Prediction of the SGEMM kernel runtime

Kashish Kothari

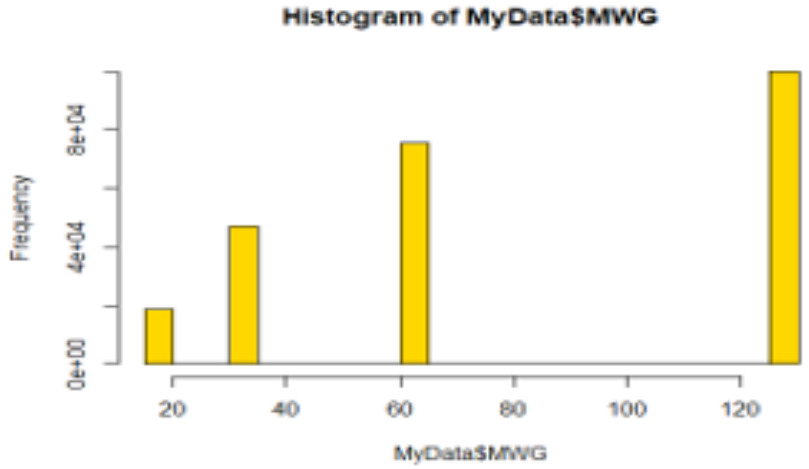
Data

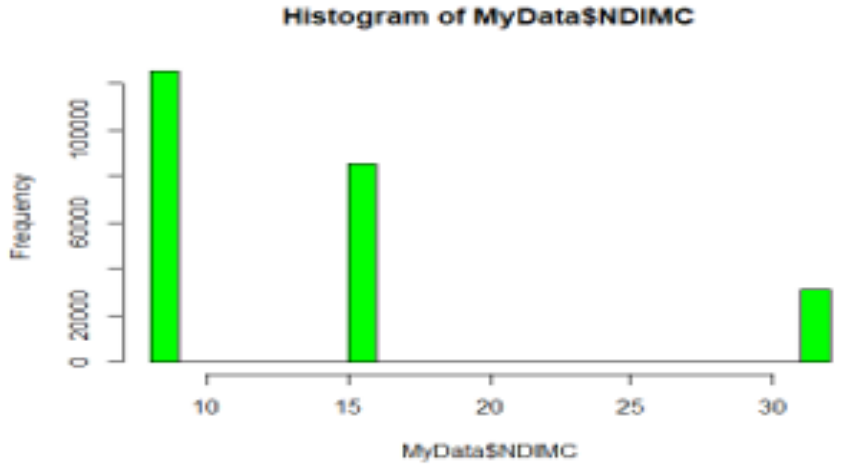
- Data consist of 250000 instances of the runtime.
- It has 14 independent variable describing different parameters of SGEMM kernel.
- 4 independent variables are Numerical.
- 10 independent variables are categorical.
- Dependent variable is Run_time in Millisecond which is numerical.

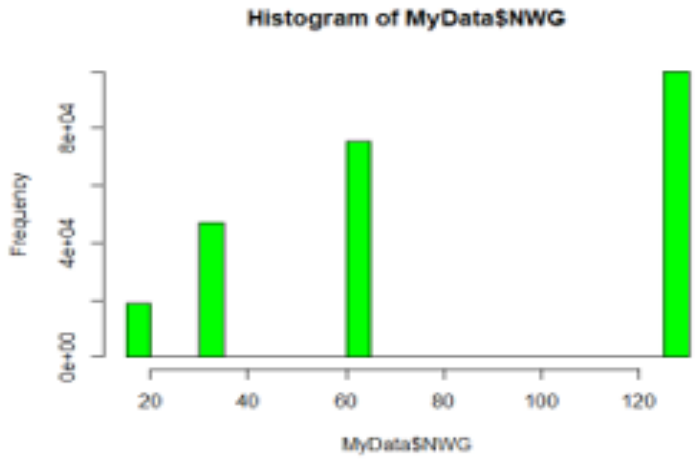
Question

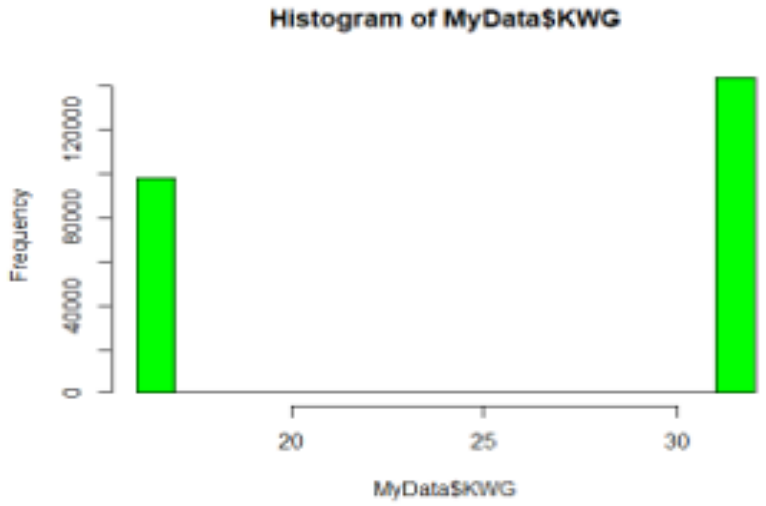
- My aim is to predict the Run time based on the given Variable values for the SGEMM Kernel
- So my question was to find out whether there is Linear Relationship between explanatory and dependent Variables.
- I wanted to select regression algorithm based on linear and non linear relationships in the data.
- Whether I should select parametric or non parametric approach.

