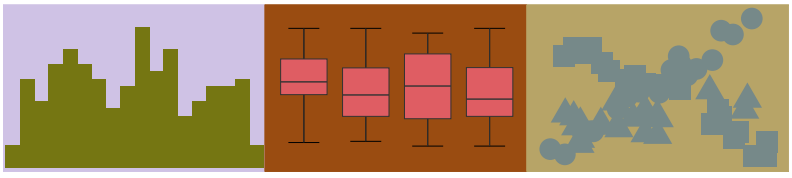
Downloadable Interpreting Descriptive and Exploratory Outputs: Plots and Summaries.

Rose Megirian

Table of contents



(a) Histograms (a) Box Plots (a) Scatter Plots

Descriptive and exploratory analysis summarises and describes the characteristics of a dataset, providing insight into distributions, central tendency, spread, and relationships between

variables. It serves two closely related purposes: as a standalone way of characterising and communicating what the data shows, and as an essential first step before formal modelling or hypothesis testing. Used in this exploratory capacity, descriptive analysis helps reveal the shape and structure of your data, identify potential issues, and inform decisions about how best to proceed with formal analysis.

This guide covers the most common outputs from descriptive and exploratory analysis that you are likely to encounter, explaining what they represent, what different patterns indicate, and what types of data each is appropriate for. It is organised into two sections: plots and figures, which covers common visualisations used to explore and describe data, and numerical summaries, which covers the key statistics used to characterise and communicate the properties of a dataset.

Plots and Figures

Visualisations are one of the most powerful tools in data analysis, making patterns, distributions, and relationships visible in ways that raw numbers alone often cannot. Plots like histograms, boxplots, and scatter plots are particularly valuable during exploratory data analysis, helping to reveal the shape and spread of data, identify outliers, and assess relationships between variables. This guide covers the most common descriptive plot types you are likely to encounter, explaining what they represent, what different patterns indicate, and what types of data each is appropriate for.

Histograms

```
//| echo: false

viewof dataset = Inputs.select(
  [
    "Normal Distribution", "Skewed Distribution", "Bimodal Distribution",
    "Uniform Distribution", "Exponential Distribution"
```

```
  ],
  {label: "Distribution", value: "Normal Distribution"}
)

viewof bin_width = Inputs.range([1, 30], {
  label: "Bin width",
  step: 1,
  value: 6
})
```

```
//| echo: false

html`<div style="margin-top: 2em;"></div>`
```

```
//| echo: false

all_data = transpose(hist_data)

filtered = all_data.filter(d => d.dataset === dataset)

annotation = ({
  "Normal Distribution": `Values cluster symmetrically around a central point.
    This bell-shaped curve is the most commonly assumed distribution in statistical
    methods. Try adjusting bin width to see how it affects the apparent shape.`,

  "Skewed Distribution": `This distribution has a long tail to the right, meaning most
    values are low but a few are very high. This is called right skew, or positive skew.
    Skewed data can violate the normality assumptions of some statistical methods.`,

  "Bimodal Distribution": `Two distinct peaks suggest two subgroups within the data.
    This is called a bimodal distribution. Seeing two humps in a histogram is often
    a signal to investigate whether the data should be analysed as separate groups.`,

  "Uniform Distribution": `Values are spread roughly evenly across the range with no
    clear peak. This uniform distribution is relatively rare in practice but useful
    for understanding how histograms represent shape.`,

  "Exponential Distribution": `A sharp peak near zero with a long right tail. Common
    in data representing waiting times or survival times. Heavily skewed data like
    this often benefits from a log transformation before analysis.`
```

```
})[dataset] || ""

Plot.plot({
  style: {background: "transparent", fontFamily: "inherit"},
  marginBottom: 50,
  x: {label: "Value"},
  y: {label: "Frequency"},
  marks: [
    Plot.rectY(
      filtered,
      Plot.binX(
        {y: "count"},
        {
          x: d => +d.value,
          thresholds: d3.range(
            d3.min(filtered, d => +d.value),
            d3.max(filtered, d => +d.value) + bin_width,
            bin_width
          ),
          fill: "#F2F2F2",
          stroke: "#414042",
          strokeWidth: 0.8
        }
      )
    ),
    Plot.ruleY([0])
  ]
})
```

```
//| echo: false

html`<div style="font-size: 0.9em; color: #555; margin-top: .5em;
  padding: 0.75em 1em; background: #f9f9f9; border-left: 3px solid #414042;">
  ${annotation}
</div>`
```

A histogram displays the distribution of a continuous variable by grouping values into intervals known as bins, and showing how many observations fall within each bin as a bar. The height of each bar represents the frequency of values in that range,

making it easy to see at a glance where values cluster, how spread out they are, and whether any unusual patterns are present.

