Data Exploration(Anamoly Detection) of Barley Yield for year 1927-1936

Rahul Raghatate

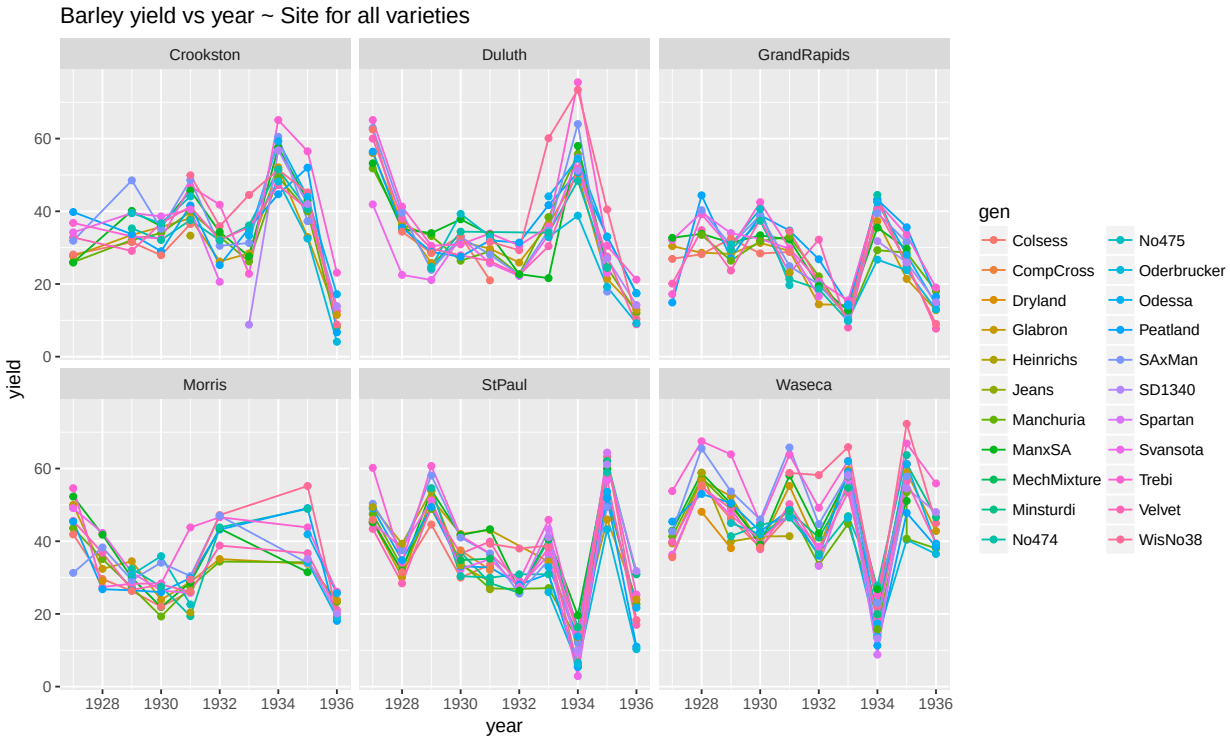
October 13, 2017

Data Import

Lets import the data and have a look at summary of it.

```
##      yield             gen           year             site
##  Min.   : 2.90   Manchuria: 58   Min.   :1927   Crookston  : 99
##  1st Qu.:26.80   Peatland : 57   1st Qu.:1929   Duluth     :107
##  Median :34.40   Trebi    : 57   Median :1932   GrandRapids:105
##  Mean   :35.63   Velvet   : 57   Mean   :1932   Morris     : 84
##  3rd Qu.:44.45   Glabron  : 56   3rd Qu.:1934   StPaul     :127
##  Max.   :75.50   ManxSA   : 51   Max.   :1936   Waseca     :125
##                  (Other)  :311
```

