The Data Driven Manager

3

Describing and Visualizing Data

Data Analysis

The first step in analysis of data has two parts:

- **Graphics**: pictures that provide a visual representation of what the numbers describe or identify
- **Numerics**: numbers and statistical calculations which summarize and describe our data

Data Analysis

- We **always** use both pictures and numbers (‘never present a picture without stats; never present stats without a picture’!)

Describing Data Graphically

Learning Objectives

- Create a run chart using RStudio and ROIStat software
- Create an ungrouped histogram using RStudio
- Create a grouped histogram using RStudio
- Create a histogram using ROIStat software
- Interpret histogram patterns

Learning Objectives

- Create a density plot using RStudio and ROIStat software
- Create a box and whisker plot using RStudio and ROIStat software
- Create a scatter plot using RStudio and ROIStat software

Common Methods

- Run Charts
- Histograms
- Density Plots
- Box and Whisker Plots
- Bar Chart
- Scatter Plot

Run Charts

Presenting data as observed through time

Run Chart

- An engineer gathered 20 consecutive computer fans from a production line, keeping track of the order in which the fans were produced.

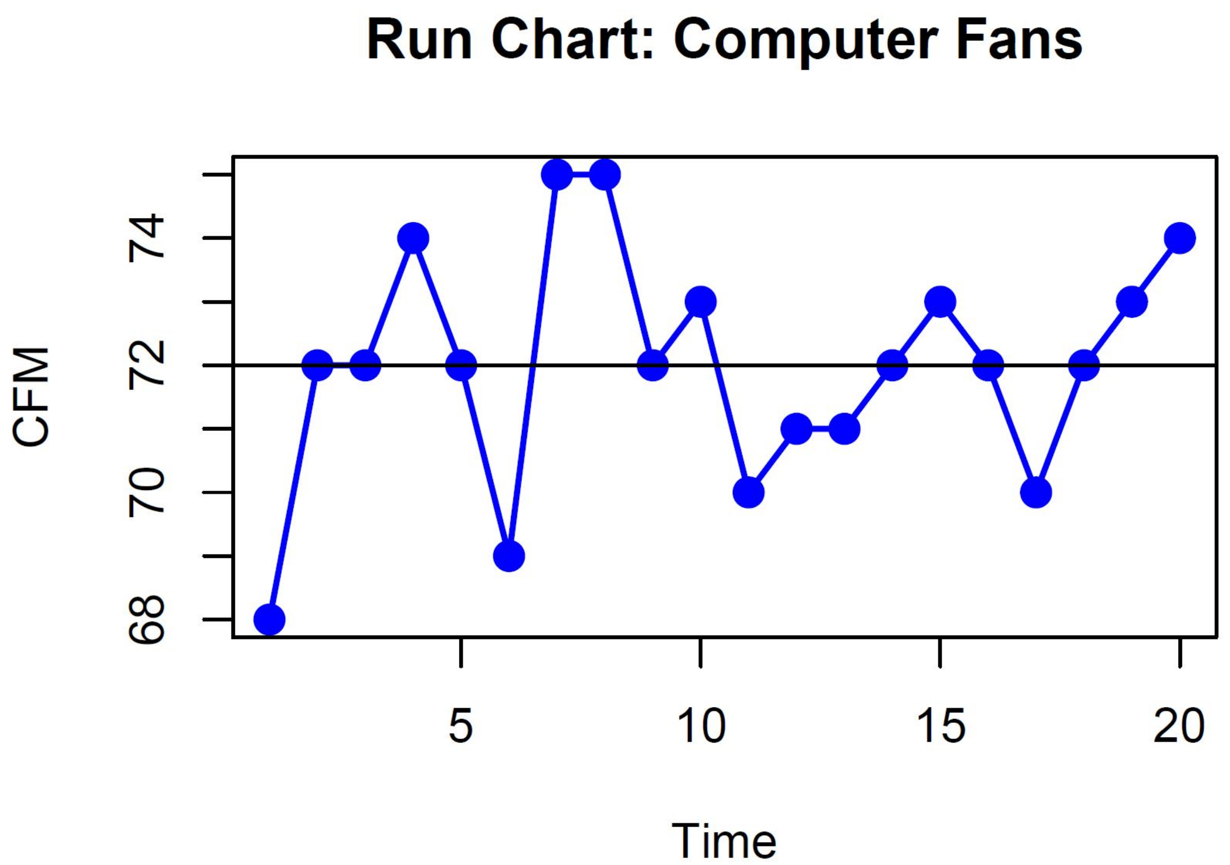
Run Chart

- Then these fans were tested for air flow in CFM. This testing produced the following data for the 20 fans, presented in time order

Fans 1-10:	68	72	72	74	72	69	75	75	72	73
Fans 10-20:	70	71	71	72	73	72	70	72	73	74

Create a Run Chart

In RStudio


Run Chart: Computer Fans
CFM
68
70
72
74
5
10
15
20
Time

1. Create the Data File

Create a Vector

```
cfm<-c(68,72,72,74,72,69,75,75,72,73,70,71,71
,72,73,72,70,72,73,74)
```

Store the Variable in a data frame

```
fans <- data.frame(cfm)
```

```
View(fans)
```

2. Create the Run Chart in RStudio

```
require(lolcat)
spc.run.chart(fans$cfm, main = "Run Chart:
Computer Fans", ylab = "CFM")
```

3. Add a Horizontal Line

```
abline(h=72)
```

Options for Customization

- Point symbol: pch = (1-25)
- Point size: cex =
- Color: col = “red” (color name or hex code)
- Line type: lty = (0-6)
- Line width: lwd =

Create a Run Chart

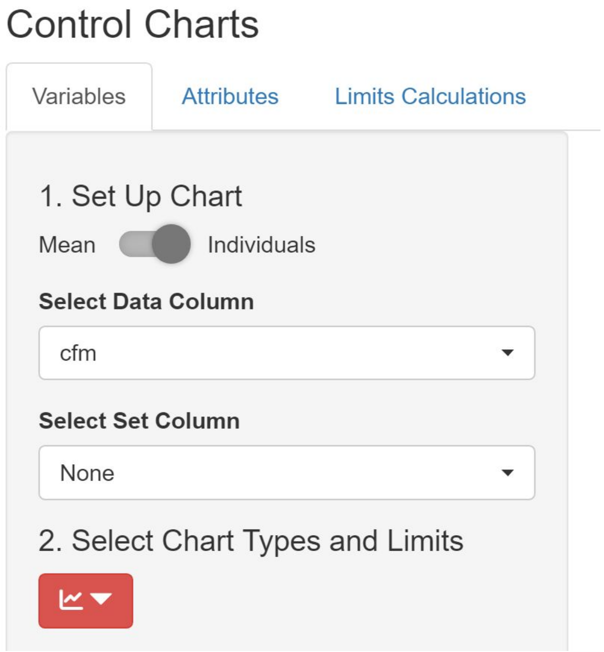
In ROIStat

1. Import the Data File

Import the file named **fans.txt**

Click on the tab labeled SPC

Under Set up Chart, select Individuals



2. Create the Run Chart in ROIStat

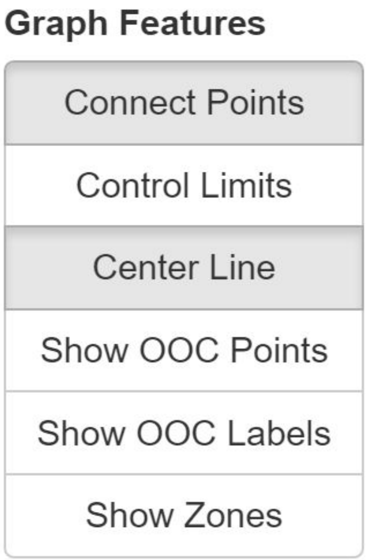
Click on the red button to create what is called an X and Moving Range Chart

Click on the green button with the gear icon to modify the chart

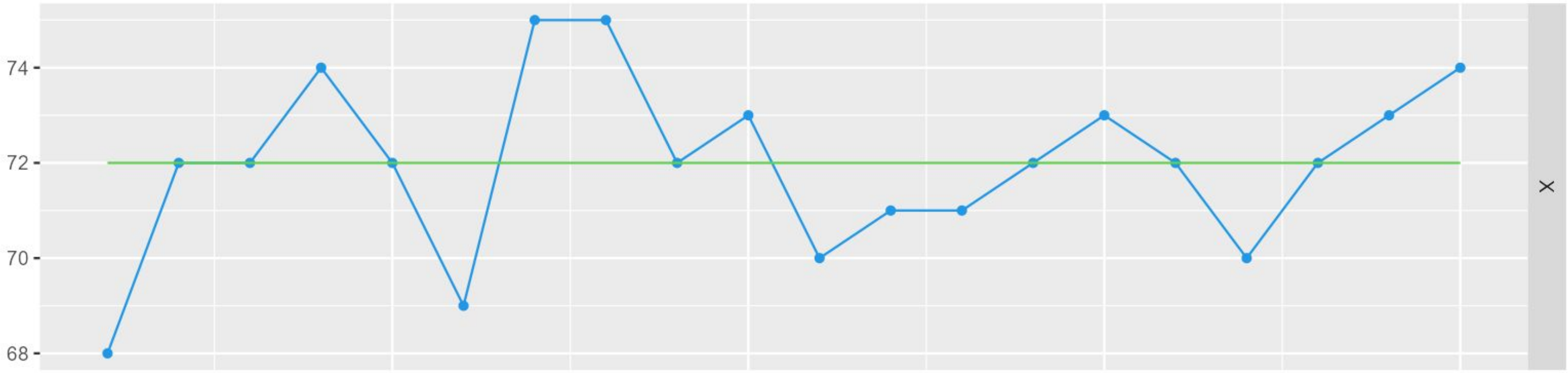
3. Create the Run Chart in ROIStat

Deselect Control Limits and
Show OOC Points

Leave Connect Points and Centerline



4. Copy the Top Chart



Histograms

Presenting data to portray the nature of a distribution

Histograms

Useful for:

- Evaluating a manufacturing or business process
- Determining machine and process capabilities
- Comparing material, vendor, operator, process and product characteristics

Ungrouped vs Grouped Histograms

- Use ungrouped when there are fewer than 20 unique data values in the data set

- Use grouped when there are more than 20 unique data values in the data set

Ungrouped Histograms

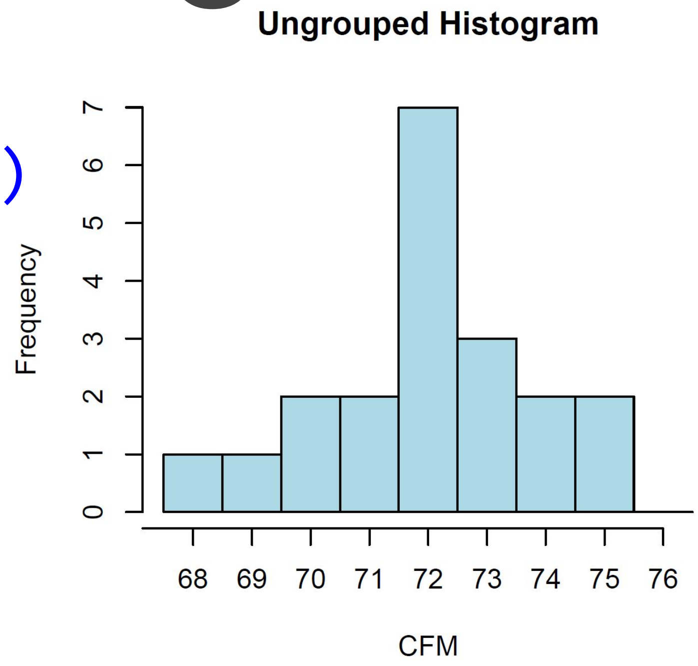
- Histograms are ungrouped when each bar, or class interval, consists of only one score, value or observation

Create an Ungrouped Histogram

In RStudio

Ungrouped Histogram

```
> hist.ungrouped(fans$cfm)
```



Create an Ungrouped Histogram

In ROIStat

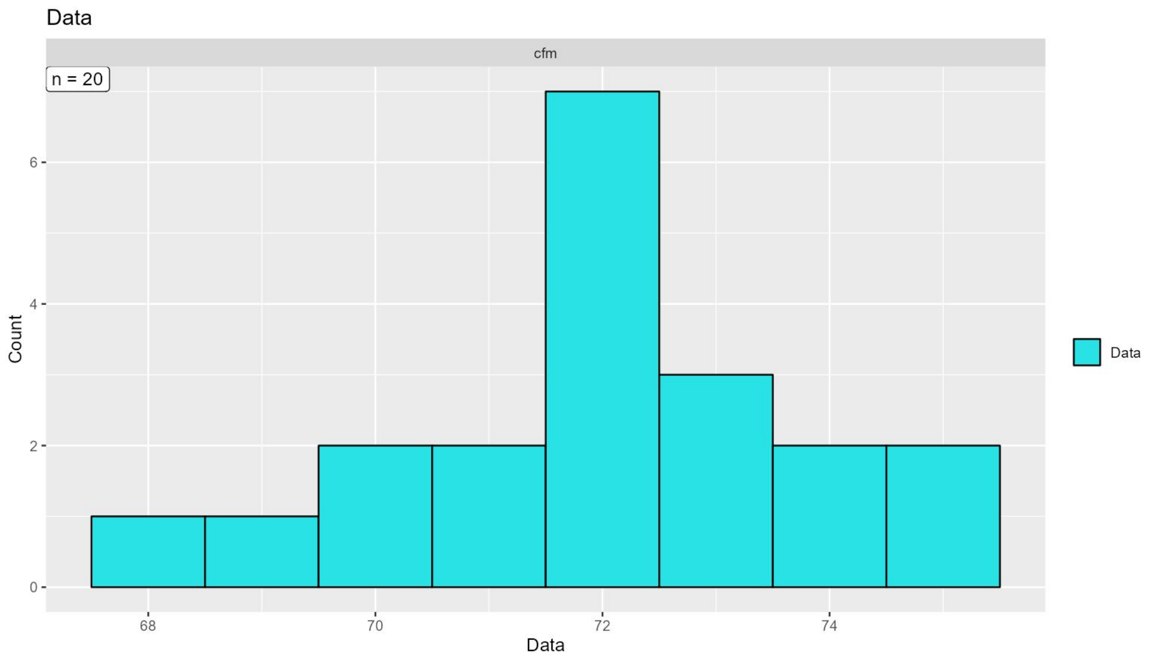
Ungrouped Histograms

- Histograms are ungrouped when each bar, or class interval, consists of only one score, value or observation

Ungrouped Histogram

1. File > Import Data
2. EDA > Data Setup > Analyze Columns
3. Select ‘cfm’
4. Click on Histograms tab

Ungrouped Histogram



Create a Grouped Histogram

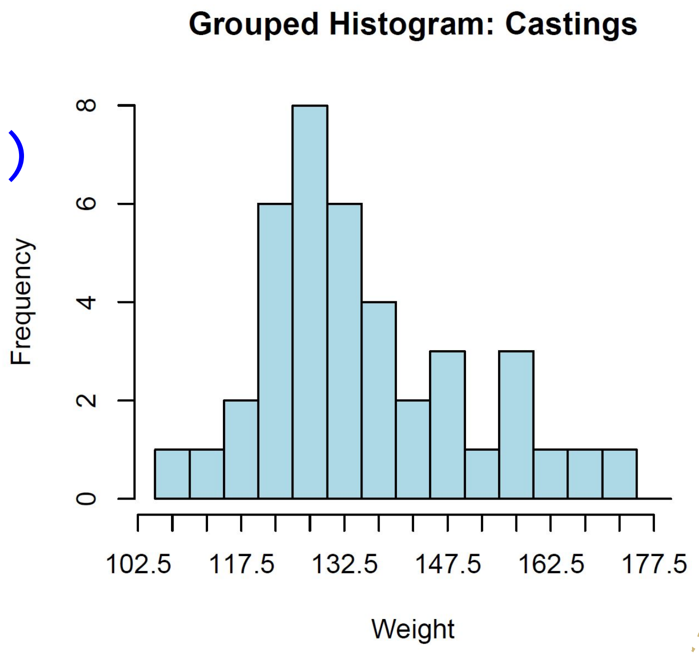
In RStudio

Grouped Histograms

- As a guideline, we'd like to create a grouped histogram with 10 class intervals.
- If this is not possible, a grouped histogram should have between 10 and 20 class intervals.

