Tutorial 09 Clustering

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 Tut09-A.ipynb and Tut09-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 well as the data explored over the last few weeks (https://tinyurl.com/ChrisCoDV/Pages/DailyHits.csv), there are a also a number of other data files in the same folder. All of these files show data averaged or totalled over the year under investigation:

- PageExitRate.csv shows, as a percentage, the average exit rate. For each page, this is the percentage of visitors who leave the site from that page (e.g. by clicking an external link or closing the browser tab).
- PageRevenue.csv shows the total revenue raised by serving adverts from each page.
- PageSize.csv shows, in Kilobytes, the size of the html content for each page.
- PageSpeed.csv shows the average time, in seconds, it takes for users to download each page.
- PageViewingTime.csv shows the average time, in seconds, that users spend looking at each page.

These page metrics are typical of website data and, for example, can be obtained for a site that you own using Google Analytics.

As in the lecture examples, the company wants understand similarities between these metrics and between the total hits for each page (which can be obtained by summing the values in DailyHits.csv) by clustering.

1. First read all the data in and compile a summary dataframe, similar to the lecture examples but changing the folder to "Pages" and the file names to the ones above. The dataframe should have 6 columns / variables (one for each file above and one for the sum of daily hits) and 167 rows (one for each page). [**Hint**: you should have already created the same dataframe for tutorial 6.]

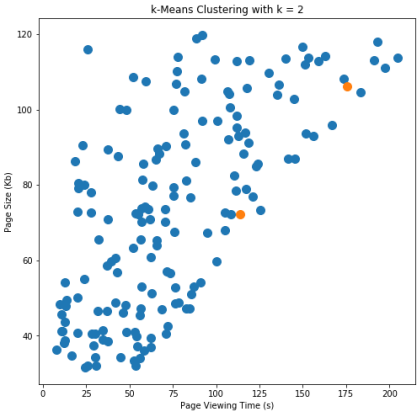
2. Use the elbow and silhouette methods to estimate the optimal number of clusters for the summary data across all 6 dimensions / variables. Do you get the same answer for both methods? [**Hint**: to use all variables you just need to set `selected = summary_data.columns`.]

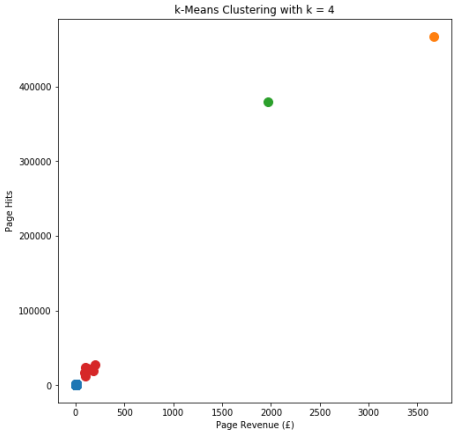
3. Using the k-means algorithm with k = 2, cluster the summary data and then view the clusters by creating a scatter plots of page revenue against total hits. What interpretation can you draw from how the clustering has segmented the data? [**Hint**: again use all variables with `selected = summary_data.columns`.]

4. Again cluster the summary data with k = 2 but this time create a scatter plots of viewing time against page size. Clearly it is not possible to interpret the clustering from this plot, but how does the clustering help you locate certain types of page in this view?

5. Now using the k-means algorithm with k = 4, cluster the summary data again and then view the clusters by creating a scatter plots of page revenue against total hits. What interpretation can you draw from how the clustering has segmented the data this time?

Two of your plots should like something like the ones below.





Exploration B

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

This week you will look clustering the dataset.

1. As previously, 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. Next, use the elbow and silhouette methods to estimate the optimal number of clusters across all 4 measurements. [**Hint**: use

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

   to select the 4 measurements.]

   The silhouette method should suggest 2 clusters and elbow method suggest 3. Given the results from tutorial 4 why might you expect either of these numbers? [**Hint**: even if you haven't done tutorial 4, exploration B yet, you can still see the final visualisation at the end of the tutorial sheet.]

3. Now using the k-means algorithm with k = 3, cluster the data and then view the clusters by creating
   a. a scatter plot of sepal length against sepal width
   b. a scatter plot of sepal length against petal length

   Comparing these plots with the known segmentation shown in the final visualisation at the end of tutorial 4, what can you conclude about k-means clustering for this dataset?

   [**Note**: if you have completed tutorial 4 you may adapt your code to show the clusters rather than the dataframe for each variety; in this case your visualisation will contain 6 subplots rather than two plots.]

   Your final plots should like something like the ones below: