Computation and Visualization for Analytics

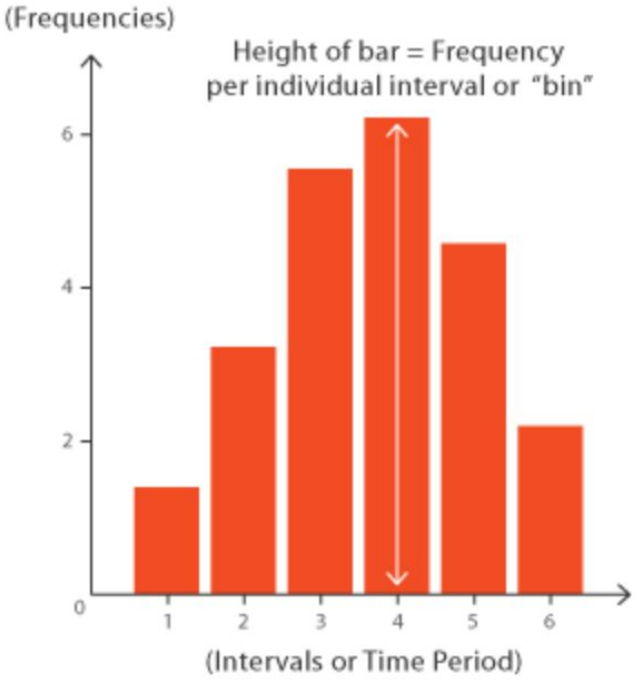
Spring 2021

Week 6.1

Visualizing Amounts

- Distributions
- Relationships

Histogram



Use

- Numerical variable distribution

Histogram Rules

- Use appropriate bin widths
- Avoid multivariable histograms that overlap

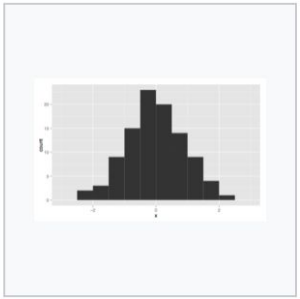
Optimum Binwidth and Number of Bins

- Freedman–Diaconis rule can be used to select the width of the bins

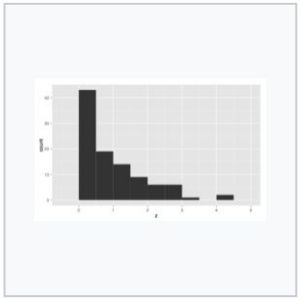
$$\text{Bin width} = 2\,\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

