SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Data Visualization
Types of Statistical Graphs

Vaibhav P. Vasani

Assistant Professor

Department of Computer Engineering

K. J. Somaiya College of Engineering

Somaiya Vidyavihar University

vaibhav.vasani@gmail.com

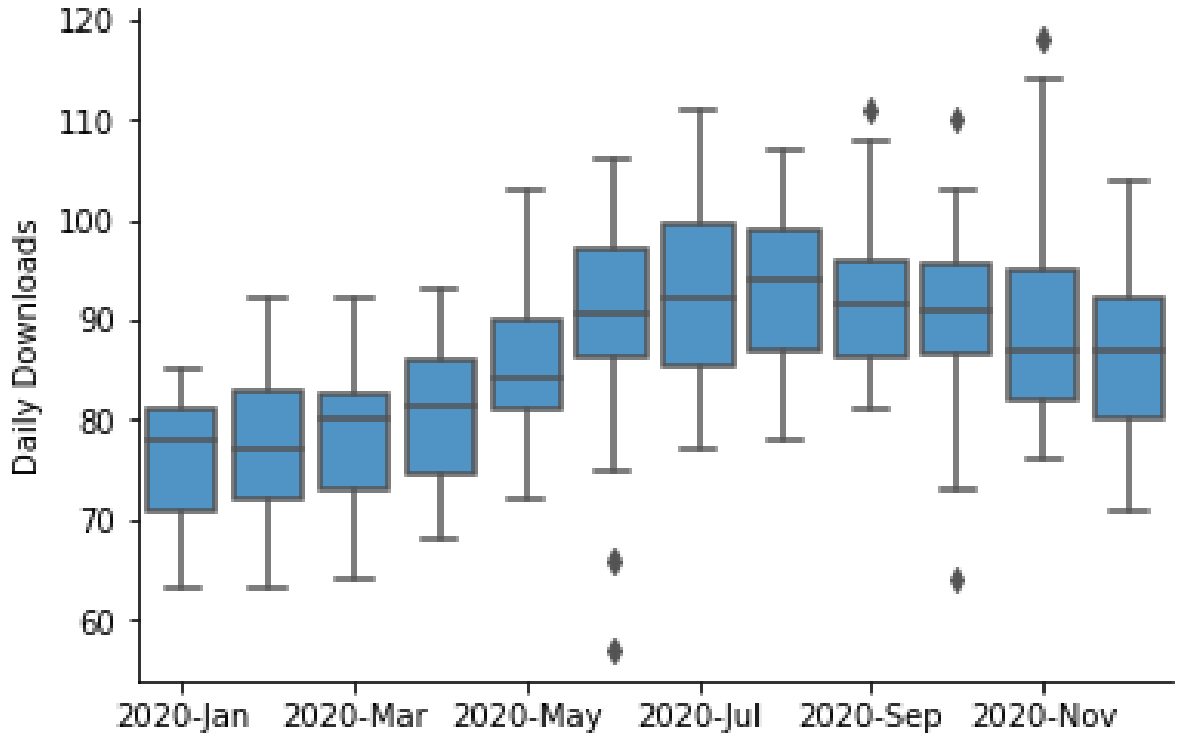
Somaiya TRUST

Types of Statistical Graphs

- Bar Charts
- Box Plots
- Histograms
- Pie Charts
- Scatter Plots

What is a box plot?

