**Chapter 7**

* A **variable** is a quantity, quality, or property that you can measure.
* A **value** is the state of a variable when you measure it.
* An **observation** is a set of measurements made under similar conditions An observation will contain several values, each associated with a different variable. I’ll sometimes refer to an observation as a data point.
* **Tabular data** is a set of values, each associated with a variable and an observation. Tabular data is tidy if each value is placed in its own “cell”, each variable in its own column, and each observation in its own row.

**Variation** is the tendency of the values of a variable to change from measurement to measurement.

This is true **even if you measure quantities that are constant**, like the speed of light. Each of your measurements will include a small amount of error that varies from measurement to measurement

**7.3.1 Visualising distributions**

* A variable is **categorical** if it can only take one of a small set of values, usually saved as factors or character vectors. To examine the distribution of a categorical variable, use a bar chart
  + The height of the bars displays how many observations occurred with each x value. You can compute these values manually with dplyr::count()
* A variable is **continuous** if it can take any of an infinite set of ordered values. To examine the distribution of a continuous variable, use a histogram
  + You can compute this by hand by combining dplyr::count() and ggplot2::cut\_width()
  + A histogram divides the x-axis into equally spaced bins and then uses the height of a bar to display the number of observations that fall in each bin.

**7.3.2 Typical values**

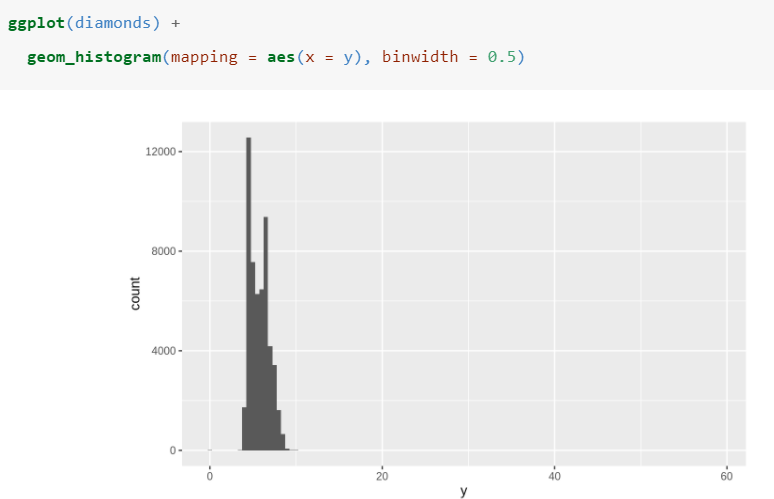
In both bar charts and histograms, tall bars show the common values of a variable, and shorter bars show less-common values.

Example with diamons: there’s more full carat diamonds and more values slightly to the right of the full carat

**7.3.3 Unusual values**

Outliers are observations that are unusual; data points that don’t seem to fit the pattern. Sometimes outliers are data entry errors; other times outliers suggest important new science.

There are so many observations in the common bins that the rare bins are so short that you can’t see them



**To check for values use ylim to focus on lower part of the histogram**

**ggplot**(diamonds) +

**geom\_histogram**(mapping = **aes**(x = y), binwidth = 0.5) +

**coord\_cartesian**(ylim = **c**(0, 50))

coord\_cartesian() also has an xlim() argument for when you need to zoom into the x-axis. ggplot2 also has xlim() and ylim() functions that work slightly differently: they throw away the data outside the limits

finding outliers that can mean false values like here at 0 or in the high 49=0’s and 60’s as they would be huge diamonds

**7.4. Missing Values**

**1.** Drop the entire row with the strange values

Not really recommended as it will filter out a lot of valuable data if applied for every missing value

diamonds2 <- diamonds %>%

**filter**(**between**(y, 3, 20))

**2.** replacing the unusual values with missing values; use mutate(); use the ifelse() function to replace unusual values with NA

diamonds2 <- diamonds %>%

**mutate**(y = **ifelse**(y < 3 | y > 20, NA, y))

To suppress that warning about missing values that weren’t included in the plot, set na.rm = TRUE:

**ggplot**(data = diamonds2, mapping = **aes**(x = x, y = y)) +

**geom\_point**(na.rm = TRUE)

**7.5 Covariation**

If variation describes the behavior within a variable, covariation describes the behavior **between variables**.

**Covariation** is the tendency for the values of two or more variables to vary together in a related way.

**7.5.1 A categorial and continuous variable**

Usually one wants to explore a continuous var. broken down into categories (cat. Variable)

In geom\_freqpoly():

The height is given by the count; this makes it hard to see the differences between the difference in distribution of the categories because the counts differ so much

**ggplot**(diamonds) +

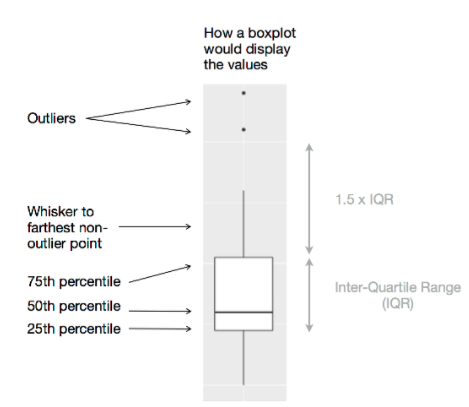
**geom\_bar**(mapping = **aes**(x = cut))

To make the comparison easier we need to swap what is displayed on the y-axis. Instead of displaying count, we’ll display **density**

**ggplot**(data = diamonds, mapping = **aes**(x = price, y = ..density..)) +

**geom\_freqpoly**(mapping = **aes**(colour = cut), binwidth = 500)

Another alternative to display the distribution of a continuous variable broken down by a categorical variable is the **boxplot**

* A box that stretches from the **25th percentile of the distribution to the 75th percentile**, a distance known as the **interquartile range (IQR).** In the middle of the box is a line that displays the **median**, i.e. 50th percentile, of the distribution.
* A line (or whisker) that extends from each end of the box and goes to the  
  farthest non-outlier point in the distribution

**7.5.2 Two categorical variables**

To visualise the **covariation between categorical variables**, you’ll need to count the number of observations for each combination (geom\_count())

**ggplot**(data = diamonds) +

**geom\_count**(mapping = **aes**(x = cut, y = color))

**or use**

diamonds %>%

**count**(color, cut)

to compute the count and then visualize it

**7.5.3 Two continuous variables**

* draw a scatterplot with geom\_point(). You can see covariation as a pattern in the points.
* Scatterplots become less useful as the size of your dataset grows, because points begin to overplot, and pile up into areas of uniform black
* using the alpha aesthetic to add transparency.

**ggplot**(data = diamonds) +

**geom\_point**(mapping = **aes**(x = carat, y = price), alpha = 1 / 100)

Previously you used geom\_histogram() and geom\_freqpoly() to bin in one dimension. Now you’ll learn how to use geom\_bin2d() and geom\_hex() to bin in two dimensions.

* divide the coordinate plane into 2d bins and then use a fill color to display how many points fall into each bin

**7.6 Patterns and models**

Patterns in your data provide clues about relationships. If a systematic relationship exists between two variables it will appear as a pattern in the data.

A scatterplot of Old Faithful eruption lengths versus the wait time between eruptions shows a pattern: longer wait times are associated with longer eruptions.

Patterns provide one of the most useful tools for data scientists because they reveal covariation

**7.7 ggplot2 calls**

The first two arguments to ggplot() are data and mapping, and the first two arguments to aes() are x and y

**ggplot**(faithful, **aes**(eruptions)) +

**geom\_freqpoly**(binwidth = 0.25)