Number of bins = (max-min) / binwidth

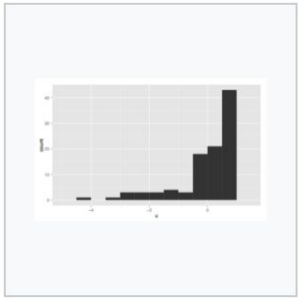
Histogram Patterns



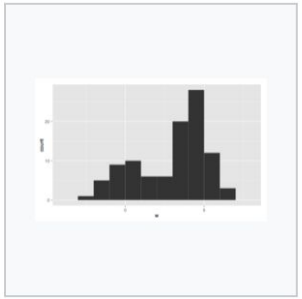
Symmetric, unimodal



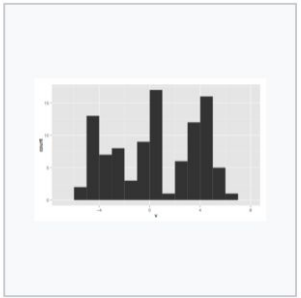
Skewed right



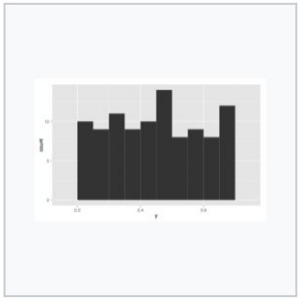
Skewed left



Bimodal

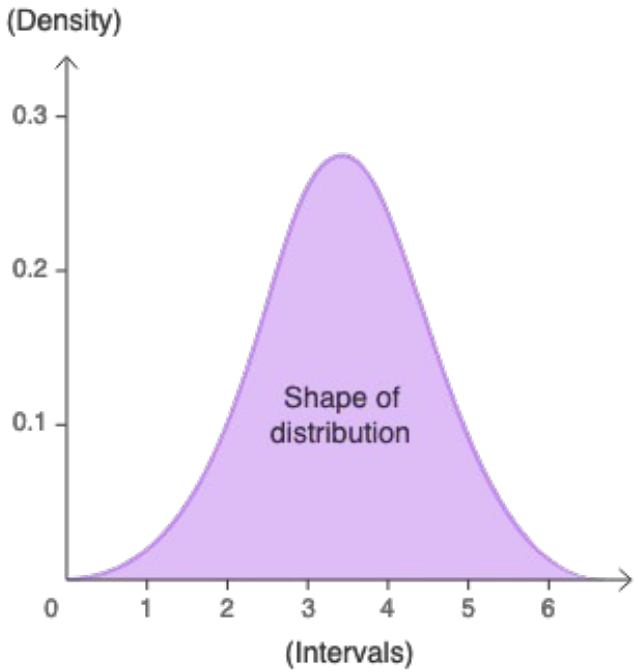


