Data Science Portfolio Project - UK Government Companies House Data

Software Requirements Specification

Author: Alan-Francis Kirby

Date of Original: 28/02/2023
Date of Latest Update: 4/03/2023

Introduction

This project is a combined data analyst and data scientist portfolio project that will access data on UK businesses via the UK governments Companies House API and perform various levels of summary and analysis. The purpose is to provide insight into the distribution of businesses throughout the UK based on geography, industry sector, profitability and the age of the company. This will be achieved by using data visualisation methods to highlight the key content within the data, providing the data analyst aspect of the portfolio project, showing the *what is it* part of the analysis. The same data will then be analysed in combination with other data sources for metrics such as population density using a variety of standard data science methods, to look for relationships and provide the *why is it* aspect of the analysis.

The end product will include a web based application that a non-technical user could use in order to obtain company data for their regions of interest. The data science analysis will then be summarised within the same web based application, ensuring a user friendly experience through a graphical user interface rather than requiring familiarity with how to use code. The summary data visualisation will be accessible via a report in PowerBI. This project is intended for use by non-technical users.

A short video overview of the project and its current state and next steps can be found on the projects README page.

Document Conventions

Terminology	Acronym
Software Requirements Specification	SRS
Application Programming Interface	API
Graphical User Interface	GUI
Amazon Web Services	AWS
Standard Industrial Classification	SIC

Table. 1: Conventions Used Within This Document.

Useful Resources

All required information on how to use the Comapies House API can be found at their official Companies House Development Portal. The SIC codes are available from the office for national statistics here.

The key components of the project are listed here within the SRS, however the road-map detailing prioritisation and which components are currently being worked on are outlined within the project wiki. The wiki also details how to get started with using the project and how the code is structured. To see any outstanding issues or to report issues and provide feedback, please make use of the projects Github Issues page.

Requirements

The project will use Python as the main coding language through which to interact with the Companies House API and with which to perform the analysis. The data visualisation will be performed using PowerBI, with a star schema for the data model. The data will be stored in a AWS cloud

database. The GUI will be constructed using either django or Flask, and the remote computing for the web-based application will be performed using a cloud computing service provided by AWS. Iterations of the project will be denoted using semantic versioning, with the version control being handled with git and Github being used for the remote repository. A single master branch will be utilised in accordance with trunk based development, with any incomplete features being disabled using feature flags.

Key Functionality

Item 1 - Identify companies for a given location

The user must be able to select a location, such as Swansea, and correctly obtain all active companies that are registered at that location.

Item 2 - Get key information on each company

The user must be able to obtain all relevant data, such as company age, revenue, etc, for all of the identified businesses.

Item 3 - Ability to save into cloud database and also as csv

The final application will work from the cloud, and must therefore have no local dependencies. Thus the company data must be saved into the cloud database of choice. A copy of the data should also be saved as a flat csv file as a backup, for those cases where a long data gathering run has completed but an issue has occurred with the cloud database service.

Item 4 - Get SIC codes to identify industry sectors

In order to analyse how different industries are represented within the UK economy, the standard industrial classification (SIC) codes that are returned by the Companies House API must be mapped to the relevant industry activity. These codes and their related business activities will be written into a flat file and loaded into PowerBI.

Item 5 - Create GUI and make accessible in web browser

The initial code will be written in Python and accessed via Jupyter notebook, however the project must then be updated to provided a GUI for use by a non-technical user. This should be kept simple for the end user, and be accessible in a web application, thus requiring no downloading and installing of software on their local machine.

Item 6 - Data science analysis

The project will have thus far involved developing the code for gathering the data and storing the data, putting the data into a web-based application with a user friendly GUI, and performing the data analyst type of data visualisation in PowerBI. Once these steps are complete the more sophisticated data scientist analysis should then be performed.

Item 7 - Obtain extra data (pop density etc)

After the analysis of the data from Companies House has been performed, this should be extended by obtaining other data sets on important metrics such as local population demographics, population density, local population health and education levels, etc, and then further data scientist analysis should be performed. This data can be taken from other stand alone data science projects.

Item 8 - Prep for publication

One of the key long-term objectives is to summarise the economic history and situation for a local area, and to turn this into an engaging and interesting piece of content for the local public for the region of interest. This would then hopefully be published in a local magazine, newspaper, or even simply written into a blog or written up on the *towards data science* blog.