BHARATIYA VIDYA BHAVAN'S
SARDAR PATEL INSTITUTE OF TECHNOLOGY
(Empowered Autonomous Institute Affiliated to University of Mumbai)
[Knowledge is Nectar]
Department of Computer Engineering

Name: Adwait Purao
UID: 2021300101
Batch: D

Aim:

Create advanced charts using Power BI / Tableau / R / Python / D3.js on the dataset - Housing data

- Advanced - Word chart, Box and whisker plot, Violin plot, Regression plot (linear and nonlinear), 3D chart, Jitter
- Write observations from each chart

Theory:

Dataset:
https://www.kaggle.com/datasets/ashydv/housing-dataset

Dataset Description:
The dataset contains information about house sales, with details on various attributes such as the house's area, number of bedrooms and bathrooms, availability of parking, and other characteristics. The dataset is structured for use in linear and logistic regression modeling, allowing analysis of the relationship between house prices and multiple predictors.

Column Descriptions:

1. **price**: The sale price of the house (target variable in the linear regression).
2. **area**: The total area of the house in square feet.
3. **bedrooms**: The number of bedrooms in the house.
4. **bathrooms**: The number of bathrooms in the house.
5. **stories**: The number of stories or floors in the house.
6. **mainroad**: A binary variable indicating if the house has direct access to a main road (1 = Yes, 0 = No).
7. **guestroom**: A binary variable indicating if the house has a guestroom (1 = Yes, 0 = No).
8. **basement**: A binary variable indicating if the house has a basement (1 = Yes, 0 = No).
9. **hotwaterheating**: A binary variable indicating if the house has hot water heating (1 = Yes, 0 = No).
10. **airconditioning**: A binary variable indicating if the house has air conditioning (1 = Yes, 0 = No).

11. **parking**: The number of parking spaces available with the house.
12. **prefarea**: A binary variable indicating if the house is in a preferred area (1 = Yes, 0 = No).
13. **furnishingstatus**: A categorical variable describing the furnishing status of the house (0 = Unfurnished, 1 = Semi-Furnished, 2 = Fully Furnished).

Charts:

1. Regression Plot - Linear Regression

A linear regression plot is used to visualize the relationship between two continuous variables and to fit a linear model that describes this relationship. The plot typically shows data points scattered on a graph, with a line of best fit that minimizes the sum of the squared differences between the observed and predicted values.

Chart:

Observation:

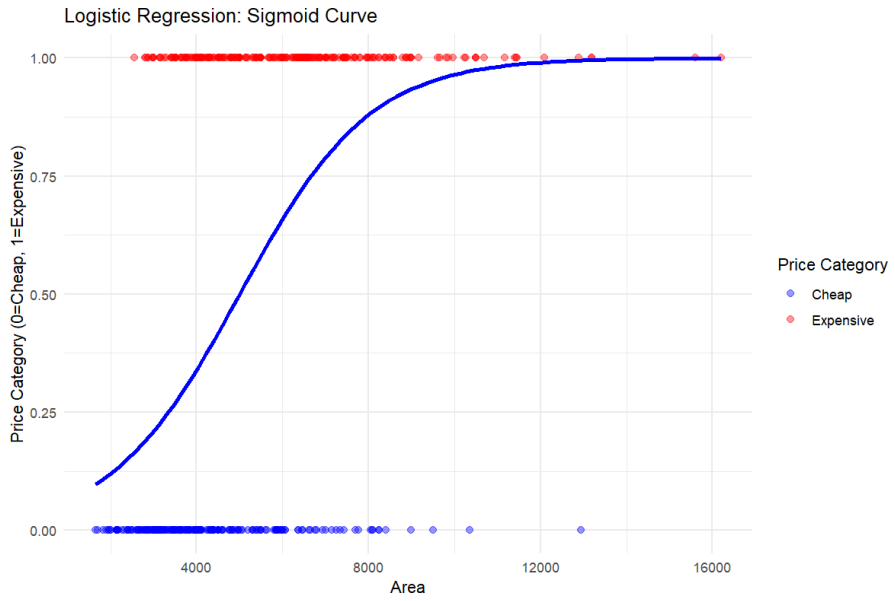
- Positive Correlation: The scatter plot shows a general upward trend, indicating a positive relationship between house price and area. As the area increases, the price tends to rise as well.
- High Variability: Although there is a positive trend, there is significant dispersion in the data points, suggesting that factors other than area also play a substantial role in determining house prices.

- Outliers: A few outliers are visible, especially for larger areas where the price seems to deviate significantly from the general trend. These could represent luxury or unique properties.

