STAT420 Final Project

Prediction for rainfall amount (in mm) in Sydney

Aaryan Bahl (aaryanb2) , Sixian Li (sixianl2), Zhixing Guo (zhixing5)

11/11/2020

Section 1 Introduction

Our group's project aims to predict the precipitation(mm) in Sydney for next day based on the previous weather data. Sydney is a metropolis with beautiful coastal line and a typical humid subtropical climate which is distinguished from the climate in Urbana and Champaign. We are especially interested in the relationship between rainfall and other weather factors such as wind, sunshine etc. And we want to use the knowledge we learned from STAT420 to build a regression model to predict the next day's precipitation(mm) and test if our model coincides with the rainfall pattern of the humid subtropical climate.

Section 1.1 Description of the data file

The weatherAUS data contains daily weather observations from numerous Australian weather stations, which are available from http://www.bom.gov.au/climate/data (http://www.bom.gov.au/climate/data). This dataset is also available via the R package rattle.data and at https://rattle.togaware.com/weatherAUS.csv (https://rattle.togaware.com/weatherAUS.csv).

The whole weatherAUS dataset contains 142193 observations of 24 variables.

- Data: The date of observation
- Location: The common name of the location of the weather station
- MinTemp: The minimum temperature in degrees celsius
- MaxTemp: The maximum temperature in degrees celsius
- Rainfall: The amount of rainfall recorded for the day in mm
- Evaporation: The so-called Class A pan evaporation (mm) in the 24 hours to 9am
- Sunshine: The number of hours of bright sunshine in the day
- WindGustDir: The direction of the strongest wind gust in the 24 hours to midnight
- WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight
- WindDir9am, WindDir3pm: Direction of the wind at 9am and 3pm respectively
- WindSpeed9am, WindSpeed3pm: Wind speed (km/hr) averaged over 10 minutes prior to 9am and 3pm respectively
- Humidity9am, Humidity3pm: Humidity (percent) at 9am and 3pm respectively
- Pressure9am, Pressure3pm: Atmospheric pressure (hpa) reduced to mean sea level at 9am and 3pm respectively
- Cloud9am, Cloud3pm: Fraction of sky obscured by cloud (oktas) at 9am and 3pm respectively

- Temp9am, Temp3pm: Temperature (degrees C) at 9am and 3 pm respectively
- RainToday: indicator for whether precipitation (mm) in the 24 hours to 9am exceeds 1mm
- RISK_MM: The amount of rainfall recorded for next day in mm
- RainTomorrow: indicator for whether precipitation (mm) in tomorrow to 9am exceeds 1mm

Section 1.2 Primary research and questions

Section 2 Exploratory Data Analysis and Methodology

```
# import data
weather <- fread("weatherAUS.csv", data.table = FALSE)
```

```
# select weather data in Sydney and its neighborhood areas
sydney.weather = weather[weather$Location %in% c("Sydney","SydneyAirport"),]
sydney.weather = sydney.weather[sydney.weather$RISK_MM > 0,]

# remove Nan
sydney.weather = na.omit(sydney.weather)
```

Section 2.1 Categorical variables

By researching from a climate statistics report by the Burea of Meteorology on Sydney we observed that February to June saw the most rainfail and was regarded as the wet season of Sydney. So, we decided to create a new categorical variable “Season” that would indicate if the data was from a wet month or a dry month in the year.

```
# the variable Date
funSeason = function(x){
  month = as.numeric(format(x, "%m"))
  if (month >= 2 & month <= 6) {"Wet"}
  else {"Dry"}
}
sydney.weather$Season <- mapply(funSeason, sydney.weather$Date)
count(sydney.weather, 'Season')
```

```
##   Season freq
## 1    Dry  873
## 2    Wet  778
```

Next, we were trying to analyze categorical variables that indicated the wind gust direction or the wind direction at 9am or 3pm. We calculated the count of each variable in a particular direction to see if we could find any trends that could be helpful.

```
# the variable WindGustDir, WindDir9am, and WindDir3pm
# count the level frequency for each of the three factor variables
df1 = count(sydney.weather, 'WindGustDir')
df1[order(df1$freq),]
```

```
##    WindGustDir freq
## 7          NNW   37
## 4            N   38
## 15         WNW   60
## 8           NW   62
## 13          SW   62
## 1            E   65
## 6          NNE   70
## 16         WSW   71
## 3          ESE   75
## 2          ENE   78
## 10          SE  105
## 5           NE  111
## 14           W  115
## 11         SSE  155
## 12         SSW  261
## 9            S  286
```

```
df2 = count(sydney.weather, 'WindDir9am')
df2[order(df2$freq),]
```

```
##    WindDir9am freq
## 2         ENE   32
## 5          NE   53
## 3         ESE   54
## 1           E   55
## 6         NNE   63
## 13         SW   70
## 16        WSW   76
## 4           N   79
## 7         NNW   79
## 11        SSE   81
## 10         SE   83
## 9           S  151
## 15        WNW  161
## 8          NW  168
## 12        SSW  187
## 14          W  259
```

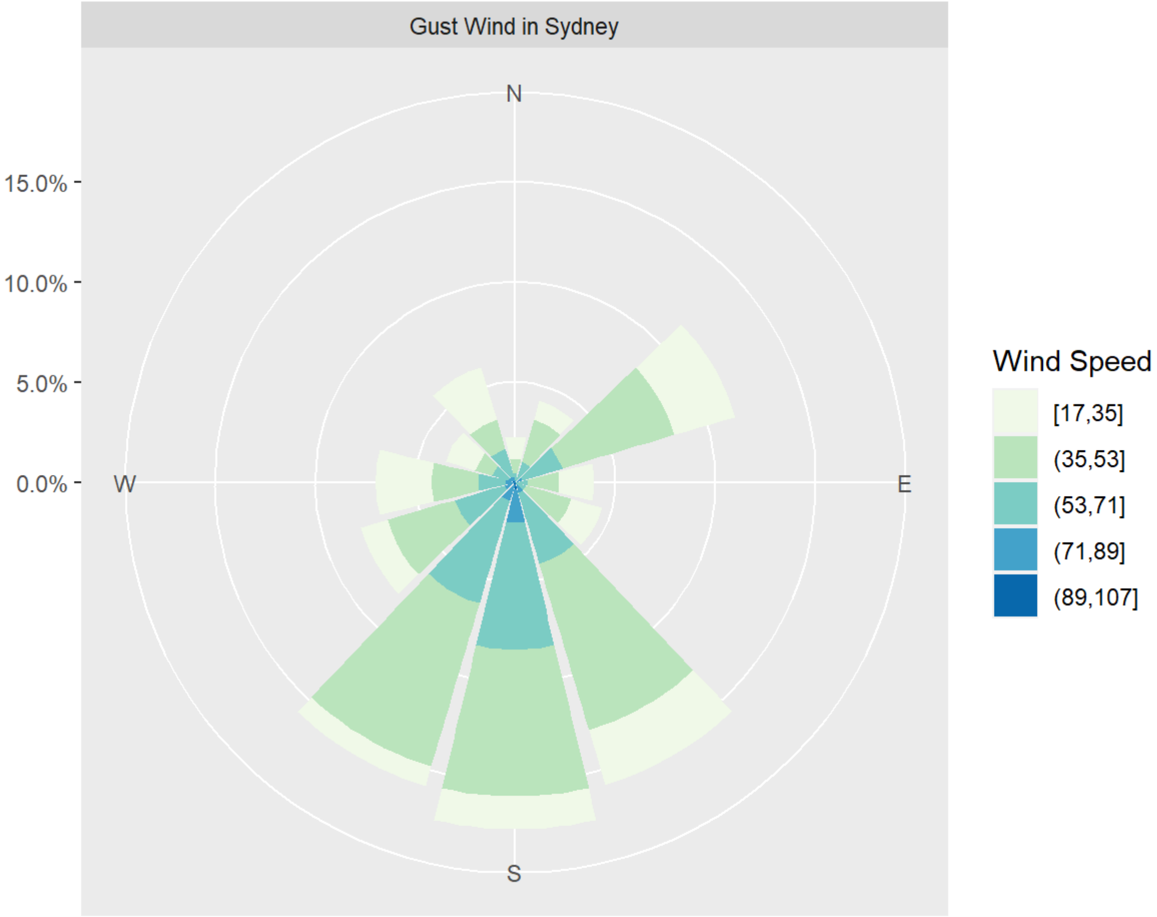
```
df3 = count(sydney.weather, 'WindDir3pm')
df3[order(df3$freq),]
```

```
##    WindDir3pm freq
## 7         NNW   34
## 8          NW   36
## 13         SW   37
## 16        WSW   39
## 4           N   41
## 15        WNW   48
## 6         NNE   57
## 14          W   57
## 3         ESE  110
## 2         ENE  120
## 5          NE  152
## 10         SE  152
## 1           E  155
## 12        SSW  176
## 11        SSE  188
## 9           S  249
```

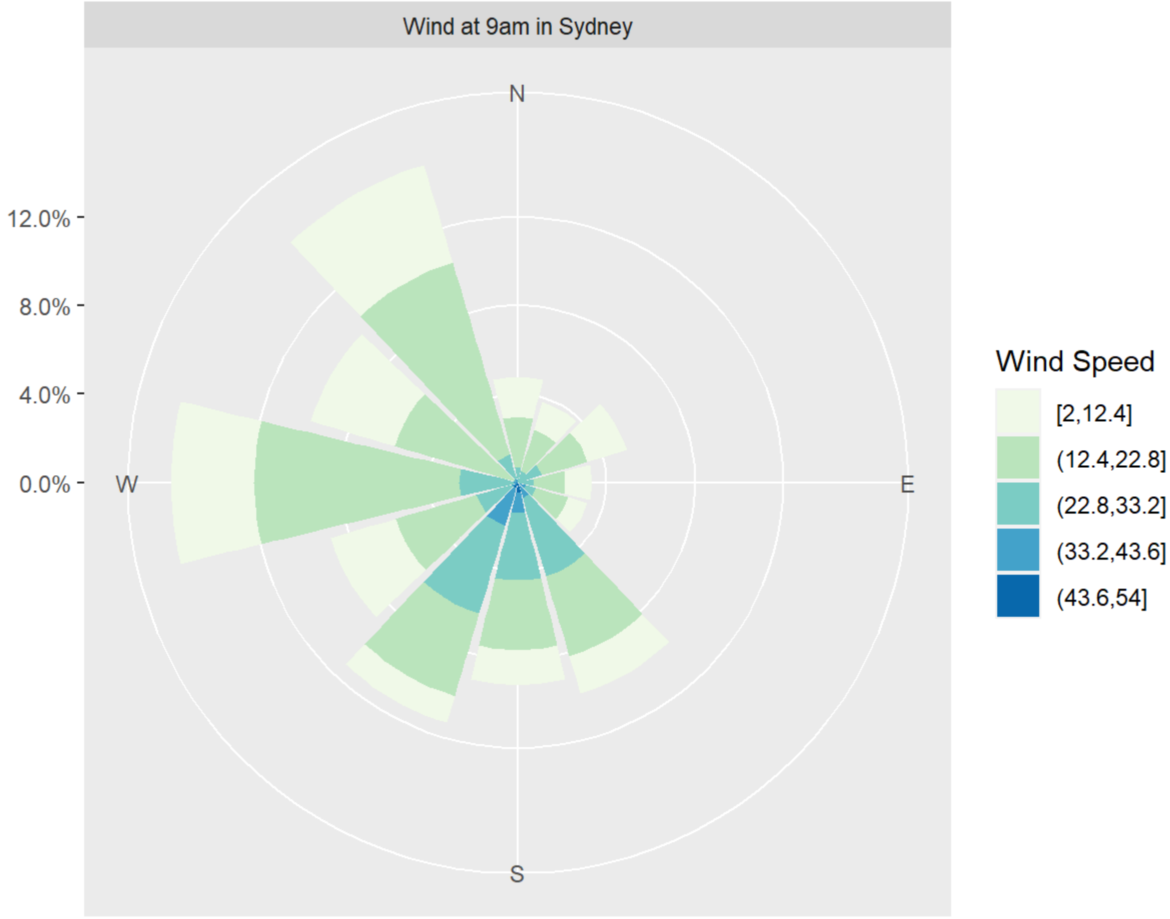
From the count above, we couldn’t find any trends in wind gust though we found some thing interesting for wind direction. We observed that some wind directions that were predominant at 9am (SSE,NNW,N,NW) were subordinate winds at 3pm. The vice versa also applied. To see if this was actually true we used plots below to visualize this.

```
# help function for printing the wind rose plot
funWind <- function(x) {
  if (x == "N") {360}
  else if (x == "NNE") {25}
  else if (x == "NE") {45}
  else if (x == "ENE") {65}
  else if (x == "E") {90}
  else if (x == "ESE") {115}
  else if (x == "SE") {135}
  else if (x == "SSE") {155}
  else if (x == "S") {180}
  else if (x == "SSW") {205}
  else if (x == "SW") {225}
  else if (x == "WSW") {245}
  else if (x == "W") {270}
  else if (x == "WNW") {295}
  else if (x == "NW") {315}
  else {335}
}
wind = data.frame(gustDir = sydney.weather$WindGustDir,
                  gustSpeed = sydney.weather$WindGustSpeed,
                  dir9am = sydney.weather$WindDir9am,
                  speed9am = sydney.weather$WindSpeed9am,
                  dir3pm = sydney.weather$WindDir3pm,
                  speed3pm = sydney.weather$WindSpeed3pm
                  )
wind$gustDir <- mapply(funWind, wind$gustDir)
wind$dir9am <- mapply(funWind, wind$dir9am)
wind$dir3pm <- mapply(funWind, wind$dir3pm)
```

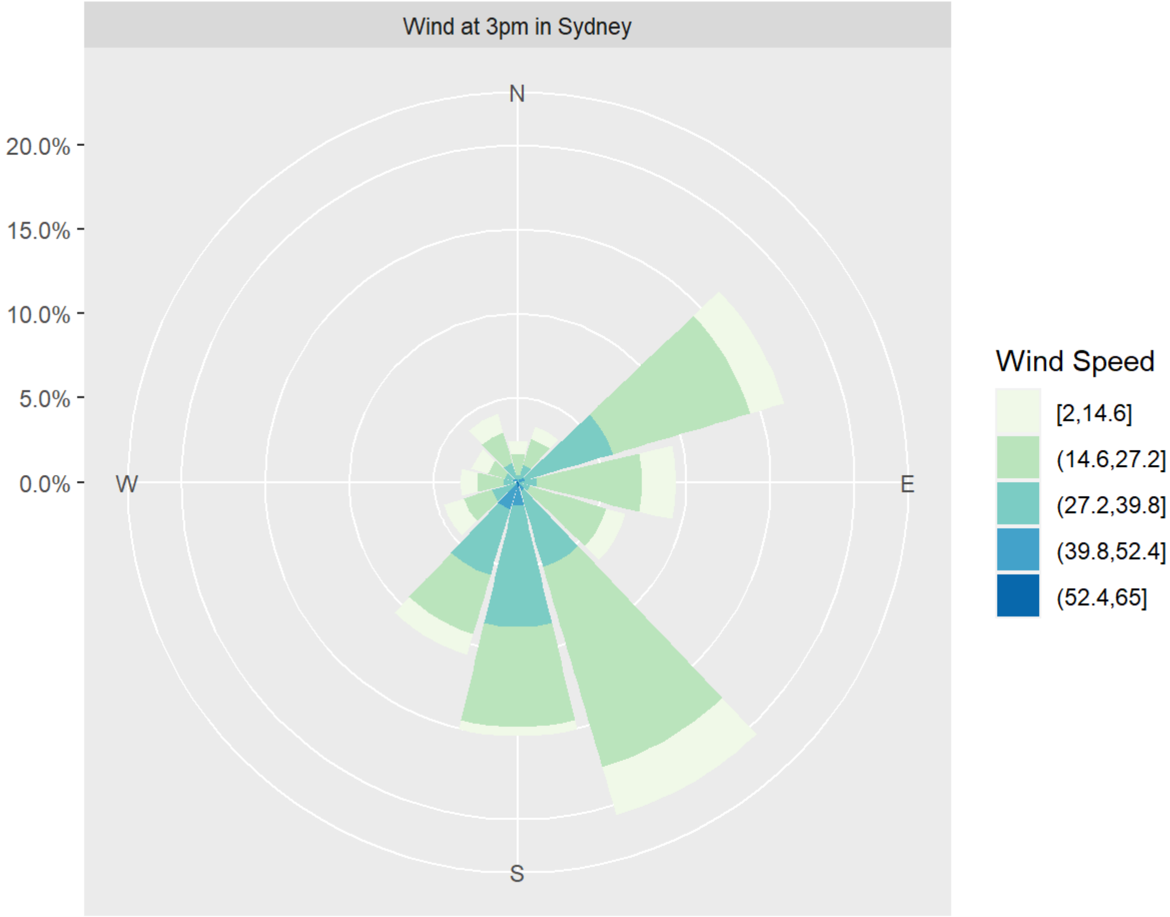
```
# wind rose plot
with(wind, windrose(gustSpeed, gustDir, "Gust Wind in Sydney"))
```



```
with(wind, windrose(speed9am, dir9am, "Wind at 9am in Sydney"))
```



```
with(wind, windrose(speed3pm, dir3pm, "Wind at 3pm in Sydney"))
```



We can see in the plots for Wind at 3pm and at 9pm that they flow in opposite direction at these times. Wind at 9am is predominant from the North West and wind at 3pm is predominat from the South East. Though we found the number of levels in these categorical variable to be too much. So we created the function below to reduce from 16 levels to 4 levels.

```
#help function for combine and reduce the level of the factor variables
funWind2 <- function(x) {
  if (x %in% c("E","ENE","ESE", "NE","SE")) {"E"}
  else if (x %in% c("W","WNW","WSW", "NW", "SW")) {"W"}
  else if (x %in% c("N","NNE","NNW")) {"N"}
  else {"S"}
}

sydney.weather$WindGustDir <- mapply(funWind2, sydney.weather$WindGustDir)
sydney.weather$WindDir9am <- mapply(funWind2, sydney.weather$WindDir9am)
sydney.weather$WindDir3pm <- mapply(funWind2, sydney.weather$WindDir3pm)

# count the level frequency for each of the three factor variables
df1 = count(sydney.weather, 'WindGustDir')
df1[order(df1$freq),]
```

```
##   WindGustDir freq
## 2           N  145
## 4           W  370
## 1           E  434
## 3           S  702
```

```
df2 = count(sydney.weather, 'WindDir9am')
df2[order(df2$freq),]
```

```
##   WindDir9am freq
## 2          N  221
## 1          E  277
## 3          S  419
## 4          W  734
```

```
df3 = count(sydney.weather, 'WindDir3pm')
df3[order(df3$freq),]
```

```
##   WindDir3pm freq
## 2          N  132
## 4          W  217
## 3          S  613
## 1          E  689
```

Now we conver our direction variables to factors and integer variables to numeric to avoid any complications in our modelling. We also remove unwanted columns like Date, Location, RainToday and RainTomorrow. RainTomorrow is removed from the table as it depends on the response we are trying to predict and we cannot have such a highly correlated variable in our model.

```
# convert the wind direction variables from char to factors
sydney.weather$WindGustDir = as.factor(sydney.weather$WindGustDir)
sydney.weather$WindDir9am = as.factor(sydney.weather$WindDir9am)
sydney.weather$WindDir3pm = as.factor(sydney.weather$WindDir3pm)
sydney.weather$Season = as.factor(sydney.weather$Season)

# convert the integer variables to numeric
sydney.weather$WindGustSpeed = as.numeric(sydney.weather$WindGustSpeed)
sydney.weather$WindSpeed9am = as.numeric(sydney.weather$WindSpeed9am)
sydney.weather$WindSpeed3pm = as.numeric(sydney.weather$WindSpeed3pm)
sydney.weather$Humidity9am = as.numeric(sydney.weather$Humidity9am)
sydney.weather$Humidity3pm = as.numeric(sydney.weather$Humidity3pm)
sydney.weather$Cloud9am = as.numeric(sydney.weather$Cloud9am)
sydney.weather$Cloud3pm = as.numeric(sydney.weather$Cloud3pm)

# remove Nan
sydney.weather = na.omit(sydney.weather)

# remove unwanted columns
sydney.weather = sydney.weather[,!(names(sydney.weather) %in% c("Date","Location", "Rain
Today", "RainTomorrow"))]

# create new columns
sydney.weather$TempDiff = with(sydney.weather, Temp3pm - Temp9am)
sydney.weather$AveTemp = with(sydney.weather, (MinTemp + MaxTemp)/2)
```

Section 2.2 Numerical variables, Transformations and Collinearity

Further, we split the refined data into a train and test dataset. The train data is 80% of the refined data that would be used for training the model that we get in the end. The test dataset will be used to fit our model on and to check how accurate our prediction will be.

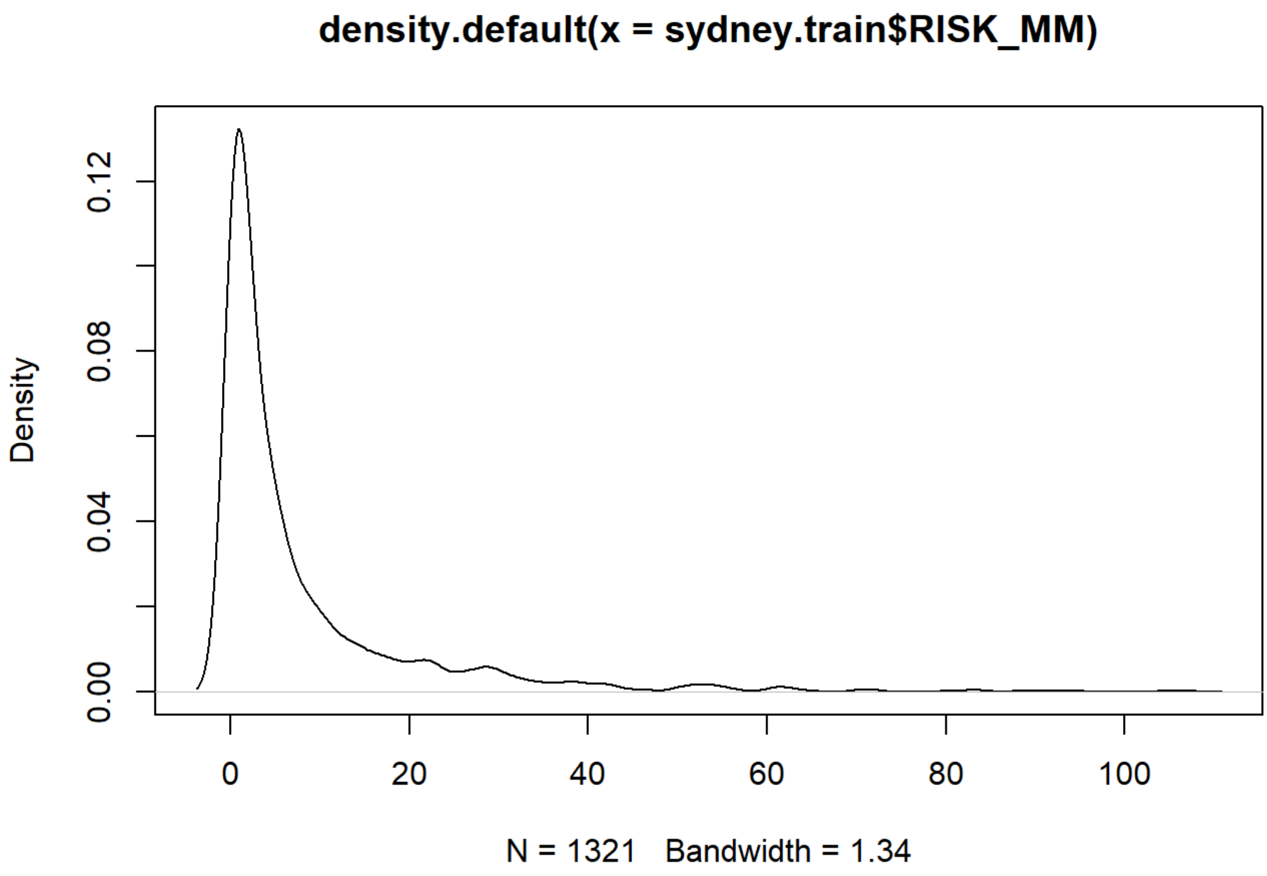
```
# split the data into train data and test
set.seed(19950922)
index = sample(nrow(sydney.weather), size = round(1651 * 0.8))
sydney.train = sydney.weather[index,]
sydney.test = sydney.weather[-index,]

str(sydney.train)
```

```
## 'data.frame':    1321 obs. of  23 variables:
##  $ MinTemp      : num  17.6 9.3 17.7 17.2 13.2 17.3 19.1 20.1 18.4 12.4 ...
##  $ MaxTemp      : num  23.5 15 23.5 25.4 17.3 28.3 31.2 28 25.1 21.1 ...
##  $ Rainfall     : num  0 12.2 0 0 0 0 5.2 51.4 23 46.6 ...
##  $ Evaporation  : num  6.4 3.2 6 2.8 4.6 5 8.6 3.4 7 5 ...
##  $ Sunshine     : num  6.6 0 3.8 9.6 7.7 5 10.7 5.9 6.4 4.4 ...
##  $ WindGustDir  : Factor w/ 4 levels "E","N","S","W": 3 3 2 1 4 2 3 1 1 4 ...
##  $ WindGustSpeed: num  56 65 67 26 59 69 69 52 41 59 ...
##  $ WindDir9am   : Factor w/ 4 levels "E","N","S","W": 4 3 1 4 4 4 4 1 2 4 ...
##  $ WindDir3pm   : Factor w/ 4 levels "E","N","S","W": 3 3 1 1 4 2 1 1 1 4 ...
##  $ WindSpeed9am : num  9 30 15 9 33 9 22 33 17 30 ...
##  $ WindSpeed3pm : num  35 39 30 15 13 39 39 35 26 35 ...
##  $ Humidity9am  : num  61 73 67 82 42 65 71 82 92 61 ...
##  $ Humidity3pm  : num  66 85 75 58 32 36 48 72 59 55 ...
##  $ Pressure9am  : num  1016 1022 1017 1026 1015 ...
##  $ Pressure3pm  : num  1016 1021 1013 1024 1016 ...
##  $ Cloud9am     : num  7 6 7 2 1 6 6 7 7 7 ...
##  $ Cloud3pm     : num  6 7 7 1 6 4 2 7 3 7 ...
##  $ Temp9am      : num  21.2 12.5 20.7 20.9 13.9 20.7 23.3 24.6 18.6 16.6 ...
##  $ Temp3pm      : num  22.3 12.7 21 24.8 16.4 26.5 29.9 25.8 24.4 17.2 ...
##  $ RISK_MM      : num  2 2.6 2 0.2 2.2 4.8 1 4.8 7.8 5.2 ...
##  $ Season       : Factor w/ 2 levels "Dry","Wet": 2 2 1 2 2 1 1 2 1 1 ...
##  $ TempDiff     : num  1.1 0.2 0.3 3.9 2.5 ...
##  $ AveTemp      : num  20.6 12.2 20.6 21.3 15.2 ...
```

Next we check if we need to transform our response or one of our predictors.

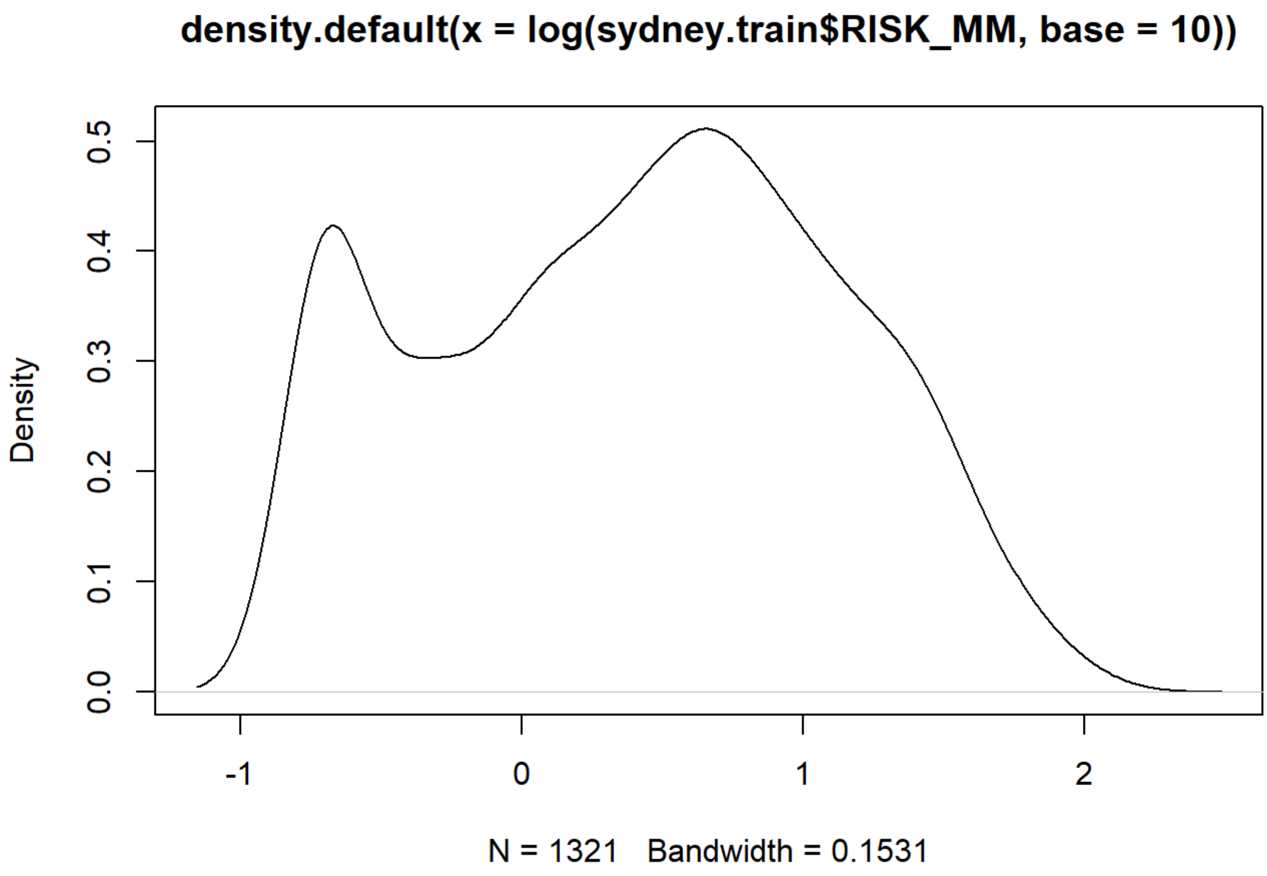
```
# response RISK_MM
plot(density(sydney.train$RISK_MM))
```



```
bestNormalize(sydney.train$RISK_MM)
```

```
## Best Normalizing transformation with 1321 Observations
##  Estimated Normality Statistics (Pearson P / df, lower => more normal):
##  - arcsinh(x): 4.9392
##  - Box-Cox: 3.1635
##  - Exp(x): 115.1185
##  - Log_b(x+a): 3.1388
##  - No transform: 24.3658
##  - orderNorm (ORQ): 3.3053
##  - sqrt(x + a): 7.6099
##  - Yeo-Johnson: 5.2054
## Estimation method: Out-of-sample via CV with 10 folds and 5 repeats
##
## Based off these, bestNormalize chose:
## Standardized Log_b(x + a) Transformation with 1321 nonmissing obs.:
##  Relevant statistics:
##  - a = 0
##  - b = 10
##  - mean (before standardization) = 0.4045364
##  - sd (before standardization) = 0.7160202
```

```
plot(density(log(sydney.train$RISK_MM, base = 10)))
```



After applying transformations to many variables we found only a tranformation for our response variable would be helpful. From the plots above, we see that the first plot shows the density of our response variable, RISK_MM, and that the data was highle skewed to the left. To normalize this data we used the bestNormalize function to find the best normalizing tranformation for this variable. The log trransformation was found to be the best and that we can clearly see from the last plot above and how it is not skewed anymore.

Now we fit a full model with log(RISK_MM) as the response.

```
# first fit a full model and apply backward selection to remove insignificant variables
fit1 = lm(log(RISK_MM, base = 10) ~., data = sydney.train)
summary(step(fit1, direction = "backward", trace = 0))
```

```
##
## Call:
## lm(formula = log(RISK_MM, base = 10) ~ MaxTemp + Rainfall + Sunshine +
##     WindGustSpeed + WindDir9am + WindDir3pm + WindSpeed3pm +
##     Humidity9am + Humidity3pm + Pressure9am + Pressure3pm + Cloud3pm +
##     Temp3pm + Season, data = sydney.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54875 -0.48514  0.00368  0.46194  1.87553
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.000533   3.292637   1.822 0.068622 .
## MaxTemp       -0.022130   0.013753  -1.609 0.107841
## Rainfall       0.007389   0.001833   4.031 5.89e-05 ***
## Sunshine      -0.028575   0.008206  -3.482 0.000514 ***
## WindGustSpeed  0.014813   0.001803   8.216 5.02e-16 ***
## WindDir9amN   -0.119755   0.068373  -1.751 0.080099 .
## WindDir9amS   -0.003600   0.058717  -0.061 0.951116
## WindDir9amW    0.061437   0.056449   1.088 0.276637
## WindDir3pmN   -0.144340   0.072291  -1.997 0.046070 *
## WindDir3pmS   -0.033165   0.049979  -0.664 0.507076
## WindDir3pmW   -0.207864   0.065857  -3.156 0.001635 **
## WindSpeed3pm  -0.008916   0.002486  -3.586 0.000348 ***
## Humidity9am   -0.004035   0.001778  -2.270 0.023372 *
## Humidity3pm    0.012641   0.001877   6.734 2.47e-11 ***
## Pressure9am    0.035893   0.010648   3.371 0.000771 ***
## Pressure3pm   -0.042462   0.010550  -4.025 6.03e-05 ***
## Cloud3pm       0.026987   0.011890   2.270 0.023390 *
## Temp3pm        0.021781   0.015282   1.425 0.154306
## SeasonWet      0.054743   0.037555   1.458 0.145168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6244 on 1302 degrees of freedom
## Multiple R-squared:   0.25,  Adjusted R-squared:  0.2396
## F-statistic: 24.11 on 18 and 1302 DF,  p-value: < 2.2e-16
```

