NAME: ARNAV HOSKOTE
UID: 2021300044
CLASS: BE COMPS A
BATCH: ADV BATCH F

ADV EXPERIMENT 5

DATASET:
Covid-19 dataset - https://www.kaggle.com/datasets/camnugent/california-housing-prices

Objectives:
1. To visualize the distribution and relationship between various features in the housing dataset.
2. To identify potential outliers and understand the spread of the data.
3. To explore the relationship between independent variables and the target variable (e.g., house prices).
4. To create informative visualizations that can guide decision-making in the housing market.

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 to 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 introductory dataset for teaching people about the basics of machine learning.

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 explanitory:

longitude

latitude

housing_median_age

total_rooms

total_bedrooms

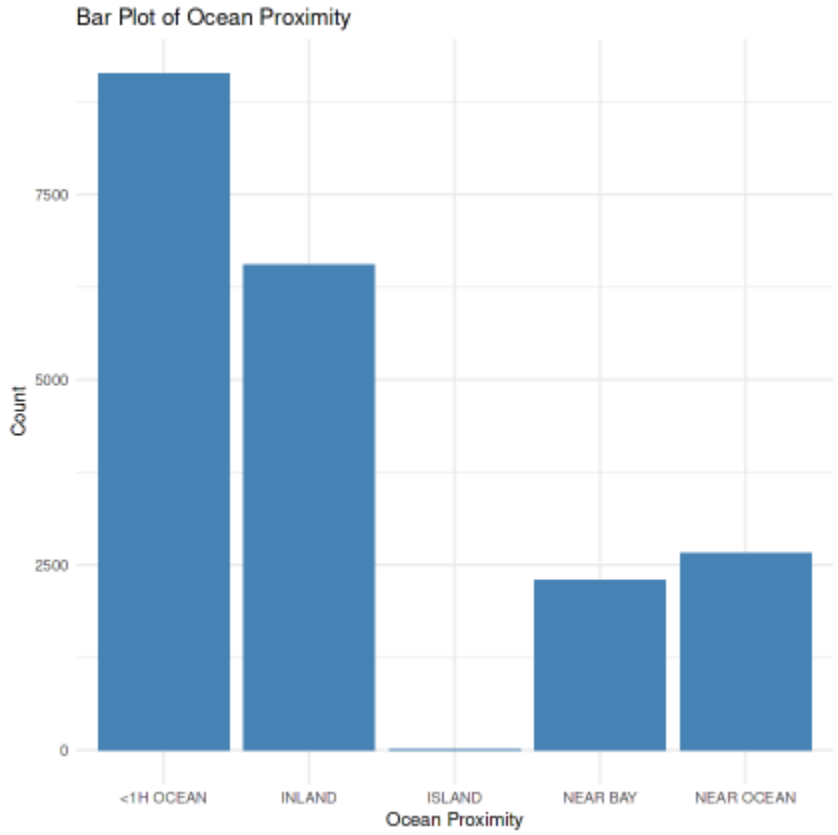
population

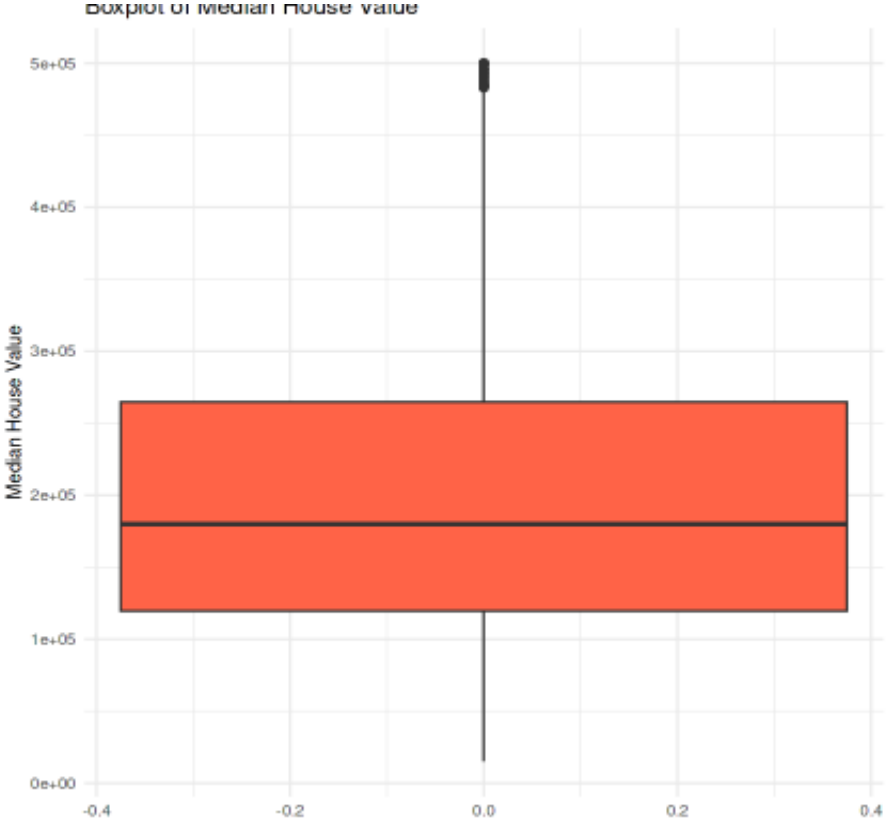
households

median_income

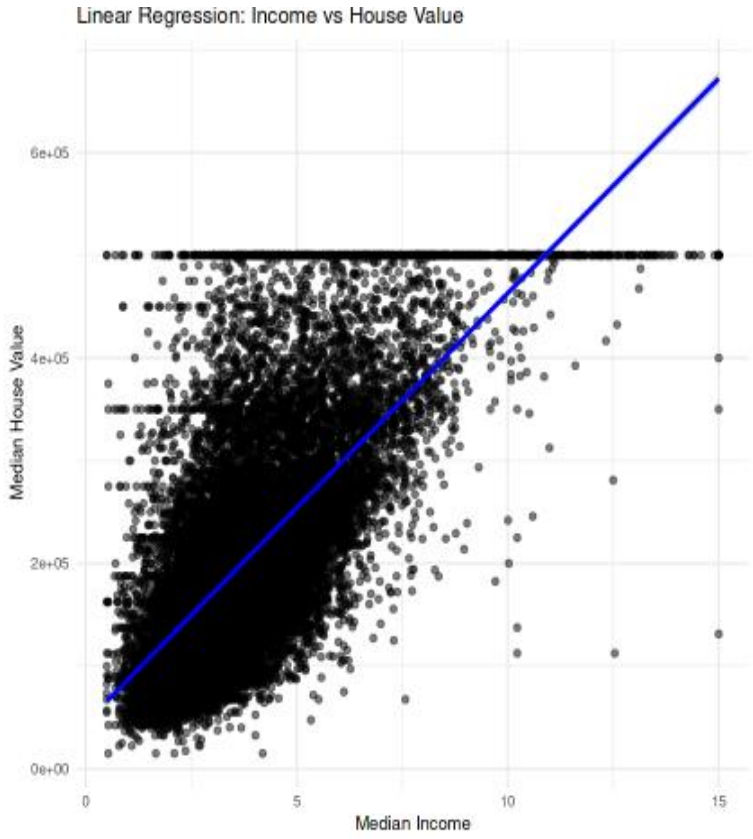
median_house_value
ocean_proximity

REPORT:

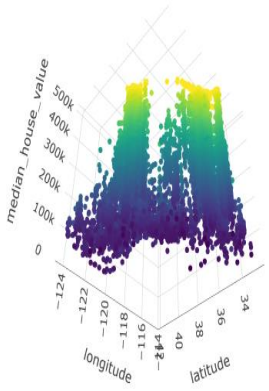

Bar Plot of Ocean Proximity
Count
7500
5000
2500
0
<1H OCEAN
INLAND
ISLAND
NEAR BAY
NEAR OCEAN
Ocean Proximity

