Report for Assessment 2

SIT 718 – Real World Analytics

Student Name: Prateek Singh

Student ID: 221218743

T1:

Histogram for all variables (X1 – X8 and Y), Scatter plot (X2 – X8, Y) vs X1

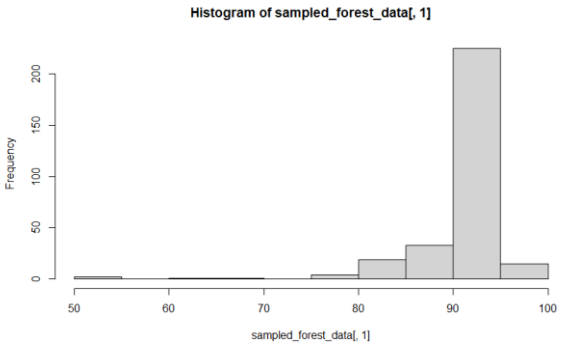
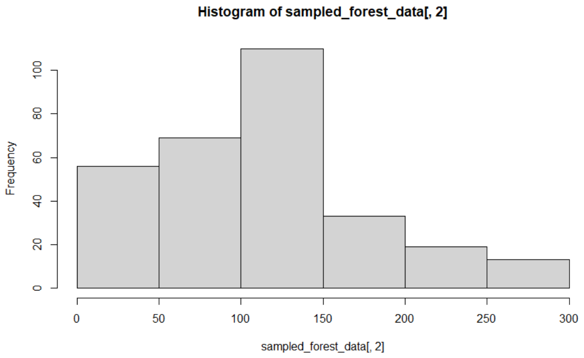


Fig: X1 – FFMC Index

The variable of interest is skewed-left. Therefore, the transformation used is normalization and not standardization.



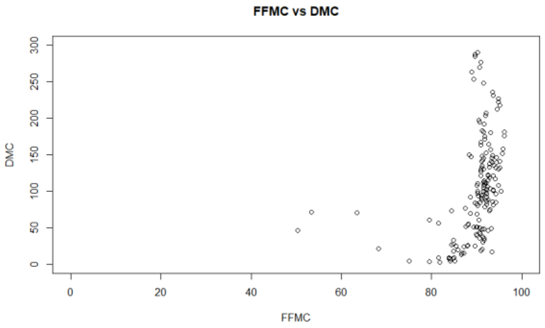
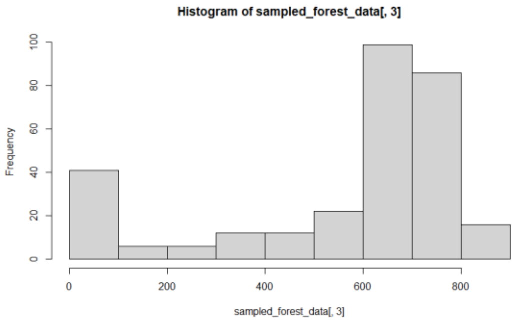


Fig: X2 - DMC index Histogram

X1 vs X2 scatter plot

X2 follows a near normal distribution.



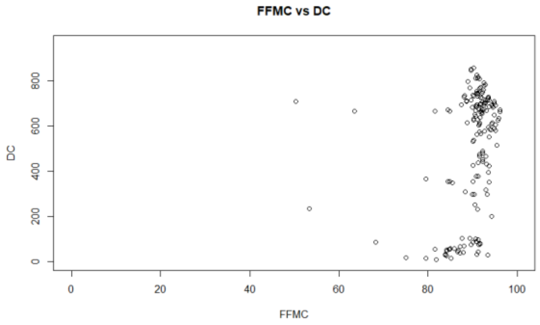
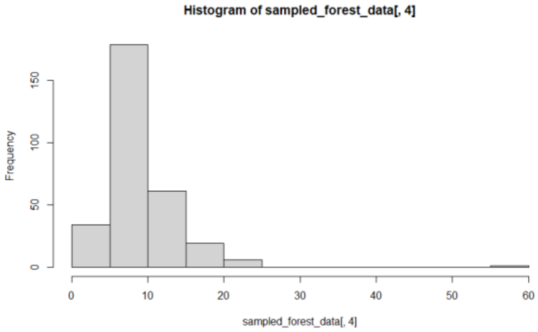


Fig: X3 Histogram

X1 vs X3 scatter plot

X3 does not follow normal distribution and is skewed.