A key feature of a histogram is that the bars touch. This reflects the fact that the underlying data is continuous, meaning there are no gaps between possible values and therefore no gaps between bins. This distinguishes a histogram from a bar chart, which is used for discrete or categorical data where each bar represents a separate group rather than a range of a continuous scale. If your data consists of counts, ordinal scores, or named categories, a bar chart is the more appropriate choice.

Histograms are commonly used during exploratory data analysis to understand the distribution of a variable before any formal modelling takes place. They help reveal how spread out values are, whether they cluster around a central point, whether there are unusually high or low values, and whether the distribution is symmetrical or skewed. This kind of early understanding of your data is important for informing decisions about which analytical methods are likely to be appropriate.

Histograms are also widely used during and after model fitting to examine residuals, the differences between observed values and the values predicted by the model. Most common analytical methods require that residuals are approximately normally distributed, and plotting a histogram of residuals is a straightforward way to assess whether this holds. It is worth noting that it is the residuals rather than the raw data that need to follow an approximately normal distribution. Raw data can be skewed or non-normal and the residuals can still be approximately normal after a model is fitted. However, heavily non-normal raw data may be an early indication that this is worth checking carefully once a model has been fitted.

Reading histograms

Look at where the bars are tallest, as this is where most values fall. A symmetrical bell-shaped curve suggests an approximately normal distribution, where values cluster evenly around a central point with the mean, median and mode all close together

(Figure 4). A longer tail on one side indicates skewness. A right-skewed distribution has a tail extending to the right, meaning most values are lower with fewer very high values pulling the mean upward. A left-skewed distribution has a tail extending to the left, meaning most values are higher with fewer very low values pulling the mean downward (Figure 5). Two distinct peaks suggest the data may contain two subgroups with different underlying characteristics, known as a bimodal distribution, which is worth investigating further before proceeding with analysis (Figure 6). Bars sitting far from the main cluster may indicate outliers worth examining.

Normal example

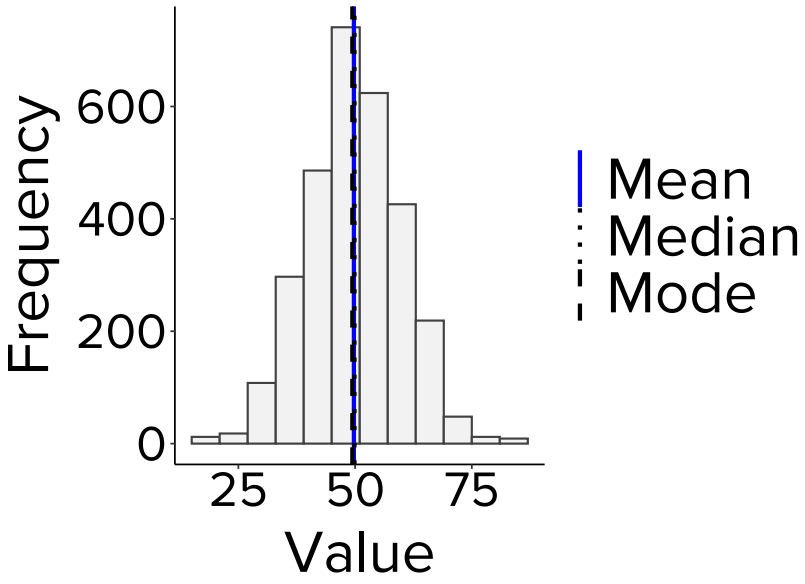


Figure 4: Histogram of an approximately normal distribution, showing the mean, median and mode clustering around the centre of a symmetrical bell-shaped curve. This is the most commonly assumed distribution shape underlying statistical methods, making it an important pattern to recognise in your data and residuals.