We calculate the RMSE for the full model.

```
# do prediction on the test data and calculate the RMSE
fit1.pre = predict(fit1, sydney.test)
```

```
## Warning in predict.lm(fit1, sydney.test): prediction from a rank-deficient fit
## may be misleading
```

```
rmse(log(sydney.test$RISK_MM), fit1.pre)
```

```
## [1] 1.578366
```

The warning above indicated that there is collinearity among a few predictor variables and we should eliminate such variables from predicting our response. Below we will create a correlation table to find variables with a high collinearity and get rid of them.

```
# based on the warning above, we will detect the collinearity among the numerical variab
les
# calculate the correlation between the variable in our selected model
round(cor(sydney.train[,c(2,3,4,5,7,11,12,13,14,15,17,19,20)]),2)
```

```
##               MaxTemp Rainfall Evaporation Sunshine WindGustSpeed WindSpeed3pm
## MaxTemp          1.00    -0.14        0.50     0.33          0.12         0.05
## Rainfall        -0.14     1.00       -0.10    -0.23          0.05         0.00
## Evaporation      0.50    -0.10        1.00     0.13          0.27         0.24
## Sunshine         0.33    -0.23        0.13     1.00          0.03         0.14
## WindGustSpeed    0.12     0.05        0.27     0.03          1.00         0.66
## WindSpeed3pm     0.05     0.00        0.24     0.14          0.66         1.00
## Humidity9am     -0.25     0.38       -0.31    -0.48         -0.20        -0.19
## Humidity3pm     -0.34     0.28       -0.14    -0.61         -0.04        -0.05
## Pressure9am     -0.46     0.09       -0.36    -0.09         -0.39        -0.22
## Pressure3pm     -0.50     0.13       -0.31    -0.13         -0.32        -0.15
## Cloud3pm        -0.09     0.15        0.03    -0.69          0.06        -0.05
## Temp3pm          0.95    -0.13        0.44     0.35          0.06         0.02
## RISK_MM         -0.12     0.19       -0.04    -0.26          0.16         0.00
##               Humidity9am Humidity3pm Pressure9am Pressure3pm Cloud3pm Temp3pm
## MaxTemp             -0.25       -0.34       -0.46       -0.50    -0.09    0.95
## Rainfall             0.38        0.28        0.09        0.13     0.15   -0.13
## Evaporation         -0.31       -0.14       -0.36       -0.31     0.03    0.44
## Sunshine            -0.48       -0.61       -0.09       -0.13    -0.69    0.35
## WindGustSpeed       -0.20       -0.04       -0.39       -0.32     0.06    0.06
## WindSpeed3pm        -0.19       -0.05       -0.22       -0.15    -0.05    0.02
## Humidity9am          1.00        0.60        0.22        0.23     0.26   -0.20
## Humidity3pm          0.60        1.00        0.15        0.20     0.52   -0.44
## Pressure9am          0.22        0.15        1.00        0.96    -0.02   -0.39
## Pressure3pm          0.23        0.20        0.96        1.00     0.01   -0.45
## Cloud3pm             0.26        0.52       -0.02        0.01     1.00   -0.17
## Temp3pm             -0.20       -0.44       -0.39       -0.45    -0.17    1.00
## RISK_MM              0.18        0.32       -0.02       -0.02     0.23   -0.15
##               RISK_MM
## MaxTemp         -0.12
## Rainfall         0.19
## Evaporation     -0.04
## Sunshine        -0.26
## WindGustSpeed    0.16
## WindSpeed3pm     0.00
## Humidity9am      0.18
## Humidity3pm      0.32
## Pressure9am     -0.02
## Pressure3pm     -0.02
## Cloud3pm         0.23
## Temp3pm         -0.15
## RISK_MM          1.00
```

By observing the table above we choose only those variables which don’t have a high collinearity and create new model with only those predictors. The new model has been created below.

```
fit2 = lm(formula = log(RISK_MM, base = 10) ~ Rainfall + Sunshine + WindGustSpeed + Wind
Dir3pm + WindSpeed3pm + Humidity3pm + Pressure3pm + Cloud3pm + Season, data = sydney.tra
in)
summary(fit2)
```

```
##
## Call:
## lm(formula = log(RISK_MM, base = 10) ~ Rainfall + Sunshine +
##     WindGustSpeed + WindDir3pm + WindSpeed3pm + Humidity3pm +
##     Pressure3pm + Cloud3pm + Season, data = sydney.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52583 -0.48879  0.00745  0.47135  1.84733
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.821906   2.816600   2.067 0.038931 *
## Rainfall       0.006410   0.001761   3.641 0.000283 ***
## Sunshine      -0.024156   0.007589  -3.183 0.001493 **
## WindGustSpeed  0.013689   0.001774   7.718 2.33e-14 ***
## WindDir3pmN   -0.101525   0.068622  -1.479 0.139250
## WindDir3pmS   -0.044424   0.044178  -1.006 0.314805
## WindDir3pmW   -0.180625   0.060838  -2.969 0.003043 **
## WindSpeed3pm  -0.009667   0.002477  -3.903 9.98e-05 ***
## Humidity3pm    0.010124   0.001440   7.031 3.29e-12 ***
## Pressure3pm   -0.006438   0.002743  -2.347 0.019061 *
## Cloud3pm       0.027543   0.011419   2.412 0.016006 *
## SeasonWet      0.049874   0.035762   1.395 0.163366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6288 on 1309 degrees of freedom
## Multiple R-squared:  0.2351, Adjusted R-squared:  0.2287
## F-statistic: 36.58 on 11 and 1309 DF,  p-value: < 2.2e-16
```

```
fit2.pre = predict(fit2, sydney.test)
rmse(log(sydney.test$RISK_MM), fit2.pre)
```

```
## [1] 1.578277
```

```
# no warnings, and RMSE is slightly smaller than the previous model
```

We can see in the model above that our R-squared has decreased but now when we calculate the RMSE no such warnings for collinearity occur. So, we proceed with the predictors chosen above.

Section 2.3 Interactions

In this section we try to find any interactions between our predictor variables that might be helpful for better model. To do this we first create a new data set that contains only the variables chosen by us in the above model. We then will check all the interactions possible to the second degree and use the most significant interactions.

