EDS : MINI PROJECT

GUIDED BY,
MADHAVI NIMKAR , DESIGNATION

PRESENTED BY,

644 OMKAR KARLEKAR
646 SARTHAK PAKHARE
654 PRATIKSHA RANMARE

❑ Motivation

1. Many statistical procedures are special cases of (or approximations to) linear regression.

2. Understanding linear regression really well will give you a deeper understanding of statistics in general.

3. Procedures that are special cases of linear regression, or can be well approximated by linear regression are as follows :

   - One/two sample t-test.
   - ANOVA
   - Correlation tests
   - Rank tests
   - Chi-square tests
   - Many others

INTRO:

- LINEAR REGRESSION :

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Understanding linear regression is important because it provides a scientific calculation for identifying and predicting future outcomes. The ability to find predictions and evaluate them can help provide benefits to many businesses and individuals, like optimized operations and detailed research materials.

Data Manipulation

- **Importing Libraries :** Begin by importing the necessary libraries, such as pandas and NumPy, for data manipulation and analysis.

- **Handling Missing Data :** Check for missing values in the dataset. If any are found, you can handle them by either removing the corresponding rows or filling in the missing values with appropriate methods like mean, median, or interpolation.

Data Visualization

- **Scatter Plots** : Scatter plots are used to visualize the relationship between two continuous variables.
- **Residual Plots:** Residual plots are used to assess the goodness of fit of a linear regression model.
- **Regression Line Plots:** Regression line plots show the fitted regression line along with the observed data points.
- **Histograms and Density Plots:** Histograms and density plots can be used to visualize the distribution of variables.

❑ Details of Dataset

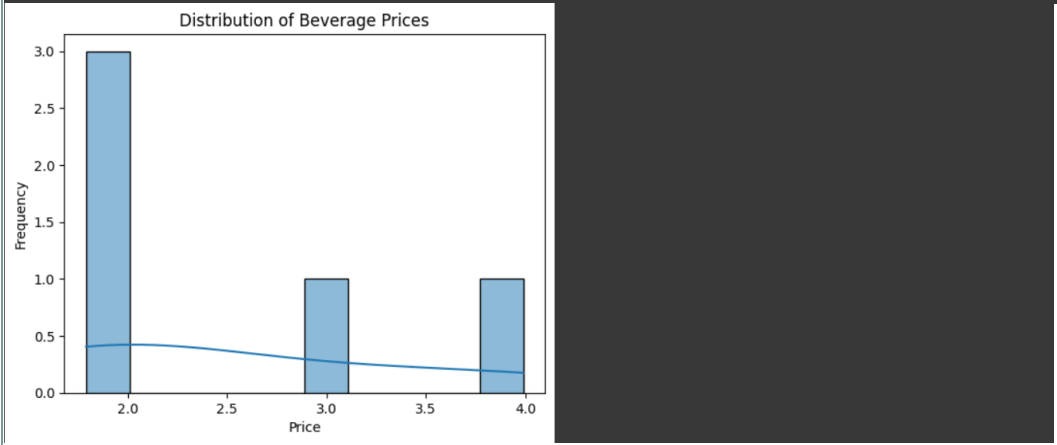
Beverage_Dataset_CSV

Name	Category	Price	Quantity
Coca-Cola	Soda	1.99	100
Pepsi	Soda	1.89	120
Sprite	Soda	1.79	80
Red Bull	Energy Drink	2.99	75
Starbucks	Coffee	3.99	70

