Statistics Cheat Sheet- 2

LEARN LOGIC AI
FOR THE FUTURE

FREQUENCY TABLES

A simple way to organize data, showing the number of occurrences (frequency) of each distinct value in a dataset.

- Example:

Satisfaction	Frequency
Very satisfied	59
Satisfied	42
Neutral	12
Dissatisfied	8
Very dissatisfied	5

- For a continuous data with too many values, it is better to use intervals (frequency bins) for the frequency table.

```
df[''].value_counts()
```

Count plot is used to Show the counts of observations in each categorical bin using bars.

```
import seaborn as sns
sns.countplot(x='')
```

Quartiles

Quartiles are the set of values that divide the data points into **four equal parts** each containing 25% of data points using three individual data points.

- **Q1: First quartile**: 25% of the data are below this value.
- **Q2: Second quartile / Median**: This value splits the data in half.
- **Q3: Third quartile**: 25% of the data are above this value.

```
import numpy as np
print(np.quantile(df[''],
[0.25,0.5,0.75]))    #quartiles
IQR = Q3 - Q1 # Calculate the IQR
```

The interquartile range (IQR): It is the difference between the first and third quartiles. IQR = Q3 – Q1.

Half of the observations fall within the interquartile range regardless of the distribution's shape.

Box Plot: Box plot is a graphical representation of the distribution of a dataset.

It displays key summary statistics such as the median, quartiles, and potential outliers in a concise and visual manner.

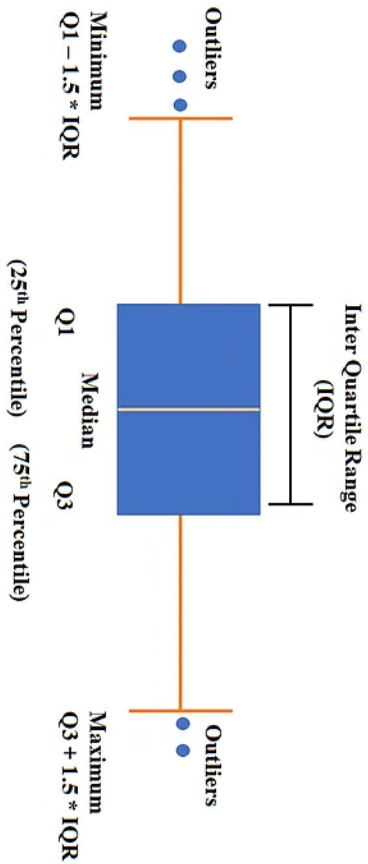
Removing outliers using Inter-Quartile Range: To identify outliers using the IQR method, we establish two boundaries:

Lower Bound: Q1–1.5 * IQR

Upper Bound: Q3 + 1.5 * IQR

Any data point that falls below the lower bound or exceeds the upper bound are considered an outlier.

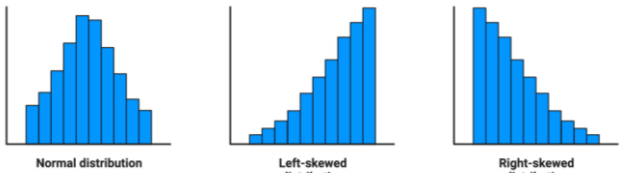
```
plt.boxplot(df[''])
df.plot(kind='box', subplots=True, layout=(, )) # Plotting boxplot
```

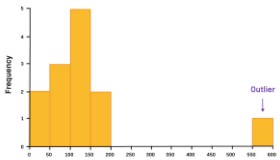


HISTOGRAM

A histogram is a chart that plots the distribution of a numeric variable's values as a series of bars.

- Histograms can be used to quickly compare 2 data sets.
- Histograms allow us to evaluate the shape of a data set:
  is it symmetrical, right-skewed, left-skewed
- Histograms allow us to evaluate the number of modes of a data set:
  is it unimodal, or multimodal.
- Histograms are useful to identify outliers in a data set.






```
import matplotlib.pyplot as
plt
#histogram of a feature
plt.hist(df[''])
# Distribution
sns.histplot(df[''],
kde=True)
# Plotting histograms for
each feature
df.hist()
```