Statistics and Trends

Name:-

Student Number:-

Github Link:-

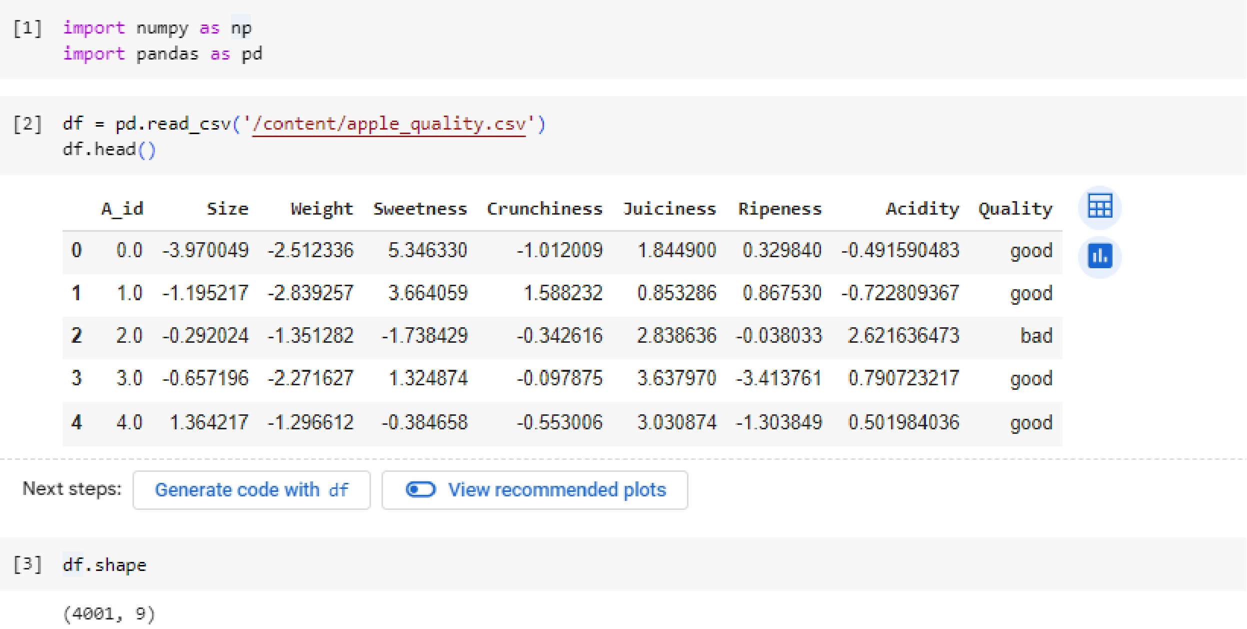
Introduction:-

The analysis aims to explore and understand the patterns within the provided dataset, focusing on various attributes related to apple quality. The dataset comprises information on Apple ID, Size, Weight, Sweetness, Crunchiness, Juiciness, Ripeness, Acidity, and Quality. The analysis involves data loading, preprocessing, and visualization to draw insights into the distribution and relationships within the data.

Data Description

Dataset Overview:-

The dataset consists of 4001 entries with nine columns, including both numerical and categorical features. The attributes range from Apple ID to Quality, with some missing values in the 'Acidity' column. The dataset was preprocessed by dropping missing values, ensuring a clean dataset for analysis.



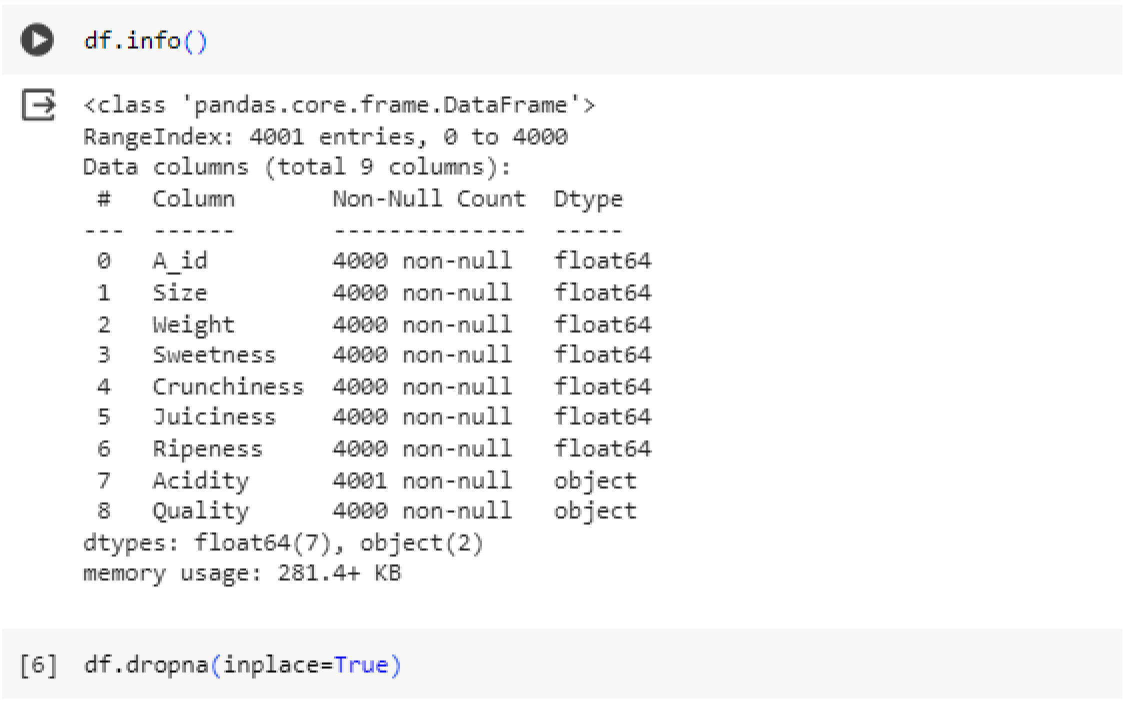
```
[1]  import numpy as np
     import pandas as pd
```

```
[2]  df = pd.read_csv('/content/apple_quality.csv')
     df.head()
```

