Box Plots

Visualizing Outliers & Differences in Distribution Between Groups

Q&A Activity: Review

- What is the **mean** of a set of data?
- What is the **median** of a set of data?

Q&A Activity: Review

- What is the **mean** of a set of data?
- What is the **median** of a set of data?

Mean is the average of all the values, i.e. (sum / count).

Median is the middle value in the sorted data.

Q&A Activity: Review

- What is a **quartile**?

Q&A Activity: Review

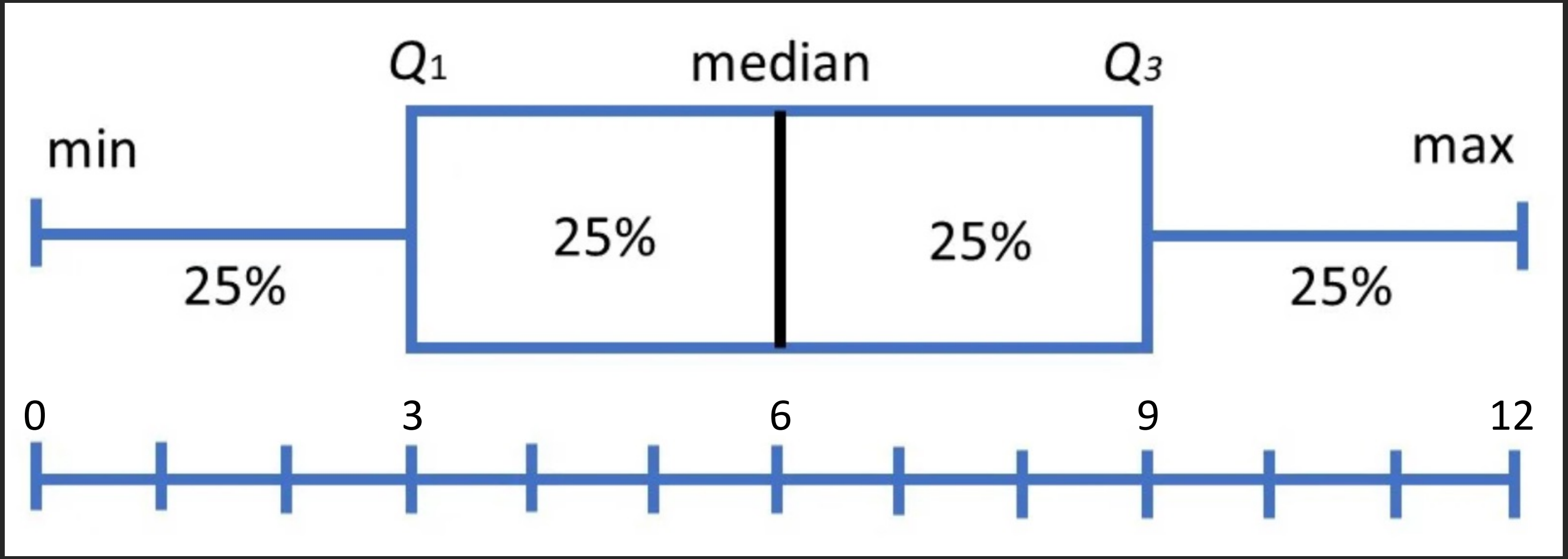
- What is a **quartile**?

4 groups of equal size into which a set of sorted data can be divided.

Q0 (0) – 0% of values are less than or equal to this number (min)
Q1 (0.25) – 25% of values are less than or equal to this number
Q2 (0.5) – 50% of values are less than or equal to this number (median)
Q3 (0.75) – 75% of values are less than or equal to this number
Q4 (1) – 100% of values are less than or equal to this number (max)