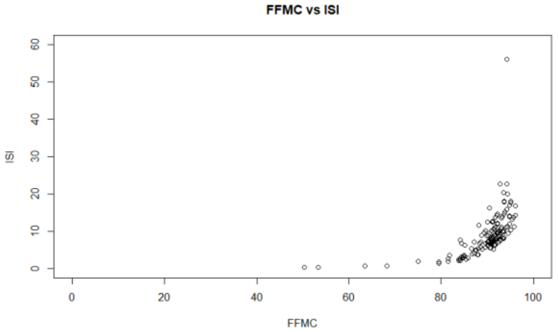
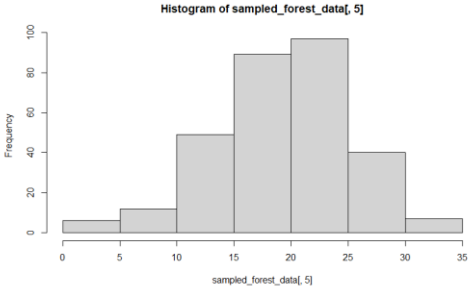


Fig: X4 – ISI Index Histogram

X1 vs X4 scatter plot

X4 does not follow normal distribution and is skewed right.



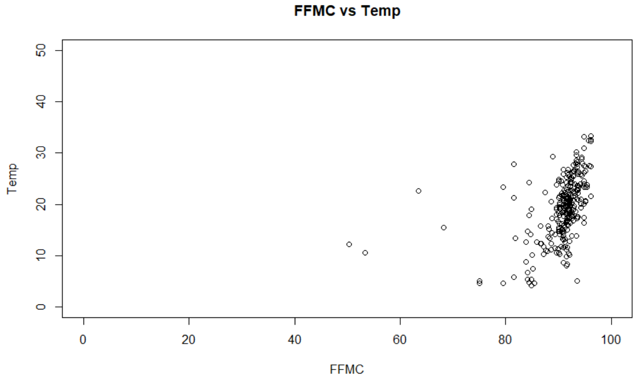
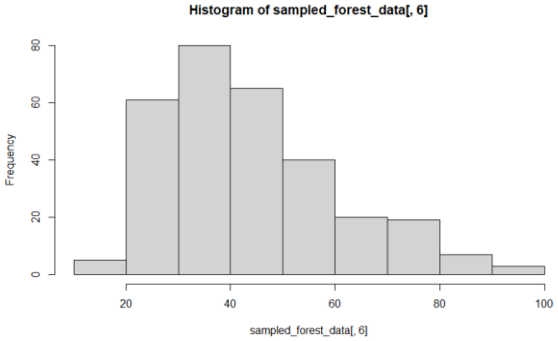


Fig: X5 - temp Histogram

X1 vs X5 scatter plot

X5 follows near normal distribution.