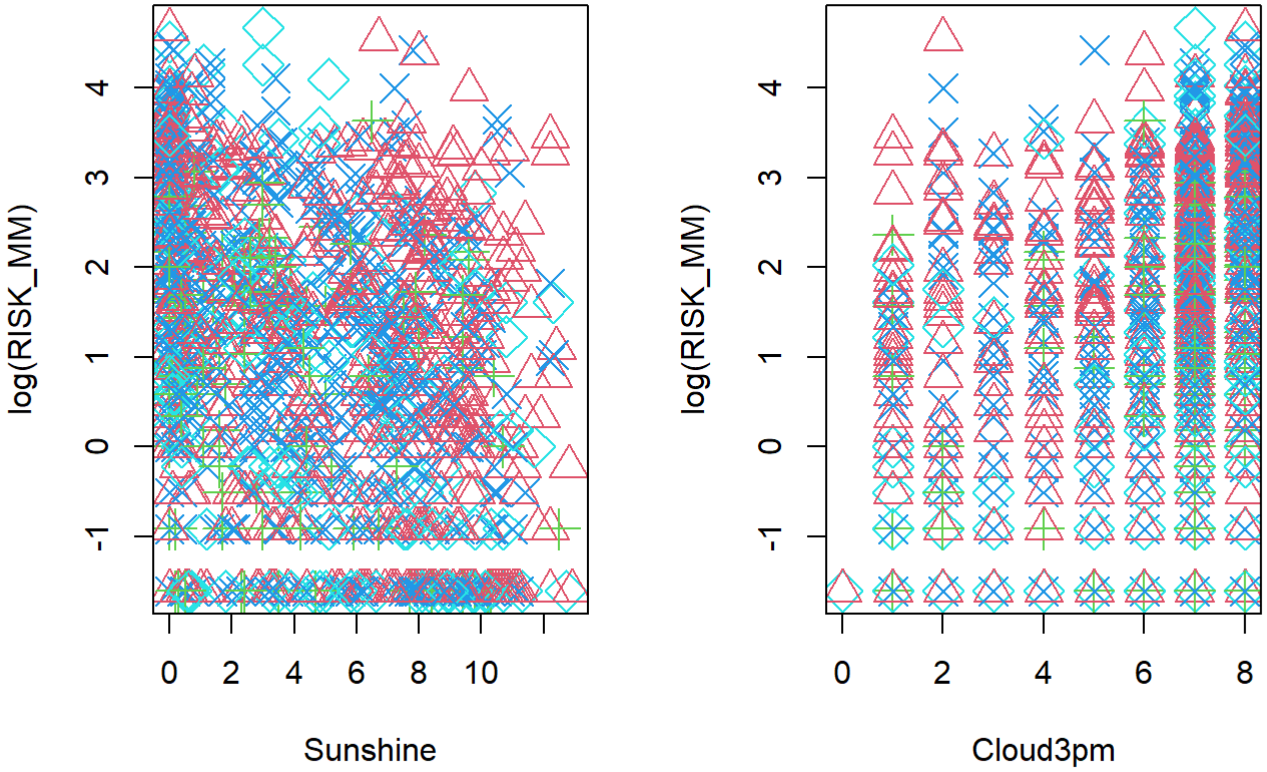
```
sydney.train2 = sydney.train[,c("RISK_MM", "Rainfall", "Humidity3pm", "WindSpeed3pm", "WindGustSpeed", "Sunshine", "WindDir3pm", "Pressure3pm", "Cloud3pm", "Season")]
fit_int2 = lm(formula = log(RISK_MM, base = 10) ~.^2 , data = sydney.train2)
summary(fit_int2)
```

```
##
## Call:
## lm(formula = log(RISK_MM, base = 10) ~ .^2, data = sydney.train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55389 -0.47437  0.01734  0.43575  1.90930
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.579e+01  2.450e+01   0.644  0.51939
## Rainfall                   -5.634e-01  3.198e-01  -1.762  0.07838 .
## Humidity3pm                -2.065e-02  2.460e-01  -0.084  0.93313
## WindSpeed3pm               -5.444e-01  4.063e-01  -1.340  0.18048
## WindGustSpeed               1.850e-01  2.722e-01   0.680  0.49692
## Sunshine                    5.922e-02  1.278e+00   0.046  0.96306
## WindDir3pmN                -1.846e+01  1.069e+01  -1.727  0.08450 .
## WindDir3pmS                -1.192e+01  7.122e+00  -1.674  0.09445 .
## WindDir3pmW                -9.598e+00  1.077e+01  -0.891  0.37284
## Pressure3pm                -1.691e-02  2.389e-02  -0.708  0.47918
## Cloud3pm                    1.226e+00  1.953e+00   0.628  0.53022
## SeasonWet                   2.205e-01  5.792e+00   0.038  0.96964
## Rainfall:Humidity3pm       -8.504e-05  1.657e-04  -0.513  0.60784
## Rainfall:WindSpeed3pm       2.236e-04  2.589e-04   0.863  0.38806
## Rainfall:WindGustSpeed     -2.286e-04  2.043e-04  -1.119  0.26352
## Rainfall:Sunshine          -2.053e-03  9.175e-04  -2.238  0.02542 *
## Rainfall:WindDir3pmN        9.164e-04  1.348e-02   0.068  0.94581
## Rainfall:WindDir3pmS        3.161e-03  4.054e-03   0.780  0.43576
## Rainfall:WindDir3pmW        3.342e-03  6.439e-03   0.519  0.60387
## Rainfall:Pressure3pm        5.704e-04  3.127e-04   1.824  0.06833 .
## Rainfall:Cloud3pm           8.701e-04  1.667e-03   0.522  0.60173
## Rainfall:SeasonWet         -2.963e-03  3.960e-03  -0.748  0.45444
## Humidity3pm:WindSpeed3pm   -2.442e-04  1.924e-04  -1.269  0.20465
## Humidity3pm:WindGustSpeed   5.186e-06  1.411e-04   0.037  0.97068
## Humidity3pm:Sunshine       -1.301e-03  5.500e-04  -2.366  0.01815 *
## Humidity3pm:WindDir3pmN    -6.848e-03  5.964e-03  -1.148  0.25111
## Humidity3pm:WindDir3pmS    -6.800e-03  3.793e-03  -1.793  0.07327 .
## Humidity3pm:WindDir3pmW    -9.404e-03  4.565e-03  -2.060  0.03960 *
## Humidity3pm:Pressure3pm     3.824e-05  2.395e-04   0.160  0.87316
## Humidity3pm:Cloud3pm        1.554e-03  9.105e-04   1.707  0.08809 .
## Humidity3pm:SeasonWet      -1.190e-03  3.032e-03  -0.393  0.69471
## WindSpeed3pm:WindGustSpeed -9.961e-05  1.283e-04  -0.777  0.43753
## WindSpeed3pm:Sunshine       6.013e-04  1.089e-03   0.552  0.58087
## WindSpeed3pm:WindDir3pmN    2.168e-03  9.792e-03   0.221  0.82480
## WindSpeed3pm:WindDir3pmS    3.220e-04  6.530e-03   0.049  0.96069
## WindSpeed3pm:WindDir3pmW   -1.150e-02  8.490e-03  -1.355  0.17578
## WindSpeed3pm:Pressure3pm    5.403e-04  3.977e-04   1.359  0.17453
## WindSpeed3pm:Cloud3pm       1.974e-03  1.756e-03   1.124  0.26115
## WindSpeed3pm:SeasonWet     -7.503e-03  5.163e-03  -1.453  0.14645
## WindGustSpeed:Sunshine     -1.466e-03  8.222e-04  -1.783  0.07487 .
## WindGustSpeed:WindDir3pmN  -5.574e-03  6.463e-03  -0.862  0.38865
## WindGustSpeed:WindDir3pmS  -5.046e-03  4.761e-03  -1.060  0.28949
## WindGustSpeed:WindDir3pmW   1.023e-02  5.514e-03   1.855  0.06390 .
## WindGustSpeed:Pressure3pm  -1.515e-04  2.670e-04  -0.567  0.57050
## WindGustSpeed:Cloud3pm     -1.835e-03  1.251e-03  -1.467  0.14261
## WindGustSpeed:SeasonWet     3.791e-03  3.641e-03   1.041  0.29788
## Sunshine:WindDir3pmN        4.097e-02  3.357e-02   1.221  0.22249
## Sunshine:WindDir3pmS        2.190e-02  1.979e-02   1.106  0.26873
## Sunshine:WindDir3pmW       -1.676e-02  2.984e-02  -0.562  0.57446
## Sunshine:Pressure3pm        1.326e-06  1.247e-03   0.001  0.99915
## Sunshine:Cloud3pm           6.645e-03  3.607e-03   1.842  0.06571 .
## Sunshine:SeasonWet          2.111e-02  1.652e-02   1.278  0.20138
## WindDir3pmN:Pressure3pm     1.798e-02  1.048e-02   1.716  0.08636 .
## WindDir3pmS:Pressure3pm     1.197e-02  6.867e-03   1.744  0.08141 .
## WindDir3pmW:Pressure3pm     9.041e-03  1.056e-02   0.856  0.39195
## WindDir3pmN:Cloud3pm        9.456e-02  4.807e-02   1.967  0.04937 *
## WindDir3pmS:Cloud3pm        2.658e-02  3.134e-02   0.848  0.39662
## WindDir3pmW:Cloud3pm        1.044e-01  3.742e-02   2.790  0.00536 **
## WindDir3pmN:SeasonWet       7.460e-02  1.439e-01   0.519  0.60417
## WindDir3pmS:SeasonWet       1.914e-01  9.005e-02   2.126  0.03374 *
## WindDir3pmW:SeasonWet      -1.643e-01  1.263e-01  -1.301  0.19357
## Pressure3pm:Cloud3pm       -1.311e-03  1.905e-03  -0.689  0.49123
## Pressure3pm:SeasonWet      -3.765e-04  5.633e-03  -0.067  0.94672
## Cloud3pm:SeasonWet          2.723e-02  2.445e-02   1.113  0.26575
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6153 on 1257 degrees of freedom
## Multiple R-squared:  0.2967, Adjusted R-squared:  0.2615
## F-statistic: 8.419 on 63 and 1257 DF,  p-value: < 2.2e-16
```

From the summary above we were able to narrow down a few interactions (WindDir3pm:Cloud3pm, Rainfall:Sunshine, Humidity3pm:Sunshine) that were significant in the model. We now try to plot the significant interactions below to notice any other interactions that might be useful.

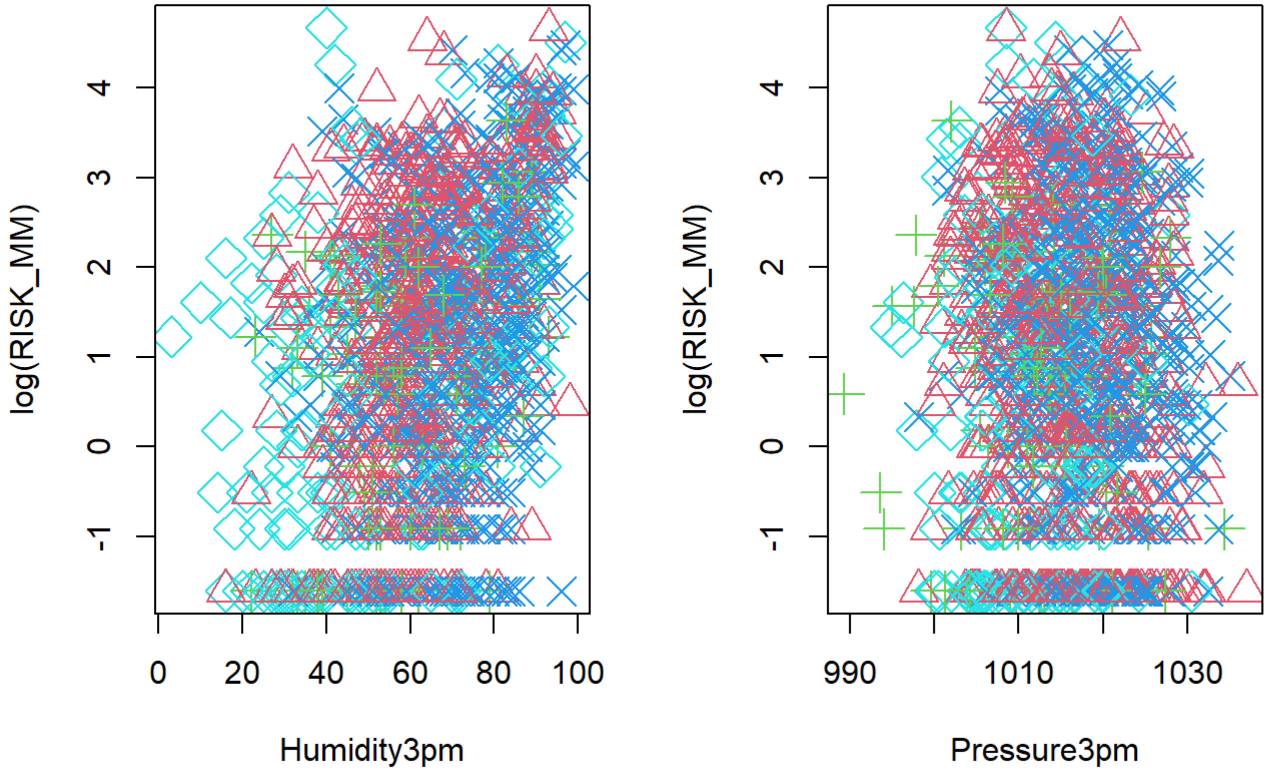
Below is a plot of Sunshine and Cloud at 3pm.

```
par(mfrow = c(1,2))
plot(log(RISK_MM) ~ Sunshine, data = sydney.train, col = as.numeric(WindDir3pm)+ 1, pch
 = as.numeric(WindDir3pm) + 1,  cex = 2)
plot(log(RISK_MM) ~ Cloud3pm, data = sydney.train, col = as.numeric(WindDir3pm)+ 1, pch
 = as.numeric(WindDir3pm) + 1,  cex = 2)
```

