Wild Blueberry Yield Prediction Using Multiple Linear and Machine Learning Regression Models.

[Wild Blueberries – Perennia]

- *Shobhakhar Adhikari*
- *Vipandeep Rataul*

Course: Statistical Methods for Data Analytics

University of North Carolina at Greensboro

Introduction: Crop Yield prediction is of great importance to global food production. Policy makers rely on accurate predictions to make timely import and export decisions to strengthen national food security. The main goal of this study is to find out how bee species composition and weather affect blueberry yield and to predict optimal bee species composition and weather conditions that achieve the best yield using computer simulation data and machine learning algorithms. Multiple linear regression (MLR), Decision trees Regressor (DTR), Random forest (RF), and Gradient boosting (GB) were evaluated as predictive tools. The techniques and models we will use on predicting Wild

Blueberry Yield can also be used on other crops Yield prediction. So, this is the main motivation for working on this project.

Data Used: Dataset Link: [Mendeley Data - Data for: Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms]

Description of Dataset: This dataset was generated by the Wild Blueberry Pollination Simulation Model, which is an open-source, spatially-explicit computer simulation program. The simulation model has been validated by the field observation and experimental data collected in Maine USA and Canadian Maritimes during the last 30 years and now is a useful tool for hypothesis testing and theory development for wild blueberry pollination research.

This dataset has/77 observations and 13 independent variables.

Response variable (Blueberry Yield (kg/ha)) is Continuous numerical variable.

Features and their description

Features	Unit	Description
Clonesize	m2	The average blueberry clone size in the field
Honeybee	bees/m2/min	Honeybee density in the field
Bumbles	bees/m2/min	Bumblebee density in the field
Andrena	bees/m2/min	Andrena bee density in the field
Osmia	bees/m2/min	Osmia bee density in the field
MaxOfUpperTRange	°F	The highest record of the upper band daily air temperature during the bloom season

MinOfUpperTRange	°F	The lowest record of the upper band daily air temperature
AverageOfUpperTRange	°F	The average of the upper band daily air temperature
MaxOfLowerTRange	°F	The highest record of the lower band daily air temperature
MinOfLowerTRange	°F	The lowest record of the lower band daily air temperature
AverageOfLowerTRange	°F	The average of the lower band daily air temperature
RainingDays	Day	The total number of days during the bloom season, each of which has precipitation larger than zero
AverageRainingDays	inch	The average of daily precipitation during the bloom season

Data Type: All features are continuous numerical except “RainningDays” feature is integer.

Data Pre-Processing:

1) Data Cleaning:

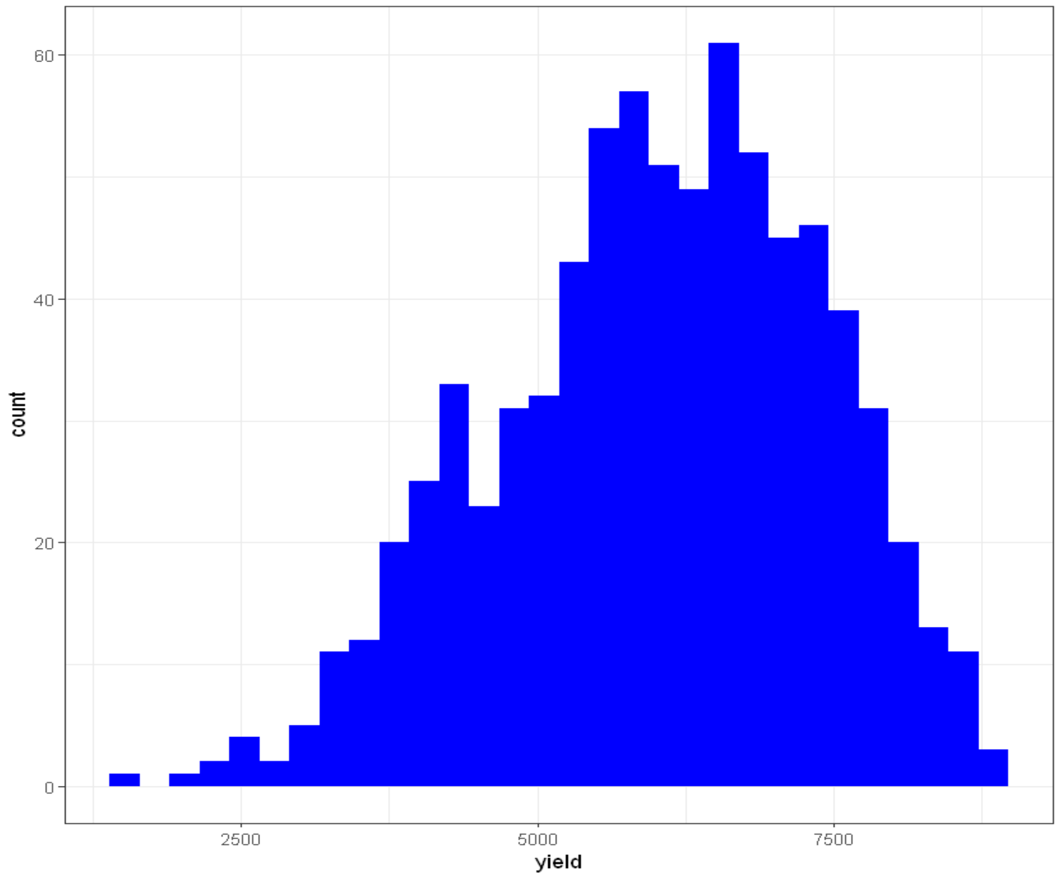
- Removed the row column from the data as it was an index for the original (raw) data, which was not required for data analysis and predictive modeling.
- Removed the columns: fruitset, fruitmass, and seeds. These variables are removed because they are not in the interest of study.
- Checked the null values in any rows and columns. There were no missing values in this dataset.
- Checked the duplicated rows and found none.

2) Data Exploration and Visualizations:

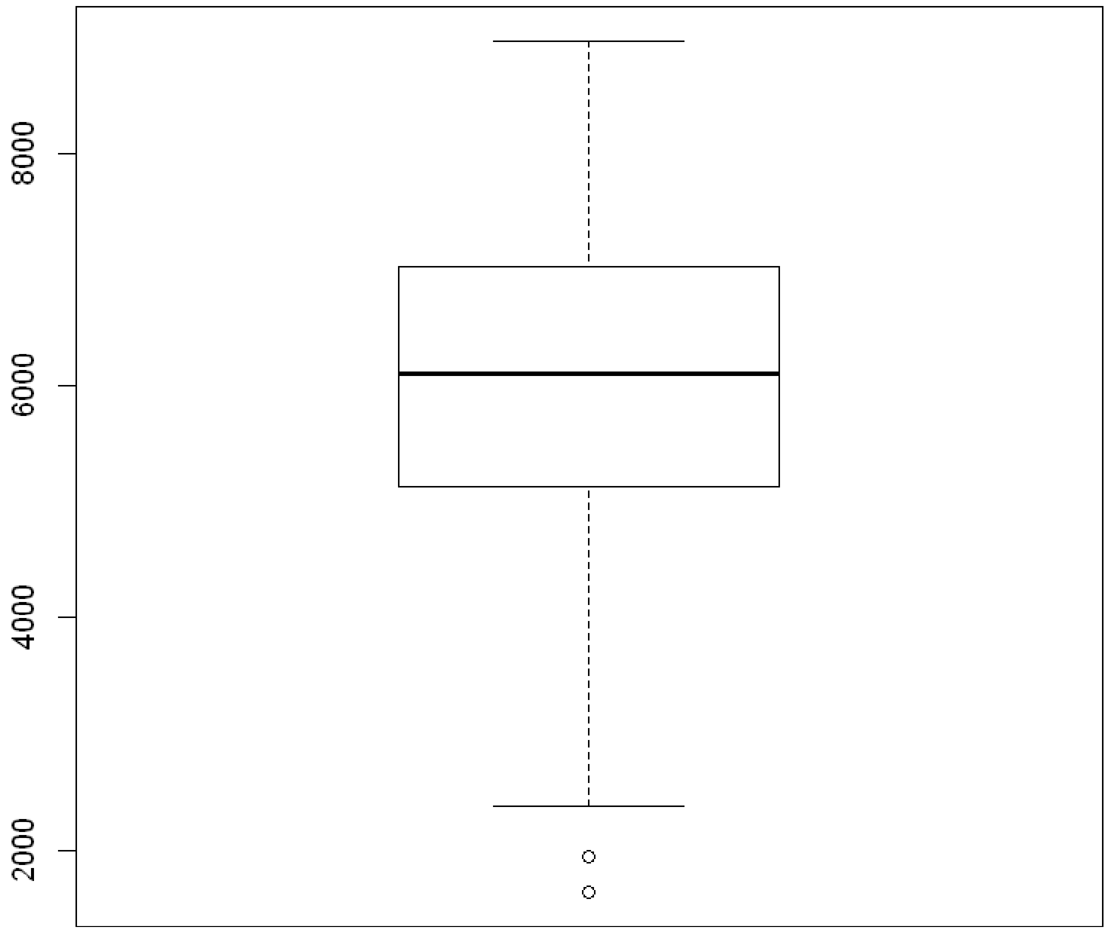
★ All the variables in our dataset are numeric.

```
'data.frame':   777 obs. of  14 variables:
 $ clonesize           : num  37.5 37.5 37.5 37.5
 $ honeybee            : num  0.75 0.75 0.75 0.75
 $ bumbles             : num  0.25 0.25 0.25 0.25
 $ andrena             : num  0.25 0.25 0.25 0.25
 $ osmia               : num  0.25 0.25 0.25 0.25
 $ MaxOfUpperTRange    : num  86 86 94.6 94.6 86
 $ MinOfUpperTRange    : num  52 52 57.2 57.2 52
 $ AverageOfUpperTRange: num  71.9 71.9 79 79 71.
 $ MaxOfLowerTRange    : num  62 62 68.2 68.2 62
 $ MinOfLowerTRange    : num  30 30 33 33 30 30 3
 $ AverageOfLowerTRange: num  50.8 50.8 55.9 55.9
 $ RainingDays         : num  16 1 16 1 24 34 24
 $ AverageRainingDays  : num  0.26 0.1 0.26 0.1 0
 $ yield               : num  3813 4948 3867 4304
```

★ Histogram of response variable(target):



★ Boxplot of target variable(yield):



After looking at both histogram and boxplot of target variable(yield), we can see that the distribution of yield looks approximately symmetric (normal distribution).

