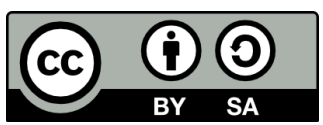
Applied Statistics

Drawing Graphs

Cesar O. Aguilar
SUNY Geneseo

Portions of these notes were created from *Learning statistics with R* by Danielle Navarro, *Learning statistics with jamovi* by David Foxcroft, and *Introduction to Statistical Thinking* by Benjamin Yakir.



Why draw graphs?

- Visualising data is one of the most important tasks facing the data analyst

Why draw graphs?

- Visualising data is one of the most important tasks facing the data analyst
- There are two main reasons for this:
  - Displaying your data in a clean, visually appealing fashion makes it easier for your reader to understand what you're trying to tell them
  - Drawing graphs helps **you** to understand the data

Why draw graphs?

- Visualising data is one of the most important tasks facing the data analyst
- There are two main reasons for this:
  - Displaying your data in a clean, visually appealing fashion makes it easier for your reader to understand what you're trying to tell them
  - Drawing graphs helps **you** to understand the data
- As you go about analysing data, it's important to draw "exploratory graphics" that help you learn about the data

Why draw graphs?

- Visualising data is one of the most important tasks facing the data analyst
- There are two main reasons for this:
  - Displaying your data in a clean, visually appealing fashion makes it easier for your reader to understand what you're trying to tell them
  - Drawing graphs helps **you** to understand the data
- As you go about analysing data, it's important to draw "exploratory graphics" that help you learn about the data
- Here is an interesting story about the importance of visualising data ...

Cholera deaths in 1854 London

- Cholera is caused by a bacteria in the digestive system that causes severe diarrhea, vomiting, muscle cramps, and can be fatal

Cholera deaths in 1854 London

- Cholera is caused by a bacteria in the digestive system that causes severe diarrhea, vomiting, muscle cramps, and can be fatal
- In 1854, an outbreak of cholera occurred in a neighborhood of London, England during the 1846-1860 worldwide cholera pandemic

Cholera deaths in 1854 London

- Cholera is caused by a bacteria in the digestive system that causes severe diarrhea, vomiting, muscle cramps, and can be fatal
- In 1854, an outbreak of cholera occurred in a neighborhood of London, England during the 1846-1860 worldwide cholera pandemic
- At the time, diseases such as cholera were believed to be caused by pollution or "bad air"; this was known as the **miasma theory**

Cholera deaths in 1854 London

- Cholera is caused by a bacteria in the digestive system that causes severe diarrhea, vomiting, muscle cramps, and can be fatal
- In 1854, an outbreak of cholera occurred in a neighborhood of London, England during the 1846-1860 worldwide cholera pandemic
- At the time, diseases such as cholera were believed to be caused by pollution or "bad air"; this was known as the **miasma theory**
- An english doctor named John Snow was skeptical of this theory and published an essay in 1849 called "On the Mode of Communication of Cholera"

Cholera deaths in 1854 London

- Cholera is caused by a bacteria in the digestive system that causes severe diarrhea, vomiting, muscle cramps, and can be fatal
- In 1854, an outbreak of cholera occurred in a neighborhood of London, England during the 1846-1860 worldwide cholera pandemic
- At the time, diseases such as cholera were believed to be caused by pollution or "bad air"; this was known as the **miasma theory**
- An english doctor named John Snow was skeptical of this theory and published an essay in 1849 called "On the Mode of Communication of Cholera"
- Snow argued that cholera was spread by contaminated water:

> *The views here explained open up to consideration a most important way in which the cholera may be widely disseminated, viz., by the emptying of sewers into the drinking water of the community*
>
> Snow, 1849, pg. 11

Cholera deaths in 1854 London

- By talking to local residents, Dr. Snow identified the source of the outbreak as one specific public water pump (Broad Street)

Cholera deaths in 1854 London

- By talking to local residents, Dr. Snow identified the source of the outbreak as one specific public water pump (Broad Street)
- Snow generated a map of the location of the water pumps throughout the affected area

