Queensland University of Technology

School of Information Systems

Science and Engineering Faculty

IFN 509 Data Manipulation

Driving Question:

How does weather affect air quality in Brisbane, Australia?

Student ID & Name:

N10105981 Wang Hsuan,

N10411186 Jen-I Wen,

N7462875 Tsung-Han Wu,

Lecturer:

Dr Laurianne Sitbon

1. Choice of technology

Data mining is the process of discovering patterns in large data sets, dealing with the real world data, involving the method of machine learning, statistics, and database systems. The main techniques for data mining include classification and prediction, clustering, outlier detection, association rules, sequence analysis, time series analysis and text mining. With these function requirements, R is used for the project because it meets the demand of all the requirement of data mining.

R is a programming language for statistical computing and graphics, it is being widely use on the area of data science, data mining and statistics and machine learning. In this project, R is used to load the data file of "southbrisbane-aq-2018.csv" and "weatherAUS.csv". R is used to clean the dirty data by removing incomplete data and removing the data which is totally worthless to use in order to meet the aim of this project. After data clean, R is also being used to integrate those two datasets and correlations between each variable are found from this merge dataset. After that, decision tree is being made based on the correlations between variables, and found out how does weather affect air quality in Brisbane.

When we used R, we found out that R has various functions to deal with data, but as the beginner learning R, we cannot use this tool fluently and need more time to learn all the related functions.

2. Data Summary

2.1. Data Quality

The quality of data is measured by things such as completeness of the data, validity, accuracy, consistency, freshness and timeliness. The original datasets provided for this project do not really have a good quality.

Firstly, there are a lot of missing data especially from "weatherAUS.csv", before the datasets are available to be use, missing data need to be remove or replace by some other values.

Secondly, there are some indicators of weather which are provided in both datasets with different measuring value. This might cause some problems with consistency of data because both datasets have indicators of temperature and wind speed. Also, data might be collected from different condition of environment, out of same indicator, a more reliable data need to be choose.

Thirdly, data of "weatherAUS.csv" has the time frame from 2008 to 2019. However, overall weather condition from 10 years ago was different from today. Take Greenhouse effect as an example, if the temperature is rising from ten years ago due to the greenhouse effect, at the same time, air pollution increase due to the population increase within 10 years, value of temperature from 10 years ago might not be likelihood to hold the correlation with air quality for today. Also, dataset of "weatherAUS.csv" and "southbrisbane-aq-2018.csv" have different data from different timeframe, so in order to merge the data, only for those data with overlapping time frame are been used for Analyzing.

2.2. Preparation

After reviewing the quality of dataset, the next step is to start to clean dirty data. Since the driving question is aimed to find out how does weather affect air quality in Brisbane, however, data frame in

"weatherAUS.csv" covers all weather data over Australia, the first thing to do is clean out the data which is not related to Brisbane and then save the data into “Brisbane.weather”.

The next step is to decide which data are going to be removed, there are few choices to handle missing data, such as remove data unknown, or fill in it automatically with some other values, or even just remove the whole variable if the variable itself is meaningless in these analysis.

Start with “Brisbane.weather”, clean the data by using the function from library (Amelia) to deal with missing data, for the reason there is still a lot of missing data from “weatherAUS.csv”. Check on if the variables are usable or not, notice that from “weatherAUS.csv”, variables such as 'Evaporation' and 'Sunshine' have a lot of missing data, but after filter out data outside of Brisbane, only less than 1% of data are missing, and it can easy to be solve my replacing the mean value. So, they are usable.

There is need on merging two datasets for due to the reason that that 'weatherAUS.csv' got weather variables such as humidity, pressure and cloud. On the other hand, 'southbrisbane-aq-2018.csv' got the quality measurement such as Nitrogen Oxide, Nitrogen Dioxide, Carbon Monoxide, PM10 and PM2.5. By merging these datasets, correlations between air quality indicators and weather indicators can be found. However, “southbrisbane-aq-2018.csv” has only the data within 2018, in order to merge it, they need to have the same time frame, in “Brisbane.weather”, only the data within 2018 are selected to use. Also, data such as wind speed, Humidity and temperature are not being chosen to use because the other dataset 'southbrisbane-aq-2018.csv already has it.