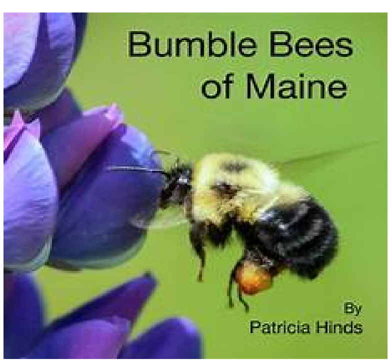
- ★ Summary of all variables:

```
   clonesize        honeybee           bumbles           andrena
 Min.   :10.00   Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:12.50   1st Qu.: 0.2500   1st Qu.:0.2500   1st Qu.:0.3800
 Median :12.50   Median : 0.2500   Median :0.2500   Median :0.5000
 Mean   :18.77   Mean   : 0.4171   Mean   :0.2824   Mean   :0.4688
 3rd Qu.:25.00   3rd Qu.: 0.5000   3rd Qu.:0.3800   3rd Qu.:0.6300
 Max.   :40.00   Max.   :18.4300   Max.   :0.5850   Max.   :0.7500
     osmia        MaxOfUpperTRange MinOfUpperTRange AverageOfUpperTRange
 Min.   :0.0000   Min.   :69.70    Min.   :39.0     Min.   :58.20
 1st Qu.:0.5000   1st Qu.:77.40    1st Qu.:46.8     1st Qu.:64.70
 Median :0.6300   Median :86.00    Median :52.0     Median :71.90
 Mean   :0.5621   Mean   :82.28    Mean   :49.7     Mean   :68.72
 3rd Qu.:0.7500   3rd Qu.:89.00    3rd Qu.:52.0     3rd Qu.:71.90
 Max.   :0.7500   Max.   :94.60    Max.   :57.2     Max.   :79.00
 MaxOfLowerTRange MinOfLowerTRange AverageOfLowerTRange  RainingDays
 Min.   :50.20    Min.   :24.30    Min.   :41.20        Min.   : 1.00
 1st Qu.:55.80    1st Qu.:27.00    1st Qu.:45.80        1st Qu.: 3.77
 Median :62.00    Median :30.00    Median :50.80        Median :16.00
 Mean   :59.31    Mean   :28.69    Mean   :48.61        Mean   :18.31
 3rd Qu.:66.00    3rd Qu.:30.00    3rd Qu.:50.80        3rd Qu.:24.00
 Max.   :68.20    Max.   :33.00    Max.   :55.90        Max.   :34.00
 AverageRainingDays      yield
 Min.   :0.06       Min.   :1638
 1st Qu.:0.10       1st Qu.:5125
 Median :0.26       Median :6107
 Mean   :0.32       Mean   :6013
 3rd Qu.:0.39       3rd Qu.:7022
 Max.   :0.56       Max.   :8969
```

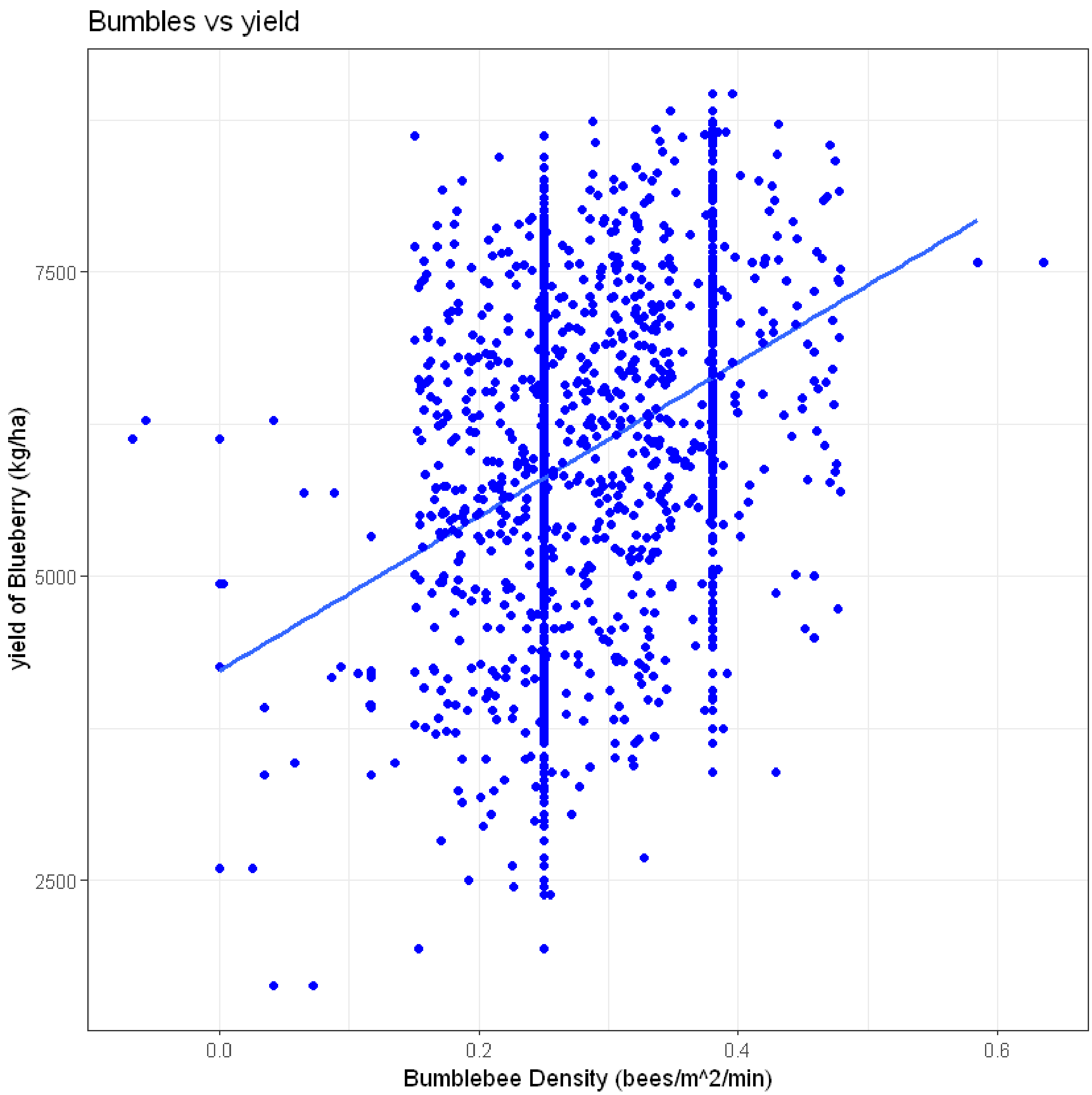
From the summary(just for an example), one can observe that the average number of rainy days per month is approximately 18 days and the average maximum temperature in the upper range is approximately 82 degree fahrenheit.

★ Scatter plot between yield and density of bumble bees:

In case you are curious, here is the image of bumble bees found in Maine, USA.

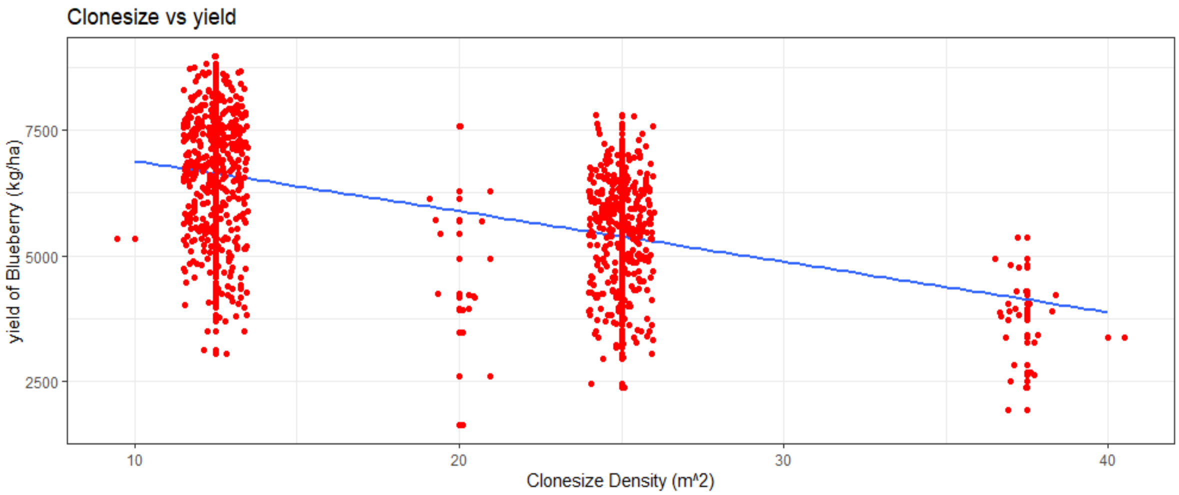


[Bumble Bees of Maine by Patricia Hinds | Blurb Books]



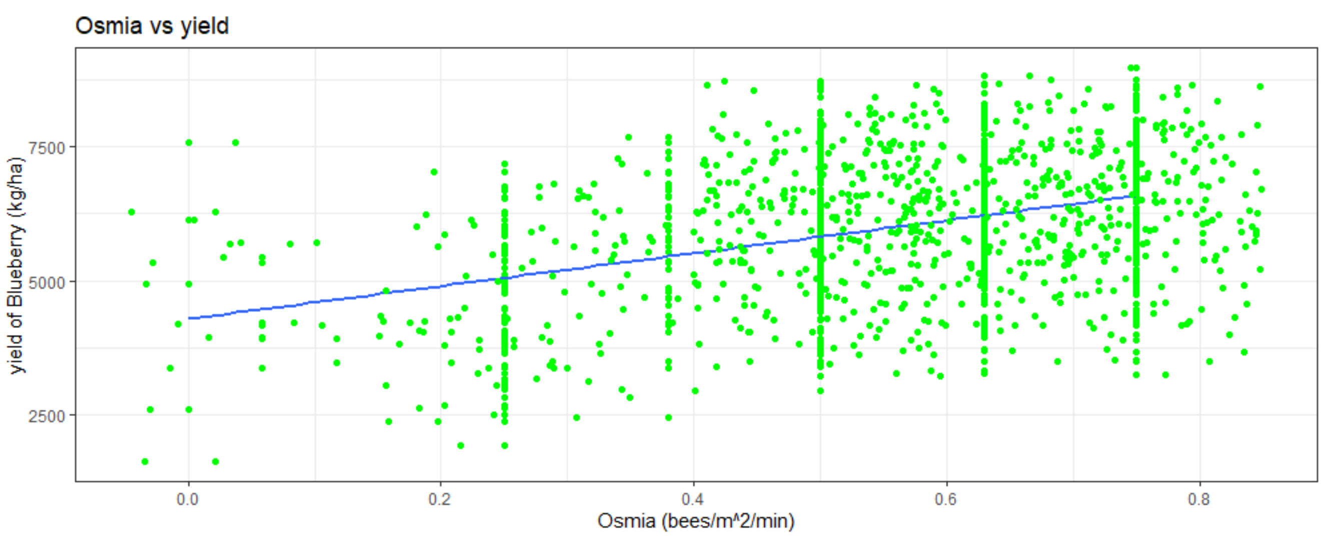
There is a slightly positive correlation between the bumblebees density and yield of wild blueberries. One can interpret this result as increasing the amount of bumble bees in the blueberry field can enhance the product of blueberry.

★ Scatter plot of Clone Size vs yield:



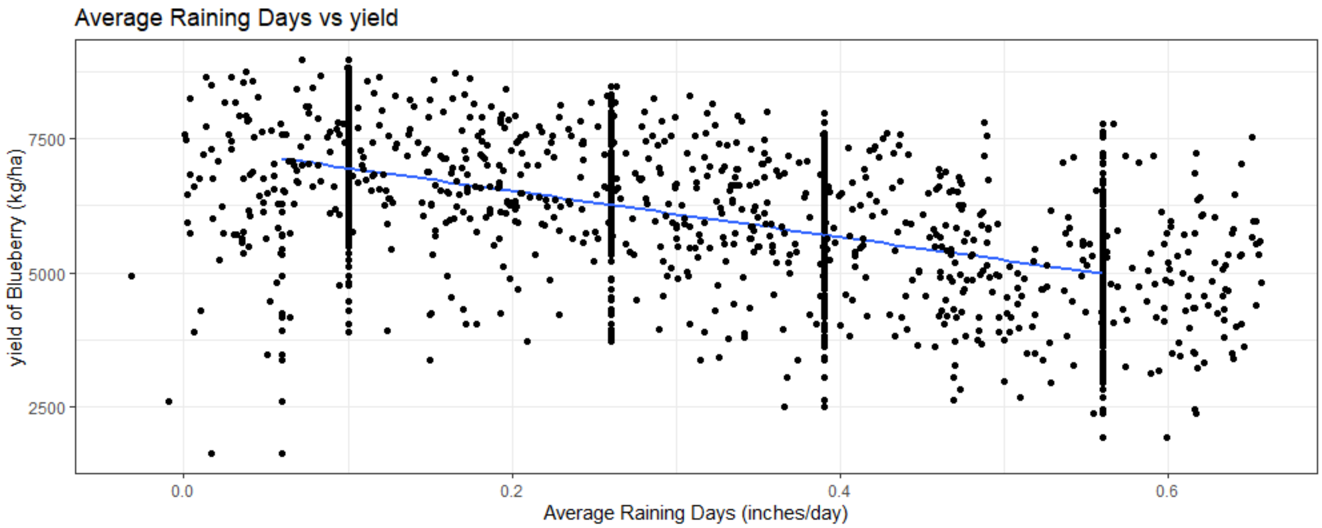
We see a negative correlation between Clonesize and Yield.One can interpret this result as decreasing the size of clones in the blueberry field can enhance the product of blueberry.