Grouped Histogram

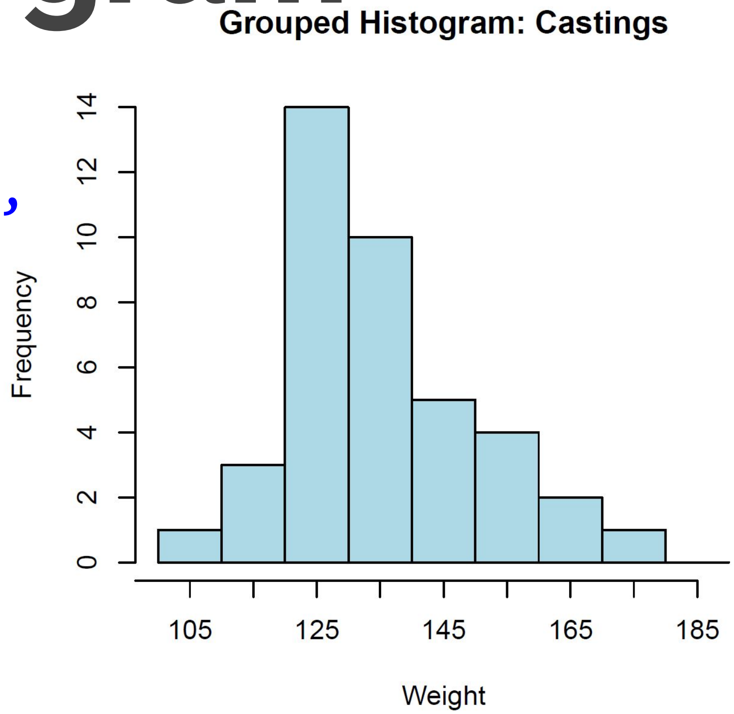
```
> hist.grouped(castings$weight)
```



Grouped Histogram

```
> hist.grouped(castings$weight,
             interval.size = 10)
```

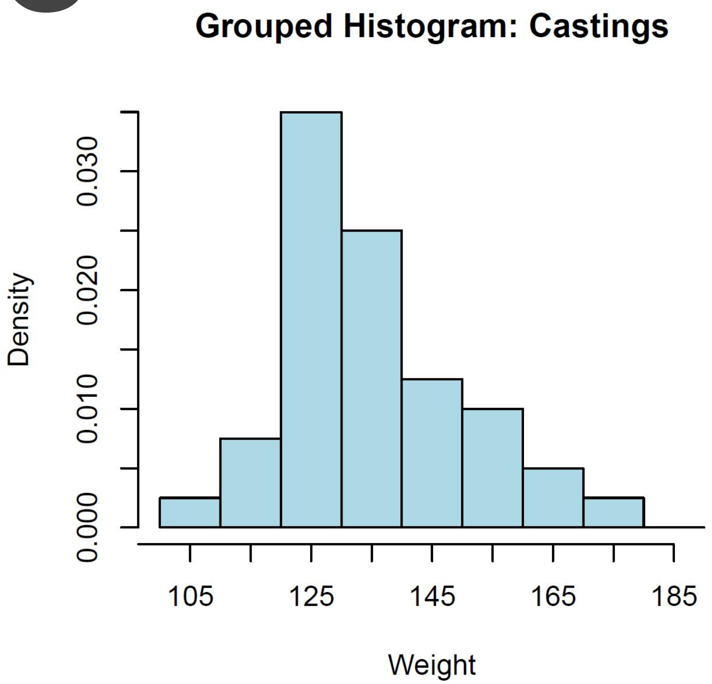
Note frequency (count) on Y axis



Grouped Histogram

```
> hist.grouped(castings$weight
               , interval.size = 10
               , freq=F)
```

Note density
on Y axis



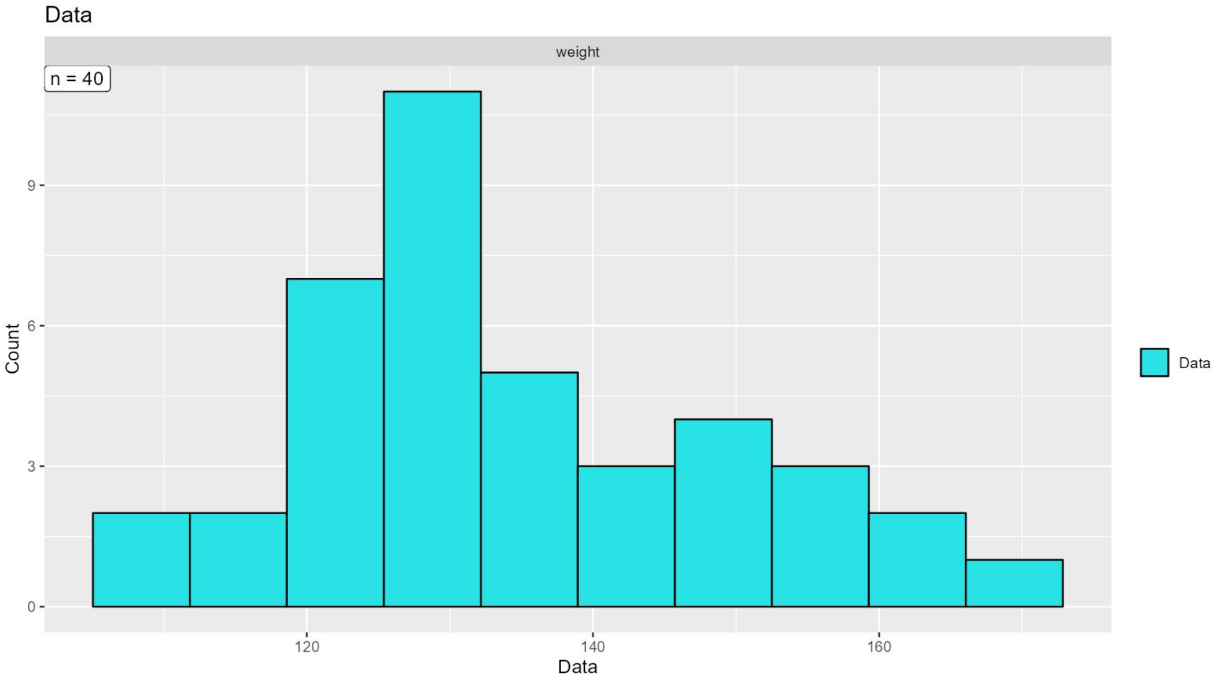
Create a Grouped Histogram

In ROIStat

Grouped Histogram

1. File > Import Data
2. EDA > Data Setup > Analyze Columns
3. Select ‘weight’
4. Click on Histograms tab

Grouped Histogram



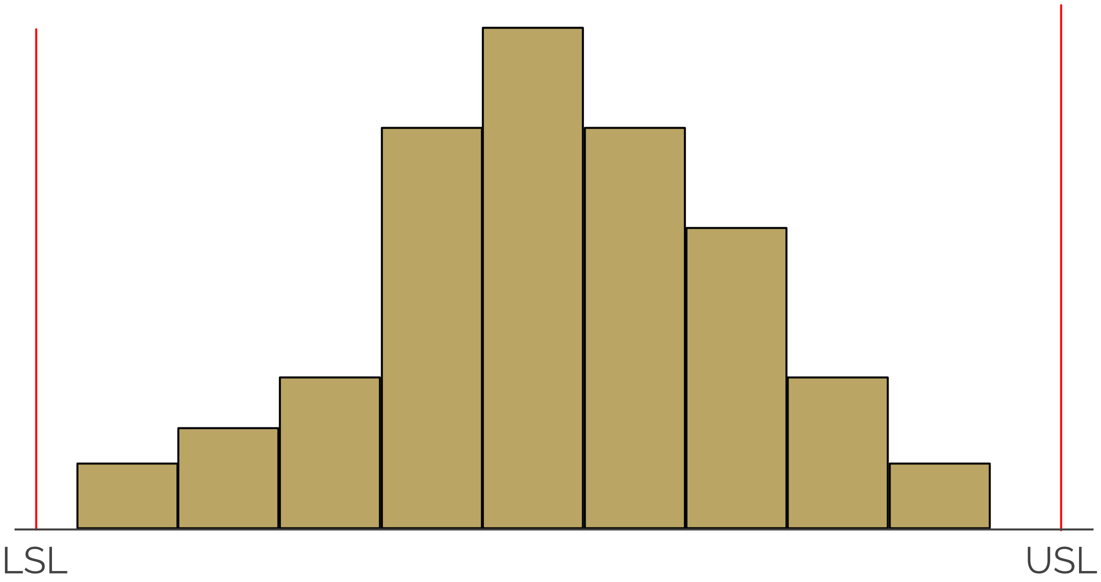
Histogram Patterns

Interpretation

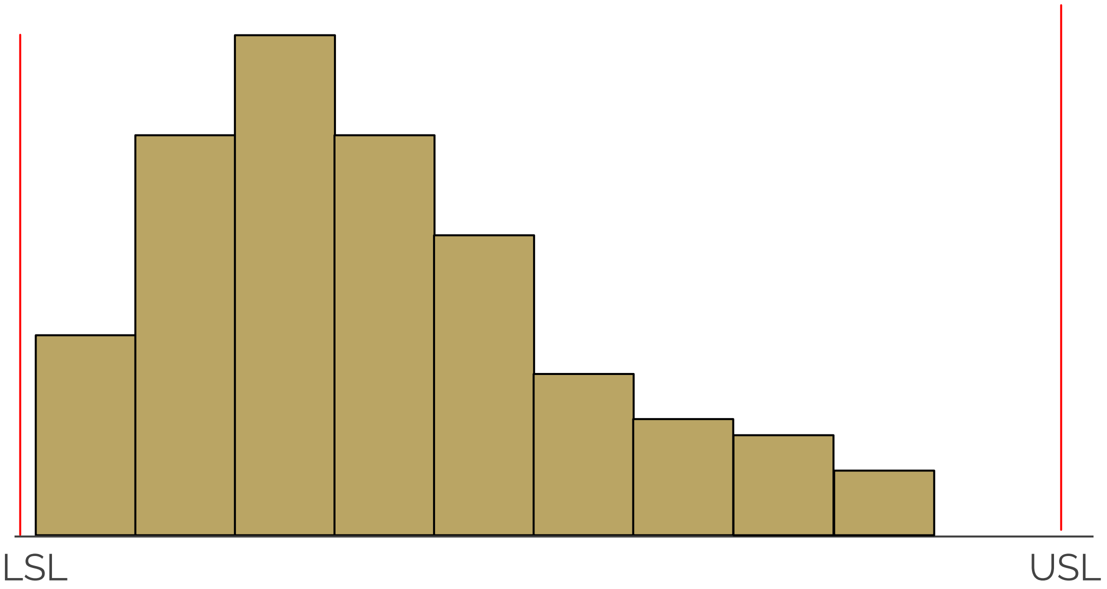
Histogram Patterns

- The center, spread and shape of a histogram can give us clues as to what the data are telling us