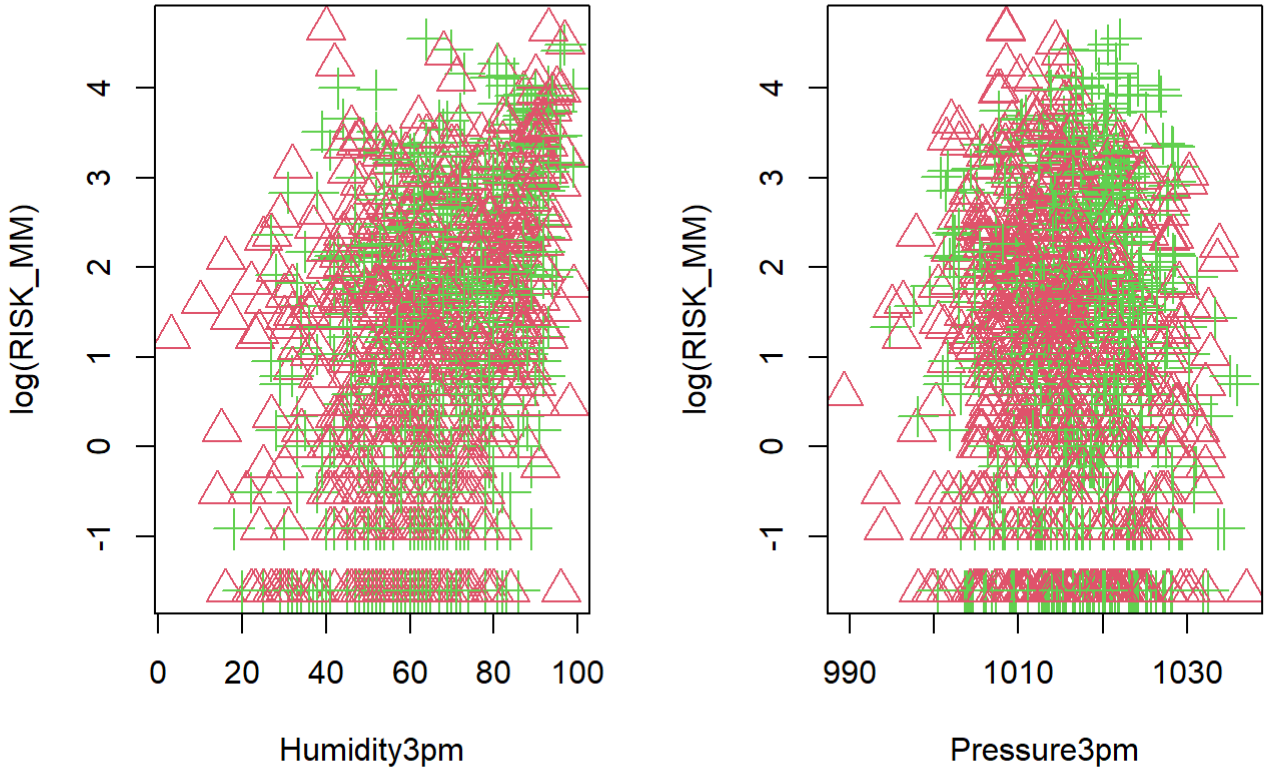


We can notice in the plot above for sunshine that there is a higher risk of rainfall the next day when there is minimal sunshine at 3pm that day. A inverse trend is seen for Cloud at 3pm where there is a higher risk of rainfall the next day when there is more cloud cover at 3pm. This makes sense logically too as more cloud cover at 3pm would result in less sunshite at 3pm. Thus, this interaction is useful.

```
par(mfrow = c(1,2))
plot(log(RISK_MM) ~ Humidity3pm, data = sydney.train, col = as.numeric(WindDir3pm)+ 1, p
ch = as.numeric(WindDir3pm) + 1,  cex = 2)
plot(log(RISK_MM) ~ Pressure3pm, data = sydney.train, col = as.numeric(WindDir3pm)+ 1, p
ch = as.numeric(WindDir3pm) + 1,  cex = 2)
```



```
par(mfrow = c(1,2))
plot(log(RISK_MM) ~ Humidity3pm, data = sydney.train, col = as.numeric(Season)+ 1, pch =
as.numeric(Season) + 1,  cex = 2)
plot(log(RISK_MM) ~ Pressure3pm, data = sydney.train, col = as.numeric(Season)+ 1, pch =
as.numeric(Season) + 1,  cex = 2)
```



THe plots above shows Humidity at 3pm to RISK_MM and Pressure at 3pm to RISK_MM. These plots do not yeild any significant trends to show an interaction.

```
fit3  = lm(formula = log(RISK_MM, base = 10) ~ Rainfall + Sunshine + WindGustSpeed + Win
dDir3pm + WindSpeed3pm + Humidity3pm + Pressure3pm + Cloud3pm + Season + Sunshine:Cloud3
pm + WindDir3pm:Humidity3pm + WindDir3pm:Pressure3pm + WindDir3pm:Cloud3pm + Rainfall:Su
nshine + Humidity3pm:Sunshine, data = sydney.train)
summary(fit3)
```

```
## 
## Call:
## lm(formula = log(RISK_MM, base = 10) ~ Rainfall + Sunshine + 
##     WindGustSpeed + WindDir3pm + WindSpeed3pm + Humidity3pm + 
##     Pressure3pm + Cloud3pm + Season + Sunshine:Cloud3pm + WindDir3pm:Humidity3pm + 
##     WindDir3pm:Pressure3pm + WindDir3pm:Cloud3pm + Rainfall:Sunshine + 
##     Humidity3pm:Sunshine, data = sydney.train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.49395 -0.48179 -0.00236  0.45588  1.83678 
## 
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)              1.393e+01  4.051e+00   3.438 0.000605 ***
## Rainfall                 1.191e-02  2.402e-03   4.958 8.07e-07 ***
## Sunshine                 5.493e-02  2.293e-02   2.396 0.016705 *  
## WindGustSpeed            1.222e-02  1.762e-03   6.937 6.32e-12 ***
## WindDir3pmN             -1.679e+01  8.615e+00  -1.949 0.051456 .  
## WindDir3pmS             -1.625e+01  5.823e+00  -2.791 0.005333 ** 
## WindDir3pmW             -1.147e+00  8.992e+00  -0.128 0.898531    
## WindSpeed3pm            -7.789e-03  2.465e-03  -3.160 0.001613 ** 
## Humidity3pm              2.270e-02  3.048e-03   7.450 1.70e-13 ***
## Pressure3pm             -1.494e-02  3.957e-03  -3.777 0.000166 ***
## Cloud3pm                -1.989e-02  2.928e-02  -0.679 0.497119    
## SeasonWet                5.334e-02  3.520e-02   1.515 0.129914    
## Sunshine:Cloud3pm        4.555e-03  3.273e-03   1.392 0.164275    
## WindDir3pmN:Humidity3pm -9.062e-03  4.859e-03  -1.865 0.062405 .  
## WindDir3pmS:Humidity3pm -8.464e-03  3.202e-03  -2.643 0.008306 ** 
## WindDir3pmW:Humidity3pm -7.032e-03  3.471e-03  -2.026 0.042980 *  
## WindDir3pmN:Pressure3pm  1.667e-02  8.554e-03   1.948 0.051607 .  
## WindDir3pmS:Pressure3pm  1.644e-02  5.715e-03   2.877 0.004079 ** 
## WindDir3pmW:Pressure3pm  7.663e-04  8.932e-03   0.086 0.931642    
## WindDir3pmN:Cloud3pm     5.701e-02  4.022e-02   1.417 0.156603    
## WindDir3pmS:Cloud3pm     8.187e-03  2.197e-02   0.373 0.709507    
## WindDir3pmW:Cloud3pm     9.417e-02  2.886e-02   3.263 0.001132 ** 
## Rainfall:Sunshine       -2.205e-03  5.897e-04  -3.739 0.000193 ***
## Sunshine:Humidity3pm    -1.603e-03  3.770e-04  -4.251 2.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.617 on 1297 degrees of freedom
## Multiple R-squared:  0.2705, Adjusted R-squared:  0.2576 
## F-statistic: 20.91 on 23 and 1297 DF,  p-value: < 2.2e-16
```

Our results with these interactions now yeild a better Multiple and Adjusted R-squared.

```
fit3.pre = predict(fit3, sydney.test)
rmse(log(sydney.test$RISK_MM), fit3.pre)
```

```
## [1] 1.55647
```

By calculating the RMSE we can also observe that the RMSE has decreased significantly and there aren’t any warning indicating correlation among the predictor variables.

Below will try to apply backward AIC step on the model again to optimize our final model.

```
fit4 = step(fit3, direction = "backward", trace = 0)
summary(fit4)
```

```
##
## Call:
## lm(formula = log(RISK_MM, base = 10) ~ Rainfall + Sunshine +
##     WindGustSpeed + WindDir3pm + WindSpeed3pm + Humidity3pm +
##     Pressure3pm + Cloud3pm + Season + WindDir3pm:Humidity3pm +
##     WindDir3pm:Pressure3pm + WindDir3pm:Cloud3pm + Rainfall:Sunshine +
##     Sunshine:Humidity3pm, data = sydney.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49929 -0.48150  0.00747  0.44796  1.86528
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.350e+01  4.041e+00   3.341 0.000859 ***
## Rainfall                 1.186e-02  2.403e-03   4.936 9.02e-07 ***
## Sunshine                 6.602e-02  2.151e-02   3.070 0.002187 **
## WindGustSpeed            1.236e-02  1.760e-03   7.022 3.51e-12 ***
## WindDir3pmN             -1.673e+01  8.618e+00  -1.941 0.052484 .
## WindDir3pmS             -1.637e+01  5.825e+00  -2.810 0.005024 **
## WindDir3pmW             -6.001e-01  8.986e+00  -0.067 0.946768
## WindSpeed3pm            -7.992e-03  2.461e-03  -3.247 0.001196 **
## Humidity3pm              2.090e-02  2.760e-03   7.574 6.83e-14 ***
## Pressure3pm             -1.464e-02  3.952e-03  -3.705 0.000220 ***
## Cloud3pm                 1.542e-02  1.462e-02   1.055 0.291821
## SeasonWet                5.611e-02  3.515e-02   1.596 0.110712
## WindDir3pmN:Humidity3pm -8.564e-03  4.848e-03  -1.767 0.077525 .
## WindDir3pmS:Humidity3pm -8.084e-03  3.191e-03  -2.533 0.011421 *
## WindDir3pmW:Humidity3pm -6.691e-03  3.463e-03  -1.932 0.053596 .
## WindDir3pmN:Pressure3pm  1.662e-02  8.557e-03   1.942 0.052336 .
## WindDir3pmS:Pressure3pm  1.654e-02  5.717e-03   2.894 0.003865 **
## WindDir3pmW:Pressure3pm  2.468e-04  8.927e-03   0.028 0.977953
## WindDir3pmN:Cloud3pm     4.873e-02  3.980e-02   1.224 0.221024
## WindDir3pmS:Cloud3pm     6.576e-03  2.195e-02   0.300 0.764544
## WindDir3pmW:Cloud3pm     8.849e-02  2.858e-02   3.096 0.002004 **
## Rainfall:Sunshine       -2.150e-03  5.886e-04  -3.653 0.000270 ***
## Sunshine:Humidity3pm    -1.326e-03  3.205e-04  -4.138 3.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6172 on 1298 degrees of freedom
## Multiple R-squared:  0.2694, Adjusted R-squared:  0.257
## F-statistic: 21.76 on 22 and 1298 DF,  p-value: < 2.2e-16
```

