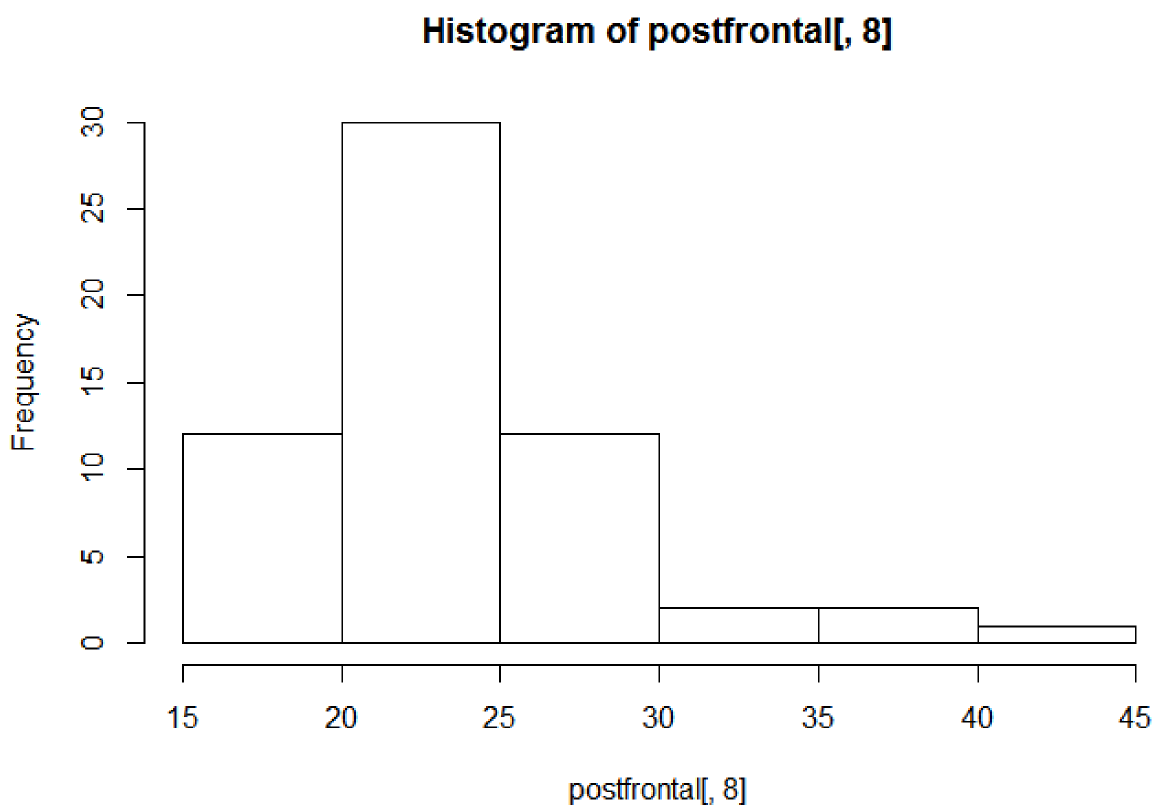
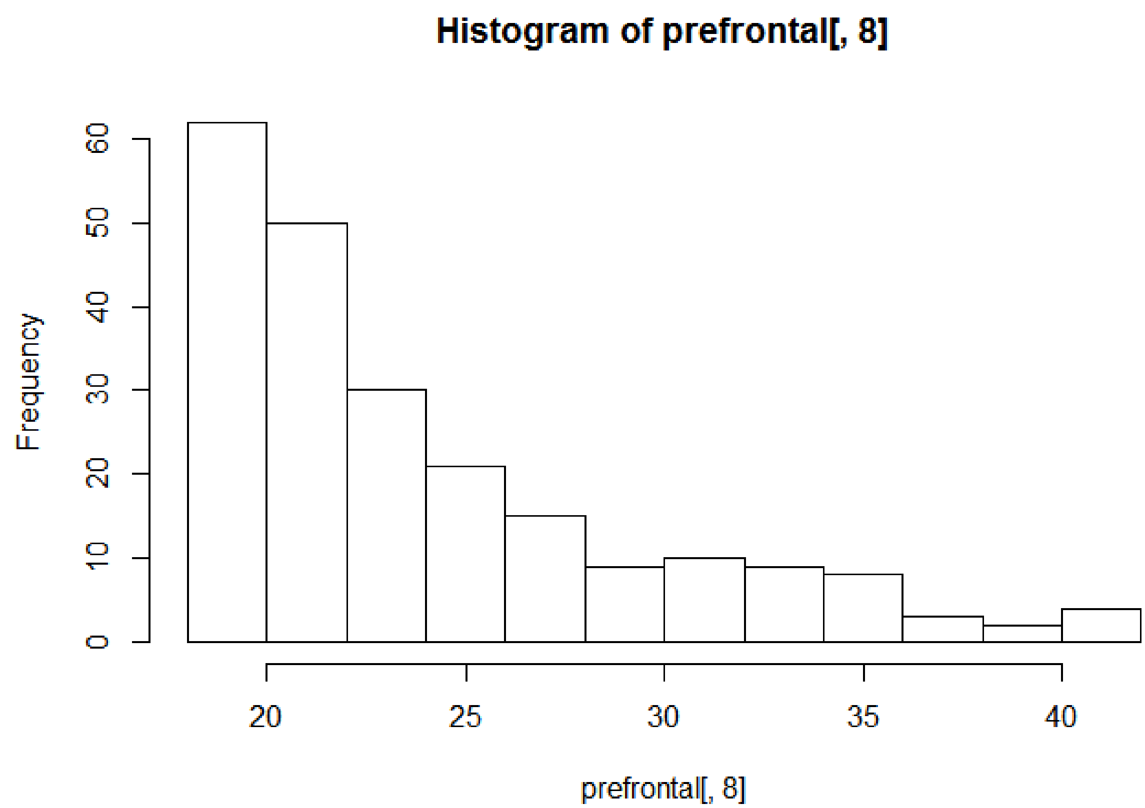
Report

