EXP 5. Housing Data Report

Name	Harshal Chawan
UID	2021300019
Dataset	Housing
Experiment no.	5

Dataset link-

https://www.kaggle.com/datasets/camnugent/california-housing-prices

Dataset Description-

This is the dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being too toyish and too cumbersome.

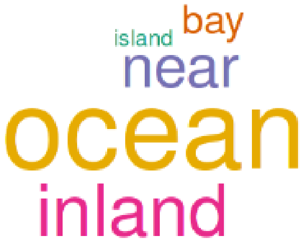
The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introduction dataset for teaching people about the basics of machine learning.

Content

The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. Be warned the data aren't cleaned so there are some preprocessing steps required! The columns are as follows, their names are pretty self explanatory:

- longitude
- latitude
- housing_median_age
- total_rooms
- total_bedrooms
- population
- households
- median_income
- median_house_value
- ocean_proximity

Report-



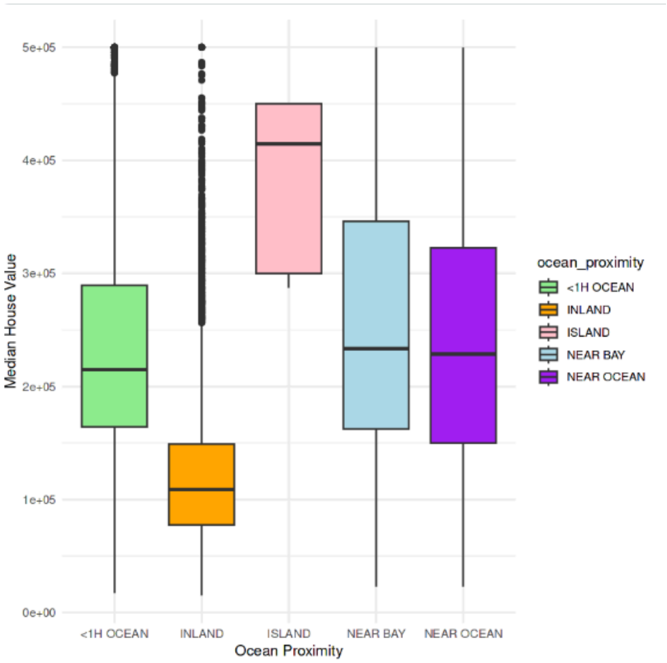
Question:

What does the word cloud plot reveal about the distribution of ocean_proximity in the housing dataset?

The word cloud shows the relative frequency of each category in the ocean_proximity column of the dataset. The size of the words indicates how often each category appears:

- "ocean" is the largest, indicating it is the most common category.
- "inland" and "near" are also prominent, suggesting they are frequent.
- "bay" and "island" appear smaller, meaning they occur less often in the dataset.

This suggests that the majority of the houses are either close to the ocean or inland, with fewer near the bay or on an island.



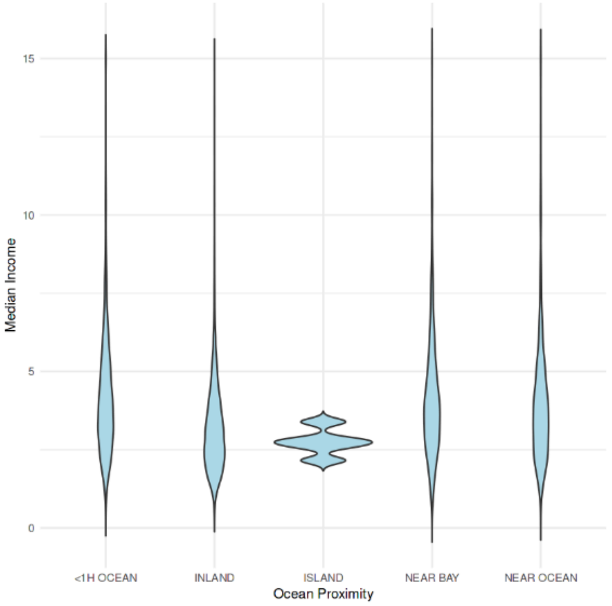
Question:

What does the box plot reveal about the relationship between ocean_proximity and median_house_value in the housing dataset?

