DSC 424 Final Project Report
Team Rain in Australia
Team Members: Divya Bhattiprolu, Yue Hou, Yawen Wang

Non-technical Summary

Data

The Rain in Australia dataset is a dataset from Kaggle. This dataset contains about 10 years of daily weather observations from 49 locations across Australia. The weatherAUS.csv file contains 145460 rows and 23 columns. The variables include both quantitative variables and categorical variables. In order to understand the Rain pattern in Australia and predict if it is going to rain tomorrow in Australia, we focused on three geographically separated locations, Perth, Melbourne Airport and Sydney Airport.

Techniques

1. **Exploratory Data Analysis**
   Exploratory Data Analysis is an approach to analyze data sets to summarize their main characteristics. In this project, we used some statistical graphs to visualize the data to perform the initial investigations for both categorical variables and numerical variables. For categorical variables, we implemented a Chi-Squared test to confirm if they are independent with the dependent variable. After that, we created bar charts to visually investigate their relationships with the dependent variable. For numerical variables, we used two graphs to figure out the multicollinearity within the dataset, and the multicollinearity issue leads us to the following steps to build parsimonious models to predict rain in the next day.

2. **Decision Tree Analysis**
   Decision tree algorithm is the graphical representation of all the possible solutions to a decision based on certain conditions. It is a relatively simple and basic model among the machine learning algorithms like Random forest, neural network, etc., and also is easier to understand. We can predict the results and the conditions for the same by just looking at the output tree diagram. In our model, we notice that in order to predict whether there will be rainfall tomorrow or not, the variable RainTomorrow will be used which is affected mainly by the components - Humidity at 3pm and Pressure at 3pm. The target outcomes are nothing but the probabilities of the *yes and no* factors of the above mentioned variables.

3. **Penalized Logistic Regression, All Subsets Regression**
   Penalized logistic regression adds a penalty to the simple logistic model, and it is useful when the dataset has too many variables. This results in shrinking the coefficients of the less contributive variables toward zero, which is also known as regularization. Due to the high multicollinearity and the number of variables in the dataset, we decided to go with lasso regression which will turn the less contributive variables to be exactly zero. The results will give us a parsimonious model for each location.

All subsets regression runs regression on all subsets of the variables and chooses the one that maximizes (or minimizes) the value of the criterion. We picked R squared as the criterion. Since we have 16 numerical variables, it is important to have a parsimonious model which is easy to explain and share most of the coefficients selected from the penalized logistic regression.

4. **Principal Component Analysis(PCA) and Common Factor Analysis(CFA)**
   They can help to build a parsimonious predictive model. In general, PCA and CFA do dimensionality reduction and find patterns that exist in latent variables. They do the same analysis which is selecting the factors/components but they use different methods. PCA and CFA are the common method for reducing the complexity of the dataset and maximizing the variance. Four components were discovered by using both techniques: 1) Temperature, 2) Wind Directions, 3) Climatic effects and 4) Wind Speed. In addition to checking if PCA/CFA is helpful to build a more parsimonious model, we performed the regression with the scores from PCA and compared with stepwise model selection by AIC. The performances of regression with the scores from PCA decrease about 20% - 30%.

Technical Summary

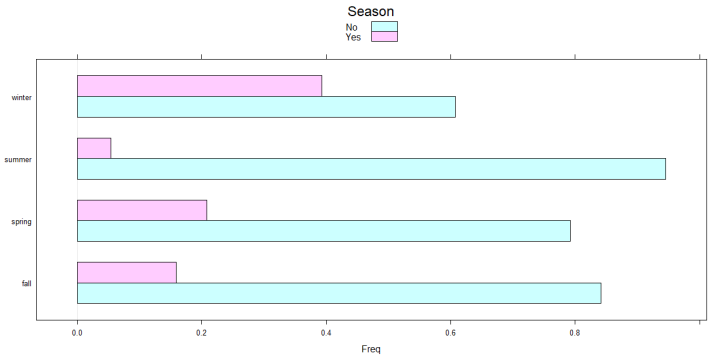
Data

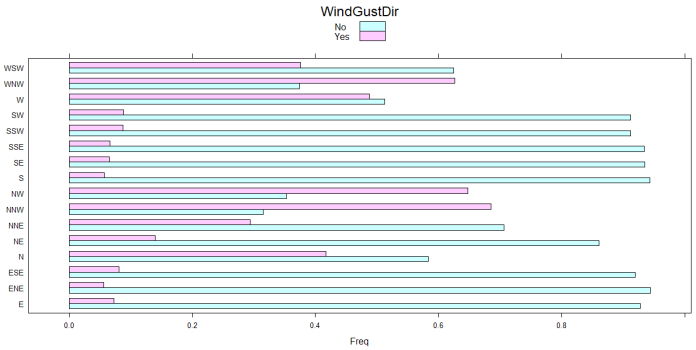
The original dataset contains roughly 10 years of daily weather observations, with 14.5k instances and 23 variables, from 49 locations in Australia. After narrowing down the research focus, we have decided to 1) Utilize penalized logistic regression to identify a parsimonious set of features that best predict rain/no-rain for “tomorrow”; 2) Feature engineer location and monthly/seasonal factors, and explain how factor predictive ability changes along those lines. Use lasso regression for feature selection; 3) Use PCA and lasso regression to build a parsimonious predictive model that predicts average accumulated rainfall by region and season. The main geographically separated locations we will be analyzing are Perth, Melbourne Airport and Sydney Airport.