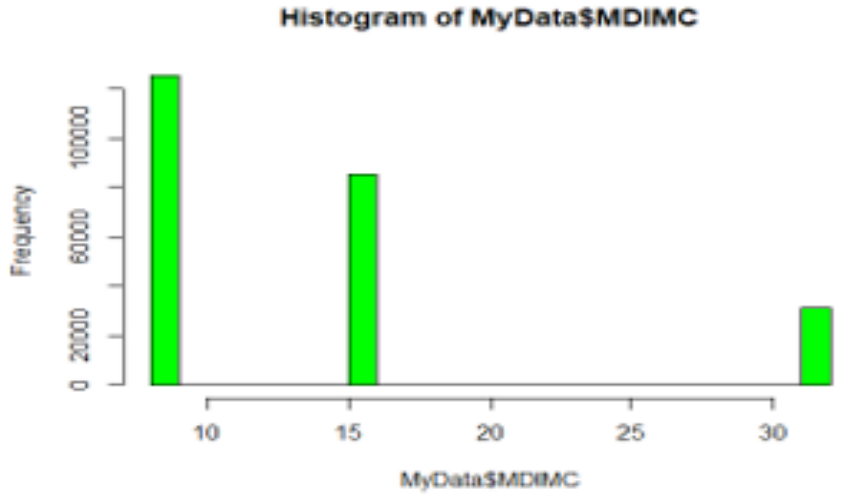
Exploratory Analysis

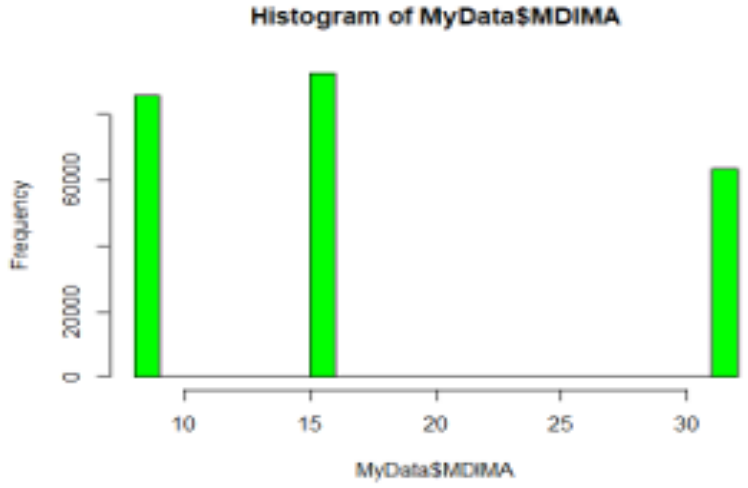


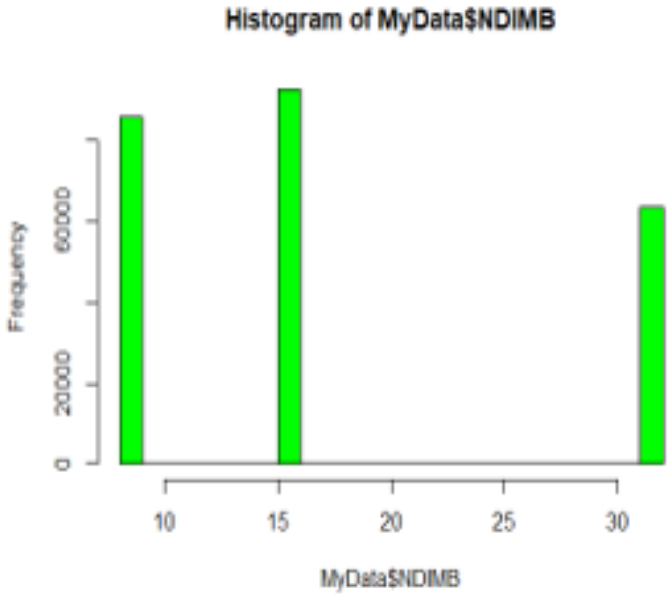


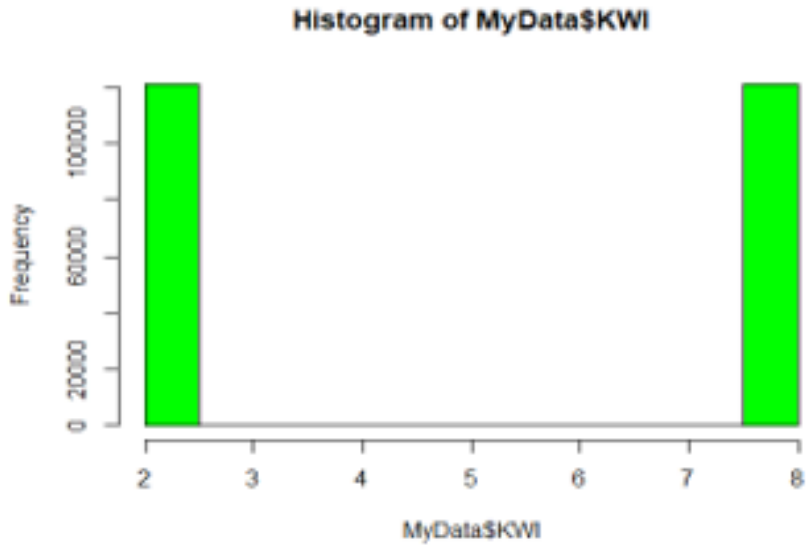


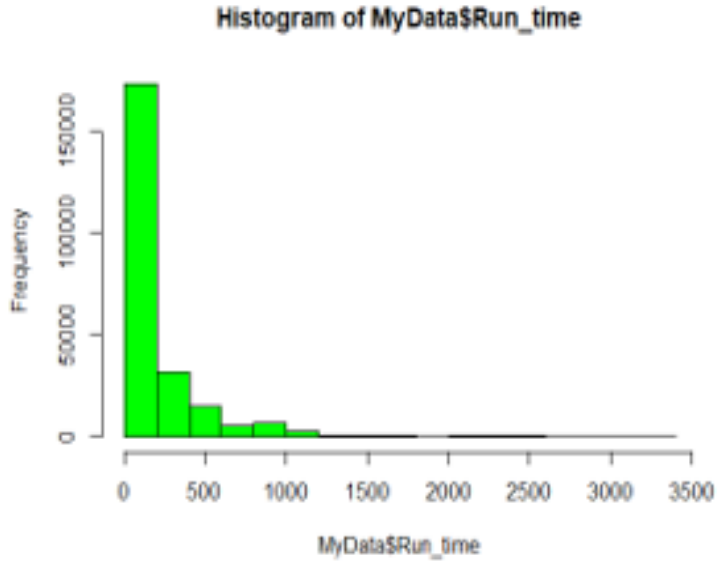


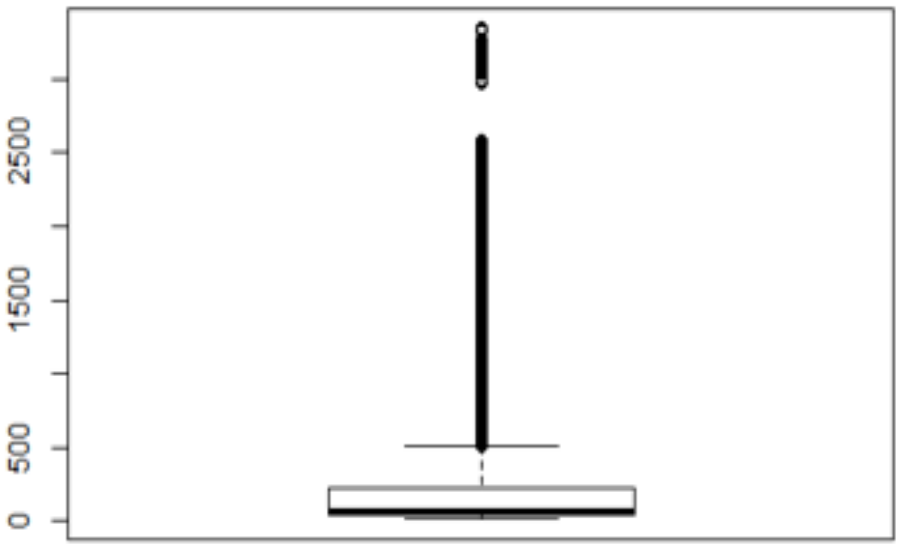





Histogram of MyData$NDIMB
