**Got It!**

**1. Histograms**

Let's explore histograms.

**2. When should you use a histogram?**

Histograms are a type of plot that takes one continuous variable as its input. It allows you to answer questions about the shape of that variable's distribution. For example, you might want to know the lowest and highest values, and which values are most common.

**3. Kings and Queens of England & Britain**

Here's a dataset on the kings and queens of England, and more recently Britain. It stretches from the current monarch, Elizabeth the second, back in time to the first king of England, Aethelstan. Let's take a look at the distribution of the ages when they ascended to the throne.

**4. Histogram of age at start of rule**

The x-axis is the variable that we are interested in - the ages. These ages are grouped into "bins", that is, intervals. In this case, the bins are zero to five years, five to ten years, and so on up to sixty to sixty five years. The y-axis is the count of monarchs who began ruling when they were in each age bin. For example, four monarchs began ruling when they were between ten and fifteen years old. Straight away, you can see that there have been no monarchs who started ruling when they were between the ages of forty five and fifty.

**5. Choosing binwidth: 1 year**

The appearance of a histogram is strongly influenced by the choice of binwidth. This is the same histogram as before, but with the binwidth changed from five years to one year. It's difficult to get much insight into the distribution, because the counts are very noisy. Choosing a binwidth that is too wide also causes problems.

**6. Choosing binwidth: 25 years**

By changing the binwidth to twenty five years, you don't see any detail in the distribution, and again it is hard to get much insight. In general, it is difficult to know the best binwidth before you draw the plot, so you'll need to experiment with several values.

**7. Modality: how many peaks?**

When interpreting histograms, the first thing to look at is the modality of the distribution. That is, how many peaks there are. A distribution with one peak is called "unimodal"; a distribution with two peaks is called "bimodal", and so on.

**8. Modality: how many peaks?**

Here, the distribution of ages is unimodal because there is one peak from twenty five to thirty five years.

**9. Skewness: is it symmetric?**

The second thing to look at is the skewness of the distribution. That's statistical jargon for whether or not it is symmetric. A left-skewed distribution has outliers, that is, the extreme values, on the left, and a right-skewed distribution has outliers on the right. The outliers look a bit pointy, so I like to remember which way round these are by imagining that the outliers are a shish-kebab skewer. In a left-skewed distribution, the skewer points leftwards, and in a right-skewed distribution, the skewer points rightwards.

**10. Skewness: is it symmetric?**

Here, the distribution is more or less symmetric.

**11. Kurtosis: how many extreme values?**

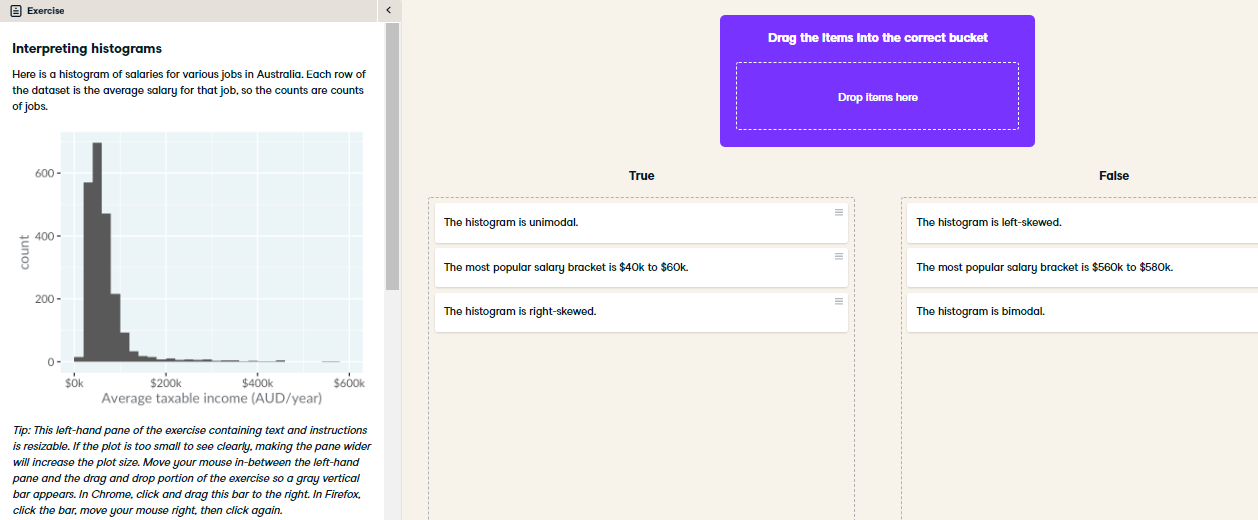
One more advanced thing you can look at is the kurtosis of the distribution, which affects the number of outliers. A mesokurtic distribution is something that looks like the bell curve from a normal distribution. A leptokurtic distribution has a narrow peak and lots of extreme values. Leptokurtic distributions are important in finance, because weird stuff happens in stock markets more often than the normal distribution would predict. A platykurtic distribution has a broad peak and few extreme values.