"southbrisbane-aq-2018.csv" is been saved as “southBrisbane”, due to the reason that out of two datasets, when there is the data indicate the same thing such as temperature, wind speed and humidity, data from "southbrisbane-aq-2018.csv" is used. As the result, there is not much change for "southbrisbane-aq-2018.csv" during the preparation of dataset.

2.3. Cleaning

For the reason that there are data missing in some of the cases from both dataset, the next step is to do the data cleaning. There are few ways to deal with missing values, Ignore the tuple or Fill in the missing value with the attribute mean, in the case of this project, missing values are filled with attribute mean, so the amount of deviation can be decreased. The first step is to remove missing values NA, such as ‘Rainfall’, ‘Evaporation’, ‘Sunshine’, ‘RainToday’ from “Brisbane.weather’ and ‘Air.Temperature’, ‘Wind.Speed’, ‘Nitrogen.Oxide’, ‘Nitrogen.Oxides’, ‘Nitrogen.Dioxide’, ‘Relative.Humidity’, ‘Carbon.Monoxide’, ‘PM10’, ‘PM2.5’, ‘Wind.Direction’ from data of “southBrisbane. After missing values are being removed, mean value for each variable above are been calculated from the rest of values. After that, mean values for each variable are used to replace missing values.

In the real world data, there always has some data points that are differ significantly from other observations. Those data points are being called as outliers. Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution. An easy way to find out outliner is by Normalize the data, if there is some value are significant different from normalize data, it would be considered as an outlier.

2.4. Integration

When both datasets are cleaned, the next step is to merge two data in to one. Library (data.table) is being used to do the merging. Before that, Date data from two datasets must be the same, they are changed to the format of day/month/year, after that, they are being combine into a Matrix based on

the date they got. In the matrix 'NewWeather', x-axis is "Brisbane.weather", y-axis is "SouthBrisbane.Weather". Correlations between all variable can be found.

3. Correlations

Function ggpairs() in library (GGally) are allowed to use for looking at the relationship between each variable in a dataset, including the correlations between variables, result are shown as figure 2 below. Figure 1 below shows that the relationship strength of the variables based on the correlation coefficient. However, there is too much variables in the figure 2, it makes the data hard to read, so another graph is made with removing all the data which is already exist, correlation with variable itself, and hightlight the correlation coeffcient with stronger relationship strength based on figure 1 below. those variables with stronger relationship strength.

TABLE 7.1

Correlation Coefficient for a Direct Relationship	Correlation Coefficient for an Indirect Relationship	Relationship Strength of the Variables
0.0	0.0	None/trivial
0.1	−0.1	Weak/small
0.3	−0.3	Moderate/medium
0.5	−0.5	Strong/large
1.0	−1.0	Perfect

Figure 1 Table of Correlation Coefficient Values and Their Interpretation

Taken from "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach". [1]

