**PROJECT OUTLINE**

1. Date
2. Project title
3. Content
4. Objective
5. Introduction
6. Methodology
7. Data gathering/source
8. Dataset description
9. Tools
10. Data cleaning and preparation
11. Data model
12. Data insights and recommendations
13. Dashboard
14. Dashboard wireframe
15. Visual colours
16. Dashboard colour
17. Work on the attached data freely and creatively to extract different insights
18. Add at least 5 questions and answer them through your analysis of the data
19. Extract statistical summaries using formulas and functions.
    1. Mean,
    2. Median,
    3. Mode,
    4. Range,
    5. Box plot,
    6. Q1,
    7. Q2,
    8. Q3,
    9. Q4,
    10. Standard deviation,
    11. Variance,
    12. Histogram,
    13. Standardization with Z-test,
    14. Hypothesis testing (Null, alternative, and T-test),
    15. Prediction
20. Apply some visual charts (we will study them later) to the data

**Data Cleaning** Before you can start the analysis, you need to ensure that your dataset is clean and ready for analysis. Here are some data-cleaning steps you might consider:

1. **Duplicate Removal**: Remove any duplicate rows in the dataset.
2. **Missing Data Handling**: Check for missing values and decide whether to impute them or remove rows/columns with too many missing values.
3. **Data Type Conversion**: Ensure that data types are appropriate for analysis (e.g., numeric fields should be of numeric data type). Ensure that the data types of each column are appropriate. For example, convert binary columns ('Yes/No') to 0/1 or Boolean values.
4. **Consistency Check**: Verify that data values are consistent and within expected ranges.
5. **Outlier Detection**: Identify and handle outliers, if any.
6. **Standardize Categories**: Ensure consistent naming conventions for categories.
7. **Data Transformation**: If necessary, transform data.

**Data Exploration and Insights:**

1. **Price Distribution**: Calculate summary statistics (mean, median, standard deviation) for the 'Price' column to understand the average and variability of house prices.
2. **Bedroom and Bathroom Analysis**: Calculate average house prices for different numbers of bedrooms and bathrooms. Do houses with more bedrooms and bathrooms tend to have higher prices?
3. **Categorical Features Analysis**: Explore how features like 'Mainroad,' 'Guestroom,' 'Basement,' 'Hot water heating,' 'Air conditioning,' 'Prefarea,' and 'Furnishing status' influence house prices. Calculate average prices for different categories within these features.
4. **Parking Analysis**: Calculate average prices for houses with different numbers of parking spaces. Are houses with more parking spaces more expensive?
5. **Area and Stories Analysis**: Analyze the relationship between the 'Area' and 'Stories' features with house prices. Does the number of stories or the area of the house affect the price?

**Key Performance Indicators (KPIs):**

1. **Average House Price**: Calculate the overall average house price as a KPI to understand the general pricing trend in the dataset.
2. **Price per Square Foot**: Calculate the price per square foot by dividing 'Price' by 'Area.' This can be a useful metric for comparing the cost efficiency of different houses.
3. **Percentage of Houses with Air Conditioning**: Calculate the percentage of houses with air conditioning as a KPI.
4. **Percentage of Houses in Preferred Areas**: Calculate the percentage of houses in preferred areas.
5. **Average Number of Bathrooms**: Calculate the average number of bathrooms in the dataset.

**Visual Charts:**

1. **Histogram**: Create histograms to visualize the distribution of 'Price' and 'Area.' This can help identify any skewed data.
2. **Bar Charts**: Create bar charts to visualize the distribution of categorical variables like 'Mainroad,' 'Guestroom,' 'Basement,' 'Hot water heating,' 'Air conditioning,' 'Prefarea,' and 'Furnishing status' in the dataset.
3. **Box Plots**: Use box plots to visualize the spread of house prices by the number of bedrooms, bathrooms, and other relevant factors.
4. **Scatter Plot**: Create a scatter plot to visualize the relationship between 'Area' and 'Price' to see if there's a correlation.
5. **Pie Charts**: Use pie charts to show the percentage of houses with air conditioning and in preferred areas.
6. **Correlation Heatmap**: Create a heatmap to visualize the correlations between numerical variables like 'Price,' 'Area,' 'Bedrooms,' 'Bathrooms,' etc.
7. **Question: What is the average price of houses in preferred areas compared to non-preferred areas?**

Answer: The average price of houses in preferred areas is higher than in non-preferred areas. On average, houses in preferred areas tend to have a higher market value.