Skewed example

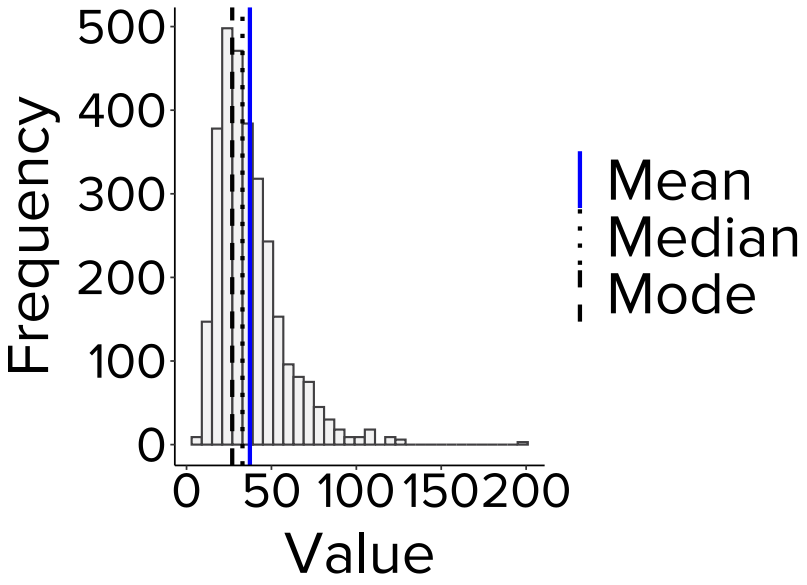


Figure 5: A histogram of a skewed distribution, where the longer tail pulls the mean away from the median and mode. This separation of the three measures is a key indicator of skewness and is important for determining which analytical methods are appropriate.

Bimodal example

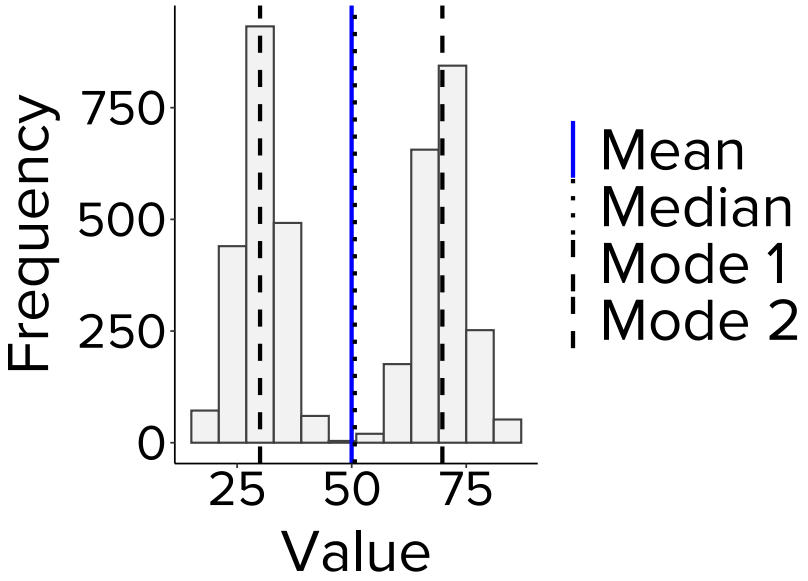


Figure 6: Histogram of a bimodal distribution showing two distinct peaks, suggesting the data may contain two subgroups with different underlying characteristics. Applying an analysis that assumes a single trend across all the data, such as comparing a single overall mean, could obscure these differences and lead to misleading conclusions.

Boxplots

A boxplot, or box-and-whisker plot, summarises the distribution of a continuous variable by displaying key statistics including the median, quartiles, and potential outliers in a compact visual form. Boxplots are particularly useful in exploratory data analysis for understanding the spread and shape of your data, and for comparing distributions across groups.

Because the boxplot is built around the median and quartiles rather than the mean, it is not strongly influenced by extreme values. A single very high or very low observation can pull the mean substantially away from the centre of the data, but has little effect on the median or the box, making boxplots a

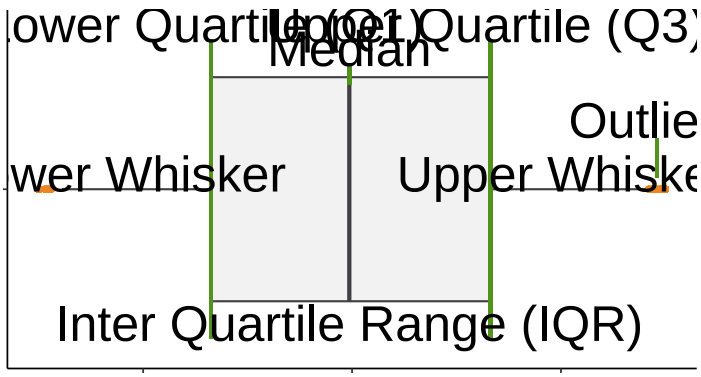


Figure 7: A boxplot provides a concise visual summary of a variable's distribution. The central box spans the interquartile range (IQR), from the first quartile (Q1) to the third quartile (Q3), representing the middle 50% of values. A line inside the box marks the median. The whiskers extend to the furthest values that fall within 1.5 times the IQR below Q1 and above Q3. values beyond this range are considered potential outliers and are plotted as individual points, indicating they may require further investigation. In this example, the relatively balanced spacing of the box and whiskers suggests that the data may follow a normal distribution.

reliable summary of your data's distribution even when outliers are present.