Multimodal



Symmetric

Density Plot



Use

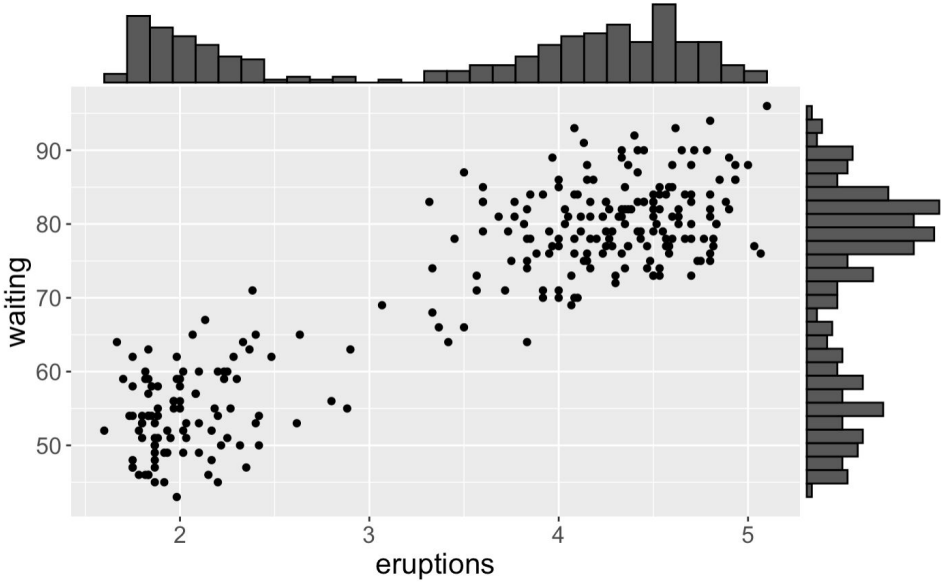
- Numerical variable distribution

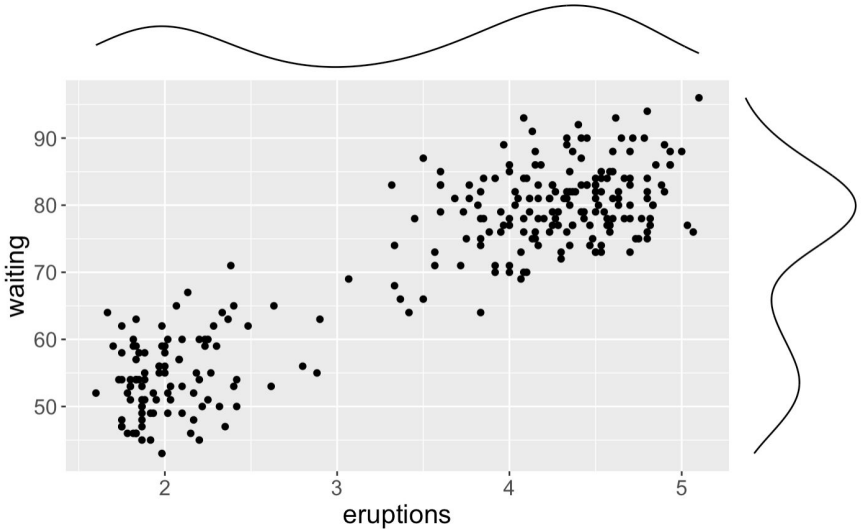
Density Plot

- Insensitive to binwidth
- Can be used for showing overlapping distributions

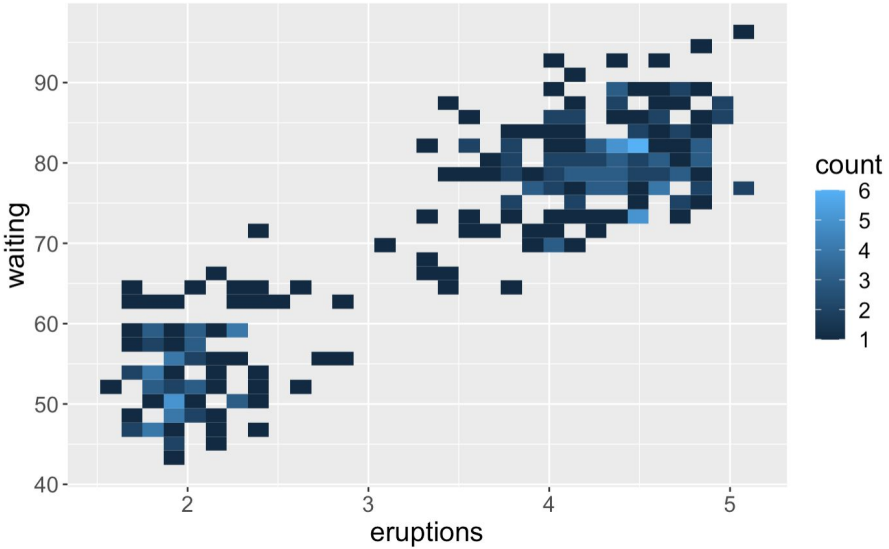
Task 1: Find the drawbacks of density plot