★ Scatter plot of Osmia Vs yield:



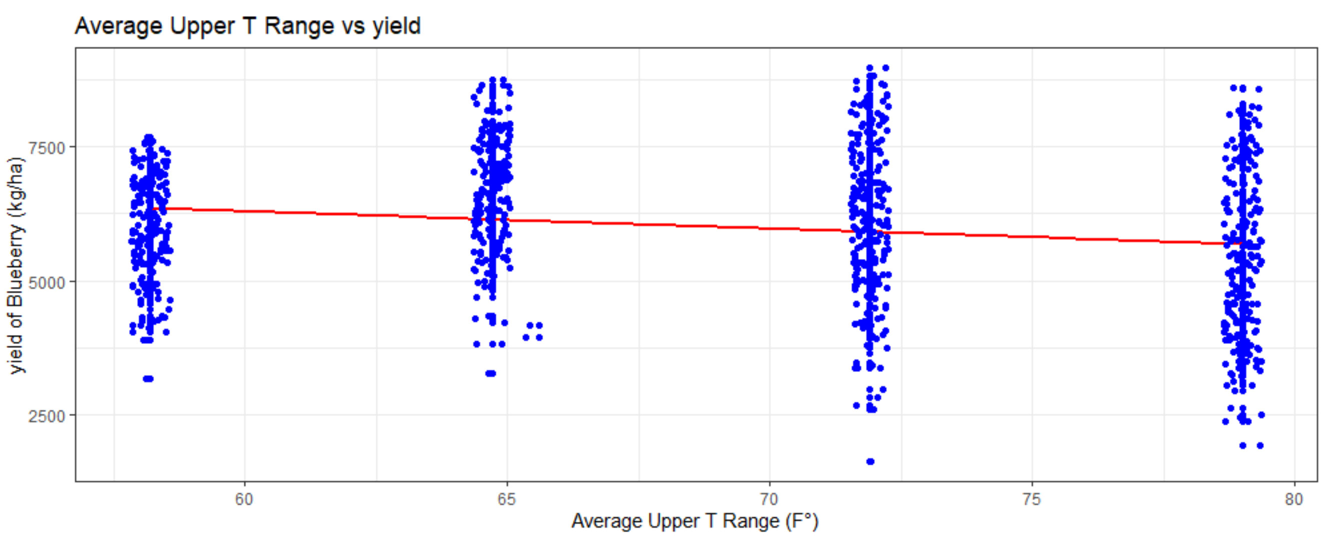
We see a positive correlation between Osmia and Yield.One can interpret this result as increasing the amount of Osmia bees in the blueberry field can enhance the product of blueberry.

★ Scatter plot of AverageRainningDays vs yield:



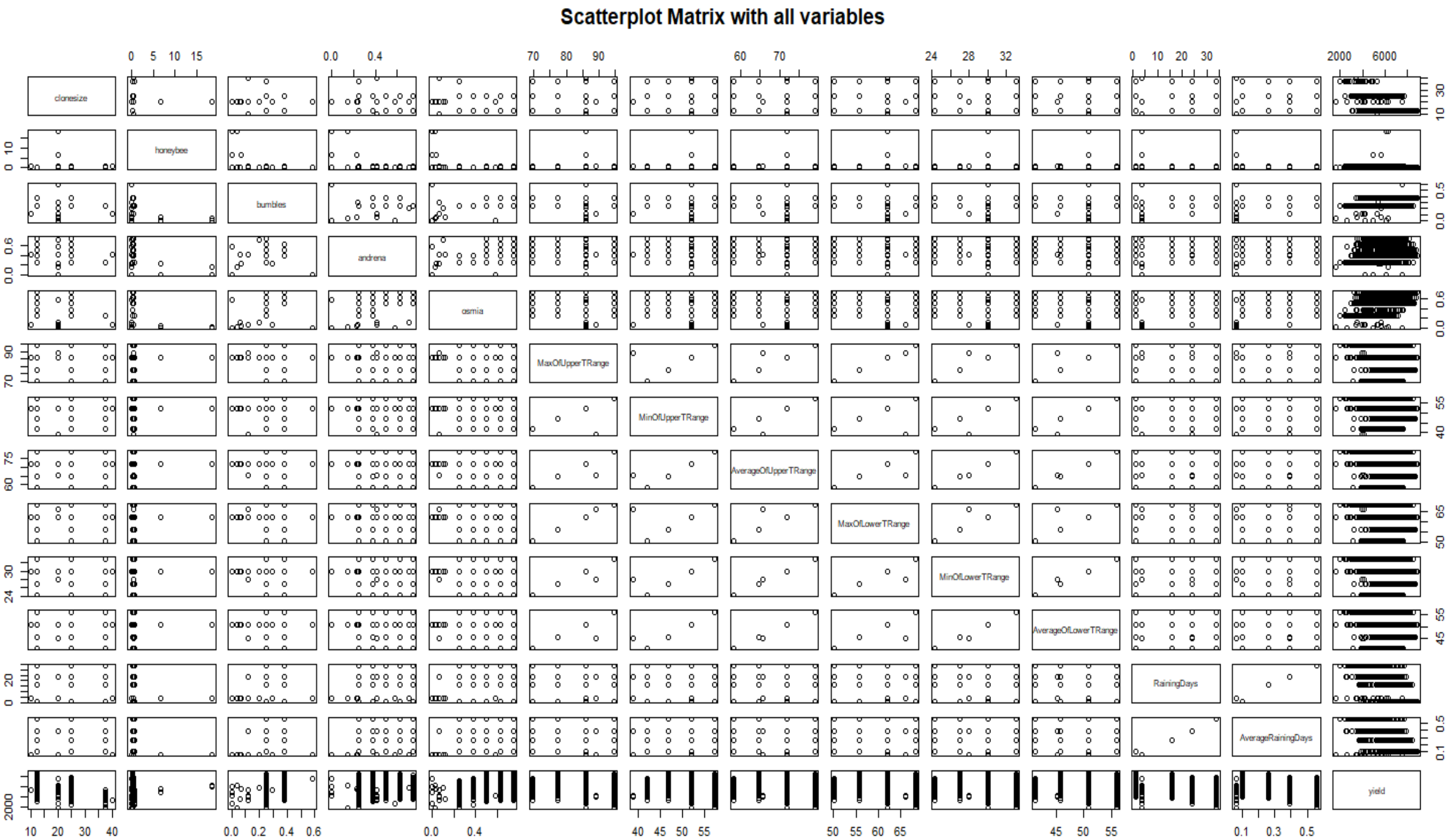
We see a negative correlation between Average Raining Days and Yield.One can interpret this result as decrease in the amount of Rainfall in the blueberry field can enhance the product of blueberry.

★ Scatter plot of Average Upper T Range V/S Yield

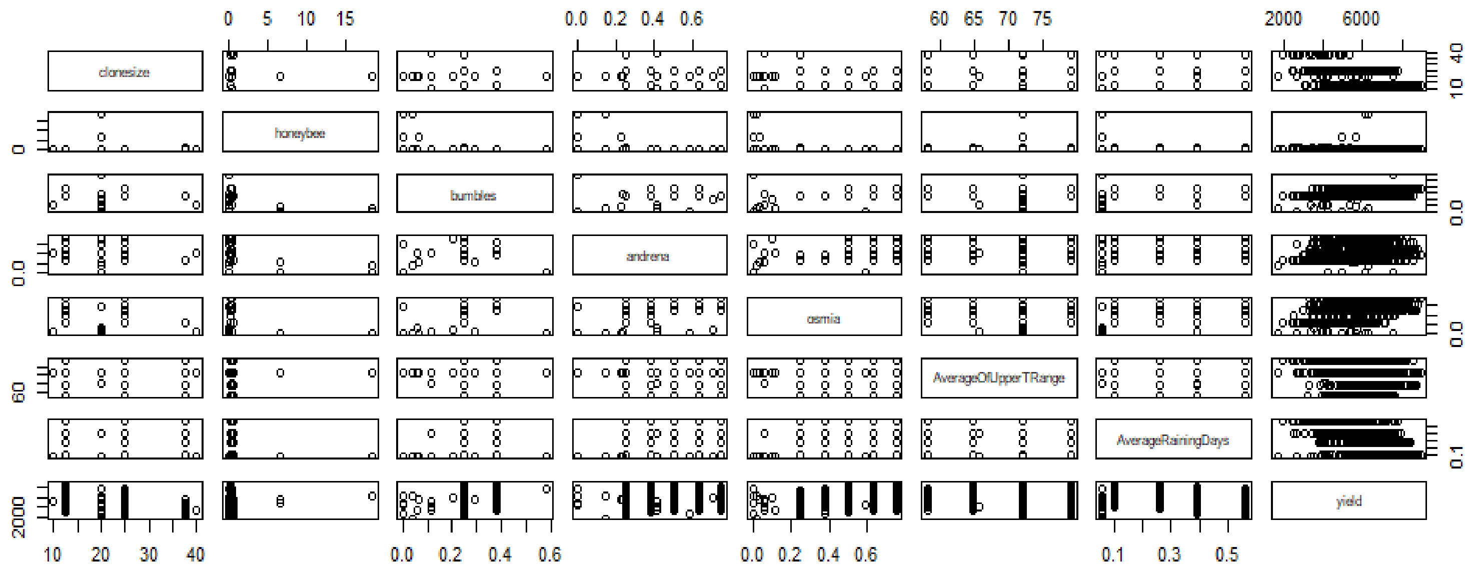


We see a negative correlation between Average Upper T Range and Yield.One can interpret cooler the temperature better the yield.

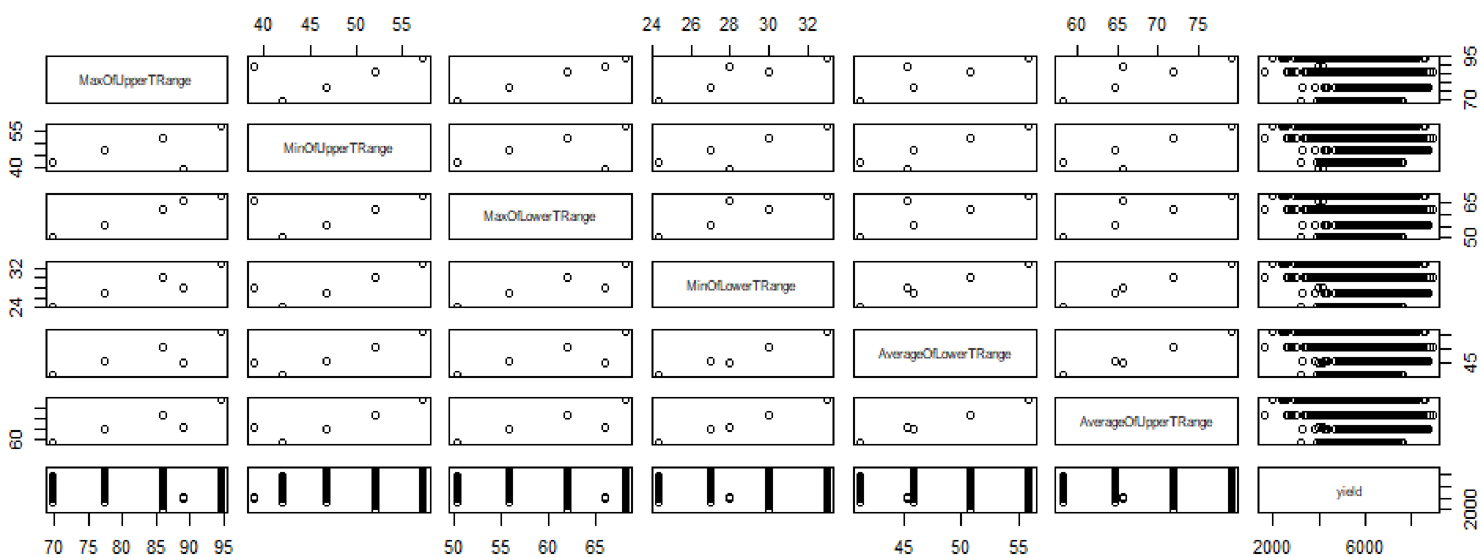
★ **Scatter plot of all variables.**



Simple Scatterplot Matrix



Simple Scatterplot Matrix



MaxOfUpperTRange, MinOfUpperTRange, MinOfLowerTRange, MaxOfLowerTRange, AverageOfUpperTRange and AverageOfLowerTRange are strongly correlated to each other.