Frequency
0
20000
60000
10
15
20
25
30
MyData$NDIMB


Histogram of MyData$KWI
Frequency
0
20000
60000
100000
2
3
4
5
6
7
8
MyData$KWI

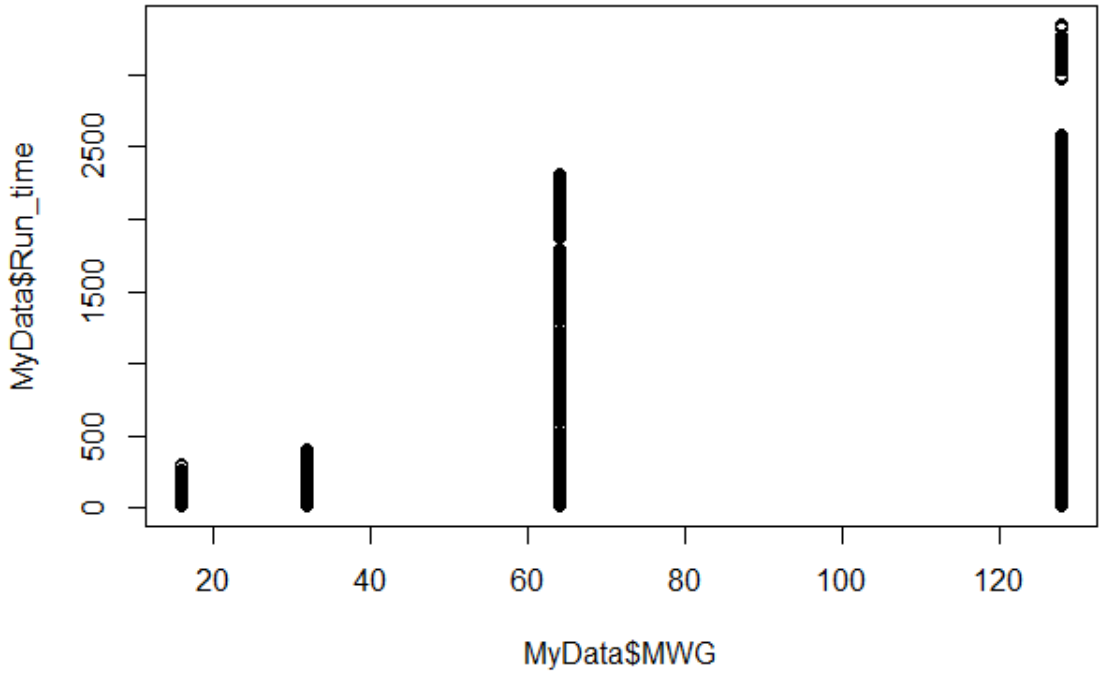

Histogram of MyData$Run_time
Frequency
0
50000
100000
150000
0
500
1000
1500
2000
2500
3000
3500
MyData$Run_time