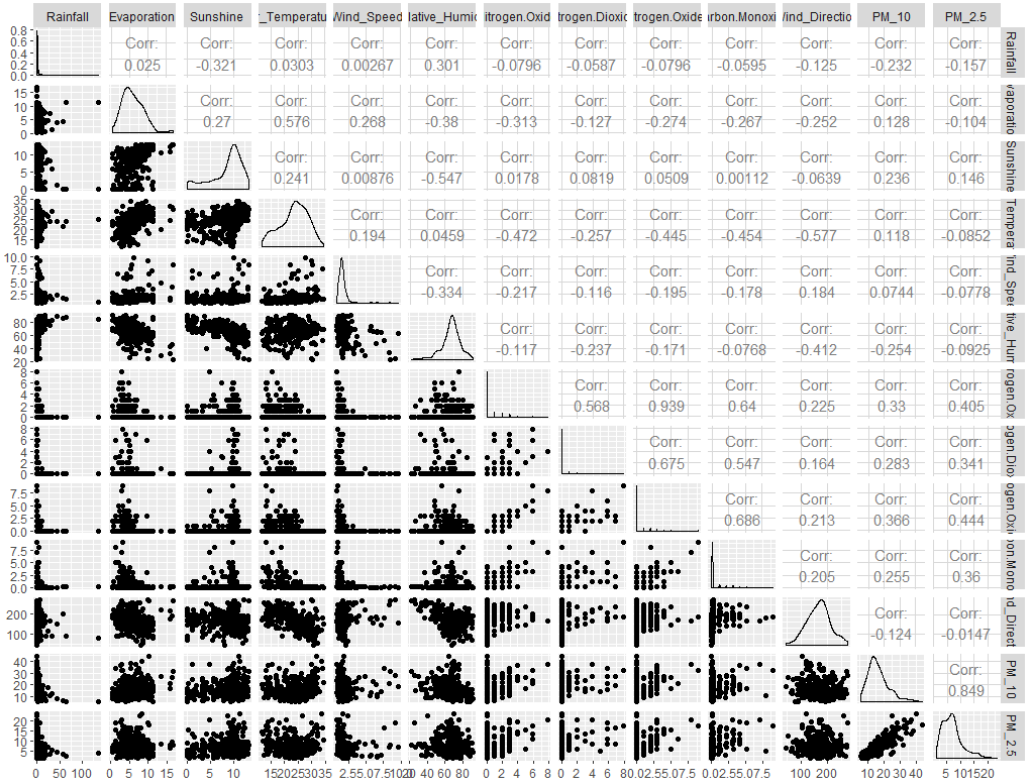

Figure 2 Correlation table between variables

	Nitrogen.Oxide	Nitrogen.Dioxide	Nitrogen.Oxides	Carbon.Monoxide	PM_10	PM_2.5
Rainfall	-0.07958369	-0.05874077	-0.07963485	-0.059450427	-0.23217840	-0.15708550
Evaporation	**-0.31273379**	-0.12667941	-0.27380280	-0.266617234	0.12798211	-0.10357722
Sunshine	0.01778577	0.08185111	0.05090905	0.001118755	0.23622519	0.14644276
Air_Temperature	**-0.47159680**	-0.25676936	**-0.44492419**	**-0.454116464**	0.11755991	-0.08521804
Wind_Speed	-0.21705118	-0.11646983	-0.19504994	-0.178381772	0.07441385	-0.07775521
Relative_Humidity	-0.11719336	-0.23686242	-0.17138148	-0.076774333	-0.25396230	-0.09249402
Wind_Direction	0.22530220	0.16353290	0.21280814	0.205088855	-0.12374505	-0.01470259

Figure 3 Correlation table with highlighting of significant values

From Table 3, it can be see that the one with highest correlation is Nitrogen Oxide and Air Temperature, that means as the air temperature increased, it is likely of get an observation of Nitrogen oxide in the air decreased. Variables with Moderate relationship strength are 'Evaporation' vs 'Nitrogen.Oxide', 'Air_Temperature' vs 'Nirtogen.Oxide', 'Air_Temperature' vs 'Nitrogen.Oxides' and 'Air_Temperature' vs 'Carbon.Monoxide'.

4. Decision Tree

According to our table, there are many factors can be air quality indicator. Thus, we follow victor government website, deciding PM10 to be our air quality indicator. Then, PM10 was divided into 3 intervals, including Very Good (infinite - 16.4), Good (16.5-33), Fair (33 - infinite).

Carbon.Monoxide <dbl>	Wind_Direction <dbl>	PM_10 <dbl>	Target <fctr>
0.135702879	150.87500	12.493861	Very Good
0.123202879	149.91667	13.509542	Very Good
0.110702879	166.20833	8.023028	Very Good
0.081536212	151.37500	13.287500	Very Good
0.064869545	120.00000	11.916667	Very Good
0.064869545	104.58333	9.444847	Very Good
0.064869545	117.08333	9.958333	Very Good
0.073202879	177.16667	11.184542	Very Good
0.089869545	124.83333	9.854167	Very Good
0.114869545	121.04167	12.012500	Very Good

1–10 of 365 rows | 10–13 of 13 col... Previous 1 2 3 4 5 6 ... 37 Next

Figure 4 Table with Target value(label)