Cholera deaths in 1854 London

- By talking to local residents, Dr. Snow identified the source of the outbreak as one specific public water pump (Broad Street)
- Snow generated a map of the location of the water pumps throughout the affected area
- Snow also performed a statistical comparison of cholera deaths based on which water companies sourced the distinct water pumps in the area

Cholera deaths in 1854 London

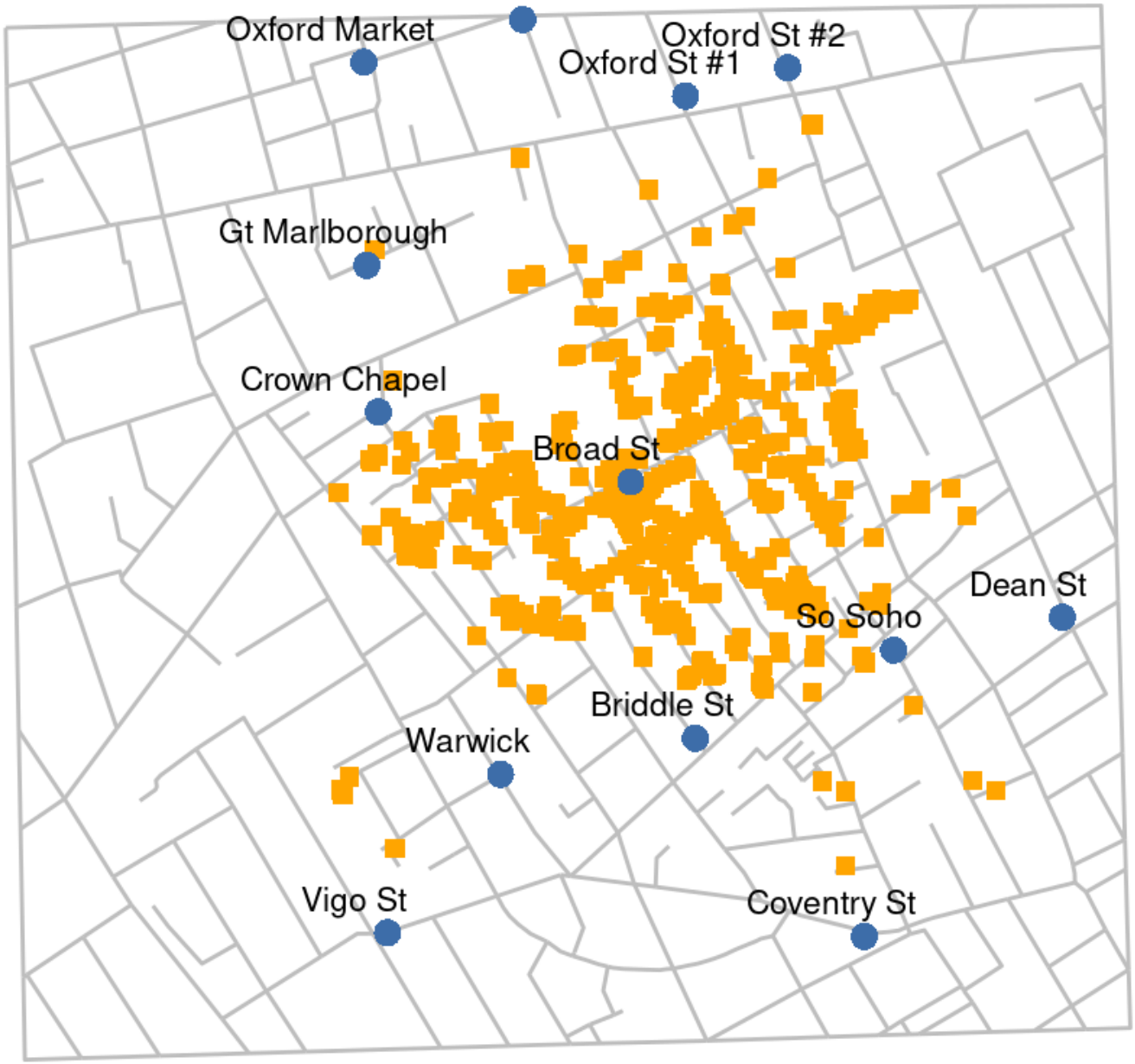
- By talking to local residents, Dr. Snow identified the source of the outbreak as one specific public water pump (Broad Street)
- Snow generated a map of the location of the water pumps throughout the affected area
- Snow also performed a statistical comparison of cholera deaths based on which water companies sourced the distinct water pumps in the area
- Snow showed that houses supplied by downriver water sources had a cholera mortality rate 14 times that of those supplied by upriver cleaner water sources