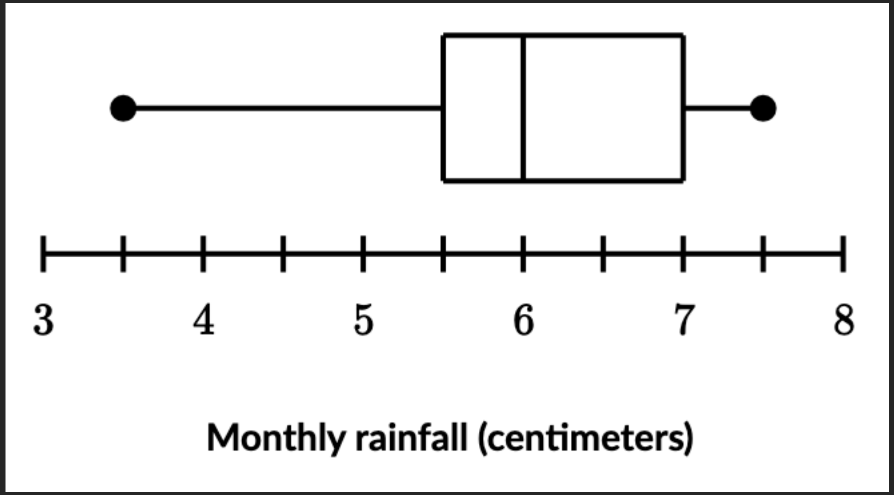
Box Plots

- This is an example of a **box plot**.
- The **interquartile range or IQR** (Q1 to Q3) contains 50% of the data.
- Here the range of each quartile is the same but usually it won’t be.



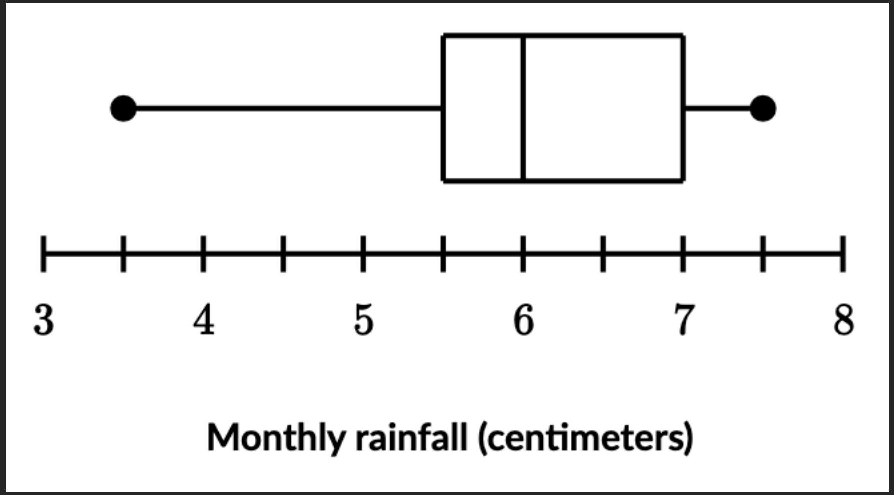
Q&A Activity: Reading Box Plots

- What is the min value?
- What is the max value?
- What is the median value?
- What is the value of Q1?
- What is the value of Q3?
- What is the IQR?
- What is the range?