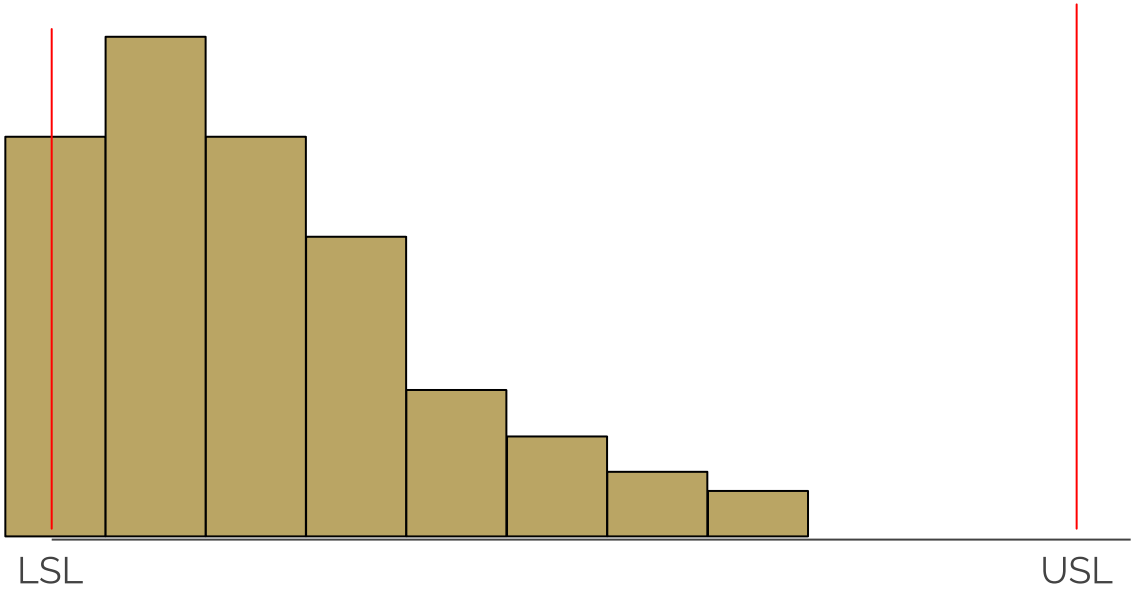
Pattern 1



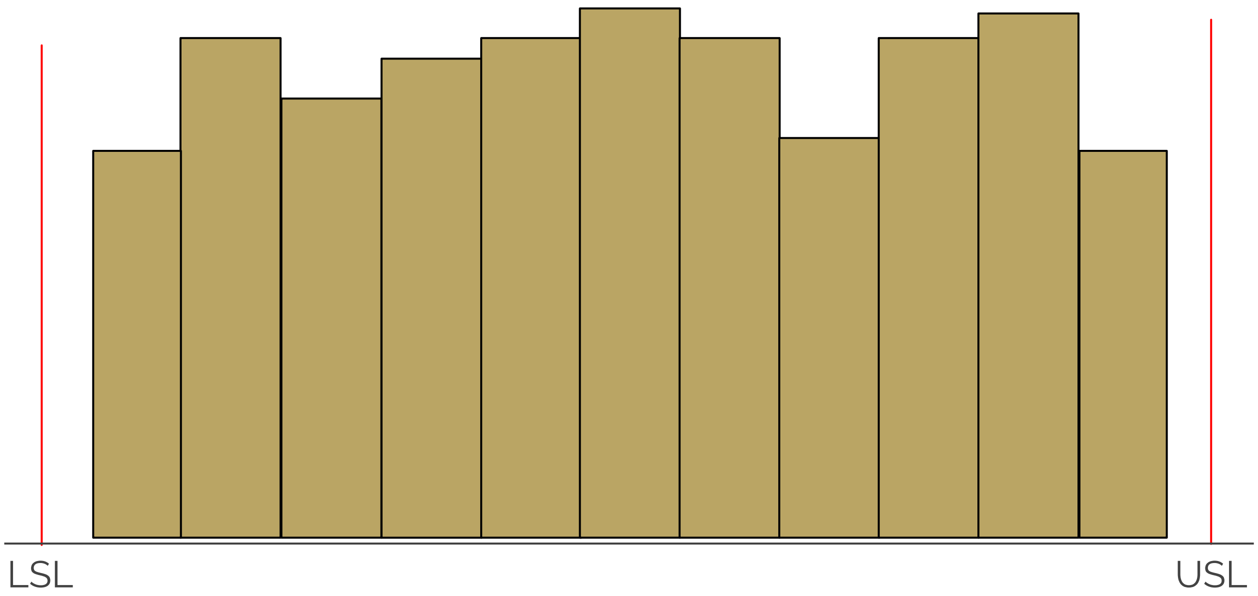
Pattern 2



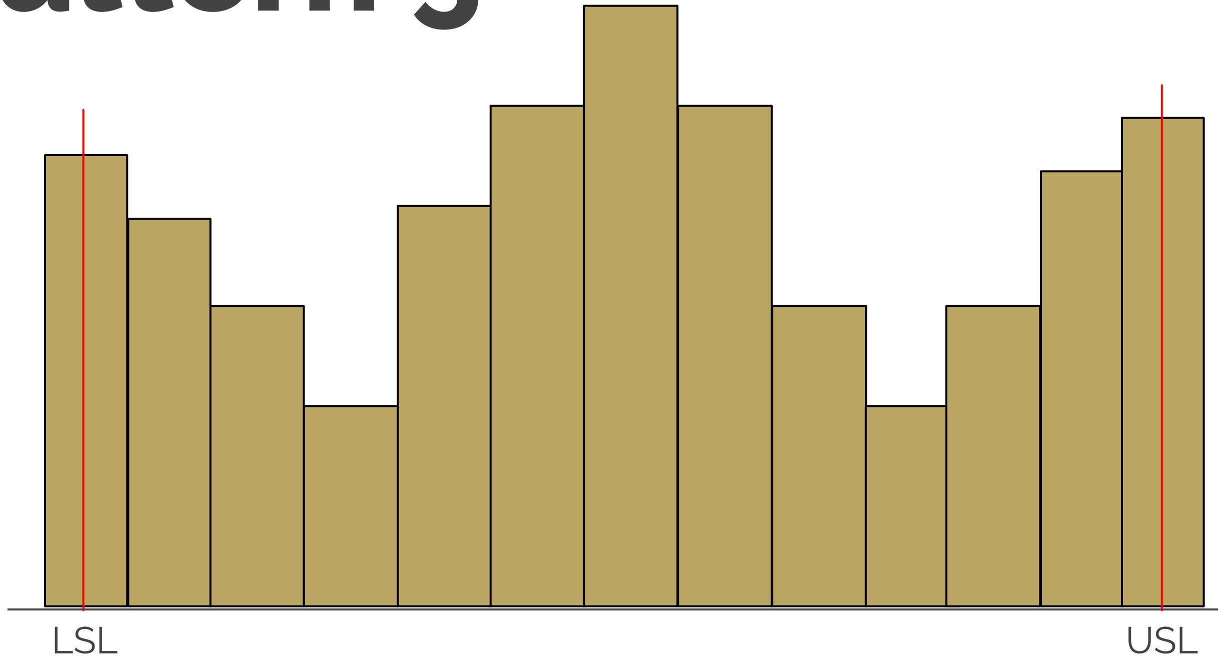
Pattern 3



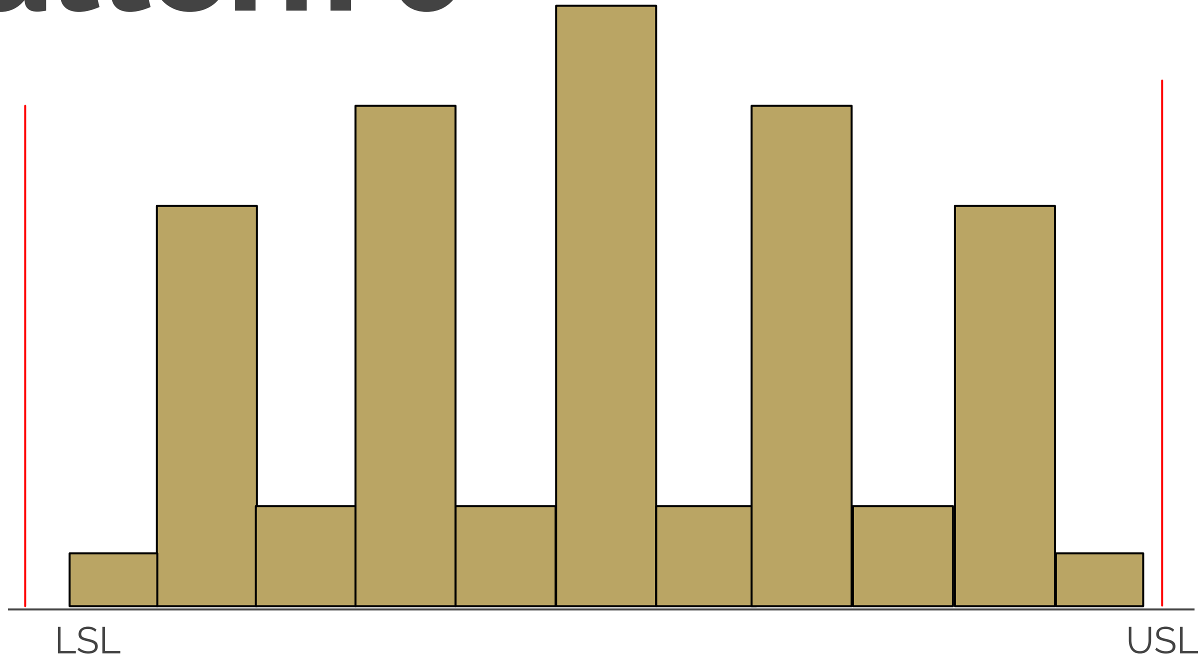
Pattern 4



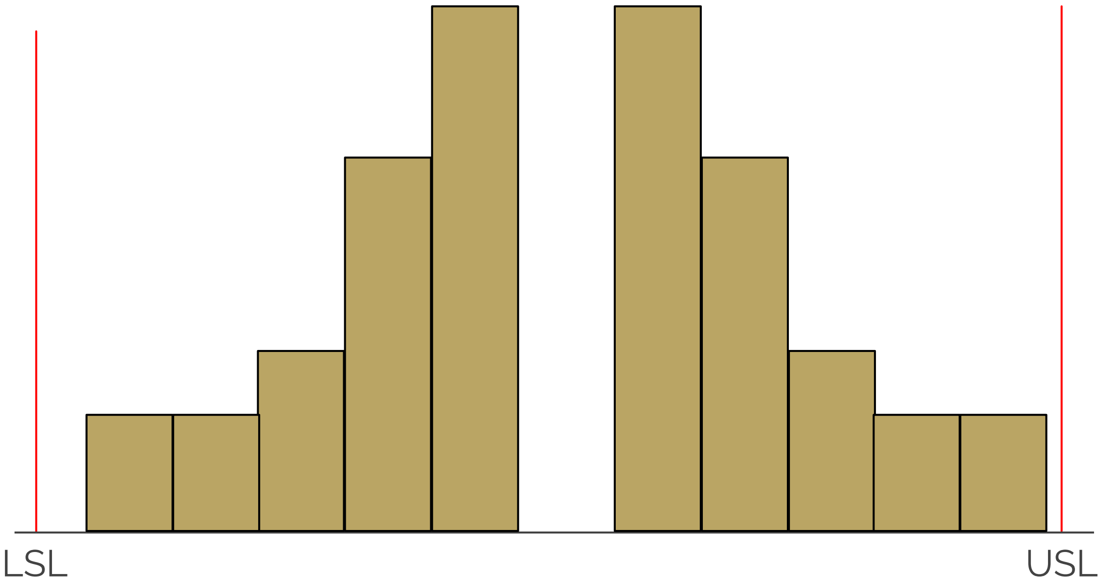
Pattern 5



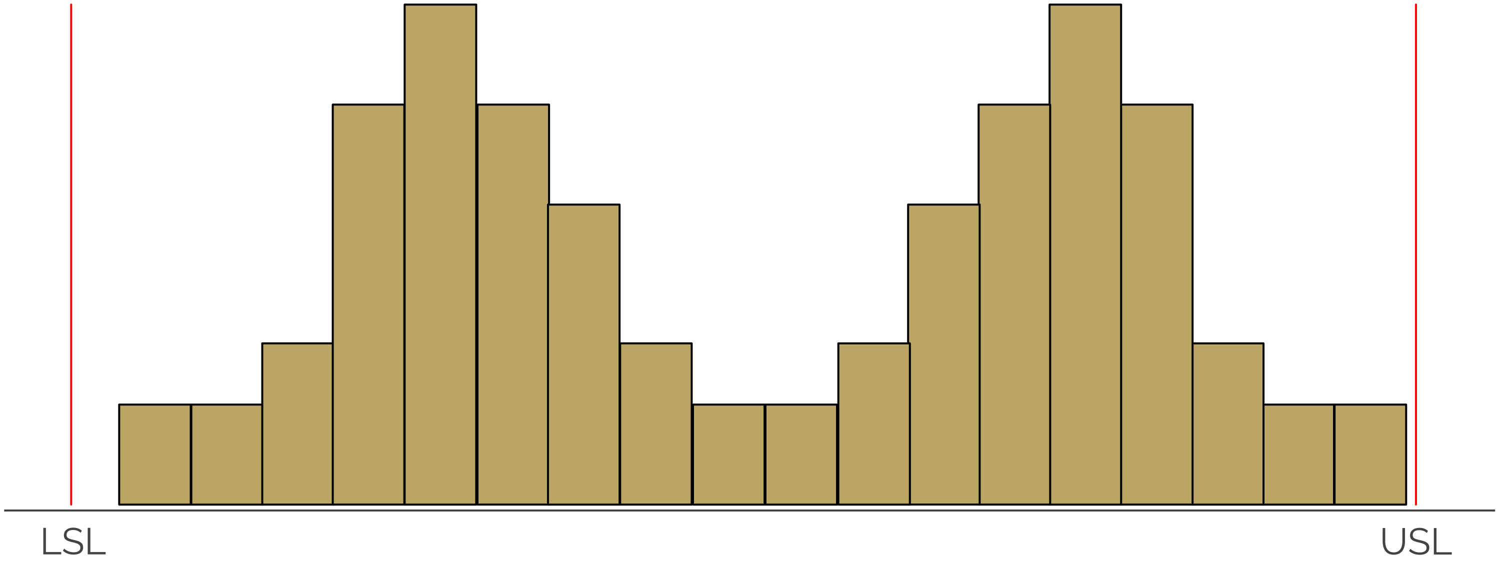
Pattern 6



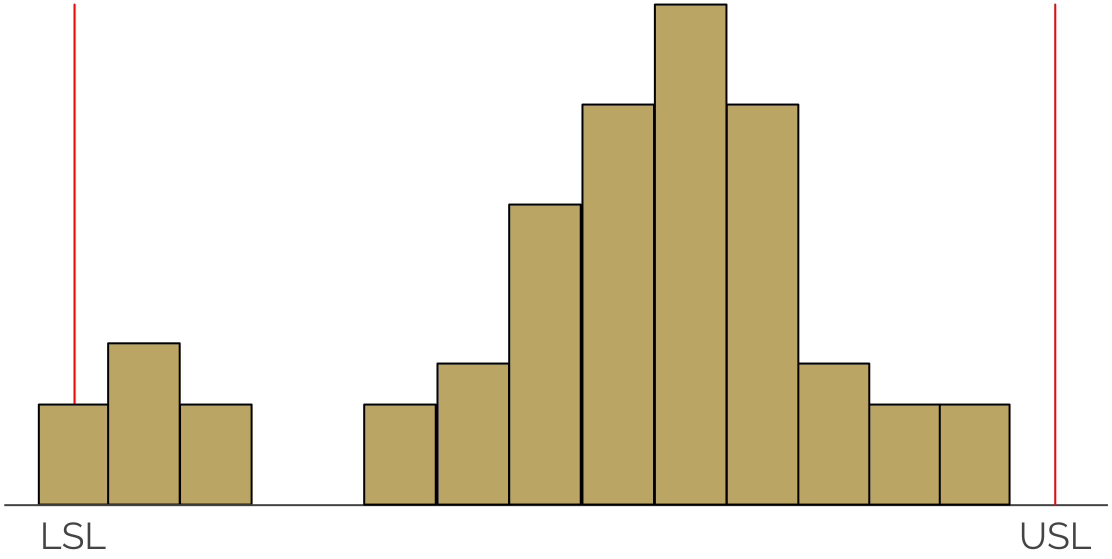
Pattern 7



Pattern 8



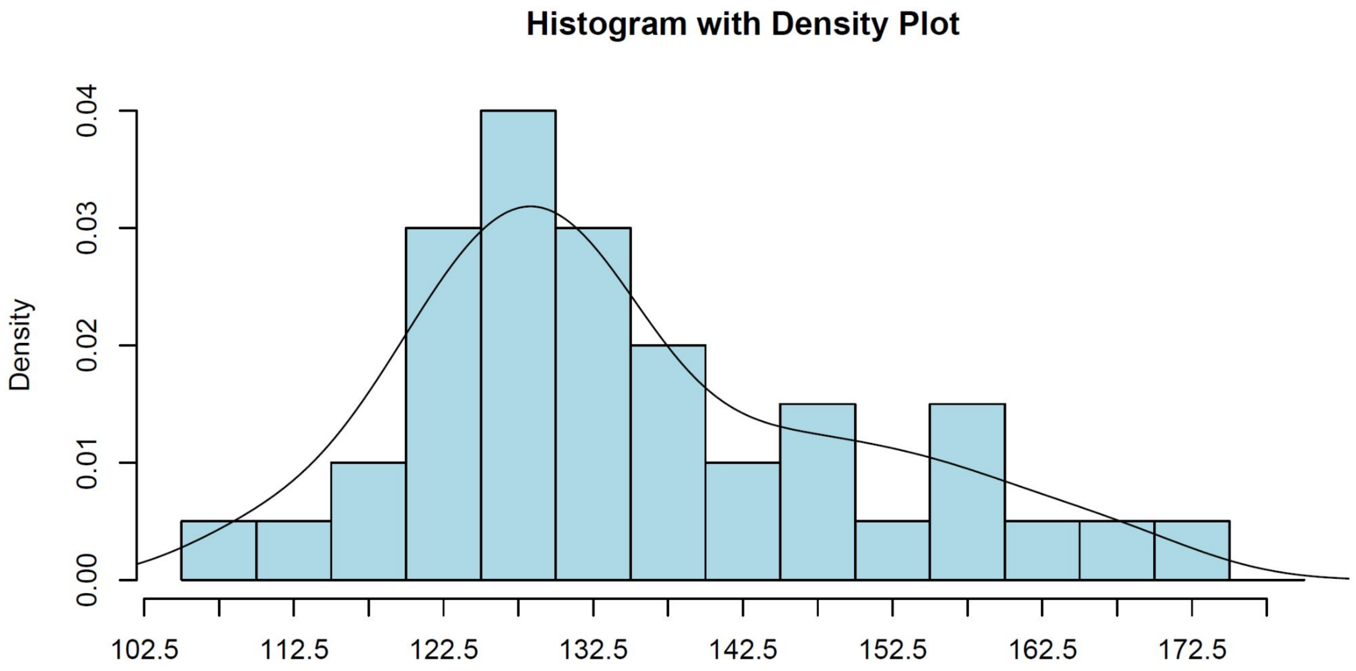
Pattern 9



Density Plot

Presenting data to portray the nature of a distribution

Density Plot



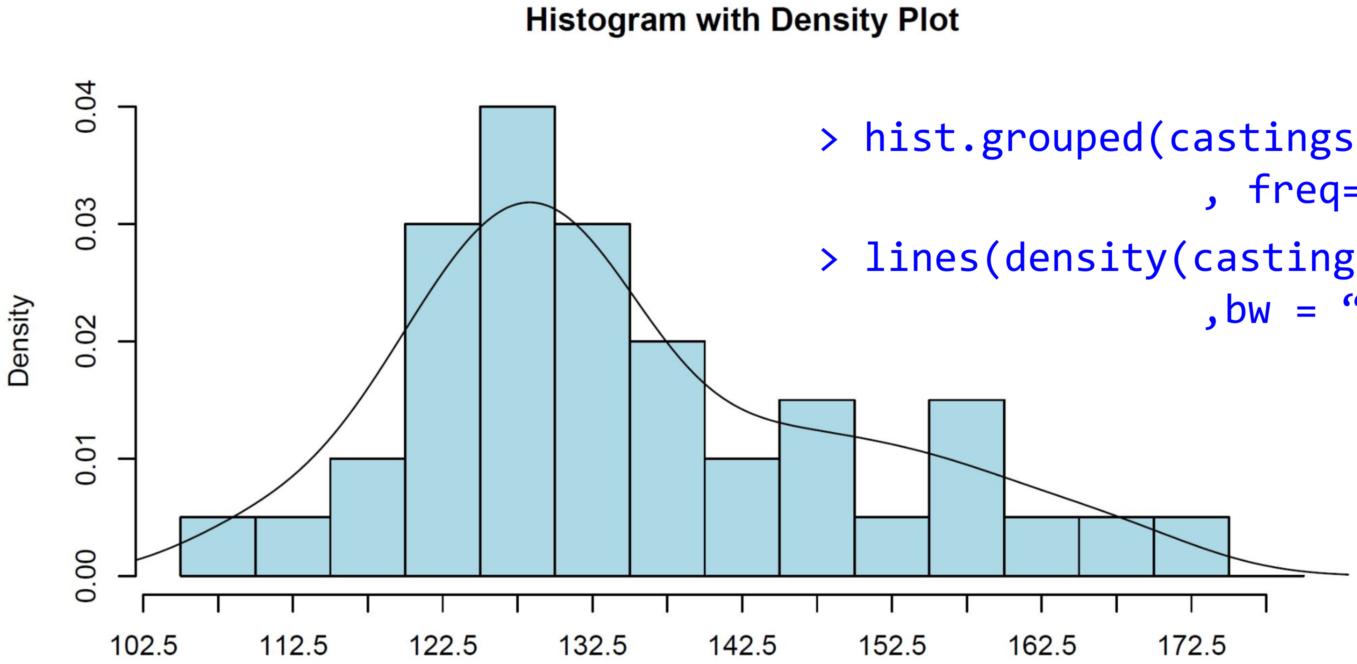
Density Plot

- Used with continuous data to visualize an underlying probability distribution
- When the data are continuous, we can use a density plot over a histogram.

Create a Density Plot

In RStudio

Density Plot




```
> hist.grouped(castings$weight
                          , freq=F)
