

Lesson 4 Exercises

Stefanie Molin

April 17, 2017

Let's do some practice problems to challenge your understanding of `ggplot2` and review the material from the prior lessons.

1. Using `ggplot2` create a scatterplot of CPC and clicks for the client of your choice daily for the last 30 days. Add red crosshairs on the graph to indicate points outside the Tukey fence (outliers), which has bounds $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$ where interquartile range (IQR) is defined as $IQR = Q3 - Q1$ and Q stands for quartile. Note you will need the `quantile()` and `IQR()` functions. Try to write a simple query and use `dplyr` to manipulate your data.
2. Pull clicks and CPC by category by day for the last 30 days on the client of your choice (choose a category level that makes sense). Create a quick base R histogram to see the distribution of CPC. Then, create a scatterplot of CPC and clicks faceted by category (be sure to try both `facet_wrap()` and `facet_grid()` to see which works best for your data). Limit to categories that have more than 1000 clicks on a given day (or a relevant threshold). Add in regression lines. Note that in this example `facet_wrap()` will most likely look the best, but you should still try `facet_grid()`.
3. Graph the percentage of deduplicated sales credited to [REDACTED] by hour between click and purchase for the last 30 days. If you can't think of a client using the deduplication parameter, use [REDACTED]
[REDACTED]
 - Write a Vertica query to pull in the amount of deduplicated sales credited to [REDACTED] by hour between click and purchase as well as the total sales for the last 30 days. Don't bucket the hours in your query—you will work with them in `dplyr`!
 - Using `dplyr` verbs, bucket the hours above 30 to the hour 31 bucket, remove any nonsense data, add a column for percentage of duplicated sales credited to [REDACTED], reduce the data into just the 2 columns needed for the graph, and sort the data by bucket. Note it is possible to use all 5 verbs and `group_by()` here and that you should use the pipe operator (`%>%`).
 - Use `ggplot2` to generate a line graph of the dedup ratio by hour.

Extra credit: format the y-axis and add labels for the axes and the title.