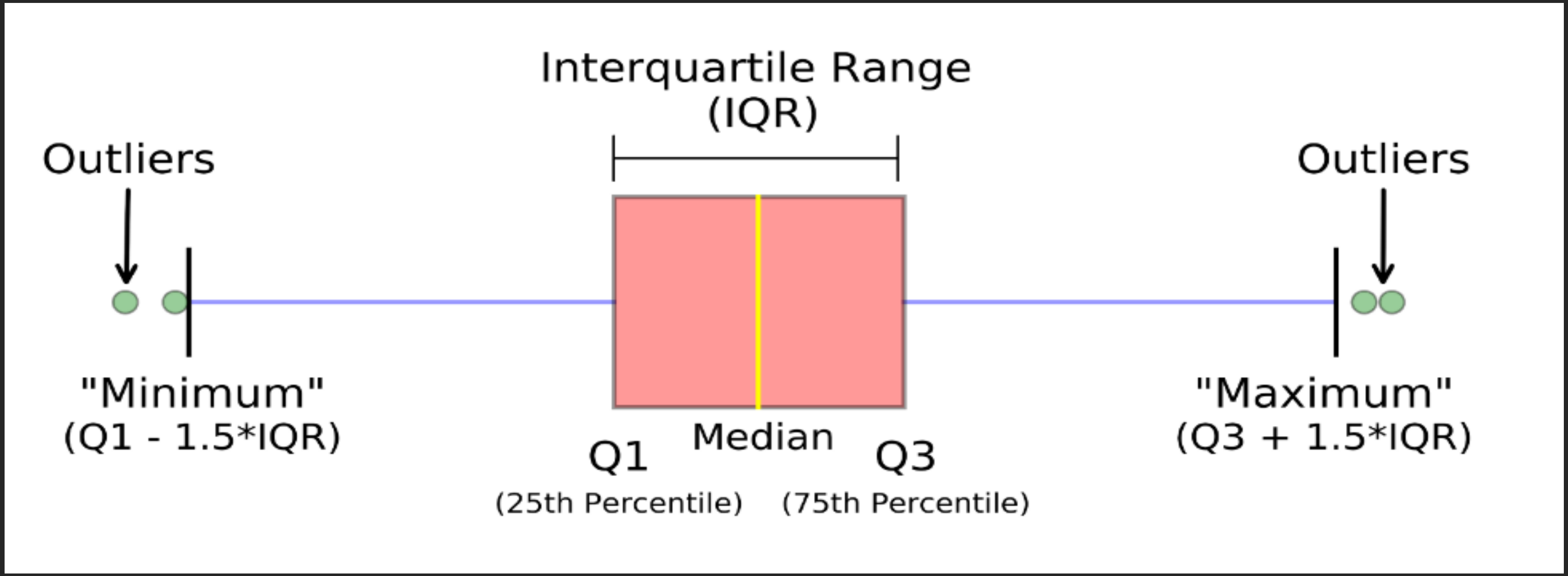
Q&A Activity: Reading Box Plots

- What is the min value? 3.5
- What is the max value? 7.5
- What is the median value? 6
- What is the value of Q1? 5.5
- What is the value of Q3? 7
- What is the IQR? 7 – 5.5 = 1.5
- What is the range? 7.5 – 3.5 = 4



