Tracey Wright

N9131302

IFN509

Data Manipulation Assignment

# Software

The entire data manipulation process was carried out using R and R Studio. Other technologies such as SQL Server (approved by course convenor) were considered, but it was decided that R and R Studio were appropriate for the entire project because:

1. There isn’t anything that R and R Studio wouldn’t be able to do for this analysis;
2. The datasets are relatively small, therefore any performance considerations associated with executing functions such as joins/complex filters in R and R Studio rather than SQL Server will be unlikely to be impactful to productivity or results;
3. Implementing solely in R and R Studio avoids any overheads associated with data compatibility between technologies and moving data between technologies, therefore optimising developer productivity;
4. The simpler architecture of using one data manipulation technology lends itself more to expansion of this project into a web application, which - given that I’m a professional web app developer - I may do at some stage (although the code would require further refactoring for this).

All functions were performed using a script in R Studio. The script is intended to be run sequentially from top to bottom. The script does everything from data exploration to data processing, analysis and visualisations.

The “rpart” package was used for decision tree implementations. “Cluster” and “factoextra” packages were used for k-means implementations.

MicroSoft Word was chosen for the report over R Markdown, as a means to increasing productivity during report writing. However, it’s noteworthy that R Markdown may have been a good choice if there wasn’t quite as much word processing required for the report.

Tables were written to csv file, then copied into Word and formatted using Word tools for the report. Figures were exported from R Studio into Word.

# Data Pre-Processing

The data provided is comprised of an Australia-wide dataset with daily weather observations and a South Brisbane dataset with hourly weather and air quality observations.

## Data Quality

Initial data exploration was performed by visualisation via histogram and by filtering and sorting of the raw data in R Studio.

It’s noteworthy for data quality purposes that there are missing and out-of-range values for many of the variables. Missing values will be imputed with the most likely values prior to running the data analyses. Further, there are out-of-range values for some of the air quality indicator variables. All options considered, it was decided that in order to preserve the consistency of natural variation throughout the dataset, these out-of-range values would remain in the hourly dataset as they were found to translate to within range during the aggregation by averages of hourly data to daily data.

## Integration

Given that the target variables are air quality and that the Australia-wide dataset doesn’t include air quality observations, we can filter out locations other than Brisbane from the Australia-wide dataset in preparation for dataset integration, as they won’t be useful for achieving the objectives of this analysis.

The two datasets were integrated by joining on dates, given that the observations from both datasets are dependent on time. Dates from each dataset required reformatting to provide consistency for the join/merge. The resultant integrated dataset was comprised of 365 unique date observations, each with 24 observations for the hours in each day.

## Imputation

Measures were taken on a per-variable basis to ensure variable consistency between the datasets during imputation. Options for imputing NA values using decision trees and correlations to determine the most likely value were explored.

Temperature variables, MinTemp and MaxTemp were imputed with consideration to a trend found with “Air.Temperature..degC.” whereby after August there is an approximately 5 degree difference between variables.

Rainfall was imputed with the most common value, zero, as only loose trends were found between this and other variables.

Sunshine and Evaporation each had just one NA value, and Wind Speed and Relative Humidity just a few each. These variables were explored for trends and imputed with what was determined to be the most likely value and/or a value that would be unlikely to disrupt any underlying trends in the data.

## Feature Selection

After integration, the merged data included redundant weather variables. Weather features were selected based on the following overarching objectives:

1. Alignment with the report objectives;
2. Omitting/merging redundant columns;
3. Dimensionality reduction (simplification of the dataset).

Temperature variables, MinTemp and MaxTemp were selected and all other temperature variables excluded.

Wind direction variables were simplified to one daily variable, based on a method which averages this mathematically-circular-based variable using trigonometry (Monforte, n.d.), for a more accurate representation of this variable.

Wind speed, wind direction, and relative humidity - originally from the South Brisbane dataset were selected from the hourly dataset and all other wind and humidity variables were excluded.

Daily variables originally from the Australia-wide dataset were selected to represent Cloud, Evaporation, Sunshine, and Pressure and all other afore-unmentioned redundant daily variables were excluded.

All air quality variables were included in the cleaned/processed dataset as they were not redundant and will all be required during the data analysis.

## Aggregation

The option was left open for either using daily or hourly data for the decision tree analysis. The hourly data was subjected to data pre-processing, then aggregated on dates using a mean function to produce one observation per day or leave the dataset at hourly observations to potentially facilitate a more precise analysis (which turns out it didn’t).