- ★ **Train-Test Split**

```
> # Train-Test Split
> set.seed(602)
> df_split <- initial_split(df, prop = 0.8)
> df_train <- training(df_split)
> df_test <- testing(df_split)
```

Since we have only 777 observations we decided to split the data into 80/20 split, where 80% of the data was used to train the model and 20% of the data was used for testing it.

3) Feature Selection:

- ★ **Forward Selection Using AIC**

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
yield ~ 1

Final Model:
yield ~ RainingDays + clonesize + osmia + bumbles + MaxOfUpperTRange +
    AverageRainingDays + honeybee + andrena + AverageOfUpperTRange +
    MaxOfLowerTRange + MinOfLowerTRange


                   Step Df  Deviance Resid. Df Resid. Dev      AIC
1                                          620 1174342541 8977.091
2         + RainingDays  1 371603214       619  802739327 8742.842
3           + clonesize  1 309118468       618  493620859 8442.873
4               + osmia  1 156162348       617  337458511 8208.691
5             + bumbles  1  64610727       616  272847784 8078.711
6    + MaxOfUpperTRange  1  25151008       615  247696776 8020.655
7  + AverageRainingDays  1  19654047       614  228042729 7971.315
8            + honeybee  1   8208540       613  219834190 7950.550
9             + andrena  1   5204697       612  214629492 7937.671
10 + AverageOfUpperTRange  1   3081352       611  211548140 7930.691
11   + MaxOfLowerTRange  1  91257729       610  120290411 7582.109
12   + MinOfLowerTRange  1   1450350       609  118840062 7576.576
```

There were 13 explanatory variables and the forward selection method suggested using 11 of those explanatory variables.

★ **Backward selection**

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
yield ~ clonesize + honeybee + bumbles + andrena + osmia + MaxOfUpperTRange +
    MinOfUpperTRange + AverageOfUpperTRange + MaxOfLowerTRange +
    MinOfLowerTRange + AverageOfLowerTRange + RainingDays + AverageRainingDays

Final Model:
yield ~ clonesize + honeybee + bumbles + andrena + osmia + MaxOfUpperTRange +
    MinOfUpperTRange + AverageOfUpperTRange + MaxOfLowerTRange +
    RainingDays + AverageRainingDays


                    Step Df Deviance Resid. Df Resid. Dev      AIC
1                                          609  118840062 7576.576
2 - AverageOfLowerTRange  0        0       609  118840062 7576.576
3     - MinOfLowerTRange  0        0       609  118840062 7576.576
```

There were 13 explanatory variables and the backward selection method suggested removing 2 of the explanatory variables as shown above.

★ **Stepwise Selection (Both Direction)**

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
yield ~ clonesize + honeybee + bumbles + andrena + osmia + MaxOfUpperTRange +
    MinOfUpperTRange + AverageOfUpperTRange + MaxOfLowerTRange +
    MinOfLowerTRange + AverageOfLowerTRange + RainingDays + AverageRainingDays

Final Model:
yield ~ clonesize + honeybee + bumbles + andrena + osmia + MaxOfUpperTRange +
    MinOfUpperTRange + MaxOfLowerTRange + RainingDays + AverageRainingDays


                    Step Df Deviance Resid. Df Resid. Dev      AIC
1                                          609  118840062 7632.441
2 - AverageOfLowerTRange  0      0.0       609  118840062 7632.441
3     - MinOfLowerTRange  0      0.0       609  118840062 7632.441
4 - AverageOfUpperTRange  1 825416.3       610  119665478 7630.084
```

There were 13 explanatory variables and the backward selection method suggested removing 3 of the explanatory variables as shown above.

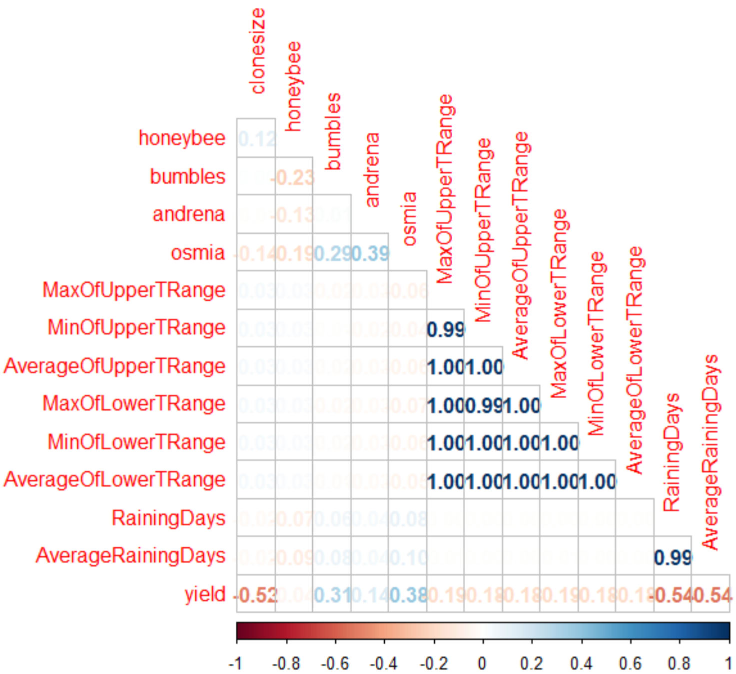
★ Subset Selection using Mallow CP Score

```
> cbind(best$which, best$cp)
```

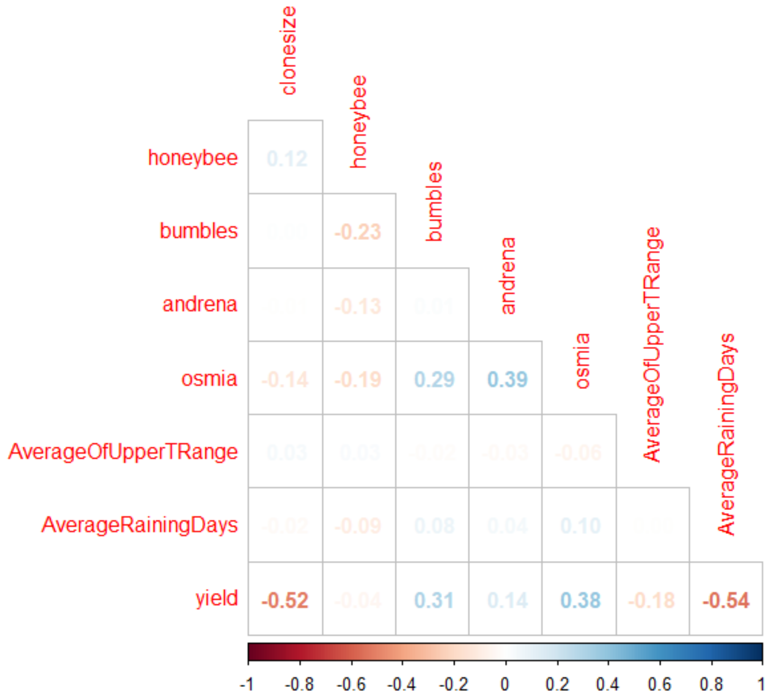
	(Intercept)	clonesize	RainingDays	honeybee	bumbles	andrena	osmia	AverageOfUpperTRange	AverageOfLowerTRange	AverageRainingDays	
3	1	1	1	0	0	0	1	0	0	0	466.21285
3	1	1	1	0	1	0	0	0	0	0	505.53790
3	1	1	1	0	0	0	0	1	0	0	881.04879
3	1	1	1	0	0	0	0	0	1	0	881.86963
3	1	1	1	0	0	1	0	0	0	0	894.12947
3	1	1	1	0	0	0	0	0	0	1	980.35709
3	1	1	1	1	0	0	0	0	0	0	991.35738
4	1	1	1	0	1	0	1	0	0	0	210.74412
4	1	1	1	0	0	0	1	0	1	0	379.05173
4	1	1	1	0	0	0	1	1	0	0	379.28471
4	1	1	1	0	1	0	0	0	1	0	402.54674
4	1	1	1	0	1	0	0	1	0	0	402.56848
4	1	1	1	0	1	1	0	0	0	0	412.16116
4	1	1	1	0	0	0	1	0	0	1	413.43471
4	1	1	1	1	0	0	1	0	0	0	461.70693
4	1	1	1	0	1	0	0	0	0	1	462.94941
4	1	1	1	0	0	1	1	0	0	0	466.94858
5	1	1	1	0	1	0	1	0	1	0	123.48532
5	1	1	1	0	1	0	1	1	0	0	124.08252
5	1	1	1	0	1	0	1	0	0	1	132.53074
5	1	1	1	1	1	0	1	0	0	0	179.36804
5	1	1	1	0	1	1	1	0	0	0	203.32956
5	1	1	1	0	1	1	0	0	1	0	313.99384
5	1	1	1	0	1	1	0	1	0	0	314.09773
5	1	1	1	0	0	0	1	0	1	1	325.63663
5	1	1	1	0	0	0	1	1	0	1	325.73668
5	1	1	1	0	1	0	0	1	0	1	358.36967
6	1	1	1	0	1	0	1	0	1	1	44.49214
6	1	1	1	0	1	0	1	1	0	1	44.94910
6	1	1	1	1	1	0	1	0	1	0	90.56254
6	1	1	1	1	1	0	1	1	0	0	91.15815
6	1	1	1	1	1	0	1	0	0	1	111.63744
6	1	1	1	0	1	1	1	0	1	0	116.43002
6	1	1	1	0	1	1	1	1	0	0	117.00734
6	1	1	1	0	1	0	1	1	1	0	118.96338
6	1	1	1	0	1	1	1	0	0	1	124.22009
6	1	1	1	1	1	1	1	0	0	0	168.43952
7	1	1	1	1	1	0	1	0	1	1	22.35108
7	1	1	1	1	1	0	1	1	0	1	22.81456
7	1	1	1	0	1	1	1	0	1	1	36.55493
7	1	1	1	0	1	1	1	1	0	1	36.99049
7	1	1	1	0	1	0	1	1	1	1	42.74472
7	1	1	1	1	1	1	1	0	1	0	79.97709
7	1	1	1	1	1	1	1	1	0	0	80.54891
7	1	1	1	1	1	0	1	1	1	0	86.19298
7	1	1	1	1	1	1	1	0	0	1	100.35810
7	1	1	1	0	1	1	1	1	1	0	112.31469
8	1	1	1	1	1	1	1	0	1	1	11.41918
8	1	1	1	1	1	1	1	1	0	1	11.85840
8	1	1	1	1	1	0	1	1	1	1	20.54928
8	1	1	1	0	1	1	1	1	1	1	35.14283
8	1	1	1	1	1	1	1	1	1	0	76.09014
8	1	1	1	1	1	1	0	1	1	1	245.19329
8	1	1	1	1	0	1	1	1	1	1	327.09933

So the Mallow CP score suggests using 8 variables (clonesize, RainingDays, honeybee, bumbles, andrena, osmia, AverageOfLowerTRange and AverageRainingDays)

★ **Correlation Heatmap.**



Before



After

As we see at the before heat map there is a major issue of high collinearity between MaxOfUpperTRange, MinOFUpperTRange, AverageOfUpperTRange, MaxOfLoweTRange, MinOfLowerTRange, AverageOfLowerTRange with each other and also between RainingDays and AverageRainingDays.

So, based on the above heat map to avoid the issue of multicollinearity we decide to choose one of each variable from all high correlating variables. We choose AverageOfUpperTRange and AverageRainingDays from each correlating group.

On the right side is the After heat map which shows the correlation between the remaining variables.

Building a new dataframe (df1) based on the selected Variables.

★ Model Building

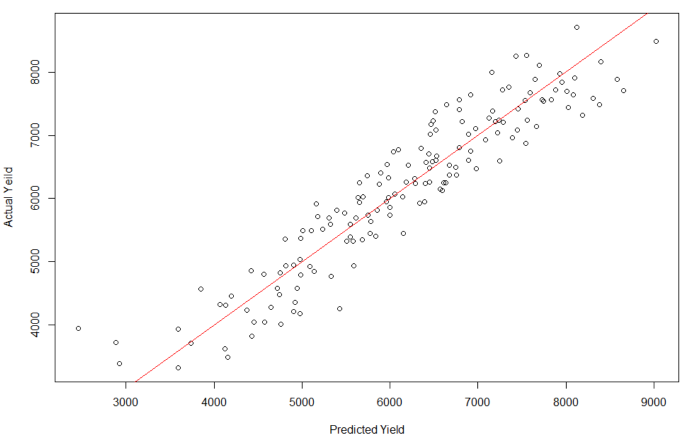
1) Linear Models

(i) Linear Model using forward selection with AIC:

```
Call:
lm(formula = yield ~ . - AverageOfUpperTRange - MinOfUpperTRange,
    data = df_train)

Residuals:
    Min      1Q  Median      3Q     Max
-2850.1  -277.4    36.8   283.2  1090.3

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          13108.764    387.569  33.823  < 2e-16 ***
clonesize              -98.725      2.598 -38.006  < 2e-16 ***
honeybee               108.449     17.673   6.136 1.52e-09 ***
bumbles               6079.182    293.782  20.693  < 2e-16 ***
andrena                588.006    123.088   4.777 2.23e-06 ***
osmia                 2274.498    122.930  18.502  < 2e-16 ***
MaxOfUpperTRange    -24462.910   1191.690 -20.528  < 2e-16 ***
MaxOfLowerTRange     25626.596   1191.484  21.508  < 2e-16 ***
MinOfLowerTRange     19898.262   2352.917   8.457  < 2e-16 ***
AverageOfLowerTRange -1746.365    849.124  -2.057 0.040144 *
RainingDays             39.709     11.535   3.442 0.000616 ***
AverageRainingDays   -7615.322    820.390  -9.283  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 441.7 on 609 degrees of freedom
Multiple R-squared:  0.8988,	Adjusted R-squared:  0.897
F-statistic: 491.7 on 11 and 609 DF,  p-value: < 2.2e-16
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        460.
```

clonesize	1.03994224041425
honeybee	1.14739669238205
bumbles	1.22047628683293
andrena	1.22985395792755
osmia	1.43469136565064
MaxOfUpperTRange	384611.508066772
MaxOfLowerTRange	196373.96702936
MinOfLowerTRange	188624.832861505
AverageOfLowerTRa…	68023.6058430748
RainingDays	60.5147062028656
AverageRainingDays	60.9627848792673



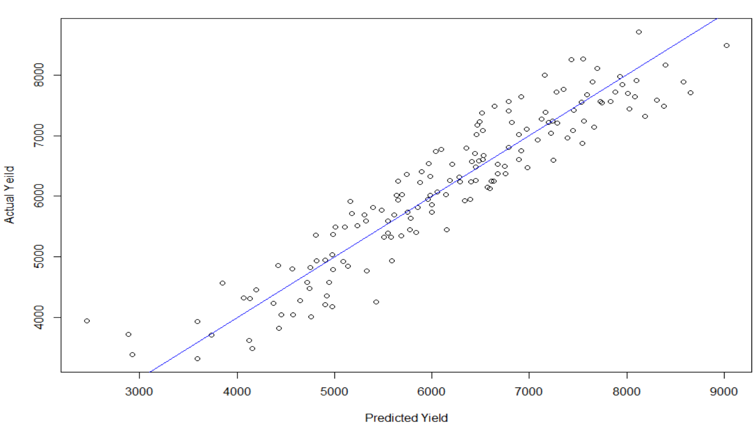

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.874
```

- All the variables are significant with root mean square error of 460.
- The values for Multiple R-Squared and Adjusted R-square are 0.8988 and 0.897 respectively.
- R-Squared from test-data is 0.874.
- Six out of 11 variables have a VIF score greater than 2.5. So, there is a major multicollinearity issue.
- There are some points in the Actual Value vs Predicted Value graph that are somewhat further from the red line.

(ii) Linear Model using backward selection with AIC:

```
Call:
lm(formula = yield ~ . - AverageOfLowerTRange - MinOfLowerTRange,
    data = df_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-2850.1  -277.4    36.8   283.2  1090.3 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          11886.309    440.060  27.011  < 2e-16 ***
clonesize              -98.725      2.598 -38.006  < 2e-16 ***
honeybee               108.449     17.673   6.136 1.52e-09 ***
bumbles               6079.182    293.782  20.693  < 2e-16 ***
andrena                588.006    123.088   4.777 2.23e-06 ***
osmia                 2274.498    122.930  18.502  < 2e-16 ***
MaxOfUpperTRange    -19450.502   1215.943 -15.996  < 2e-16 ***
MinOfUpperTRange      1906.656    699.373   2.726 0.006590 ** 
AverageOfUpperTRange  1746.365    849.124   2.057 0.040144 *  
MaxOfLowerTRange     23266.567   1881.019  12.369  < 2e-16 ***
RainingDays             39.709     11.535   3.442 0.000616 ***
AverageRainingDays   -7615.322    820.390  -9.283  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 441.7 on 609 degrees of freedom
Multiple R-squared:  0.8988,	Adjusted R-squared:  0.897 
F-statistic: 491.7 on 11 and 609 DF,  p-value: < 2.2e-16
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        460.
```



clonesize	1.03994224050367
honeybee	1.1473966926959
bumbles	1.22047628697997
andrena	1.22985395810147
osmia	1.43469136584669
MaxOfUpperTRange	359714.685994654
MinOfUpperTRange	47816.3377447188
AverageOfUpperTRan…	136605.184792084
MaxOfLowerTRange	457520.719795583
RainingDays	60.5147061997944
AverageRainingDays	60.9627848761812

VIF Scores

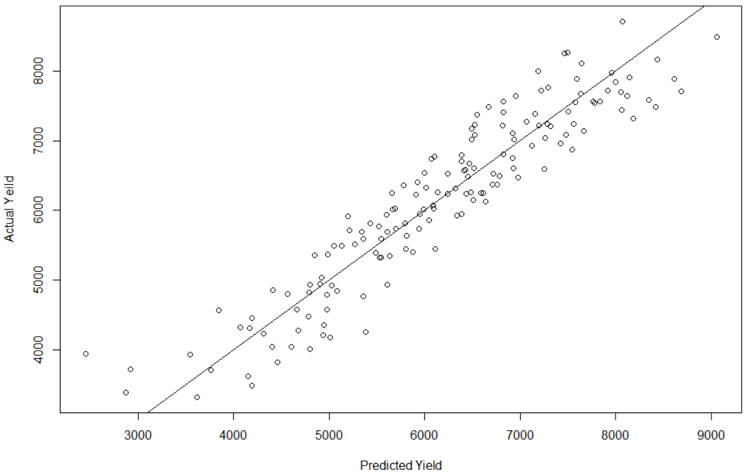
```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.874
```

- All the variables are significant with root mean square error of 460.
- The values for Multiple R-Squared and Adjusted R-square are 0.8988 and 0.897 respectively.
- The R-Squared for test-data is 0.874.
- Six out of 11 variables have a VIF score greater than 2.5. So, there is a major multicollinearity issue.
- There are some points in the Actual Value vs Predicted Value graph that are somewhat further from the blue line.
- This model performance is similar to the Linear Model using forward selection with AIC.

(iii) Linear Model using both direction selection with BIC:

```
Call:
lm(formula = yield ~ . - AverageOfLowerTRange - MinOfLowerTRange -
    AverageOfUpperTRange, data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max
-2814.42  -280.21    30.83   287.44  1054.32

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         12572.695    287.589  43.718  < 2e-16 ***
clonesize             -98.852      2.604 -37.965  < 2e-16 ***
honeybee              110.562     17.690   6.250 7.71e-10 ***
bumbles              6057.377    294.366  20.578  < 2e-16 ***
andrena               593.770    123.381   4.812 1.88e-06 ***
osmia                2256.316    122.936  18.354  < 2e-16 ***
MaxOfUpperTRange   -21000.271    956.831 -21.948  < 2e-16 ***
MinOfUpperTRange     3311.179    151.238  21.894  < 2e-16 ***
MaxOfLowerTRange    26251.898   1199.518  21.885  < 2e-16 ***
RainingDays            40.736     11.555   3.525 0.000455 ***
AverageRainingDays  -7697.918    821.573  -9.370  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 442.9 on 610 degrees of freedom
Multiple R-squared:  0.8981,	Adjusted R-squared:  0.8964
F-statistic: 537.6 on 10 and 610 DF,  p-value: < 2.2e-16
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        462.
```



clonesize	1.03948311114509
honeybee	1.14079442304643
bumbles	1.21924914873651
andrena	1.22982796714172
osmia	1.43143251630803
MaxOfUpperTRange	229210.384290042
MinOfUpperTRange	2088.09507351557
MaxOfLowerTRange	188440.352634334
RainingDays	60.4039965905744
AverageRainingDays	60.8460120640412

VIF Score

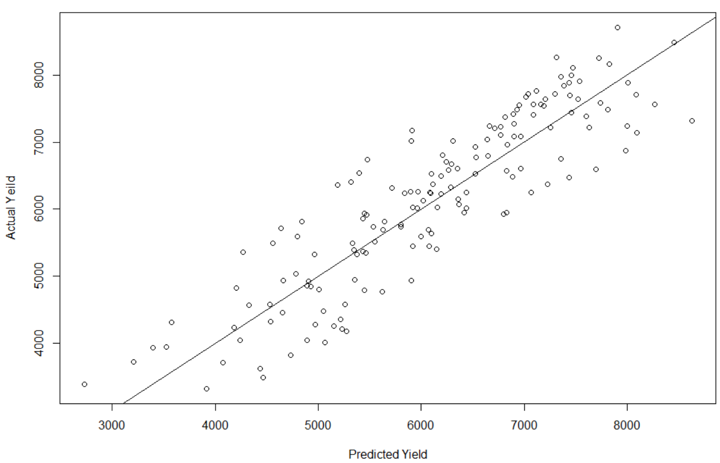
```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.874
```

- All the variables are significant with a root mean square error of 462.
- The values for Multiple R-Squared and Adjusted R-square are 0.8981 and 0.8964 respectively.
- The R-Squared for test-data is 0.874
- Five out of 10 variables have a VIF score greater than 2.5. So, there is a major multicollinearity issue.
- There are some points in the Actual Value vs Predicted Value graph that are somewhat further from the black line.