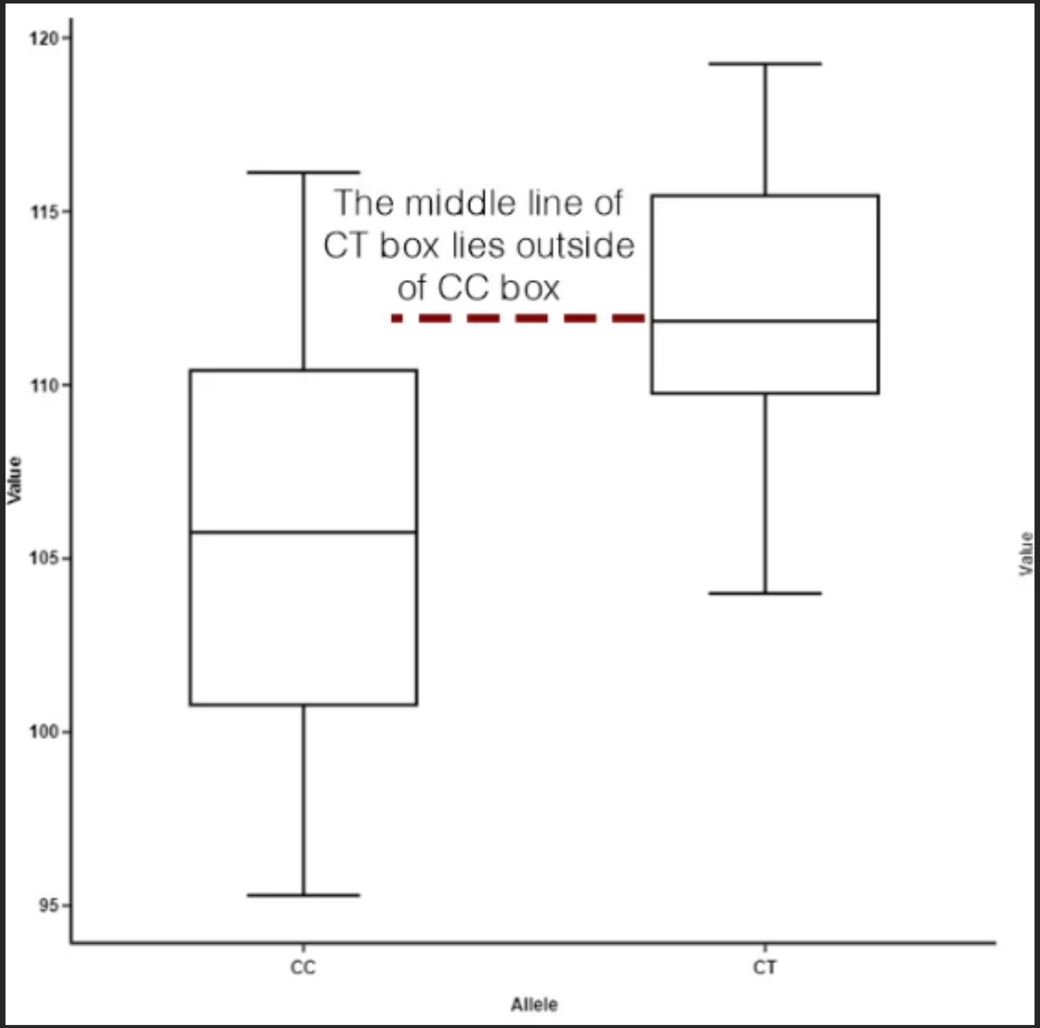
Outliers

- Sometimes the range is ridiculously large compared to the IQR.
- This may indicate we have **outliers** so we redefine “minimum” and “maximum” based on the IQR according to these formula.



Comparing Distributions

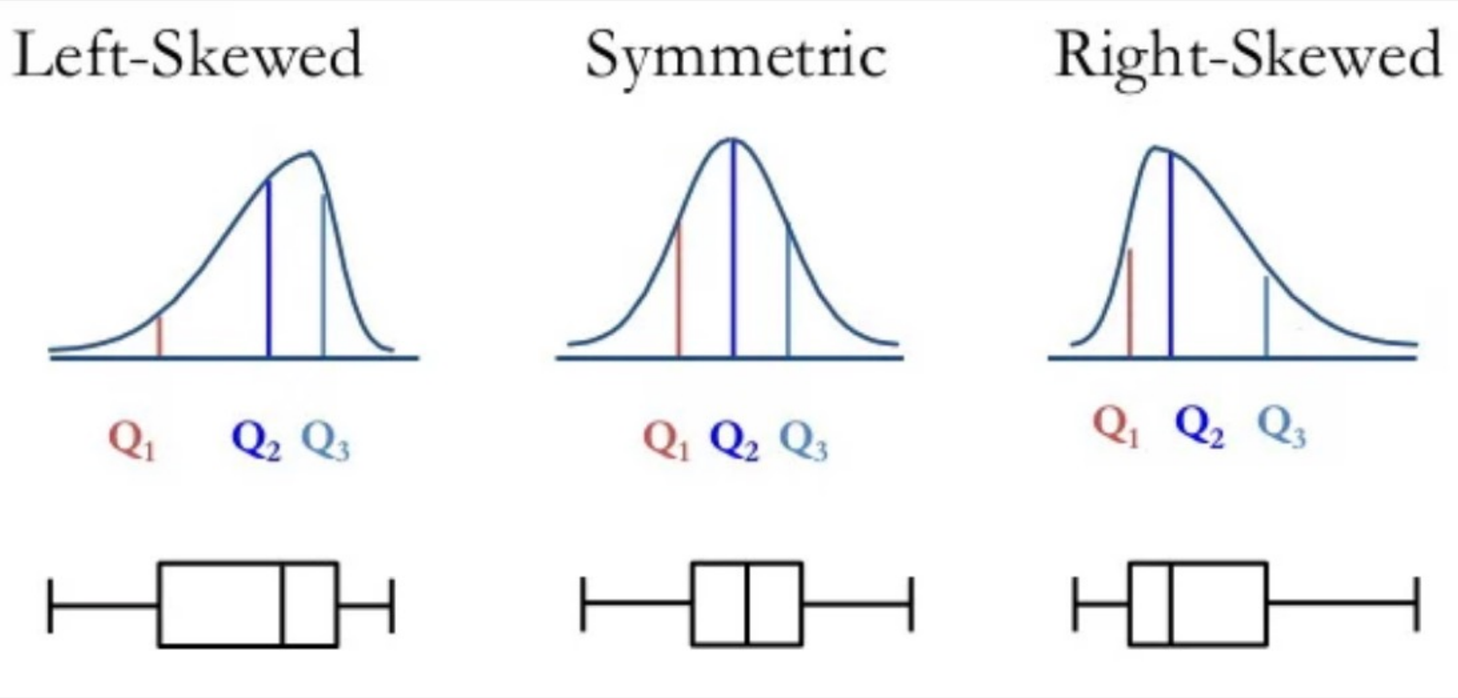
- When you have multiple categories in your dataset, you can use box plots to compare the values of one variable for each category.
- Compare different metrics:
  - Medians
  - IQR and whisker ranges
  - Outliers
  - Skewness



