IFN509 Assignment 2: Project

Assignment type: Project (applied)

Topics: Data preparation (cleaning, integration, transformation), data analysis and mining.

Weight: 25%

Group or Individual: You will work on this assignment in groups of two or three. Note that only a single submission is required from each group, however make sure the report submitted contains the name and student number of each student in the group.

Due date: Sunday June 2nd, 23:50pm (end of week 13)

Driving question

How does weather affect air quality in Brisbane, Australia?

Data provided :

southbrisbane-aq-2018: Daily weather observations in various cities in Australia from July 2008 until March 2019[1].

weatherAUS.csv: South Brisbane (South East Queensland) 2018 hourly air quality and meteorological data[2].

Analysis required:

a) Investigate whether there are any direct correlations between air quality indicators and either rain, humidity, wind, or temperatures. Explain what the correlations mean (you may use visualization if you wish).
b) Use decision trees to see if at least one of the average daily air quality indicators can be predicted on the basis of any or all of the weather indicators provided. Explain what patterns the best decision tree highlights.
c) Cluster the days and demonstrate through visualization how the clusters are organized. Explain what patterns they reveal.

[1] Observations were drawn from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data. Copyright Commonwealth of Australia 2010, Bureau of Meteorology. Available as a package at https://rattle.togaware.com/weatherAUS.csv

[2] Environment and Science, Queensland Government, South Brisbane (South East Queensland) 2018 hourly air quality and meteorological data API, licensed under Creative Commons Attribution 4.0 sourced on 16 May 2019. Available at https://data.qld.gov.au/dataset/air-quality-monitoring-2018/resource/f28488d1-44fc-4fda-aeff-291039d30f70

Preparing the data:

a) Investigate the quality of the datasets, clean the data where required (address missing data and outliers). Explain what decisions you made for cleaning the data and why.
b) Integrate the two provided datasets. You may use manual methods, or external sources to do so. Explain your approach[3].
c) Provide the final, clean dataset you have used.

Software:

Explain which technology/ies you have chosen to use for this project. Explain the limitations and benefits of your approach. If you are not using MySQL, R, please consult with the unit coordinator for approval.

Submission Requirements

You are to submit the following files:

1. Through the Blackboard link: a) CSV files containing your clean data that you used for analysis (***import1.csv***, ***import2.csv*** *etc.*) b) Source code. You will need to compress several files into a .zip file.
2. Through the Turnitin link: An 8 pages (maximum) report which contains the following sections (***report.pdf***):
    a. Title, Group members (including student numbers)
    b. Description of how software was used (1/2 page)
    c. Data Summary (quality, cleaning, preparation, integration) (2 pages)
    d. Correlations (1 page)
    e. Decision Tree (1.5 pages)
    f. Clusters (1.5 pages)

[3] You can proceed with the rest of the assignment without integrating the two datasets, and only work with the air quality dataset. If you chose to do so, you will lose marks in the data quality analysis/and data preparation sections of the marking scheme, but not in other sections.

Marking Scheme

Marking is based only on the report, code and dataset are for verification purposes only. ß

Task	Marks	Criteria
Explain choice of technology	3	- The student described the software they used in the report : The student described how they used software in the report, and this evidences that the packages were used effectively (e.g. MySQL might be used for complex joins, Excel for particular graphs, R for association analysis, another program for a bubble plot etc.) - The student critically reflected on the benefits and limitations of their chosen approach
Data quality analysis	3	- The student explains and illustrates the steps required and applied to prepare the dataset. - Data quality issues are all considered and discussed appropriately.
Data preparation	4	- The student correctly cleans the data provided so that the entire set can be imported into a statistical package for analysis - The student organizes data in a way that is sensible for the software they use - Redundant data is removed - Data types are correctly applied - Outliers are removed using a commonly accepted rule - Methods for cleaning and outlier exclusion are described in the report
Visuals/summary for correlations	2	The student was able to represent the effectively data to explain.
Correlations	2	The student appropriately identified trends in the data.
Decision tree	2	The decision tree produced supports the analysis required
Analysis of decision tree	2	The interpretation of the decision tree is correct and well explained
Visualisation of clusters	2	- The visualization provided is effective to illustrate the nature of the clusters. - A variety of layers are used appropriately.
Analysis of clusters	2	Appropriate conclusions were drawn based on the data set.
Presentation	3	- No spelling or grammatical errors were found in the report - The report is presented professionally - Files are appropriately titled