The box plot shows the distribution of median_house_value for different categories of ocean_proximity:

- "ISLAND" properties have the highest median house values with relatively little variability, indicating that island houses are more expensive on average.
- "NEAR OCEAN" and "NEAR BAY" also have relatively high median house values, but with greater variability compared to "ISLAND".
- "<1H OCEAN" has moderately high median house values, with some outliers, showing a wide range of house prices.
- "INLAND" properties have the lowest median house values, with a narrower distribution and several outliers below the lower quartile, indicating that inland properties tend to be less expensive.

Overall, houses closer to the ocean or bay tend to have higher median values compared to inland properties, with island homes being the most expensive.



Question:

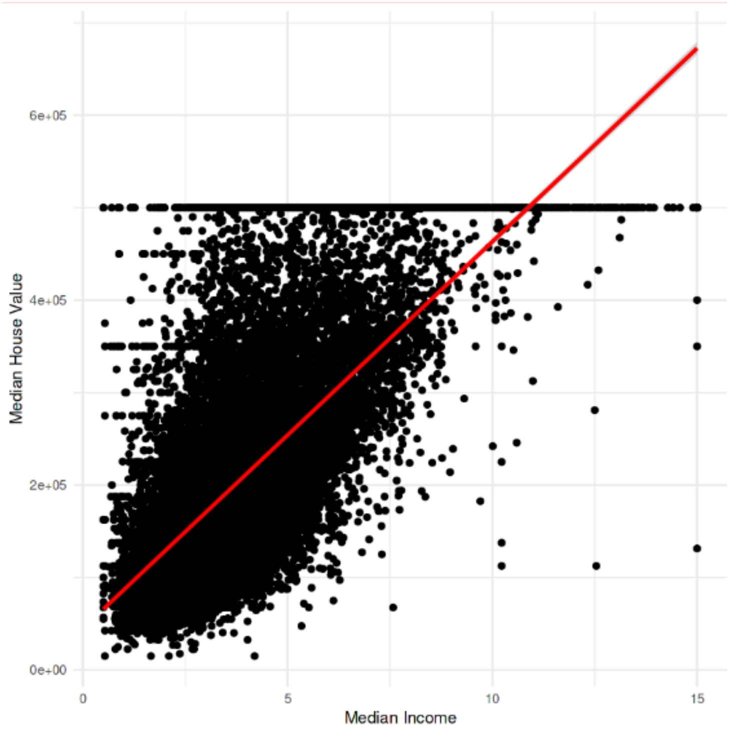
What does the violin plot reveal about the distribution of median_income across different ocean_proximity?

The violin plot shows the distribution of median_income for each ocean_proximity category:

- "<1H OCEAN", "INLAND", "NEAR BAY", and "NEAR OCEAN" all exhibit a wide range of median_income values with similarly shaped distributions. These categories have a long, slim shape, suggesting that incomes are spread across a wide range, with many households having lower to mid-level incomes and fewer at the higher end.
- "ISLAND" has a very distinct distribution with a much narrower range of median_income, indicating that income levels for island residents are more

concentrated around a median value. The compact shape suggests less variability in income for houses on islands.

In summary, most proximity categories show wide income variability, but island homes tend to have more uniform income levels, with incomes clustering around a central value.