The middle line of CT box lies outside of CC box

Skewness

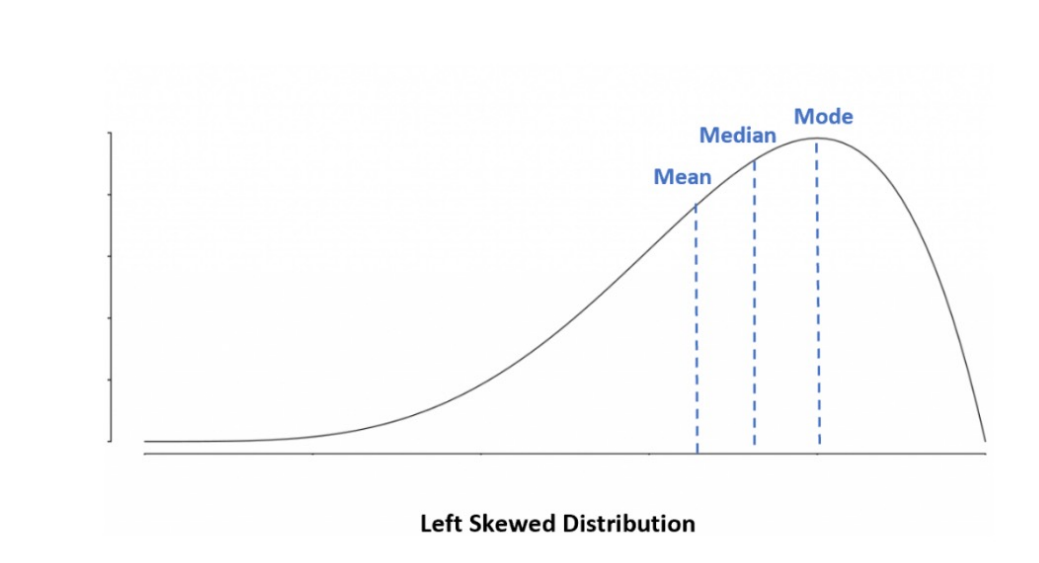
- **Skewness** is a way to describe the symmetry of a distribution.
- Consider the position of the median and the size of the **tail**.
- Symmetric distribution is also called a **normal distribution**.