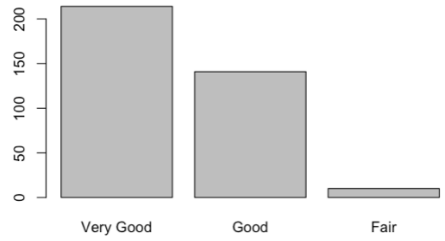

Figure 5 Number of Target values

After calculating from program, root node is relative humidity. And then, root was separated to two paths. One which is lower than 75.483 was pointed to Node 7 directly. This means air quality is very

good. Second path was assigned to wind direction node which judges whether wind direction is lower than 143.292 or not. If the answer is yes, way will assign to Node 3 which represents air quality is very good. If the answer is no, way will point to relative humidity which will determine their value is higher or lower than 63.72. If the value is lower, it will achieve Node 5 which represents air quality is Good. Finally, if the value is higher than 63.279, this means air quality is very good.

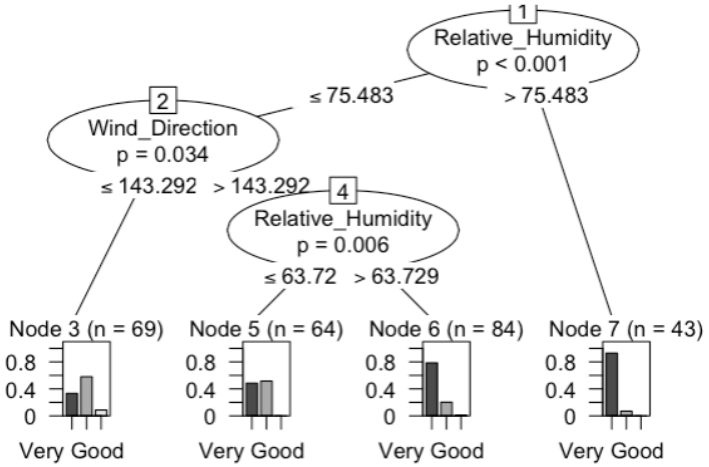

Figure 6 Decision Tree

5. Clustering

5.1. Preparing data for clustering

We decide to select 6 variables—Rainfall, evaporation, air temperature, wind speed, relative humidity and wind direction in the dataset. Furthermore, we use target variables, "Very good", "Good" and "Fair" according the air quality requirement of measurement of PM10 concentration.

5.2. Determine K

K is the number of the cluster. We use the Elbow method to set the K value (King, Richards, Zuckerman, Blasier, Dillman, Friedman & Woo, 1999). In this method, the plot is drawn in Figure7 to find the K. As shown in the Figure7, we decide the K value between 3 and 5. Therefore, we use other method, silhouette, to decide K value precisely. The silhouette score of 0.424 for K = 3, 0.421 for K =4 and 0.37 for K = 5. Thus, we choose the K value to be 3.

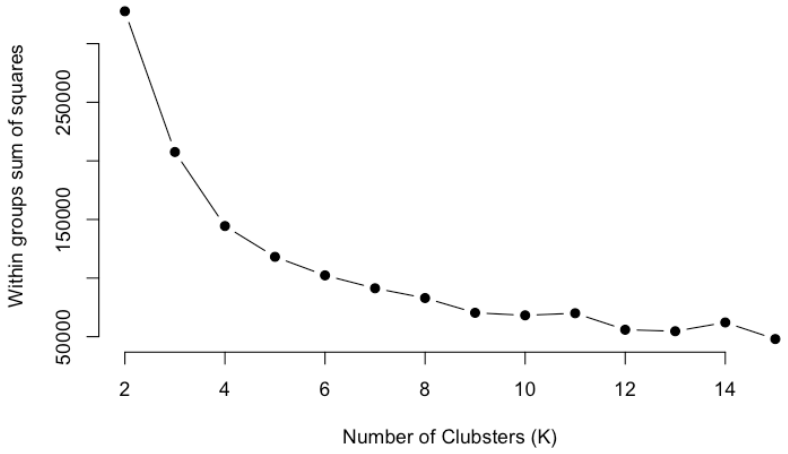

Figure 7 Elbow Chart

5.3. Visualize the cluster model

According to decision tree analysis, we choose relative humidity and wind direction as our objectives of this clustering process. The Figure 8 is draw below to show that the wind direction highly affects the cluster.