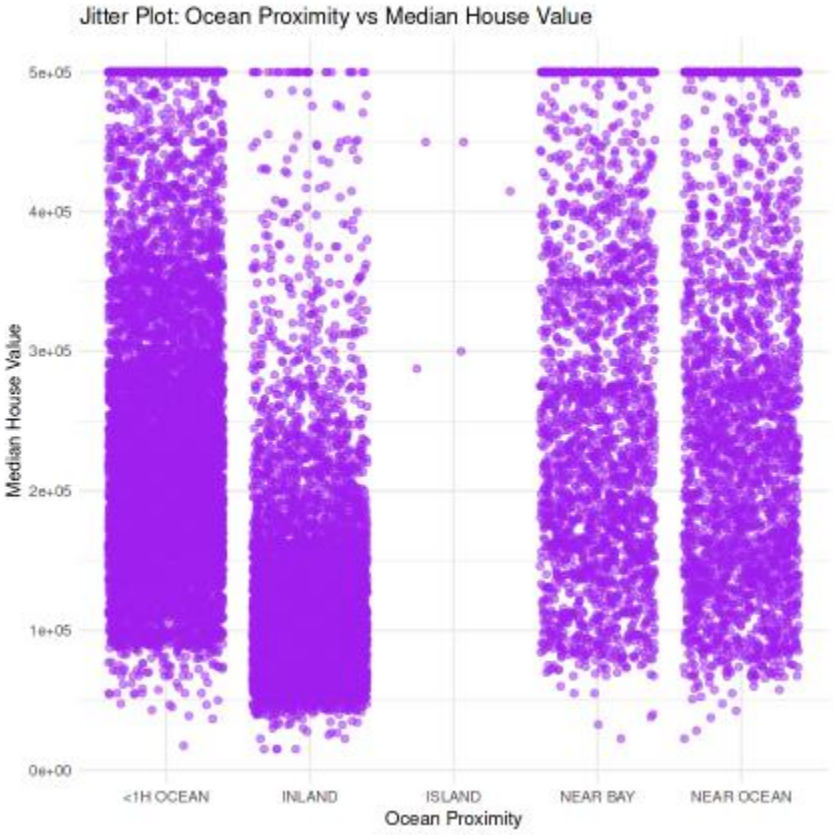

Median House Value
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
-0.4
-0.2
0.0
0.2
0.4


Violin Plot of Median Income
Median Income
15
10
5
0
1
factor(1)


Linear Regression: Income vs House Value
Median House Value
6e+05
4e+05
2e+05
0e+00
0
5
10
15
Median Income

3D Scatter Plot: House Value by Location





1. Bar Plot of Ocean Proximity
Question:
What is the distribution of houses based on their proximity to the ocean?
Answer:
The bar plot shows the frequency of houses based on different categories of ocean proximity. Most of the houses are either `NEAR OCEAN`, `INLAND`, or `NEAR BAY`, while fewer houses fall into the categories of `ISLAND` or `NEAR OCEAN`.

2. Box and Whisker Plot for Median House Value
Question:
What is the spread and distribution of median house values in the dataset?
Answer:
The box plot displays the range of median house values, highlighting the interquartile range (middle 50% of the data). The box extends from the 25th to the 75th percentile, with the median (50th percentile) represented by a line inside the box. The whiskers show variability outside the upper and lower quartiles, and any points outside this range are considered outliers. The distribution shows a skew towards higher house values with a few outliers in the upper range.

3. Violin Plot for Median Income
Question:
How is the distribution of median income spread across households?
Answer:
The violin plot displays the distribution of median income in the dataset. It provides a symmetrical visualization of the income distribution's density, showing how most households are concentrated around certain income ranges, while fewer households are either very high or very low income.

The plot reveals that most incomes are concentrated in the lower and middle ranges, with a taper towards the higher ends.

4. Regression Plot (Linear) for Median Income vs. Median House Value
Question:
Is there a linear relationship between median income and median house value?
Answer:
The linear regression plot shows a positive correlation between median income and median house value, meaning as the median income increases, the median house value also tends to rise. The blue line represents the best fit linear regression line, indicating a strong positive relationship.

5. Nonlinear Regression Plot (Polynomial) for Median Income vs. Median House Value
Question:
Is there a nonlinear relationship between median income and median house value?
Answer:
The nonlinear regression plot (using a polynomial fit) shows a more complex relationship between median income and median house value. The red curve indicates that house values increase sharply with income at lower ranges and level off at higher income ranges, suggesting a diminishing return on house value at higher income levels. This is a better fit for the data compared to a simple linear model, as it captures the curvilinear trend.

6. 3D Scatter Plot for Longitude, Latitude, and Median House Value
Question:
How are house values distributed geographically based on their location (longitude and latitude)?
Answer:
The 3D scatter plot provides a geographic visualization of house values based on their latitude and longitude. Houses located near the coast or in specific regions tend to have higher median house values, while houses further inland or in certain areas show lower values. This helps to visually understand how house values are related to location and proximity to certain regions.

7. Jitter Plot for Ocean Proximity vs. Median House Value
Question:
What is the variation of house values based on proximity to the ocean?
Answer:
The jitter plot shows how median house values vary by the different ocean proximity categories. Houses closer to the ocean (`NEAR BAY`, `NEAR OCEAN`, `ISLAND`) tend to have higher median values, while those `INLAND` have lower values. This plot helps in visualizing the spread and variability of house values within each proximity category, with a visible concentration of higher values for homes near the ocean.

CONCLUSION: I have successfully plotted advanced graphs using R language and answered all questions regarding the dataset.