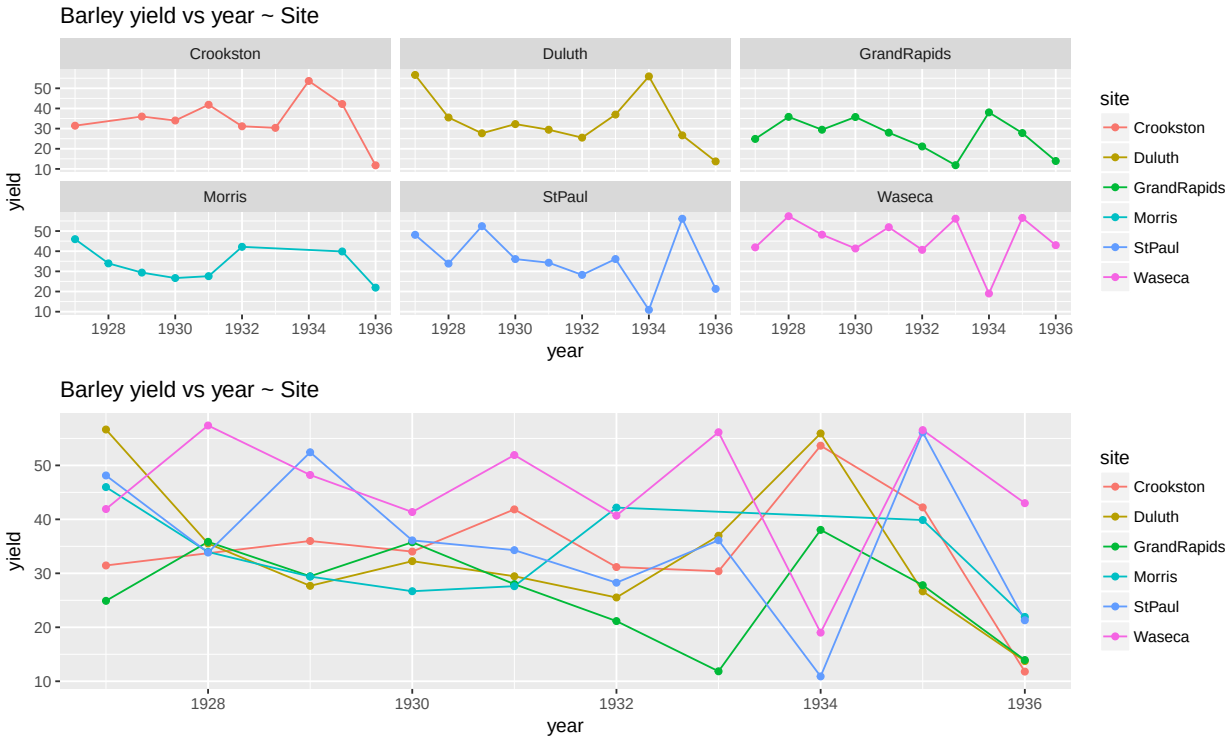
Plots for barley yields varied by gen(variety) and year at each site



There is no pattern in above plot. Sites StPaul, Duluth and are showing some extreme skewness.There has been a decrease of yield during 1935-36. The pattern remains irregular.

Pattern Exploration for yield at diff sites and time

Consider the mean yields (average over all the varieties) in each year at each site.