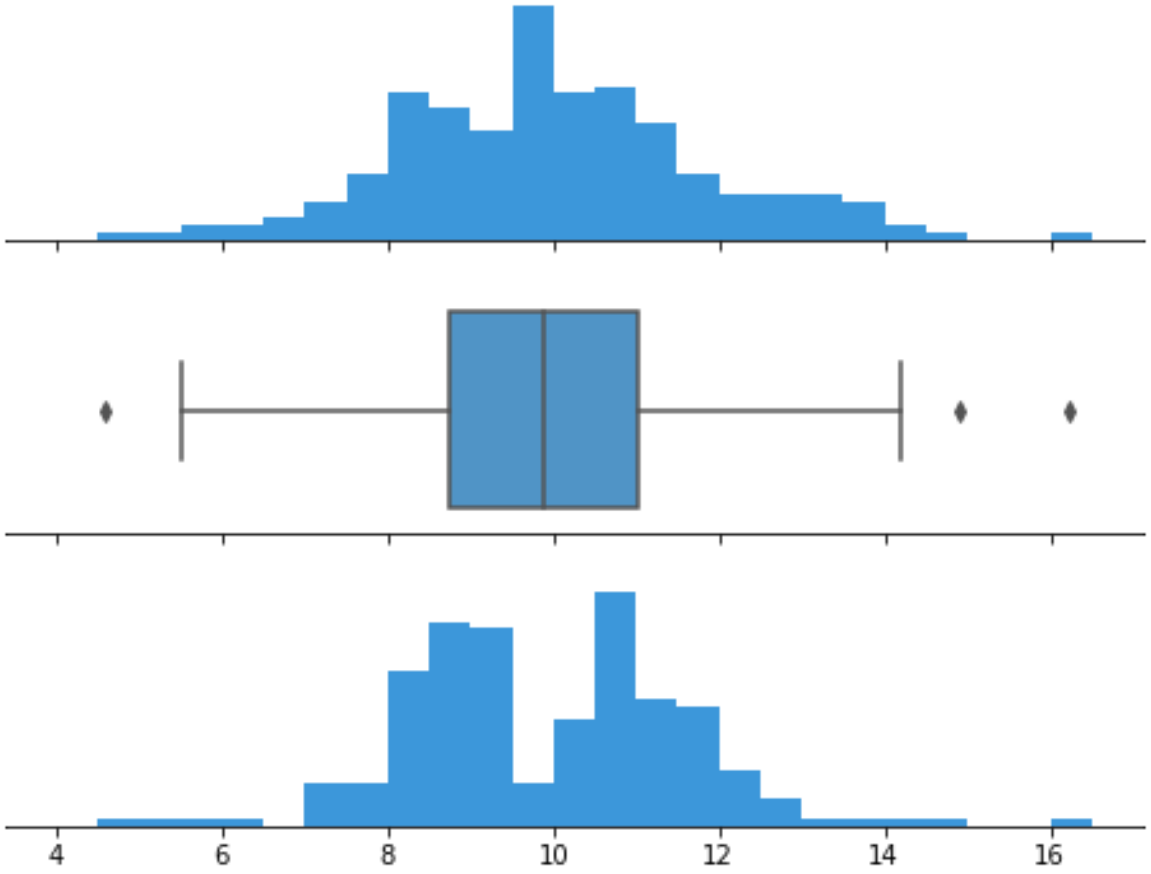
- A box plot (aka box and whisker plot) uses boxes and lines to depict the distributions of one or more groups of numeric data.
- Box limits indicate the range of the central 50% of the data, with a central line marking the median value.
- Lines extend from each box to capture the range of the remaining data, with dots placed past the line edges to indicate outliers.


Daily Downloads
120
110
100
90
80
70
60
2020-Jan
2020-Mar
2020-May
2020-Jul
2020-Sep
2020-Nov

