
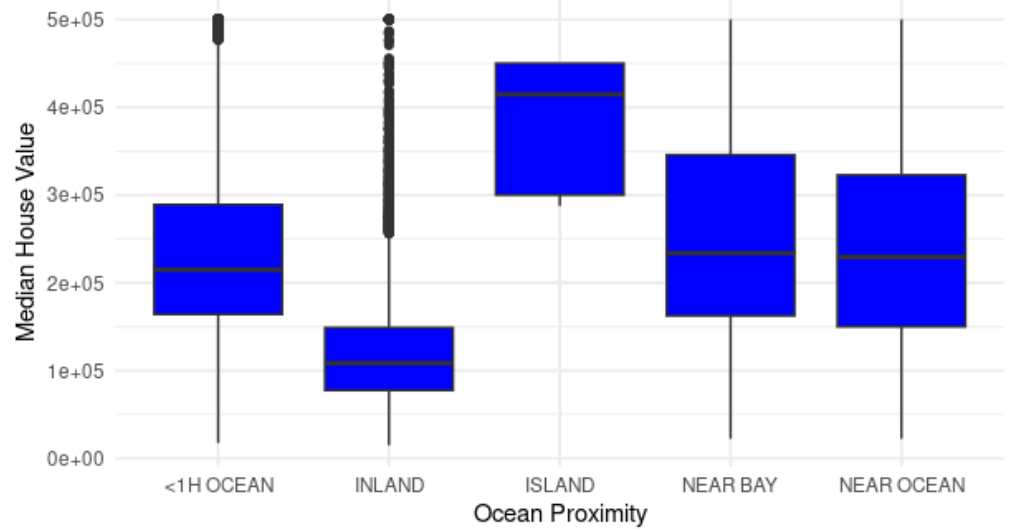


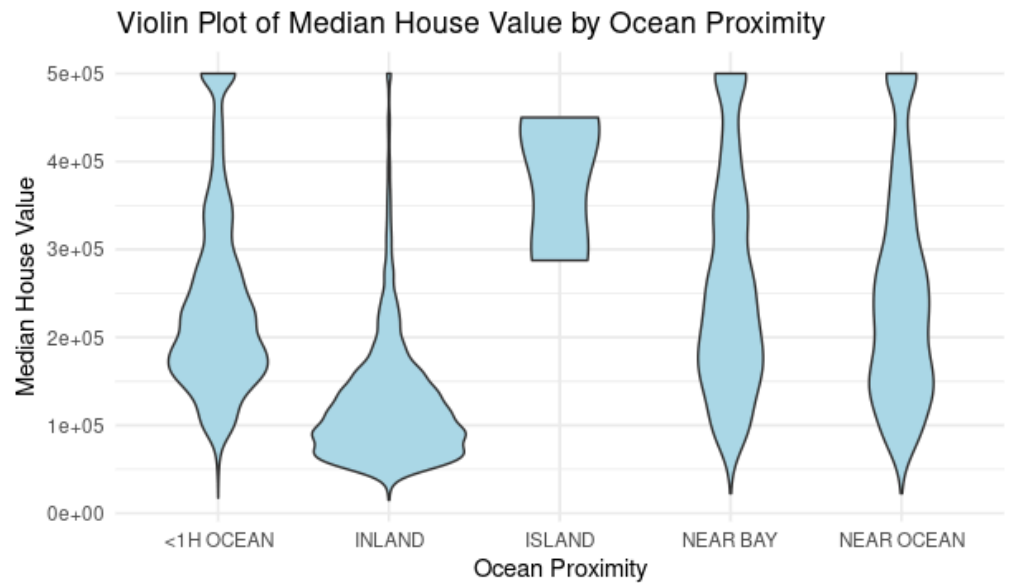
Name:	Prashil Deepak Kadam
UID:	2021600031
Experiment No:	5
Aim:	To perform data preprocessing and EDA a Housing dataset in RStudio using R.
Dataset link:	https://www.kaggle.com/datasets/camnugent/california-housing-prices The California Housing Prices dataset is a popular dataset derived from the 1990 U.S. Census, used for regression tasks in machine learning. It contains information on various features like median income, housing age, and the number of rooms in different California districts, with the target variable being the median house value. This dataset is often used to build models that predict housing prices based on these features, making it a valuable resource for exploring real-world regression problems.
Code:	<pre> library(ggplot2) library(wordcloud) data <- housing head(data) ocean_counts <- table(data\$ocean_proximity) wordcloud(names(ocean_counts), freq = ocean_counts, min.freq = 1, max.words = 100, random.order = FALSE) ggplot(data, aes(x = ocean_proximity, y = median_house_value)) + geom_boxplot() + labs(title = "Box and Whisker Plot of Median House Value by Ocean Proximity", x = "Ocean Proximity", y = "Median House Value") + theme_minimal() ggplot(data, aes(x = ocean_proximity, y = median_house_value)) + geom_violin(fill = "lightblue") + labs(title = "Violin Plot of Median House Value by Ocean Proximity", x = "Ocean Proximity", y = "Median House Value") + theme_minimal() ggplot(data, aes(x = median_income, y = median_house_value)) + geom_point(alpha = 0.5) + geom_smooth(method = "lm", color = "red") + labs(title = "Linear Regression Plot: Median Income vs Median House Value", x = "Median Income", y = "Median House Value") + </pre>