(iv) Linear Model using CP Mellow with Subset Selection:

```
Call:
lm(formula = yield ~ clonesize + RainingDays + honeybee + bumbles +
    andrena + osmia + AverageOfLowerTRange + AverageRainingDays,
    data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max
-2601.84  -370.60    46.03   415.15  1455.88

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           7969.639    265.755  29.989  < 2e-16 ***
clonesize              -97.503      3.486 -27.966  < 2e-16 ***
RainingDays             37.344     15.432   2.420 0.015815 *
honeybee               122.928     23.665   5.195 2.80e-07 ***
bumbles               6145.213    393.571  15.614  < 2e-16 ***
andrena                633.698    165.198   3.836 0.000138 ***
osmia                 2246.765    164.045  13.696  < 2e-16 ***
AverageOfLowerTRange   -36.896      4.394  -8.398 3.17e-16 ***
AverageRainingDays   -7415.268   1096.762  -6.761 3.20e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 593.2 on 612 degrees of freedom
Multiple R-squared:  0.8166,	Adjusted R-squared:  0.8142
F-statistic: 340.6 on 8 and 612 DF,  p-value: < 2.2e-16
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        584.
```



clonesize	1.03898182991798
RainingDays	60.2244493124665
honeybee	1.13754342676976
bumbles	1.20928243855015
andrena	1.22774216286004
osmia	1.40601269344496
AverageOfLowerTRa...	1.00667301838998
AverageRainingDays	60.659018584147

VIF Score

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.791
```

- All the variables are significant with a root mean square error of 584.
- The values for Multiple R-Squared and Adjusted R-square are 0.8166 and 0.8142 respectively.
- The R-Squared for test-data is 0.791.

- Two out of 8 variables have a VIF score greater than 2.5. So, there is a major multicollinearity issue.
- There are some points in the Actual Value vs Predicted Value graph that are somewhat further from the black line.

For the further model development selected variables after handling multicollinearity using both VIF scores and correlation heat map will be used.

★ **Train-Test Split based on new dataframe (df1).**

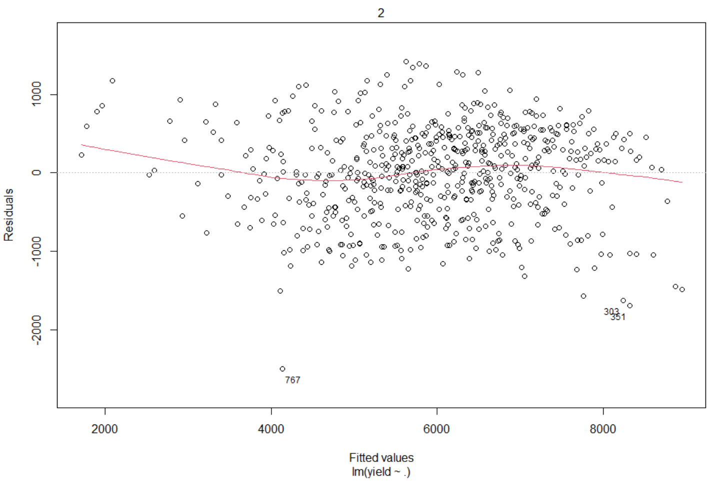
```
> set.seed(602)
> df1_split <- initial_split(df1, prop = 0.8)
> df1_train <- training(df1_split)
> df1_test <- testing(df1_split)
```

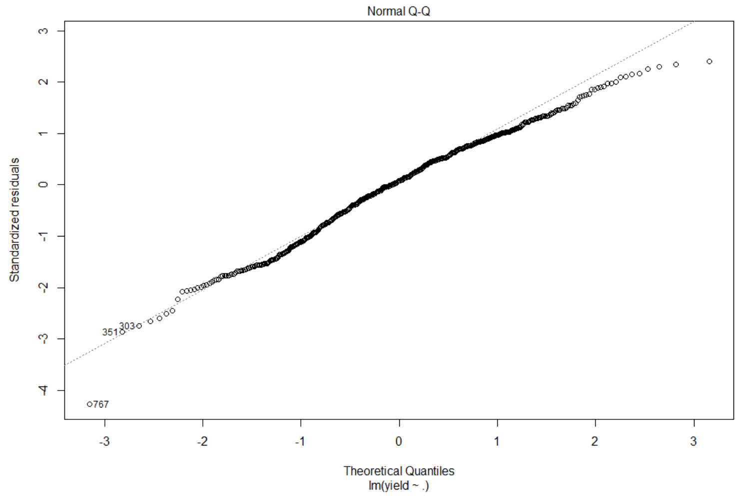
Since we have only 777 observations we decided to split the data into 80/20 split, where 80% of the data was used to train the model and 20% of the data was used for testing it.

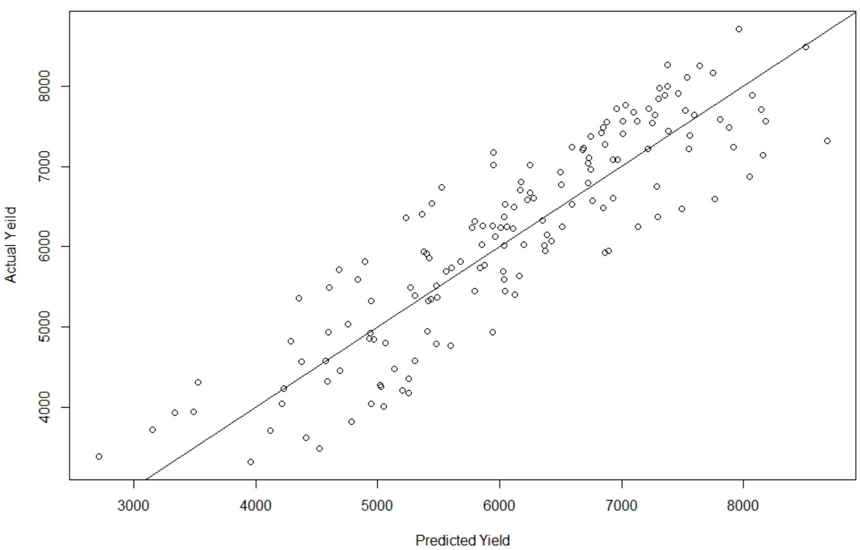
(V) Linear Regression Model:

```
Call:
lm(formula = yield ~ ., data = df1_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-2495.80  -394.01    46.58   442.94  1419.13 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           7849.211    262.423  29.911  < 2e-16 ***
clonesize              -97.456      3.501 -27.836  < 2e-16 ***
honeybee               129.178     23.624   5.468 6.63e-08 ***
bumbles               6074.034    394.202  15.408  < 2e-16 ***
andrena                640.210    165.877   3.860 0.000126 ***
osmia                 2199.526    163.656  13.440  < 2e-16 ***
AverageOfUpperTRange   -26.077      3.113  -8.378 3.69e-16 ***
AverageRainingDays   -4783.675    140.385 -34.075  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 595.7 on 613 degrees of freedom
Multiple R-squared:  0.8147,	Adjusted R-squared:  0.8126 
F-statistic: 385.1 on 7 and 613 DF,  p-value: < 2.2e-16
```

clonesize	1.03891203251679
honeybee	1.12278256498536
bumbles	1.20270890404981
andrena	1.2276826792534
osmia	1.3951501169476
AverageOfUpperTRan...	1.00675604506569
AverageRainingDays	1.01984548221808








```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        596.
```

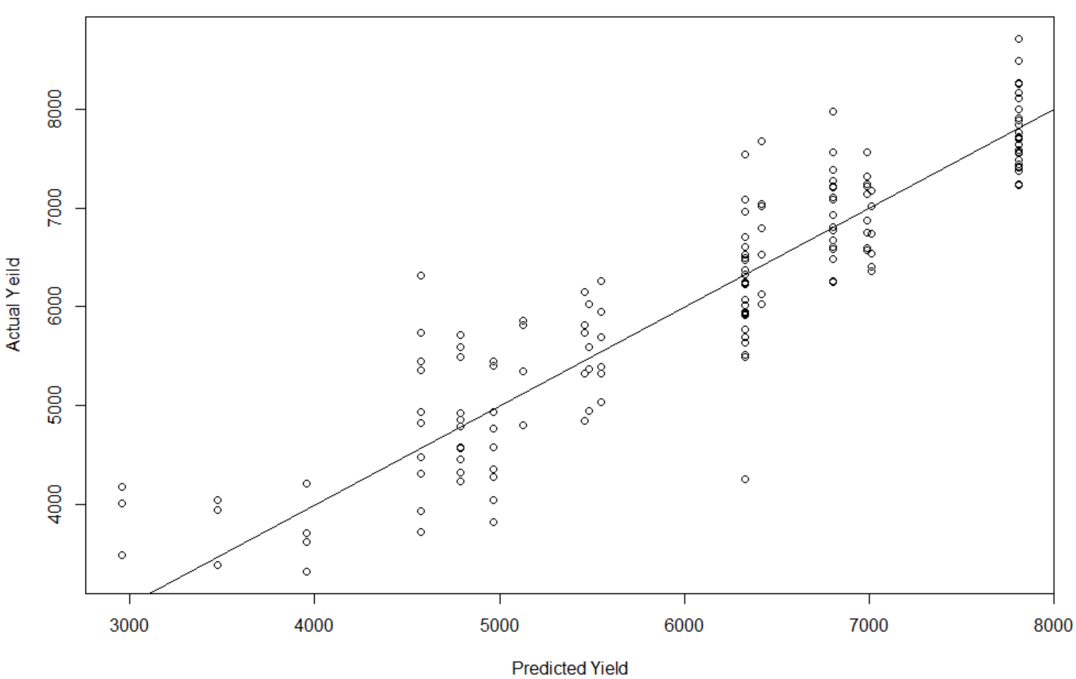
```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.783
```

- All the variables are significant with a root mean square error of 596.
- The values for Multiple R-Squared and Adjusted R-square are 0.8147 and 0.8126 respectively.
- The R-Squared from test-data is 0.783.
- All variables that have a VIF score are smaller than 2.5. So, there is no multicollinearity between variables.
- Residual V/S fitted plot looks good but we have a concern for point 767 which is further down than the horizontal line. We suspect it to be an outlier which is visible in the Normal Q-Q plot as well.

- There are some points in the Actual Value vs Predicted Value graph that are somewhat further from the black line.

(Vi) Decision Tree:

```
Call:
rpart(formula = yield ~ ., data = df1_train, method = "anova")
  n= 621

           CP nsplit rel error    xerror       xstd
1  0.28015645      0 1.0000000 1.0022197 0.05198585
2  0.14421999      1 0.7198435 0.7240662 0.04523218
3  0.09637701      2 0.5756236 0.5800563 0.03644040
4  0.07005610      3 0.4792465 0.5039189 0.03132275
5  0.06339236      4 0.4091904 0.4674378 0.02789269
6  0.02673472      5 0.3457981 0.3675077 0.02160530
7  0.02608808      6 0.3190634 0.3309049 0.02015999
8  0.02150293      7 0.2929753 0.3114418 0.01892339
9  0.02080589      8 0.2714724 0.3002551 0.01814547
10 0.01505789      9 0.2506665 0.2944826 0.01784623
11 0.01466188     10 0.2356086 0.2747744 0.01645451
12 0.01412569     11 0.2209467 0.2747744 0.01645451
13 0.01177829     12 0.2068210 0.2578151 0.01624320
14 0.01052790     13 0.1950427 0.2350858 0.01584289
15 0.01000000     15 0.1739869 0.2241690 0.01570458

Variable importance
  AverageRainingDays            honeybee           clonesize               osmia AverageOfUpperTRange             andrena
                  25                  20                  19                  14                  10                   9
             bumbles
                   3
```



