100.0	1.782896	1.855723	1.258152

4

New derived variable ‘household density’

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

- 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

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

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**

	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

- **Highest median house value compared to average house value of each region**
  - 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'

Data Analysis (Continued)

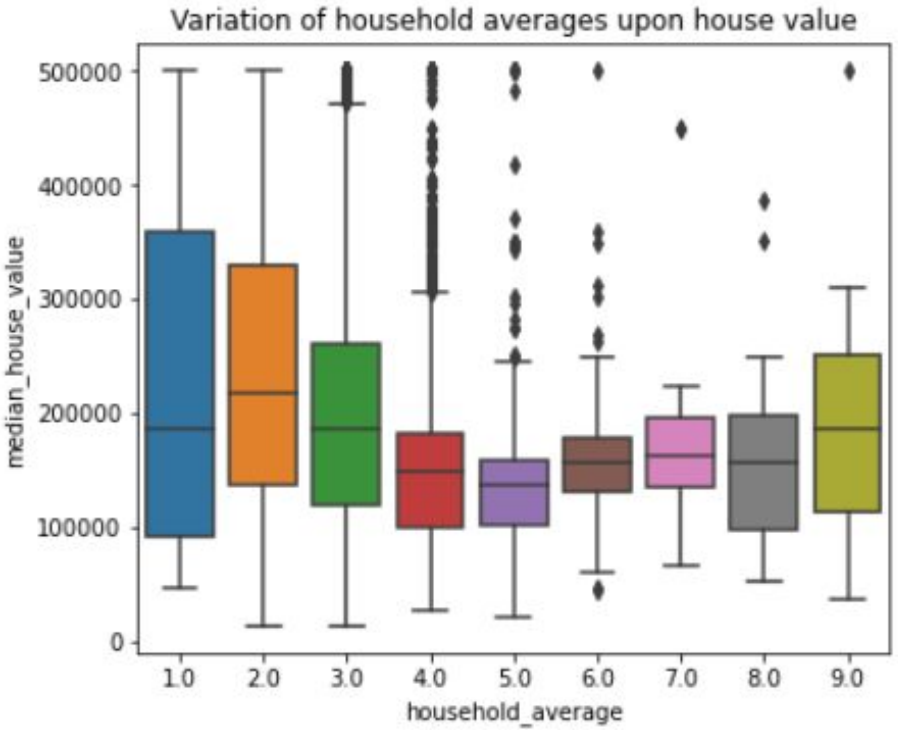
- **Median household income for each no. of people in household**
  - 1 people households: $27,058
  - 2 people households: $36,543
  - 3 people households: $42,449
  - 4 people households: $32,276
  - 5 people households: $27,989
  - 6 people households: $31,731
- **Median house value for each no. of 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.**

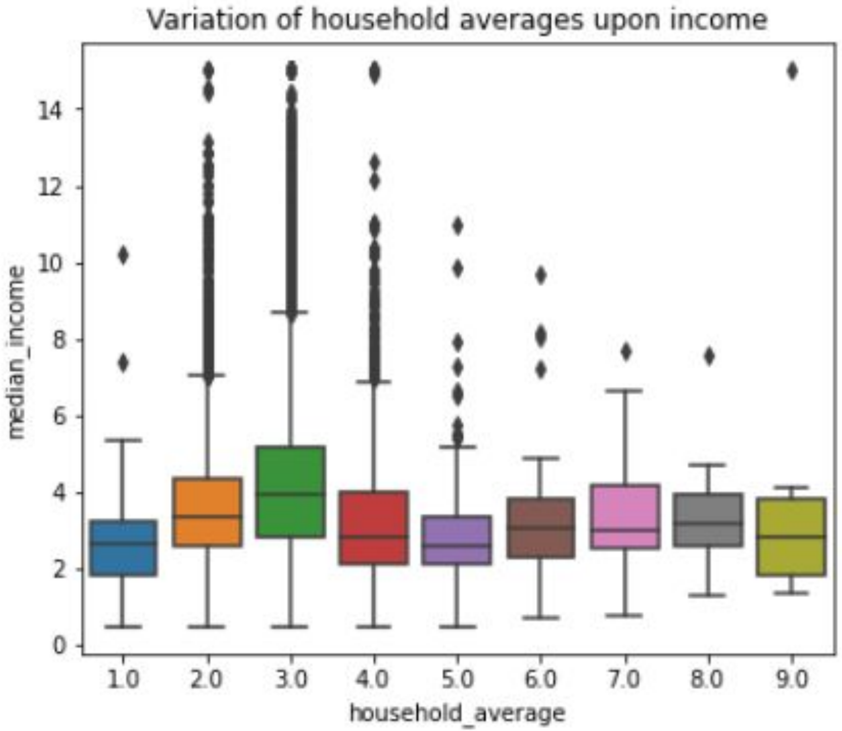
```
In [200]: cali_cleaned['median_income'].corr(cali_cleaned['median_house_value'])
Out[200]: 0.6895984666143862
```

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

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

Median income and median house value against household average





Modelling

Why was the model created?