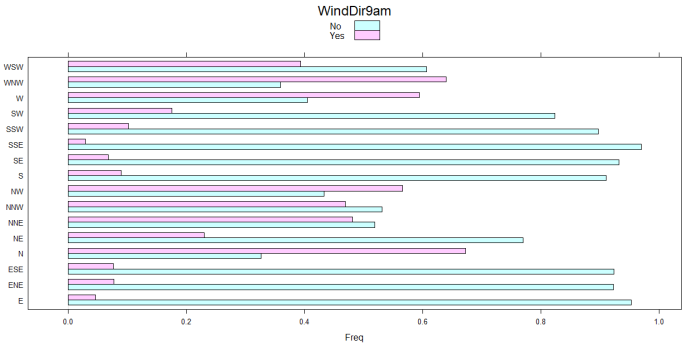
Exploratory Data Analysis

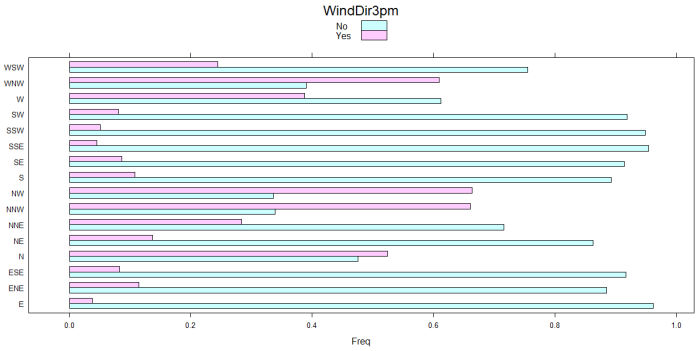
1. Categorical Variables Analysis
   There are 4 categorical variables, Season, WindGustDir, WindDir9am, WindDir3pm, other than the dependent variable, RainTomorrow. Season has 4 levels, and WindGustDir, WindDir9am and WindDir3pm has 16 levels each. Firstly, we implemented the Chi-Squared test as the image shown below and results show all four categorical variables are statistically significant, so we can reject the null hypothesis that RainTomorrow does not depend on these four categorical variables.

Next we used bar charts for 4 categorical variables vs RainTomorrow to find the relationships within these variables. Below are plots for Perth.





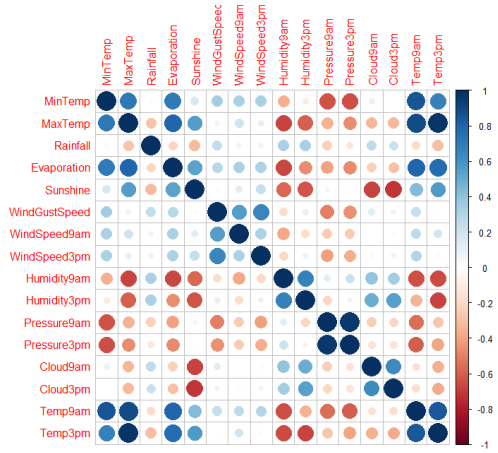


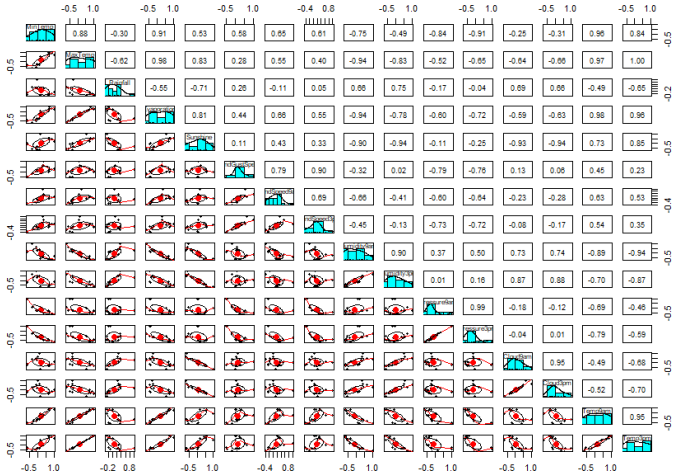


We computed the same process for three locations, and have the following conclusions:
1) Season: In Perth, winter is the season with more rains, and summer days are mostly without rain. In Melbourne Airport, the chance of rain the next day is not as high as Perth, meanwhile the chance of raining the next day in summer is higher than Perth.