> lines(density(castings$weight
                          ,bw = “SJ”))
```

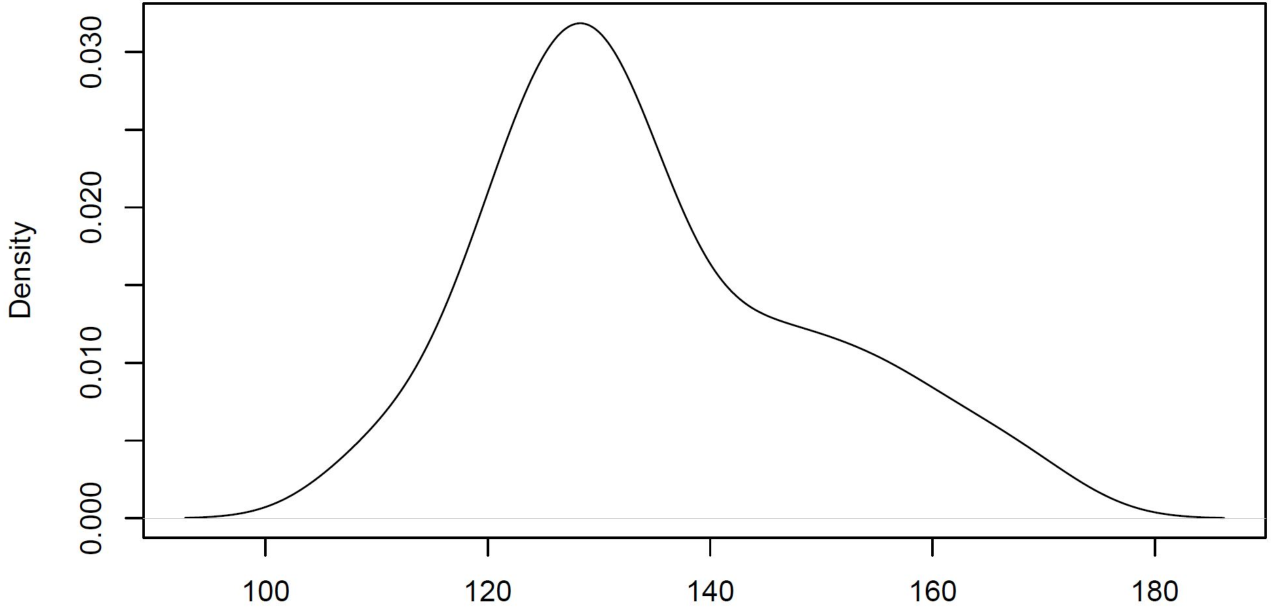
Density Plot

- The density plot can also be plotted without a histogram:

```
> plot(density(castings$weight))
```

Density Plot

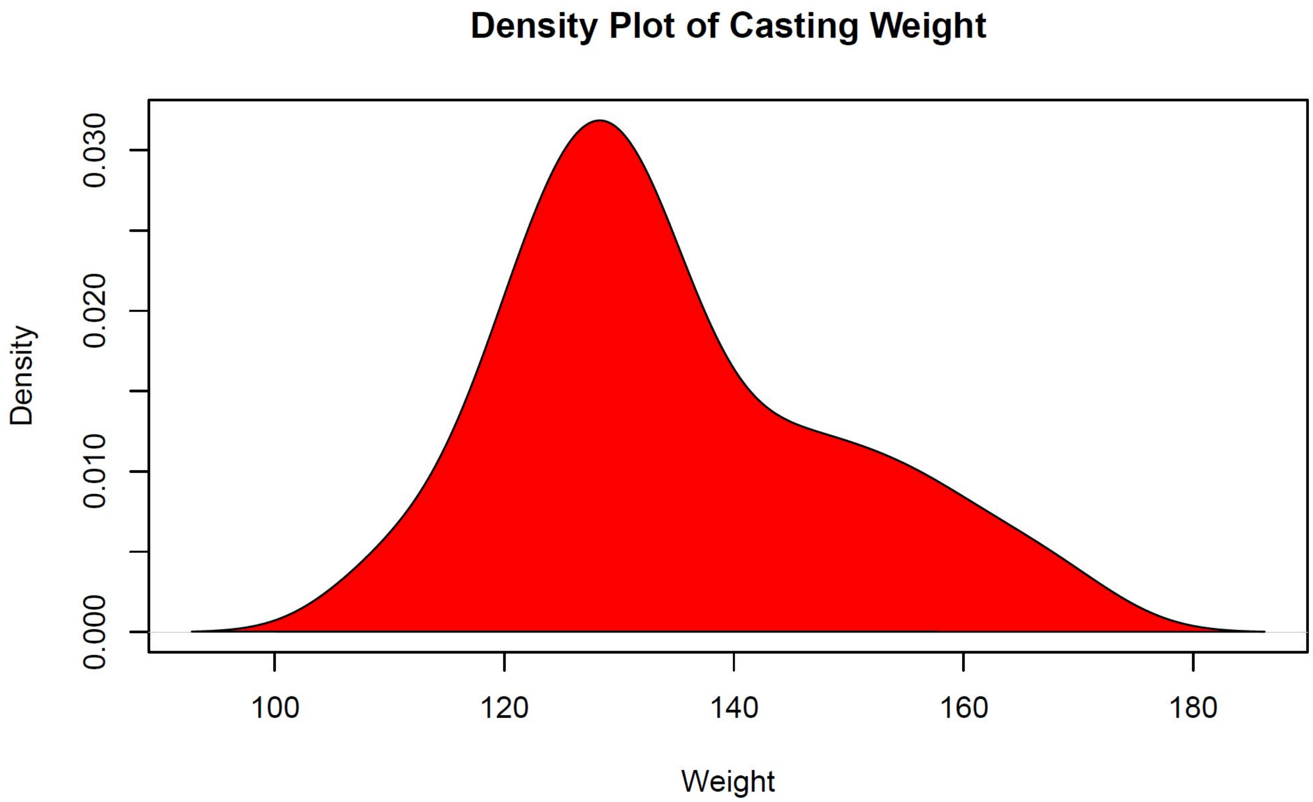
```
> plot(density(castings$weight)
```



Density Plot

To fill a density plot with color:

```
> dp<-density(castings$weight)
> plot(dp, main="Density Plot of Casting
        Weight", xlab="Weight")
> polygon(dp, col="red", border="black")
```


Density Plot of Casting Weight
Density
0.000
0.010
0.020
0.030
100
120
140
160
180
Weight

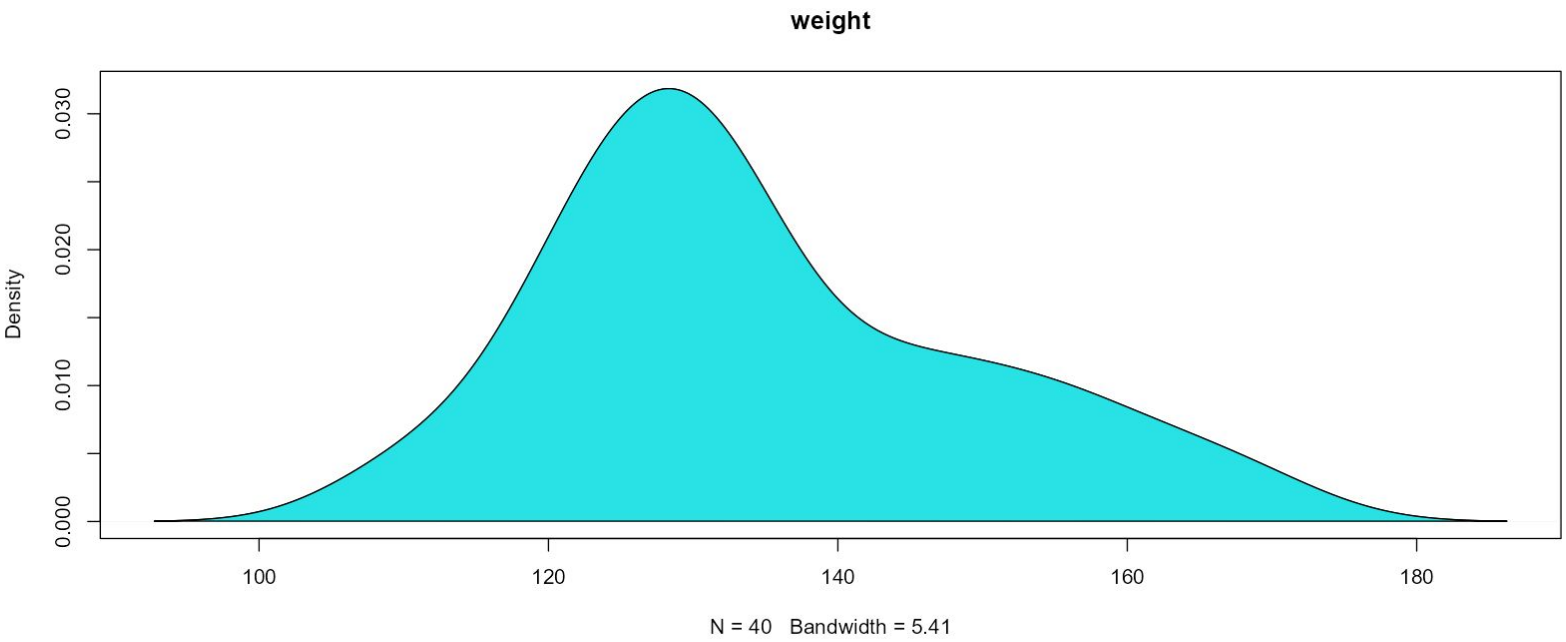
Create a Density Plot

In ROIStat

Density Plot

1. File > Import Data
2. EDA > Data Setup > Analyze Columns
3. Select ‘weight’
4. Click on Histograms tab
5. Select Kernel Density