When you should use a box plot

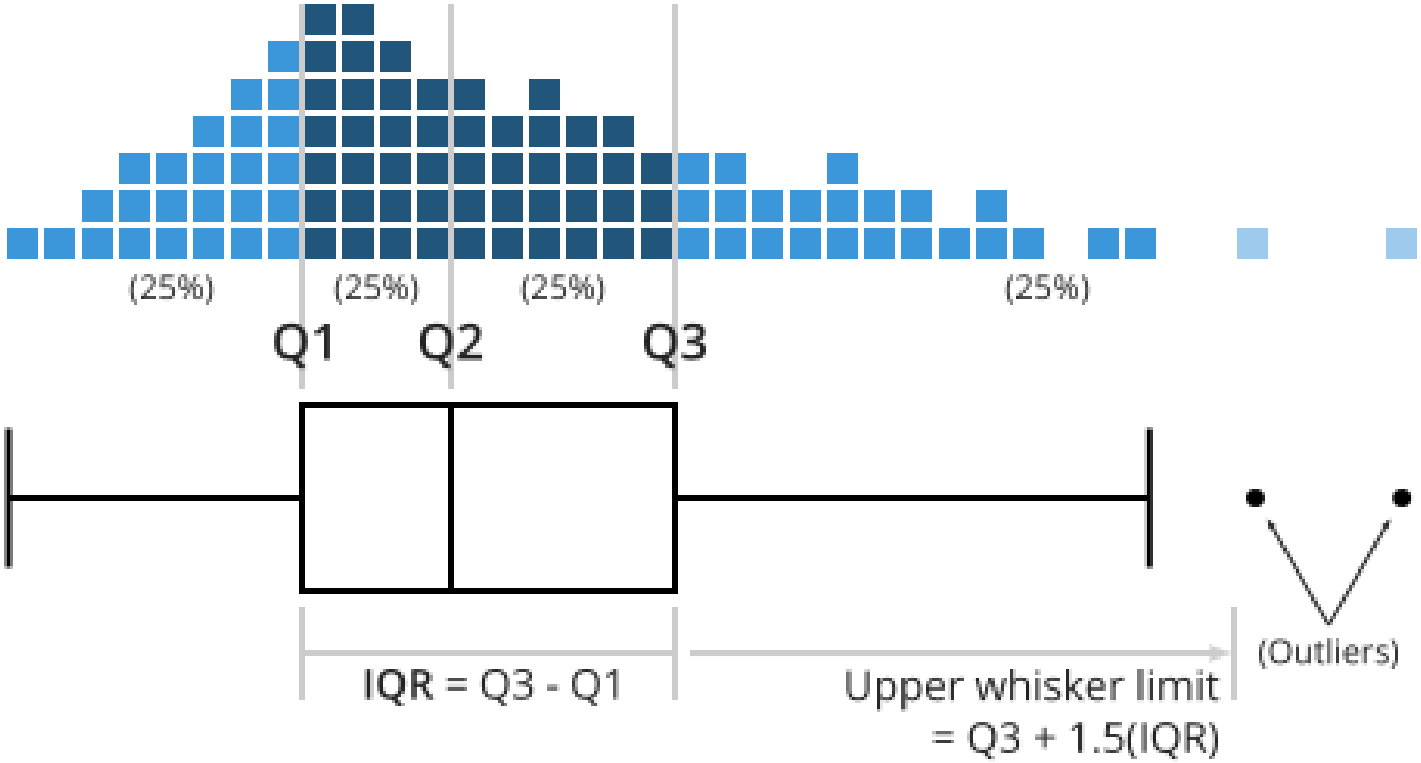
- Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups.
- They are built to provide high-level information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers.
- It is easy to see where the main bulk of the data is, and make that comparison between different groups.

- On the downside, a box plot's simplicity also sets limitations on the density of data that it can show.
- With a box plot, we miss out on the ability to observe the detailed shape of distribution, such as if there are oddities in a distribution's modality (number of 'humps' or peaks) and skew.


