Introduction to Exploratory Data Analysis (EDA)

Notes by Mannan Ul Haq (BDS-3C)

Exploratory Data Analysis (EDA) is a critical initial step in the data analysis process that involves the exploration, visualization, and summary of data to uncover patterns, trends, anomalies, and insights.

Key Objectives of EDA:

1. **Understand the Data**: EDA helps you get a comprehensive understanding of your dataset, including its structure, size, and the variables it contains.
2. **Detect Patterns and Relationships**: EDA aims to uncover relationships and patterns within the data, such as correlations between variables, trends over time, and clusters of similar data points.
3. **Identify Anomalies and Outliers**: EDA helps you spot data points that deviate significantly from the norm, which could be errors or noteworthy observations.
4. **Prepare Data for Modeling**: EDA assists in data preprocessing by revealing data quality issues, missing values, and helping with feature engineering.

Common Techniques in EDA:

1. **Descriptive Statistics**: Calculating summary statistics like mean, median, variance, and quartiles to understand the central tendency and variability of your data.
2. **Data Visualization**: Creating charts, graphs, and plots to visualize data distributions, relationships between variables, and trends. Common visualization tools include histograms, scatter plots, bar charts, and box plots.
3. **Correlation Analysis**: Examining correlations between variables to understand how they are related.
4. **Outlier Detection**: Identifying outliers using various methods, such as the Z-score, the IQR (Interquartile Range), or visualization techniques.

Distributions and Frequency plotting (Histograms)

Distributions and frequency plotting, particularly through histograms, are essential components of Exploratory Data Analysis (EDA). These tools help you understand how data is distributed, identify patterns, and visualize the central tendencies and variabilities in your dataset. Here's an explanation of distributions and how to create histograms:

Distributions:

- A distribution is a representation of how data is spread or arranged in a dataset.
- It describes the frequency of different values or ranges of values.
- Understanding the distribution of data is fundamental in EDA because it helps you identify the characteristics and properties of your dataset.

Frequency Plotting (Histograms):

- A histogram is a graphical representation of the distribution of a dataset.
- It divides the data into discrete intervals or "bins" and shows the number of data points that fall into each bin.

Here's how you can create a histogram in Python using the Matplotlib library:

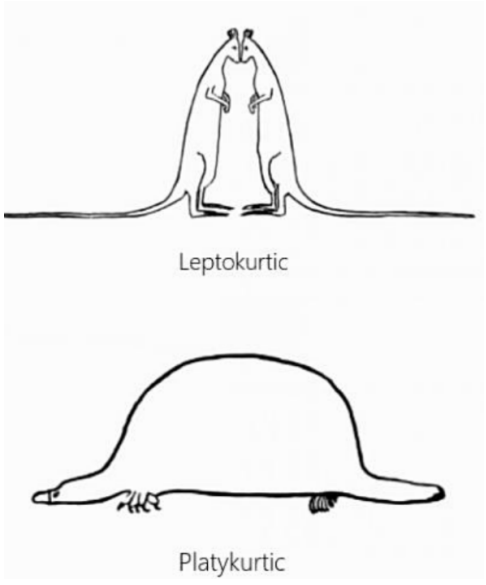
```
import matplotlib.pyplot as plt
import numpy as np

# Generate some example data
data = np.random.randn(1000)  # Generate 1000 random data points

# Create a histogram
plt.hist(data, bins=20, edgecolor='black')  # 'bins' specifies the number of intervals
plt.title('Histogram of Random Data')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.grid(True)

# Show the histogram
plt.show()
```

Common observations you can make from a histogram include:

- **Shape**: Is the distribution symmetric or skewed (left or right)?

