Outliers in the dataset

Define:

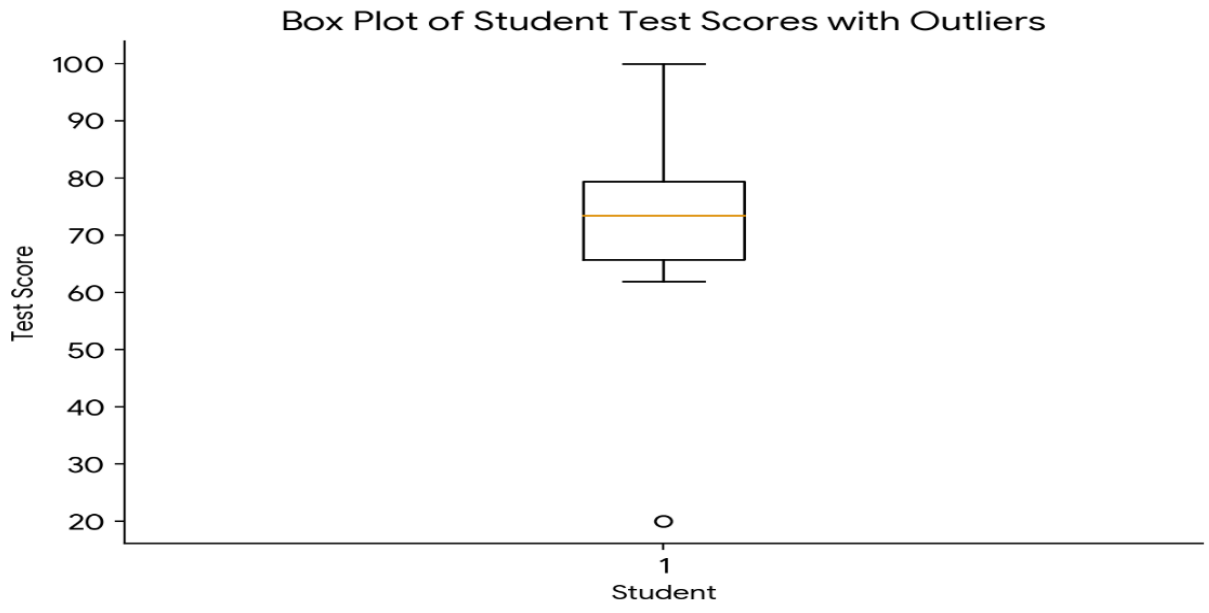
Outliers are data points that fall significantly outside the typical pattern of the rest of the data. They can be extremely high or extremely low compared to many of the observations.

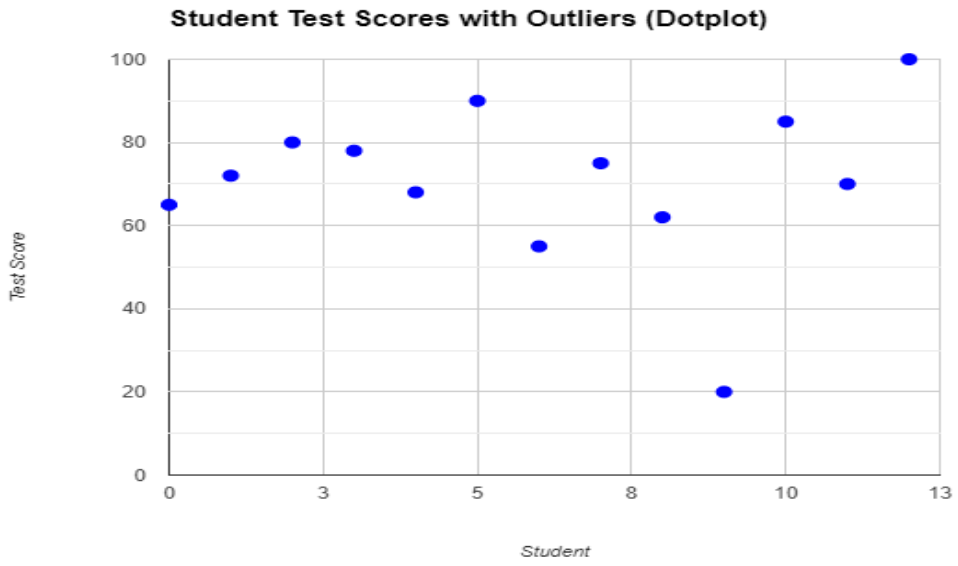
- Thing or object that is different from the rest of the things.
- Outliers have an impact on the data structure it makes normal data skewed.

