Tutorial 04 Correlation

This tutorial consists of two data explorations. The first uses data that is relatively similar to the data in the lecture notes and examples. The second is more unstructured and open ended.

Notebooks: 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.

Naming: For each exploration you should create a notebook and save it when you have finished. You should name the two notebooks Tut04-A.ipynb and Tut04-B.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 A

The company who supplied the Products data in the lecture notes also want an investigation into their website. As last week, the data (https://tinyurl.com/ChrisCoDV/Pages/DailyHits.csv) shows the number of page hits per day.

As in the lecture examples the company wants to explore correlations between page hits over the year. In tutorial 02 you should have identified the high volume and medium volume pages, with the low volume being all the others. It was your choice how high, medium and low were defined, but for the purposes of this exploration pick the top 2 as high volume and the next 8 as medium volume (this is very likely to be the choice you made originally).

1. Create a scatter plot of the two high volume pages against each other. Can you detect any correlation visually?

2. Create a single visualisation showing scatter subplots for all pairs of medium volume pages against each other. Which two pages are most strongly correlated with each other? Which other pages seem to have some correlation, visually?

3. Now confirm your findings by creating a heatmap showing the correlations between all the high and medium volume pages. Comment on the strongest positive and inverse correlations.

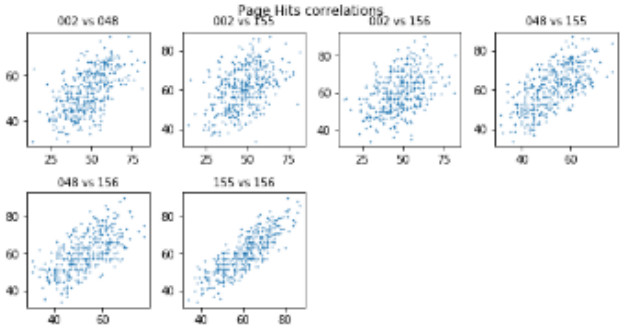
4. After seeing the heatmap, the company has decided it is interested in all positive correlations where the Pearson coefficient is greater than 0.5. Identify the corresponding pages and create a visualisation showing scatter subplots for all such pairs of pages.

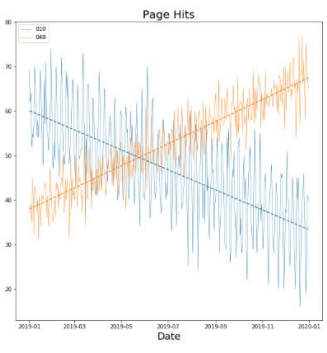
5. Now draw line plots with trendlines for the positively correlated pages you selected above.

6. The company is also interested in all inverse correlations where the Pearson coefficient is less than –0.55. Identify the corresponding pages and create a visualisation showing scatter subplots for all such pairs of pages.

7. Now draw line plots with trendlines for the inversely correlated pages you selected above.

   Your visualisations for items 4 and 7 should look something like those below.





Exploration B

The file iris.csv (available at https://tinyurl.com/ChrisCoDV/iris.csv) is a well-known dataset created by the statistician and biologist Ronald Fisher for a paper published in 1936. It contains measurements of 3 different varieties of iris flower, setosa, versicolor and virginica. For each variety there are 50 measurements each of sepal length & width and petal length & width.

It is well-known because the correlations of different measurements make it reasonably easy to identify the variety of an iris based just on their relative proportions. This makes it a good test case for machine learning algorithms.

For this exploration you will explore the different correlations.

1. The data in this dataset has no obvious index column (recall that an index column usually has a unique identifier in each row). So first read the data in and check it is in the right format – there should be 5 columns (4 for the measurements and 1 for the variety) and 150 rows. [**Hint**: don't include the parameter `index_col=0` when you read it in.]

2. Create a single visualisation showing scatter subplots for all pairs of measurements. You may find the following line of code useful:
   selected = ['sepal.length', 'sepal.width', 'petal.length', 'petal.width']
   Comment on the strongest correlation, visually.

3. Now confirm your findings by creating a heatmap showing the correlations between all the measurements. Which 3 pairs of measurements are most strongly correlated?

4. Next the aim is to investigate correlations within the different varieties. First of all create a dataframe containing just the data for the Setosa variety with the following line of code:
   data_setosa = data[data['variety'] == 'Setosa']
   Now create a single visualisation showing scatter subplots for all pairs of measurements for Setosa irises. Which is the strongest correlation, visually?

5. A heatmap gives a more precise measure than scatter subplots, so create a heatmap showing the correlations between all pairs of measurements for Setosa irises. Comment on the strongest positive correlation. Is there any evidence of inverse correlations?

6. Repeat item 5 for Versicolor irises and then for Virginica irises.

7. Clearly different varieties of iris have different measurement correlations. It would be nice to summarise this in one single visualisation with different colours for each variety. To achieve this, repeat your code from item 2. However, this time rather than using sub.scatter() to visualise the whole dataset for each pair of measurements, use it repeatedly to visualise each variety. [**Hint**: replace the line of code containing sub.scatter() with 3 lines, each containing sub.scatter(), but each visualising a different variety].

   Your final visualisation should look something like the one below.