Compared to Perth, Melbourne has less raining days in winter, and more raining days in summer. In Sydney Airport, four seasons share similar frequencies of raining days.
2) WindGustDir: In Perth, NW, NNW, and WNW directions suggest higher chance of rain the next day. In Melbourne Airport, E and ENE directions suggest higher frequencies of rain in the next day. In Sydney Airport, ESE and SE directions suggest higher frequencies of rain in the next day.
3) WinDir9am: In Perth, N and WNW directions suggest higher chance of rain the next day. In Melbourne Airport, NNE and ENE directions are related with higher chances of rain the next day. In Sydney Airport, E and SSW are related with higher chances of rain the next day.
4) WinDir3pm: In Perth, NW, NNW and WNW directions suggest higher chance of rain the next day. In Melbourne Airport, NE, SSW and WNW directions are related with higher chances of rain the next day. In Sydney Airport, SSW and SW directions are related with higher chances of rain in the next day.

2. Numerical Variables Analysis
For quantitative variables, the initial analysis is based on two figures below. The first figure is the correlation plot between numerical variables. The second figure is the scatter plots and correlation coefficient between variables. From the first figure, we can tell some strong correlations that will lead to multicollinearity such as MinTemp and MaxTemp, Humidity9am and Temp3pm. In the second figure, similarly, we can tell some multicollinearity among the variables.

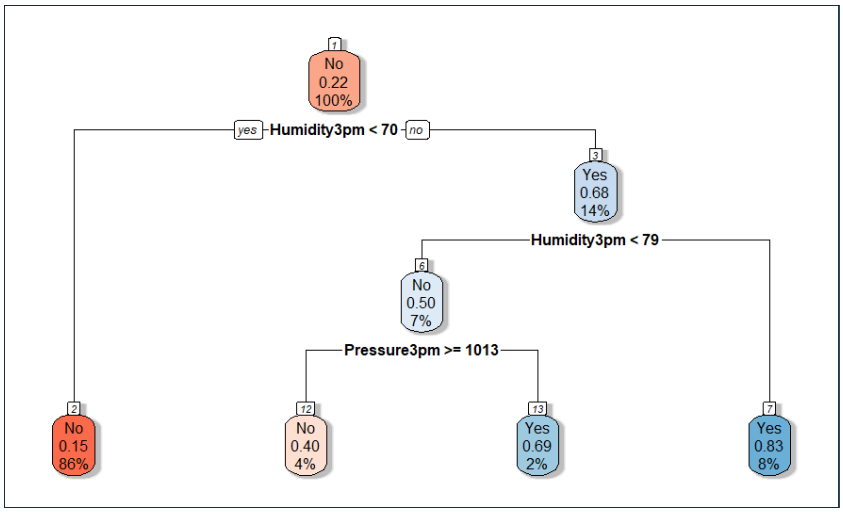




Decision Tree Analysis

The decision tree mainly has - decision node, leaf node, splitting, branch/subtree. The decision node or the root node is the starting point of the algorithm with the entire data set, leaf node is the end of the algorithm and from this node it cannot be separated further, splitting is based on the condition and the branch/subtree is the part after splitting the tree or data.

The library function ‘rpart’ is used to build a decision tree model to predict if there will be rainfall the day after it has rained or not. The function ‘rpart.plot’ is used to draw a tree diagram to show the rules that are built into the model. The output obtained shows that the variable RainTomorrow is affected mainly by the components Humidity and Pressure at 3pm. The decision tree diagram is as shown below:

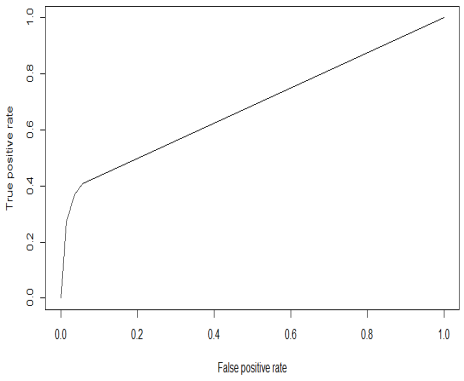


Looking at the decision tree analysis, we can say that there is an 86% chance of no rainfall the next day when the humidity at 3pm is less than 70 today in the locations - Melbourne Airport,

Perth and Sydney Airport. Whereas, there is a 14% chance of rainfall tomorrow when the humidity at 3pm is more than 70 today, but if the humidity is more than 70 and less than 79 then the chances fall to 8% of there being a rainfall tomorrow. On the other hand, there's a 7% chance that it will not rain when the humidity at 3pm is less than 79, but in case the pressure at 3pm today would be greater than or equal to 1013 then there's a 2% chance that it will rain tomorrow and a 4% chance that it will not rain tomorrow.

Next, obtaining the confusion matrix for the decision tree analysis we can see that there is a 83.7% accuracy level and on the other hand the value of sensitivity is too low.