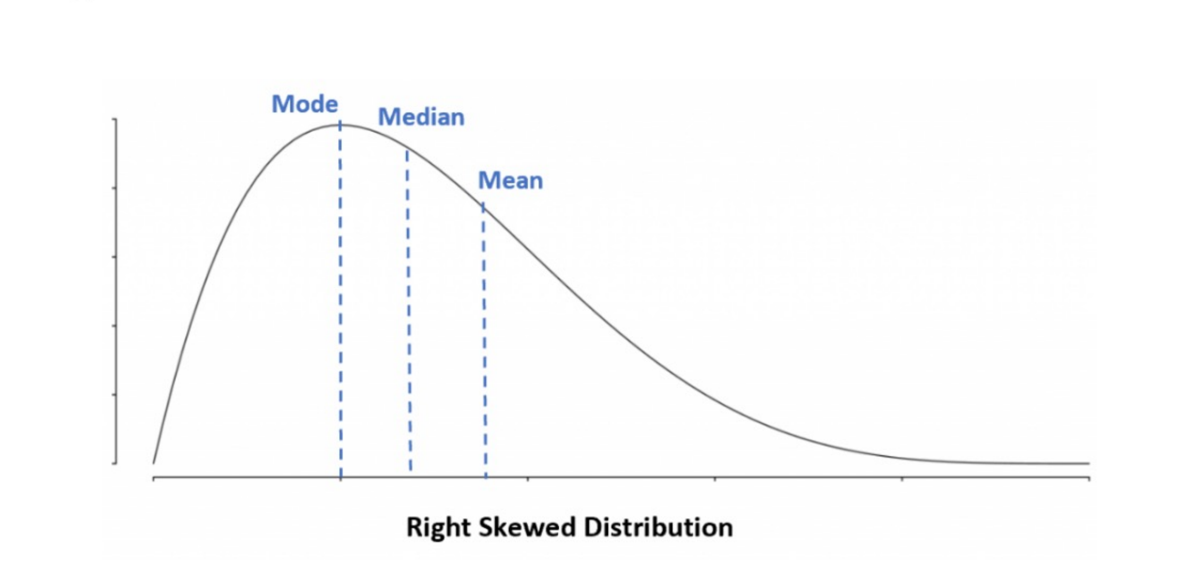
Skewness

- You can also use the position of the mean and median to determine skew.

