**Video Script: Section 2 Video 5 – Using boxplots**

|  |  |  |  |
| --- | --- | --- | --- |
| 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 will see how to use a common statistical plot: boxplots. |
| 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. |  | Histograms and density plots show us the whole distribution of a quantity.  They take a lot of space and can be difficult to compare the more distributions you have.  Boxplots have less information, but are more concise, so you can compare lots of them easily. |
| 3 | Guidance (How to do it and how it works): | Open activity01\_05.R in the editor.  Highlight and run:  library(ggplot2)  head(diamonds)  summary(diamonds)  > 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  > summary(diamonds)  carat cut color clarity depth table price  Min. :0.2000 Fair : 1610 D: 6775 SI1 :13065 Min. :43.00 Min. :43.00 Min. : 326  1st Qu.:0.4000 Good : 4906 E: 9797 VS2 :12258 1st Qu.:61.00 1st Qu.:56.00 1st Qu.: 950  Median :0.7000 Very Good:12082 F: 9542 SI2 : 9194 Median :61.80 Median :57.00 Median : 2401  Mean :0.7979 Premium :13791 G:11292 VS1 : 8171 Mean :61.75 Mean :57.46 Mean : 3933  3rd Qu.:1.0400 Ideal :21551 H: 8304 VVS2 : 5066 3rd Qu.:62.50 3rd Qu.:59.00 3rd Qu.: 5324  Max. :5.0100 I: 5422 VVS1 : 3655 Max. :79.00 Max. :95.00 Max. :18823  J: 2808 (Other): 2531  x y z  Min. : 0.000 Min. : 0.000 Min. : 0.000  1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910  Median : 5.700 Median : 5.710 Median : 3.530  Mean : 5.731 Mean : 5.735 Mean : 3.539  3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040  Max. :10.740 Max. :58.900 Max. :31.800 | Open activity\_01\_05.R in RStudio.  Run the first few lines to load ggplot2 and display some numerical information about the dataset. |
| 4 |  | Highlight and run:  # Example 01  # box plot of a diamond's price given its cut  ggplot(diamonds, aes( x = cut, y = price)) + geom\_boxplot() + ggtitle("box plot of a diamond's price given its cut")A description... | Run example 01 in the console. The graph shows the distribution of the price for each value of cut, in the form of box plots. |
| 5 |  | Point to the middle bar | The middle bar of the box plot is the median value: 50% of the prices are below this quantity, 50% are above it. |
| 6 |  | Point to top and bottom of the boxes | The bottom and top values of the box are the 25% and 75% quartiles. |
| 7 |  | Point to whiskers | The whiskers represent where rarer values would lie. |
| 8 |  | Point to dots. | Dots are extreme points, or outliers. |
| 9 |  |  | Box plots are very good for summarizing a dataset with a few numbers but can give an impression that the distribution has one single peak. It is good practice to check whether this is the case , for example, with a density plot. |
| 10 |  |  |  |
| 11 |  |  |  |
| 12 |  | Select and run:  # Example 02  # boxplot of a diamond's price given its cut  p + geom\_boxplot(notch=TRUE) + ggtitle("boxplot of diamond prices given their cut, with notch")  A description... | You can add an extra piece of information to the boxplot with the parameter ‘notch’, which indents the box around the median. This is used for comparing two boxplots: if the notches don’t overlap, the two medians are very likely to be different.  Add a notch to the previous plot with:  geom\_boxplot(notch=TRUE)  For example, diamonds of fair and good cuts have very close medians, but since their notches don’t overlap, their difference is statistically significant. |
| 13 |  | Select and run:  # Example 03  # boxplot with notch,  p + geom\_boxplot(aes(fill = clarity), notch=TRUE) + ggtitle("box plot of diamond prices given their cut and their clarity")  A description... | As with any other geom, you can change many more aesthetics. For example, you can map the filling color with ‘fill’. If we map it to clarity, a property of the diamonds, we can compare the price of diamonds on the basis of their cut and clarity.  Run the third example. |
| 14 |  |  | We can easily compare 40 distributions in one graph. |
| 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 boxplots to summarise and compare multiple distributions.  In the next section, we’ll see how to use group and color for conveying more information in a graph. |