Over view

**Understanding Customer Preferences and Pricing Strategies**

This analysis explores the factors influencing pizza prices and identifies key customer preferences based on a dataset of 129 pizza orders.

**Key Findings:**

* **Price Distribution:** Pizza prices are skewed to the right, with a majority falling between 20,000 and 80,000 units (assuming local currency). The most frequent price is around 40,000 units.
* **Diameter and Topping:** Larger pizzas and certain toppings like chicken and mozzarella tend to have higher prices. Sausage and onion pizzas are generally less expensive.
* **Company Variation:** Prices vary across companies, with Company A charging the most and Company E charging the least.
* **Customer Preferences:** The "classic" pizza variant is most popular, followed by "meat lovers," "double\_mix," "crunchy," and "new\_york." Medium-sized pizzas are the most popular, while extra sauce and extra cheese are common add-ons.

**Recommendations:**

* **Pricing Strategy:** Consider adjusting prices for pizzas with less popular toppings or smaller sizes to attract more customers.
* **Menu Optimization:** Promote "classic" and "medium" pizzas as popular choices. Offer more variety in toppings to cater to different preferences.
* **Customer Segmentation:** Analyze customer preferences by demographic or location to tailor marketing and promotions.
* **Pricing Analysis:** Continuously monitor pricing trends and adjust strategies accordingly to remain competitive.

By understanding these insights, businesses can optimize their pricing, menu offerings, and marketing efforts to increase sales and customer satisfaction.

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.

Findings

**FINDINGS**

1. **Data Characteristics**:
   * The dataset contains pizza-related features such as company, diameter, topping, variant, size, extra sauce, extra cheese, and extra mushrooms.
   * The dataset required significant preprocessing to clean and format the data correctly.
   * **The dataset has 129 rows and 9 columns.**

**UNIVARIATE ANALYSIS**

1. **Price Distribution:**
   * Prices were adjusted from Indonesian Rupiah to Kes, showing a varied distribution of pizza prices.
   * The price distribution is skewed to the right, with a majority of prices falling between **20,000 and 80,000.**
   * The **mode is around 40,000**, and there are a few **outliers above 120,000**. Overall, most items are affordable, with fewer expensive items.
2. **The diameter distribution:**
   * The diameter distribution is skewed to the right, with most pizzas being small **(8.5-14 inches).**
   * There are a few outliers **(20-22 inches).**
3. **Topping distribution**
   * Overall, chicken, mushrooms, and mozzarella are the dominant toppings with 29, 22, and 21 occurrences, respectively in the dataset.
   * Less popular toppings: Tuna, vegetables, and meat are less popular, with 9, 9, and 8 occurrences, respectively.
4. **Variant distribution**
   * The classic pizza variant is the most popular, followed by "meat lovers," "double\_mix," "crunchy," and "new\_york." Other variants have fewer occurrences.
5. **Size distribution**
   * Most popular size is medium size with 41 occurrences.
   * Other popular sizes: "Small," "large," and "reguler" are also relatively popular, with 22, 21, and 20 occurrences, respectively.Less popular sizes: "Jumbo" and "XL" are less popular, with 15 and 10 occurrences, respectively.
6. **Extra Sauce & Extra Cheese distribution**
   * Over 60% of customers preferred extra sauce and **Extra Cheese** on the pizza.

# **BIVARIATE ANALYSIS**

### **Comparing the target variable (Price) with company,topping and size**

1. **Price By Company:**
   * Company A has the highest average pizza price, while Company E has the lowest. Companies B, C, and D have similar prices.
   * This information can help understand pricing strategies and make informed purchasing decisions.
2. **Price By Toppings:**
   * The price of a pizza is influenced by the topping, with chicken and mozzarella being more expensive options.
   * Sausage and onion pizzas have the lowest prices.
   * There is some variation in prices within each topping category, as indicated by the size of the boxes and whiskers.
   * Black pepper and beef pizzas have some outliers with very high prices.
3. **Price By Size:**
   * Jumbo pizzas tend to have the highest prices, followed by regular and large pizzas.
   * Small and medium pizzas have lower prices, with small pizzas having the lowest average price.
   * There is some variation in prices within each size category, as indicated by the size of the boxes and whiskers.
   * Large and XL pizzas have some outliers with very high prices**.**

**These findings can be useful for understanding the pricing structure,customer preferences and making informed decisions about pricing strategies or target markets.**

1. **Find the most expensive pizza**
   * The most expensive pizza in the dataset is from Company A, costs 162,092.
   * Has a diameter of 18 inches, is topped with mozzarella, has the double\_signature variant, is jumbo sized, has extra sauce, but no extra cheese or extra mushroom**s.**
2. **Feature Importance Relationships:**
   * Diameter and topping showed noticeable relationships with the price of the pizzas.
3. **Model Performance:**
   * The Gradient Booster Regressor model provided the best R2 score of 93%, indicating it as the most accurate model for predicting pizza prices.
   * Feature importance analysis highlighted diameter and topping as significant contributors to the model's predictions.

14. Price Prediction with the saved model

I tested the model to predict the price of the pizza with the following parameter

* + 'company':1,
  + 'diameter':22.0,
  + 'topping':2,
  + 'variant':8,
  + 'size':1,
  + 'extra\_sauce':1,
  + 'extra\_cheese':1,
  + 'extra\_mushrooms':1

**The model predicted the price to be 151,782.52**

**LINK TO THE DATA**

<https://www.kaggle.com/datasets/knightbearr/pizza-price-prediction?select=pizza_v2.csv>