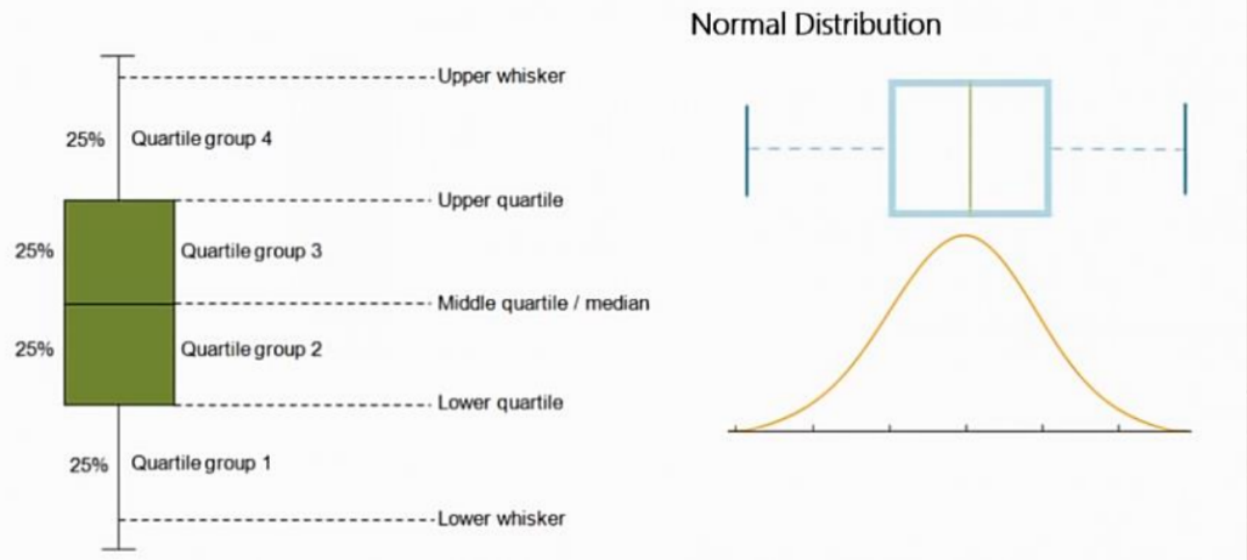
- **Central Tendency**: Where is the peak of the distribution, often indicated by the mode, median, or mean?
- **Spread**: How spread out are the data points?
- **Kurtosis:** Is a statistical measure that quantifies the shape of the probability distribution of a dataset, specifically how "heavy-tailed" or "light-tailed" the distribution is compared to a normal distribution. It provides information about the presence of outliers and the degree of peakedness in the distribution. There are typically three common types of kurtosis:

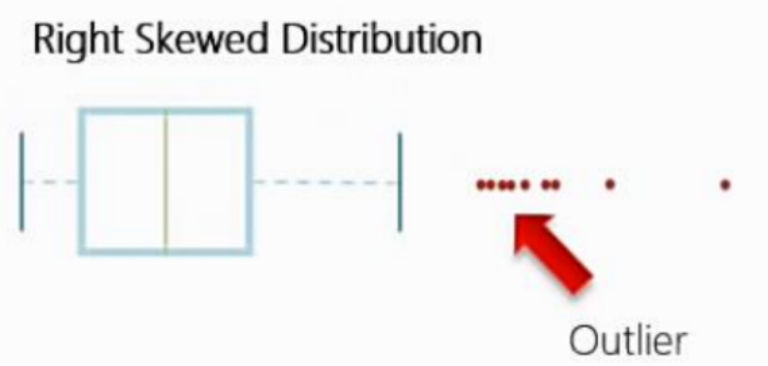
1. **Mesokurtic (Kurtosis = 3)**: The distribution has kurtosis equal to 3, which is the kurtosis of a normal distribution. It indicates that the distribution is neither heavily tailed nor too peaked.
2. **Leptokurtic (Kurtosis > 3)**: A leptokurtic distribution has positive kurtosis, indicating heavy tails and a higher peak than a normal distribution. It implies that the dataset has more outliers and is more "pointy."
3. **Platykurtic (Kurtosis < 3)**: A platykurtic distribution has negative kurtosis, meaning it has lighter tails and is flatter than a normal distribution. It implies that the dataset has fewer outliers and is less peaked.



Data Spread, Range, and Outlier Analysis (Box-and-Whisker plots)

- A box-and-whisker plot is a graphical representation of the spread and distribution of a dataset. It displays the central tendency, data spread, and identifies potential outliers.
- The key components of a box-and-whisker plot include:
  - A rectangular box, which represents the interquartile range (IQR), with the lower boundary being the first quartile (Q1) and the upper boundary being the third quartile (Q3).
  - A horizontal line inside the box, which represents the median (Q2), also known as the second quartile.
  - Whiskers extending from the box, which can help identify potential outliers.
  - Individual data points beyond the whiskers, which are considered potential outliers.