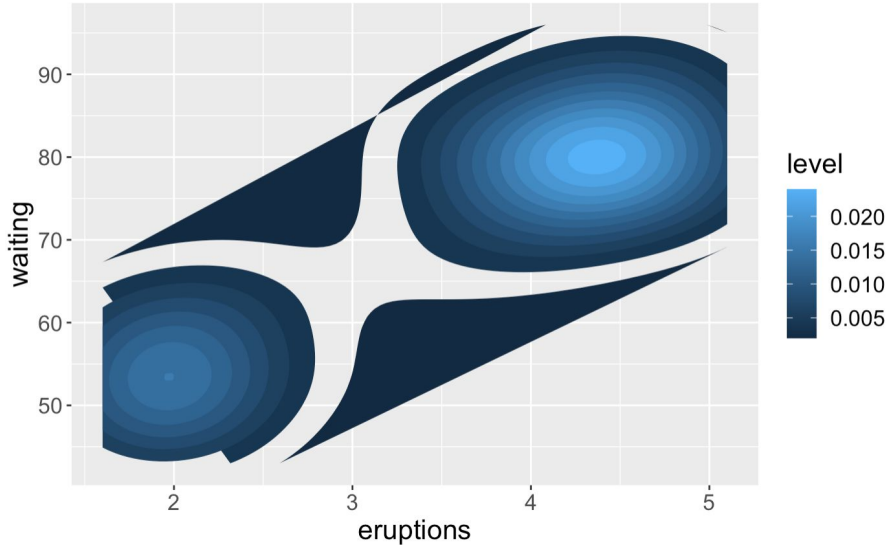
Marginal Plots



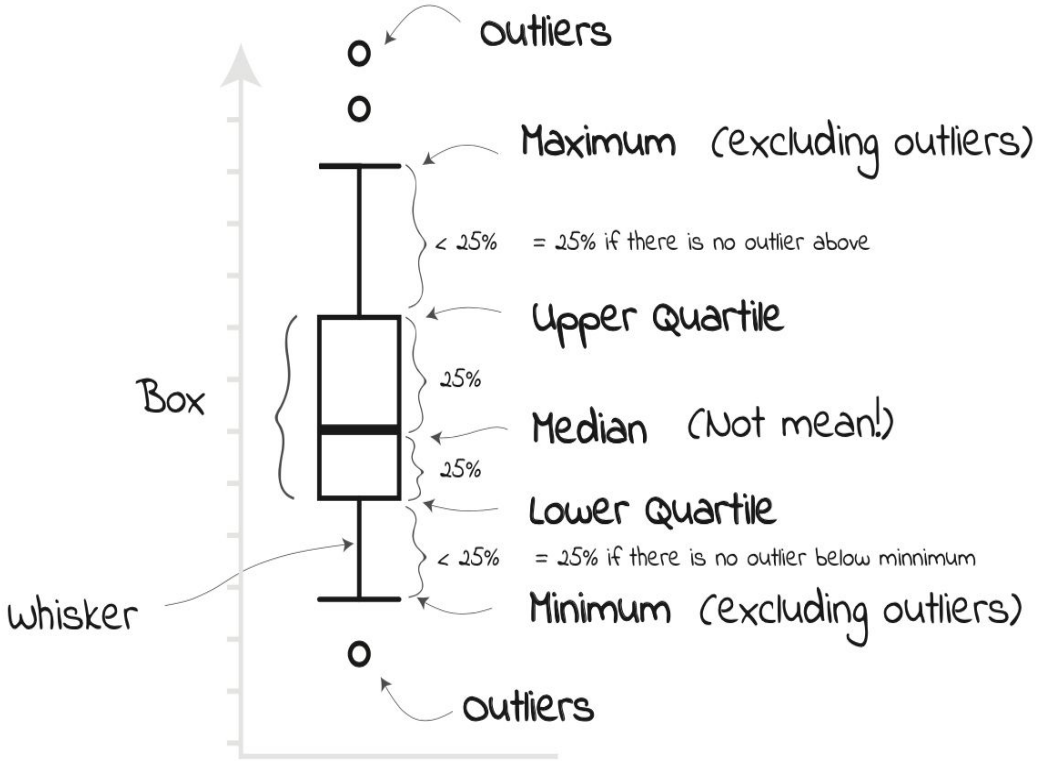


2D Histogram and Density Plot





Box Plot



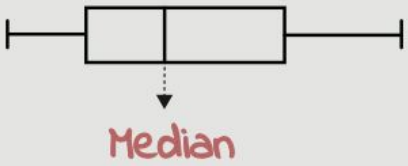
Use

- Numerical variable distribution

https://visualizationcheatsheets.github.io/boxplot.html

Box Plot Example

The 'median' splits the data set into two equal groups.



45 58 88 111 125 140 155 168 170 195 211 224 235 255 260 285 320 333 380

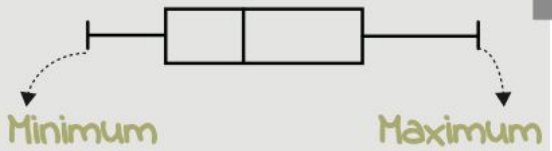
https://visualizationcheatsheets.github.io/boxplot.html

The 'lower quartile' and 'upper quartile' are the median values of lower half and higher half respectively.