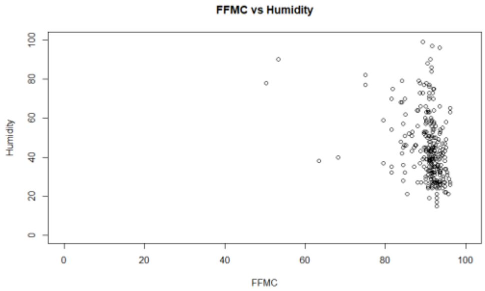
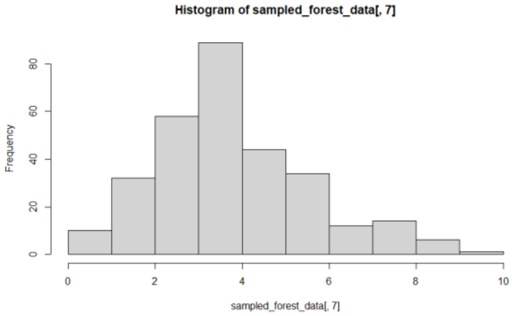


Fig: X6 - RH Histogram

X1 vs X6 scatter plot

X6 follows near normal distribution.



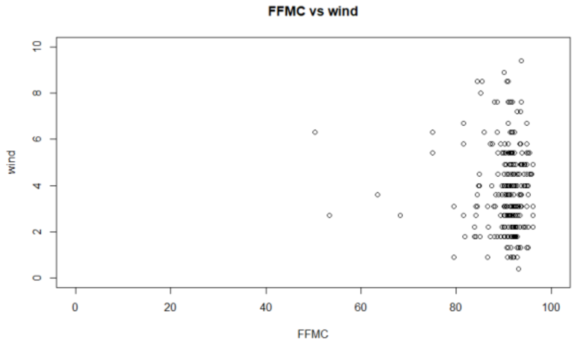
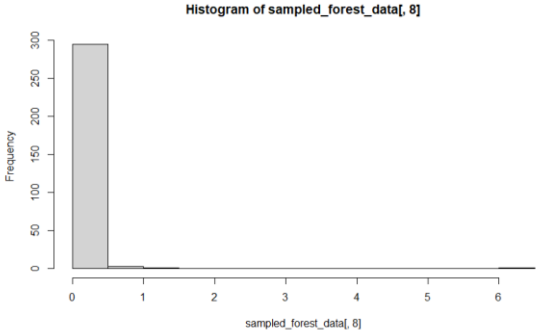


Fig: X7 -wind Histogram

X1 vs X7 scatter plot

X7 follows near normal distribution.



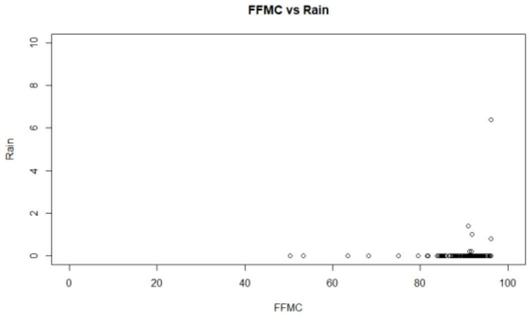
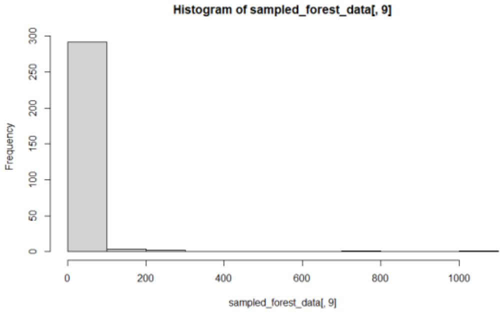


Fig: X8 - rain Histogram

X1 vs X8 scatter plot

X8 does not follow normal distribution and is skewed.



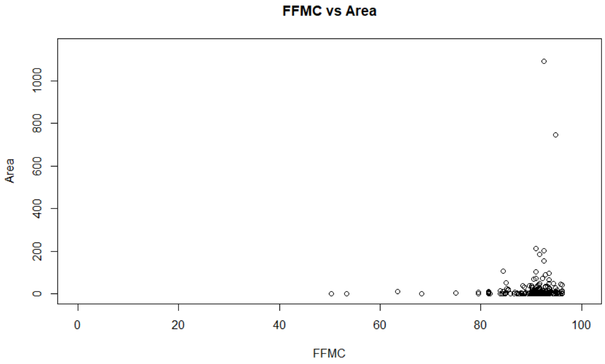


Fig: Y -area Histogram

X1 vs Y scatter plot

Y does not follow normal distribution and is skewed.