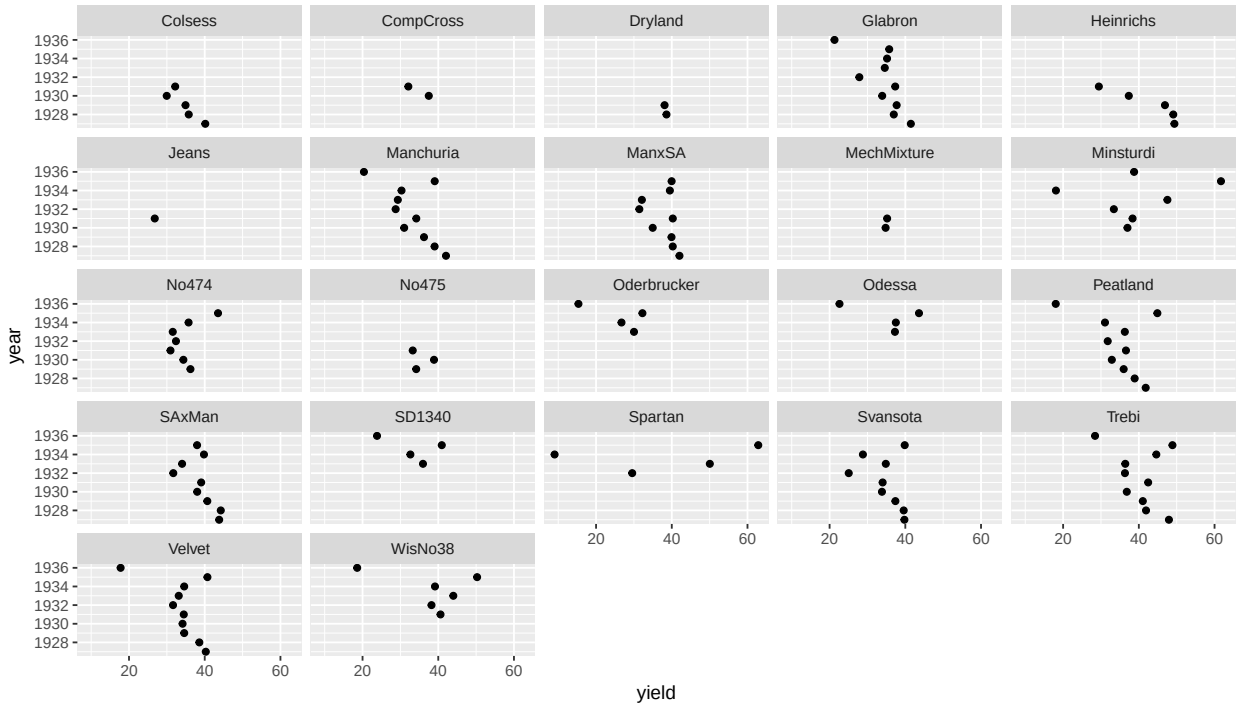


Looking at the graph, it is difficult to comment on the pattern of barley yield per year with respect to the location. On an average compared to the yield for any location during 1927, there has been a decrease of yield during 1935-36. The pattern remains irregular, it is more common for the yields to increase at some locations and decrease at others.