Density Plot



Box and Whisker Plots

Presenting data to portray the nature of a distribution

Box and Whisker Plots

- Used to display data corresponding to percentiles, and typically from two or more sources or process streams, simultaneously

Box and Whisker Plots

Advantages

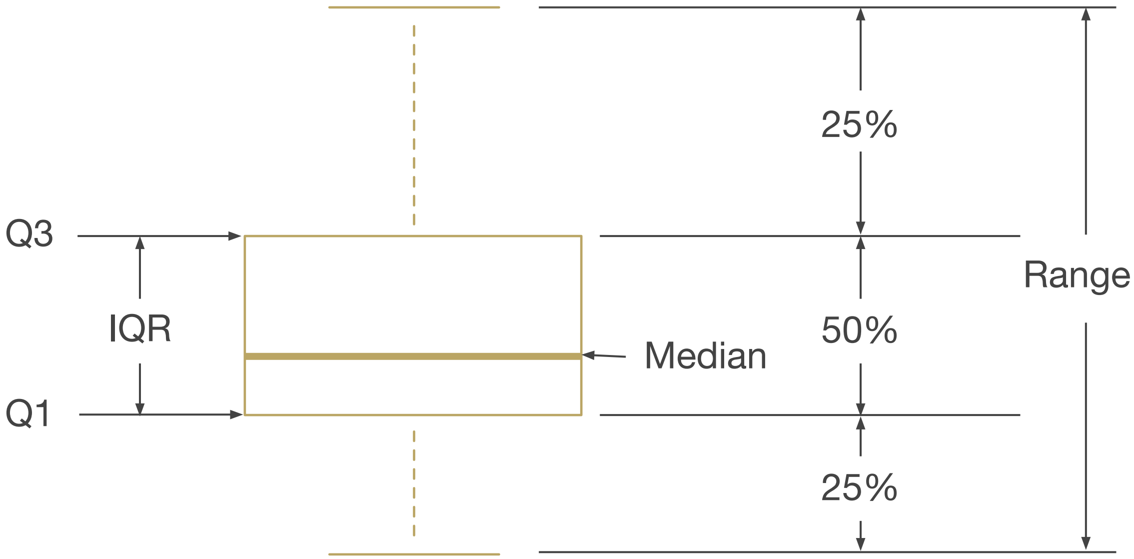
- Two sample data sets do not have to possess the same shape, but are directly comparable nonetheless
- Can display outliers, which may represent Special Causes of Variation

5 Number Summary

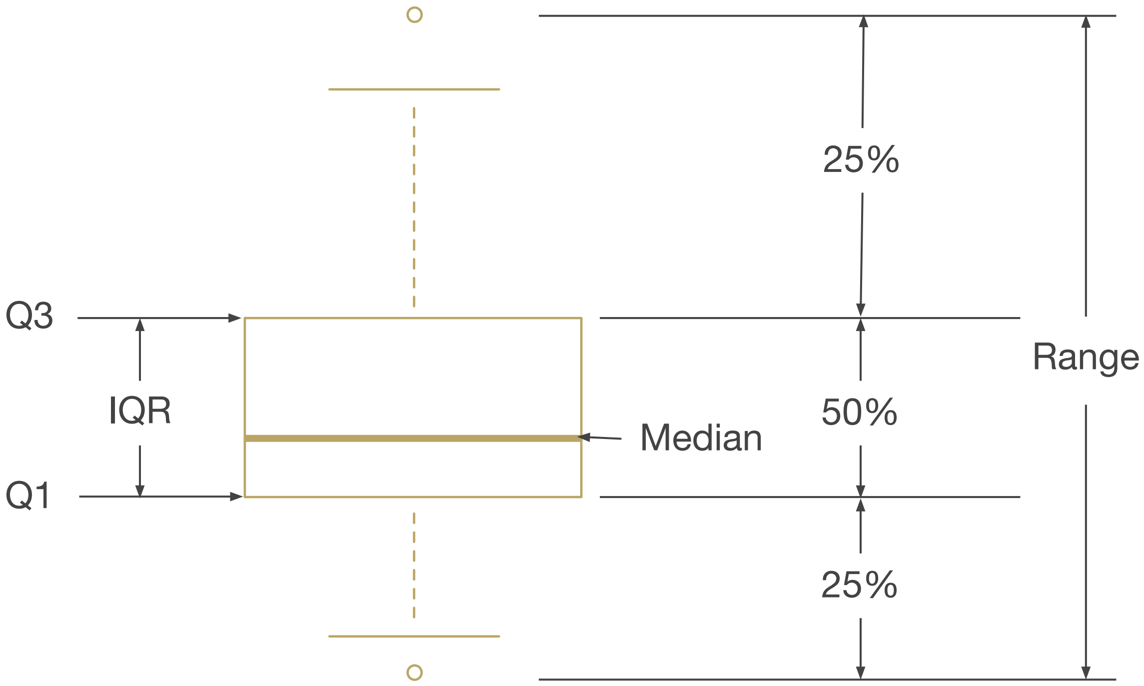
- Maximum
- Q3 (3rd Quartile)
- Median (Q2) (2nd Quartile)
- Q1 (1st Quartile)
- Minimum

```
> summary(castings$weight)
```

Box and Whisker Plot



With Outliers

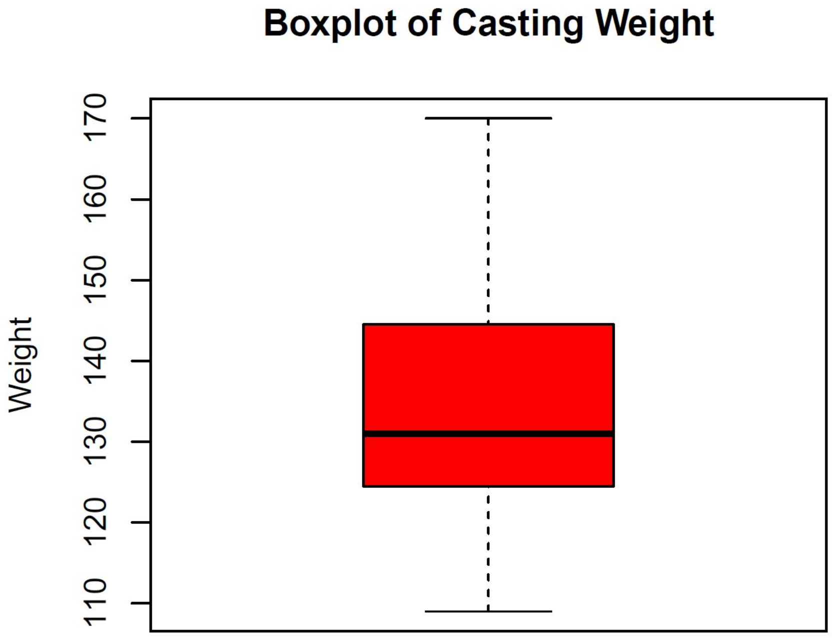


Create a
Box and Whisker Plot

In RStudio

Box and Whisker Plot

```
> boxplot(castings$weight)
```



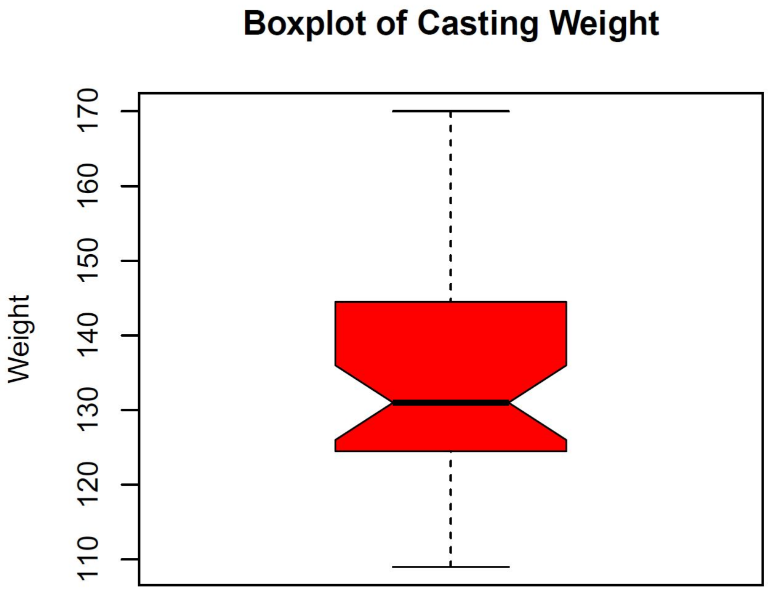
Notched Box and Whisker Plot

- A notched Box and Whisker plot shows the 95% confidence interval of the median.

```
> boxplot(castings$weight, notch=T)
```

Box and Whisker Plot

```
> boxplot(castings$weight,
          notch=T)
```



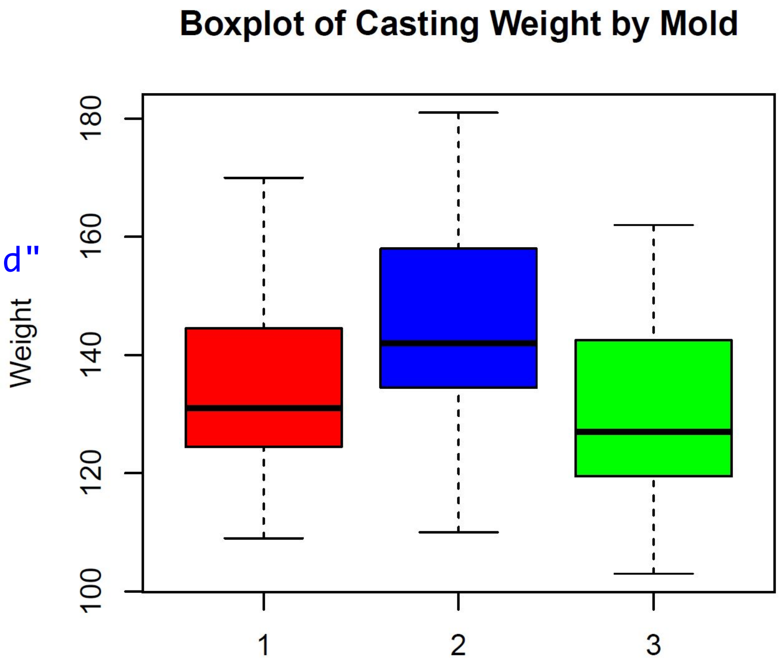
Boxplot To Compare Groups

```
> boxplot(y ~ x, data = data.frame)

> boxplot(weight ~ mold, data = castings3)
```

Boxplot To Compare Groups

```
boxplot(weight ~ mold
      , data = castings3
      , main="Boxplot of Casting Weight by Mold"
      , ylab="Weight"
      , col = c("red","blue","green"))
```



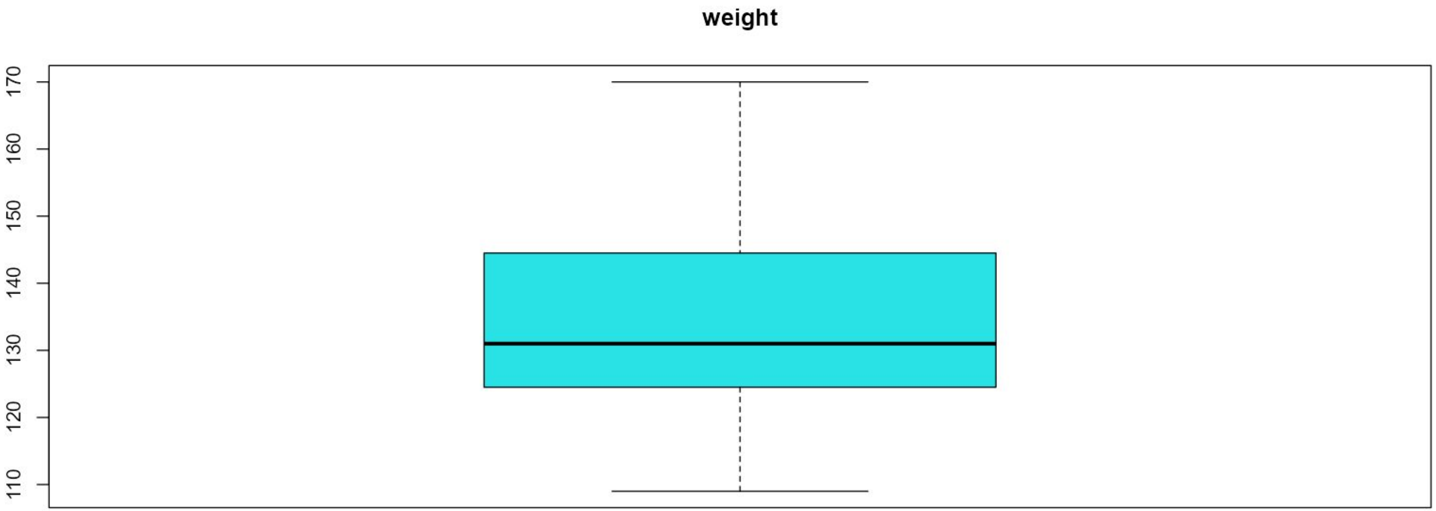
Create a Box and Whisker Plot

In ROIStat

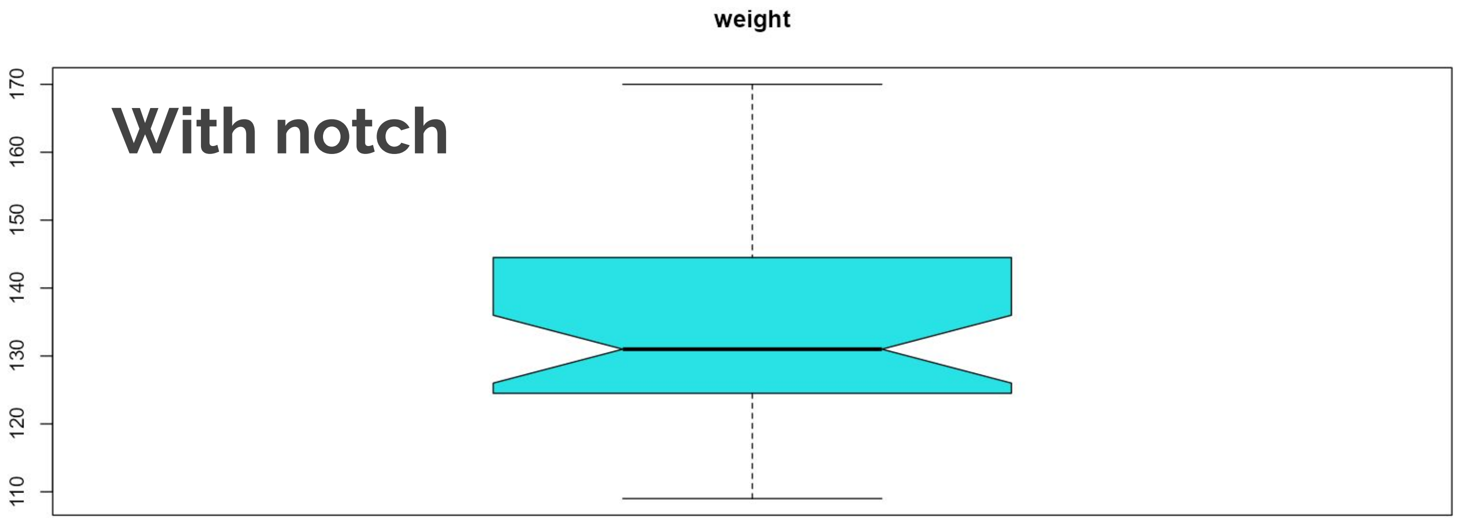
Box and Whisker Plot

1. File > Import Data
2. EDA > Data Setup > Analyze Columns
3. Select ‘weight’
4. Click on Boxplots tab
5. For Notched, select ‘Use notch?’

