Outliers: Data Science Mein Ghair Mutawaqa Mehman

Data ki duniya mein, outliers woh ghair mutawaqa mehman hain jo aap ki dinner party mein baghair dawat ke aa jate hain. Jab aap samajhte hain ke sab kuch sahi tareeqe se set hai, toh woh aakar harmony ko disturb kar dete hain. Lekin yeh outliers akhir hain kya? Inki ahmiyat kya hai? Aur sab se zaroori, hum inhe kaise handle karte hain? Chaliye, shuru karte hain.

Outliers Kya Hain? 🔍

Outliers are data points that deviate significantly from the rest of the observations in a dataset. Imagine plotting the ages of a group of high school students, and among those teenagers, you find an age of 85. That 85 is an outlier—it doesn't fit the general trend or expectation.

Outliers Ko Kaise Pehchaanen?

Visual Tareeqa: Tools jaise scatter plots, box plots, aur histograms behtareen visual madadgar hote hain. Maslan, box plot mein, jo data points whiskers ke bahar hote hain, woh outliers maane ja sakte hain, but we need to know about IQR.

Statistical Tareeqe: Z-score aur IQR (Interquartile Range) method do maqbool statistical tareeqe hain. Z-score batata hai ke ek data point mean se kitne standard deviations door hai. Aam taur par, agar Z-score > 3 ya < -3 ho toh woh outlier maana jata hai.

Inhe Kaise Hatayen? 🛠

1. **Truncation ya Capping:** High outliers ke liye, kisi threshold se upar ki value ko maximum cap par set kiya ja sakta hai. Isi tarah, low outliers ke liye, kisi threshold se neeche ki values ko minimum cap par set kiya ja sakta hai.
2. **Transformation:** Kabhi-kabhi, logarithms jaise mathematical transformations se outliers ko control kiya ja sakta hai.
3. **Imputation:** Outlier ko mean, median, ya mode jaise central tendency measures se replace karein.
4. **Deletion:** Agar outlier data entry errors ki wajah se hai ya clear hai ke woh value add nahi karega, toh behtar hai ke aap usey remove kar dein.

Python code to remove outliers

```
import seaborn as sns
import pandas as pd
# Load the Titanic dataset
titanic = sns.load_dataset('titanic')
# Display the first few rows of the dataset
print(titanic.head())

# Calculate the IQR for the 'age' column
Q1 = titanic['age'].quantile(0.25)
Q3 = titanic['age'].quantile(0.75)
IQR = Q3 - Q1
```

```
# Define bounds for the outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
# Remove outliers
titanic_no_outliers = titanic[(titanic['age'] >= lower_bound) & (titanic['age'] <= upper_bound)]
# Display the first few rows of the dataset without outliers
print(titanic_no_outliers.head())
```

⚠️ Outliers Ko Ignore Karne Ka Asar

Tircha Analysis: Outliers descriptive aur inferential statistics dono ko skew kar sakte hain, jisse data ka distorted view milta hai.

Machine Learning Models Par Asar: Algorithms, khaas kar linear models, outliers se sensitive hote hain. Woh coefficient estimates aur predictions par drastic asar daal sakte hain.

Gumrahi Paida Karne Wale Assumptions: Data assumptions, jaise normality, outliers ki presence ki wajah se violated ho sakte hain, jisse galat nataij milte hain.