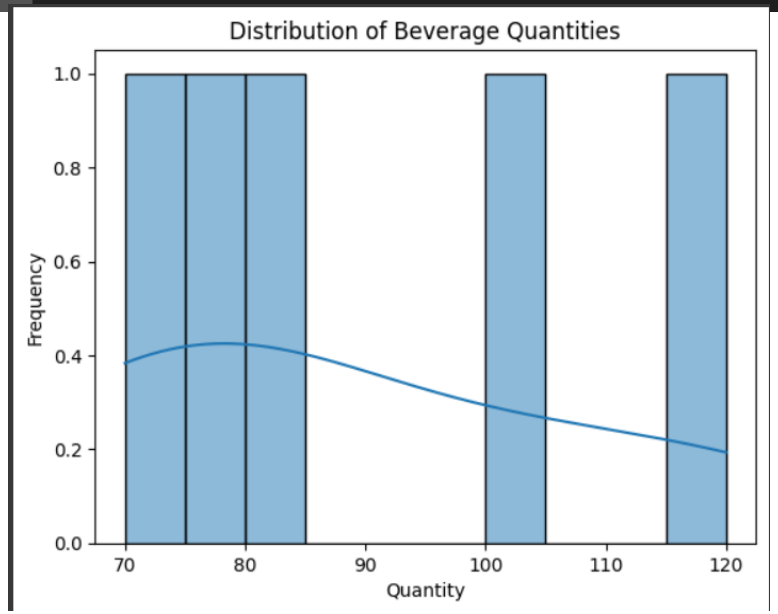
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
df = pd.read_csv('/content/Name,Category,Price,Quantity.csv')
```

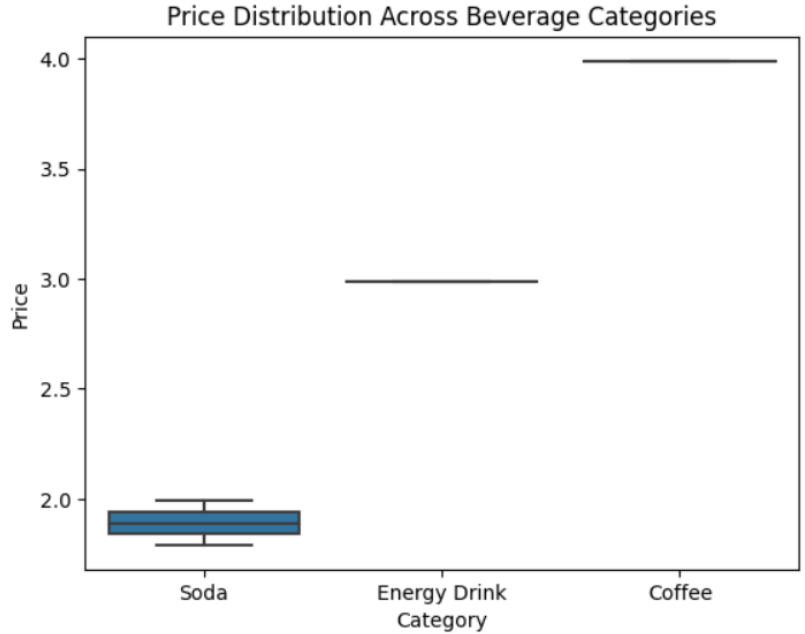
```python
#What is the distribution of beverage prices in the dataset?
# Create a histogram of beverage prices
sns.histplot(data=df, x='Price', bins=10, kde=True)

# Set plot labels and title
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.title('Distribution of Beverage Prices')

# Display the plot
plt.show()
```



```python
#What is the distribution of beverage quantities in the dataset?
# Create a histogram of beverage quantities
sns.histplot(data=df, x='Quantity', bins=10, kde=True)

# Set plot labels and title
plt.xlabel('Quantity')
plt.ylabel('Frequency')
plt.title('Distribution of Beverage Quantities')

# Display the plot
plt.show()
```



```python
#How does the price vary across different beverage categories?
# Create a box plot of prices for each category
sns.boxplot(data=df, x='Category', y='Price')

# Set plot labels and title
plt.xlabel('Category')
plt.ylabel('Price')
plt.title('Price Distribution Across Beverage Categories')

# Display the plot
plt.show()
```



```
[ ]  #What is the average price for each beverage category?
     # Calculate the average price for each category
     average_price = df.groupby('Category')['Price'].mean()

     # Print the average prices
     print(average_price)
```

```
Category
Coffee          3.99
Energy Drink    2.99
Soda            1.89
Name: Price, dtype: float64
```

```
#What is the total quantity sold for each beverage category?
# Calculate the total quantity sold for each category
total_quantity = df.groupby('Category')['Quantity'].sum()

# Print the total quantities
print(total_quantity)
```

```
Category
Coffee           70
Energy Drink     75
Soda            300
Name: Quantity, dtype: int64
```

❑ Application

Linear regression is a statistical measure that establishes the relationship between variables that businesses use to develop forecasts and make informed decisions. It has applications in finance, business planning, marketing, health and medicine.

❑ Conclusion

Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship.

THANK YOU!