Box and Whisker Plot



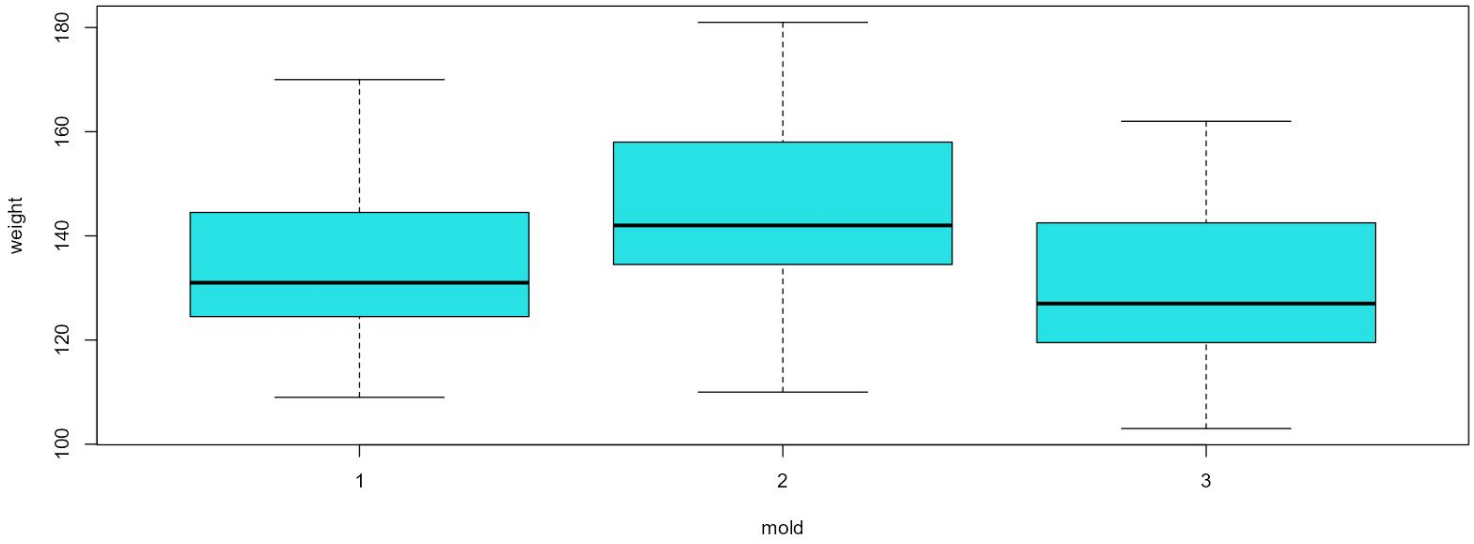
Box and Whisker Plot



Boxplot To Compare Groups

1. File > Import Data
2. EDA > Data Setup > Analyze by Factors
3. Select factor 'mold' (click to right)
4. Select data 'weight' (click to right)
5. Click on Boxplots tab

Boxplot To Compare Groups



Scatter Plots

Presenting data to portray the nature of a relationship between two variables

Scatter Plot

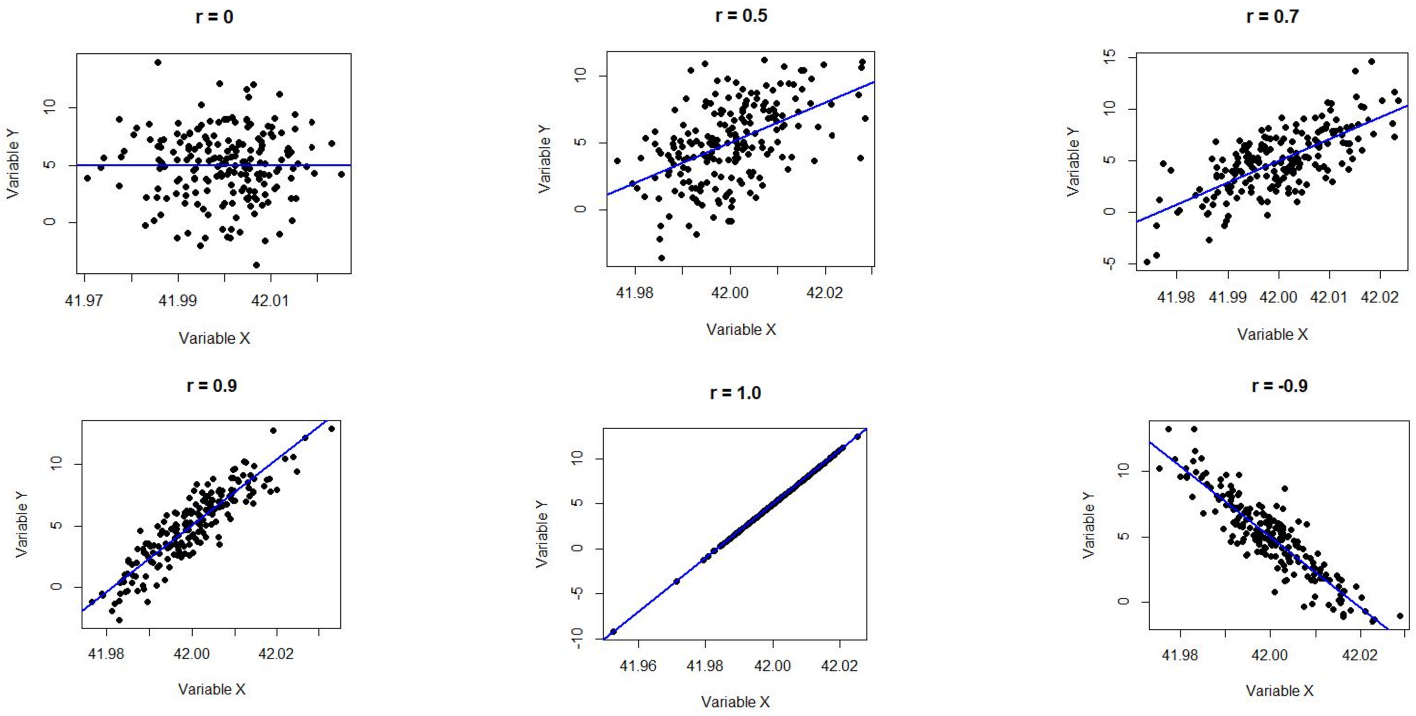
- Shows the relationship between two variables
- Shape of points tell whether relationship is:
  - Positive
  - Negative
  - Special
  - No Relationship

Scatter Plot

Strength can be "seen" on a scatter plot

- 0 : no relationship, circular shape
- ~ 0.4 - 0.5 : shape of a "football"
- ~ 0.7 - 0.8 : shape of a Zeppelin
- ~ 0.9 : shape of a cigar
- 1.0 : perfect line

Scatter Plots



Scatter Plot

- IMPORTANT - When a scatter plot shows a correlation between two variables, there is not necessarily a cause and effect relationship.

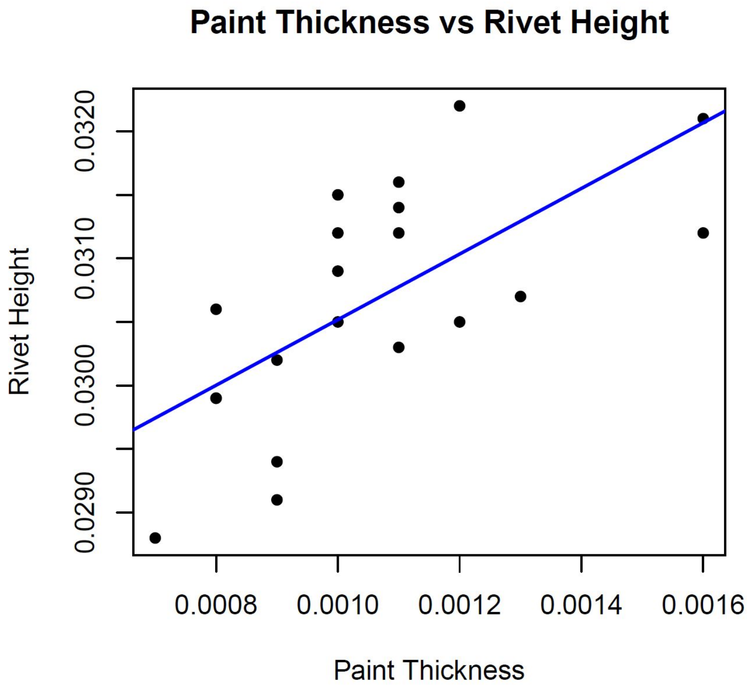
Create a Scatter Plot

In RStudio

Scatter Plot

```
plot(x = Rivet$paint
    , y = Rivet$rivet
    , xlab = "Paint Thickness"
    , ylab = "Rivet Height"
    , pch = 19, cex = 0.8
    , main = "Paint Thickness vs
      Rivet Height")

abline(lm(Rivet$rivet~Rivet$paint)
      ,col="blue", lwd=2)
```



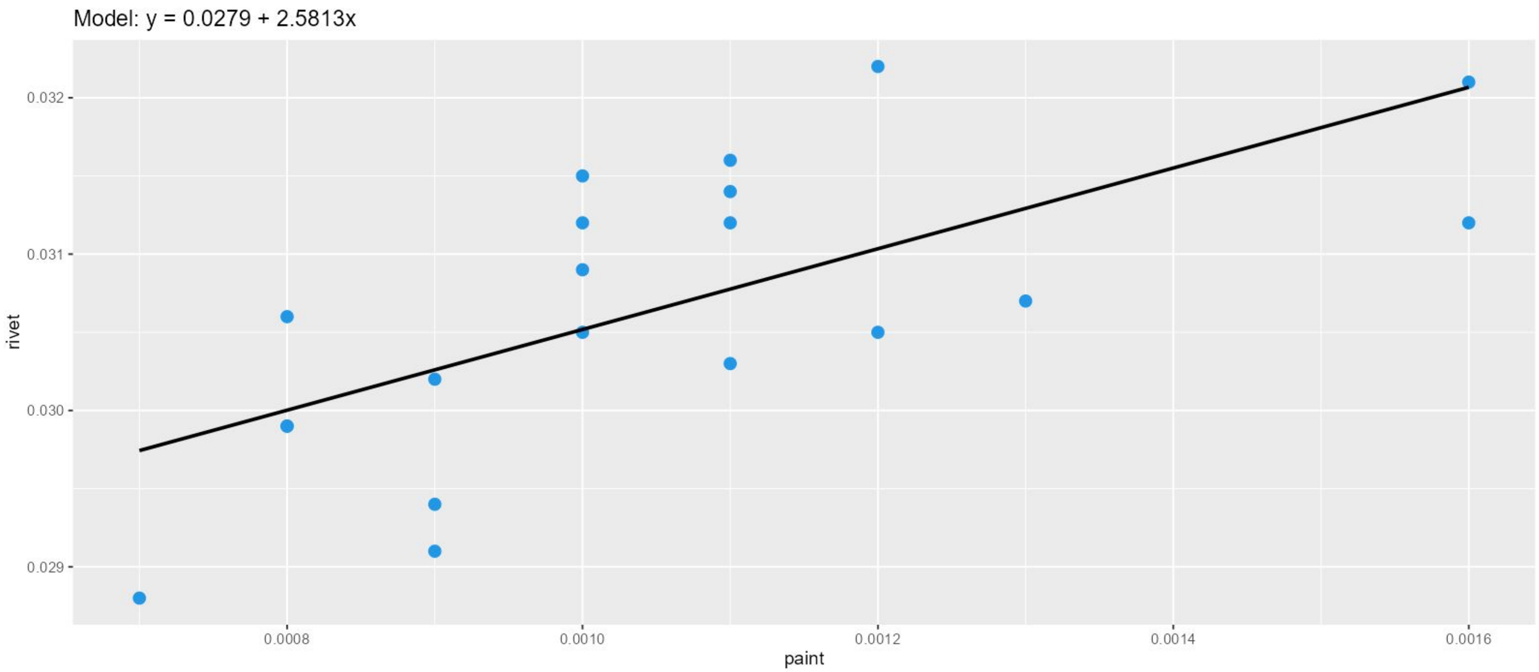
Create a Scatter Plot

In ROIStat

Scatter Plot

1. File > Import Data
2. Correlation and Association > Scatterplot
3. Select x as ‘paint’
4. Select y as ‘rivet’
5. Fit model as ‘Linear’

Scatter Plot



Describing Data Numerically

Learning Objectives

- Calculate measures of central tendency for a dataset in RStudio and ROIStat
- Calculate measures of dispersion in RStudio and ROIStat

Learning Objectives

- Discriminate between skewness and kurtosis
- Calculate the sample skewness and kurtosis
- Discriminate between correlation & association
- Calculate correlation for two continuous variables