4
6
8
10
12
14
16

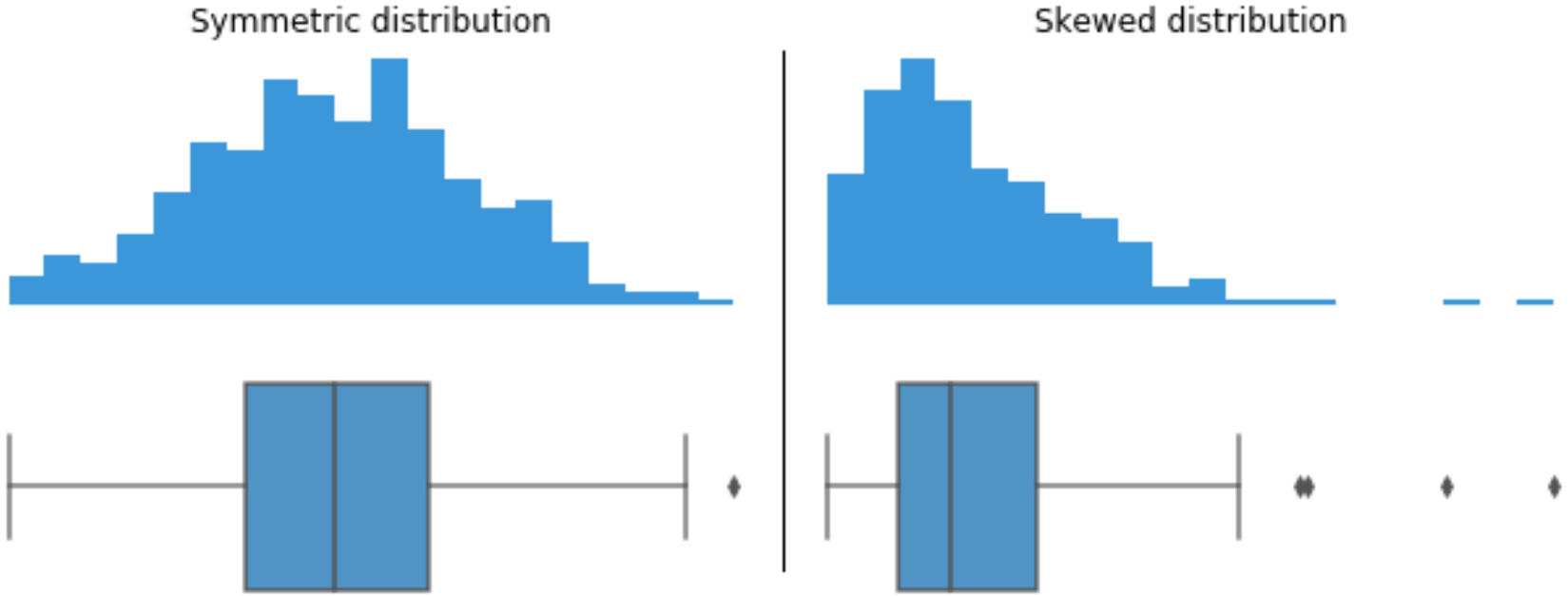
Interpreting a box and whiskers

- Construction of a box plot is based around a dataset's quartiles, or the values that divide the dataset into equal fourths.
- The first quartile (Q1) is greater than 25% of the data and less than the other 75%.
- The second quartile (Q2) sits in the middle, dividing the data in half. Q2 is also known as the median.
- The third quartile (Q3) is larger than 75% of the data, and smaller than the remaining 25%.
- In a box and whiskers plot, the ends of the box and its center line mark the locations of these three quartiles.


(25%)
(25%)
(25%)
(25%)
Q1
Q2
Q3
(Outliers)
IQR = Q3 - Q1
Upper whisker limit
= Q3 + 1.5(IQR)

- The distance between Q3 and Q1 is known as the interquartile range (IQR) and plays a major part in how long the whiskers extending from the box are.
- Each whisker extends to the furthest data point in each wing that is within 1.5 times the IQR.
- Any data point further than that distance is considered an outlier, and is marked with a dot.

- When a data distribution is symmetric, you can expect the median to be in the exact center of the box: the distance between Q1 and Q2 should be the same as between Q2 and Q3.
- Outliers should be evenly present on either side of the box.
- If a distribution is skewed, then the median will not be in the middle of the box, and instead off to the side.
- An imbalance in the whisker lengths, where one side is short with no outliers, and the other has a long tail with many more outliers.