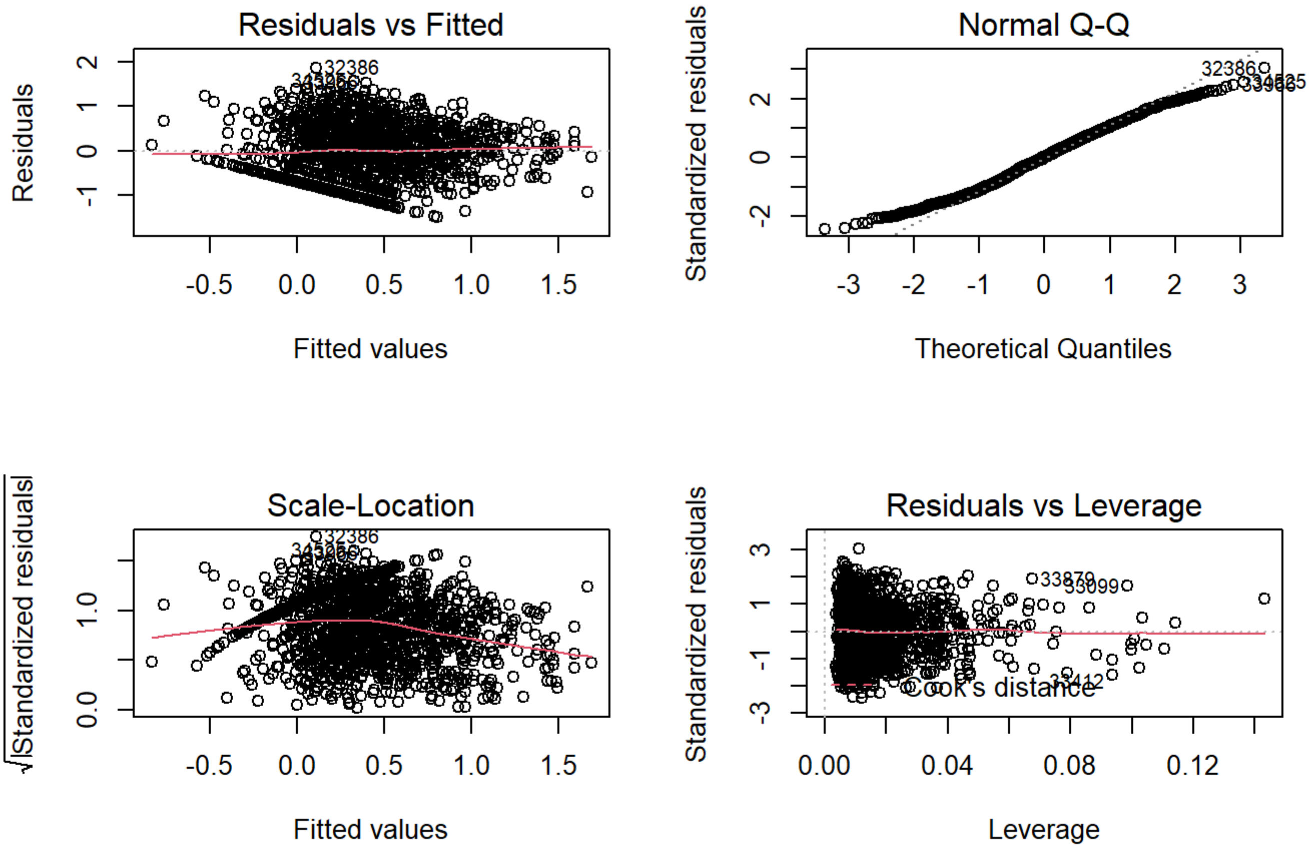
```
fit4.pre = predict(fit4, sydney.test)
rmse(log(sydney.test$RISK_MM), fit4.pre)
```

```
## [1] 1.557931
```

By applying step transformation to the final model we observe a very slight decrease in the R-squared which can be ignored.

Section 3 Results and Discusstion

```
par(mfrow = c(2,2))
plot(fit4)
```



Section 4 Appendix

At the beginning, we choose RISK_MM (The amount of rainfall recorded for next day in mm) as the response and all other variables as the predictors to fit the full model

```
lm.full = lm(RISK_MM~., data = sydney.train)
summary(lm.full)
```

```
## 
## Call:
## lm(formula = RISK_MM ~ ., data = sydney.train)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -22.197  -6.495  -2.674   2.486  88.933 
## 
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   94.76849   65.35925   1.450 0.147311    
## MinTemp        0.45387    0.31455   1.443 0.149295    
## MaxTemp       -0.28806    0.28271  -1.019 0.308429    
## Rainfall       0.12031    0.03555   3.384 0.000736 ***
## Evaporation   -0.04354    0.16765  -0.260 0.795120    
## Sunshine      -0.17059    0.18606  -0.917 0.359395    
## WindGustDirN  -2.81032    1.40793  -1.996 0.046137 *  
## WindGustDirS  -2.49450    0.99652  -2.503 0.012429 *  
## WindGustDirW  -2.55261    1.17650  -2.170 0.030214 *  
## WindGustSpeed  0.29199    0.03545   8.237 4.29e-16 ***
## WindDir9amN   -0.37059    1.32519  -0.280 0.779791    
## WindDir9amS    1.82770    1.16019   1.575 0.115419    
## WindDir9amW    1.26273    1.16035   1.088 0.276696    
## WindDir3pmN   -4.44326    1.42757  -3.112 0.001896 ** 
## WindDir3pmS   -1.47490    1.02073  -1.445 0.148715    
## WindDir3pmW   -0.85693    1.29563  -0.661 0.508475    
## WindSpeed9am  -0.00822    0.05038  -0.163 0.870429    
## WindSpeed3pm  -0.20799    0.05041  -4.126 3.93e-05 ***
## Humidity9am   -0.11581    0.04598  -2.519 0.011891 *  
## Humidity3pm    0.22387    0.04473   5.005 6.36e-07 ***
## Pressure9am    0.67856    0.22806   2.975 0.002981 ** 
## Pressure3pm   -0.77455    0.22352  -3.465 0.000547 ***
## Cloud9am       0.18477    0.21305   0.867 0.385972    
## Cloud3pm       0.18748    0.22756   0.824 0.410176    
## Temp9am       -0.55873    0.35522  -1.573 0.115978    
## Temp3pm        0.17687    0.32840   0.539 0.590261    
## SeasonWet      1.98113    0.74920   2.644 0.008284 ** 
## TempDiff            NA         NA      NA       NA    
## AveTemp             NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 11.83 on 1294 degrees of freedom
## Multiple R-squared:  0.1943, Adjusted R-squared:  0.1781 
## F-statistic:    12 on 26 and 1294 DF,  p-value: < 2.2e-16
```

```
lm.step = step(lm.full,trace = FALSE)
summary(lm.step)
```

```
##
## Call:
## lm(formula = RISK_MM ~ MinTemp + Rainfall + Sunshine + WindGustDir +
##     WindGustSpeed + WindDir3pm + WindSpeed3pm + Humidity9am +
##     Humidity3pm + Pressure9am + Pressure3pm + Temp9am + Season,
##     data = sydney.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.174  -6.567  -2.597   2.778  89.221
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   90.98841   64.09747   1.420 0.155983
## MinTemp        0.51391    0.28044   1.832 0.067106 .
## Rainfall       0.11885    0.03483   3.412 0.000665 ***
## Sunshine      -0.32002    0.12787  -2.503 0.012443 *
## WindGustDirN  -2.90295    1.39901  -2.075 0.038183 *
## WindGustDirS  -2.19103    0.96127  -2.279 0.022810 *
## WindGustDirW  -2.57710    1.14479  -2.251 0.024542 *
## WindGustSpeed  0.28506    0.03438   8.290 2.78e-16 ***
## WindDir3pmN   -4.67551    1.40854  -3.319 0.000927 ***
## WindDir3pmS   -1.16340    1.00616  -1.156 0.247779
## WindDir3pmW   -0.71029    1.28415  -0.553 0.580275
## WindSpeed3pm  -0.20478    0.04758  -4.304 1.80e-05 ***
## Humidity9am   -0.10205    0.03726  -2.739 0.006248 **
## Humidity3pm    0.22217    0.02950   7.532 9.32e-14 ***
## Pressure9am    0.63362    0.20712   3.059 0.002264 **
## Pressure3pm   -0.72350    0.20057  -3.607 0.000321 ***
## Temp9am       -0.74893    0.27843  -2.690 0.007241 **
## SeasonWet      1.90909    0.69579   2.744 0.006157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.82 on 1303 degrees of freedom
## Multiple R-squared:  0.1903, Adjusted R-squared:  0.1798
## F-statistic: 18.02 on 17 and 1303 DF,  p-value: < 2.2e-16
```

We find the most significant predictors after doing forward AIC.

Though we note that we can’t use RainTomorrow as a predictor since it highly correlated with RISK_MM. From the definition of RainTomorrow we understand why this so, since a high RISK_MM results in Yes for RainTomorrow. Thus RainTomorrow is a response for RISK_MM instead. So we continue finding the best model without RainTomorrow.

```
sydney.train3 = sydney.train[,c("RISK_MM", "Rainfall", "Evaporation", "Humidity3pm", "Te
mp9am", "WindSpeed3pm", "WindGustSpeed",
    "WindSpeed9am")]
newmod = lm(formula = RISK_MM ~.,
    data = sydney.train3)
summary(newmod)
```

```
##
## Call:
## lm(formula = RISK_MM ~ ., data = sydney.train3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.200  -6.627  -2.789   2.003  89.599
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -12.798274   2.206421  -5.800 8.28e-09 ***
## Rainfall        0.111897   0.033431   3.347  0.00084 ***
## Evaporation    -0.070021   0.147242  -0.476  0.63447
## Humidity3pm     0.222120   0.020394  10.891  < 2e-16 ***
## Temp9am        -0.057122   0.091212  -0.626  0.53125
## WindSpeed3pm   -0.237407   0.047360  -5.013 6.10e-07 ***
## WindGustSpeed   0.270421   0.032254   8.384  < 2e-16 ***
## WindSpeed9am    0.006893   0.045113   0.153  0.87859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12 on 1313 degrees of freedom
## Multiple R-squared:  0.1587, Adjusted R-squared:  0.1542
## F-statistic: 35.38 on 7 and 1313 DF,  p-value: < 2.2e-16
```

We observe he R-squared is very low now. We try to take an interaction to the power two to see if it improves.

```
newmod_int2 = lm(RISK_MM ~.^2,
    data = sydney.train3)
summary(newmod_int2)
```

```
## 
## Call:
## lm(formula = RISK_MM ~ .^2, data = sydney.train3)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -28.098  -6.173  -2.630   2.358  88.127 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                -1.791e+01  9.635e+00  -1.858 0.063354 .  
## Rainfall                   -2.678e-01  2.812e-01  -0.952 0.341066    
## Evaporation                 1.762e+00  9.913e-01   1.778 0.075701 .  
## Humidity3pm                 9.308e-02  1.053e-01   0.884 0.376809    
## Temp9am                    -3.580e-01  5.226e-01  -0.685 0.493466    
## WindSpeed3pm               -5.687e-01  3.010e-01  -1.890 0.059046 .  
## WindGustSpeed               8.319e-01  2.078e-01   4.004 6.59e-05 ***
## WindSpeed9am               -1.904e-01  3.180e-01  -0.599 0.549410    
## Rainfall:Evaporation        1.232e-02  1.298e-02   0.949 0.342766    
## Rainfall:Humidity3pm        8.830e-03  2.438e-03   3.621 0.000305 ***
## Rainfall:Temp9am           -2.734e-02  9.033e-03  -3.027 0.002521 ** 
## Rainfall:WindSpeed3pm       3.446e-03  5.102e-03   0.675 0.499549    
## Rainfall:WindGustSpeed      2.245e-03  3.697e-03   0.607 0.543694    
## Rainfall:WindSpeed9am      -2.640e-03  4.242e-03  -0.622 0.533799    
## Evaporation:Humidity3pm    -2.317e-02  8.860e-03  -2.615 0.009017 ** 
## Evaporation:Temp9am         2.851e-02  2.971e-02   0.960 0.337400    
## Evaporation:WindSpeed3pm    2.307e-03  1.997e-02   0.116 0.908058    
## Evaporation:WindGustSpeed  -3.092e-02  1.471e-02  -2.102 0.035730 *  
## Evaporation:WindSpeed9am    2.451e-02  1.907e-02   1.285 0.198895    
## Humidity3pm:Temp9am         1.121e-02  5.606e-03   2.000 0.045722 *  
## Humidity3pm:WindSpeed3pm   -8.082e-04  2.668e-03  -0.303 0.762044    
## Humidity3pm:WindGustSpeed  -1.369e-03  1.880e-03  -0.728 0.466863    
## Humidity3pm:WindSpeed9am    5.051e-03  2.613e-03   1.933 0.053419 .  
## Temp9am:WindSpeed3pm        2.199e-02  1.177e-02   1.868 0.062006 .  
## Temp9am:WindGustSpeed      -1.486e-02  7.841e-03  -1.895 0.058308 .  
## Temp9am:WindSpeed9am       -1.180e-02  1.209e-02  -0.975 0.329539    
## WindSpeed3pm:WindGustSpeed -1.271e-03  2.486e-03  -0.511 0.609108    
## WindSpeed3pm:WindSpeed9am   3.134e-04  5.058e-03   0.062 0.950599    
## WindGustSpeed:WindSpeed9am -9.030e-04  3.794e-03  -0.238 0.811904    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 11.82 on 1292 degrees of freedom
## Multiple R-squared:  0.1962, Adjusted R-squared:  0.1788 
## F-statistic: 11.26 on 28 and 1292 DF,  p-value: < 2.2e-16
```

From researching online we observe that WindGustSpeed should not be affecting rainfall the next day. SO we have to reconsider our model and eliminate other predictors.