It is worth noting that the 1.5 times IQR rule used to define the whiskers and flag potential outliers is a convention rather than a formal statistical test. Points plotted beyond the whiskers are not necessarily problematic but are worth examining in the context of your data. It is also worth being aware that boxplots do not convey sample size, meaning a box based on ten observations looks the same as one based on ten thousand.

Reading boxplots

```
//| echo: false

viewof box_dataset = Inputs.select(
  [
    "Normal Distribution", "Skewed Distribution", "Bimodal Distribution",
    "Uniform Distribution", "Exponential Distribution"
  ],
  {label: "Distribution", value: "Normal Distribution"}
)
```

```
//| echo: false

html`<div style="margin-top: 2em;"></div>`
```

```
//| echo: false

box_all_data = transpose(box_data)

box_filtered = box_all_data.filter(d => d.dataset === box_dataset)

box_values = box_filtered.map(d => +d.value)

// Compute summary stats
box_sorted = [...box_values].sort((a, b) => a - b)
```

```
box_q1 = d3.quantile(box_sorted, 0.25)
box_q3 = d3.quantile(box_sorted, 0.75)
box_median = d3.quantile(box_sorted, 0.5)
box_iqr = box_q3 - box_q1
box_lower_fence = box_q1 - 1.5 * box_iqr
box_upper_fence = box_q3 + 1.5 * box_iqr
box_lower_whisker = d3.min(box_values.filter(d => d >= box_lower_fence))
box_upper_whisker = d3.max(box_values.filter(d => d <= box_upper_fence))
box_outliers = box_values.filter(d => d < box_lower_fence || d > box_upper_fence)

box_annotation = ({
  "Normal Distribution": `The median sits centrally within the box and the whiskers
    are roughly equal in length, indicating a roughly symmetric distribution with
    values spread evenly around the centre. This is consistent with a normal
    distribution. There are few or no outliers.`,

  "Skewed Distribution": `The median is pulled toward the left of the box and the
    right whisker is considerably longer than the left, indicating right skew. A
    small number of very high values extend the upper whisker and appear as outliers.
    Compare this with the histogram view - the long right tail is visible in both
    plot types but manifests differently.`,

  "Bimodal Distribution": `A bimodal distribution is one of the harder patterns to
    detect in a boxplot. The box and whiskers may appear relatively wide, reflecting
    the spread between two clusters, but the two peaks are not visible the way they
    are in a histogram. This is an important limitation of boxplots - they summarise
    distribution shape but can obscure underlying structure. A histogram or density
    plot is better for detecting bimodality.`,

  "Uniform Distribution": `The median sits near the centre of the box and both
    whiskers are long and roughly equal, reflecting the even spread of values across
    the full range. There are few outliers because no values fall unusually far from
    the rest - everything is equally spread.`,

  "Exponential Distribution": `The median sits close to the left edge of the box and
    the right whisker is much longer than the left, with many high-value outliers
    visible beyond the upper whisker. This strongly right-skewed pattern is typical
    of exponential data and is a signal that standard analytical methods assuming
    normality may not be appropriate without transformation.`
```

```
})[box_dataset] || ""

// Combine whisker and box data for Plot
box_summary = [{
  x1: box_lower_whisker,
  x2: box_upper_whisker,
  q1: box_q1,
  q3: box_q3,
  median: box_median
}]

Plot.plot({
  style: {background: "transparent", fontFamily: "inherit"},
  marginBottom: 40,
  marginLeft: 40,
  height: 160,
  x: {label: "Value"},
  y: {axis: null},
  marks: [
    // Whisker line
    Plot.ruleX(box_summary, {
      x1: d => d.x1,
      x2: d => d.x2,
      y: 0,
      stroke: "#414042",
      strokeWidth: 1.5
    }),
    // Box
    Plot.rect(box_summary, {
      x1: d => d.q1,
      x2: d => d.q3,
      y1: -0.4,
      y2: 0.4,
      fill: "#F2F2F2",
      stroke: "#414042",
      strokeWidth: 1.5
    }),
    // Median line
    Plot.ruleX(box_summary, {
      x: d => d.median,
```

```
      y1: -0.4,
      y2: 0.4,
      stroke: "#414042",
      strokeWidth: 2
    }),
    // Outliers
    Plot.dot(box_outliers.map(v => ({v})), {
      x: d => d.v,
      y: 0,
      fill: "#ec8525",
      r: 3
    }),
    // Lower whisker (from lower whisker end to Q1)
Plot.link(box_summary, {
  x1: d => d.x1,
  x2: d => d.q1,
  y1: 0,
  y2: 0,
  stroke: "#414042",
  strokeWidth: 1.5
}),
// Upper whisker (from Q3 to upper whisker end)
Plot.link(box_summary, {
  x1: d => d.q3,
  x2: d => d.x2,
  y1: 0,
  y2: 0,
  stroke: "#414042",
  strokeWidth: 1.5
})
  ]
})
```

```
//| echo: false
html`<div style="font-size: 0.9em; color: #555; margin-top: 0.5em;
  padding: 0.75em 1em; background: #f9f9f9; border-left: 3px solid #414042;">
  ${box_annotation}
</div>`
```