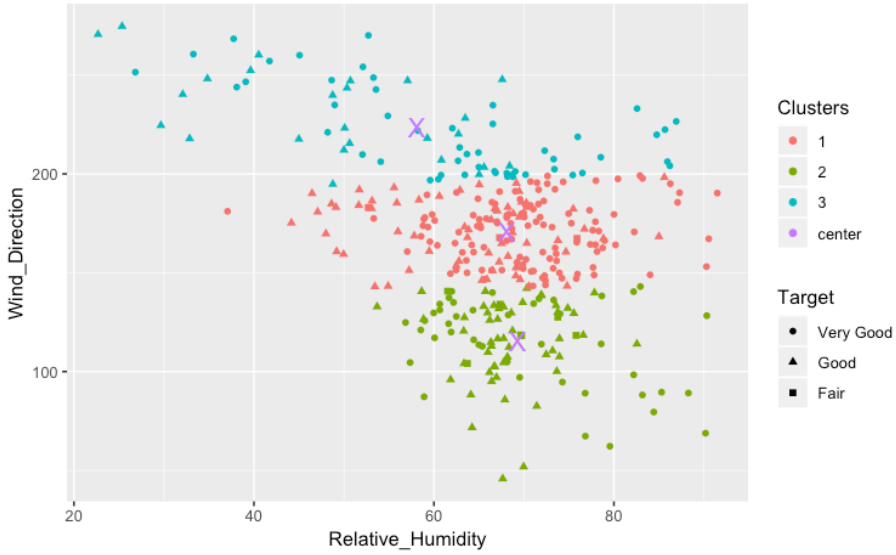

Figure 8

We choose two factors, air temperature and evaporation, to cluster data again in Figure9. In Figure9, we can find that cluster 3 more likely to be in lower air temperature and low evaporation environment. On the other hand, cluster 2 is more likely to appear in higher evaporation and higher air temperature environment.

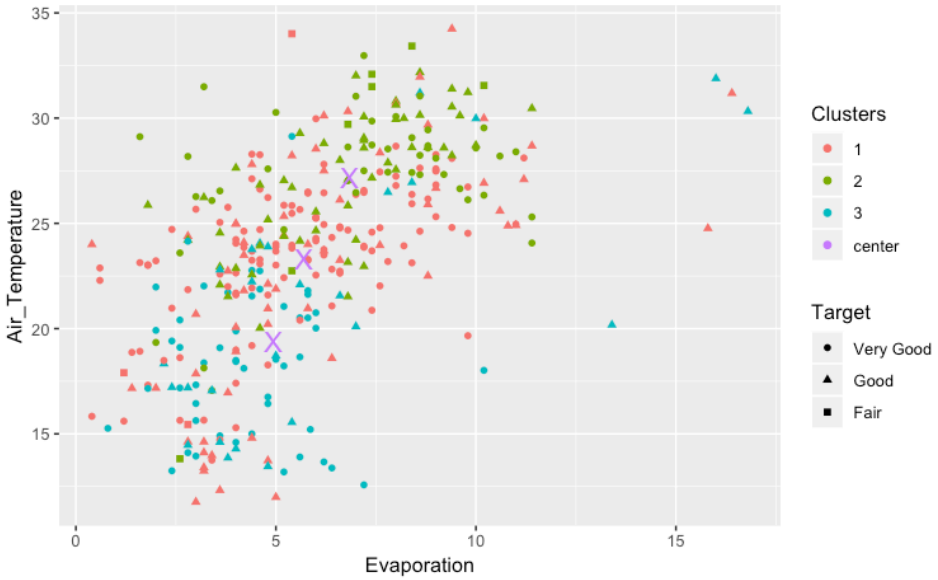

Figure 9

Reference

[1] Gregory W. Corder., & Dale I. Foreman (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach* (2nd ed.). Retrieved from https://onlinelibrary.wiley.com/doi/book/10.1002/9781118165881

King, G. J., Richards, R. R., Zuckerman, J. D., Blasier, R., Dillman, C., Friedman, R. J., ... & Woo, S. L. (1999). A standardized method for assessment of elbow function. Journal of shoulder and elbow surgery, 8(4), 351-354.