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Quality
0	0.0	-3.970049	-2.512336	5.346330	-1.012009	1.844900	0.329840	-0.491590483	good
1	1.0	-1.195217	-2.839257	3.664059	1.588232	0.853286	0.867530	-0.722809367	good
2	2.0	-0.292024	-1.351282	-1.738429	-0.342616	2.838636	-0.038033	2.621636473	bad
3	3.0	-0.657196	-2.271627	1.324874	-0.097875	3.637970	-3.413761	0.790723217	good
4	4.0	1.364217	-1.296612	-0.384658	-0.553006	3.030874	-1.303849	0.501984036	good

```
[3]  df.shape
```

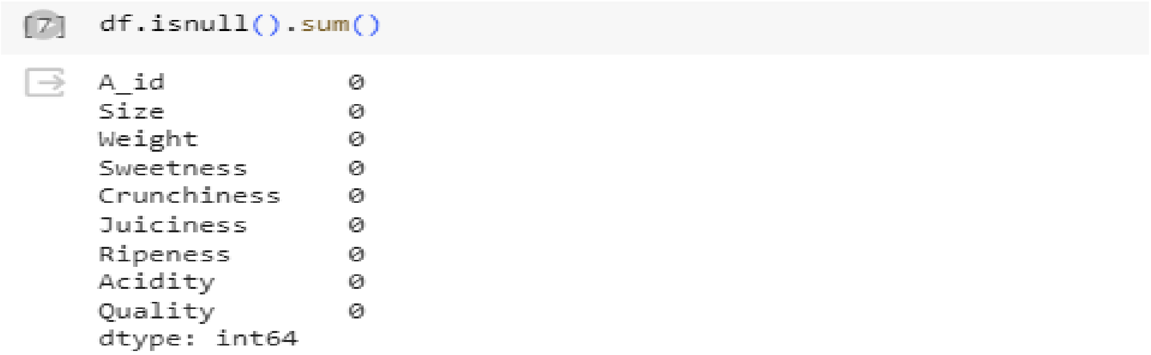
```
(4001, 9)
```



```
df.info()
```

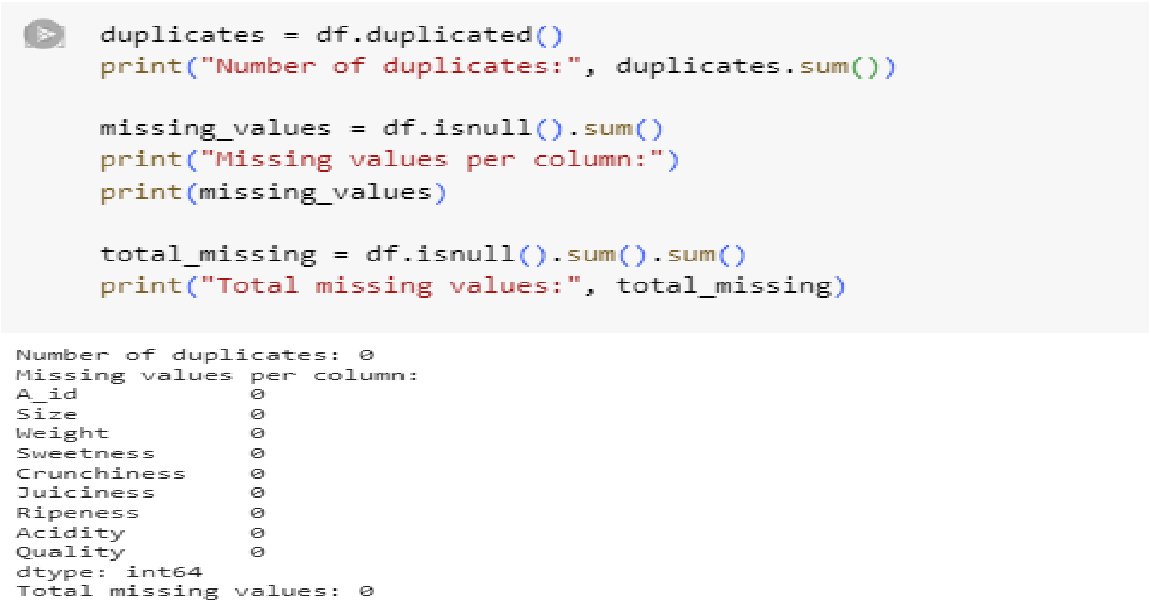
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4001 entries, 0 to 4000
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   A_id         4000 non-null   float64
 1   Size         4000 non-null   float64
 2   Weight       4000 non-null   float64
 3   Sweetness    4000 non-null   float64
 4   Crunchiness  4000 non-null   float64
 5   Juiciness    4000 non-null   float64
 6   Ripeness     4000 non-null   float64
 7   Acidity      4001 non-null   object 
 8   Quality      4000 non-null   object 
dtypes: float64(7), object(2)
memory usage: 281.4+ KB
```