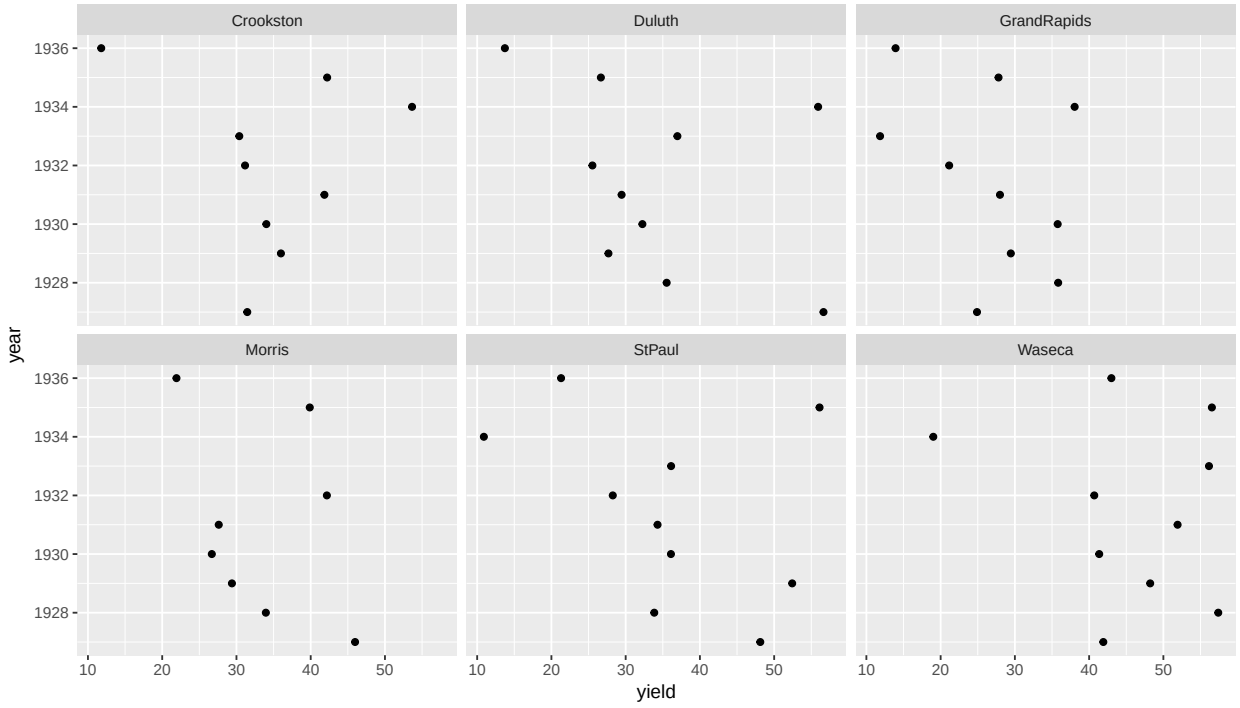
Interaction among the variables

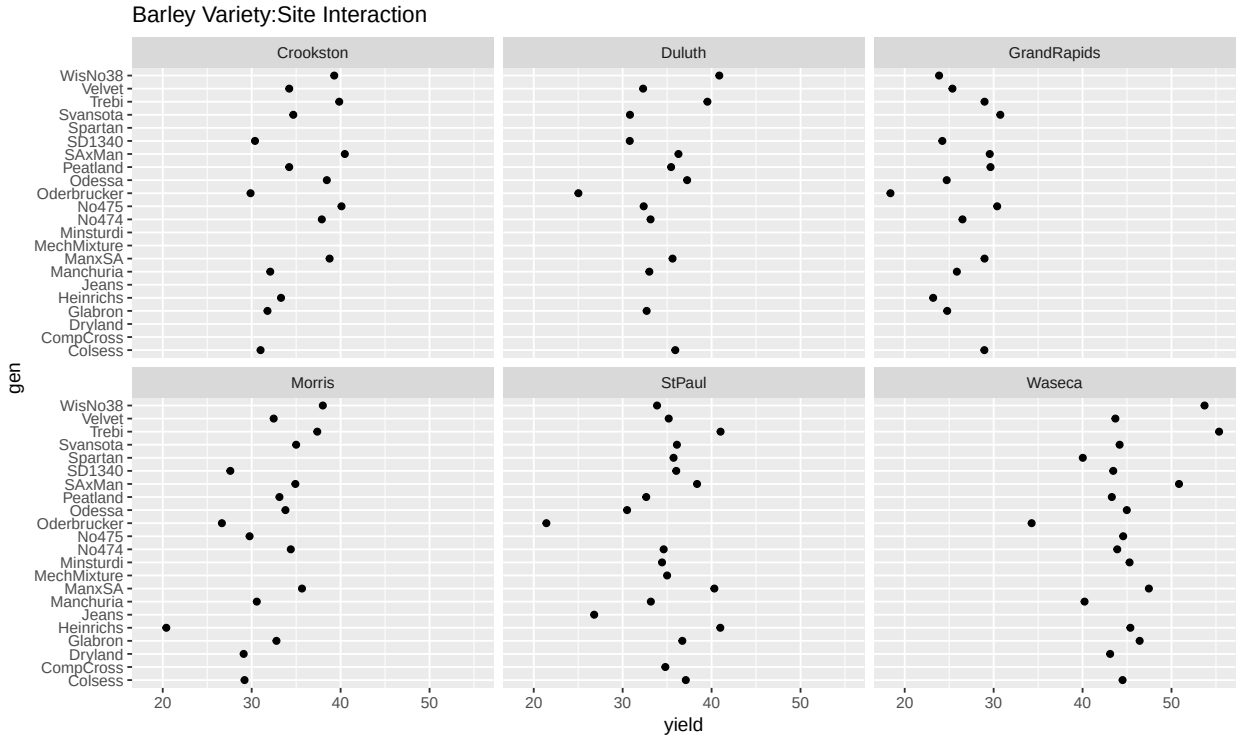
let's plot the graph of their interactions and check which is relatively highly scattered i.e. if the year and variety variation here is small (i.e. in each panel dots are scattered close to vertical line), then perhaps we can do without such interactions.