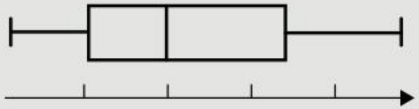
45 58 88 111 125 140 155 168 170 195 211 224 235 255 260 285 320 333 380

https://visualizationcheatsheets.github.io/boxplot.html

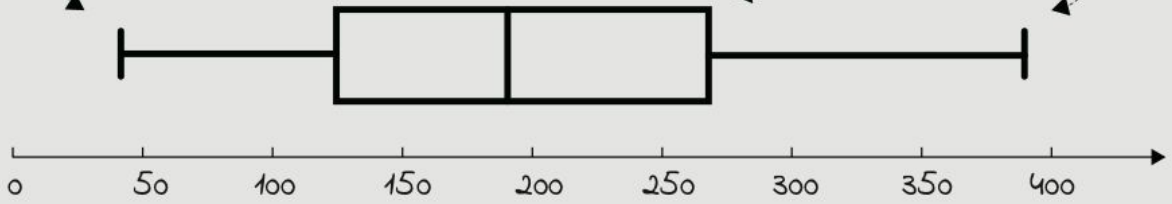
Find the 'minimum' and 'maximum'.



45 58 88 111 125 140 155 168 170 195 211 224 235 255 260 285 320 333 380



The last step, put these 5 values on the axis, draw a 'box' between 'lower quartile' and 'upper quartile' and link two 'whiskers' to 'minimum' and 'maximum'.

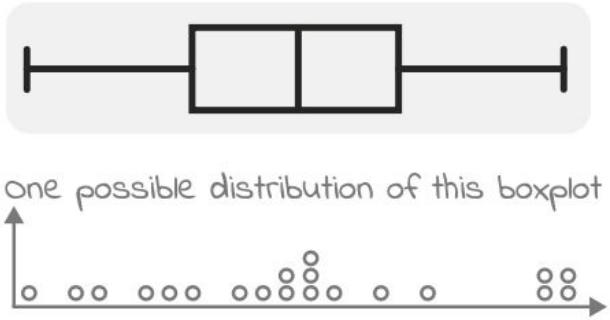


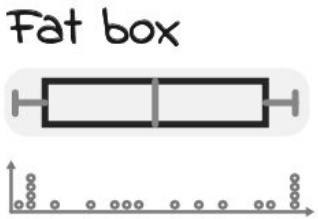
https://visualizationcheatsheets.github.io/boxplot.html

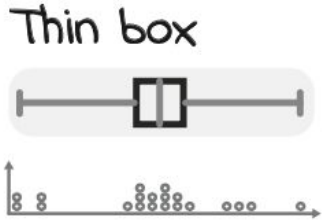
Box Plot Patterns

Ballanced

If the box plot will look **symmetric**, the distribution will be **normal**, there are few exceptionally large or small values. The mean will be about the same as the median.