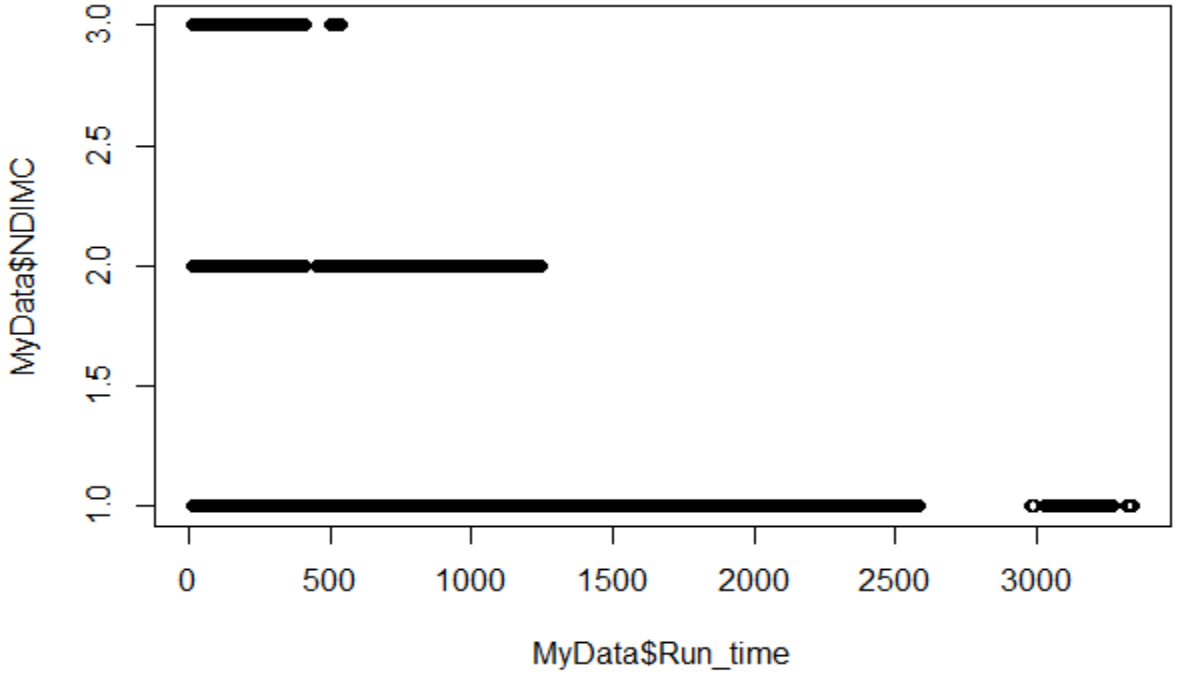

0
500
1500
2500

Potential Problems to be checked before applying Linear Regression

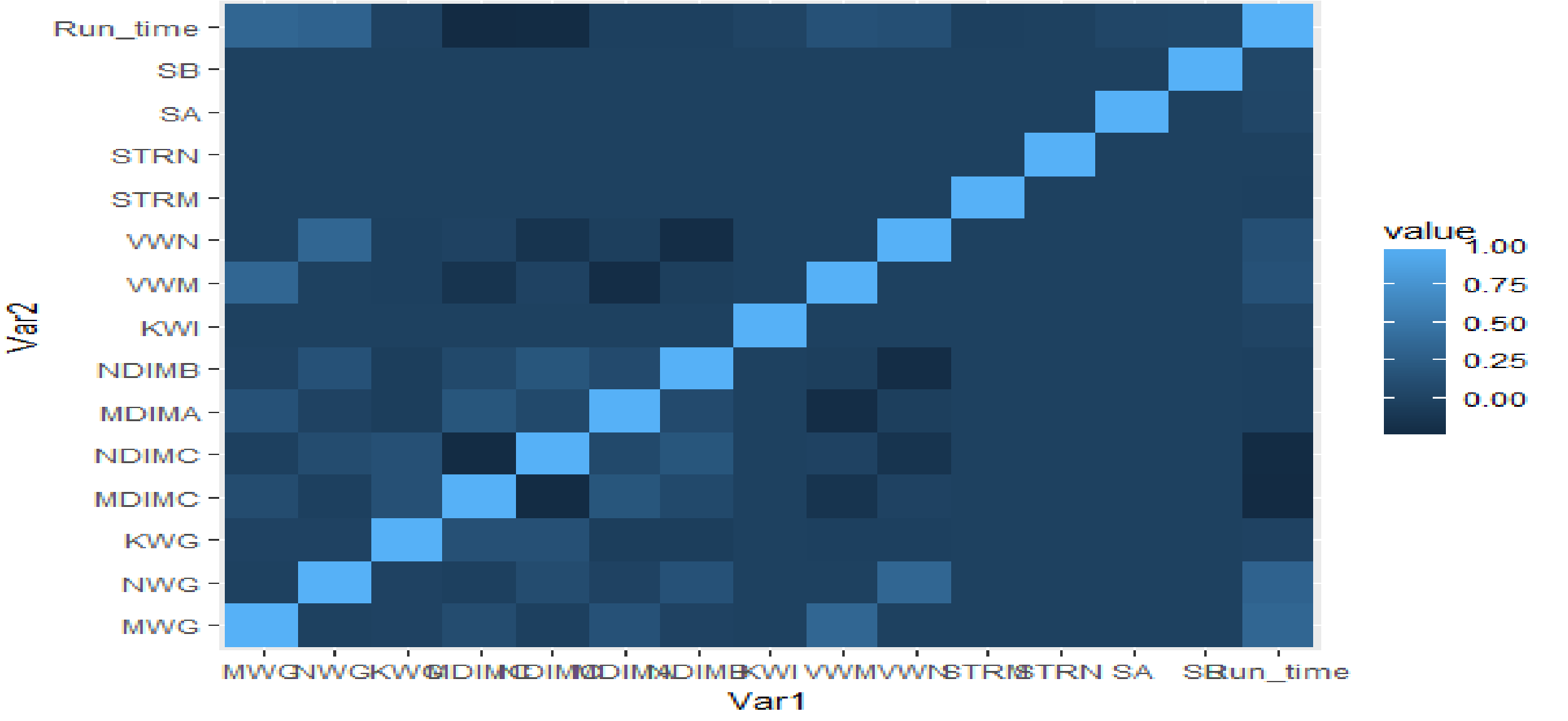
- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms.
- Outliers.
- High-leverage points.
- Collinearity.

