Tutorial 07 Interaction

This tutorial consists of just one data exploration and uses data that is relatively similar to the data in the lecture notes and examples.

Notebooks: For the tutorials we will be using Colab to create Python Notebooks on your Google Drive (although you may also use Jupyter). Read the document Tutorials.pdf from Moodle to see how to set these up.

For this tutorial you also need to make some minor changes to the code (see the lecture notes, slide 22).

Naming: For the exploration you should create a Python notebook and save it when you have finished. You should name the notebook Tut07.ipynb.

Structure: Every numbered item in the exploration should have a code section and a markdown section underneath where you discuss your findings. There should also be a code section at the top of the notebook with the imports.

Exploration

The company who supplied the Products data in the lecture notes also want an investigation into their website and in particular want some interactive plots of the data you have been exploring over the last few tutorials which they can use to example the data more closely.

1. First read in the DailyHits.csv data (https://tinyurl.com/ChrisCoDV/Pages/DailyHits.csv) and create an interactive visualisation showing line plots of all the pages on the site.

2. Next create an interactive visualisation showing just the 2 high volume pages as line plots. Almost every day one of the pages has more hits than the other … except for one day. Which day was that? [**Hint**: use the box zoom tool.]

3. Now create an interactive visualisation showing superimposed histograms of the distributions of high volume pages with customised bin sizes. [**Hint**: this is a straightforward adaptation of one of the lecture examples – however you will need to change the key parameters x_min and x_max or nothing will show up – start with a wide range, x_min small and x_max large and then narrow it down. You should also adapt the bin width so that there are somewhere between 10 to 50 bins.]

4. Next create an interactive visualisation showing a heatmap of time-series correlations between all pages. Using the box zoom confirm that the strongest positive correlations are between pages 155 and 156. Where are the strongest negative correlations?

5. In fact you should have previously found in Tutorial 04 that the 3 most closely correlated pages, in terms of hits, are 048, 155 & 156. Create an interactive visualisation containing 3 scatter plot comparisons (as subplots) for each pair of these pages. [**Hint**: adapt one of the lecture examples and change the x and y limits so that the scatter plots fit nicely in the axes.]

   On the day that 155 and 156 have 34 and 35 hits, how many does 048 have? [**Hint**: use the hover tool.]

6. Now create an interactive visualisation showing the pages 048, 155 & 156 and consisting of 3 line subplots. Using the zoom possibilities, zoom into a region covering about a month's worth of data – which 2 pages appear to match each other most closely in terms of page hits and why?

7. Finally create a bubble plot that uses the summary data you constructed in last week's tutorial and shows annual hits vs revenue, with the bubble sizes determine by viewing time. Use the guidelines in the lecture to set the bubble scaling factor.

   Comment on whether the hits seems to be correlated with revenue and how interaction allows you to investigate this. [**Hint**: use the box zoom to investigate the bottom left corner.]

Certain of your solutions should look something like the following (the last picture has been zoomed in):