```
[6] df.dropna(inplace=True)
```



```
[7] df.isnull().sum()
```

```
A_id           0
Size           0
Weight         0
Sweetness      0
Crunchiness    0
Juiciness      0
Ripeness       0
Acidity        0
Quality        0
dtype: int64
```

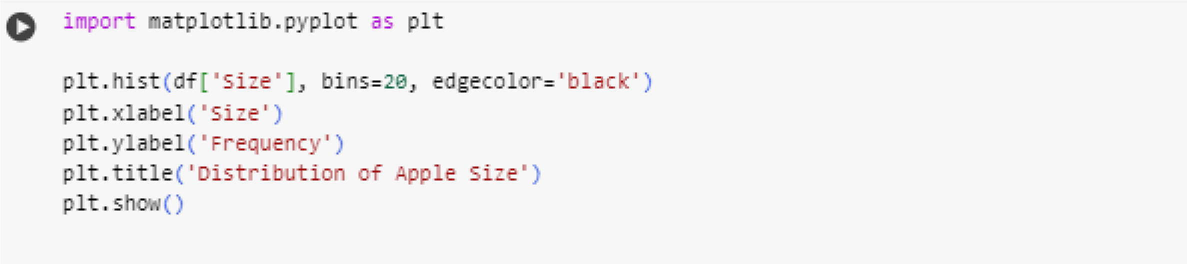


```
duplicates = df.duplicated()
print("Number of duplicates:", duplicates.sum())

missing_values = df.isnull().sum()
print("Missing values per column:")
print(missing_values)

total_missing = df.isnull().sum().sum()
print("Total missing values:", total_missing)
```

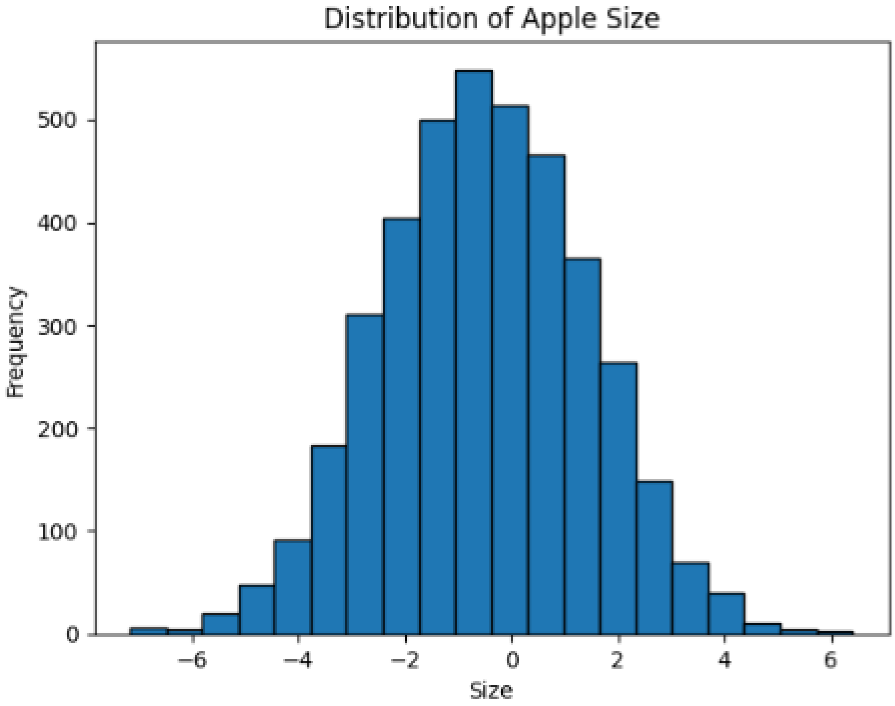
```
Number of duplicates: 0
Missing values per column:
A_id           0
Size           0
Weight         0
Sweetness      0
Crunchiness    0
Juiciness      0
Ripeness       0
Acidity        0
Quality        0
dtype: int64
Total missing values: 0
```

Visualization:-

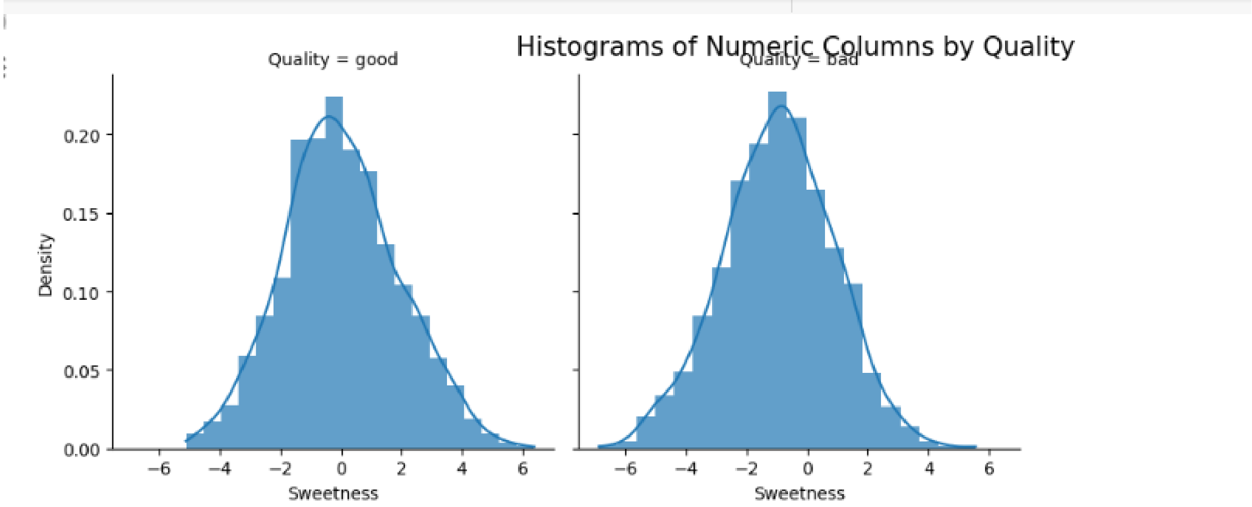
Histogram:-