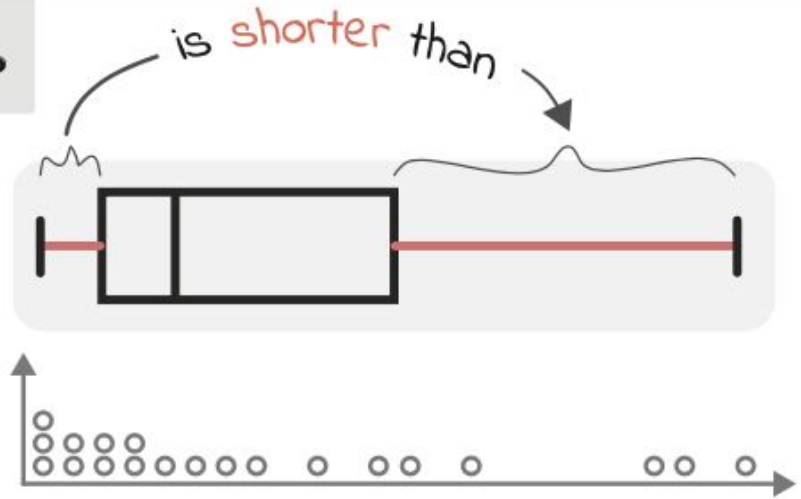


https://visualizationcheatsheets.github.io/boxplot.html

Box Plot Patterns

Possitive Skewness

A distribution with a **positive skew** would have a **longer** whisker in the **positive direction** than in the negative direction.

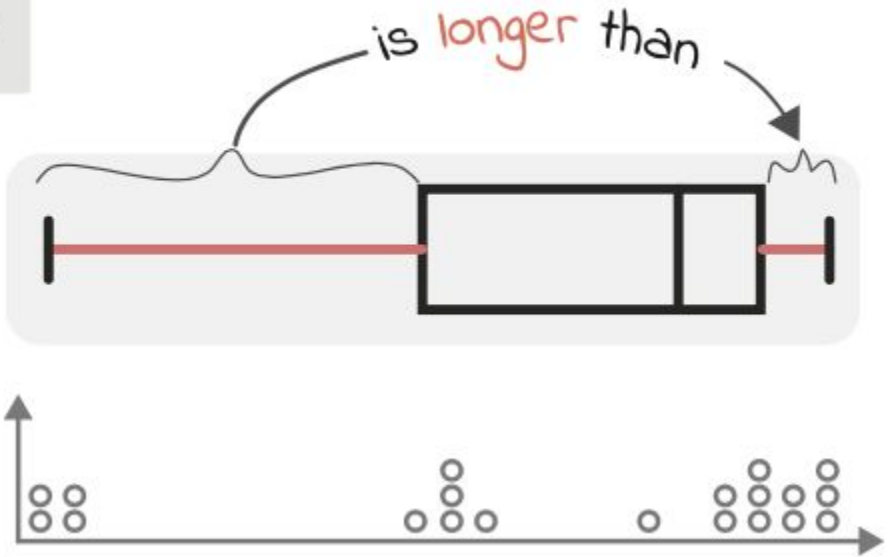


https://visualizationcheatsheets.github.io/boxplot.html

Box Plot Patterns

Negative Skewness

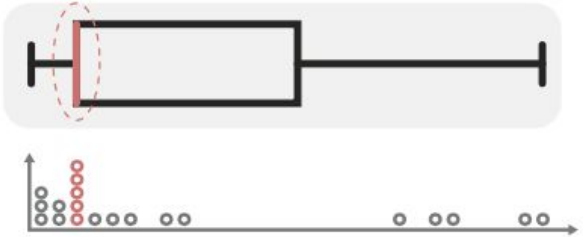
A distribution with a **negative skew** would have a **longer** whisker in the **negative direction** than in the positive direction.



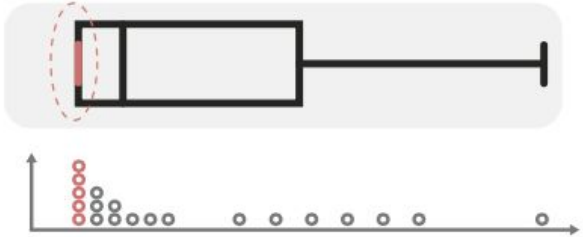
https://visualizationcheatsheets.github.io/boxplot.html

Box Plot Patterns

overlapping



The lower quartile and the median are **overlapped**, this occurs when the 25% of values are same between the lower quartile and the median.

