EDA-Box Plot Analysis

Box Plot Analysis

A box plot, also known as a box-and-whisker plot, is a type of graphical representation that displays the distribution of a dataset. It's commonly used in Exploratory Data Analysis (EDA) to understand the central tendency and variability of a variable. Here's a summary of box plot analysis with an example:

Components of a Box Plot:

1. **Box**: The box represents the interquartile range (IQR), which is the difference between the 75th percentile (Q3) and the 25th percentile (Q1).
2. **Median**: The line inside the box represents the median, or the middle value of the dataset.
3. **Whiskers**: The lines extending from the box are called whiskers. They represent the range of the data, excluding outliers.
4. **Outliers**: Data points that fall outside the whiskers are considered outliers.

Interpretation:

1. **Symmetry**: If the median is close to the center of the box and the whiskers are relatively equal on both sides, the distribution is likely symmetric.
2. **Skewness**: If the median is shifted towards one end of the box or the whiskers are not balanced, the distribution may be skewed.
3. **Outliers**: The presence of outliers can indicate unusual patterns in the data.
4. **Spread**: The length of the box and whiskers can give an idea about the spread of the data.

Example:

Suppose we have a dataset of exam scores for 20 students:

Score	
70	
75	
80	
85	
90	
95	
92	
88	
78	
76	
94	
91	
89	
77	
84	
82	
96	(outlier)
98	(outlier)

Box Plot:

The box plot for this dataset would show:

- Median: 85
- Box length: approximately 20-30 units
- Whiskers: extending from 70 to 98
- Outliers: 96 and 98

In this example, the distribution appears to be slightly skewed to the right (towards higher scores), with a few outliers. The box plot helps us quickly identify patterns in the data and understand the central tendency and variability of exam scores.

Code Example:

Here's an example code snippet in Python using the `matplotlib` library to create a box plot:

```
 import matplotlib.pyplot as plt

# assume 'scores' is a list of exam scores
plt.boxplot(scores)
plt.show()
```

This will generate a basic box plot showing the distribution of exam scores. You can customize the appearance and add more features, such as mean and standard deviation, using various options available in the `matplotlib` library.

Curated by Brajesh Kumar