Python & Qlik Assignment

By Nathan Carney and Tom Grant

Datasets:

The three datasets that were chosen for this assignment were:

- VSA 41 Marriages Registered by Year, Region of Cermony and Statistic
    - Found at: 'https://ws.cso.ie/public/api.restful/PxStat.Data.Cube_API.ReadDataset/VSA41/JSON-stat/1.0/'
- SIA 51 Income and Poverty Rates by Region, Year and Statistic
    - Found at: 'https://ws.cso.ie/public/api.restful/PxStat.Data.Cube_API.ReadDataset/SIA51/JSON-stat/1.0/'
- VSA 15 Births (Number) by Age Group of Mother, Year and Region
    - Found at: 'https://ws.cso.ie/public/api.restful/PxStat.Data.Cube_API.ReadDataset/VSA15/JSON-stat/1.0/'

Methodology Script 1:

Using Tkinter a Graphical User Interface (GUI) a file explorer was used to allow the user to choose a file in the formats json, json-stat and csv. When the user selects a file the program saves that file's filepath as the variable 'filepath.

When the file has been chosen filepath is no longer 'None'. This causes the code to run through a series of try, except statements, The first attempts to read the chosen file in as a csv file. If this fails the program then tries to process the file in the format typical of the json files on data.gov.ie. If this fails the program then tries to read the file in as files oriented 'index', or 'column'. Lastly the program tries to read in the file as a json stat file, it does this using the pyjstat library.

The final try, except statement attempts to convert the dataframe 'dataframe' to an excel file. If any of the previous attempts to read in the data have been successful and the file will be saved as 'exported_data.xlsx'. If not the program will return an error message. This error message informs the user that the program was unable to open the file.

Methodology Script 2:

These three datasets were loaded into the python environment using the URLs listed above. The datasets were then converted into the pandas dataframe format. Then the datasets were subsetted to the years which were common to all. The column 'Region of Cermony'

was changed to ‘Region’ to match with the other two datasets. Lastly all three datasets were transformed pivot tables with the two indexes ‘Region’ and ‘Year’. This step served to reduce data duplication within the datasets.

Then the three datasets were joined on their indexes. The final dataset was called ‘full_data’ this dataset was then exported to an excel spreadsheet as it was detailed in the assignment that the dataset should be saved in a suitable format.

The dataset was then loaded into plotly and drop down menus were used to allow the user to choose a ‘Region’ and a ‘Statistic’. This displays two graphs, the first is the values for this statistic which have been observed in the user’s chosen ‘Region’, the second is an overview of the chosen ‘Statistic’ in all regions in the most recent year in the dataset.

Difficulties Script1:

The difficulties which were experienced during the writing of Script1.py were mainly due to the various formats which json can be saved in. After researching possible solutions online and multiple attempts, the json_normalize function was used as it offers the functionality of automatically reading in json formats without specifying the keys.

The other problem experienced during the writing of Script1.py was that multiple libraries exist for dealing with json-stat files. After multiple attempts to load jsonstat.py it was not possible to download. The pyjstat library was then chosen and used to load any json-stat files.

Difficulties Script2:

The difficulties which arose for Script 2 were pivoting the datasets and using dropdown menus to choose the criteria for the graphs to display. The first was done after a while was spent considering how best to join the datasets in a manner similar to the movies dataset. It was decided that the best way to do that was to create a unique key for each entry which was done by converting ‘Region’ and ‘Year’ to indexes.

The dropdown menu problem was solved by searching for a similar entry on Stack Overflow. There were no exact matches so multiple solutions were examined and parts were drawn from all of them in order to finally lead to the final solution. The solution allows the user to choose a ‘Region’ from a drop down menu and a ‘Statistic’ from a dependent drop down menu. This displays two graphs, the first is the values for this statistic which have been observed in the user’s chosen ‘Region’, the second is an overview of the chosen ‘Statistic’ in all regions in the most recent year in the dataset.