Diamond Price Analysis Assignment

Analyze the provided dataset to uncover insights into diamond prices, quality, and customer perceptions. This will involve handling missing data, performing exploratory data analysis, and drawing meaningful conclusions from the data.

Dataset

The dataset contains information about diamonds, including their physical characteristics, prices, and customer perceptions. The columns are as follows:

- ID: Anonymous number assigned to a diamond.
- carat: The weight of the diamond.
- cut: Quality of the cut (Ideal, Premium, Good, etc.).
- colour: Color of the diamond (scale from 0 (colorless) to 10 (light yellow or brown)).
- clarity: Clarity of the diamond, referring to the absence of faults (inclusions).
- depth: Height of the diamond.
- price: The price of the diamond.
- x: Price per carat.
- y: Price per depth.
- P: Perceptions on prices (Positive, Negative, etc.).
- PC: Change in perceptions (Positive, Negative, Somewhat Positive, etc.).

Tasks

1. Data Loading and Exploration

- Load the dataset into a pandas DataFrame.
- Display the first 10 rows of the dataset.
- Print a summary of the dataset, including the data types of each column and the number of non-null entries per column.

2. Handling Missing Values

- Identify and display the total number of missing values in each column.
- Replace missing values in the `carat`, `price`, `x`, and `y` columns with the mean of the respective column.
- Replace missing values in categorical columns (`cut`, `colour`, `clarity`, `P`, `PC`) with the mode.
- Verify that all missing values have been handled.

3. Data Analysis

Descriptive Statistics:

- Calculate and display the mean, median, and mode for `carat`, `price`, `x`, and `y`.

Categorical Data Analysis:

- Create a bar plot to show the distribution of different `cut` qualities.
- Create a count plot to show the distribution of diamond `colour` grades.

Correlation Analysis:

- Calculate and display the correlation matrix for `carat`, `depth`, `price`, `x`, and `y`.
- Create a heatmap to visualize the correlation matrix.

4. Data Visualization

Price Distributions:

- Plot histograms for `carat`, `price`, `x`, and `y`.

Cut Quality vs Price:

- Create a box plot to compare diamond prices across different `cut` qualities.

Impact of Clarity on Price:

- Create a bar plot showing the average price for each `clarity` level.

Customer Perceptions:

- Create a pie chart or bar plot to show the distribution of customer perceptions (`P`) regarding prices.

5. Advanced Analysis

Perception Change Analysis:

- Analyze and visualize how customer perceptions (`P`) change (`PC`) based on diamond characteristics.

Overall Diamond Value:

- Create a new column `total_value` that combines `price`, `x`, and `y` to represent the overall value of a diamond.
- Plot a histogram to analyze the distribution of `total_value` among diamonds.

Submission Guidelines

- Submit the Jupyter notebook (`your_name.ipynb` file) with your analysis via email to mrbasiima@gmail.com (Add your name to the Subject of the email)
- Make sure to include comments in your code explaining each step.
- Ensure that all visualizations are clearly labeled with titles and axis labels.

Evaluation Criteria

- Correctness: Accuracy of the data analysis and handling of missing values.
- Code Quality: Readability, use of comments, and adherence to best practices.
- Visualizations: Clarity and relevance of the plots, including proper labeling and

interpretation.
- Insights: Depth of analysis and ability to extract meaningful insights from the data.

Due Date

The assignment is due at 2pm on Friday 30th August, 2024 from the date of assignment.