When reading a boxplot, start with the median line. Its position

within the box indicates whether the data is roughly symmetric or skewed. A median sitting centrally suggests an even distribution on both sides, while a median closer to one end of the box indicates that values are more concentrated on that side. The relative length of the whiskers tells a similar story; a longer whisker on one side suggests the data extends further in that direction.

A wide box indicates high variability in the middle 50% of values, while a narrow box suggests values are tightly clustered. The interactive examples above allow you to see how these features change across different distribution shapes, and how the same patterns identified in histograms appear differently when summarised as a boxplot.

Comparing groups

Where boxplots are most powerful is in side-by-side comparison (Figure 8). Differences in median position indicate differences in central tendency between groups. Differences in box width and whisker length indicate differences in spread. Overlapping boxes suggest groups may not differ meaningfully, while clearly separated boxes suggest real differences worth investigating further.

Scatter Plots

A scatter plot displays the relationship between two numeric variables by plotting individual observations as points, with one variable on each axis. It is most appropriate when both variables are measured on a continuous scale, making it easy to see at a glance whether a relationship exists, what direction it takes, and how strong it appears to be.

Scatter plots are particularly valuable during exploratory data analysis as an early step before formal modelling. Seeing the shape of the relationship between two variables helps inform which analytical methods are likely to be appropriate.

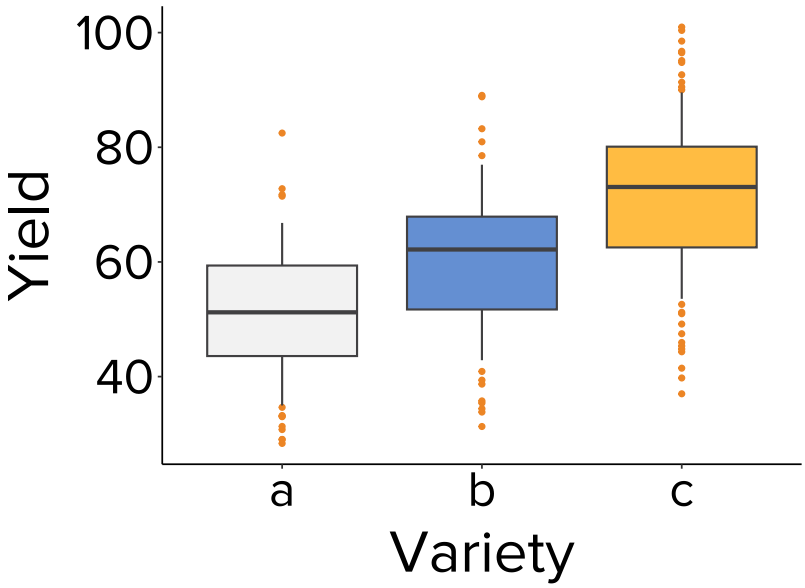


Figure 8: Boxplots comparing yield distributions across three crop varieties. Differences in the position of the median line indicate differences in central tendency between groups, while differences in box width and whisker length indicate differences in spread. Overlapping boxes suggest the groups may not differ meaningfully, while clearly separated boxes suggest real differences worth investigating further.