2. Regression Plot - Logistic Regression

Logistic regression is used to model the probability of a binary outcome (0 or 1, true or false) as a function of one or more independent variables. The logistic regression plot typically shows the probability of the outcome on the y-axis and the independent variable on the x-axis, with a sigmoid curve representing the probability function.

Chart:



Observation:

- Clear Separation of Categories: The logistic regression curve shows a clear distinction between cheap (0) and expensive (1) houses based on area. As the area increases, houses are more likely to fall into the "expensive" category.
- S-Shaped Sigmoid Function: The sigmoid curve properly fits the data, indicating that small area houses are more likely to be cheap, while larger area houses have a higher probability of being expensive. The transition occurs between approximately 4000 to 8000 square feet.
- Prediction Accuracy: There is a relatively good alignment between the observed data points and the sigmoid curve, suggesting the model is capturing the

relationship between area and price category effectively. However, a few misclassified points (in red and blue) are visible, which may indicate areas where other factors impact price.

3. Word Cloud

A word cloud is a visual representation of text data, where the size of each word indicates its frequency or importance within the dataset. Words that appear more frequently are displayed in larger, bolder fonts, while less frequent words are smaller. Word clouds are often used for quickly identifying the most prominent terms in a large corpus of text.

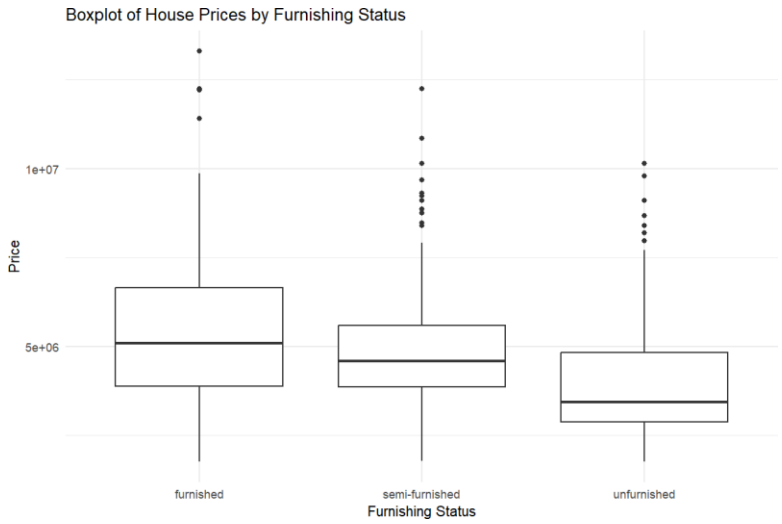
Chart:

Observations:
The word cloud represents the relative frequencies of the terms 'semi-furnished', 'furnished' and 'unfurnished'. Based on the size and prominence of each word, it is seen that semi-furnished is the most common furnishing type among the data points. This suggests that a significant portion of the housing units in the dataset are partially furnished, indicating a preference for this option among potential renters or buyers.

4. Box and Whisker Plot

A box and whisker plot (or box plot) is a graphical representation of the distribution of a dataset. It displays the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values. The "box" represents the interquartile range (IQR), containing the middle 50% of the data, while the "whiskers" extend to the minimum and maximum values within 1.5 times the IQR. Outliers are often shown as individual points beyond the whiskers.

Chart:



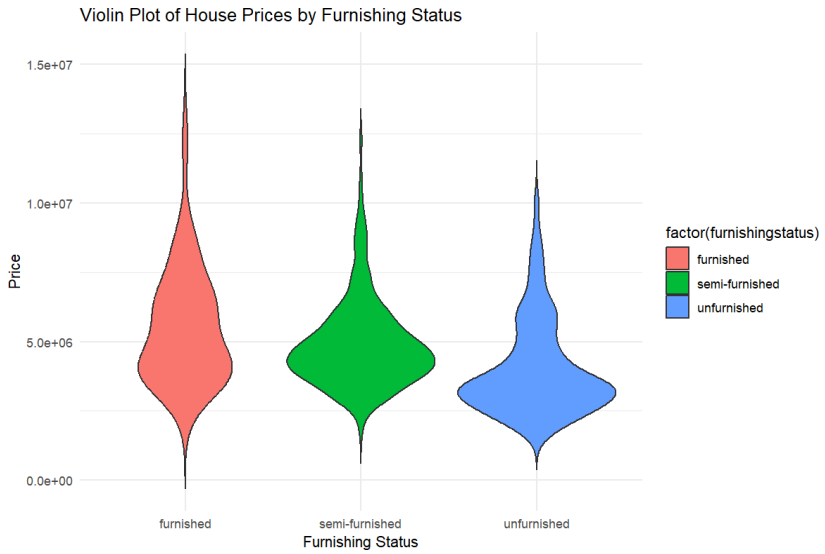
Observations:

- **Higher Prices for Furnished Houses**: Houses that are fully furnished tend to have the highest median prices, as indicated by the higher position of the box for this category, suggesting that furnishing status positively impacts house prices.
- **Wider Price Range for Semi-Furnished and Unfurnished**: Both semi-furnished and unfurnished houses exhibit a wider interquartile range (IQR) compared to furnished houses. This indicates greater variability in prices for these categories, likely due to other factors affecting the price, such as location or size.
- **Outliers**: All categories show several outliers, but semi-furnished and unfurnished houses have more extreme high-value outliers, suggesting that even though these categories generally have lower prices, there are some exceptions where semi-furnished or unfurnished houses are sold at higher-than-expected prices.

5. Violin Plot

A violin plot combines elements of a box plot and a kernel density plot. It shows the distribution of the data across different values by depicting the density of the data at different values along the y-axis. The width of the "violin" at different y-values indicates the density of the data, while the inner box plot shows the interquartile range and median.

Chart:



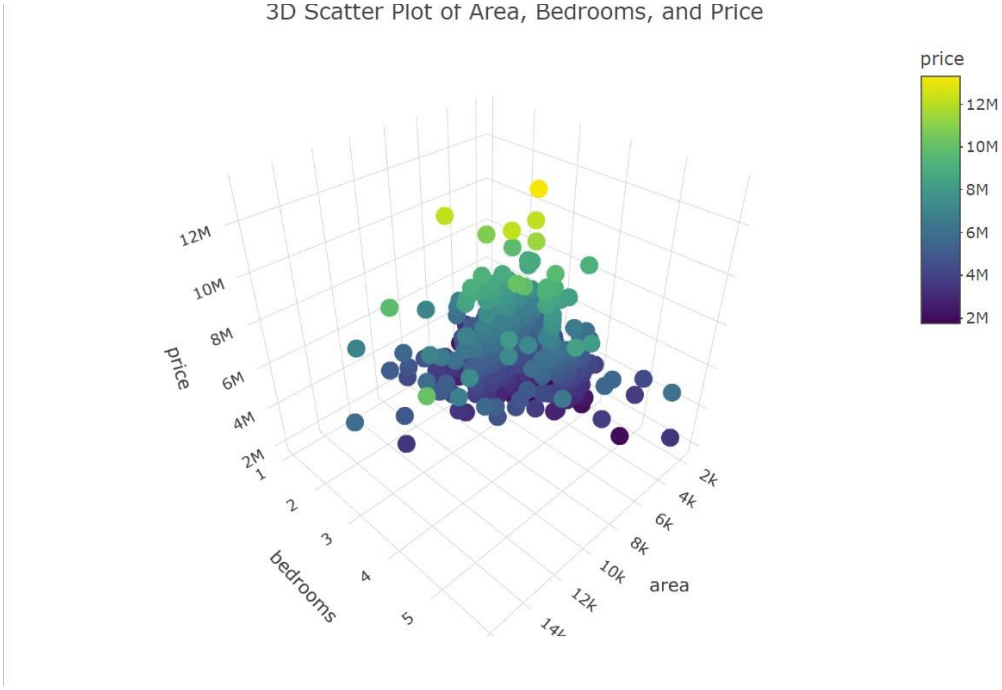
Observations:

- Furnishing Status and Price: The violin plot clearly indicates a strong relationship between the furnishing status of a house and its sale price. Fully furnished houses tend to have the highest median price, while unfurnished houses have the lowest. This suggests that furnishing can significantly influence a home's perceived value.
- Price Distribution: The violin plots reveal the distribution of prices within each furnishing category. While the median price is highest for fully furnished houses, there's also a wider range of prices, suggesting that other factors besides furnishing can contribute to price variation.
- Outliers: The plots also highlight potential outliers, particularly in the fully furnished category. These might represent unique properties with exceptionally high or low prices, which could be worth investigating further to understand their influencing factors.

5. 3D Scatter Plot

A 3D scatter plot is a graphical representation used to display data points in three dimensions, allowing the visualization of relationships between three variables. Each point in the plot represents an observation with coordinates (x, y, z), corresponding to its values on the three axes. This type of plot is useful for identifying patterns, clusters, or trends in complex datasets, such as the correlation between multiple variables or the distribution of data points in space. It's often used in data analysis, machine learning, and scientific research to explore multidimensional data visually.