1. **Question: How does the number of stories in a house relate to its price?**

Answer: The number of stories in a house has a positive correlation with its price. Generally, houses with more stories tend to be priced higher.

1. **Question: Do fully furnished houses have a higher price compared to semi-furnished or unfurnished houses?**

Answer: Yes, fully furnished houses tend to have a higher average price compared to semi-furnished or unfurnished houses. Buyers are willing to pay more for houses that come with complete furnishings.

1. **Question: What is the average price per square foot in this dataset?**

Answer: The average price per square foot is obtained by dividing the average 'Price' by the average 'Area' of houses. This metric provides an estimate of the cost efficiency of the houses in the dataset.

1. **Question: Is there a difference in house prices based on whether they have a guest room or a basement?**

Answer: Houses with a guest room tend to have a slightly higher average price compared to houses without one. However, the presence of a basement does not seem to have a significant impact on house prices in this dataset.

**Descriptive Statistics for 'Price' (House Prices)**:

* + Mean (Average) Price: Calculate the average house price.
  + Median Price: Calculate the middle value, which represents the price at the 50th percentile.
  + Standard Deviation of Price: Measure the spread or variability in house prices.
  + Minimum and Maximum Price: Find the lowest and highest prices in the dataset.
  + Interquartile Range (IQR): Calculate the range between the 25th and 75th percentiles, which shows the middle 50% of prices.

1. **Price per Square Foot**:
   * To calculate the price per square foot, divide 'Price' by 'Area' for each house.
2. **Percentage of Houses with Air Conditioning**:
   * Calculate the percentage of houses with 'Air conditioning' set to 'Yes' using the count of such houses and the total count of houses.
3. **Percentage of Houses in Preferred Areas**:
   * Calculate the percentage of houses in preferred areas ('Prefarea' set to 'Yes') using the count of such houses and the total count of houses.
4. **Average Number of Bathrooms**:
   * Compute the average number of bathrooms across all houses in the dataset.
5. **Average Price for Houses with a Guest Room and without a Guest Room**:
   * Calculate the average price for houses with 'Guestroom' set to 'Yes' and 'No' separately.
6. **Average Price for Fully Furnished, Semi-Furnished, and Unfurnished Houses**:
   * Calculate the average price for houses with different furnishing statuses: 'Fully Furnished,' 'Semi-Furnished,' and 'Unfurnished.'
7. **Price Statistics by Number of Stories**:
   * Calculate average prices for houses with different numbers of stories (e.g., 1 story, 2 stories, etc.).
8. **Price Statistics by Parking Availability**:
   * Calculate average prices for houses with different numbers of parking spaces (e.g., 0 parking spaces, 1 parking space, etc.).
9. **Histogram for House Prices**:
   * Create a histogram to visualize the distribution of house prices. This will show how prices are distributed across the dataset.
10. **Bar Chart for Categorical Features**:
    * Create bar charts for categorical features like 'Mainroad,' 'Guestroom,' 'Basement,' 'Hot water heating,' 'Air conditioning,' 'Prefarea,' and 'Furnishing status.' These charts will show the distribution of categories and how they relate to house prices.
11. **Box Plot for Bedrooms and Bathrooms**:
    * Use box plots to visualize the distribution of house prices for different numbers of bedrooms and bathrooms. This can help identify outliers and understand the spread of prices within each category.
12. **Scatter Plot for Price vs. Area**:
    * Create a scatter plot to visualize the relationship between house prices and the total area of the house. This can help identify patterns, correlations, and outliers.
13. **Pie Charts for Air Conditioning and Preferred Areas**:
    * Use pie charts to show the percentage of houses with air conditioning and the percentage of houses in preferred areas. This provides a visual representation of the prevalence of these features.
14. **Bar Chart for Number of Stories and Parking Availability**:
    * Create a bar chart to visualize the average prices for houses with different numbers of stories and parking spaces. This can help assess the impact of these features on house prices.
15. **Correlation Heatmap**:
    * Create a heatmap to visualize the correlations between numerical variables (e.g., 'Price,' 'Area,' 'Bedrooms,' 'Bathrooms,' etc.). This can help identify strong and weak correlations between variables.
16. **Box Plot for Furnishing Status**:
    * Use box plots to visualize the spread of house prices for different furnishing statuses, such as 'Fully Furnished,' 'Semi-Furnished,' and 'Unfurnished.' This can help understand how furnishing affects prices.
17. **Line Chart for Price Trends Over Time (if applicable)**:
    * If you have a time-related variable (e.g., date of sale), you can create a line chart to show how house prices have changed over time.
18. **Correlation Matrix Plot**:
    * Create a correlation matrix plot to visualize the relationships between all numerical variables in the dataset. This can help identify multiple variable interactions.