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  4633  840
       Yes  167  537

               Accuracy : 0.837
                 95% CI : (0.8275, 0.8461)
    No Information Rate : 0.7771
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4301

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.38998
            Specificity : 0.96521
         Pos Pred Value : 0.76278
         Neg Pred Value : 0.84652
             Prevalence : 0.22292
         Detection Rate : 0.08694
   Detection Prevalence : 0.11397
      Balanced Accuracy : 0.67759

       'Positive' Class : Yes
```

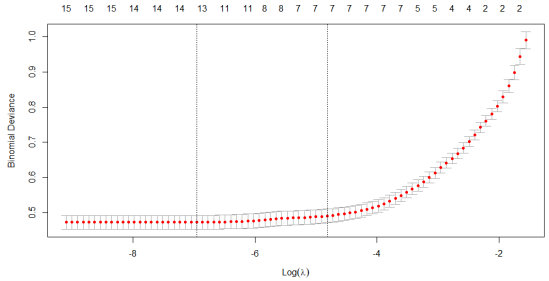



```
> performance(P_Test, "auc")@y.values
[[1]]
[1] 0.6828006
```

From the confusion matrix, we can see that there are 4633 cases of no rainfall which is a true negative while the remaining 840 are wrongly classified as the cases of rainfall which is the false positive. There are 537 cases of a true positive which is the chances of there being a rainfall tomorrow while 167 cases of false negative which are wrongly classified as the times that there will be no rainfall. The output by plotting the false positive rate and True positive rate is also shown. The ROC AUC value for the decision tree method is 0.68 which is calculated using the performance function from the ROCR library function.

Penalized Logistic Regression, All Subsets Selection

1. Penalized Logistic Regression (Perth)
   The penalized logistic regression model will use Lasso regression to select features, due to the high multicollinearity in this dataset. The dataset was split into training (75%) and testing (25%) set as follows, and 10 fold cross validation was used to fit the model (see the plot below for using cv.glmnet with 10 fold cross validation). I will go with the coefficient based on lambda.1se, which gives 7 features with betas not equal to zero.




```
> coef(cv.fit, s="lambda.1se")
17 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept)   187.37273550
MinTemp        -0.11843762
MaxTemp         .
Rainfall        .
Evaporation    -0.17331667
Sunshine       -0.20639888
WindGustSpeed   0.05256767
WindSpeed9am    .
WindSpeed3pm    .
Humidity9am     .
Humidity3pm     0.04678459
Pressure9am     .
Pressure3pm    -0.18666708
Cloud9am        .
Cloud3pm        0.05344651
Temp9am         .
Temp3pm         .
```

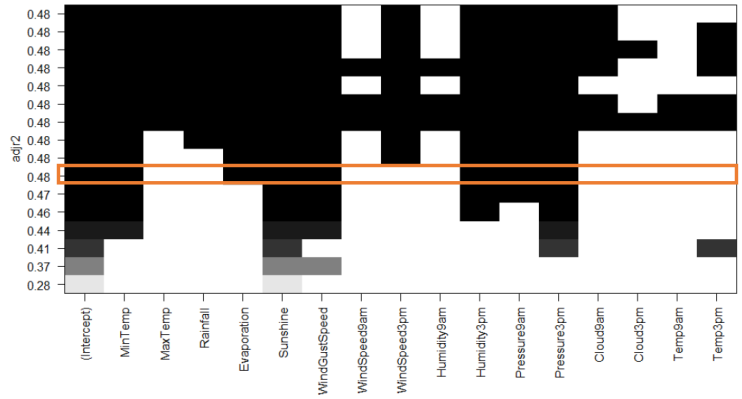
Next, we predicted the RainTomorrow both on the training set and testing set. Based on the confusion tables below, the accuracy on the training set is 90%, and the accuracy on the testing set is 89%. The difference is only around 1 percent, so this lasso penalized logistic regression model performs pretty good.

```
> table(predTr, yTrain)
      yTrain
predTr   No  Yes
   No  1757  156
   Yes   64  291
```

```
> table(pred, yTest)
     yTest
pred   No Yes
  No  569  67
  Yes  19 102
```

2. All Subsets Regression (Perth)

All subsets selection is also applied for this dataset. The adjusted R squared was selected as criterion, and the regression chooses the variables that maximizes the value of adjusted R squared. For a dataset with a large number of variables, choosing a parsimonious model is important. The orange box marked below is the parsimonious model that I chose, which is easy to explain and share most of the coefficients with the lasso penalized logistic regression.



3. Comparison of coefficients from penalized logistic regression and all subsets selection (Perth)
   Based on the table below, we can tell these two methods gave two similar models with almost identical variables (6 out of 7).

	Penalized logistic regression	All subsets selection
MinTemp	√	√
MaxTemp		
Rainfall		
Evaporation	√	√
Sunshine	√	√
WindGustSpeed	√	√
WindSpeed9am		
WindSpeed3pm		
Humidity9am		
Humidity3pm	√	√
Pressure9am		√
Pressure3pm	√	√
Cloud9am		
Cloud3pm	√	
Temp9am		
Temp3pm		

4. Comparison of logistic regression models for three locations
   Based on the comparison below, we can conclude that in all three locations, RainTomorrow is negatively correlated with Sunshine and Pressure3pm, and is positively correlated with WindGustSpeed, Humidity3pm and Cloud3pm. Evaporation is also an important variable in most locations (Perth and Sydney Airport). The accuracy for models of Perth, Melbourne Airport, and Sydney Airport on testing set is 89%, 85%, and 83% for each location.

Perth:

```
> coef(cv.fit, s="lambda.1se")
17 x 1 sparse Matrix of class '
                       1
(Intercept)   187.37273550
MinTemp        -0.11843762
MaxTemp         .
Rainfall        .
Evaporation    -0.17331667
Sunshine       -0.20639888
WindGustSpeed   0.05256767
WindSpeed9am    .
WindSpeed3pm    .
Humidity9am     .
Humidity3pm     0.04678459
Pressure9am     .
Pressure3pm    -0.18666708
Cloud9am        .
Cloud3pm        0.05344651
Temp9am         .
Temp3pm         .
```

Melbourne Airport:

```
> coef(cv.fit, s="lambda.1se")
17 x 1 sparse Matrix of class "
                       1
(Intercept)   89.91451499
MinTemp         .
MaxTemp         .
Rainfall        .
Evaporation     .
Sunshine       -0.07990532
WindGustSpeed   0.01184071
WindSpeed9am    .
WindSpeed3pm    .
Humidity9am     .
Humidity3pm     0.03633565
Pressure9am     .
Pressure3pm    -0.09236445
Cloud9am        .
Cloud3pm        0.07812323
Temp9am         .
Temp3pm         .
```

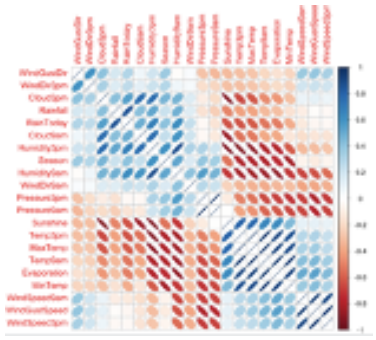
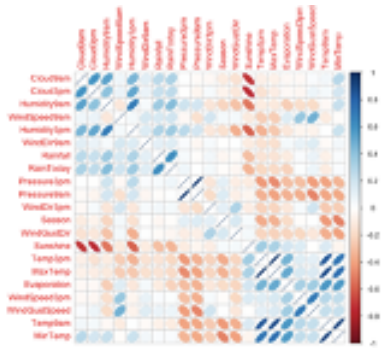
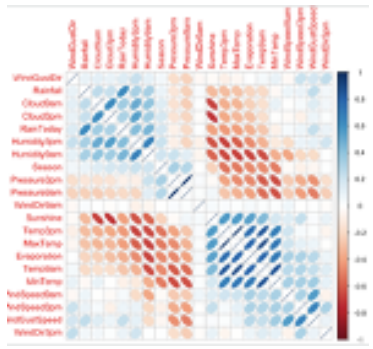
Sydney Airport:

```
> coef(cv.fit, s="lambda.1se")
17 x 1 sparse Matrix of class "
                       1
(Intercept)   -3.2512431202
MinTemp         .
MaxTemp         .
Rainfall        0.0239957514
Evaporation    -0.0278462077
Sunshine       -0.1325874909
WindGustSpeed   0.0253877867
WindSpeed9am    .
WindSpeed3pm    .
Humidity9am     .
Humidity3pm     0.0350391146
Pressure9am     .
Pressure3pm    -0.0007677225
Cloud9am        .
Cloud3pm        0.1212198661
Temp9am         .
Temp3pm         .
```

Principal Component and Common Factor Analysis

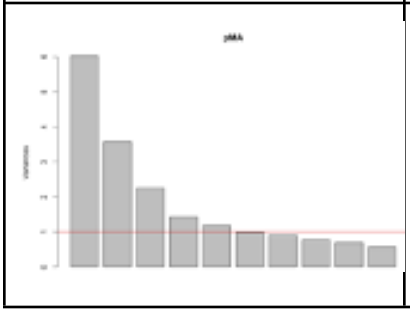
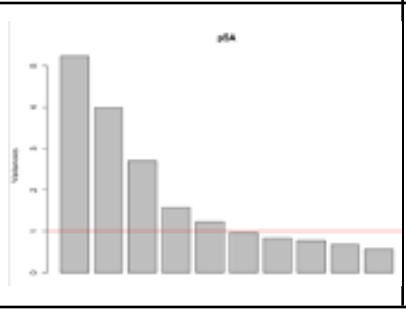
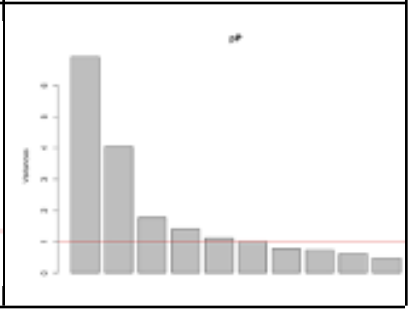
First, corrplot() was used to visualize the ordered correlation matrix for all remaining variables of three locations(Figure: correlation matrix). By comparing three correlation matrices, some variables have different levels of correlation and some variables have the same levels among three locations. It is very interesting to see variables play different roles for different locations.

Figure: correlation matrix

Melbourne Airport	Sydney Airport	Perth
		

Second, we utilized prcomp() to find the ideal number of factors/components by looking at the knee and var = 1 criteria (Figure: variance). For knees, where the scree plots “bend” to become more horizontal, they are 4 for Melbourne Airport and Sydney Airport and 3 for Perth. Those components capture about 64-67% cumulative proportion. For var = 1 criteria, all three scree plots suggest 5 components, and PC 5 capture 68-73%. Due to the parsimonious model, we picked PC 4.

Figure: variance

Melbourne Airport	Sydney Airport	Perth
		

Next, we used PC 4 and 0.4 cutoff to run the PCA. In general, RC1s are consistent among the three locations. All RC1s mainly have Season, MinTemp, MaxTemp, Evaporation, Temp9am, Temp3pm, Pressure9am, Pressure3pm, and Humidity9am. This component mainly measures

the temperature. RC2s, RC3s, and RC4s are different for different locations, but they are formed by the same variables. They are about the 1)wind direction, 2)humidity, rainfall, and sunshine 3)wind speed. Wind direction variables have weak factor loadings in PCA and CFA.

Additionally, CFA produced similar types of factors but more variables were dropped from the result because of the weak factor loadings. CFA makes the result more easier to interpret but it captures less variance than PCA does.

Finally, we ran the regression with the scores from PCA and compared with stepwise model selection by AIC. For Melbourne Airport, the adjusted R-squared for stepwise model is 0.3169 and for PCA scores is 0.2228. For Sydney Airport, the adjusted R-squared for stepwise model is 0.3509 and for PCA scores is 0.2845. For Perth, the adjusted R-squared for stepwise model is 0.4915 and for PCA scores is 0.3739. Overall, the performances of regression with the scores from PCA decrease about 20% - 30%.

Four components were discovered: 1) Temperature, 2) Wind Directions, 3) Climatic effects and 4) Wind Speed. But based on the analysis results, we would recommend removing wind direction variables from the regression analysis.

Conclusion

For exploratory data analysis, we figured out the multicollinearity issue in the dataset. We also confirmed that all 4 categorical variables are dependent with the dependent variable, RainTomorrow. Through the bar charts analysis, we can conclude 1) Season: In Perth, winter is the season with more rains, and summer days are mostly without rain. Compared with Perth, Melbourne has less raining days in winter, and more raining days in summer. In Sydney Airport, four seasons share similar frequencies of raining days; 2) Wind related variables: In Perth, RainTomorrow is related with winds from NW related directions. In Melbourne Airport, it is related with E related directions. In Sydney airport, it is related with E and SW related directions.

From the decision tree analysis, there seems to be a higher probability of there being no rainfall tomorrow when Humidity at 3pm is less than 70 or 79 and Pressure at 3pm is greater than or equal to 1013 today in Melbourne Airport, Perth and Sydney Airport.

For the penalized logistics regression, among the selected 3 locations, RainTomorrow is negatively correlated with Sunshine and Pressure3pm, and is positively correlated with WindGustSpeed, Humidity3pm and Cloud3pm. Evaporation is also an important variable in most locations (Perth and Sydney Airport).

For the principal component and common factor analysis, four components were found, temperature, wind directions, climatic effects, and wind speed. Those components capture about 64-67% cumulative proportion. We also conducted the regression with the scores from

PCA and compared with stepwise model selection by AIC. The result shows the regression with the scores from PCA decreased 20% - 30% of Adjusted R-squared.

Appendix

Individual Reports

A. Individual Report - Divya Bhattiprolu

Summary of my analysis

Initially our dataset had 145460 rows and 23 columns. With the head function I noticed that there are quite a few columns that have a lot of missing values; Evaporation and Sunshine having around 98% of the missing values. Here, cleaning the dataset and deciding the way to proceed with these NA or missing values was a challenge. The discussions we had regarding this provided an insight into working with real-time data as there will be data discrepancies with any data set out in the world. My contribution to this project is decision tree analysis. Understanding the structure of a general decision tree and then interpreting the outcome values and their probabilities with respect to our dataset proved to be a bit confusing at first. It was interesting to find out that humidity and pressure at 3pm only is what plays an important role in determining whether it is going to rain the next day. This might be due to that fact that these humidity readings are at a certain time of day and not an average of the day.

Apart from this, I have performed some exploratory analysis as well and also explored the neural network and k nearest neighbours' techniques. For the neural network technique, I have chosen the variables RainTomorrow, RainToday, Rainfall, Humidity3pm, MaxTemp and Pressure3pm. The output of the neural network is as shown below.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 16886  2923
         1   695  2080

               Accuracy : 0.8398
                 95% CI : (0.8349, 0.8446)
    No Information Rate : 0.7785
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4475

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9605
            Specificity : 0.4158
         Pos Pred Value : 0.8524
         Neg Pred Value : 0.7495
             Prevalence : 0.7785
         Detection Rate : 0.7477
   Detection Prevalence : 0.8771
      Balanced Accuracy : 0.6881

       'Positive' Class : 0
```