```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        533.
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.828
```

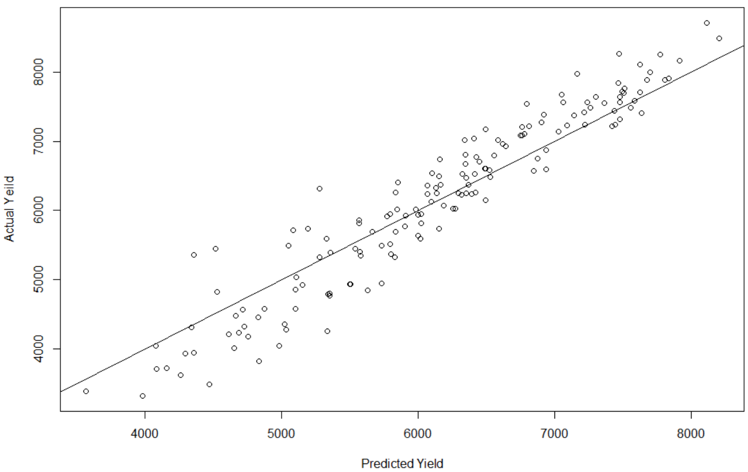
RMSE **RSQ**

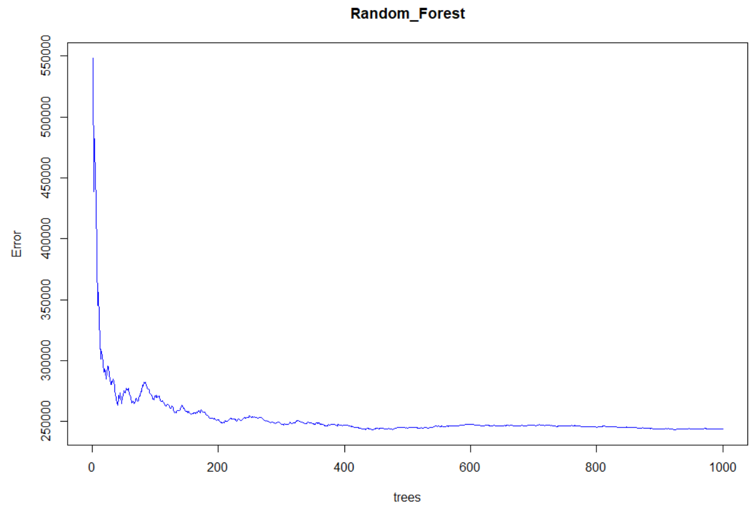
- The most important variable is AverageRainingDays.
- The Root Mean Square Error for the test data is 533.
- The R-Squared for the test data is 0.828.
- There are some points in the Actual Value vs Predicted Value graph that are somewhat further from the black line.
- This model performance is better as compared to the Linear model.

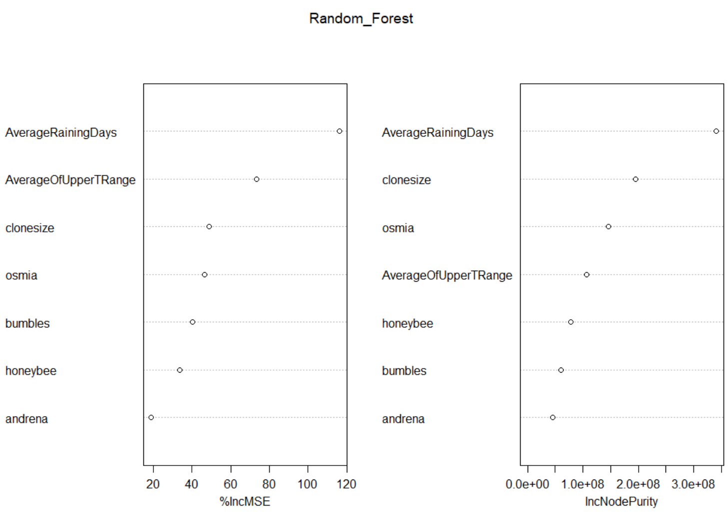
(Vii) Random Forest:

```
                Length Class  Mode
call               5   -none- call
type               1   -none- character
predicted        621   -none- numeric
mse             1000   -none- numeric
rsq             1000   -none- numeric
oob.times        621   -none- numeric
importance        14   -none- numeric
importanceSD       7   -none- numeric
localImportance    0   -none- NULL
proximity          0   -none- NULL
ntree              1   -none- numeric
mtry               1   -none- numeric
forest            11   -none- list
coefs              0   -none- NULL
y                621   -none- numeric
test               0   -none- NULL
inbag              0   -none- NULL
terms              3   terms  call
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        412.
```



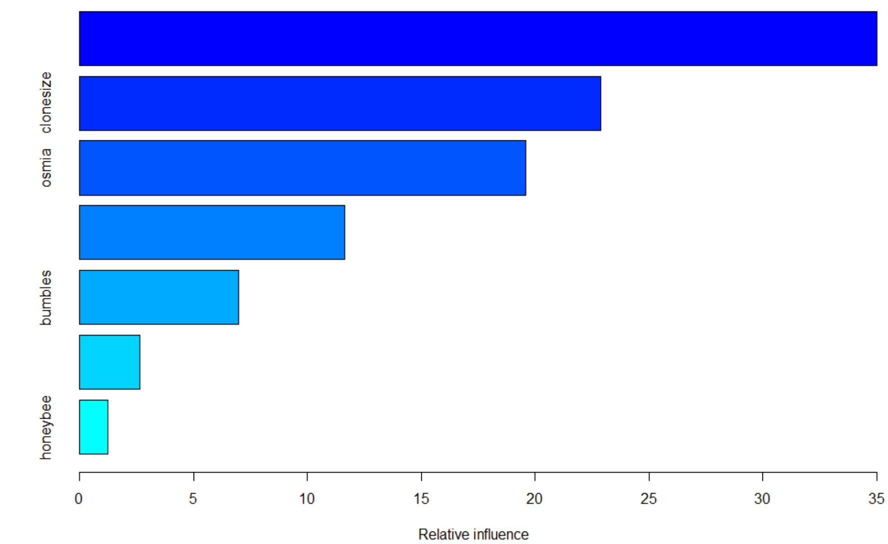





```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.914
```

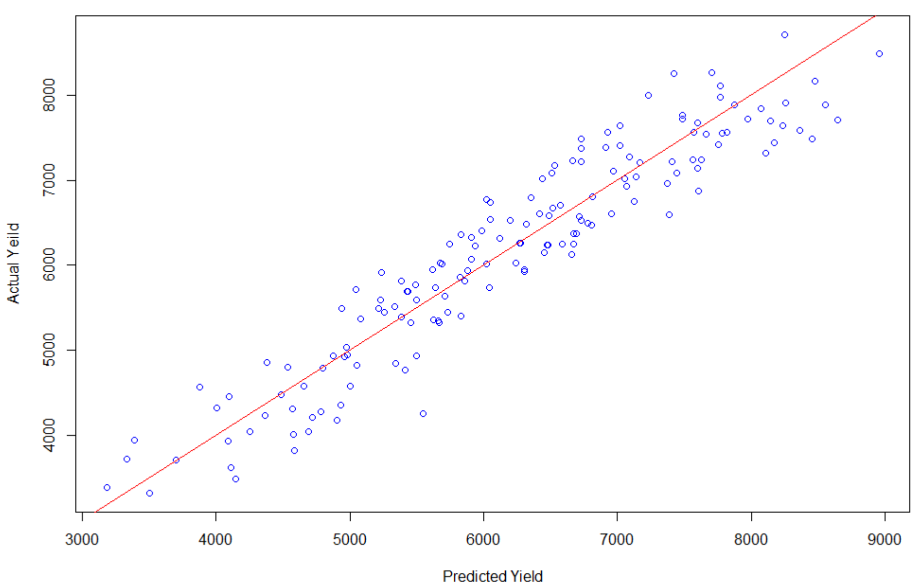
- The most important variable in terms of accuracy and Gini impurity is AverageRainingDays.
- The Root Mean Square Error for the test data is 412.
- The R-Squared for the test data is 0.914.
- Most of the points in the Actual Value vs Predicted Value graph are somewhat closer to the black line.
- This model performance is better as compared to the previous models.

(Viii) Gradient Boosted:




```
                                           var   rel.inf
AverageRainingDays       AverageRainingDays 35.004346
clonesize                         clonesize 22.875396
osmia                                 osmia 19.586503
AverageOfUpperTRange AverageOfUpperTRange 11.642518
bumbles                             bumbles  7.000006
andrena                             andrena  2.637065
honeybee                           honeybee  1.254167
```

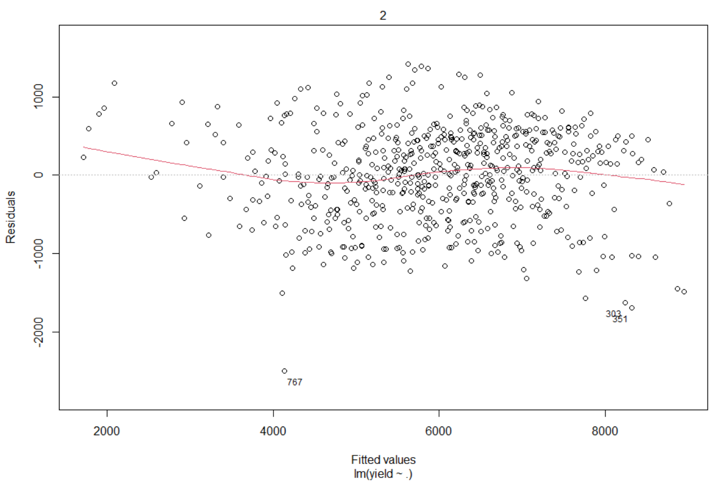
```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        429.
```

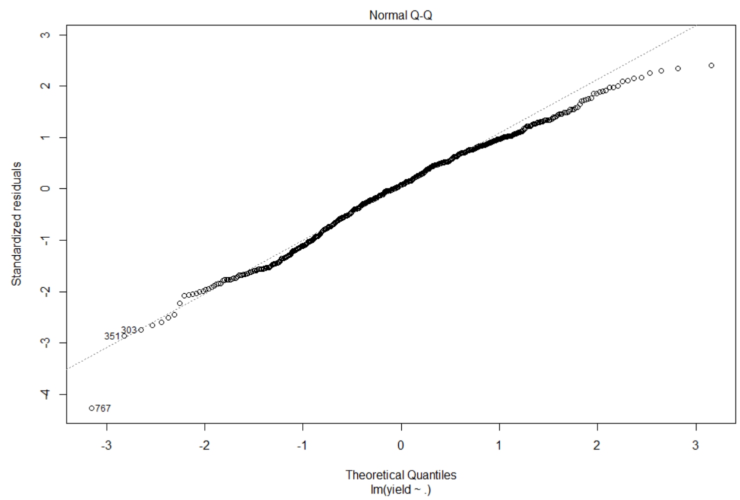