These visualizations will make it easier to explore and communicate insights from the dataset, allowing you to identify patterns, trends, and relationships between variables.

**Descriptive Statistics:**

1. **Summary Statistics for House Prices:**
   * Use the functions AVERAGE, MEDIAN, STDEV, MIN, MAX, and QUARTILE on the 'Price' column to calculate statistics.
2. **Calculate Price per Square Foot:**
   * Create a new column 'Price per Sqft' and divide the 'Price' by 'Area' for each row.

**Data Visualization:**

1. **Histogram for House Prices:**
   * Use the 'Insert' > 'Charts' > 'Histogram' feature to visualize the distribution of house prices.
2. **Bar Chart for Bedrooms vs. Price:**
   * Create a bar chart using 'Insert' > 'Charts' to show the average price for different numbers of bedrooms.

**Correlation Analysis:**

1. **Correlation Matrix:**
   * Utilize the 'Data Analysis' tool (if enabled) or manually calculate correlations between numerical variables like 'Price,' 'Area,' 'Bedrooms,' 'Bathrooms,' etc.
2. **Scatter Plot for Price vs. Area:**
   * Use the 'Insert' > 'Charts' > 'Scatter Plot' to visualize the relationship between house prices and the total area.

**Categorical Analysis:**

1. **Bar Chart for Furnishing Status vs. Price:**
   * Create a bar chart to show the average price for different furnishing statuses ('Fully Furnished,' 'Semi-Furnished,' 'Unfurnished').
2. **Pie Chart for Air Conditioning:**
   * Use a pie chart to display the percentage of houses with air conditioning ('Yes' and 'No').

**Advanced Insights:**

1. **Regression Analysis (if needed):**
   * Utilize Excel's regression analysis tools to understand how different factors (like bedrooms, area, etc.) impact house prices.
2. **Trend Analysis (if applicable):**
   * Use Excel's trendline feature to identify trends over time (if a time-related variable is available).

Data cleaning in Excel involves preparing the dataset for analysis by addressing issues such as missing data, duplicates, inconsistent formats, and errors. Here's a step-by-step guide on how to clean data in Excel:

1. **Open the Excel File**: Open the Excel file containing your dataset.
2. **Identify and Address Missing Data**:
   * Identify and locate missing data, which may be represented as blank cells or placeholders like "N/A" or "Not available."
   * Decide how to handle missing data. Options include:
     + Removing rows with missing data: Select the row(s) with missing data and delete them.
     + Fill in missing data: Use Excel's 'Find & Replace' function or the 'IF' function to populate missing values.
3. **Remove Duplicates**:
   * Check for duplicate rows and eliminate them if necessary.
   * Go to the 'Data' tab and use the 'Remove Duplicates' function to select and remove duplicate rows based on specific columns.
4. **Standardize Data Formats**:
   * Ensure that data in each column follows a consistent format. For example, dates should be in the same date format, and text should be in the same case.
   * Use Excel's text functions like 'PROPER,' 'UPPER,' and 'LOWER' to standardize text case.
5. **Identify and Handle Inconsistent Data**:
   * Check for inconsistent values that represent the same information (e.g., "Yes" and "yes").
   * Use Excel's 'Find & Replace' function to replace inconsistent values with a consistent format.
6. **Check for Outliers**:
   * Identify potential outliers by plotting data on a scatter plot, box plot, or by calculating z-scores.
   * Decide whether to keep or remove outliers based on the context of your analysis.
7. **Data Validation**:
   * Apply data validation rules to ensure that the data entered meets specific criteria. For example, you can restrict the values entered in a cell to a specific range or set of values.
8. **Create a Data Cleaning Log**:
   * Keep a record of the changes made during the data cleaning process. This will help in maintaining transparency and accountability.
9. **Organize and Document Your Changes**:
   * Clearly document the data cleaning steps you've taken, including the rationale for each decision. This documentation will be helpful for reproducibility and auditing.
10. **Save a Cleaned Copy**:
    * Once you've completed the data cleaning process, save a copy of the cleaned dataset separately from the original, so you have both versions for reference.
11. **Test Data Quality**:
    * Conduct sanity checks to ensure that the data is now clean and ready for analysis.

Project Title: "Housing Market Analysis and Price Prediction"

Content: This data analysis project aims to explore and analyze the factors affecting housing prices in a specific area. It will investigate how various features, such as the size of the house, the number of bedrooms and bathrooms, the presence of amenities like air conditioning, and the location, influence the price of houses. The project will also include a predictive model for estimating house prices based on these features.

Objective: The primary objective of this project is to understand the key determinants of housing prices in a particular area and to build a predictive model to estimate house prices accurately. By doing so, we aim to provide valuable insights for homeowners, buyers, and real estate professionals to make informed decisions.

Introduction: The real estate market is a vital part of any economy, and housing prices are influenced by a multitude of factors. This project seeks to gain insights into the relationship between various house features and their impact on housing prices. By analyzing the provided dataset, we can develop a predictive model to help individuals and real estate stakeholders understand the price dynamics better.

Methodology:

1. Data Collection: Collect and clean the dataset, ensuring it is free from missing values and outliers.
2. Data Exploration: Perform descriptive statistics and data visualization to gain initial insights into the dataset.
3. Feature Engineering: Create new features if necessary, and encode categorical variables.
4. Data Analysis: Explore relationships between the features and the target variable (Price).
5. Predictive Modeling: Build regression models (e.g., linear regression, random forest) to predict house prices using the selected features.
6. Model Evaluation: Assess model performance using appropriate metrics (e.g., RMSE, MAE) and cross-validation techniques.
7. Interpretation: Interpret the results to understand the impact of different factors on housing prices.

Data Source: The dataset used in this project is obtained from [Data Source Name]. It comprises the following fields:

* Price: The price of the house.
* Area: The total area of the house in square feet.
* Bedrooms: The number of bedrooms in the house.
* Bathrooms: The number of bathrooms in the house.
* Stories: The number of stories in the house.
* Mainroad: Whether the house is connected to the main road (Yes/No).
* Guestroom: Whether the house has a guest room (Yes/No).
* Basement: Whether the house has a basement (Yes/No).
* Hot water heating: Whether the house has a hot water heating system (Yes/No).
* Air conditioning: Whether the house has an air conditioning system (Yes/No).
* Parking: The number of parking spaces available within the house.
* Prefarea: Whether the house is located in a preferred area (Yes/No).
* Furnishing status: The furnishing status of the house (Fully Furnished, Semi-Furnished, Unfurnished).

This project will help stakeholders in the real estate industry, prospective buyers, and sellers to make informed decisions regarding house prices based on the features and amenities of the property, ultimately contributing to a more transparent and efficient housing market.