Since categorical variables don’t lend themselves well to aggregation, wind direction categories were re-applied after the aggregation.

## Processed Data

The daily processed data (which after the analysis was determined to be the preferred data set) comprises of 365 observations of weather and air quality indicator variables. There are no NA values. The cleaned data file is named “cleaned.csv”.

# Correlations

This section investigates whether there are any direct correlations between air quality indicators and either rain, humidity, wind, or temperatures and explains what the correlations mean.

## Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *N.Oxide* | *N.Dioxide* | *N.Oxides* | *C.Monoxide* | *PM2.5* | *PM10* |
| MinTemp | -0.58 | -0.77 | -0.67 | -0.57 | -0.37 | -0.12 |
| MaxTemp | -0.47 | -0.53 | -0.5 | -0.43 | -0.08 | 0.08 |
| Rainfall | -0.03 | -0.13 | -0.08 | -0.04 | -0.14 | -0.15 |
| Relative.Humidity | -0.08 | -0.14 | -0.12 | -0.15 | -0.07 | -0.24 |
| Wind.Speed | -0.28 | -0.42 | -0.36 | -0.42 | -0.29 | 0 |

All oxides and particulate matter units are micro grams per meters cubed (μg/m3).

Wind Direction has been omitted from this analysis as it is a categorial variable, which is not supported by Pearson correlation.

## Analysis of Results

For this analysis:

“Oxides” refers to nitrogen oxide (N.Oxide), nitrogen dioxide (N.Dioxide), nitrogen oxides (N.Oxides) and carbon monoxide (C.Monoxide).

“Particulate Matter” refers to particulate matter 2.5 (PM2.5) and particulate matter 10 (PM10).

There are negative correlations between oxides and temperatures (MinTemp and MaxTemp). PM2.5 has a weak negative correlation with MinTemp, and the other particulate matters have little to no correlation with temperatures. This means that, in general, when there is an increase in MinTemp there is a decrease in the Oxides and PM2.5, and vice-versa.

Rainfall correlation results indicate that there is little to no correlation between rainfall and the air quality indicators. This means that there has been no linear trend detected between rainfall and the air quality indicators.

Relative Humidity results indicate little to no correlation with oxides and PM2.5, and maybe a very weak negative correlation with PM10. This means that there has been little to no linear trend detected between relative humidity and the air quality indicators.

Wind speed results indicate a weak to moderate negative correlation with the oxides and PM2.5, and no correlation with PM10. This means that when there is an increase in wind speed there is a small to moderate decrease in oxides and PM2.5, and vice versa. There has been no linear trend detected between wind speed and the PM10.

# Decision Trees

The general approach towards implementing the decision trees was to group air quality indicator values into levels as per national standard air quality information derived from the Victorian EPA (Victorian Environmental Protection Agency, n.d.), whom states that it derives this information from the National Standards (but formats it much more nicely). This worked well in terms of providing an additional dimension to the results for interpretability and usefulness.

The decision tree analysis was also implemented with hourly data, and it was found that this was not any more accurate than the daily data set. This is probably because many of the variables are averaged anyway, particularly the variables which produced the most information gain.

All decision trees were trained on 70% of the data, and the models were validated on the remaining 30% of the data. Data was randomised prior to input into the algorithms.

The use of minimum splits was investigated, but it was found that this didn’t significantly change the results.

Only PM10 and PM2.5 provided results, as carbon monoxide and nitrogen dioxide values were all grouped into the “Very Good” class which broke the library (and wouldn’t have provided insightful results anyway). For nitrogen oxide and nitrogen oxides, no reliable standard values were able to be obtained by research, therefore categories were unable to be determined so these variables were unfortunately omitted from the analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Fair* | *Good* | *Poor* | *Very Good* |
| Fair | 0 | 3 | 0 | 1 |
| Good | 0 | 25 | 0 | 10 |
| Poor | 0 | 1 | 0 | 1 |
| Very Good | 0 | 5 | 0 | 64 |

**Table 2: Particulate Matter 10 μg/m3** Approximately 80% of the test data has been classified correctly by the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Fair* | *Good* | *Poor* | *Very Good* |
| Fair | 0 | 2 | 0 | 5 |
| Good | 0 | 12 | 0 | 10 |
| Very Good | 0 | 3 | 0 | 79 |

**Table 3: Particulate Matter 2.5 μg/m3** Approximately 83% of the test data has been classified correctly by the model.

These models could be used to predict air quality for unseen data and would correctly classify the outcome approximately 80% - 83% of the time.