	<pre>theme_minimal() ggplot(data, aes(x = median_income, y = median_house_value)) + geom_point(alpha = 0.5) + geom_smooth(method = "loess", color = "blue") + labs(title = "Non-linear Regression Plot: Median Income vs Median House Value", x = "Median Income", y = "Median House Value") + theme_minimal() ggplot(data, aes(x = ocean_proximity, y = median_house_value)) + geom_jitter(alpha = 0.5, width = 0.2) + labs(title = "Jitter Plot of Median House Value by Ocean Proximity", x = "Ocean Proximity", y = "Median House Value") + theme_minimal()</pre>
Results / Outputs:	<div><p>Word Chart of the ocean proximity of the house</p><p>Box and Whisker Plot of Median House Value by Ocean Proximity</p></div>

Box and Whisker Plot of Median House Value by Ocean Proximity

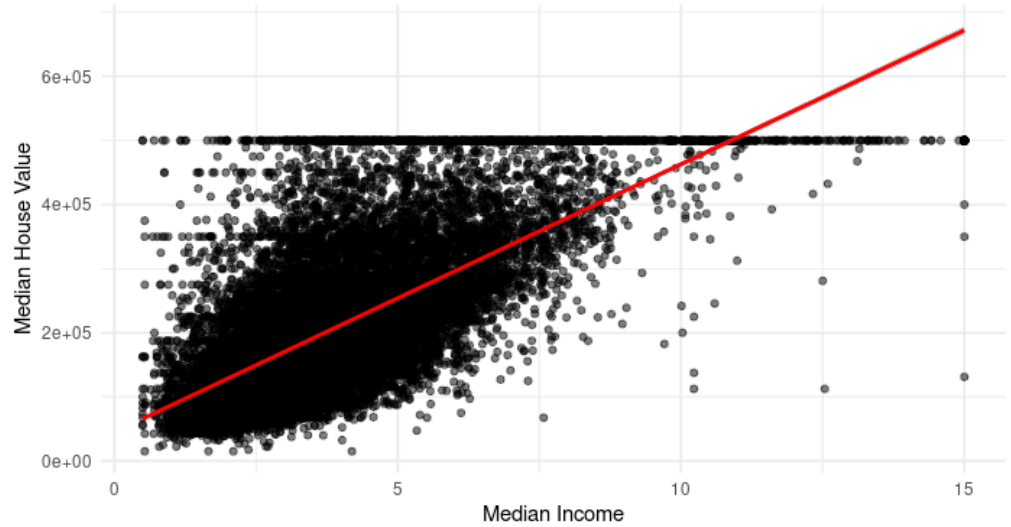


Violin Plot of Median House Value by Ocean Proximity



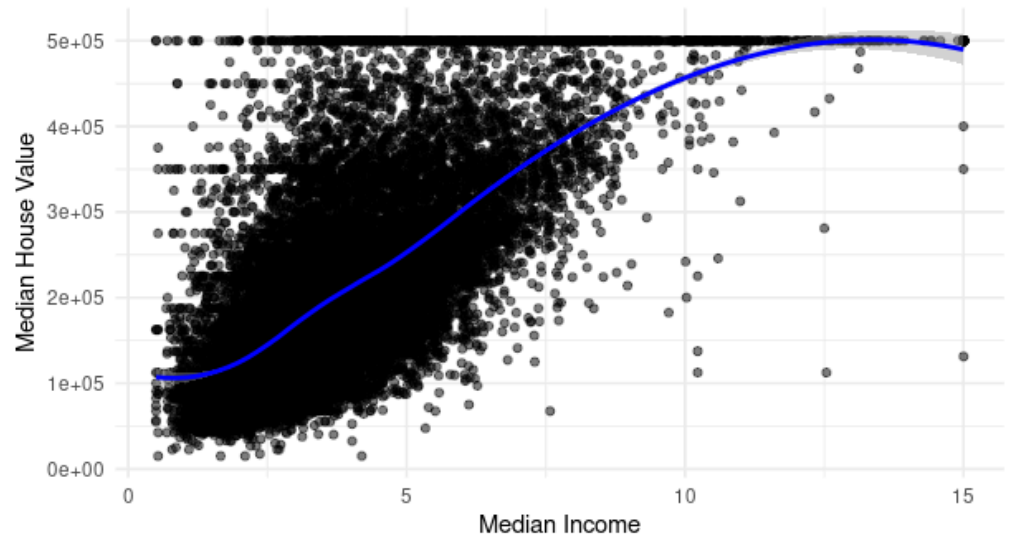
Linear Regression Plot of Median House Value vs Median Income

Linear Regression Plot: Median Income vs Median House Value



Non Linear Regression Plot for Median House vs Median Income

Non-linear Regression Plot: Median Income vs Median House Value



Jitter Plot of Median House Value vs Ocean Proximity

	<div><p>Jitter Plot of Median House Value by Ocean Proximity</p><p>A jitter plot showing the distribution of median house values for five categories of ocean proximity: <1H OCEAN, INLAND, ISLAND, NEAR BAY, and NEAR OCEAN. The y-axis represents the median house value, ranging from 0e+00 to 5e+05. The <1H OCEAN, INLAND, NEAR BAY, and NEAR OCEAN categories show a high density of points, with values ranging from approximately 50,000 to 500,000. The ISLAND category has a much smaller number of points, with values ranging from approximately 300,000 to 450,000.</p></div>
Conclusion	<div><p>1. Word Chart of Ocean Proximity of Houses The word chart highlights that most houses are located inland, with relatively fewer houses near the ocean, bay, or on an island. This suggests a majority of properties in the dataset are not located near water bodies, potentially indicating that ocean proximity is a less frequent attribute.</p><p>2. Box and Whisker Plot of Median House Value by Ocean Proximity This plot shows the distribution of house values based on proximity to the ocean. Houses closer to the ocean, especially those within an hour (<1H OCEAN), generally have higher median values compared to inland properties. There is also a larger spread of house values for homes near the ocean, indicating greater variability in prices. Inland properties have lower median values with less variability, suggesting that proximity to water is a significant factor in housing prices.