**12. Kurtosis: how many extreme values?**

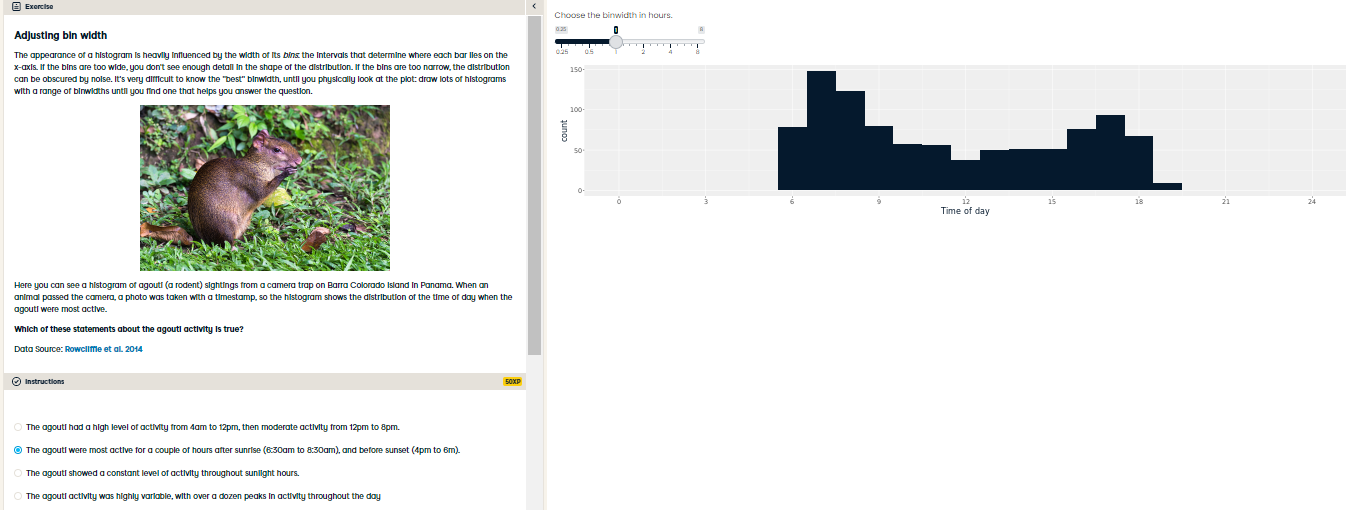
Here, the distribution of ages is slightly platykurtic.

**13. Let's practice!**

Time to draw some histograms.



Hooray for histograms! Counting the number of peaks, assessing the skewness (lack of symmetry), and identifying the most popular areas are useful things to look at when interpreting a histogram.



Brilliant binwidth adjustment! Choosing an appropriate binwidth make it easier to interpret the histogram.