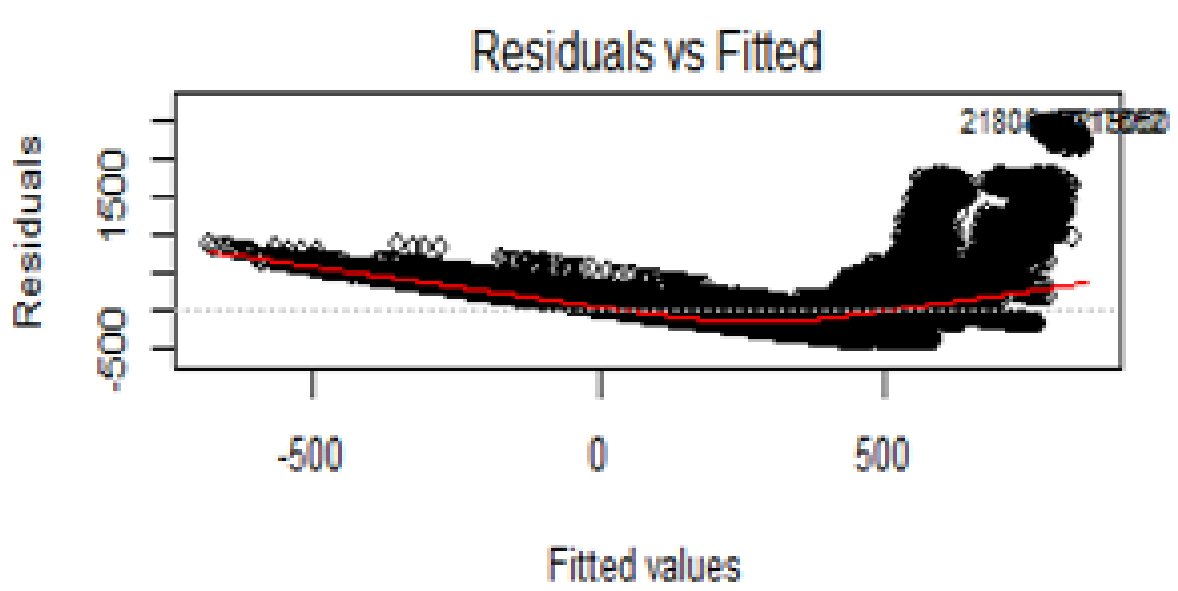
Some Linear Association

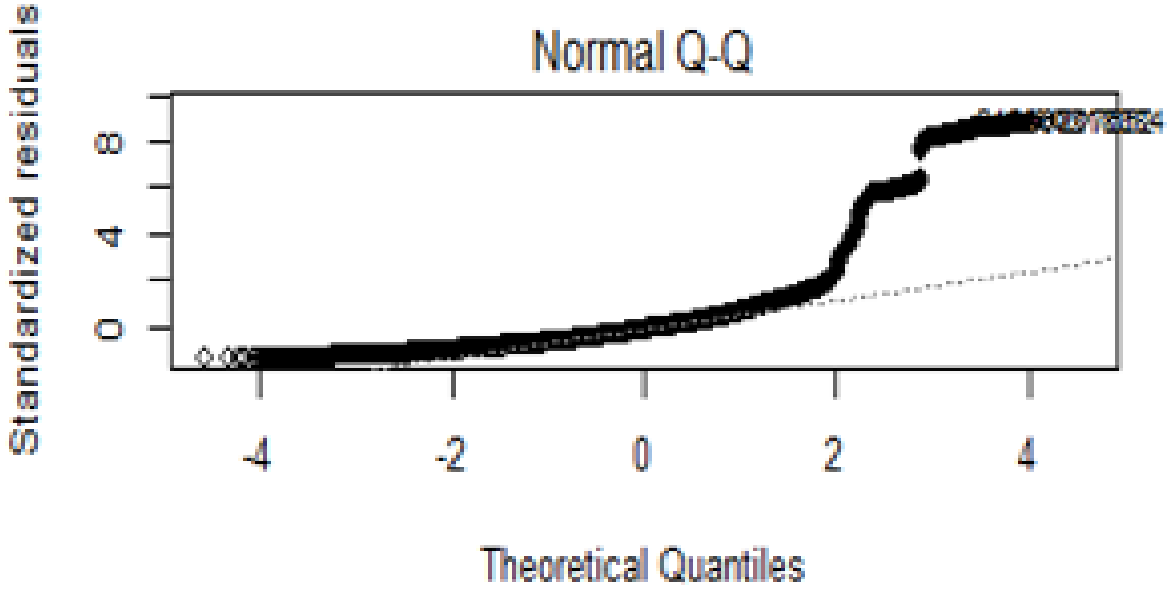


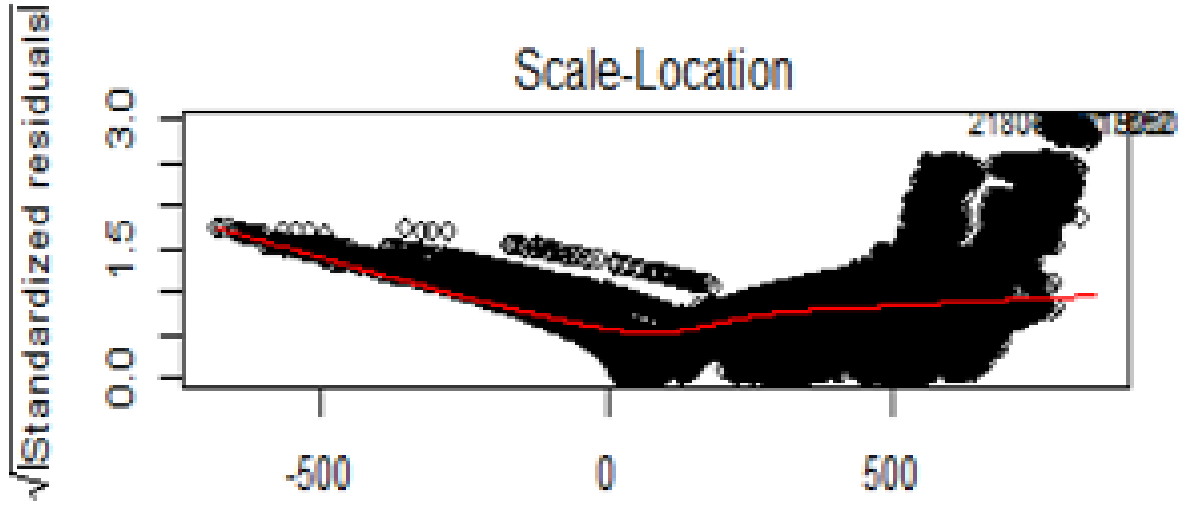


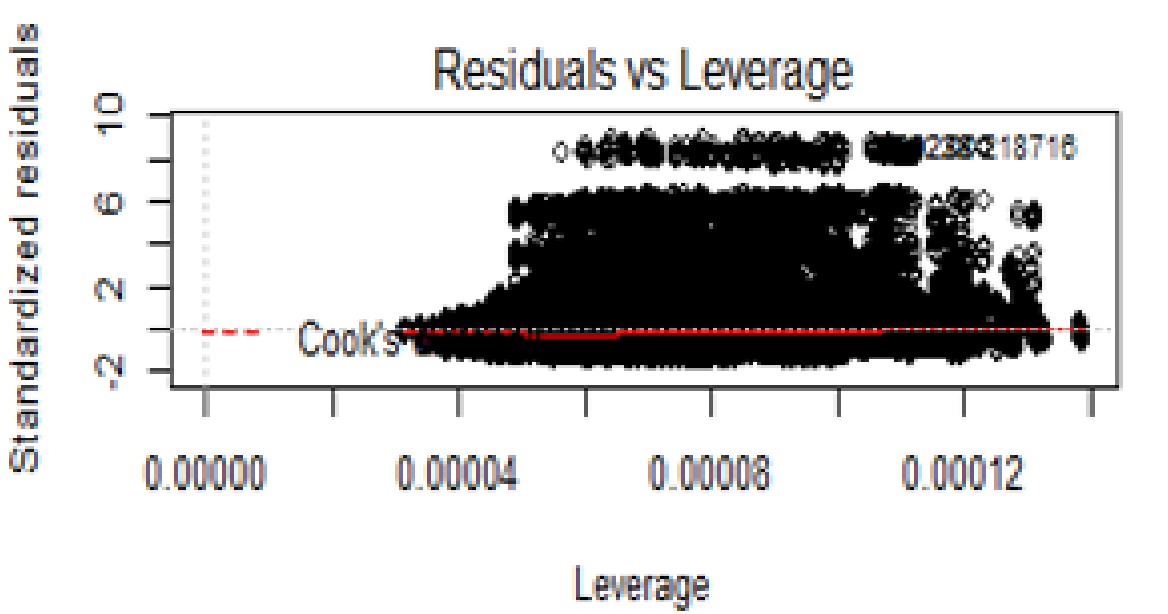
Checking for Collinearity




Residuals vs Fitted
Residuals
1500
-500
-500
0
500
Fitted values


Normal Q-Q
Standardized residuals
8
4
0
-4
-2
0
2
4
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
3.0
1.5
0.0
-500
0
500
Fitted values


Residuals vs Leverage
Standardized residuals
10
6
2
-2
Cook's
0.00000
0.00004
0.00008
0.00012
Leverage

What went Wrong?

- Linearity Assumptions were violated.
- Normal Distribution Conditions were Violated
- As there were nonlinear distribution in data, I opted for Non linear regression method such as GAM.
- I selected GAM because even though in non linear space it is easy for interpretation which is important.
- I applied smoothing spline to 4 numerical variables.

Why GAM Did not work

- AS majority of Independent variables are categorical and rest of the variables are also integers with only 4 unique values.
- Therefore predictors are not able to cover the variance of the predicted variable.
- As the distribution of the data is not known and majority of variables are categorical .

