**Guiding questions**

**Step-by-Step Analysis of Pizza Price Prediction**

1. **Import Libraries**:
2. **Load Dataset**:
3. **Initial Data Inspection**:

A . \*\*First 5 Rows\*\*: What do the first five rows of the dataset look like?

* + Displayed the first five rows of the dataset using pizza\_df.head().

B. \*\*Shape of the Data Set\*\*:

* + What is the shape of the dataset? (i.e., number of rows and columns)
  + Checked the shape of the dataset using pizza\_df.shape.

C. \*\*Get Information About Our Dataset\*\*:

What information can we get about our dataset, such as the total number of rows, total number of columns, data types of each column, and memory requirement?

1. **Check for Missing Values**: **- Are there any null values in the dataset?**
   * Checked for null values in the dataset using pizza\_df.isnull().sum().
2. **Data Preprocessing**:
   * Renamed the column price\_rupiah to price.
   * Removed 'Rp' and commas from the price column and converted it to integers.
   * Converted the price column from Indonesian Rupiah to Kes.
   * Removed 'inch' from the diameter column and converted it to floats.
   * Displayed the first five rows after preprocessing using pizza\_df.head().
3. **Data Analysis (Univariate Analysis)**:

A. \*\*Univariate Analysis - Company\*\*:

- What are the counts of each company in the dataset?

B. \*\*Univariate Analysis - Price\*\*:

- What is the distribution of prices in the dataset?

C. \*\*Univariate Analysis - Diameter\*\*:

- What are the counts of each diameter size in the dataset?

1. **Bivariate Analysis**:

A. \*\*Bivariate Analysis - Diameter vs. Price\*\*:

- What is the relationship between diameter and price?

B. \*\*Bivariate Analysis - Company vs. Price\*\*:

- What is the relationship between company and price?

C. Create a scatter plots to analyze relationships between price and other numerical features (diameter, topping, size).

1. **Label Encoding**: **Label encoding is a crucial preprocessing step in machine learning that ensures categorical data is transformed into a numerical format suitable for algorithm processing.**

**NB:machine learning algorithms can only work with numerical data that is why we have to change/encode text data to numerical data.**

* + Apply label encoding to categorical features like company, variant, and others using LabelEncoder.

1. **Correlation Analysis**:**\*\*Correlation Heatmap\*\***
2. What are the correlations between different features in the dataset?
3. Compute and plot the correlation matrix to understand relationships between features.
4. **Train-Test Split**:
   * Split the data into training and testing sets using train\_test\_split from sklearn.model\_selection.
5. **Feature Scaling**:Feature scaling is a crucial preprocessing step in machine learning that ensures that the features of the dataset are within a comparable range. This step is essential for many machine learning algorithms to function correctly and efficiently.
   * **Scale the features using StandardScaler.**
6. **Model Building**:**\*\*Model Training\*\*:**

**- How do we train different regression models on the dataset?**

* + Built and trained multiple regression models:
    - Linear Regression (LinearRegression).
    - Support Vector Regressor (SVR).
    - Random Forest Regressor (RandomForestRegressor).
    - Gradient Boosting Regressor (GradientBoostingRegressor).
    - XGBoost Regressor (XGBRegressor).

1. **13. \*\*Model Prediction\*\* and** . \*\*Model Evaluation\*\*:

- How do the trained models predict prices on the test set?

- How do we evaluate the performance of different models using the R² score?

* + Made predictions on the test set using each model.
  + Evaluated the models using r2\_score from sklearn.metrics.
  + Compared model performance by plotting the R2 scores.

1. **Feature Importance**:
   * Extracted and plotted feature importance for RandomForestRegressor, GradientBoostingRegressor, and XGBRegressor.
2. **Saving the Best Model**:
   * Identified the best-performing model (XGBoost in this case).
   * Saved the trained model using joblib.
3. **Model Deployment**:
   * Loaded the saved model.
   * Demonstrated making a prediction with new data.