- **Impact:** They can skew statistical analyses and distort the true representation of the data.
- **Identification:** No strict rules exist, but common methods include:
    - **Visual inspection:** Looking for points far away from the main group in plots like scatter plots or box plots.
    - **Statistical methods:** Using techniques like Interquartile Range (IQR) to define thresholds for identifying outliers.

Example (Numerical):

Imagine a dataset of student test scores with most scores ranging from 60 to 80. An outlier could be a score of 20 (significantly lower) or 100 (significantly higher) compared to the majority.





The code of the above plot is given below:

```
import matplotlib.pyplot as plt

# Sample student test scores with outliers
scores = [65, 72, 80, 78, 68, 90, 55, 75, 62, 20, 85, 70, 100]
# Create the dot plot
plt.figure(figsize=(8, 6))
plt.scatter(range(len(scores)), scores, s=50, color='blue', alpha=0.7) # Adjusts for dot size and alpha for transparency
# Label the axes
plt.xlabel("Student")
plt.ylabel("Test Score")
# Title the plot
plt.title("Student Test Scores with Outliers (Dotplot)")
# Show the plot
plt.grid(True)
plt.show()
```

It's important to remember that outliers can be due to various reasons, including:

- Errors in data collection or measurement.
- Natural variations in the population.
- Unusual or rare events.

Types of the outliers in the dataset

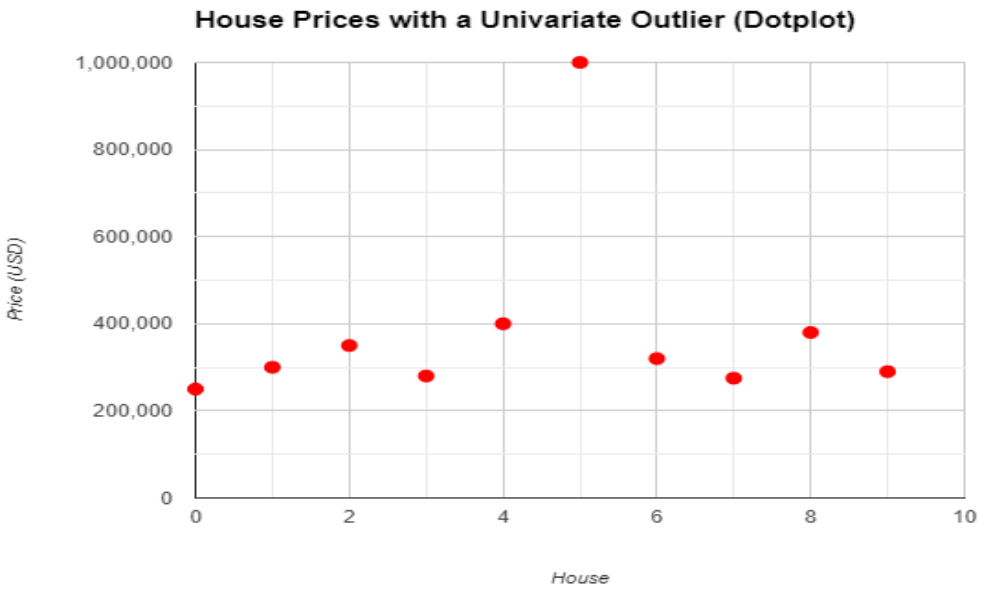
1. Univariate outliers.
2. Multivariate outliers.
3. Global outliers
4. Contextual outliers
5. Collective outliers

Univariate outliers

Define:

A data point that significantly deviates from the overall pattern in a single variable.

- **Important Point:** Easy to detect visually in histograms, dot plots, or box plots.
- **Real-life Example:** In a dataset of house prices, a house priced at $1 million might be a univariate outlier compared to the rest priced between $200,000 and $400,000.