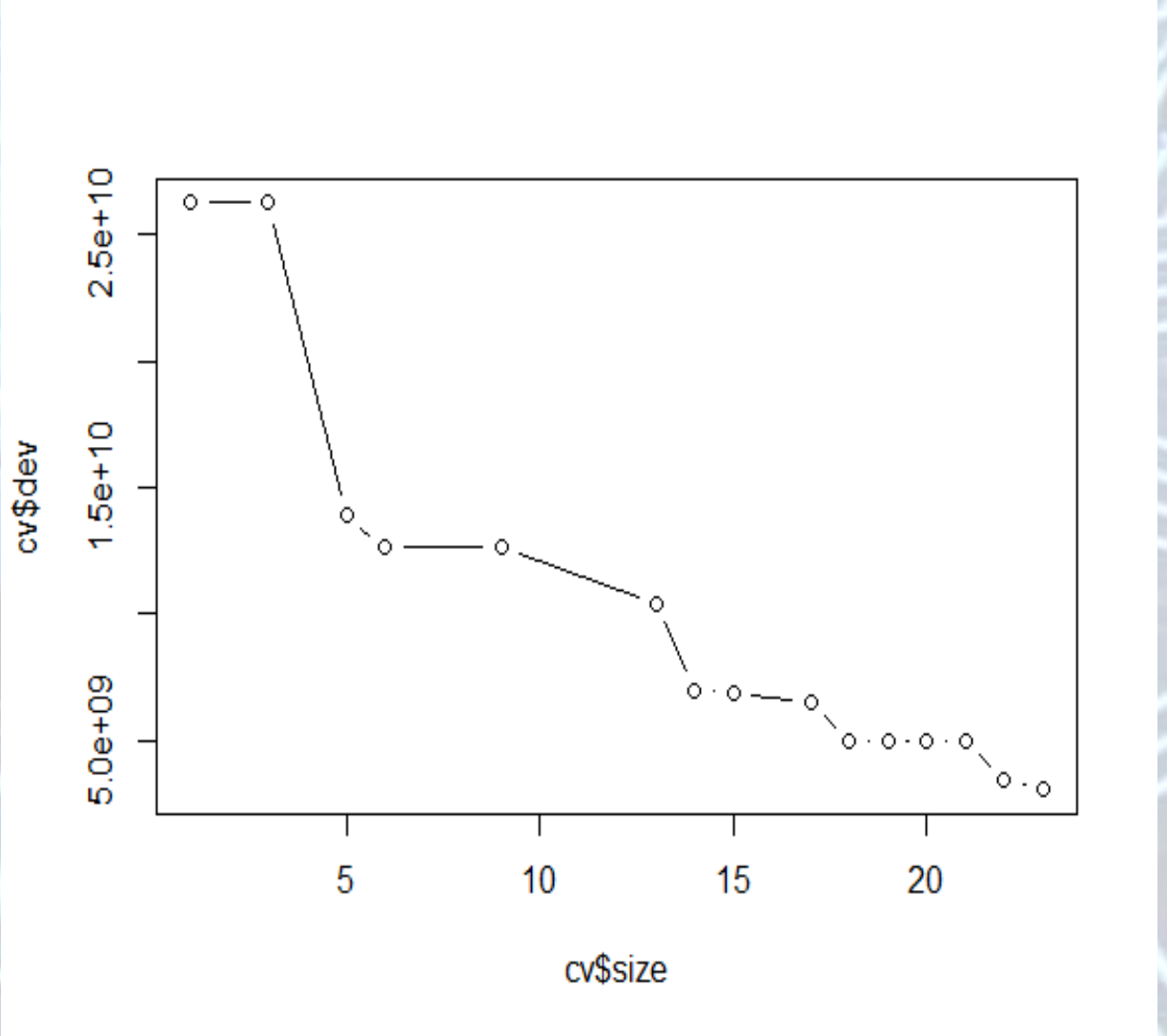
Non Parametric Approach

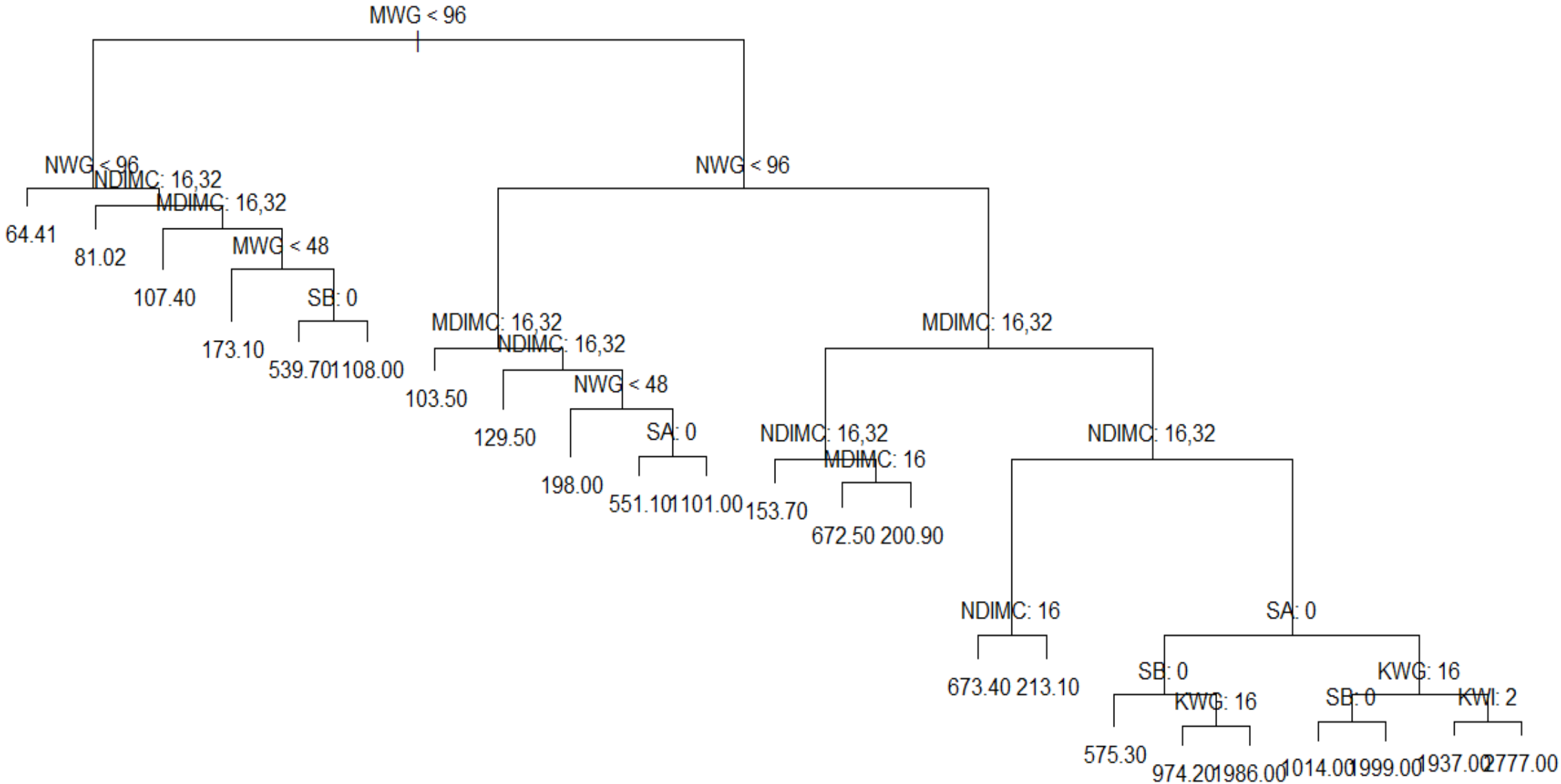
- I have selected Regression tree because I wanted to predict non linearity in the data.
- Regression tree is non linear model which can segregate predictor space into simple regions
- They are easy to interpret
- They generally work better when majority of independent variables are categorical
- There is less effort in data preparation and feature selection.
- They are robust to outliers and leverage points.