We observe he R-squared is very low now so forward AIC is not the best step method the use maybe.

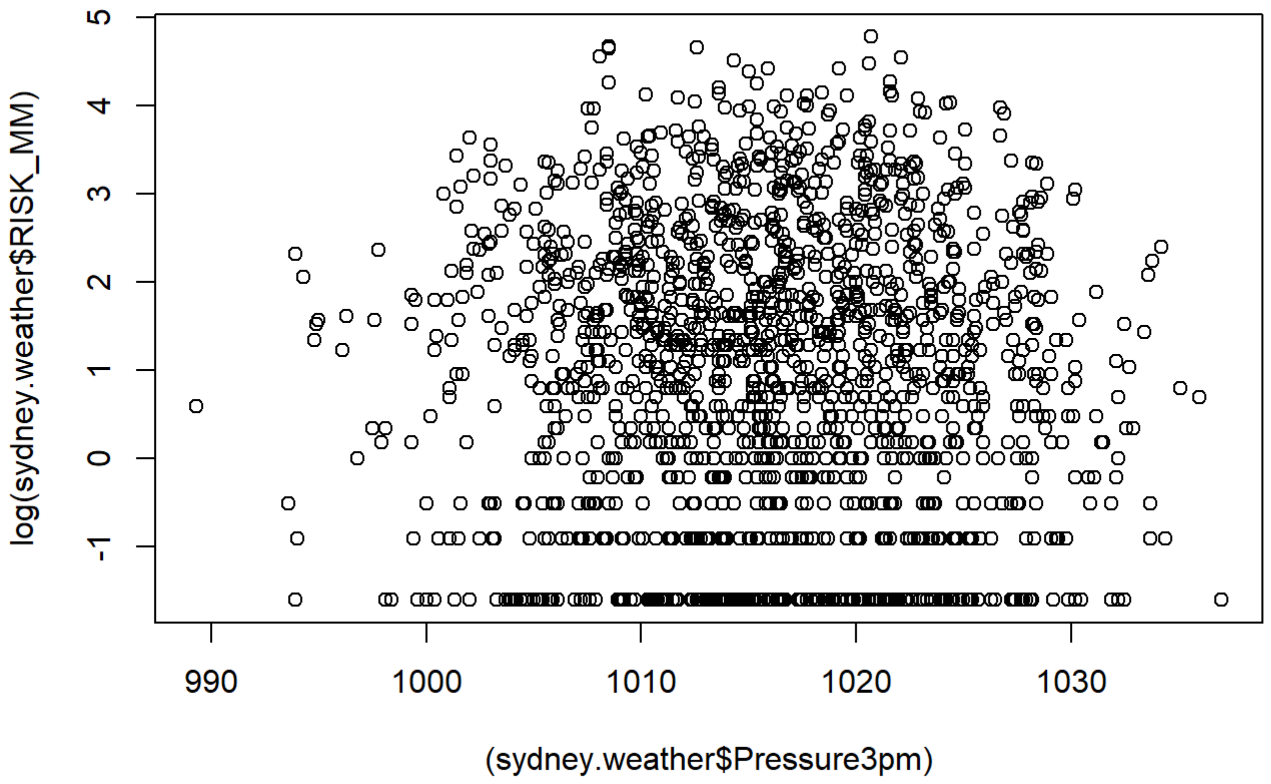
We found backward AIC to be the best step method and then came up with a new dataset - sydney.train2 with only variables chosen for the best model. This here is the summary for an interaction model of the power 3. We found this interaction useless as none of the third degree interactions signifcantly increased our R squared so we ignored it.

```
fit_int2 = lm(formula = log(RISK_MM, base = 10) ~.^2 , data = sydney.train2)
summary(fit_int2)
```

```
##
## Call:
## lm(formula = log(RISK_MM, base = 10) ~ .^2, data = sydney.train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55389 -0.47437  0.01734  0.43575  1.90930
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.579e+01  2.450e+01   0.644  0.51939
## Rainfall                   -5.634e-01  3.198e-01  -1.762  0.07838 .
## Humidity3pm                -2.065e-02  2.460e-01  -0.084  0.93313
## WindSpeed3pm               -5.444e-01  4.063e-01  -1.340  0.18048
## WindGustSpeed               1.850e-01  2.722e-01   0.680  0.49692
## Sunshine                    5.922e-02  1.278e+00   0.046  0.96306
## WindDir3pmN                -1.846e+01  1.069e+01  -1.727  0.08450 .
## WindDir3pmS                -1.192e+01  7.122e+00  -1.674  0.09445 .
## WindDir3pmW                -9.598e+00  1.077e+01  -0.891  0.37284
## Pressure3pm                -1.691e-02  2.389e-02  -0.708  0.47918
## Cloud3pm                    1.226e+00  1.953e+00   0.628  0.53022
## SeasonWet                   2.205e-01  5.792e+00   0.038  0.96964
## Rainfall:Humidity3pm       -8.504e-05  1.657e-04  -0.513  0.60784
## Rainfall:WindSpeed3pm       2.236e-04  2.589e-04   0.863  0.38806
## Rainfall:WindGustSpeed     -2.286e-04  2.043e-04  -1.119  0.26352
## Rainfall:Sunshine          -2.053e-03  9.175e-04  -2.238  0.02542 *
## Rainfall:WindDir3pmN        9.164e-04  1.348e-02   0.068  0.94581
## Rainfall:WindDir3pmS        3.161e-03  4.054e-03   0.780  0.43576
## Rainfall:WindDir3pmW        3.342e-03  6.439e-03   0.519  0.60387
## Rainfall:Pressure3pm        5.704e-04  3.127e-04   1.824  0.06833 .
## Rainfall:Cloud3pm           8.701e-04  1.667e-03   0.522  0.60173
## Rainfall:SeasonWet         -2.963e-03  3.960e-03  -0.748  0.45444
## Humidity3pm:WindSpeed3pm   -2.442e-04  1.924e-04  -1.269  0.20465
## Humidity3pm:WindGustSpeed   5.186e-06  1.411e-04   0.037  0.97068
## Humidity3pm:Sunshine       -1.301e-03  5.500e-04  -2.366  0.01815 *
## Humidity3pm:WindDir3pmN    -6.848e-03  5.964e-03  -1.148  0.25111
## Humidity3pm:WindDir3pmS    -6.800e-03  3.793e-03  -1.793  0.07327 .
## Humidity3pm:WindDir3pmW    -9.404e-03  4.565e-03  -2.060  0.03960 *
## Humidity3pm:Pressure3pm     3.824e-05  2.395e-04   0.160  0.87316
## Humidity3pm:Cloud3pm        1.554e-03  9.105e-04   1.707  0.08809 .
## Humidity3pm:SeasonWet      -1.190e-03  3.032e-03  -0.393  0.69471
## WindSpeed3pm:WindGustSpeed -9.961e-05  1.283e-04  -0.777  0.43753
## WindSpeed3pm:Sunshine       6.013e-04  1.089e-03   0.552  0.58087
## WindSpeed3pm:WindDir3pmN    2.168e-03  9.792e-03   0.221  0.82480
## WindSpeed3pm:WindDir3pmS    3.220e-04  6.530e-03   0.049  0.96069
## WindSpeed3pm:WindDir3pmW   -1.150e-02  8.490e-03  -1.355  0.17578
## WindSpeed3pm:Pressure3pm    5.403e-04  3.977e-04   1.359  0.17453
## WindSpeed3pm:Cloud3pm       1.974e-03  1.756e-03   1.124  0.26115
## WindSpeed3pm:SeasonWet     -7.503e-03  5.163e-03  -1.453  0.14645
## WindGustSpeed:Sunshine     -1.466e-03  8.222e-04  -1.783  0.07487 .
## WindGustSpeed:WindDir3pmN  -5.574e-03  6.463e-03  -0.862  0.38865
## WindGustSpeed:WindDir3pmS  -5.046e-03  4.761e-03  -1.060  0.28949
## WindGustSpeed:WindDir3pmW   1.023e-02  5.514e-03   1.855  0.06390 .
## WindGustSpeed:Pressure3pm  -1.515e-04  2.670e-04  -0.567  0.57050
## WindGustSpeed:Cloud3pm     -1.835e-03  1.251e-03  -1.467  0.14261
## WindGustSpeed:SeasonWet     3.791e-03  3.641e-03   1.041  0.29788
## Sunshine:WindDir3pmN        4.097e-02  3.357e-02   1.221  0.22249
## Sunshine:WindDir3pmS        2.190e-02  1.979e-02   1.106  0.26873
## Sunshine:WindDir3pmW       -1.676e-02  2.984e-02  -0.562  0.57446
## Sunshine:Pressure3pm        1.326e-06  1.247e-03   0.001  0.99915
## Sunshine:Cloud3pm           6.645e-03  3.607e-03   1.842  0.06571 .
## Sunshine:SeasonWet          2.111e-02  1.652e-02   1.278  0.20138
## WindDir3pmN:Pressure3pm     1.798e-02  1.048e-02   1.716  0.08636 .
## WindDir3pmS:Pressure3pm     1.197e-02  6.867e-03   1.744  0.08141 .
## WindDir3pmW:Pressure3pm     9.041e-03  1.056e-02   0.856  0.39195
## WindDir3pmN:Cloud3pm        9.456e-02  4.807e-02   1.967  0.04937 *
## WindDir3pmS:Cloud3pm        2.658e-02  3.134e-02   0.848  0.39662
## WindDir3pmW:Cloud3pm        1.044e-01  3.742e-02   2.790  0.00536 **
## WindDir3pmN:SeasonWet       7.460e-02  1.439e-01   0.519  0.60417
## WindDir3pmS:SeasonWet       1.914e-01  9.005e-02   2.126  0.03374 *
## WindDir3pmW:SeasonWet      -1.643e-01  1.263e-01  -1.301  0.19357
## Pressure3pm:Cloud3pm       -1.311e-03  1.905e-03  -0.689  0.49123
## Pressure3pm:SeasonWet      -3.765e-04  5.633e-03  -0.067  0.94672
## Cloud3pm:SeasonWet          2.723e-02  2.445e-02   1.113  0.26575
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6153 on 1257 degrees of freedom
## Multiple R-squared:  0.2967, Adjusted R-squared:  0.2615
## F-statistic: 8.419 on 63 and 1257 DF,  p-value: < 2.2e-16
```

We try to see if there is any correlation between Pressure at 3pm and the log of RISK_MM.

```
plot(x = (sydney.weather$Pressure3pm), y = log(sydney.weather$RISK_MM))
```



Next, we also try to see any correlation using vif.

```
car::vif(newmod)
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
##      Rainfall   Evaporation   Humidity3pm       Temp9am  WindSpeed3pm 
##      1.111696      1.577987      1.106994      1.558344      1.966264 
## WindGustSpeed  WindSpeed9am 
##      1.893951      1.451933
```