5 Aspects of Data

- Location or Central Tendency
- Spread or Dispersion (Variability)
- Shape
- Time Sequence
- Relationship

Sample Data

- Preforms for a compression molding process were randomly sampled
- Sample size (n) is 10
- Each Preform was then weighed on a gram scale

Sample Data

- Suppose the resultant data appeared as:

  65 67 36 37 36 57 53 39 38 58

- We will use this sample data set to demonstrate the calculation of various statistics

Create the Data File

Create a Vector

```
weight <-c(65,67,36,37,36,57,53,39,38,58)
```

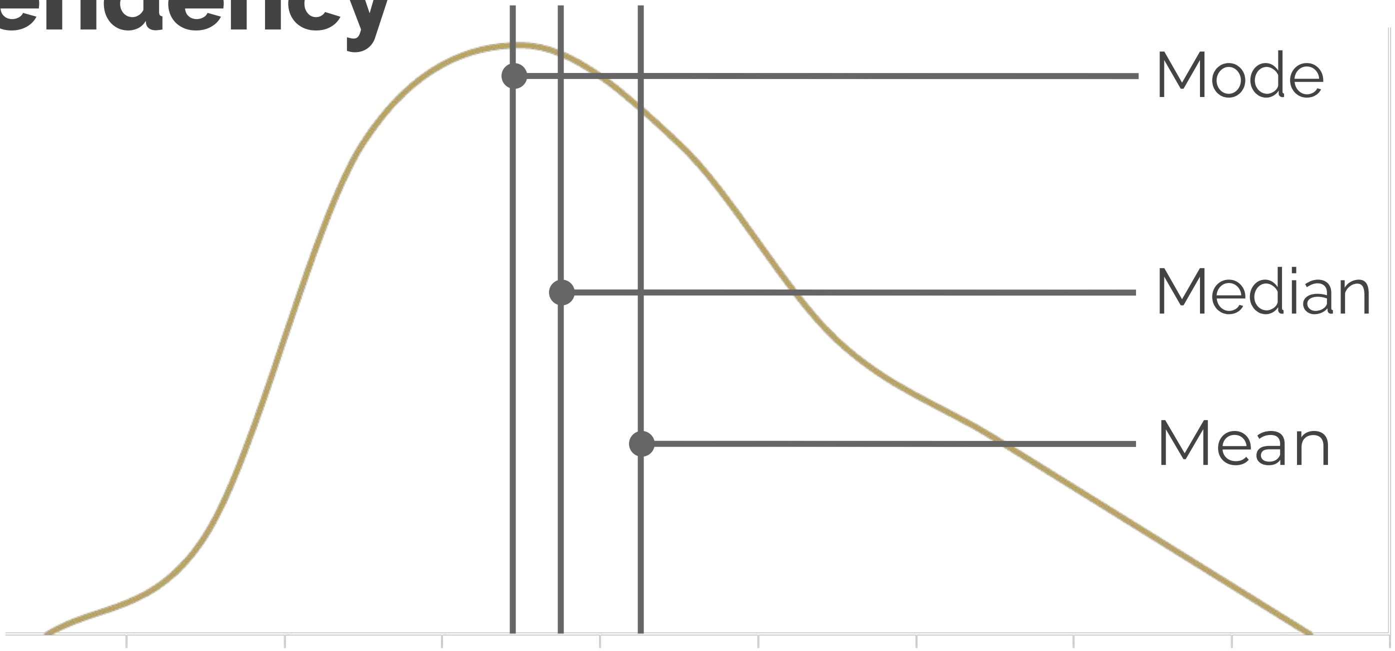
Store the Variable in a data frame

```
preform <- data.frame(weight)
View(preform)
```

Measures of Central Tendency

Describing the middle or central point of a distribution

Measures of Central Tendency



The Mean

$$\overline{X} = \frac{\sum X}{n}$$

- Arithmetic average
- Can be thought of as the "center of gravity" of the frequency distribution
- The value in which the sum of all deviations from this value are zero
- Symbols: population (μ) and sample ($\overline{X}$)

Advantages and Disadvantages

- Advantages
  - Easy to understand
  - Simple to calculate
  - Every data set possesses an arithmetic mean
- Disadvantages
  - Affected by extreme measures or values

Mean: Calculations

For our ungrouped preform data set, the calculation for the mean is as follows:

$$\overline{X} = \frac{\sum X}{n} = \frac{486}{10} = 48.6$$

The Median

- The median is the value at or below which 50% of the data fall, or at or above which 50% of the data fall
- The median is a measure of position and is the middle value in a sorted array of data
- Symbols: population (M) and sample ($\tilde{X}$)

Median: Example

For our ungrouped preform data set:

- First, the data set is sorted from low to high
- We note the median may be found in the (n + 1)/2th position, or (10 + 1)/2 = 5.5 position

36 36 37 38 39 53 57 58 65 67

- The median is found as the average of the 5th and 6th value, or 39 and 53
- The median is 46

Advantages and Disadvantages

Advantages

- Easy to understand
- Not affected by extreme values

Disadvantages

- The median does not take the relative magnitude of the values into account

The Mode

- The mode is the most frequently occurring value in a data set
- For a population, the mode is the peak of the population distribution curve
- Symbols: population (M_o) and sample (X_{mode})

Mode: Example

- For our preform data set (sorted)

  36 36 37 38 39 53 57 58 65 67

- The mode is 36

Advantages and Disadvantages

Advantages

- Not affected by extreme values
- Can be used with categorical data

Advantages and Disadvantages

Disadvantages

- The data set may not have a modal value. For example, it is possible that no two values are alike
- The data set may contain too many modal values to be useful

How To Calculate Central Tendency

In RStudio

In RStudio

```
mean(preform$weight)
median(preform$weight)
sample.mode(preform$weight)

summary.continuous( )
summary.impl( )
```

How To Calculate Central Tendency

In ROIStat

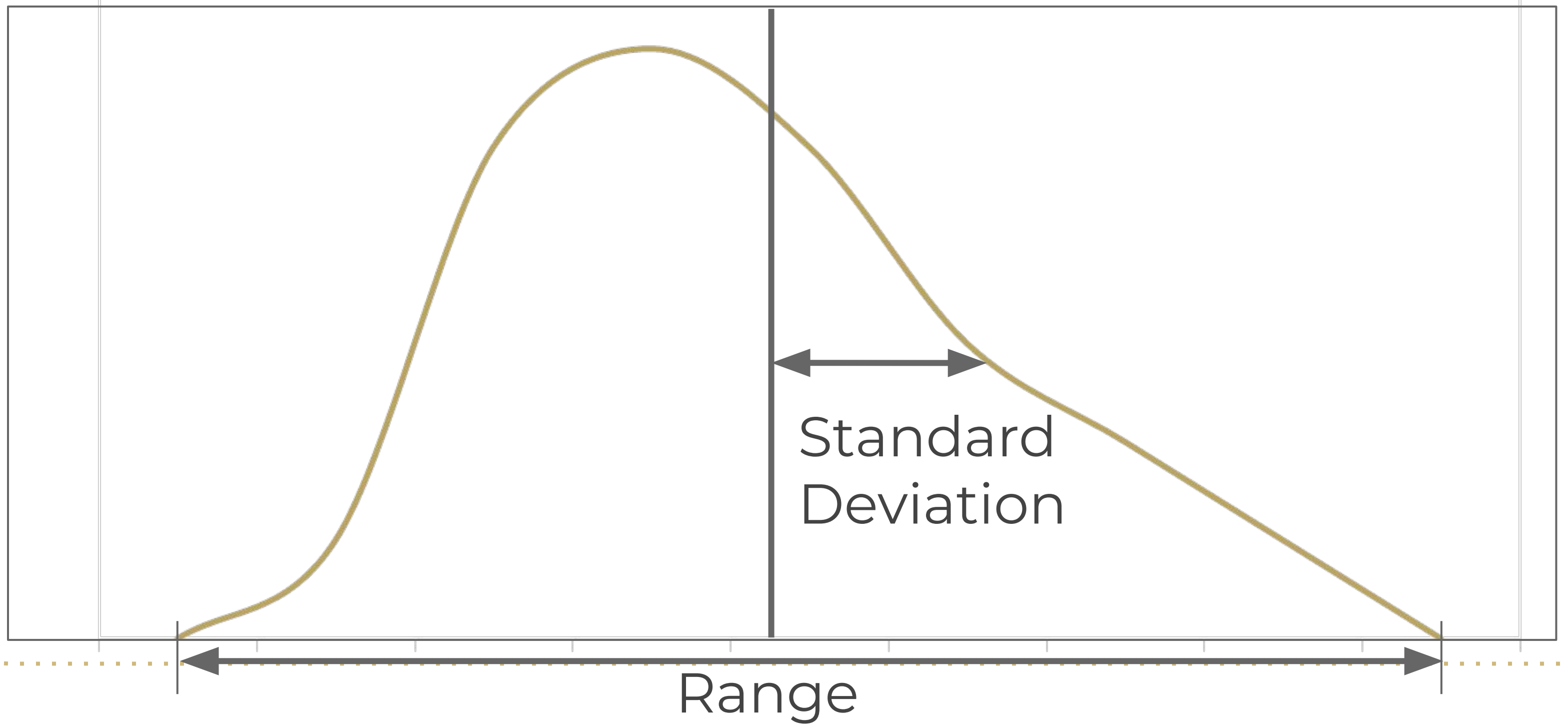
In ROIStat

- File > Import Data
- EDA > Data Setup > Analyze Columns
- Select ‘weight’
- Click on Descriptives tab
- Select statistics of interest

Measures of Dispersion

Describing the extent to which the data are scattered or distributed

Measures of Spread



Measures of Dispersion

- Measures of dispersion reflect the variation or spread in a data set or distribution
- Some of the common measures of dispersion are:
  - Range
  - Standard Deviation
  - Variance

The Range

- The range is the difference between the highest and lowest value in a data set
- Symbols:
    - Population (generally does not exist)
    - Sample (R)

The Range

- Calculations: $R = X_H - X_L$
- Example:
  - For our sample data set, the low is 36 and the high is 67
  - The range is: $R = 67 - 36 = 31$

Advantages and Disadvantages

Advantages

- Depends on only two values - Maximum minus minimum
- Easy to understand

Disadvantages

- Extremely sensitive to “outliers”

The Standard Deviation

- The standard deviation is a measure of variation that includes all data values in its calculation

- The standard deviation is the square-root of the average squared distance values fall from the mean

Standard Deviation: Calculations

For a sample

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}}$$

Standard Deviation: Example

For our sample data set, with a mean of 48.6

$$s = \sqrt{\frac{\sum(X - \overline{X})^2}{n - 1}} = \sqrt{\frac{\sum(X - 48.6)^2}{9}} = \sqrt{\frac{1442.40}{9}} = 12.66$$

Standard Deviation: Example 2

65 67 36 37 36 57 53 39 38 58

- Calculate the mean: 48.6
- Calculate deviations from the mean for each value
- 16.4 18.4 -12.6 -11.6 -12.6 8.4 4.4 -9.6 -10.6 9.4

Standard Deviation: Example 2

- Square each deviation

269.96	338.56	158.76	134.56	158.76
70.56	19.36	92.16	112.36	88.36

- Sum the squared deviations: 1442.40
- Divide the sum of the squared deviations by (n – 1) and then take the square root of this value
- s = 12.66

The Variance

- The variance is the square of the standard deviation
- The variance is the average squared distance values fall from the mean
- Symbols: Population (σ^2) and Sample (s^2)

Variance: Calculation

For a sample

$$s^2 = \frac{\sum(X - \overline{X})^2}{n - 1}$$

Variance: Calculation

- For our sample preform data set, in which the standard deviation is 12.6596 (using four decimal places), the variance is:
- $s^2 = (12.6596)^2 = 160.27$

How To Calculate Dispersion

In RStudio

In RStudio

```
range(preform$weight)
sd(preform$weight)
var(preform$weight)

summary.continuous( )
summary.impl( )
```

How To Calculate Dispersion

In ROIStat

In ROIStat

- File > Import Data
- EDA > Data Setup > Analyze Columns
- Select ‘weight’
- Click on Descriptives tab
- Select statistics of interest

Measures of Shape

Describing the symmetry, peak and tails of a distribution

Measures of Shape

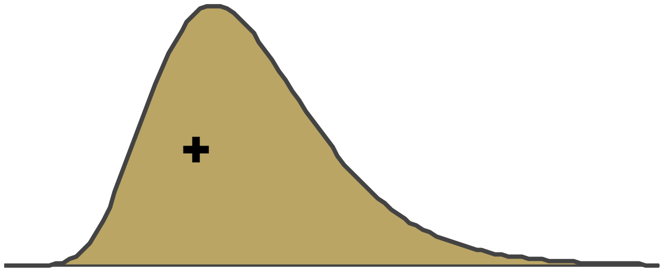
- Measures of shape reflect the type of distribution sampled
- Skewness is concerned with the symmetrical nature of the distribution
- Kurtosis is concerned with the tails as compared to the peak of the distribution

Skewness

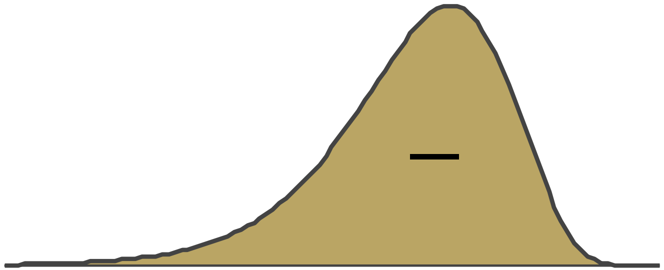
- Skewness is the degree of departure from symmetry of a distribution
- Symbols
  - Population (γ_3) and
  - Sample (g_3)

Skewness

- Basically measures "lopsidedness."
- Symmetric distributions have zero skewness.