Regression Tree

- I have used cross validation for selecting the effective size of the tree.
- Tree with 23 terminal nodes has lowest error .
- Therefore this tree was used for the prediction
-

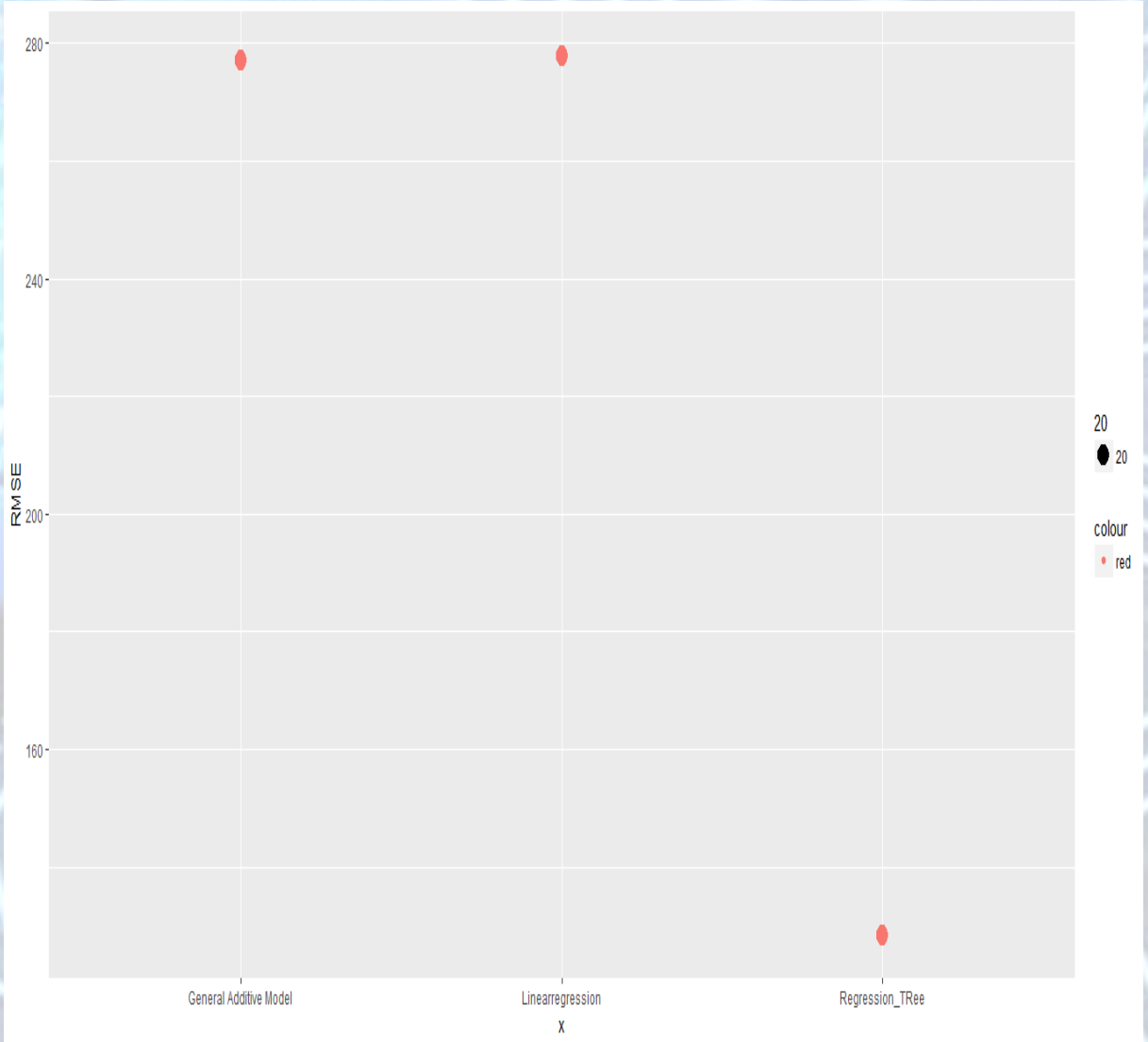
```
Regression tree:
tree(formula = Run_time ~ ., data = training)
Variables actually used in tree construction:
[1] "MWG"   "NWG"   "NDIMC" "MDIMC" "SB"    "SA"    "KWG"   "KWI"
Number of terminal nodes:  23
Residual mean deviance:  16270 = 3.145e+09 / 193300
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1031.00   -50.19   -22.36     0.00    28.10  1272.00
```




MWG < 96
NWG < 96
NDIMC: 16,32
MDIMC: 16,32
MWG < 48
SB: 0
64.41
81.02
107.40
173.10
539.70
1108.00
NWG < 96
MDIMC: 16,32
NDIMC: 16,32
NWG < 48
SA: 0
103.50
129.50
198.00
551.10
1101.00
MDIMC: 16,32
NDIMC: 16,32
MDIMC: 16
153.70
672.50
200.90
NDIMC: 16,32
NDIMC: 16
SA: 0
673.40
213.10
SB: 0
KWG: 16
KWG: 16
SB: 0
KWI: 2
575.30
974.20
1986.00
1014.00
1999.00
1937.00
2777.00

Result of prediction on testing dataset

- Regression Tree performed better in this data with lowest RMSE compare to other two algorithm.
- There was no difference in the RMSE when I applied Linear Model and GAM model on test data.




Thank You!