Barley Year:Variety Interaction



Barley Year:Site Interaction





We can observe that the year:site interaction graph shows the dots to be reasonably spaced out, thus indicating interaction required to be considered.

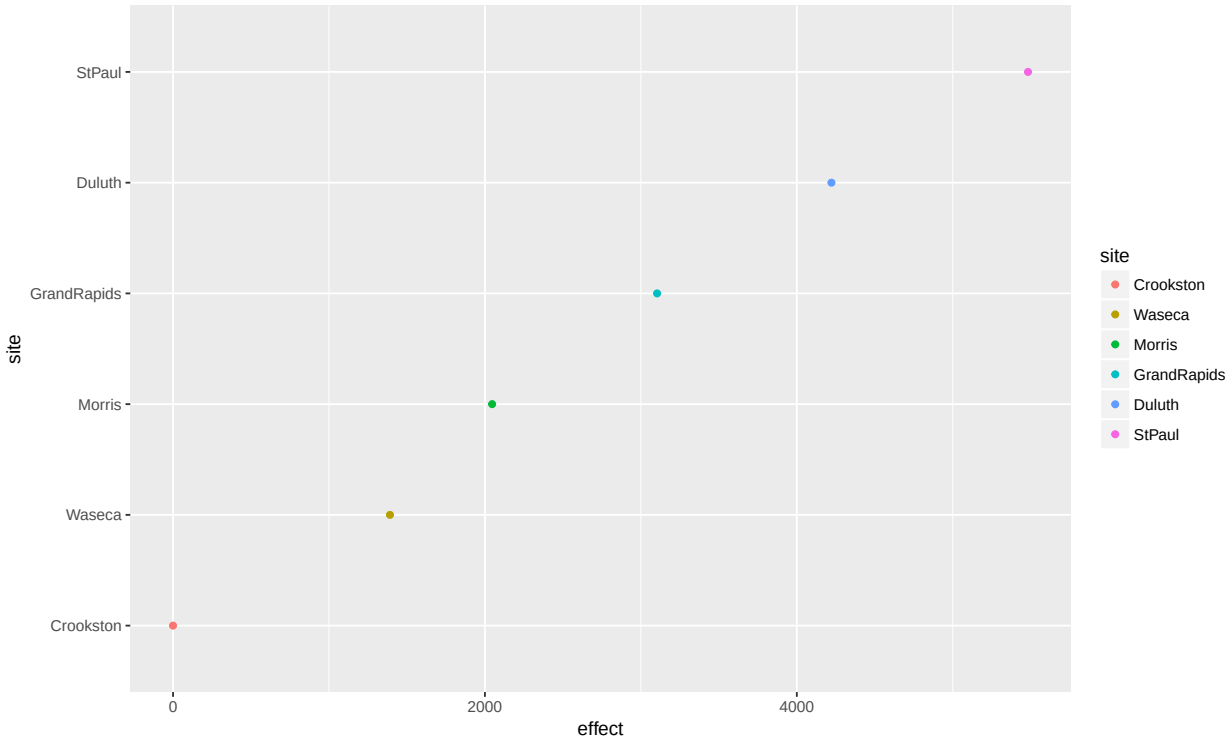
Model Selection

As the data consist of outliers and considering interaction also, least squares will be potentially misleading, I choose an outlier-resistant alternative i.e rlm() with bisquare to minimize the impact.