Chart:



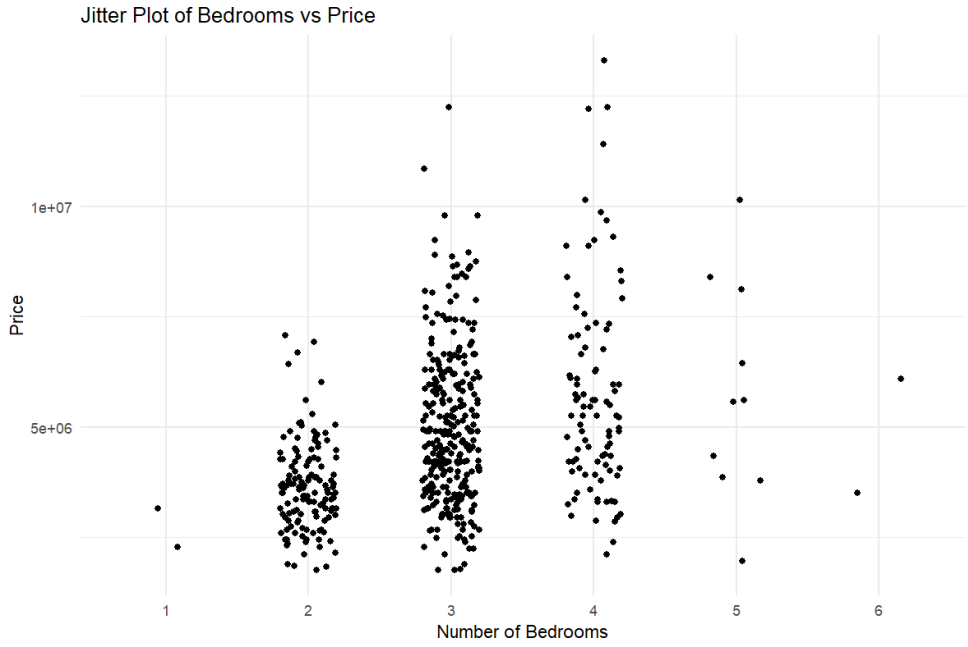
Observations:

- Positive Correlation with Area and Price: As expected, there's a clear positive correlation between the area of a house and its price. Larger houses tend to have higher prices, which is reflected in the upward trend along the area axis.
- Bedrooms and Price: While there's a general trend for houses with more bedrooms to have higher prices, the relationship is less pronounced than with area. This suggests that other factors, such as the size of the bedrooms or the overall layout, might influence the price more significantly.
- Clustering: The scatter plot shows distinct clusters of data points, indicating potential groupings based on similar combinations of area and bedrooms. These clusters might represent different market segments or price ranges.

5. Jitter Plot

A jitter plot is a variation of a scatter plot that introduces slight random noise (or "jitter") to the position of data points to prevent overplotting. Overplotting occurs when multiple data points overlap, making it difficult to distinguish them, especially in datasets with many identical or similar values. By adding jitter to the points, the plot spreads out the overlapping data slightly, improving visibility without altering the actual values. Jitter plots are particularly useful for displaying categorical data, where many points might fall on the same axis value, and for exploring distributions and patterns in densely packed data.

Chart:



Observations:

- Positive Correlation: The jitter plot confirms a positive correlation between the number of bedrooms and the price of a house. As the number of bedrooms increases, the overall trend is for prices to rise.
- Price Variation Within Bedroom Categories: While there's a general upward trend, the plot also highlights significant price variation within each bedroom category. This indicates that other factors, such as the size of the bedrooms, overall living space, or location, play a crucial role in determining the final price.
- Outliers: A few outliers are visible, particularly for houses with fewer bedrooms but relatively high prices. These might represent unique properties with exceptional features or located in highly desirable areas.

Conclusion:

In this experiment, we explored both linear and logistic regression models to analyze house prices based on various features such as area, bedrooms, bathrooms, and additional amenities. The linear regression model identified significant predictors of house prices, with area being a key driver, as visualized in the "Price vs Area" plot. The logistic regression model categorized houses as "cheap" or "expensive," providing a probabilistic understanding of price categories based on the same set of features. Complementing these models, visualizations such as the word cloud, box plots, and violin plots provided deeper insights into the impact of furnishing status on price. Finally, the 3D scatter plot and jitter plot helped visualize the relationships between area, bedrooms, and price. Overall, the experiment highlights the strong influence of multiple factors on house pricing, confirming that features like area, number of bedrooms, and amenities significantly affect a home's market value.