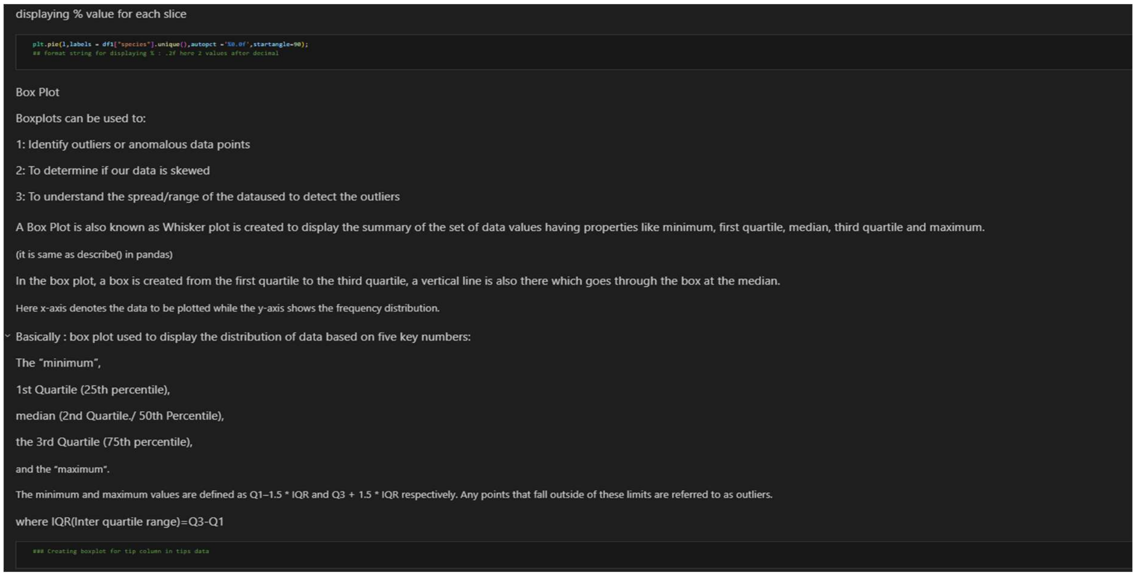
15-11-2024

Training Day – 43

November 15, Friday

- *Topic:* Revisiting Data Cleaning Techniques

- Practiced handling outliers and formatting columns.
- Example: Removed outliers using the interquartile range (IQR).



displaying % value for each slice

```
plt.pie(l,labels = df1["species"].unique(),autopct ='%0.0f',startangle=90);
## format string for displaying % : .2f here 2 values after decimal
```

Box Plot

Boxplots can be used to:

1: Identify outliers or anomalous data points

2: To determine if our data is skewed

3: To understand the spread/range of the dataused to detect the outliers

A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum.

(it is same as describe() in pandas)

In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median.

Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution.

Basically : box plot used to display the distribution of data based on five key numbers:

The "minimum",

1st Quartile (25th percentile),

median (2nd Quartile./ 50th Percentile),

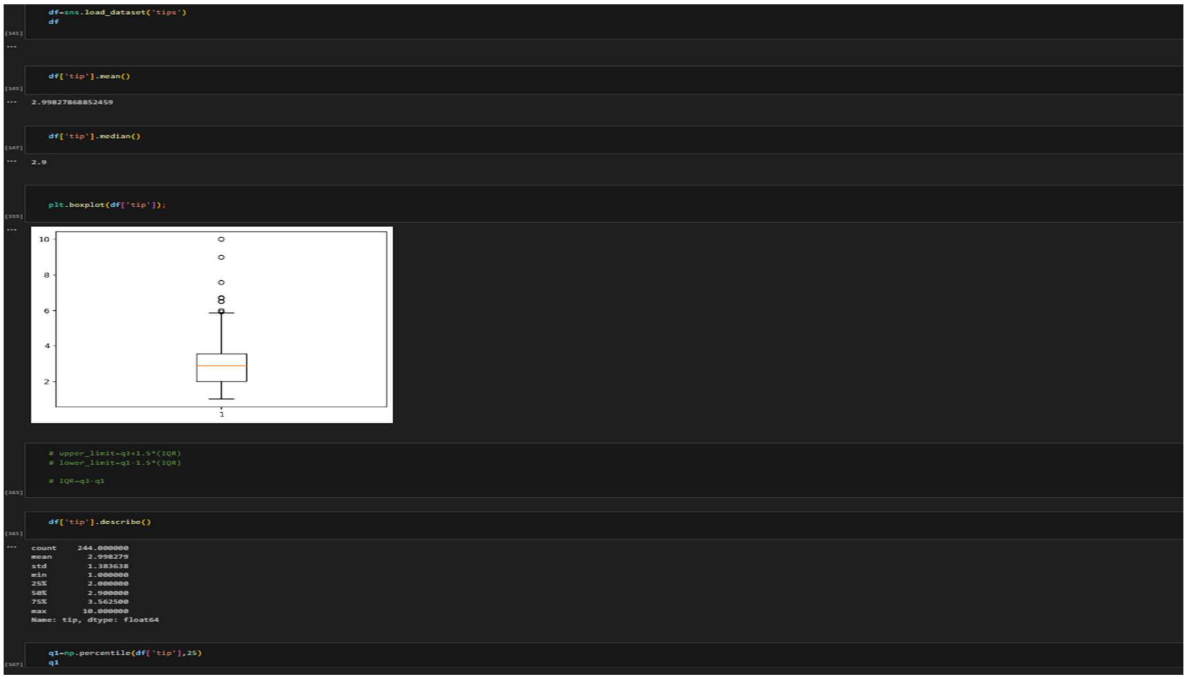
the 3rd Quartile (75th percentile),

and the "maximum".

The minimum and maximum values are defined as Q1–1.5 * IQR and Q3 + 1.5 * IQR respectively. Any points that fall outside of these limits are referred to as outliers.

where IQR(Inter quartile range)=Q3-Q1

```
### Creating boxplot for tip column in tips data
```



```
df=sns.load_dataset('tips')
df
```

```
df['tip'].mean()
```

```
2.99827868852459
```

```
df['tip'].median()
```

```
2.9
```

```
plt.boxplot(df['tip']);
```

```
# upper_limit=q3+1.5*(IQR)
# lower_limit=q1-1.5*(IQR)

# IQR=q3-q1
```

```
df['tip'].describe()
```

```
count    244.000000
mean       2.998279
std        1.383638
min        1.000000
25%        2.000000
50%        2.900000
75%        3.562500
max       10.000000
Name: tip, dtype: float64
```

```
q1=np.percentile(df['tip'],25)
q1
```