Tutorial 05 Distribution

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. If using Colab you need to update pandas & matplotlib for the box plots to work: type "!pip install -U matplotlib pandas" in the first code cell.

Naming: For each exploration you should create a notebook and save it when you have finished. You should name the two notebooks Tut05-A.ipynb and Tut05-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 deviations and distributions of the page hits data. 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 visualisation showing box plots for the 2 high volume pages. Comment briefly on the distribution of hits per day – which is more tightly distributed?

2. Using the box plots in item 1 to determine max and min x limits, create a single visualisation showing histogram subplots of the high volume pages. You should use the guidelines in the lecture to decide on the bin width (and therefore how many bins) and ensure that the 2 subplots share the same x and y limits. Comment briefly on what these show.

3. Now create a visualisation showing box plots for the 8 medium volume pages. Comment briefly on the distribution of hits per day – do the medium volume pages have similar distributions?

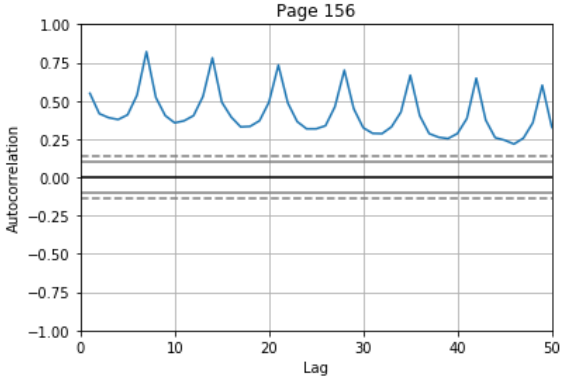
4. Again, using the box plots in item 2 to determine max and min x limits, create a single visualisation showing histogram subplots of the medium volume pages. Again use the guidelines to decide on the bin width (and therefore how many bins) and ensure that the 8 subplots share the same x and y limits. Comment briefly on what these show – does this confirm your comments in 3 above?

5. Now create a single visualisation showing line subplots for all high and medium volume pages. Can you detect any quarterly (or other large-scale) seasonality?

6. In fact the company strongly suspects that some of the medium volume pages exhibit weekly seasonality. Check this by producing an autocorrelation plot for each medium volume page.

7. For those pages which do appear to exhibit weekly seasonality, limit the x axis to produce a zoomed in autocorrelation plot and comment on the seasonality.

   One of your plots should like something like the one to the right.



Exploration B

Once again the second exploration looks at the Iris dataset (see Tutorial 04 for more details).

This week you will look at the distributions of the data.

1. As last week, 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.

2. Create 3 separate box plot visualisations, one for each variety, showing the distribution of the 4 measurement types (so there should be 3 visualisations each with 4 box plots).

   In order to do this, you will need to create a dataframe for each iris variety using the code in last week's tutorial. You will also need to select the columns to turn into box plots, i.e.:

   ```
   selected = ['sepal.length', 'sepal.width', 'petal.length', 'petal.width']
   ```

   If you can, create a loop which generates these 3 visualisations; if not, duplicate the code.

   Which two varieties of iris are most similar in terms of their distributions and why?

3. Now create 3 separate histogram visualisations, one for each variety, showing the distribution of the 4 measurement types all superimposed on the same plot (so again there should be 3 visualisations each with 4 histograms). See the picture below for an example.

   None of the lecture examples create visualisations like this so the code you need is as follows: assuming you are using a dataframe called `data_setosa` and have `selected` the four columns as above, then use:

   ```
   for label in selected:
       plt.hist(data_setosa[label], bins, alpha=0.5, label=label)
   ```

   (Obviously you will need to construct the bins and also set up a matplotlib figure with a title, etc.)

   Note that the alpha parameter makes the histogram columns partially transparent so they can be superimposed.

   The 3 plots should share the same bin widths and limits for the x and y axes so that they are easy to compare. Again, if you can, create a loop which generates these 3 separate visualisations; if not, duplicate.

   Finally comment on how the histograms confirm your answer in item 2 above and also which variety has the smallest spread of measurements. One of your plots should like something like the one below.