Left Skewed Distribution: Mean < Median < Mode

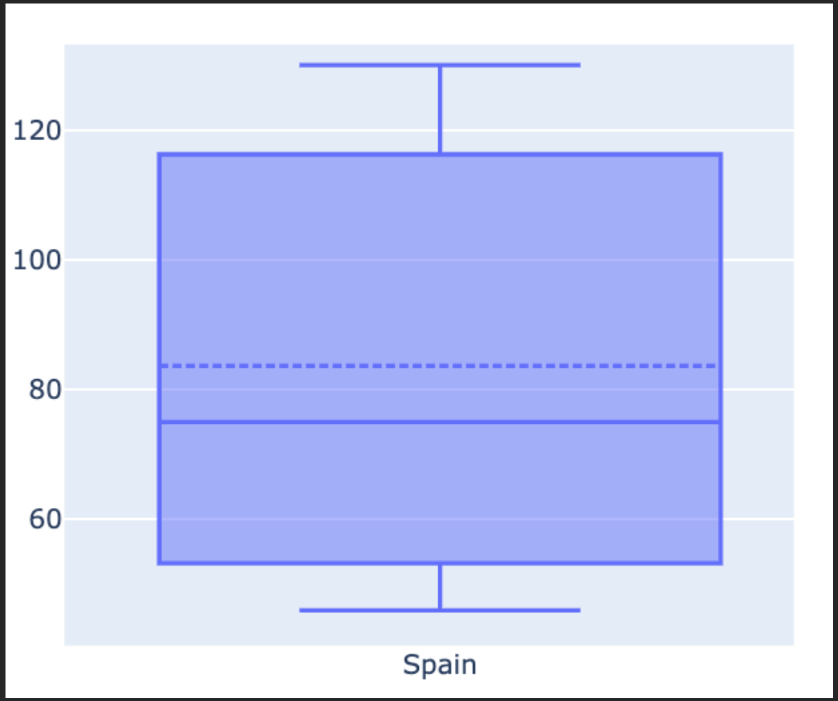


Right Skewed Distribution: Mode < Median < Mean



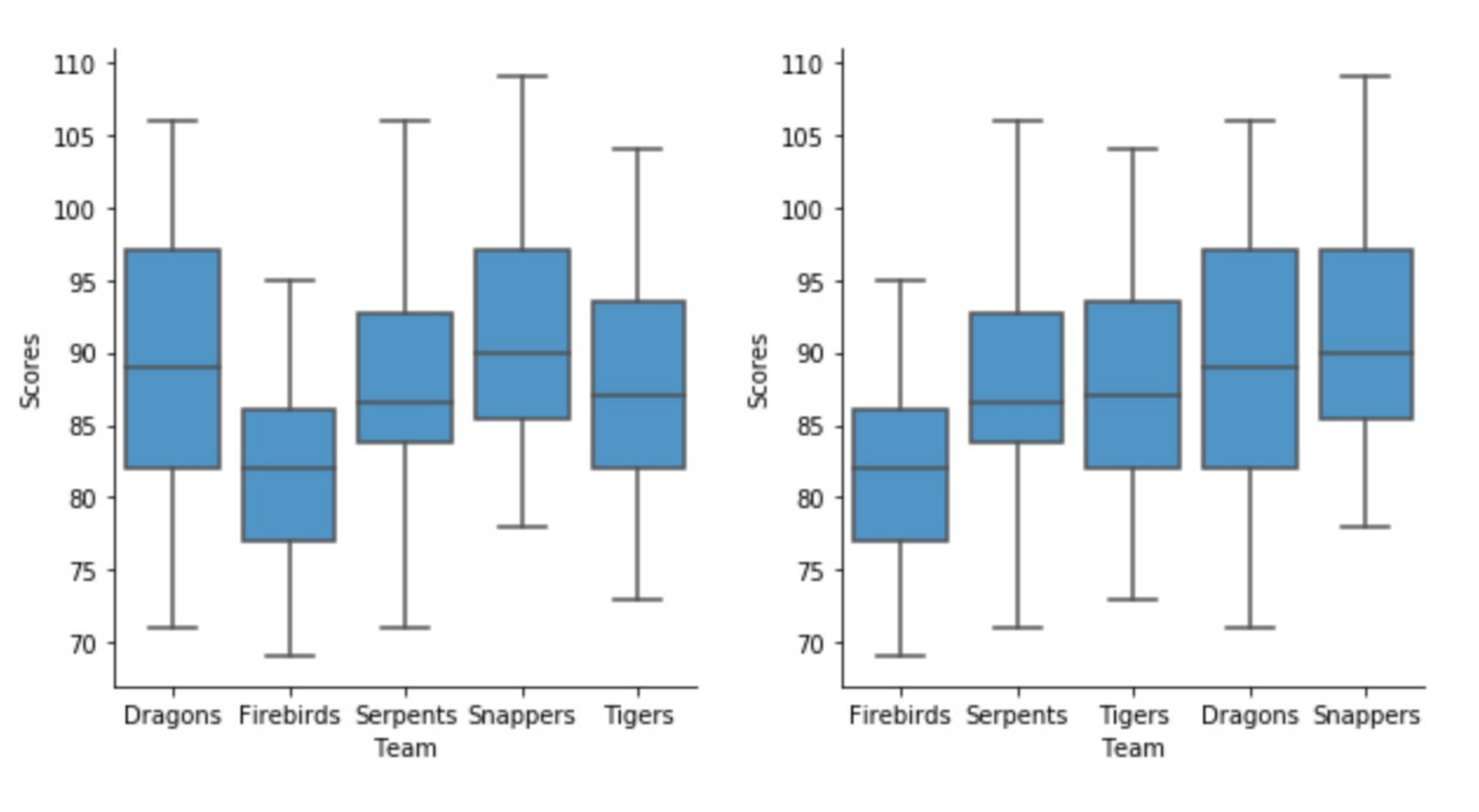
Skewness

- We can also draw the mean on the box plot.
- Median is represented by the solid line.
- Mean is represented by the dotted line.



Ordering Categories in Box Plots

- Sorting the categories by the median value can improve readability.



How-to: Make Box Plots in Jupyter

1. Read CSV Data into Pandas Dataframe
   - Import Pandas Library
   - Read CSV data and Save in Variable
   - Display Dataframe Contents
2. Generate Plotly Box Plot
   - Import Plotly Express Library
   - Set Columns as x and y
   - Set Additional Plot Options (Category Order)
   - Generate Chart (with Means)

Summary

- Quartiles & IQR
- Box Plots
- Outliers
- Skewness
- Ordering Plots
- Plotly Box Plots in Jupyter