The code of the plot above is given below:

```
import matplotlib. pyplot as plt

# Sample house prices with a univariate outlier
house_prices = [250000, 300000, 350000, 280000, 400000, 1000000, 320000, 275000, 380000, 290000]
```

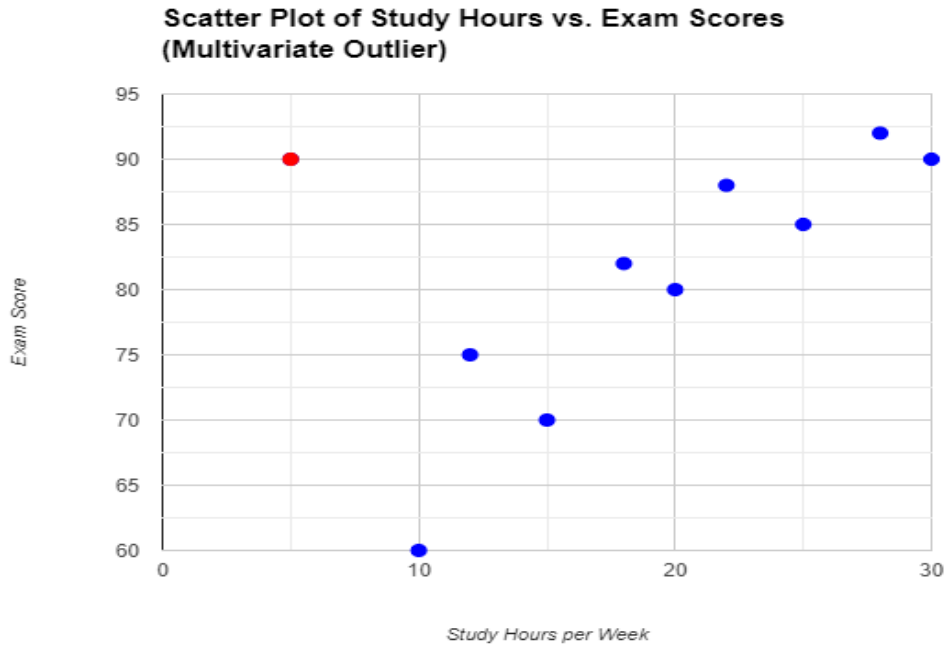
```
# Create the dotplot
plt.figure(figsize=(8, 6))
plt.scatter(range(len(house_prices)), house_prices, s=50, color='red', alpha=0.7) # Adjust s for dot size and alpha for transparency
# Label the axes
plt.xlabel("House")
plt.ylabel("Price (USD)")
# Title the plot
plt.title("House Prices with a Univariate Outlier (Dotplot)")
# Show the plot
plt.grid(True)
plt.show()
```

Multivariate Outlier:

Define:

A data point that deviates significantly from the overall pattern in multiple variables considered together.

- **Important Point:** Requires analyzing relationships between variables. Often detected using statistical methods like Mahalanobis distance.
- **Real-life Example**: In a dataset of customer purchases, a customer buying a large amount of both diapers and expensive electronics might be a multivariate outlier.



The code of the above graph is given below:

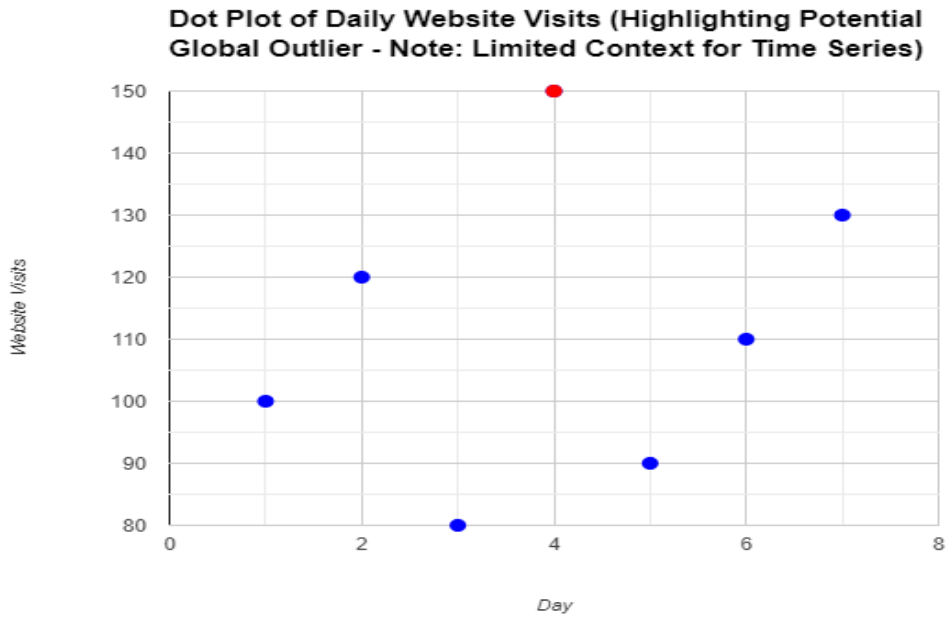
```
# This code generates a scatter plot to visualize a multivariate outlier
import matplotlib.pyplot as plt

# Sample data with a multivariate outlier
study_hours = [10, 15, 20, 25, 30, 5, 12, 18, 22, 28]
exam_scores = [60, 70, 80, 85, 90, 90, 75, 82, 88, 92]
# Identify potential outlier (low study hours, high exam score)
outlier_index = study_hours.index(5)
# Create the scatter plot
plt.figure(figsize=(8, 6))
plt.scatter(study_hours, exam_scores, s=50, alpha=0.7, color='blue')  # Adjusts for dot size and alpha for transparency
plt.scatter(study_hours[outlier_index], exam_scores[outlier_index], s=100, alpha=1, color='red')
# Highlight outlier
# Label the axes
plt.xlabel("Study Hours per Week")
plt.ylabel("Exam Score")
# Title the plot
plt.title("Scatter Plot of Study Hours vs. Exam Scores (Multivariate Outlier)")
# Show the plot
plt.grid(True)
plt.show()
```

Global Outlier (Point Anomaly):

Define:

A data point that deviates significantly from the entire dataset, regardless of any specific context. Essentially a univariate or multivariate outlier depending on the number of variables considered.

- **Important Point:** Often easy to detect with standard outlier detection techniques.
- **Real-life Example:** In a dataset of daily website visits, a day with ten times the usual traffic could be a global outlier. (This could be a univariate outlier if only visits are considered, or a multivariate outlier if visits are analyzed with other factors like time of day)



The below is the code of the above graph:

```
import matplotlib.pyplot as plt
# Sample data with a global outlier (day 4 with 10 times the usual traffic)
days = [1, 2, 3, 4, 5, 6, 7]
visits = [100, 120, 80, 150, 90, 110, 130]
# Highlight the outlier day
outlier_day = days[visits.index(max(visits))]
# Create the dot plot
plt.figure(figsize=(8, 6))
plt.scatter(days, visits, s=100, alpha=0.7, color='blue')  # Adjust s for dot size and alpha for transparency
plt.scatter(outlier_day, max(visits), s=150, alpha=1, color='red')  # Highlight outlier
# Label the axes
plt.xlabel("Day")
plt.ylabel("Website Visits")
# Title the plot (with a note about limitations)
plt.title("Dot Plot of Daily Website Visits (Highlighting Potential Global Outlier - Note: Limited Context for Time Series)")
# Annotate the outlier
plt.annotate("Potential Outlier", (outlier_day, max(visits)), xytext=(outlier_day + 0.2, max(visits) + 10),  arrowprops=dict(facecolor='red', shrink=0.05))
# Show the plot
plt.grid(True)
plt.show()
```

Contextual Outlier (Conditional Anomaly):

Define:

A data point that deviates from the expected pattern within a specific context.

- **Important Point**: Requires additional information or conditions. Often found in time-series data.
- **Real-life Example:** In a dataset of daily temperatures, a temperature of 80°F in December might be a contextual outlier, while the same temperature in July wouldn't be.

Collective Outlier:

Define:

A group of data points that collectively deviates from the overall pattern, even though individual points might not be outliers on their own.

- **Important Point:** Can be harder to detect than other types. Often identified using clustering algorithms.
- **Real-life Example:** In a dataset of customer purchases, a group of customers all buying a new, rarely purchased product together might be a collective outlier.