1. Descriptive statistics for the 0-12 hour peak wind speed variable

Statistic	Prefrontal	Postfrontal
Minimum	18.00	18.00
1st quartile	19.50	20.60
Median	21.60	22.60
Mean	23.82	23.87
3rd quartile	26.80	26.50
Maximum	41.20	41.20

To further study the distribution of the variable, histograms for the 0-12 hour wind speed variable we plotted histograms for the variable in both datasets.





From the above histograms we can see that the variable is rightly skewed for both datasets. The histograms are also asymmetric and since there are no secluded values it shows that the variable does not have outliers in both datasets. In both cases, the variable distribution somehow forms a bell-shaped pattern showing that the variable could be from a normal distribution.

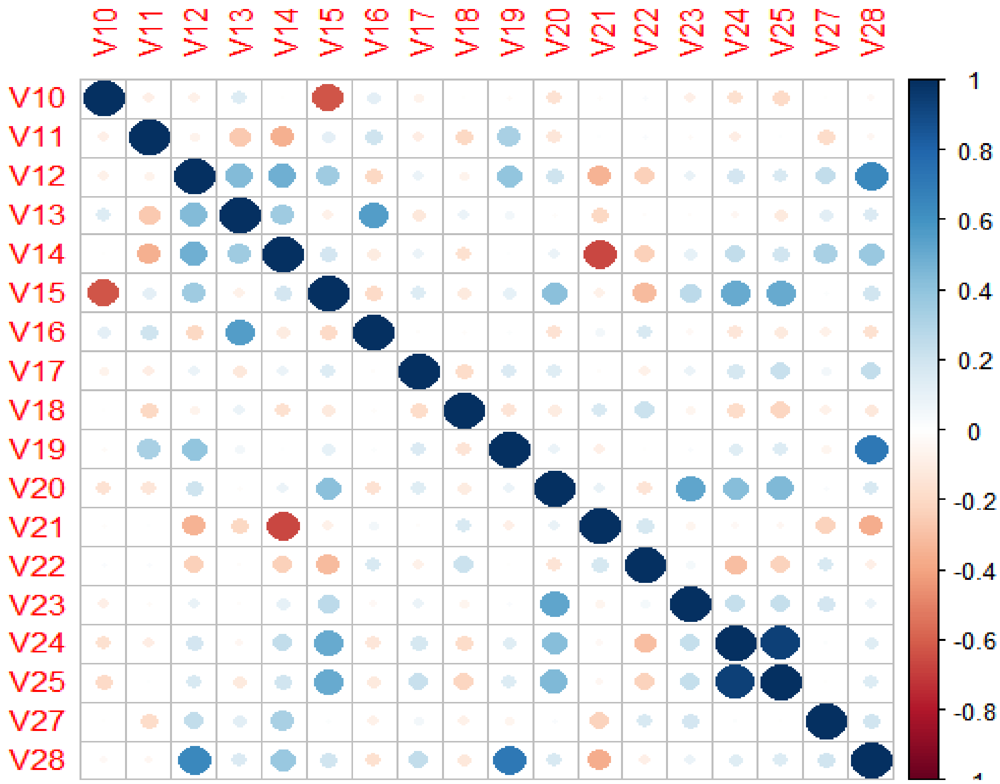
2. Bootstrap confidence intervals

After attaining bootstrap confidence intervals on the wind speeds, a paired sample t test was run on the data and the p-value obtained was 0.1382 which meant that there was no statistical significance between the means of the 0-12 wind speed for the two samples. Other statistics obtained include the 95% CI for the mean difference between the sample observations and the mean of the differences.

```
95 percent confidence interval:
 -9.967638  1.506382
sample estimates:
mean of the differences
              -4.230628
```

3. Correlation

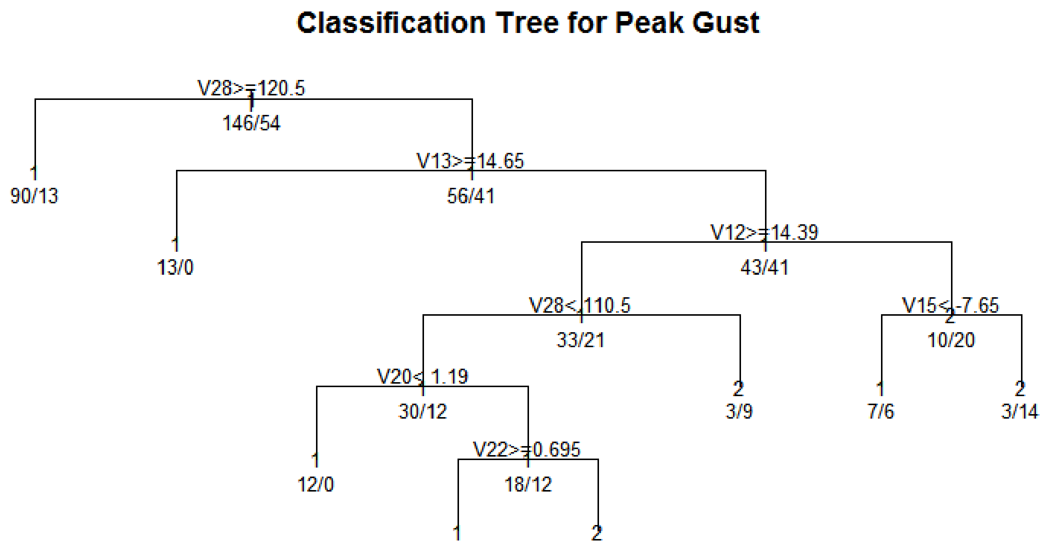
Correlation for the different predictor variables was obtained and the results were as shown in the correlation plot below.



From the above we were able to sort out the variables which had a correlation of (22/324)*100 which means that 6.8% of the predictors have a correlation that is greater than 0.7. The low percentage means that the variables have a low association with each other. This is also means that there is little linear relationship between the various variables of interest.

4. Classification problem

To solve the classification we first formed a factor variable which we then classified using a decision tree. The tree obtained is as shown below.



From the above tree we can see that the sangster parameter is the most important predictor in terms of predicting the 0-12 wind speed for the prefrontal dataset. 10% of the original classified data was then used for testing and contingency tables of frequencies developed to check on different probabilities before and after the prediction. The results are as shown below

Contingency table for fitted values

```
1   2
15  8
```

```
Contingency table for predicted values
```

```
1   2
20  3
```

From the above contingency tables we can see that the predicted values had more major storms compar ed to the fitted values. Probability for a prediction of major storm based on the model is 20/23 which is greater than actual value of 15/23 as the results indicate.