</p><p>3. Violin Plot of Median House Value by Ocean Proximity The violin plot also reinforces that houses near the ocean or bay tend to have higher median values. The shape of the violins shows that houses <1H from the ocean have a wider distribution, suggesting variability in prices. Inland homes have a lower median value and a narrower distribution, indicating consistency in the prices of properties located further from water bodies.</p><p>4. Linear Regression Plot of Median Income vs. Median House Value The linear regression plot indicates a positive correlation between median income and median house value. As median income increases, the house value also tends to increase. This suggests that areas with higher incomes are more likely to have higher property values, aligning with economic principles where wealthier areas tend to have more expensive housing.</p><p>5. Non-Linear Regression Plot for Median House Value vs Median Income The non-linear regression plot shows that the relationship between median income and median house value is not strictly linear. Initially, as median income increases, house values rise at a slower rate, but beyond a certain income level, house values start increasing more rapidly. This suggests that in higher-income regions, the house prices escalate sharply, indicating a possible luxury housing market effect or increased demand for premium properties.</p><p>6. Jitter Plot of Median House Value vs Ocean Proximity</p></div>

	<p>The jitter plot indicates that there is a wide spread of house values at various levels of ocean proximity. However, homes closer to the ocean, especially those categorized as <1H OCEAN, tend to have higher median values. The spread of house values is more concentrated for inland properties, which generally have lower median values. This suggests that while ocean proximity does influence house values, there is significant variability in prices for properties near water.</p> <p>The six plots collectively show that ocean proximity and median income are significant factors influencing house values. Most homes are inland, and those closer to the ocean tend to have higher and more variable prices. Both linear and non-linear regression plots confirm a positive relationship between median income and house values, with wealthier areas showing steeper price increases. The jitter plot reinforces the variability of house values near the ocean. Overall, homes near water and in high-income areas are associated with higher prices.</p>
--	--