Positively skewed



Negatively skewed

Skewness: Calculations

- The most important group of measures of skewness and kurtosis use the third and fourth moments about the mean

- Moments about the mean are the average of the deviations from the mean raised to some power

Skewness: Calculations

The r^{th} moment about the mean is:

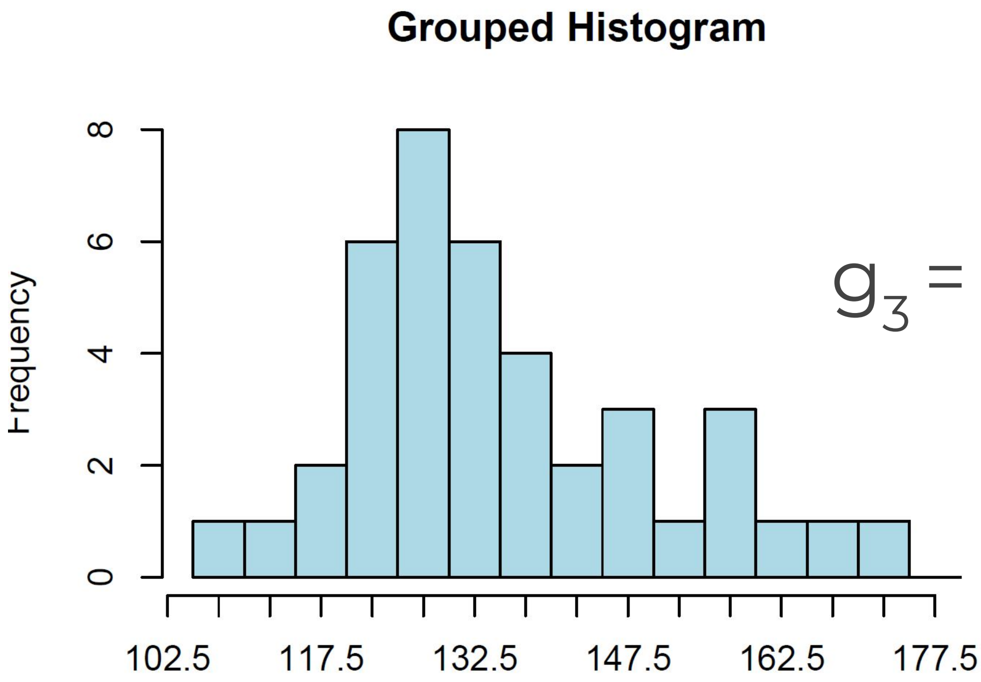
$$m_r = \frac{\sum (X - \overline{X})^r}{n}$$

Skewness: Calculations

- A measure of skewness may then be calculated as follows
- The sign displays the direction of skewness

$$g_3 = \left[\frac{\sqrt{n(n-1))}}{n-2} * \frac{m_3}{m_2^{3/2}} \right]$$

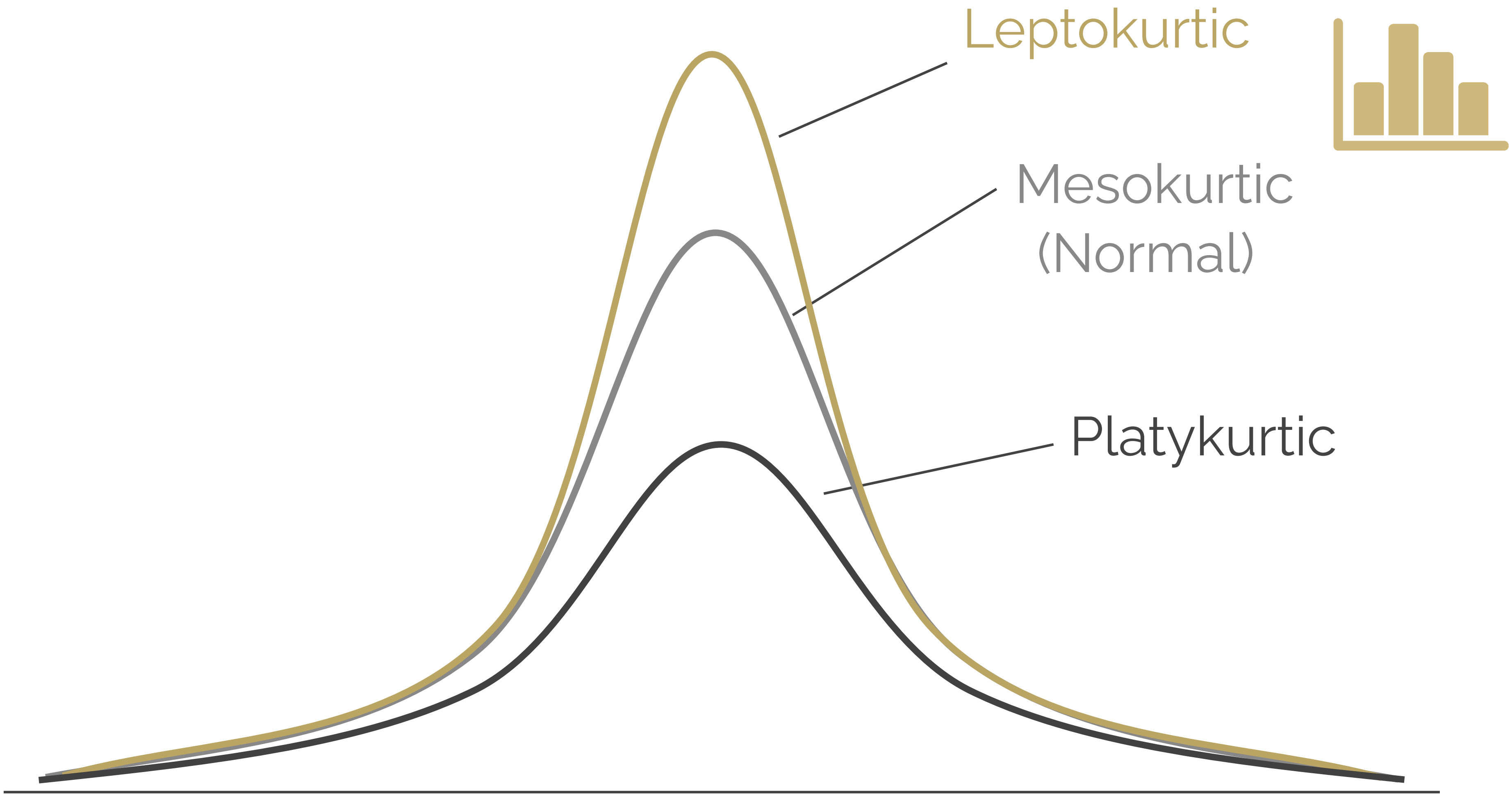
Skewed Distributions



$g_3 = 0.643$

Kurtosis

- Kurtosis is the degree of “tailedness” of a distribution
- An intermediate distribution, with zero kurtosis, is known as a mesokurtic distribution
- A symmetrical leptokurtic distribution has a higher peak and has heavier tails, and has positive kurtosis
- A symmetrical platykurtic distribution has a lower peak and lighter tails, and has negative kurtosis

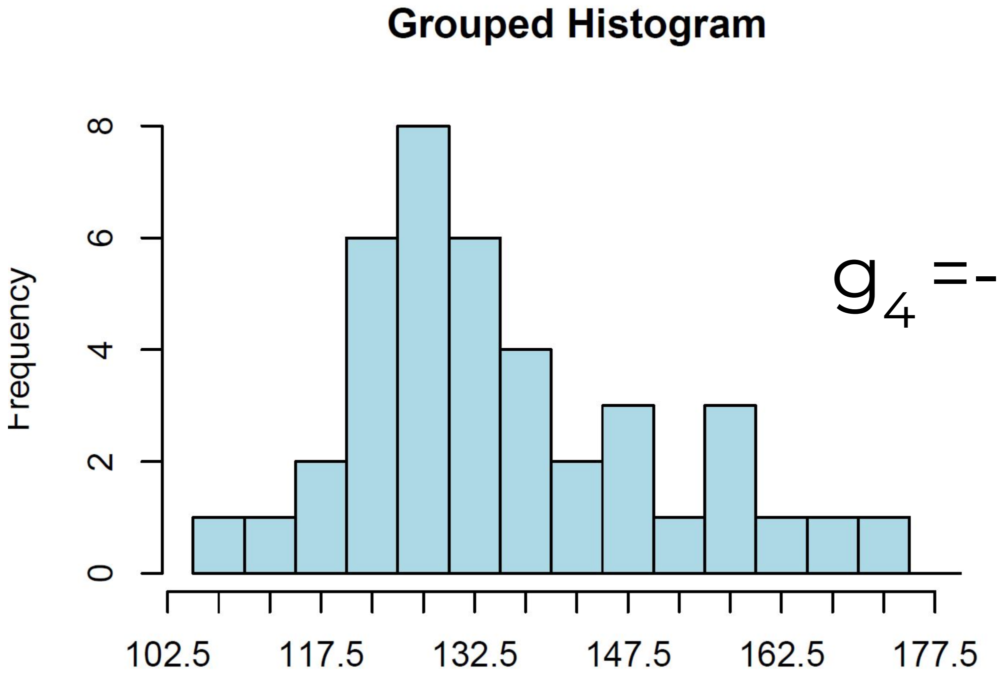

Leptokurtic
Mesokurtic
(Normal)
Platykurtic

Kurtosis: Calculations

- Symbols
  - Population (γ_4) and
  - Sample (g_4)

$$g_4 = \left[\frac{(n-1)(n+1)}{(n-2)(n-3)}\right] * \frac{m_4}{m_2^2} - 3\left[\frac{(n-1)^2}{(n-2)(n-3)}\right]$$

Kurtotic Distributions



$g_4 = -0.169$

How To Calculate Skewness and Kurtosis

In RStudio

In RStudio

```
lolcat::skewness(preform$weight)
lolcat::kurtosis(preform$weight)

summary.continuous( )
summary.impl( )
```

How To Calculate Skewness and Kurtosis

In ROIStat

In ROIStat

- File > Import Data
- EDA > Data Setup > Analyze Columns
- Select ‘weight’
- Click on Descriptives tab
- Select statistics of interest

Measures of Relationship

Describing the strength of a relationship between two variables

Measures of Relationship

- Correlation and association are measures of the strength of a relationship between two variables

Measures of Relationship

- Before we calculate statistics related to relationship, we must first properly classify each variable.
    - Nominal
    - Ordinal
    - Continuous

Correlation vs Association

- Where both variables are continuous, the statistic employed to measure the relationship may be referred to as a Coefficient of Correlation

Correlation vs Association

- Where both variables are nominal, the statistic employed to measure the relationship may be referred to as a Coefficient of Association

Correlation and Association

- Coefficients of Correlation and Association can vary given all possible combinations of nominal, ordinal, and continuous data that can occur

Coefficient of Correlation

- The most frequently used coefficient of correlation used is the Pearson Product-Moment Coefficient of Correlation.
- Symbols
  - Population: ρ_{xy}
  - Sample: r_{xy}

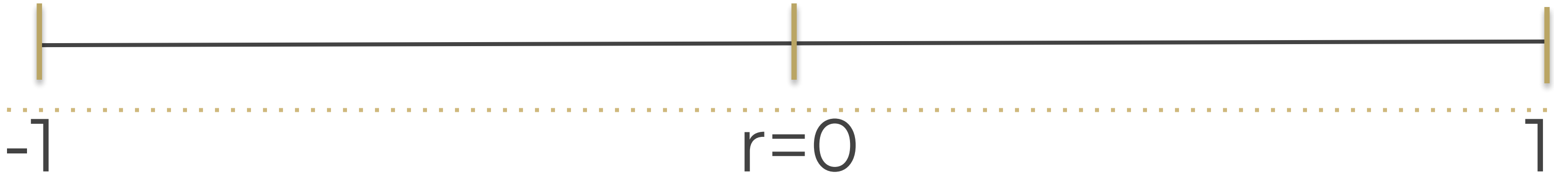
Product Moment Coefficient

- Two components:
    - Sign (+ or -)
    - Numeric Value

$$r_{xy} = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \sum(Y - \overline{Y})^2}}$$

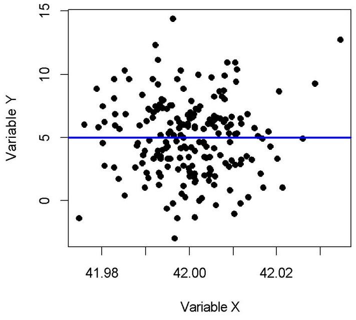
Product Moment Coefficient

- Sign (+ or -) gives the direction of the relationship
  - Positive: As one variable increases in magnitude, the other variable increases
  - Negative: As one variable increases in magnitude, the other variable decreases

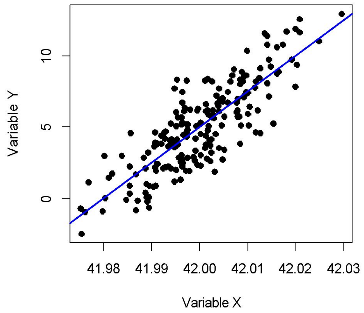


Scatterplot Examples

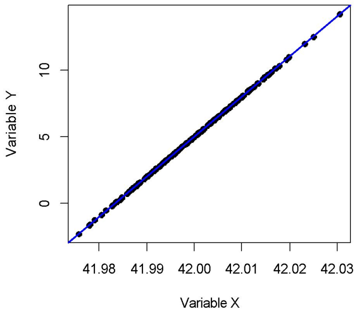
r = 0.00



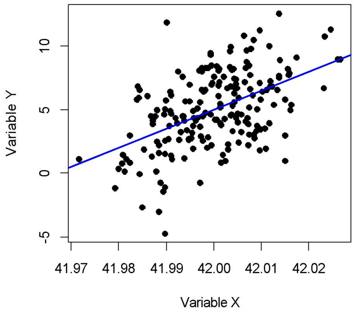
r = 0.83



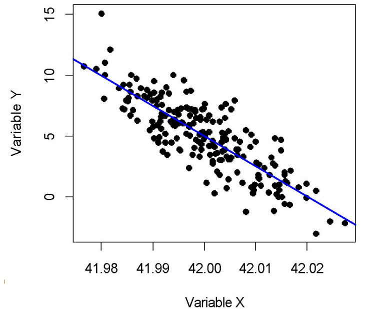
r = +1.00



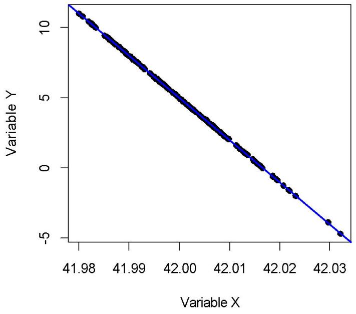
r = 0.50



r = - 0.83



r = -1.00



How To Calculate Correlation

In RStudio

In RStudio

```
cor(Rivet)

cor(x = Rivet$paint, y = Rivet$rivet)
```

How To Calculate Correlation

In ROIStat

In ROIStat

- File > Import Data
- Correlation and Association > Use Data
- Select x as ‘paint’
- Select y as ‘rivet’

Parameters and Statistics

	Sample	Population
Definitions	Subgroup or portion of the population chosen for evaluation or study	Collection of all items produced or considered
Characteristics	Statistics	Parameters
Size	n	N
Mean	$\bar{X}$	μ
Median	$\tilde{X}$	M
Standard Deviation	s	σ
Variance	s^2	σ^2
Skewness	g_3	γ_3
Kurtosis	g_4	γ_4
Proportion	p	π
Rate	$\bar{c}$	λ