```
import matplotlib.pyplot as plt

plt.hist(df['Size'], bins=20, edgecolor='black')
plt.xlabel('Size')
plt.ylabel('Frequency')
plt.title('Distribution of Apple Size')
plt.show()
```



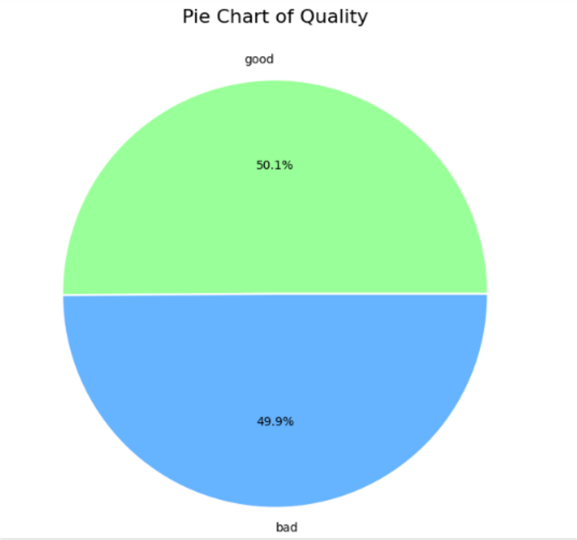
In the histogram, we visualized the distribution of apple sizes in the dataset. The histogram displays the frequency of different size ranges, divided into 20 bins. The x-axis represents the size of apples, while the y-axis shows the corresponding frequency of each size range. The uniform distribution observed suggests that apple sizes