Question:

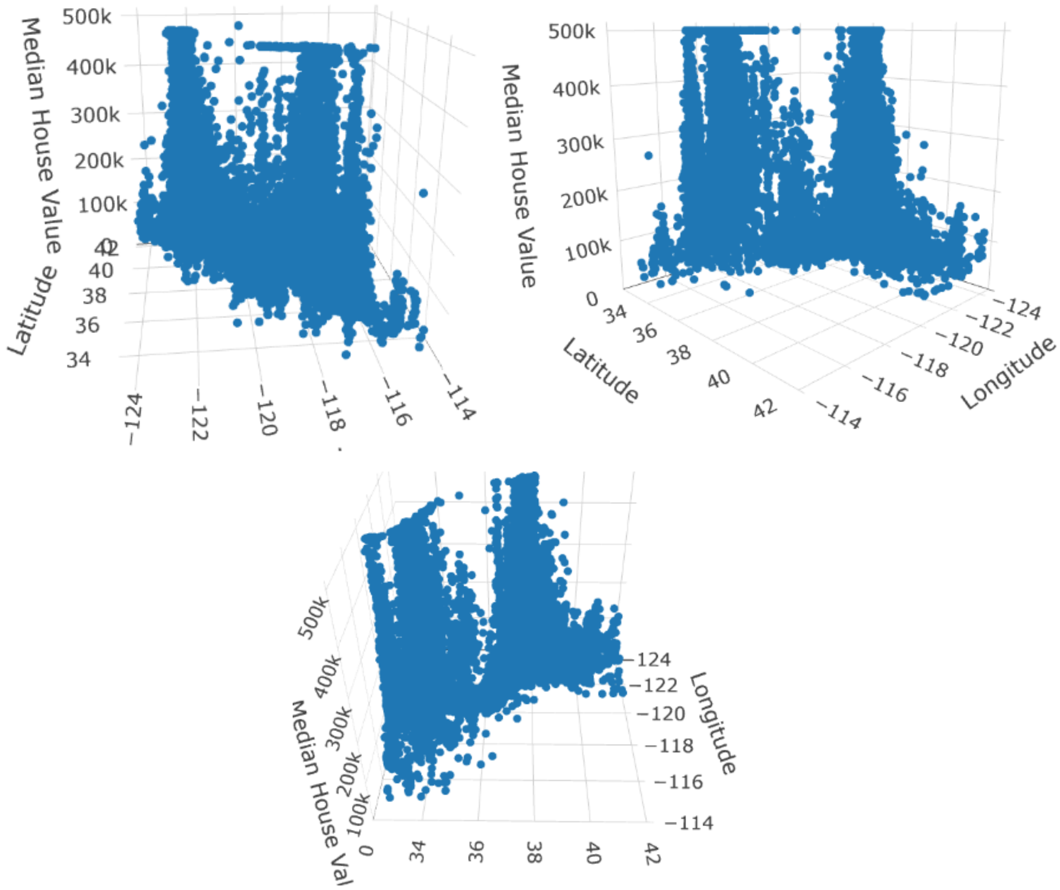
What does the regression reveal about the distribution of median_house_value across different median_income?

The scatterplot with a regression line suggests that median house value increases as median income increases. Here's what can be inferred:

- Positive Correlation: The red regression line shows a positive slope, indicating a positive linear relationship between median income and median house value. As median income rises, the median house value tends to increase as well.
- Wide Dispersion: There is a significant amount of scatter around the line, suggesting a lot of variability in house values for similar income levels. The plot has many data points that are far from the regression line, implying that income is not the sole predictor of house value—other factors likely contribute to the variability.
- Upper Capping: The points seem to flatten at the upper right of the graph (around 500,000), indicating that median house values hit a maximum threshold,

regardless of how high median income goes beyond a certain level. This could be due to a cap on house prices in the dataset.

In conclusion, while the regression line shows a clear trend that higher incomes tend to correspond to higher house values, there are outliers and some flattening of house values at high income levels. This implies that while income is a key factor, it is not the only factor influencing house values.



Question:

What does the regression reveal about the distribution of median_house_value across different latitudes and longitudes?

The 3D plot shows the median house value across California represented by latitude and longitude.

There exists 3 different spikes in the data showing the 3 major cities. The surrounding area has lower house values showing suburban areas with the remaining being rural with the least value.