Model Implementation

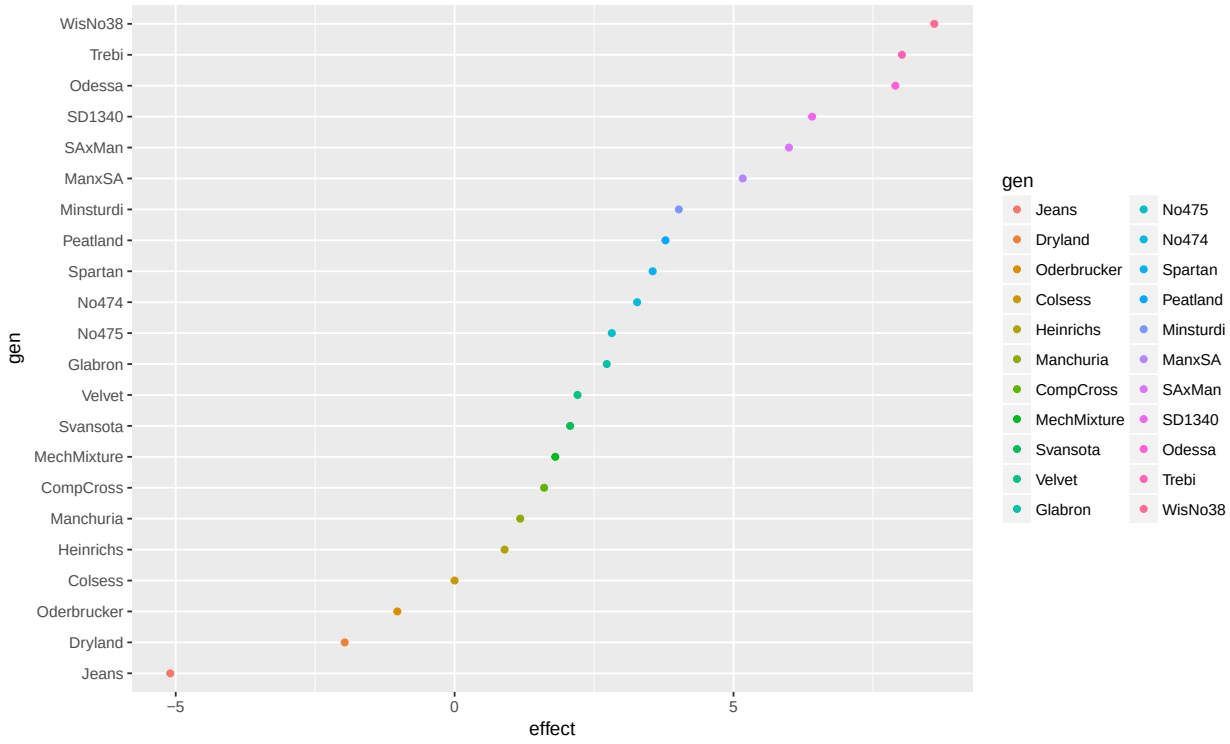
Applying RLM Model,with the goal of determining whether Morris 1931-1932 is an anomaly.

Plot to show the site effects.