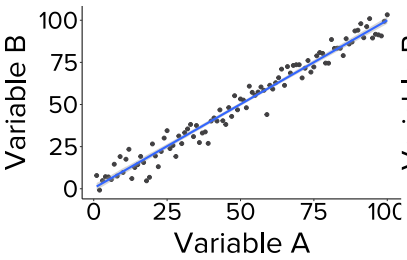


Figure 9: A scatter plot showing a strong positive linear relationship between two variables. Points cluster tightly around the fitted line, indicating that values of one variable are highly predictable from the other. The shaded band represents the confidence interval around the fitted line, reflecting uncertainty in the estimated relationship.

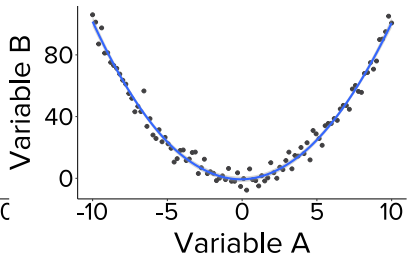


Figure 10: A scatter plot showing a non-linear relationship between two variables. The curved fitted line indicates that the rate of change is not constant across the range of values, meaning a straight line would poorly describe the relationship.

Reading scatter plots

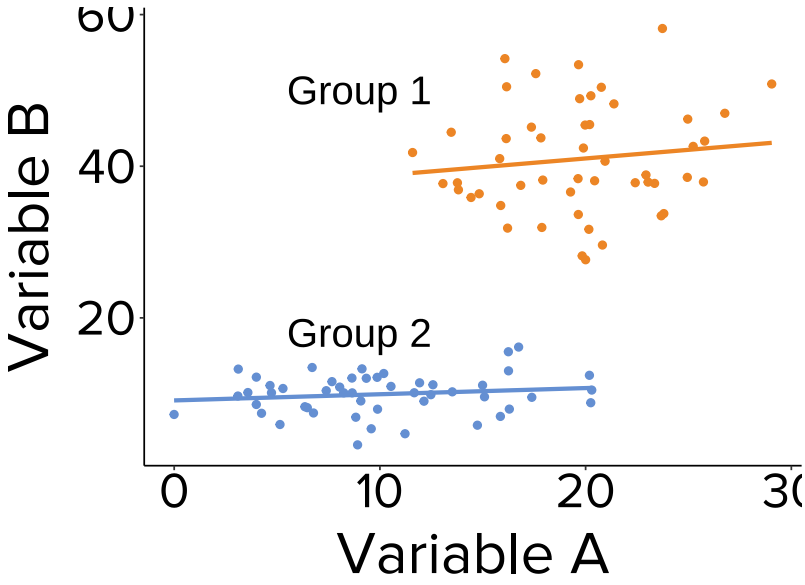


Figure 11: A scatter plot showing two distinct groups with different relationships between the same two variables. The steeper cluster of points indicates a more rapid rate of change, while the shallower cluster indicates a slower rate. When groups show different relationships like this, analysing them together as a single trend can be misleading, and the grouping variable should be accounted for in any subsequent analysis.

Start with the overall pattern. Points forming an upward trend indicate a positive relationship, where higher values of one variable tend to coincide with higher values of the other. A downward trend indicates a negative relationship. A random scatter with no clear direction suggests little or no relationship between the variables.

The closeness of points to an imaginary or fitted line reflects the strength of the relationship. Points clustering tightly around a clear trend indicate a strong relationship, while a more dispersed pattern suggests a weaker one. It is also worth looking at

whether the relationship is linear, following a straight line, or non-linear, following a curve, as this affects which analytical methods are appropriate (Figure 9, Figure 10).

Where points fall into distinct clusters or groups, analysing them as a single trend can be misleading. If a grouping variable is present in your data, it is worth plotting groups separately or using colour to distinguish them, as different groups may show quite different relationships (Figure 11).

With large datasets, points can overlap heavily, making it difficult to see the true density and distribution of observations. If a scatter plot looks unusually sparse or clustered in a way that seems inconsistent with the data, overplotting may be obscuring the true pattern.

Points sitting far from the main cluster may be outliers worth investigating. It is also important to remember that a relationship visible in a scatter plot indicates association, not causation. Two variables can be strongly correlated without one causing the other, and apparent relationships can sometimes reflect the influence of a third variable not included in the plot.

Numerical Summaries

Means

Standard Deviations

Frequency Tables