**Video Script: Section 2 Video 4 – plotting histograms and density plots.**

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Description | Action on screen | Narration |
| 1 | Introduction  (Outcome and why it is desirable)  This should give the viewer an idea of the outcome of the task at the beginning of the videos and set the stage and expectations of the viewer. | Opening slides. | **In this video**, we are going to take a look at histograms and density plots for visualizing the distribution of the data. |
| 2 | Context(Problem/Solution)  Present the viewer with a real-world solution and how the situation would pose as a challenge. It always helps to draw the viewer's attention using a use-case. Metadata template can be used here. |  | It’s often interesting to visualize the range and the most frequently occurring values of your data. For this, we can use an histogram or a density plot. Ggplot2 offers both of these options. |
| 3 | Guidance (How to do it and how it works): | Switch to RStudio,  Activity01\_04.R appears in the editor. | Let's see a few examples. Open activity02\_04.R. (CTRL+O) in RStudio. |
| 4 | Consider a simple scatter plot. | Highlight and run:  library(ggplot2)  head(diamonds)  carat cut color clarity depth table price x y z  1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  4 0.29 Premium I VS2 62.4 58 334 4.20 4.23 2.63  5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48 | Run the first 2 lines of the code (select , CTRL+ENTER).  We will be using the diamonds dataset again. Type ?diamonds in the console if you want to know more about it. |
| 5 |  | Highlight and run:  # Example 01  # distribution of price.  ggplot(diamonds) + geom\_histogram( aes(x = price)) + ggtitle("Histogram of price (counts)") | Let’s see what is the distribution of diamond prices in the dataset.  Run example 01. |
| 6 |  | A description... | Geom\_histogram() has one required aesthetics : x  The range of values for x will be divided in equally-sized bins (30 bins by default) and the histogram shows the number of rows for which x (here the price of the diamonds) falls into the bin.  This gives us a clear picture of the range of values spanned by the price of a diamond. For example, most diamonds cost less than 5000$ and most expensive diamonds are rarer. |
| 7 |  |  | If you add all the bins together, you’ll find the total number of diamonds in the dataset. |
|  |  | Select and run:  # Example 02  # changing the binwidth  ggplot(diamonds) + geom\_histogram( aes(x = price,fill=cut), binwidth=100) + ggtitle("Histogram of price (counts), smaller binwidths")  A description... | You can easily change the number of bins with binwidth. For example:  binwidth=100  Note how we have also used the aesthetics ‘fill’ to get a separate histogram for each value of cut. |
| 8 |  |  | If you have a large number of data points, you can use a density plot instead.  The density plot does two things:   1. It smoothes out the distribution. 2. The area under the curve sums to 1, so it’s easier to compare distributions even they don’t have the same total number of items; doubling the number of items doesn’t make the curve twice higher. |
| 9 |  | Highlight and run:  # Example 03  # density of price  ggplot(diamonds) + geom\_density( aes(x = price, fill = cut)) + ggtitle("Density plot of price, by cut (normalised and smoothed out)")  A description... | Run example 03 to see the results.  Using geom\_density instead of geom\_histogram made the presentation of the distribution smoother on the eye.  You should only use density plots when you have a large number of items: the smoothing can lead to misleading artefacts, bumping up values of low frequency. |
| 10 |  |  |  |
| 11 |  | Highlight and run:  # Example 04  # density of price, by cut  ggplot(diamonds) + geom\_density( aes(x = price, fill = cut), alpha=0.3) + ggtitle("Density plot of price, by cut (normalised and smoothed out)")  A description... | The distribution for diamonds of ‘ideal’ cut masks the other distributions. We can make the colouring slightly transparent by using the parameter ‘alpha’, a number between 0 and 1, from completely transparent to completely opaque, as in example 4 |
| 12 |  |  |  |
| 13 |  |  |  |
| 14 |  |  |  |
| 15 |  |  |  |
| 16 | Conclusion:The video concludes by showing the viewer that the goal has been achieved, and reminding them why they should be happy about that. A PowerPoint summary slide with the key points emphasized would make it easier for the viewer to remember what was covered in the video | Last slide of the PPT | We’ve seen how to use histograms and density plots for visualizing the distribution of a quantity.  In the next video, we will see how to represent the data more succinctly with boxplots. |