```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.889
```

- The most important variable is AverageRainingDays.
- The Root Mean Square Error for the test data is 429.
- The R-Squared for the test data is 0.889.
- Most of the points in the Actual Value vs Predicted Value graph are somewhat closer to the red line.
- This model performance is better as compared to the Linear model and Decision tree but inferior in performance as compared to the Random Forest.

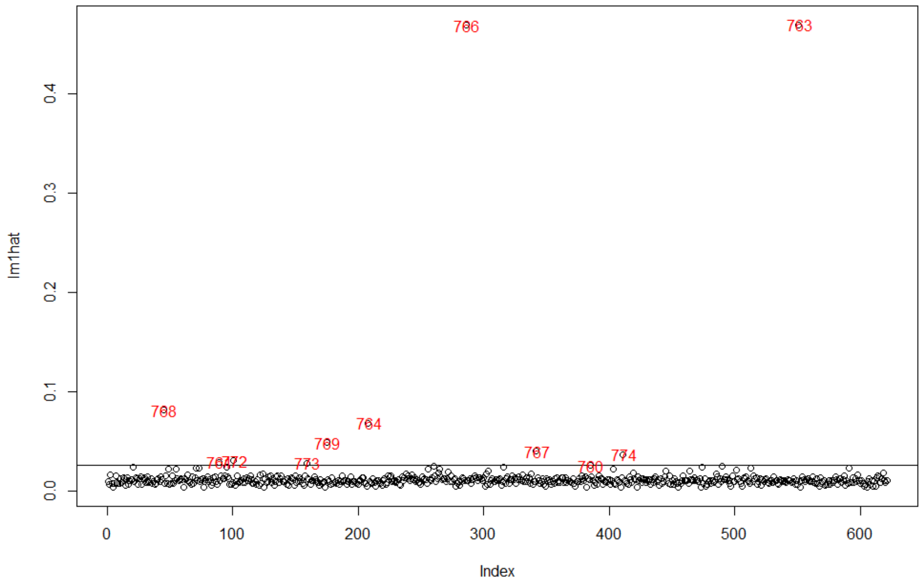
★ Outliers and influential observations





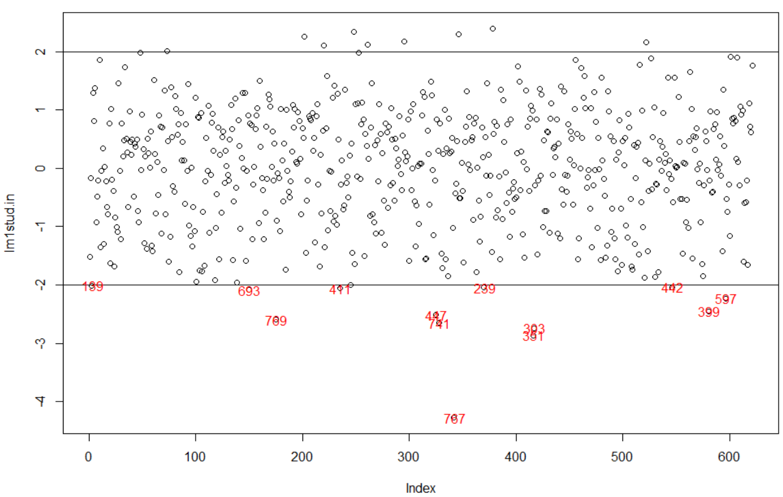
Above given Residual plot and Q-Q plot are from the linear model which shows potential outliers or influential points which we will investigate in this portion of the report.

(I) Leverage or Hat Values: -

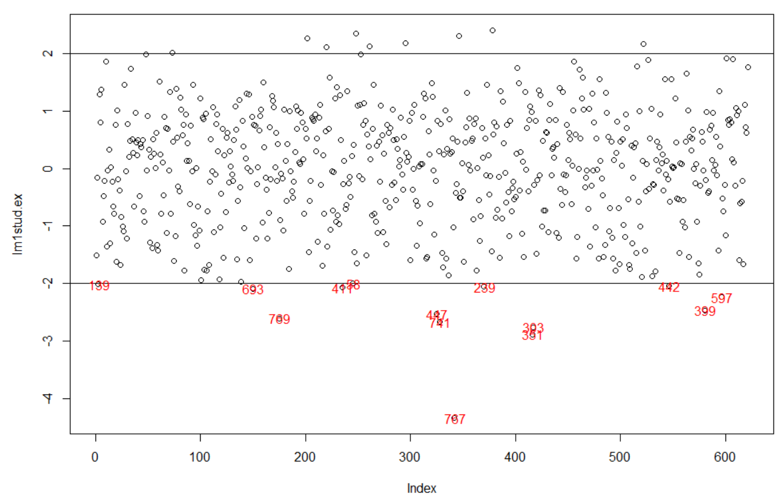


From the above plot we see there are two observations (766 and 763) are flagged as the influential points according to the hat values. Since their hat values are relatively large as compared to 2p/n.

(II) Studentized test: -



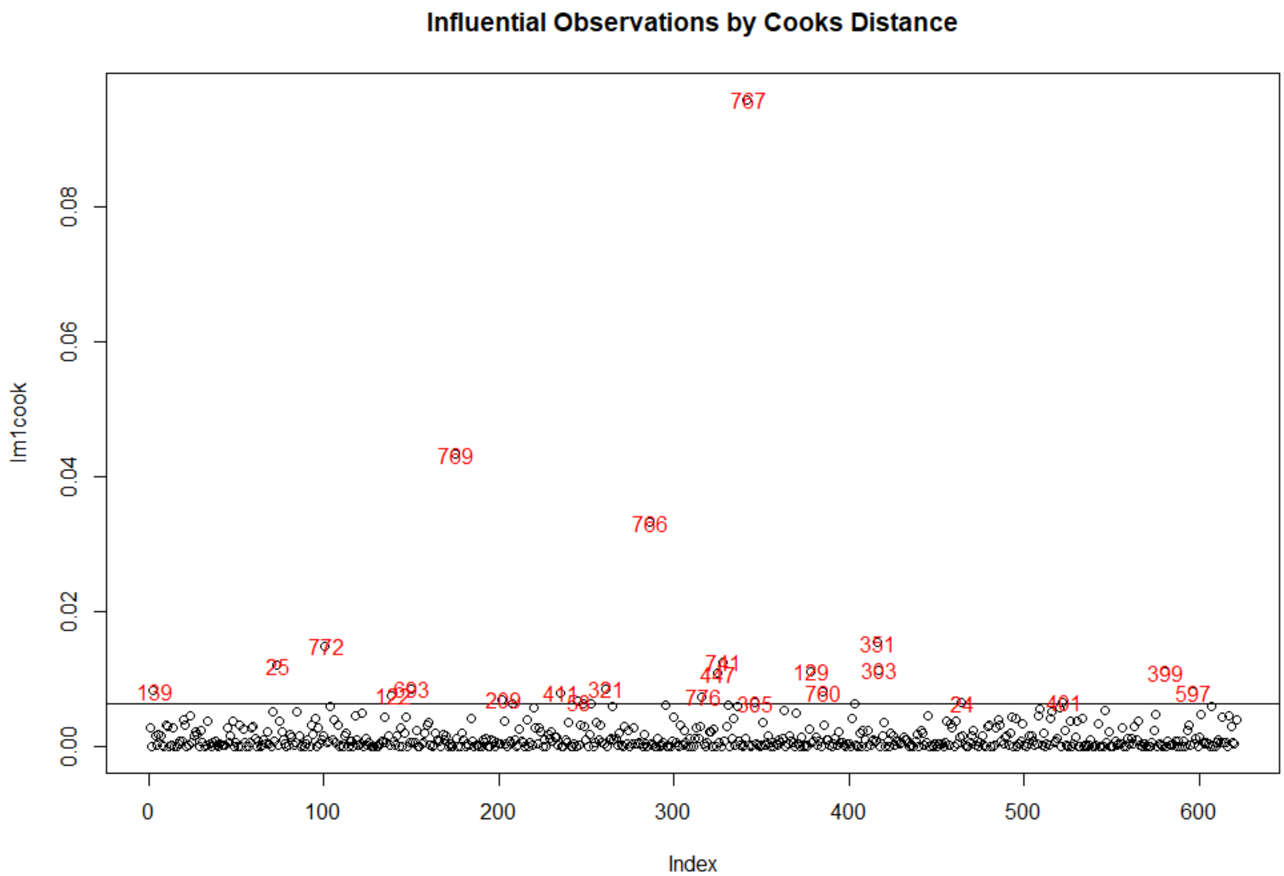
Internal



External

From our data observation 767 is relatively outside of -2 and 2, both for internally and externally studentized residuals.

(III) Cook's Distance: -



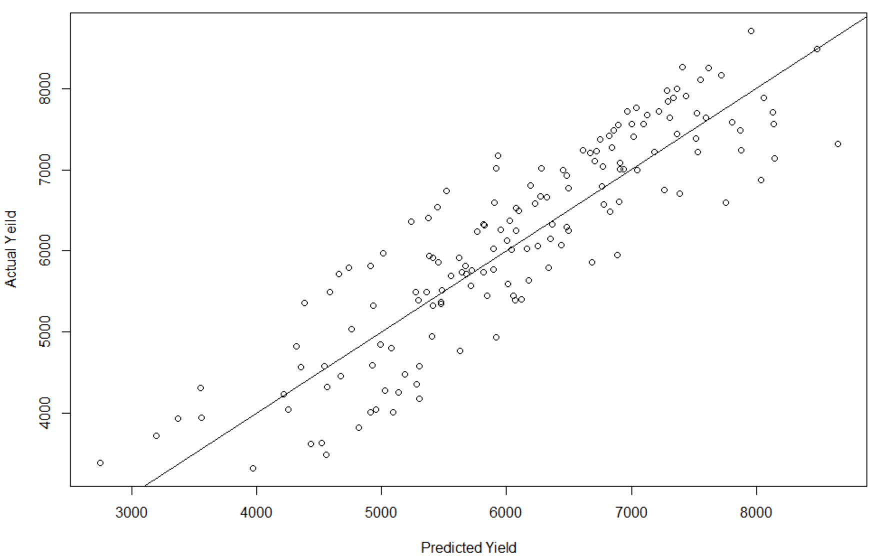
Again Observations 767 stand's out.

★ Refined Linear Model

```
> df2 <- df1[-c(767), ]
> set.seed(602)
> df2_split <- initial_split(df2, prop = 0.8)
> df2_train <- training(df2_split)
> df2_test <- testing(df2_split)
> ln_modeldf2 = lm(yield ~ . ,data = df2_train) # reduced model after handling multicollinearity
```

```
Call:
lm(formula = yield ~ ., data = df2_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1669.86  -399.24    52.03   431.33  1411.08 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           7919.531    259.597  30.507  < 2e-16 ***
clonesize              -97.926      3.448 -28.405  < 2e-16 ***
honeybee               122.824     23.009   5.338 1.33e-07 ***
bumbles               5824.271    389.959  14.936  < 2e-16 ***
andrena                572.013    164.730   3.472 0.000552 ***
osmia                 2146.059    163.155  13.154  < 2e-16 ***
AverageOfUpperTRange   -24.927      3.059  -8.149 2.07e-15 ***
AverageRainingDays   -4800.966    137.891 -34.817  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 586.9 on 612 degrees of freedom
Multiple R-squared:  0.8178,	Adjusted R-squared:  0.8157 
F-statistic: 392.5 on 7 and 612 DF,  p-value: < 2.2e-16
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        595.
```




```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.781
```

After removing one of the influential points we see slight improvement in this linear model.

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        448.
```

We also ran it through our best performing model (Random forest) and its performance degraded as shown above.

★ Model Comparison

Model Name	R-Squared	RMSE
Forward Selection Linear Model	0.874	460
Backward Selection Linear Model	0.874	460
Both direction Linear Model	0.874	462
Subset Selection Linear Model	0.791	584
Linear Model (dataframe = df1)	0.783	596
Decision Tree	0.828	533
Random Forest	0.914	412
Gradient Boost	0.889	429
Refined Linear Model	0.781	595

From the above Model Comparison we can say that the Random forest is the best performing model for this particular data.

★ Limitations

- We don’t have the information regarding duration of sunlight, we believe it is an important variable for the yield.
- This is a simulated data derived from the data collected from Maine, US over the time span of 30 years.
- This data is collected from a single location due to which the data has very limited variability.
- We do not know the exact reason why most of our observations have limited values, however, we assume this could be due to the following reasons: -
    a) The data is not actual data, it is simulated data.
    b) The original data is collected from a single state of Maine, US.
    c) The data is collected only during the blooming season.
    d) We do not have enough domain knowledge to understand why we have limited values.

<u>Thank - You.</u>