T2:

The chosen variables all follow nearly normal distribution and hence they are standardized to make them mean centred.

The variable of interest X1 is skewed and hence it is normalized to squash the range between in [0,1].

T3:

Error Table

	WAM	Power Mean (p = 0.5)	Power Mean (p = 2)	Choquet Integral
RMSE	1.04	NA	0.45	0.619
Avg Absolute Error	0.92	NA	0.33	0.5
Pearson Correlation	0.38	NA	-0.36	-0.02
Spearman Correlation	0.53	NA	-0.21	0.2

Table: Summary of all error measures of models

Summary Table

Iterations /Models	WAM	Power Mean (p = 0.5)	Power Mean (p = 2)	Choquet Integral
1	0.237202467102734	0	0.160354192860803	0.0826774903604669
2	0.435622965824268	0.654239686414314	0.177326181307446	0.457874514885869
3	0.195513190577415	0.078966942224904	0.297400346553952	0.227194168800777
4	0.13166137649558	0.266793371360782	0.364919279277798	0.232253825952854

Table: Summary of all weights learned by models

a.) Here we can observe that the power mean (p=2) outperforms every other model in the table. Whereas the Choquet integral ranks 2nd after power mean.

b.) Here we have chosen the variables which are nearly normally distributed, this is of utmost importance if we want the model to be robust and generalised on real world data.

c.) The variables are not related to each other.

d.) Better models favour inputs that are mean centred and tend to ignore inputs that can be termed as outliers. In vase of number of variables, better models tend to learn from lower number of variables as they can establish/map relationship between dependent and independent variables by lesser number of features.

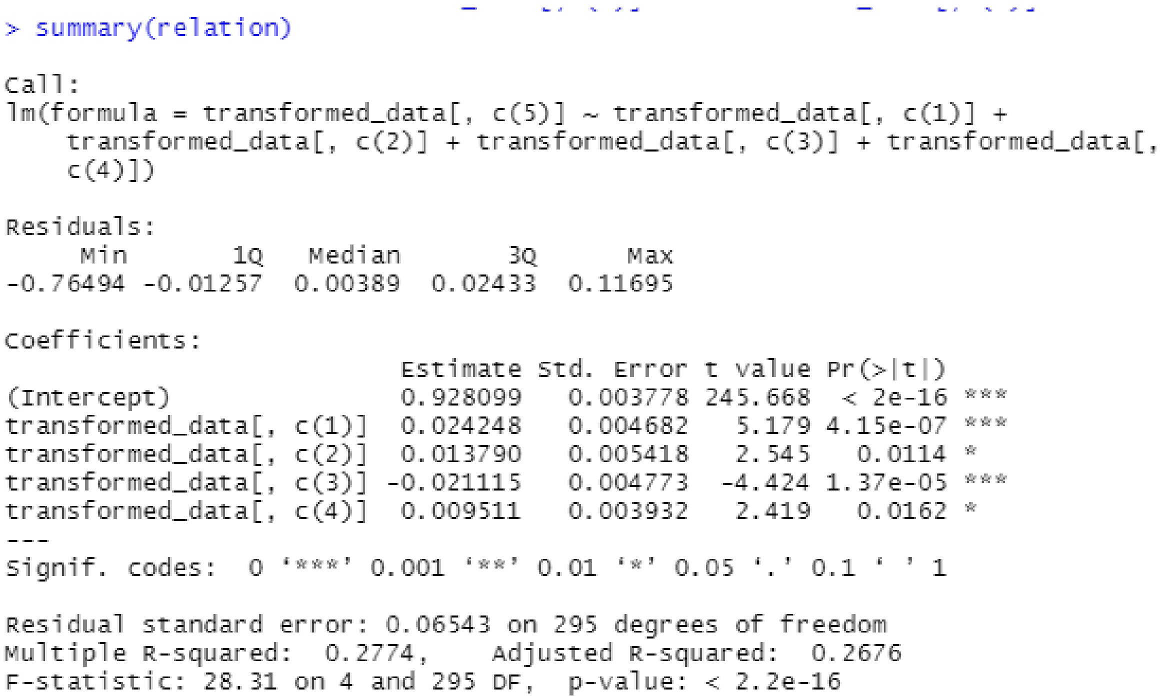
T4:

For the given input, the predicted value with power mean (p=2) is 89.09629. This prediction denotes that power mean (p=2) can be used effectively to predict FFMC index as the error margin is less.

Ideal conditions in which a large FFMC index will be obtained is when all the constituent variables have high values in their respective ranges.

T5:

Following information contains statistics related to the trained linear regression model.



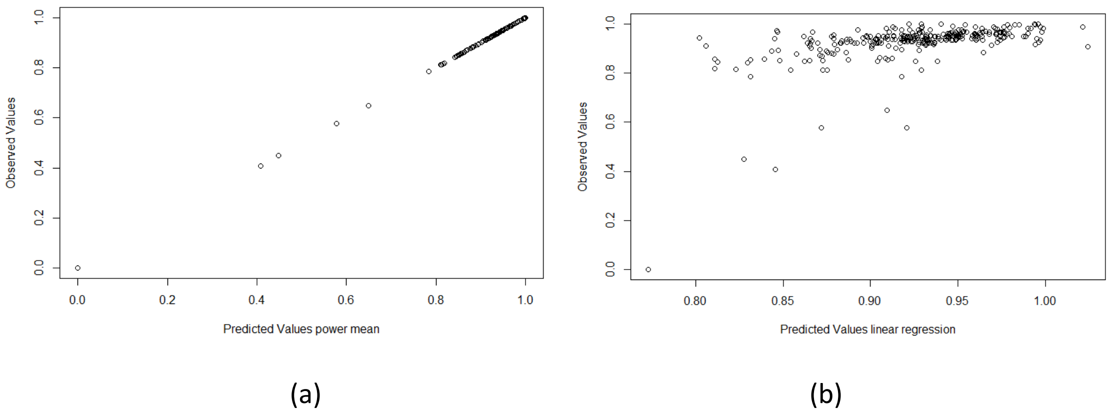
```
> summary(relation)

Call:
lm(formula = transformed_data[, c(5)] ~ transformed_data[, c(1)] + 
    transformed_data[, c(2)] + transformed_data[, c(3)] + transformed_data[, 
    c(4)])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.76494 -0.01257  0.00389  0.02433  0.11695 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.928099   0.003778 245.668  < 2e-16 ***
transformed_data[, c(1)]  0.024248   0.004682   5.179 4.15e-07 ***
transformed_data[, c(2)]  0.013790   0.005418   2.545   0.0114 *  
transformed_data[, c(3)] -0.021115   0.004773  -4.424 1.37e-05 ***
transformed_data[, c(4)]  0.009511   0.003932   2.419   0.0162 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06543 on 295 degrees of freedom
Multiple R-squared:  0.2774,	Adjusted R-squared:  0.2676 
F-statistic: 28.31 on 4 and 295 DF,  p-value: < 2.2e-16
```



(a)

(b)

Fig: (a): Predicted values from power mean, (b) predicted values from linear regression.

We can see that the aggregation function (power mean [p=2]) is fit very well as compared to the trained linear regression model. This shows how aggregation function can be useful in finding patterns on datasets. On the other hand, linear regression tries to fit the best fit line on the transformed data and builds a generalised model which can be useful for unseen data. This is a good example of the difference between aggregation and approximation.