Hello, my name is Holly Tibble. In this demonstration, we're going to go through how to plan, create, and describe a simple visualisation - Respiratory deaths in the UK throughout the year, between 2011 and 2012. First, let's define our research question. In 2012 there were more respiratory deaths in the UK than in the previous year, despite a bad flu season in 2011. Were there differences throughout the year in the death rate, or was 2012 consistently higher? We're going to recreate Figure A above in R, but adding in the 2012 data. This video demonstrates how we can do this using four useful R packages. Before we plot, we're going to do a little bit of data manipulation using the dplyr package. Dplyr is a powerful R package to transform and summarise tabular data with rows and columns. In particular, we want to use the mutate function to make a new variable, and the filter function to select a subset of the full dataset. But we're also going to use the dplyr pipe operator to make this date manipulation super efficient. The second package is lubridate, which includes useful functions for formatting dates and times. We are using the year function to extract the year from a date variable. The third package is NHS R datasets, a free collaborative datasets package for the NHS R community when learning or teaching R. Finally for the plotting itself, we are using ggplot 2.  ggplot 2 and dplyr, along with several other useful packages, are part of the tidyverse, which can be installed and loaded in one go rather than individually. Efficiency! The dataset we're using in this demonstration is called ONS_Mortality - mortality records compiled by the Office of National Statistics. The dataset was pulled together from the weekly Excel spreadsheets provided by the Office of National Statistics or ONS, from 2010 to 2019 for training purposes and is a static dataset. The dataset includes provisionally-registered deaths for England and Wales, for all calls, regions of usual residence, age bands, sex, and by underlying cause, in particular all respiratory diseases. We're going to focus on the latter summaries: deaths caused by respiratory diseases. Respiratory diseases, including asthma, influenza, and chest infections, are a major source of unscheduled care in the UK. So better understanding annual trends enables us to better prepare resource allocation. Cause of death is recorded using the International Classification of Diseases or ICD. We're currently using version ten. It has ICD-10 for short. Respiratory diseases are coded starting with a J and then the numbers 00 to 99. J-45, for example, is an asthma attack. As a final note, be careful of the difference between the pipe operator in dplyr and the plus in ggplot. They may have similar purposes but they cannot be interchanged. In this example, I have kept all of the manipulation separate to the plotting to keep it clear. Although you can also do manipulation in ggplot functions when you get competent.

Okay, let's jump in. First, we're going to load the tidyverse package, which I've already installed. Next, we're going to load up the NHSR dataset library, which contains some health care datasets we can play about with. Finally, we're going to do a very basic bit of manipulation with dates. So, we're going to load the lubridate package. So, let's load up the dataset ONS mortality, with mortality records compiled by the Office of National Statistics. We can see it contains over 18,000 rows and has five variables, one of which is date. We're going to go a little bit further than the plot I showed you above. I'm going to plot multiple years of data. So first, let's make a new variable called year using the year function from lubridate. We're also using the dplyr package here for both the mutate function and the piping. This dataset is quite complicated in layout. But what we want is to select the rows relating to respiratory diseases with the coding version from 2010 and data in 2011 or 2012. The first bit of plotting we're going to do is as simple as can be. First, we call the function ggplot and tell it which dataset we're using. Then we choose the geom, which is the graphing object. Here we are using geom_line, which is a line plot. Inside the GM object, we give it the basic aesthetics, or aes, which variables on the x-axis, which is on the y-axis. We can see that something isn't quite right here - the lines look very odd. This is because we have one value each week for the two years in the

dataset. We're going to add another option to the aesthetics section to group the lines by the variable 'year'. Great. Now we can see two separate lines, but we don't know which is which. We can add the option to colour by year 2, so the lines can be differentiated. We can see that on the right-hand side it has added a little key with a gradient. By stating this variable should be treated as a factor, we tell R it can't have values that are not integers and changes the key to being distinct values. We can also take out the group option in the aesthetics, because when we're colouring by year, it knows that the points are grouped by year already. Now let's get customising. On the y-axis, we can see that there are data points that go below the lowest number. This makes it a little hard to estimate what these values actually are. Furthermore, almost always rocks axis should start at 0. We can change this by setting the limits of the axis. We also want to change the x axis a bit. There is no Week 0 and the highest week is 52. It doesn't look perfect, but it starts and ends in a logical place. Awesome. Let's add some more labels to the plot to make it super clear what we're showing. We can add a title. We can change the title of the legend. And finally, we can change the theme, which alters loads of specifics, of the appearance of the plot to a nice preset theme, which is very printer friendly.

Woohoo! We've made a beautiful plot. So, what can we tell about this plot? While deaths were higher over the entirety of 2012? This plot adds a lot of information about where there were differences. We can see that between the 25th week, which is summertime, until the end of the year, there was very little difference. For the first two weeks of January, they're almost 2000 more deaths in 2011 than in 2012. But then the deaths stayed higher for longer in 2012. Maybe the flu season started earlier at the end of 2010. This plot suggests that respiratory deaths might be better viewed from summer to summer rather than by calendar year. An interesting insight.