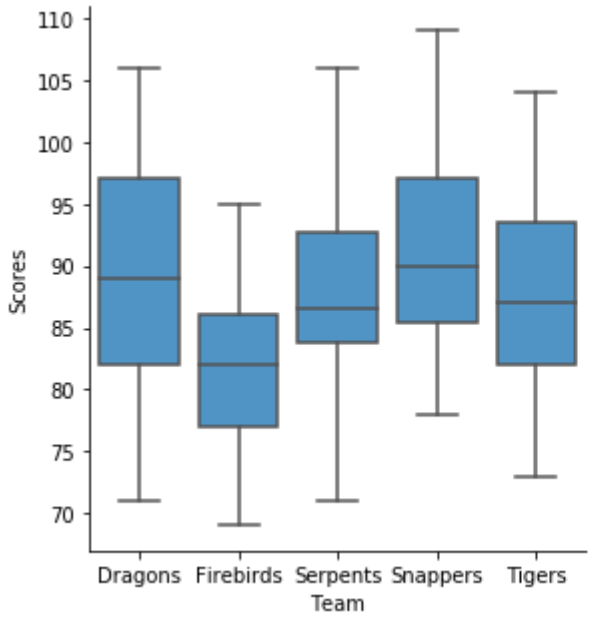


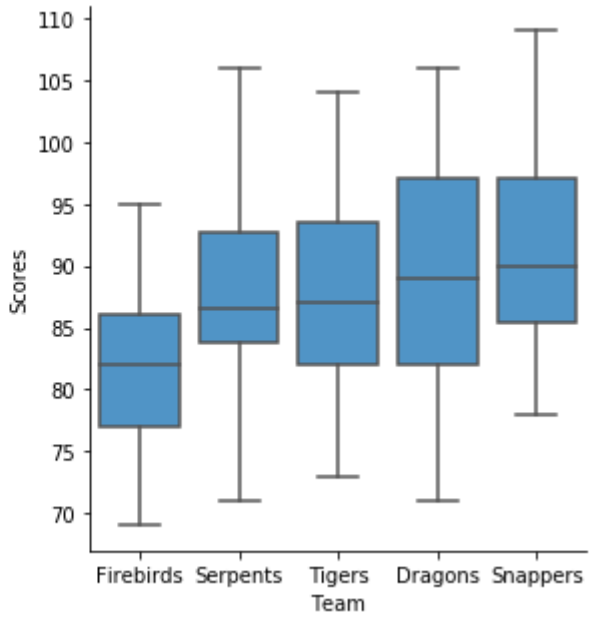
DATE	...	MONTH	DOWNLOADS
2020-01-30	...	2020-01	81
2020-01-31	...	2020-01	78
2020-02-01	...	2020-02	76
2020-02-02	...	2020-02	79
...	...	...	...

Best practices for using a box plot

- Compare multiple groups
  - Box plots are at their best when a comparison in distributions needs to be performed between groups.
  - They are compact in their summarization of data, and it is easy to compare groups through the box and whisker markings' positions.
  - It is less easy to justify a box plot when you only have one group's distribution to plot.
  - Box plots offer only a high-level summary of the data and lack the ability to show the details of a data distribution's shape.
  - With only one group, we have the freedom to choose a more detailed chart type like a histogram or a density curve.

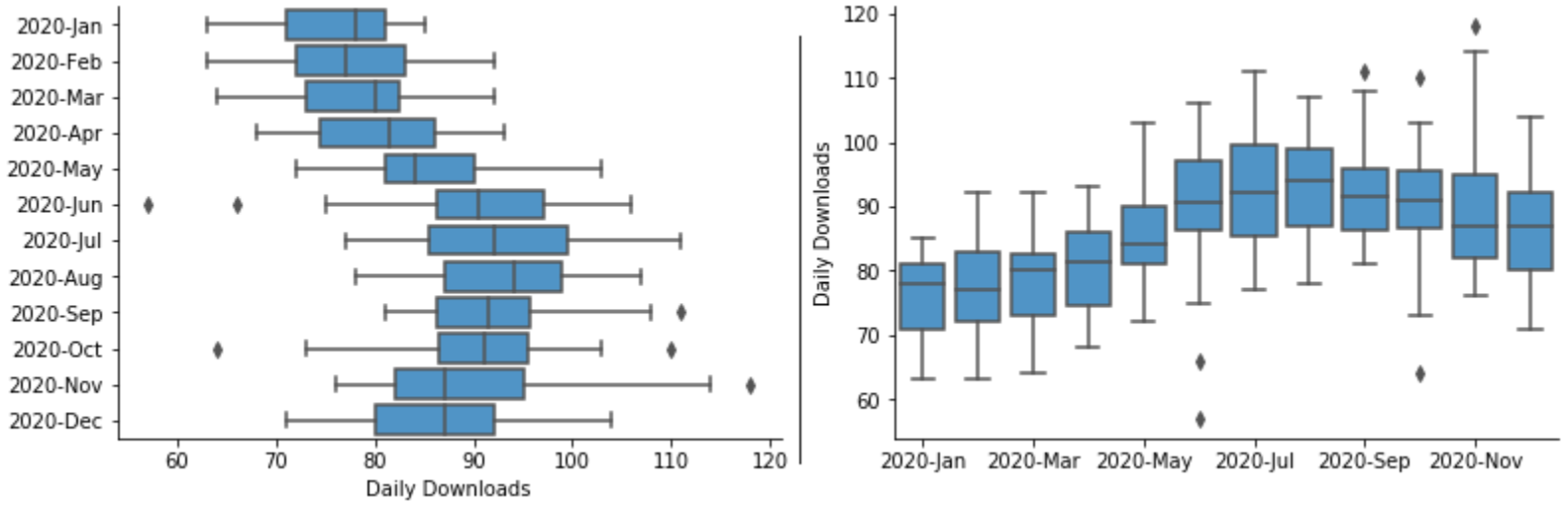
- Consider the order of groups
  - If the groups plotted in a box plot do not have an inherent order, then you should consider arranging them in an order that highlights patterns and insights.
  - One common ordering for groups is to sort them by median value.