The accuracy in this model is 84.37% which is quite similar to the value obtained from the decision tree technique. As we have an imbalanced data set when it comes to class proportion using accuracy might not be ideal. So, we could choose a metric such as improving the precision maybe for an improved better model.

The output obtained by dividing the data set into train and test, fitting K-NN to the training set and predicting the test set results is as shown below.

```
library(class)
kn_pred = knn(train = training_set[, -105],
              test = test_set[, -105],
              cl = training_set[, 105],
              k = 5,
              prob = TRUE)
```

```
> confusionMatrix(factor(kn_pred), factor(test_set$RainTomorrow),positive="1")
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 79755 10634
         1  2982 13274

               Accuracy : 0.8723
                 95% CI : (0.8703, 0.8743)
    No Information Rate : 0.7758
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5858
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5552
            Specificity : 0.9640
         Pos Pred Value : 0.8166
         Neg Pred Value : 0.8824
             Prevalence : 0.2242
         Detection Rate : 0.1245
   Detection Prevalence : 0.1524
      Balanced Accuracy : 0.7596

       'Positive' Class : 1
```

We have an accuracy of 87.2% which is near to 88% and it is almost always said that it will not rain tomorrow, this model might not be considered as being much informative.

My Takeaways

Throughout this course there were quite a few interesting techniques - decision tree analysis and correspondence analysis being the ones that I really enjoyed working on. Coming from a business analytics major I personally feel that correspondence analysis would help a great deal in exploring relationships among categorical variables. And decision tree analysis is something that I might be using a lot as it helps in decision making by clearly laying out the problem where all the options are challenged. This helps in analyzing all the possible outcomes by quantifying the values and the probabilities of achieving them. The course has been really informative and I got to understand and implement the techniques better by working on this project with my teammates.

B. Individual Report - Yue Hou

As a team member in Team Rain in Australia, I'm taking the following duties: 1) Set up a Google folder to enable us working as a group and starting new documents for milestones; 2) Communicate with professor to set up meeting time; 3) Set up recurring Zoom meeting for us to discuss milestones.

My analysis in the final project include the following parts: 1) Exploratory data analysis on the dataset; 2) Penalized logistic regression; 3) All subsets regression.

Summary of My Analysis

1.Exploratory Data Analysis

There are two main issues in this dataset: 1) NA values across the dataset; 2) time series issue.

To solve the first problem, I deleted NA values of this dataset and compared the results of each location. Location Perth has 3193 rows before cleaning, and 3025 rows afterwards, which would be a good location to start with. Melbourne Airport and Sydney Airport are selected based on this same criterion.

As for the time series issue, I chose to ignore the influence of day-to-day change for my analysis, and deleted the column “Date”. Meanwhile, considering the influence of seasons on rainfall, I created a new column “Season”. In Australia, summer is from December to February, fall is from March to May, winter is from June to August, and spring is from September to November.

I also noticed that when column RainFall is 1mm or more, the column RainToday equals “Yes”, so I deleted the column RainToday.

There are 4 categorical variables, Season, WindGustDir, WindDir9am, WindDir3pm, other than the dependent variable, RainTomorrow. Season has 4 levels, and WindGustDir, WindDir9am and WindDir3pm have 16 levels each. Firstly, I implemented the Chi-Squared test and results show all four categorical variables are statistically significant, so we can reject the null hypothesis that RainTomorrow does not depend on these four categorical variables. Next, I used bar charts for these categorical variables to test their relationship with RainTomorrow.

For quantitative variables, the initial analysis is based on two figures. The first figure is the correlation plot between numerical variables. The second figure is the scatter plots and correlation coefficient between variables. The multicollinearity issue was recognized through the two plots, so the next step would be building the logistic regression model using lasso regression for numerical variables and all subsets regression, and then compare the results of these two methods.

2.Penalized Logistic Regression

The penalized logistic regression model used Lasso regression to select features, due to the high multicollinearity in this dataset. The dataset was split into training (75%) and testing (25%) set as follows, and 10 fold cross validation was used to fit the model. I chose the coefficients based on lambda.1se, which gives 7 features with betas not equal to zero for location Perth.

Next, I predicted the RainTomorrow both on the training set and testing set. Based on the confusion table, the accuracy on the training set is 90%, and the accuracy on the testing set is 89% for Perth. The difference is only around 1 percent, so this lasso penalized logistic regression model performs pretty good. Then I repeated this process for Sydney Airport and Melbourne Airport.