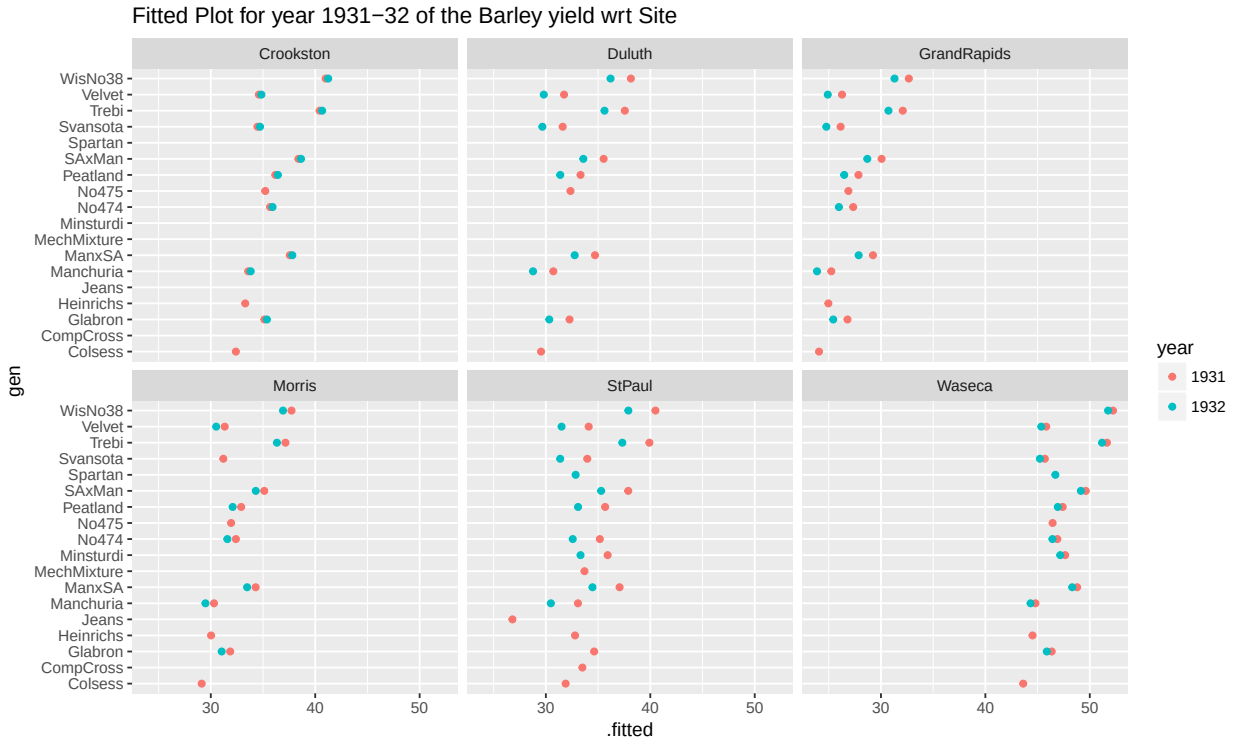


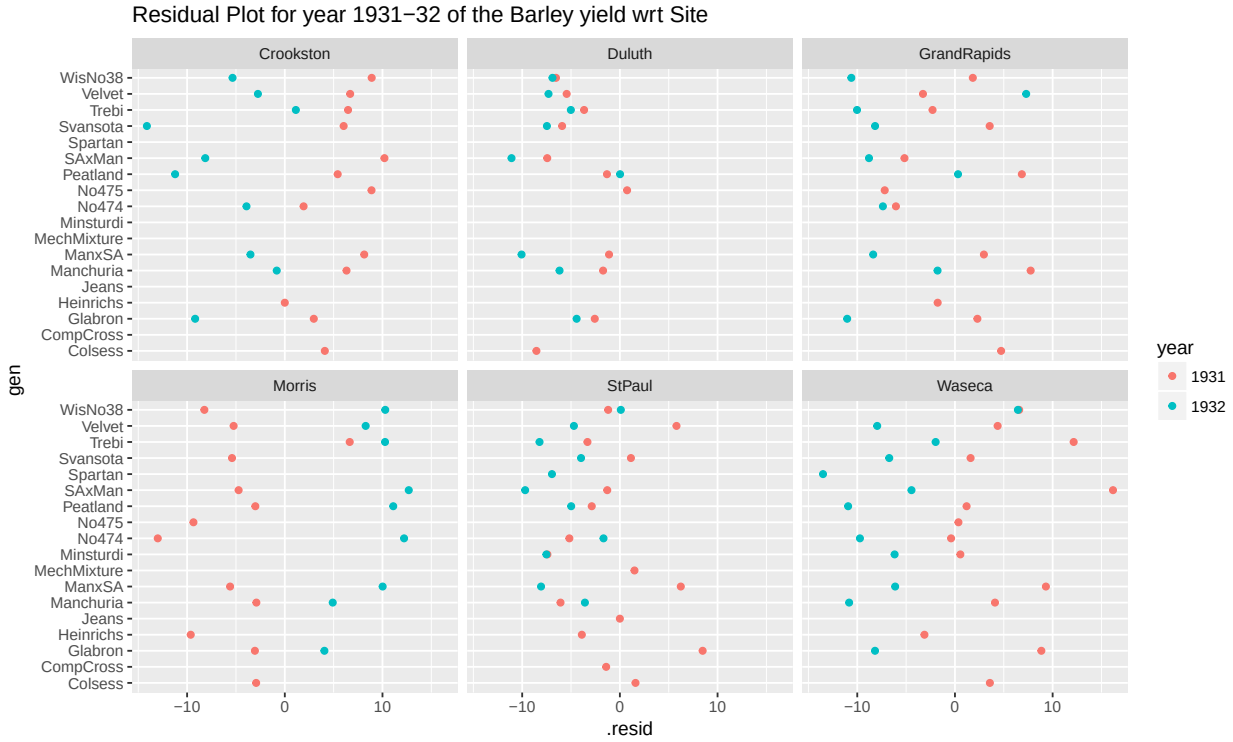
As we can see StPaul has the highest site effect where Crookston has the lowest effect. The other four approximately equally spaced from each other.