Here's an example of creating a box-and-whisker plot in Python using Matplotlib:

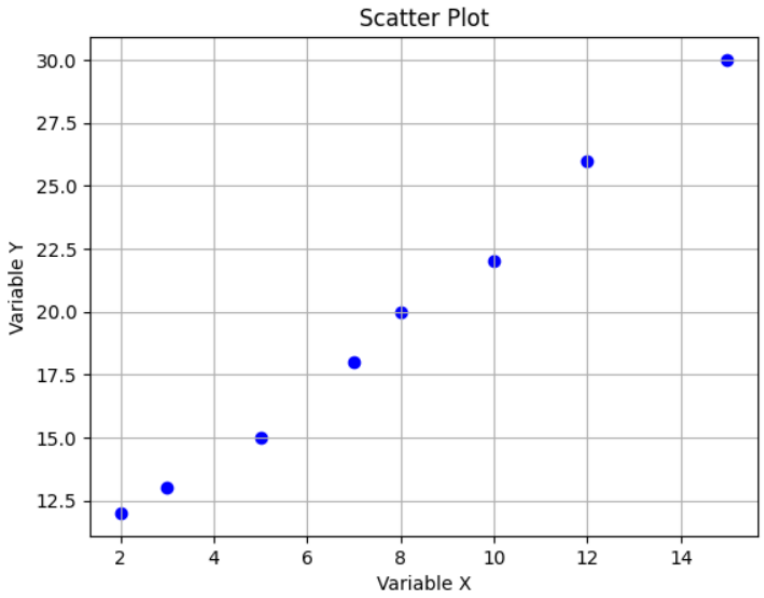
```
import matplotlib.pyplot as plt

data = [10, 15, 20, 25, 30, 35, 40, 100]

# Create a box-and-whisker plot
plt.boxplot(data)

plt.title('Box-and-Whisker Plot')
plt.ylabel('Value')
plt.show()
```

Correlation Analysis (Scatter charts)

Correlation analysis is a fundamental technique in exploratory data analysis (EDA) that helps you understand the relationships between variables in your dataset. Scatter charts, also known as scatter plots or scatter diagrams, are commonly used to visualize these relationships.

Scatter Charts:

- A scatter chart is a graphical representation of data points in a Cartesian coordinate system. Each data point is represented as a dot on the chart.
- Scatter charts are used to visualize the relationship between two continuous variables, making them suitable for assessing correlation.
- In correlation analysis, you create scatter plots to visually inspect how data points are distributed and whether there's a discernible pattern in their arrangement.



Here's an example of creating a scatter plot in Python using Matplotlib:

```
import matplotlib.pyplot as plt

# Sample data for two variables
x = [2, 3, 5, 7, 8, 10, 12, 15]
y = [12, 13, 15, 18, 20, 22, 26, 30]

# Create a scatter plot
plt.scatter(x, y, marker='.', color='red')

plt.title('Scatter Plot')
plt.xlabel('Variable X')
plt.ylabel('Variable Y')
plt.grid(True)
plt.show()
```

In this example, we have two variables, 'x' and 'y,' and we create a scatter plot to visualize the relationship between them. The scatter plot shows how data points are distributed and whether there's an apparent trend in the data. If there's a strong linear relationship between the variables, the points in the scatter plot will tend to form a pattern.

Here are the common types of correlation that can be assessed through scatter plots:

1. **Positive Correlation**:

In a positively correlated relationship, as one variable increases, the other variable tends to increase as well. When plotted on a scatter plot, data points tend to form an upward-sloping pattern from the bottom left to the top right.

2. **Negative Correlation**:

   In a negatively correlated relationship, as one variable increases, the other variable tends to decrease. On a scatter plot, data points tend to form a downward-sloping pattern from the top left to the bottom right.

3. **No Correlation (Zero Correlation)**:

   When there is no correlation, the two variables do not show any consistent pattern or trend on the scatter plot. Data points are scattered randomly, without forming any noticeable direction.

4. **Linear Correlation**:

   Linear correlation indicates that the relationship between the two variables can be represented by a straight line on the scatter plot. The data points form a roughly straight-line pattern, either upward or downward.

5. **Non-Linear Correlation**:

   Non-linear correlation suggests that the relationship between the variables is best described by a curve rather than a straight line. Data points may form curved patterns on the scatter plot.

6. **Perfect Correlation**:

   - A perfect correlation indicates that all data points fall exactly on a straight line, forming a perfect pattern with no variation. This is quite rare in practice.

7. **Strong and Weak Correlations**:

   - In addition to the type of correlation, you can assess the strength of the relationship. A strong correlation indicates a tight and consistent pattern in the scatter plot, while a weak correlation suggests a less consistent or scattered pattern.