Cholera deaths in 1854 London

- By talking to local residents, Dr. Snow identified the source of the outbreak as one specific public water pump (Broad Street)
- Snow generated a map of the location of the water pumps throughout the affected area
- Snow also performed a statistical comparison of cholera deaths based on which water companies sourced the distinct water pumps in the area
- Snow showed that houses supplied by downriver water sources had a cholera mortality rate 14 times that of those supplied by upriver cleaner water sources

> *On proceeding to the spot, I found that nearly all the deaths had taken place within a short distance of the (Broad Street) pump . . . With regard to the deaths occurring in the locality belonging to the pump, there were 61 instances in which I was informed that the deceased persons used to drink the pump-water from Broad Street, either constantly or occasionally ..... .*

Cholera deaths in 1854 London



Cholera deaths in 1854 London

New location of the replica pump, the handle of which John Snow had removed (source: Wikimedia Commons)

Cholera deaths in 1854 London

John Snow pub, near the new location of the pump (source: Wikimedia Commons)

Types of graphs

- Histograms
- Boxplots
- Bar graphs
- Scatterplots

Histograms

- **Histograms** are one of the simplest and most useful ways of visualising data

Histograms

- **Histograms** are one of the simplest and most useful ways of visualising data
- They make most sense when you have an interval or ratio scale variable

Histograms

- **Histograms** are one of the simplest and most useful ways of visualising data
- They make most sense when you have an interval or ratio scale variable
- To create a histogram, divide up the possible data values into **bins** and then count the number of observations that fall within each bin

Histograms

- **Histograms** are one of the simplest and most useful ways of visualising data
- They make most sense when you have an interval or ratio scale variable
- To create a histogram, divide up the possible data values into **bins** and then count the number of observations that fall within each bin
- The bins, or buckets, are non-overlapping numerical intervals that cover the range of the data

Histograms

- **Histograms** are one of the simplest and most useful ways of visualising data
- They make most sense when you have an interval or ratio scale variable
- To create a histogram, divide up the possible data values into **bins** and then count the number of observations that fall within each bin
- The bins, or buckets, are non-overlapping numerical intervals that cover the range of the data
- The bins are often of equal width, but they don't have to be

Histograms

- **Histograms** are one of the simplest and most useful ways of visualising data
- They make most sense when you have an interval or ratio scale variable
- To create a histogram, divide up the possible data values into **bins** and then count the number of observations that fall within each bin
- The bins, or buckets, are non-overlapping numerical intervals that cover the range of the data
- The bins are often of equal width, but they don't have to be
- The count is referred to as the **frequency** or **density** of the bin and is displayed as a vertical bar

Example: Time spent on social media

- Suppose that we have a data set $X_1, X_2, \ldots, X_N$ measuring time spent on social media (hours), with min $= 2.5$ and max $= 9.5$

Example: Time spent on social media

- Suppose that we have a data set $X_1, X_2, \ldots, X_N$ measuring time spent on social media (hours), with min $= 2.5$ and max $= 9.5$
- We divide the range of the data into 7 bins of equal width:

$$[2.5, 3.49), [3.5, 4.49), [4.5, 5.49), \ldots, [7.5, 8.49), [8.5, 9.5]$$

Example: Time spent on social media

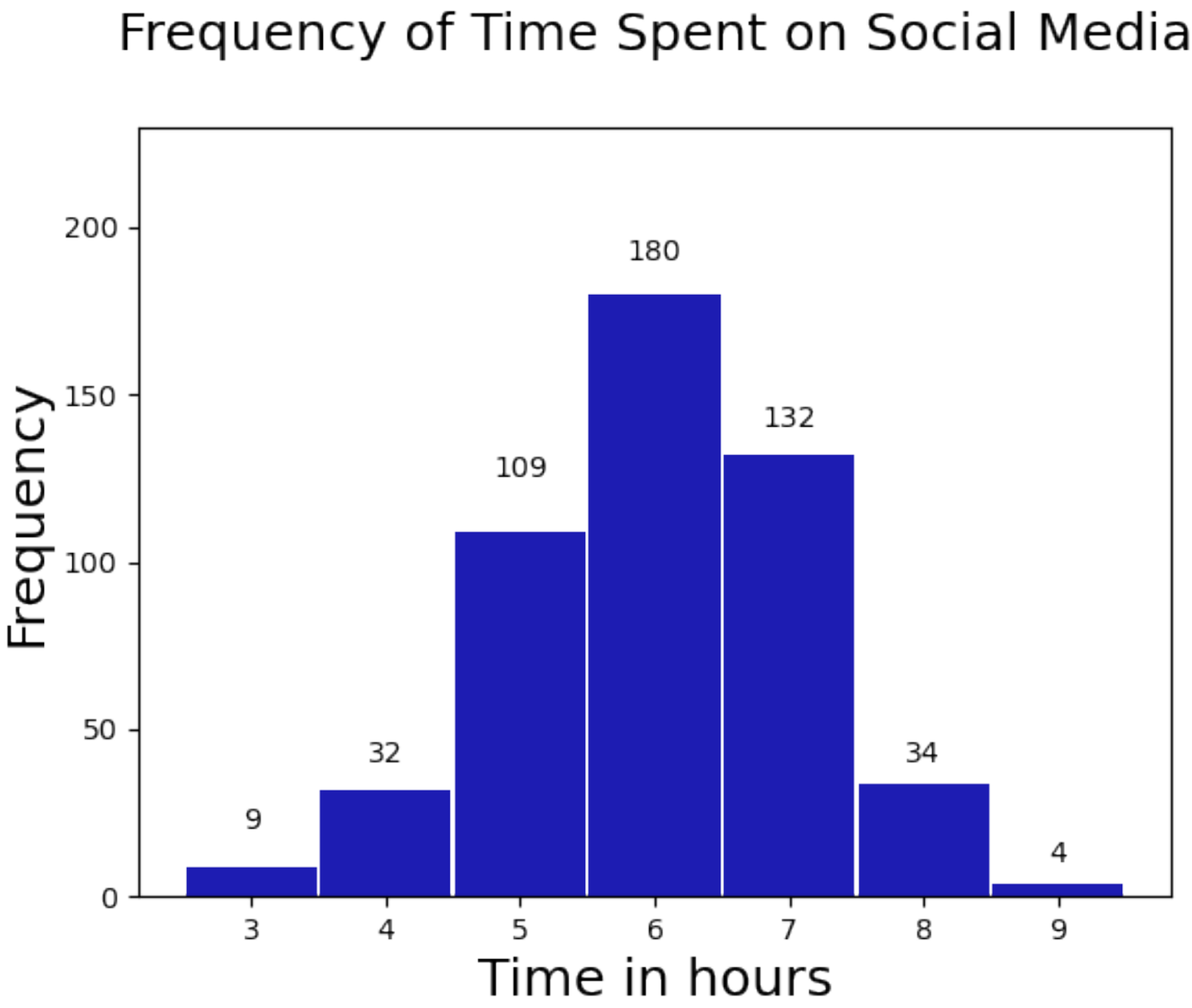
- Suppose that we have a data set $X_1, X_2, \ldots, X_N$ measuring time spent on social media (hours), with min $= 2.5$ and max $= 9.5$
- We divide the range of the data into 7 bins of equal width:

$$[2.5, 3.49), [3.5, 4.49), [4.5, 5.49), \ldots, [7.5, 8.49), [8.5, 9.5]$$

- We then count the number of observations that fall within each bin:

Bin/Interval	**Count/Frequency**
$[2.5, 3.49)$	9
$[3.5, 4.49)$	32
$[4.5, 5.49)$	109
$[5.50, 6.49)$	180
$[6.50, 7.49)$	132
$[7.5, 8.49)$	34
$[8.5, 9.5]$	4

Example: Time spent on social media



Skewness

- **Skewness** is a measure of the asymmetry of a distribution of data

Skewness

- **Skewness** is a measure of the asymmetry of a distribution of data
- A distribution of data is said to be **skewed** if it is not symmetric about its mean

Skewness

- **Skewness** is a measure of the asymmetry of a distribution of data
- A distribution of data is said to be **skewed** if it is not symmetric about its mean
- The formula for skewness is

$$\text{skewness}(X) = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^{N} (X_i - \bar{X})^3$$

Skewness

- **Skewness** is a measure of the asymmetry of a distribution of data
- A distribution of data is said to be **skewed** if it is not symmetric about its mean
- The formula for skewness is

$$\text{skewness}(X) = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^{N} (X_i - \bar{X})^3$$

- If there are relatively more values X_i that are greater than $\bar{X}$, then the skewness is **positive** (right-skewed) and the histogram will have a tail to the right

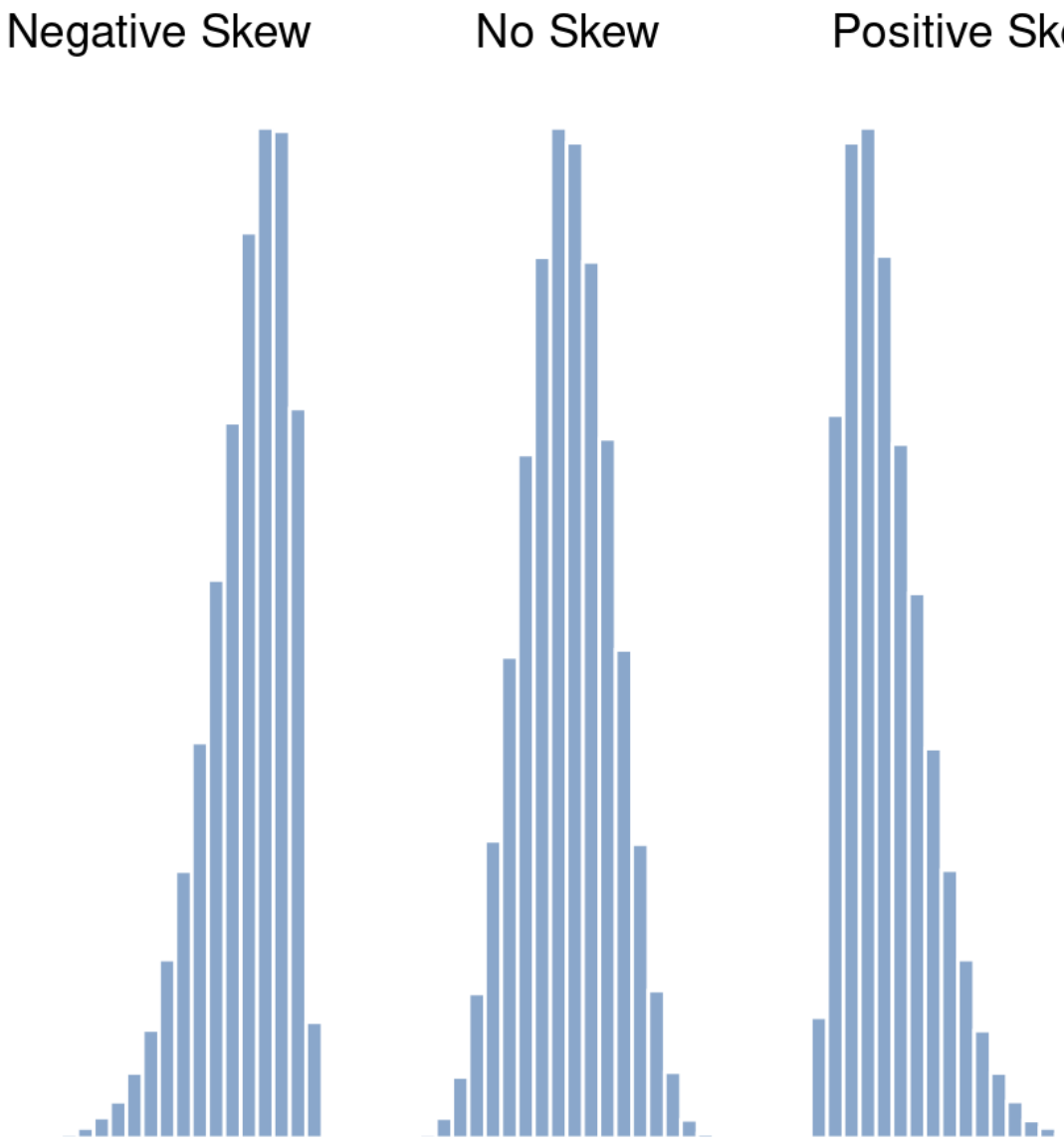
Skewness

- **Skewness** is a measure of the asymmetry of a distribution of data
- A distribution of data is said to be **skewed** if it is not symmetric about its mean
- The formula for skewness is

