GEOG5995 Programming for Social Scientists: Core Skills

Independent project software documentation

The intention of this software was to develop some code that may be relevant and useful to my PhD research, i.e. looking at social and psychological influences on people's dietary choices. For my research I will be working with a large panel survey data set from YouGov. As I will mostly be working with survey data, it made sense to use this assessment as an opportunity to gain familiarity with the Pandas software library for the purpose of data exploration and statistical analysis. I chose the British Social Attitudes survey from 2014 (NatCen Social Research, 2014) as a data source for this project as it is freely available and features a great deal of demographic information as well as a range of questions about meat-eating behaviour and attitudes among the British public (these questions were sponsored by the partner organisation for my PhD project, The Vegetarian Society). Linear and logistic regression were chosen as appropriate statistical techniques for examining relationships between variables.

The software design process was a trade-off between developing some code that could be potentially useful for my PhD project and developing something manageable and functional within the specified timeframe given the limitations of my knowledge of survey data analysis and processing using Python.

Once a suitable data source had been identified, I investigated how to open it within Python using the Pandas library, and then carried out some basic descriptive statistics, originally using the Matplotlib library. After examining the shape of the data and the distribution of variables, I wanted to undertake some form of analysis. At first, I intended to carry out a Random Forest

Classification using the Sci-kit Learn library, as a way of examining which variables were the strongest predictors of certain dietary choices or attitudes. Unfortunately, this proved to be problematic without copying large chunks of code from other sources which left little room for original contribution. Multivariate linear regression was chosen as a suitable compromise for this, and the Seaborn library was found to have good functionality for this purpose. Logistic regression, also available as a function within Seaborn, was then applied across different variables to investigate relationships of probability within binary outcomes.

In order to utilise these methods, it was necessary to overcome several kinds of data processing errors. Largely, these were caused by a variable's data type not being correct for the process I was attempting, such as a string object needing to be converted to a float to enable numerical analysis. Other issues were caused by the coding of the source data leading to problems of missing values, such as a '9' indicating non-response rather than a particular value on an ordinal scale. This required the development of functions to recode specific values within certain variables as np.nan (a floating-point value that represents non-numerical values). This was particularly important for variables I was interested in using as binary independent variables for logistic regression, whereby zeros and ones represented Boolean responses.