5. Regression model performance

Below are the summaries obtained from the regression models performed for the different datasets.

Prefrontal dataset

```
Call:
lm(formula = train$`prefrontal$V8` ~ V10 + V12 + V13 + V14 +
    V15 + V18 + V19 + V24 + V27, data = train)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-11.5989  -3.3469  -0.7769   2.7202  15.7062 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.18616    3.33268   6.357 1.49e-09 ***
V10         -0.41241    0.10831  -3.808 0.000189 ***
V12          0.31060    0.10724   2.896 0.004218 ** 
V13          0.10434    0.07361   1.418 0.157970    
V14          0.16604    0.06125   2.711 0.007327 ** 
V15         -0.20730    0.07283  -2.846 0.004908 ** 
V18         -0.03062    0.02150  -1.424 0.156108    
V19          0.02297    0.01329   1.728 0.085581 .  
V24         -0.52939    0.26131  -2.026 0.044178 *  
V27         -0.01717    0.01231  -1.395 0.164595    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.701 on 190 degrees of freedom
Multiple R-squared:  0.3186,	Adjusted R-squared:  0.2863 
F-statistic: 9.871 on 9 and 190 DF,  p-value: 2.295e-12
```

Postfrontal dataset

```
Call:
lm(formula = train$`postfrontal$V8` ~ V12 + V19, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9276 -2.0665 -0.5107  1.6376 10.0655 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.90869    2.35363   3.360  0.00150 ** 
V12          0.83923    0.16052   5.228 3.37e-06 ***
V19          0.07222    0.02235   3.231  0.00219 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.897 on 50 degrees of freedom
Multiple R-squared:  0.4926,	Adjusted R-squared:  0.4723 
F-statistic: 24.27 on 2 and 50 DF,  p-value: 4.296e-08
```

Root mean square errors for the two models were also obtained and the results were as follows;
Prefrontal - 4.581579
Postfrontal - 3.785455

From the above results we can see that both models are significant since their p-values are less than 0.05. The regression model for the postfrontal dataset also proved to be more reliable than the regression model for the prefrontal dataset since the adjusted R-squared for the latter was smaller compared to that of the former i.e. for the prefrontal dataset the model covered for 28.63% of the variation while for the postfrontal dataset the model covers for 47.23 percent of the variation. Looking at the values of RMSE, the prefrontal model has a greater error compared to the postfrontal model. This means that the postfr ontal model is more accurate compared to the prefrontal model.

Regression analysis for all the variables in both of the datasets was also done and the results obtained ar e as indicated below.

```
Call:
lm(formula = train$`newD$V8` ~ V10 + V12 + V13 + V14 + V15 +
    V18 + V20 + V23 + V24 + V27 + V28, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0633  -3.1217  -0.9161   2.7321  16.2447

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.660353   3.159416   6.223 2.15e-09 ***
V10         -0.119399   0.064511  -1.851  0.06542 .
V12          0.311666   0.105374   2.958  0.00341 **
V13          0.175973   0.070216   2.506  0.01287 *
V14          0.113672   0.058901   1.930  0.05480 .
V15         -0.060220   0.027588  -2.183  0.03002 *
V18         -0.035476   0.017954  -1.976  0.04930 *
V20         -0.465383   0.294073  -1.583  0.11484
V23          0.939085   0.491721   1.910  0.05735 .
V24         -0.443138   0.249874  -1.773  0.07742 .
V27         -0.032679   0.011939  -2.737  0.00666 **
V28          0.022196   0.007509   2.956  0.00343 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.809 on 241 degrees of freedom
Multiple R-squared:  0.3078,   Adjusted R-squared:  0.2762
F-statistic: 9.742 on 11 and 241 DF,  p-value: 1.486e-14
```

The model formed from the linear regression of the joined datasets is statistically significant since it has a p-value which is less than 0.05. The model accounts for 27.62% of the variation within the prediction. It has an RMSE value of 4.693639 which is higher than the other regression models developed showing t hat it is least accurate model used for the prediction of wind speed.