Plot to show the variety (gen) effects.



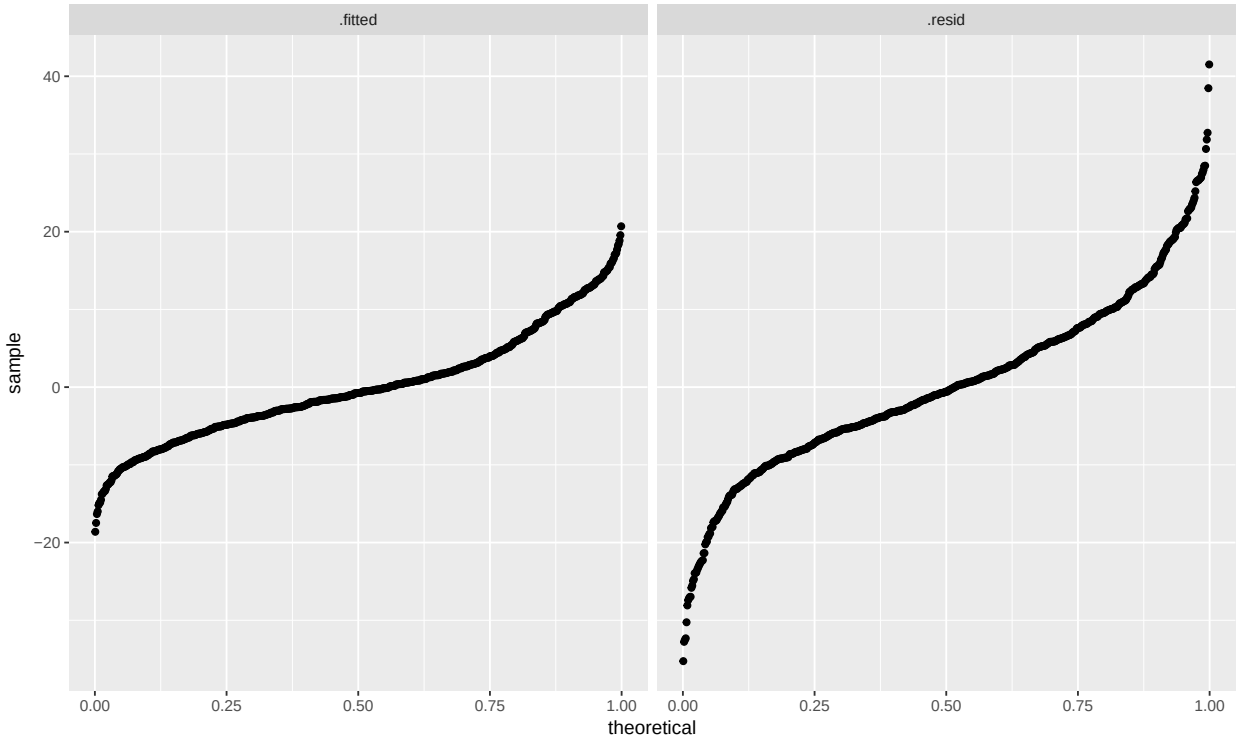
Residual and Fitted Values plot for anamoly detection





After proceeding with the RLM to model the data, it is evident from the residual plot that Morris during the year 1931-32 appears as an anomaly. The plot indicates a general trend in the residuals across all sites that

1931 appears to be on the positive end and 1932 appears to be on the negative end. But, Morris seems to defy the trend. Thus exhibiting an anomaly.



- Thus from above plots it is evident that Morris fails to follow the pattern of Barley yield that is exhibited across all sites.
- But, this should not be treated as an anomaly. Morris could have been under the influence of some sort of famine during 1931, to address the low barley yield issue.
- The yield vs year ~ Site from Question:1 clearly states that there was no pattern observed over the years in the Barley yield across sites.
- The ~650 Observations are convincing to believe that the Morris 1931-32 data should not be considered as a mistake.

Thus, this should be considered as a natural variation.