are evenly spread across the dataset, indicating a diverse representation of apple sizes in the given data.



Numeric Columns by Quality

A facet grid of histograms further explores numeric attributes concerning the 'Quality' category. The histograms are color-coded for 'good' and 'bad' qualities.

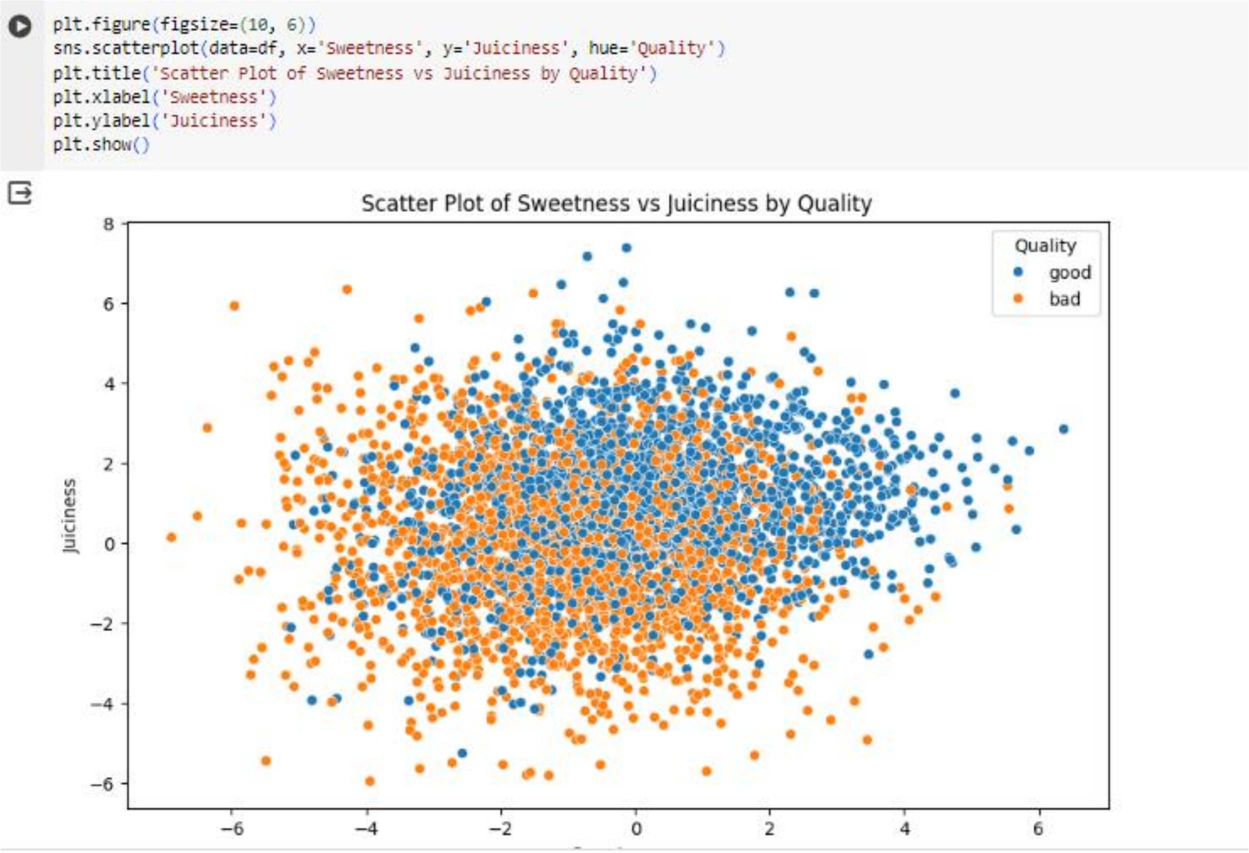
Pie chart:-



Pie Chart of Quality

A pie chart illustrates the distribution of 'Quality' values in the dataset, indicating the proportion of 'good' and 'bad' apples.

Scatter Plot:-

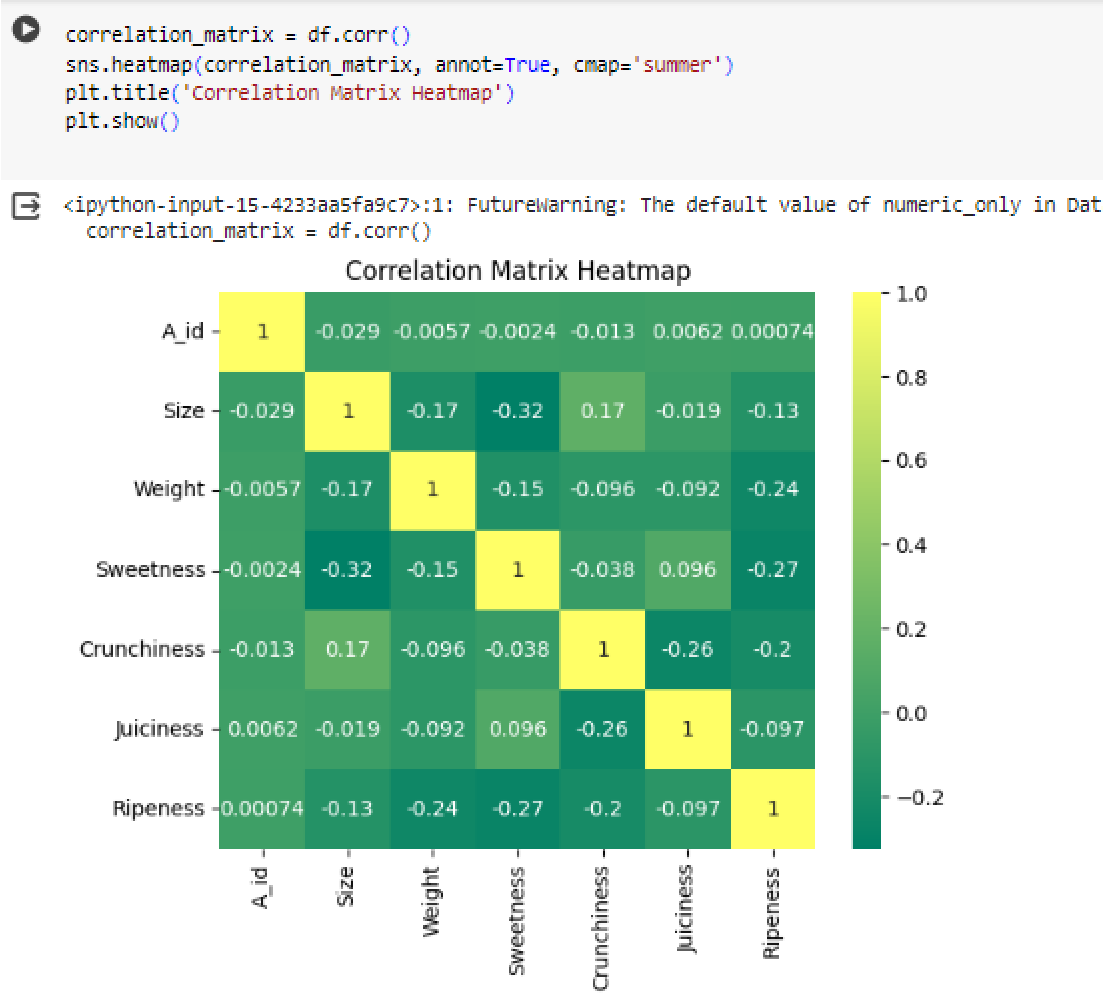