3. All Subsets Regression

In order to further explore this dataset, and confirm my analysis in logistic regression is on the right track, I implemented all subsets regression on the numerical variables. I chose adjusted R squared as the criterion. Since the dataset has 16 numerical variables, I chose the model based on interpretability and similarity to the lasso penalized logistic regression model. The results confirm my analysis in the logistic regression model that 6 out of 7 features are common features.

Conclusions

For the penalized logistics regression and all subsets regression, among the selected 3 locations, RainTomorrow is negatively correlated with Sunshine and Pressure3pm, and is positively correlated with WindGustSpeed, Humidity3pm and Cloud3pm. Evaporation is also an important variable in most locations (Perth and Sydney Airport).

From the analysis of categorical variables, we can observe different patterns of weather in 3 locations. For Perth, winter is the season with more rains, and summer days are mostly without rain. RainTomorrow is related to winds from NW related directions. For Melbourne airport, it has less raining days in winter, and more raining days in summer. RainTomorrow is related to winds from E related directions. For Sydney airport, four seasons share similar frequencies of raining days. RainTomorrow is related to winds from E and SW related directions.

My Takeaways

The analysis of the dataset provides me the opportunity to practice how to deal with a real life dataset. I started from the data cleaning by deleting the NA values, deleting columns, creating new features, to initial analysis of the dataset to select locations. Next I applied the penalized logistic regression and all subsets regression which was covered by DSC 424. Through the milestone, I improved my understanding of this dataset, as well as the methodology applied on it. I also enjoyed the process of working with my team members, and learning from them how to work properly in a team.

C. Individual Report - Yawen Wang

As a teammate, I found the suitable dataset, set up the group chat by using WhatsApp, cleaned up the data set for the final analysis, initiated some of the files for milestones. For the data analysis, I performed the principal component and common factor analysis. Additionally, built the regression model by using the scores from the principal component analysis to check if it would be a more parsimonious model.

Summary of My Analysis

I used prcomp() to find the suitable number of factors for the further analysis. The knee and var = 1 criteria show different results. In order to keep the model or analysis as simple as it could be, I chose four factors. Next step, I set nfactors = 4 and cutoff = 0.4 to run the principal component and common factor analysis. The results are similar but more variables were dropped by the common factor analysis. Four components were discovered: 1) Temperature, 2) Wind Directions, 3) Climatic effects and 4) Wind Speed.

Additionally, I ran the regression with the scores from PCA and compared with stepwise model selection by AIC. Overall, the performances of regression with the scores from PCA decrease about 20% - 30%. Some of the key graphs are included in this report.

My Takeaways

Due to the complexity of the dataset, I had learned some new functions and expressions to clean the data and create new variables for better analysis. Using creating a seasonal variable as an example, I formatted the date variables first then took out the month value and assigned the season based on the value of the month. My focus for this project is to utilize the principal component and common factor analysis to discover the latent variables or identify new variables. I really enjoyed the process because it gave me the chance to apply what I have learned from class on the real project/dataset and to work as a team. I always love team projects. They provide the opportunity to learn from others.

Variables

Variables	Description
Date	The date of observation
Location	The common name of the location of the weather station
MinTemp	The minimum temperature in degrees centigrade
MaxTemp	The maximum temperature in degrees centigrade
Rainfall	The amount of rainfall recorded for the day in millimeters
Evaporation	Class A pan evaporation (in millimeters) during 24 h
Sunshine	The number of hours of bright sunshine in the day
WindGustDir	The direction of the strongest wind gust in the 24 h to midnight
WindGustSpeed	The speed (in kilometers per hour) of the strongest wind gust in the 24 h to midnight
WindDir9am	The direction of the wind gust at 9 a.m.
WindDir3pm	The direction of the wind gust at 3 p.m.
WindSpeed9am	Wind speed (in kilometers per hour) averaged over 10 min before 9 a.m.

WindSpeed3pm	Wind speed (in kilometers per hour) averaged over 10 min before 3 p.m.
Humidity9am	Relative humidity (in percent) at 9 am
Humidity3pm	Relative humidity (in percent) at 3 pm
Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9 a.m.
Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3 p.m.
Cloud9am	Fraction of sky obscured by cloud at 9 a.m. This is measured in ”oktas,” which are a unit of eighths. It records how many eighths of the sky are obscured by cloud. A 0 measure indicates completely clear sky, while an 8 indicates that it is completely overcast
Cloud3pm	Fraction of sky obscured by cloud at 3 p.m.
Temp9am	Temperature (degrees C) at 9 a.m.
Temp3pm	Temperature (degrees C) at 3 p.m.
RainToday	Integer 1 if precipitation (in millimeters) in the 24 h to 9 a.m. exceeds 1 mm, otherwise 0
RainTomorrow	The binary target variable whether it rains or not during the next day