110
105
100
95
90
85
80
75
70
Scores
Dragons
Firebirds
Serpents
Snappers
Tigers
Team


110
105
100
95
90
85
80
75
70
Scores
Firebirds
Serpents
Tigers
Dragons
Snappers
Team

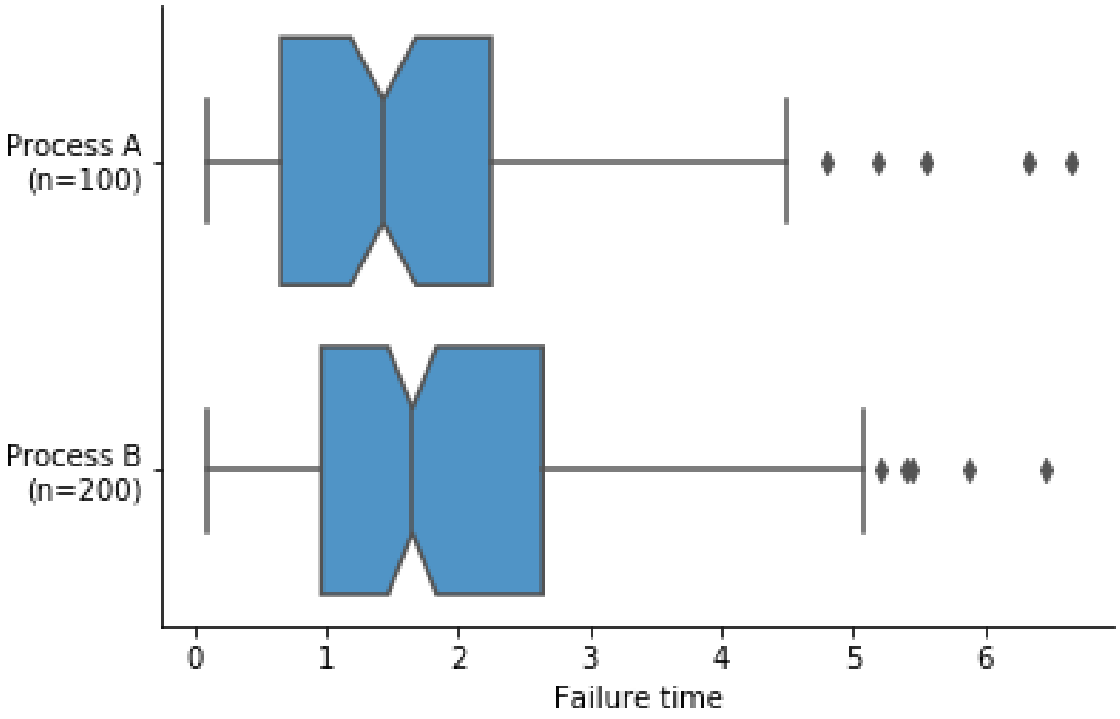
Common box plot options

- Vertical vs. horizontal box plot

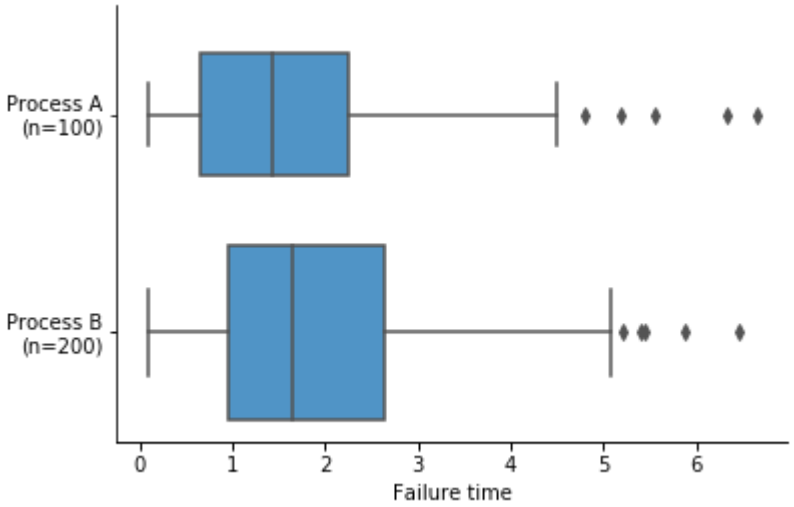


Variable box width and notches

- Certain visualization tools include options to encode additional statistical information into box plots.
- This is useful when the collected data represents sampled observations from a larger population.
- Notches are used to show the most likely values expected for the median when the data represents a sample.
- When a comparison is made between groups, you can tell if the difference between medians are statistically significant based on if their ranges overlap.
- If any of the notch areas overlap, then we can't say that the medians are statistically different; if they do not have overlap, then we can have good confidence that the true medians differ.


Process A
(n=100)
Process B
(n=200)
0
1
2
3
4
5
6
Failure time

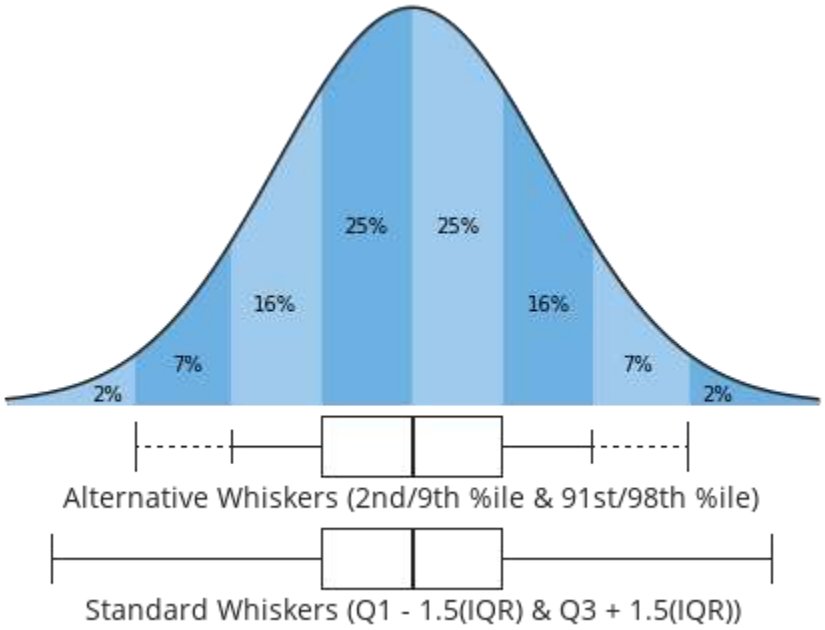
- Box width can be used as an indicator of how many data points fall into each group.
- Box width is often scaled to the square root of the number of data points, since the square root is proportional to the uncertainty (i.e. standard error) we have about true values. Since interpreting box width is not always intuitive, another alternative is to add an annotation with each group name to note how many points are in each group.


Process A
(n=100)
Process B
(n=200)
0
1
2
3
4
5
6
Failure time

Whisker range and outliers

- There are multiple ways of defining the maximum length of the whiskers extending from the ends of the boxes in a box plot.
- As noted above, the traditional way of extending the whiskers is to the furthest data point within 1.5 times the IQR from each box end.
- Alternatively, you might place whisker markings at other percentiles of data, like how the box components sit at the 25th, 50th, and 75th percentiles.

- Common alternative whisker positions include the 9th and 91st percentiles, or the 2nd and 98th percentiles.
- These are based on the properties of the normal distribution, relative to the three central quartiles.
- Under the normal distribution, the distance between the 9th and 25th (or 91st and 75th) percentiles should be about the same size as the distance between the 25th and 50th (or 50th and 75th) percentiles, while the distance between the 2nd and 25th (or 98th and 75th) percentiles should be about the same as the distance between the 25th and 75th percentiles.
- This can help aid the at-a-glance aspect of the box plot, to tell if data is symmetric or skewed.