```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Sweetness', y='Juiciness', hue='Quality')
plt.title('Scatter Plot of Sweetness vs Juiciness by Quality')
plt.xlabel('Sweetness')
plt.ylabel('Juiciness')
plt.show()
```

Scatter Plot of Sweetness vs Juiciness by Quality

A scatter plot visualizes the relationship between Sweetness and Juiciness, differentiating 'good' and 'bad' qualities.

Correlation Matrix Heatmap:-



```
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='summer')
plt.title('Correlation Matrix Heatmap')
plt.show()
```

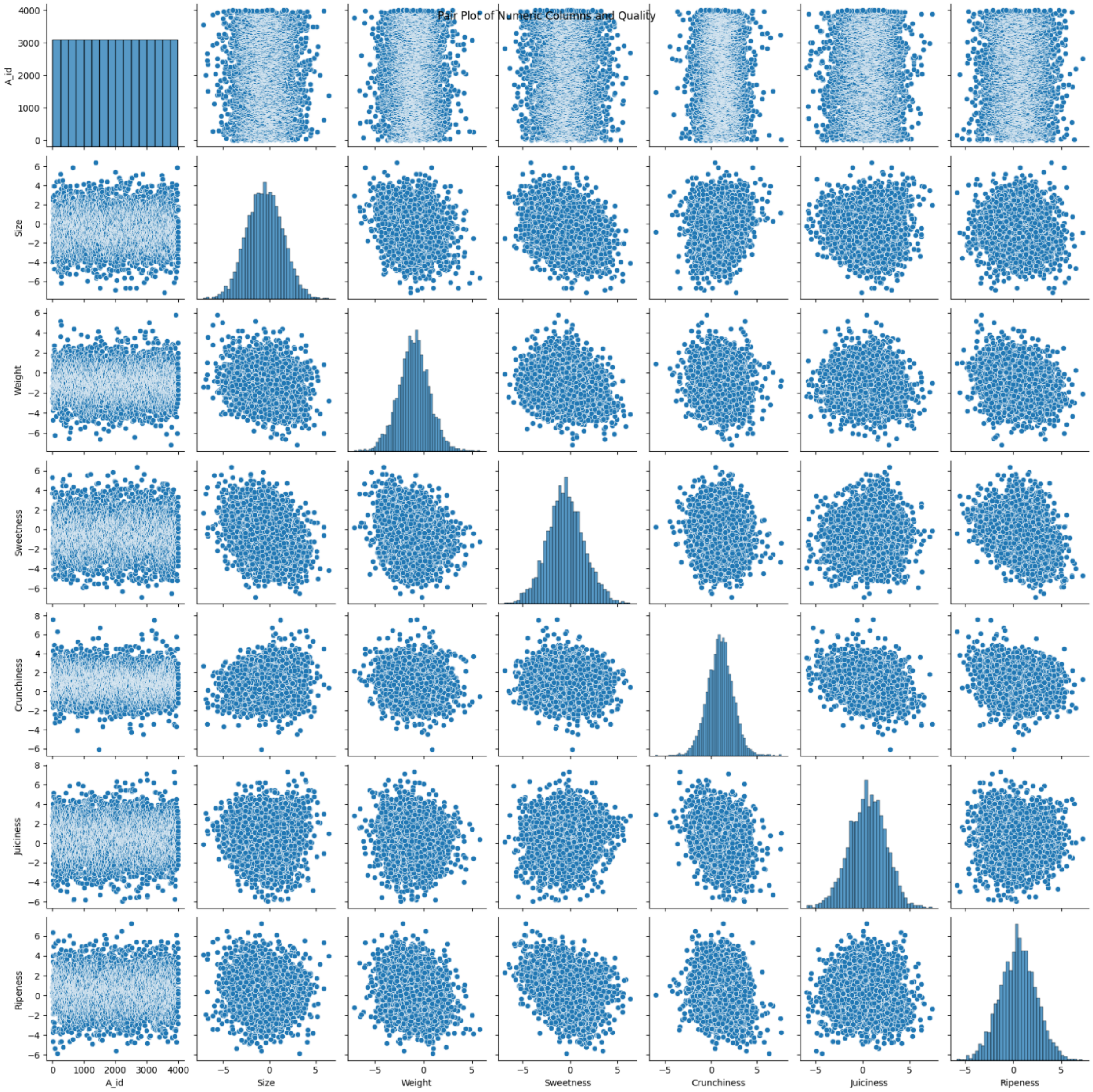
```
<ipython-input-15-4233aa5fa9c7>:1: FutureWarning: The default value of numeric_only in Dat
  correlation_matrix = df.corr()
```

Correlation Matrix Heatmap

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness
A_id	1	-0.029	-0.0057	-0.0024	-0.013	0.0062	0.00074