**Figure 1: Particulate Matter 10 μg/m3 Decision Tree** A graphical representation of the factors contributing to PM10 air quality. The highest information gain is derived from the MinTemp variable.



**Figure 1: Particulate Matter 2.5 μg/m3 Decision Tree** A graphical representation of the factors contributing to PM2.5 air quality. The highest information gain is derived from the MinTemp variable.

The data partitioning rules are transparent for decision trees, as seen in these graphs, which can easily be translated to code during software development.

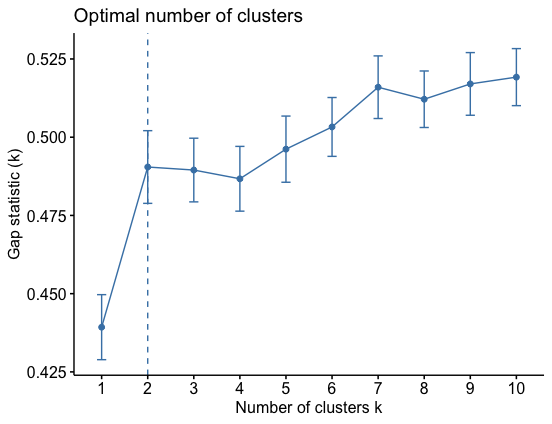
# Clustering

Variables were scaled based on z-scores prior to implementing the clustering algorithm.

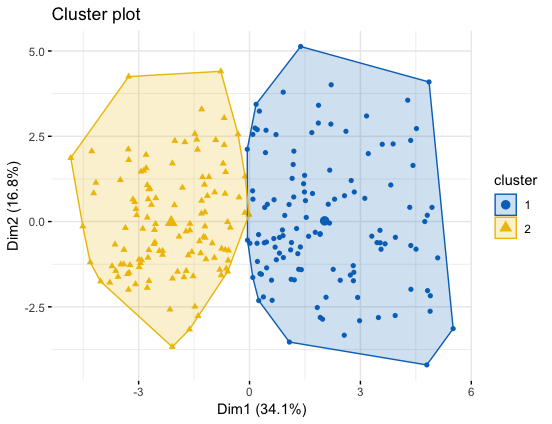
The k in k-means refers to the number of centroids or clusters. To start the process, we can either arbitrarily select a value for k or use another algorithm to compute a suggested value for k.

Next, the k-means algorithm selects the next closest value from the dataset to the centroid/s and assigns that value to the cluster. Each time a new value is assigned to a cluster, the centroid is re-approximated to the average of the updated data points for each cluster. This continues until all data points have been assigned to a cluster.

Prior to running the algorithm, outliers were removed from the data (any value with a z-scale value of < negative 2 or > 2) as the mean algorithm of the centroid is sensitive to outliers and can cause the rest of the data to be bias when choosing data points to form the cluster.



**Figure Number of Clusters (k):** Two clusters are optimal for this dataset according to the algorithm, although we could have also chosen three as the gap statistic doesn’t differ considerably.



**Figure Cluster Graph:** There is not a great deal of separation between the 2 clusters, however, there is some (very little) aggregation of data towards the center of each cluster and a (very) small amount of definition around the borders. It’s arguable that there may be just one cluster for this data, which would indicate that the trends within and between the clusters are probably not particularly distinct.

To gain insight from these results, or from k-means results in general, further investigation of any trends within the clusters or between the clusters could be performed. However, this is beyond the scope of this assignment.

# Bibliography

Kassambara, A. (2018). *Types of Clustering Methods: Overview and Quick Start Code*. Retrieved from Types of Clustering Methods: Overview and Quick Start Code: https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/

Monforte, F. A. (n.d.). *Computing a Weighted Average Wind Speed and Wind Direction Across Multiple Weather Stations*. Retrieved from Itron: http://blogs.itron.com/computing-a-weighted-average-wind-speed-and-wind-direction-across-multiple-weather-stations/

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining (Second Edition).* Minnesota: Pearson.

Victorian Environmental Protection Agency. (n.d.). *Victorian EPA Air Pollution*. Retrieved from Victorian EPA: https://www.epa.vic.gov.au/your-environment/air/air-pollution