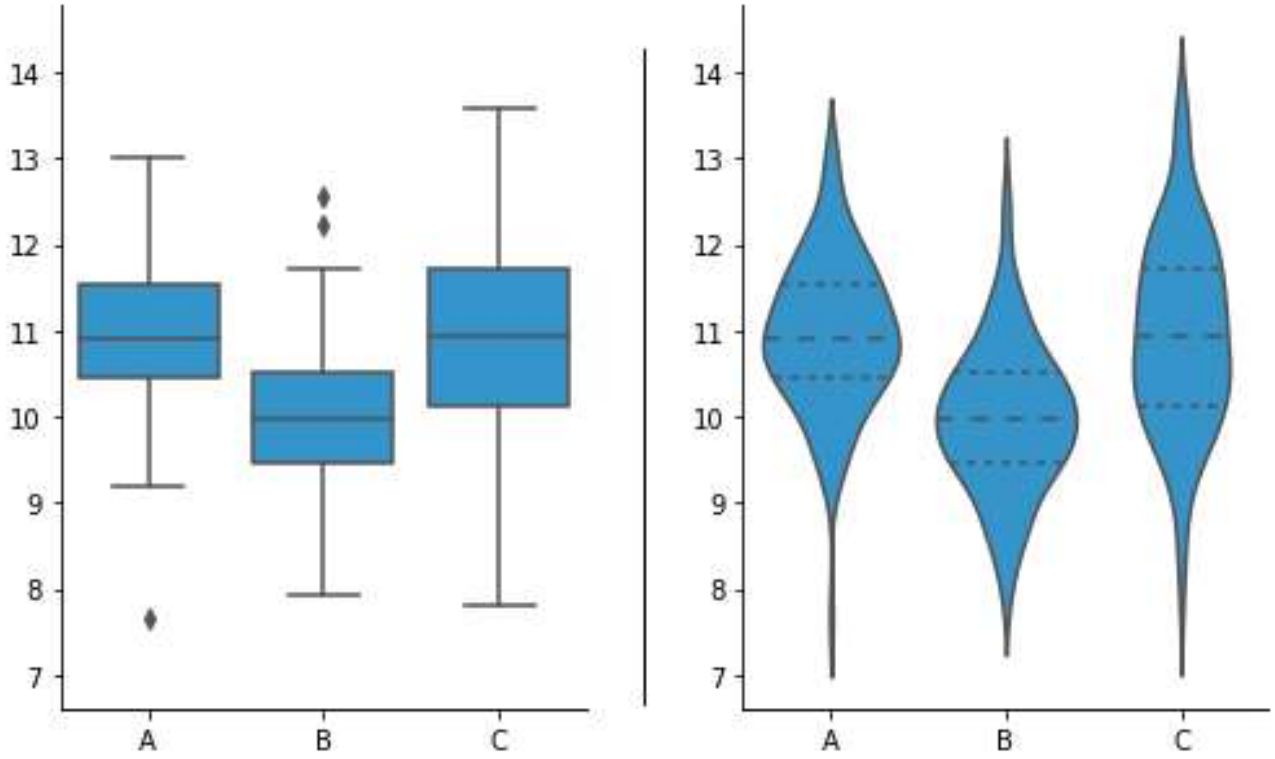

25%
25%
16%
16%
7%
7%
2%
2%
Alternative Whiskers (2nd/9th %ile & 91st/98th %ile)
Standard Whiskers (Q1 - 1.5(IQR) & Q3 + 1.5(IQR))

Letter-value plots

- developed by Hofmann, Kafadar, and Wickham, letter-value plots are an extension of the standard box plot.
- Letter-value plots use multiple boxes to enclose increasingly-larger proportions of the dataset.
- The first box still covers the central 50%, and the second box extends from the first to cover half of the remaining area (75% overall, 12.5% left over on each end).
- The third box covers another half of the remaining area (87.5% overall, 6.25% left on each end), and so on until the procedure ends and the leftover points are marked as outliers.

Read Yourself

- Violin plot
- One alternative to the box plot is the violin plot. In a violin plot, each group's distribution is indicated by a density curve.
- In a density curve, each data point does not fall into a single bin like in a histogram, but instead contributes a small volume of area to the total distribution.
- Violin plots are a compact way of comparing distributions between groups. Often, additional markings are added to the violin plot to also provide the standard box plot information, but this can make the resulting plot noisier to read.


14
13
12
11
10
9
8
7
A
B
C
14
13
12
11
10
9
8
7
A
B
C

Question ?