- Median house value of blocks at or above $500000 were capped at $500000
- The data therefore did not accurately reflect the median house price of blocks in California
    - This was an issue as the main idea we wanted to investigate was what factors have the biggest impact on Median house Value of blocks
- Made a multiple linear regression model to predict the real median house value of blocks that were capped at $500000

```
caps
```

```
caps = cali_cleaned[(cali_cleaned.median_house_value >= 500001)]
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
89	-122.27	37.80	52.0	249.0	78.0	396.0	85.0	1.2434	500001.0
459	-122.25	37.87	52.0	609.0	236.0	1349.0	250.0	1.1696	500001.0
493	-122.24	37.86	52.0	1668.0	225.0	517.0	214.0	7.8521	500001.0
494	-122.24	37.85	52.0	3726.0	474.0	1366.0	496.0	9.3959	500001.0
509	-122.23	37.83	52.0	2990.0	379.0	947.0	361.0	7.8772	500001.0
...	...	...	...	...	...	...	...	...	...
20422	-118.90	34.14	35.0	1503.0	263.0	576.0	216.0	5.1457	500001.0
20426	-118.69	34.18	11.0	1177.0	138.0	415.0	119.0	10.0472	500001.0
20427	-118.80	34.19	4.0	15572.0	2222.0	5495.0	2152.0	8.6499	500001.0
20436	-118.69	34.21	10.0	3663.0	409.0	1179.0	371.0	12.5420	500001.0
20443	-118.85	34.27	50.0	187.0	33.0	130.0	35.0	3.3438	500001.0

953 rows × 20 columns

Modelling continued

How was the model created?

- Trained on all blocks with a median house value that was not capped (lower than $500000)
- Used all variables as X (explanatory variables)
- Used Median House Value as Y (response variable)
- Training score: 0.6288
- Testing score: 0.6127

- Applied model to the blocks with a capped median house 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]
```

Training on all median house values below 500000

```
caps = cali_cleaned[(cali_cleaned.median_house_value >= 500001)]
predict_X = caps[['longitude', 'latitude', 'housing_median_age', 'total_rooms', 'total_bedrooms', 'population', 'househ
            'median_income', '<1H OCEAN', 'INLAND', 'ISLAND', 'NEAR BAY', 'NEAR OCEAN']]
```

```
caps['predicted value'] = linear1.predict(predict_X)
actual_predict = linear1.predict(predict_X)
```

Modelling continued

What were the results?

- Some predictions were good, but majority were below $500000, which is incorrect
  - 75% of the data is below $454766
- Attempts to fix:
  - Training model on original data set
  - Only using X variables with strong correlation with 'Median House Value'
- Could be due to external factors which are not included in the X variables used for the model
- Decided to not apply the model as the values that were predicted to be below $500000 would be inaccurate

```
caps
```

e	median_house_value	ocean_proximity	household_average	zmedianincome	zhouseage	zhousevalue	<1H OCEAN	INLAND	ISLAND	NEAR BAY	NEAR OCEAN	predicted value
4	500001.0	NEAR BAY	5.0	-1.383549	1.855723	2.539394	0	0	0	1	0	132873.374433
5	500001.0	NEAR BAY	5.0	-1.422405	1.855723	2.539394	0	0	0	1	0	116426.488502
1	500001.0	NEAR BAY	2.0	2.096013	1.855723	2.539394	0	0	0	1	0	390446.993931
9	500001.0	NEAR BAY	3.0	2.908842	1.855723	2.539394	0	0	0	1	0	444711.301740
2	500001.0	NEAR BAY	3.0	2.109228	1.855723	2.539394	0	0	0	1	0	390022.300208
.	...	...	...	...	...	...	...	...	...	...	...	...
7	500001.0	<1H OCEAN	3.0	0.671060	0.505639	2.539394	1	0	0	0	0	282793.257276
2	500001.0	<1H OCEAN	3.0	3.251760	-1.400363	2.539394	1	0	0	0	0	433783.503504
9	500001.0	<1H OCEAN	3.0	2.516064	-1.956280	2.539394	1	0	0	0	0	401768.687968
0	500001.0	<1H OCEAN	3.0	4.565302	-1.479779	2.539394	1	0	0	0	0	522965.295466
3	500001.0	<1H OCEAN	4.0	-0.277662	1.696890	2.539394	1	0	0	0	0	217387.395978

Predicted values added to corresponding capped median house value

```
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
```

75% of the data is below $454766

Conclusions

Findings:

- Lower population density regions are worth more than regions with higher population density
- Positive correlation between median income and median house value, areas with higher median income generally have expensive houses
- Ocean proximity impacts house price: 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.
- Households with more rooms have higher house value
- 2-4 people generally earn more and have higher average median incomes as well as high-valued houses, which likely reflect stable families in California, most frequent average number of people per household.
- 1 person household have the lowest average income, but the second highest median house value. Plausible reason: 1 person households are tenants living on lease, justifying how they may manage to live in a property with a high house value with low income.
- Households with 4+ people have lower median income, most likely impacted by external factors

Conclusions (continued)

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

Overall, a variety of factors impact median house value within each of the addressed regions in the California, however the most impactful variables that impacted median house value in each block were:

- Ocean Proximity
- Median income
- Population density

From the analysis, these 3 variables were seen to have a clear pattern in impacting median house value, and are minimally impacted by external variables that aren't addressed in the dataset.

Future suggestion :

- For more accurate analysis of house values, capped values will need to be either extrapolated or recorded accurately
- Investigating nature of houses in each block such as no. of rooms and bedrooms per household will give better insight into what type of houses are valued more

Thank You

for

Listening!