$$\text{skewness}(X) = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^{N} (X_i - \bar{X})^3$$

- If there are relatively more values X_i that are greater than $\bar{X}$, then the skewness is **positive** (right-skewed) and the histogram will have a tail to the right
- If there are relatively more values X_i that are less than $\bar{X}$, then the skewness is **negative** (left-skewed) and the histogram will have a tail to the left

Skewness



Box plots

- Also called box and whisker plots, **box plots** are a graphical representation of the five-number summary

Box plots

- Also called box and whisker plots, **box plots** are a graphical representation of the five-number summary
- Recall that the five-number summary consists of the minimum, Q_1, Q_2 (median), Q_3, and maximum

Box plots

- Also called box and whisker plots, **box plots** are a graphical representation of the five-number summary
- Recall that the five-number summary consists of the minimum, Q_1, Q_2 (median), Q_3, and maximum
- In a box plot, a box is drawn from Q_1 to Q_3 with a horizontal line drawn in the middle to denote the median (Q_2)

Box plots

- Also called box and whisker plots, **box plots** are a graphical representation of the five-number summary
- Recall that the five-number summary consists of the minimum, Q_1, Q_2 (median), Q_3, and maximum
- In a box plot, a box is drawn from Q_1 to Q_3 with a horizontal line drawn in the middle to denote the median (Q_2)
- Thin vertical lines called the **whiskers** extend from the box to a lower and upper value