Size	-0.029	1	-0.17	-0.32	0.17	-0.019	-0.13
Weight	-0.0057	-0.17	1	-0.15	-0.096	-0.092	-0.24
Sweetness	-0.0024	-0.32	-0.15	1	-0.038	0.096	-0.27
Crunchiness	-0.013	0.17	-0.096	-0.038	1	-0.26	-0.2
Juiciness	0.0062	-0.019	-0.092	0.096	-0.26	1	-0.097
Ripeness	0.00074	-0.13	-0.24	-0.27	-0.2	-0.097	1

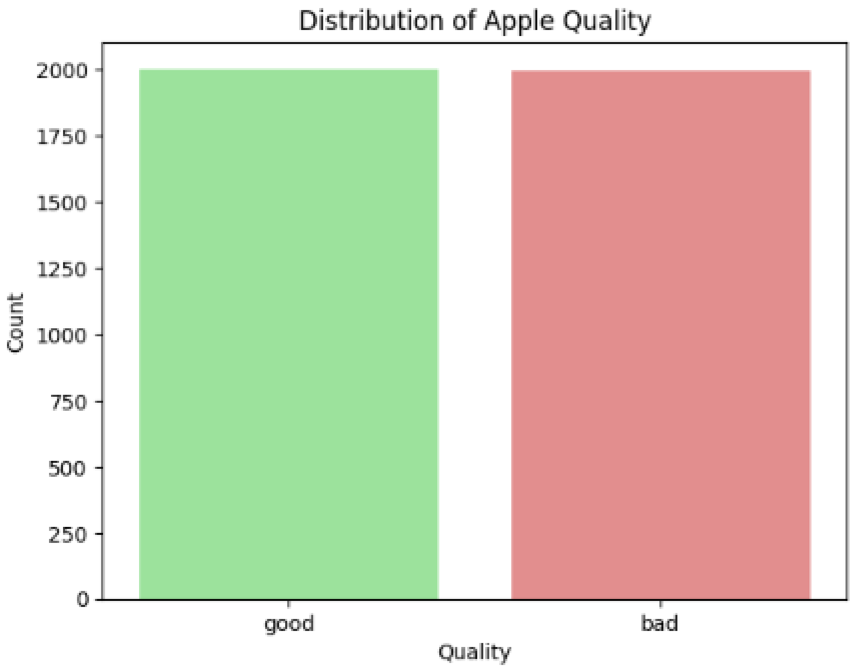
A heatmap provides a graphical representation of the correlation matrix, making it easy to identify strong correlations.

Pair Plot:-



A pair plot visualizes pairwise relationships among numeric columns and quality, offering insights into potential patterns

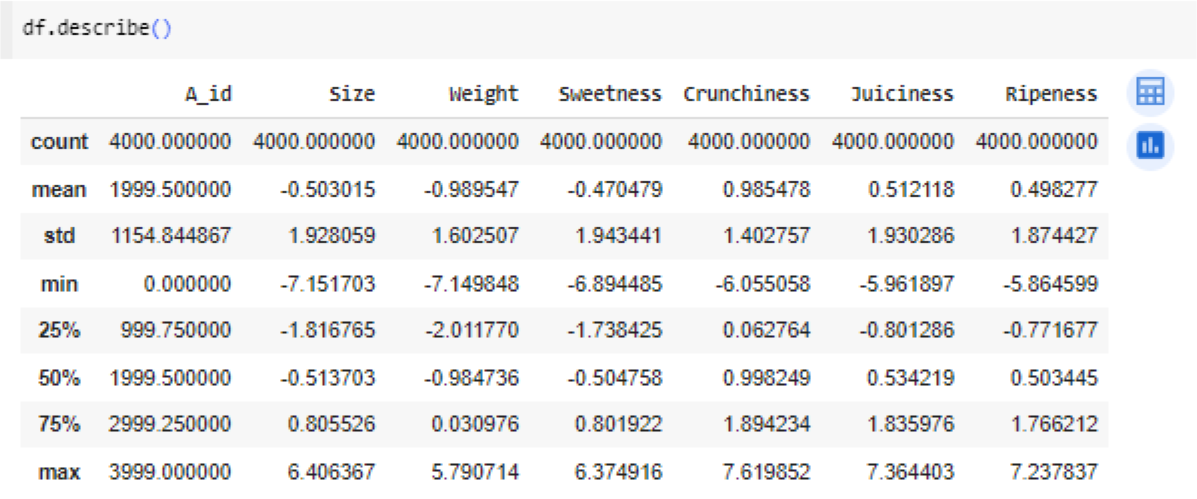
Count Plot:-



Count plot displays the distribution of apple quality, using light colors to distinguish between 'good' and 'bad' qualities.

Statistics:-

Descriptive Statistics:-

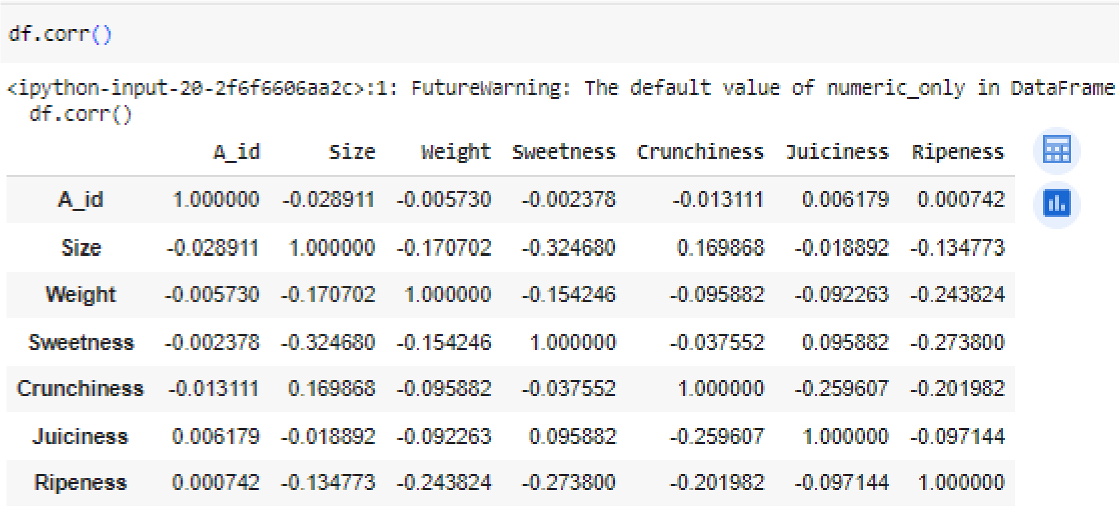