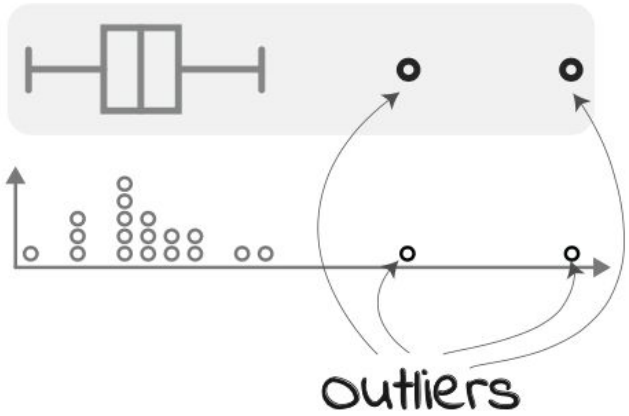


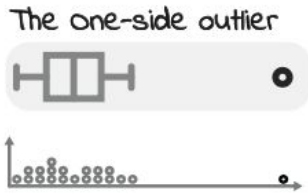
The minimum and the lower quartile are **overlapped**, this occurs when the 25% of values are same between the minimum and the lower quartile.

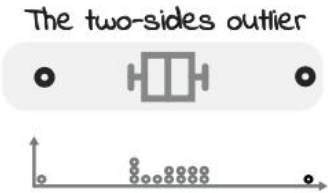
Box Plot Patterns

outlier(s)

when reviewing a boxplot, an outlier is defined as a data point that is located **outside the whiskers** of the boxplot (e.g: outside 1.5 times the interquartile range above the upper quartile and bellow the lower quartile).

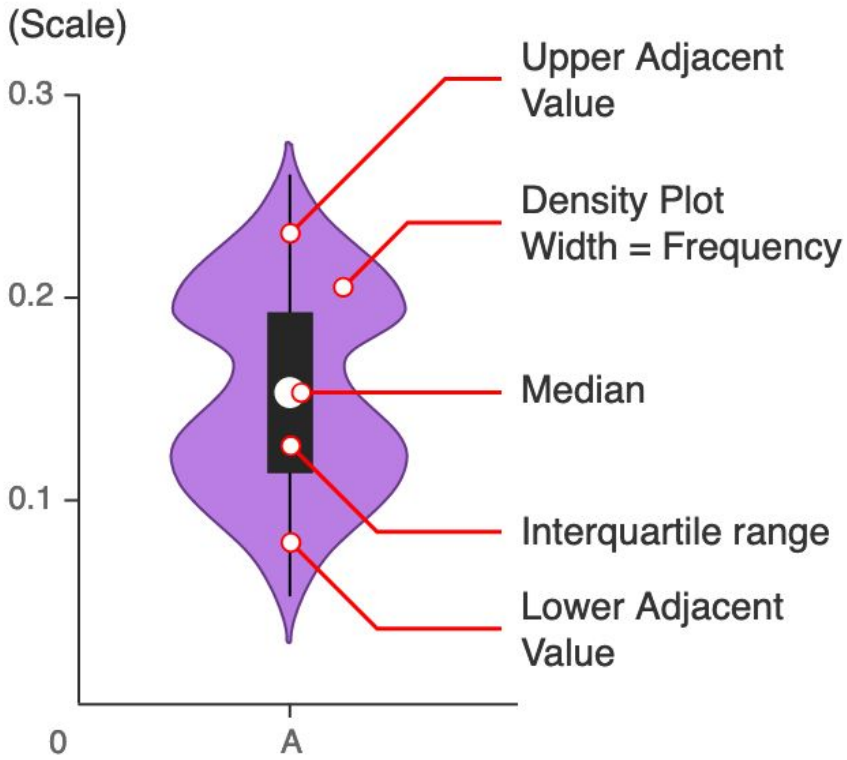






https://visualizationcheatsheets.github.io/boxplot.html

Violin Plot



Use

- Numerical variable distribution

Why Use Violin Plot?