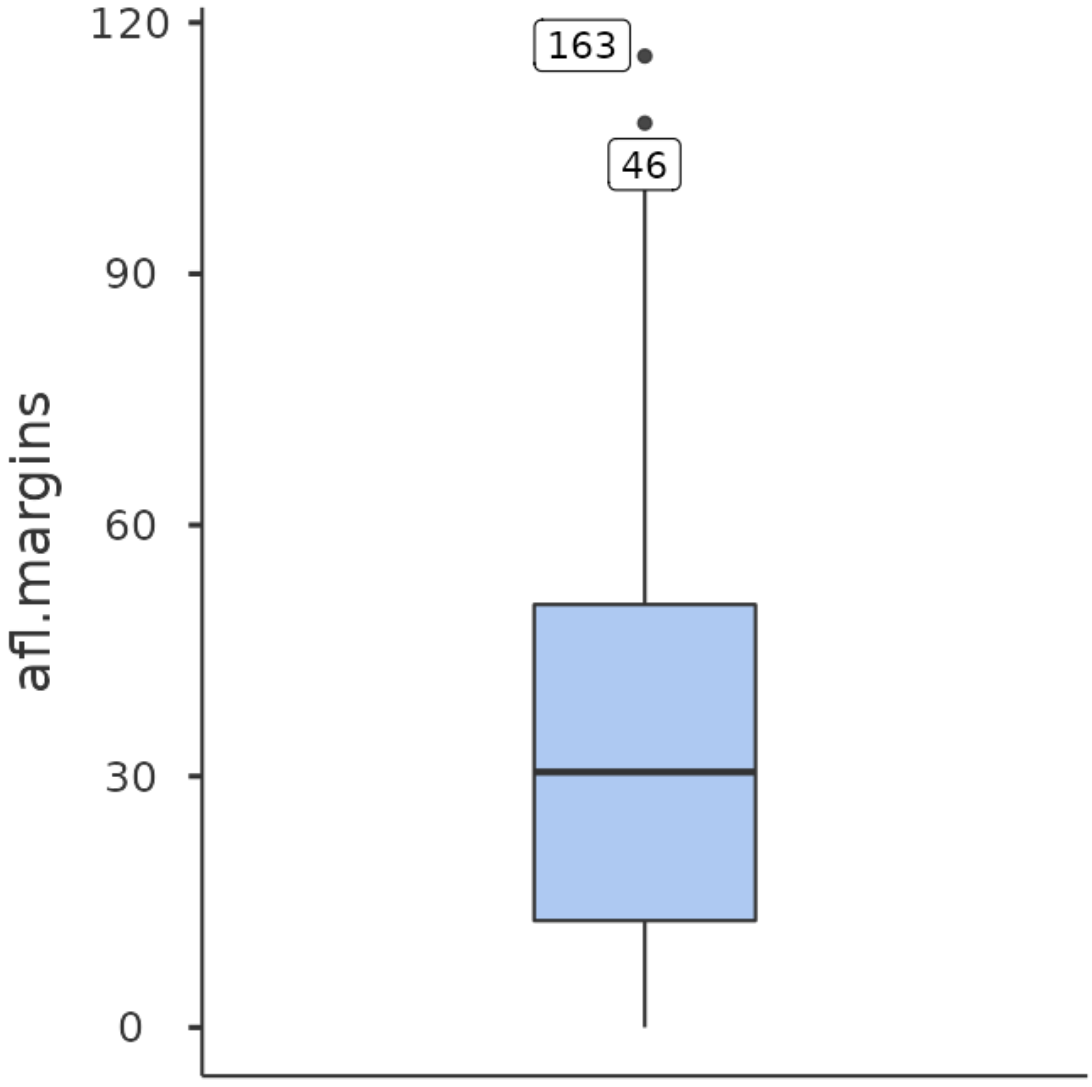
Box plots

- Also called box and whisker plots, **box plots** are a graphical representation of the five-number summary
- Recall that the five-number summary consists of the minimum, Q_1, Q_2 (median), Q_3, and maximum
- In a box plot, a box is drawn from Q_1 to Q_3 with a horizontal line drawn in the middle to denote the median (Q_2)
- Thin vertical lines called the **whiskers** extend from the box to a lower and upper value
- The lower and upper values could be:
  - lower = min and upper = max, or
  - lower = $Q_1 - 1.5 \times IQR$ and upper = $Q_3 + 1.5 \times IQR$ (the default)
  - lower = 9th percentile and upper = 91st percentile

Box plots

- Also called box and whisker plots, **box plots** are a graphical representation of the five-number summary
- Recall that the five-number summary consists of the minimum, Q_1, Q_2 (median), Q_3, and maximum
- In a box plot, a box is drawn from Q_1 to Q_3 with a horizontal line drawn in the middle to denote the median (Q_2)
- Thin vertical lines called the **whiskers** extend from the box to a lower and upper value
- The lower and upper values could be:
  - lower = min and upper = max, or
  - lower = $Q_1 - 1.5 \times IQR$ and upper = $Q_3 + 1.5 \times IQR$ (the default)
  - lower = 9th percentile and upper = 91st percentile
- Any values outside the whiskers are plotted as individual points and are called **outliers**

Example: Boxplots



Bar graphs

- A **bar chart** or **bar graph** is a graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent

Bar graphs

- A **bar chart** or **bar graph** is a graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent
- The bars can be plotted vertically or horizontally

Example: Bar graphs