```
df.describe()
```

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness
count	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000
mean	1999.500000	-0.503015	-0.989547	-0.470479	0.985478	0.512118	0.498277
std	1154.844867	1.928059	1.602507	1.943441	1.402757	1.930286	1.874427
min	0.000000	-7.151703	-7.149848	-6.894485	-6.055058	-5.961897	-5.864599
25%	999.750000	-1.816765	-2.011770	-1.738425	0.062764	-0.801286	-0.771677
50%	1999.500000	-0.513703	-0.984736	-0.504758	0.998249	0.534219	0.503445
75%	2999.250000	0.805526	0.030976	0.801922	1.894234	1.835976	1.766212
max	3999.000000	6.406367	5.790714	6.374916	7.619852	7.364403	7.237837

The descriptive statistics summary provides a comprehensive overview of the dataset's central tendencies and variability. It includes key metrics such as mean, standard deviation, and quartiles for each numerical attribute, offering insights into the distribution and range of values within the dataset.

Correlation Matrix:-



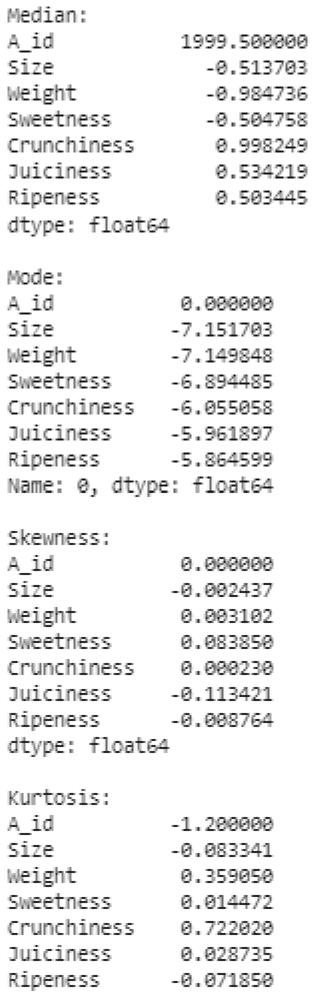
```
df.corr()
```

```
<ipython-input-20-2f6f6606aa2c>:1: FutureWarning: The default value of numeric_only in DataFrame
  df.corr()
```

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness
A_id	1.000000	-0.028911	-0.005730	-0.002378	-0.013111	0.006179	0.000742
Size	-0.028911	1.000000	-0.170702	-0.324680	0.169868	-0.018892	-0.134773
Weight	-0.005730	-0.170702	1.000000	-0.154246	-0.095882	-0.092263	-0.243824
Sweetness	-0.002378	-0.324680	-0.154246	1.000000	-0.037552	0.095882	-0.273800
Crunchiness	-0.013111	0.169868	-0.095882	-0.037552	1.000000	-0.259607	-0.201982
Juiciness	0.006179	-0.018892	-0.092263	0.095882	-0.259607	1.000000	-0.097144
Ripeness	0.000742	-0.134773	-0.243824	-0.273800	-0.201982	-0.097144	1.000000

The correlation matrix illustrates the relationships between numeric attributes. Positive and negative correlations help identify patterns, indicating how attributes change concerning each other.

Statistical Measures:-



```
Median:
A_id           1999.500000
Size             -0.513703
Weight           -0.984736
Sweetness        -0.504758
Crunchiness       0.998249
Juiciness         0.534219
Ripeness          0.503445
dtype: float64

Mode:
A_id           0.000000
Size          -7.151703
Weight        -7.149848
Sweetness     -6.894485
Crunchiness   -6.055058
Juiciness     -5.961897
Ripeness      -5.864599
Name: 0, dtype: float64

Skewness:
A_id           0.000000
Size          -0.002437
Weight         0.003102
Sweetness      0.083850
Crunchiness    0.000230
Juiciness     -0.113421
Ripeness      -0.008764
dtype: float64

Kurtosis:
A_id          -1.200000
Size          -0.083341
Weight         0.359050
Sweetness      0.014472
Crunchiness    0.722020
Juiciness      0.028735
Ripeness      -0.071850
```

Computed median, mode, skewness, and kurtosis to capture central tendency, peak, asymmetry, and tail characteristics, providing deeper insights into the distributional properties of the numeric features.

Conclusion: - This analysis aligns with the objectives outlined in the question paper, covering essential components such as descriptive statistics and correlation matrices. The visualizations, including histograms and scatter plots, fulfill the requirement of presenting information in a graphical format.