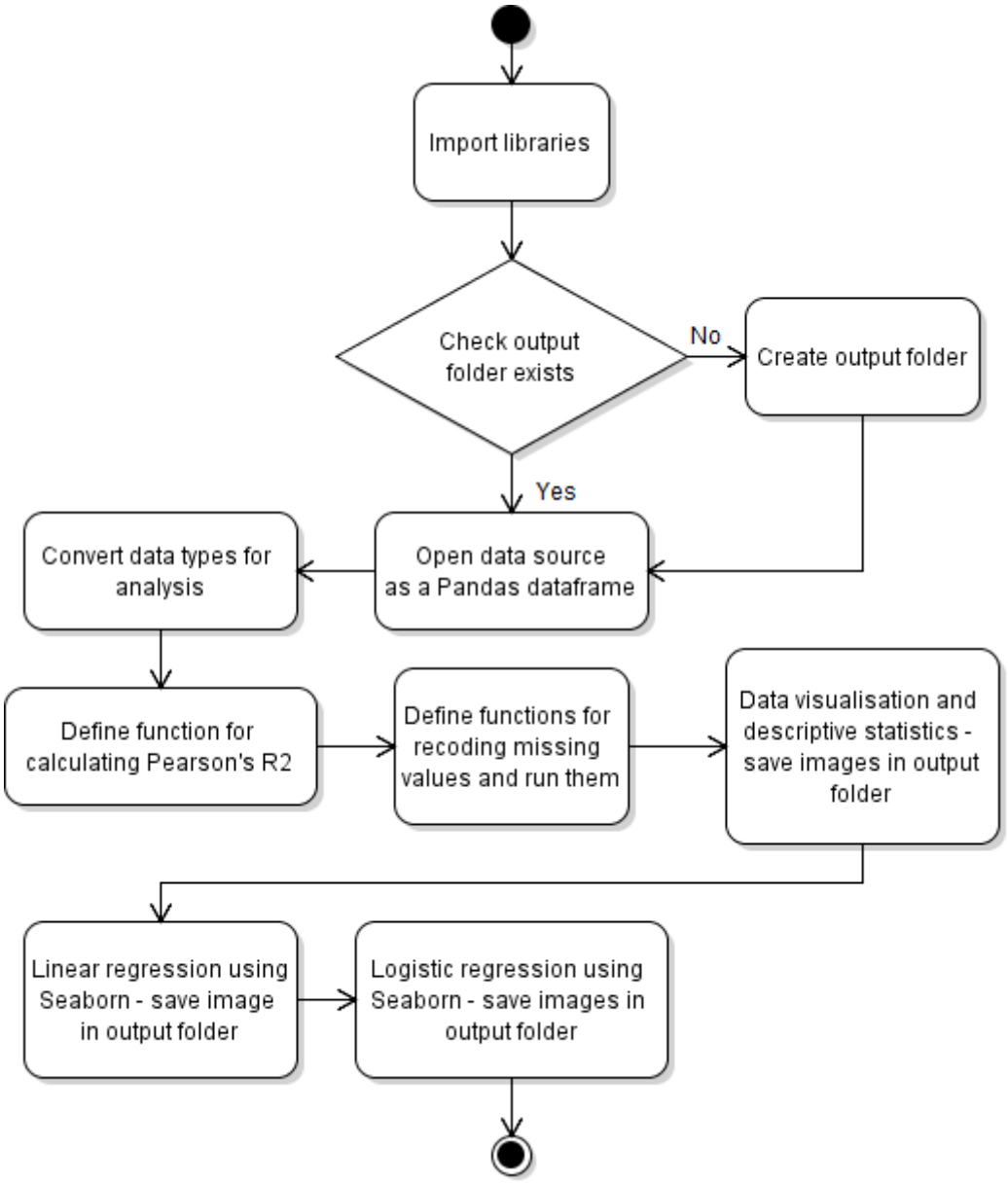
The documentation for the Python library Seaborn was consulted extensively during the software development process (Waskom, 2017). Another useful source of information was Downey (2014), *Think Stats: Probability and Statistics for Programmers*, as it contained details of methods for recoding variables within Python. These sources are fully attributed within the code for this assessment, which is available freely at https://github.com/ChrisDNewton/GEOG5995_Assessment2.

After working solutions were implemented for data processing and analysis, two internal tests were incorporated to account for errors that could occur when the program was used on different systems. A test was introduced for creating a folder for image outputs in the working directory if one was not already present, and a try/except loop was written to check that the correct data file was present within the working directory, with instructions printed on where to obtain it if not. These were tested by removing the file or folder in question from the working directory and running the main program file.

For further development, I would like to explore the creation of new data frames by selecting specific variables from the original file, as this could be more efficient in terms of memory use. Developing functions for carrying out factor analysis and classification, looking at which variables might be the strongest predictors of certain dietary behaviours or attitudes, could also be useful in that they could help address issues of multicollinearity across variables. I would also like the explore the Bokeh library's capability for generating data visualisations in HTML format for use on webpages.

(745 words)

UML activity diagram



References

Downey, A.B. (2014). *Think Stats - Exploratory Data Analysis in Python*. Green Tea Press. [online]. Available at: http://greenteapress.com/wp/think-stats-2e/

NatCen Social Research (2014). *British Social Attitudes Survey* [computer file]. 2nd Edition. Colchester, Essex: UK Data Archive [distributor], May 2016. SN: 7809, http://dx.doi.org/10.5255/UKDA-SN-7809-2

Waskom, M. (2017) Seaborn: Statistical Data Visualization. Available at: https://seaborn.pydata.org/ (Accessed: 12 November 2017)