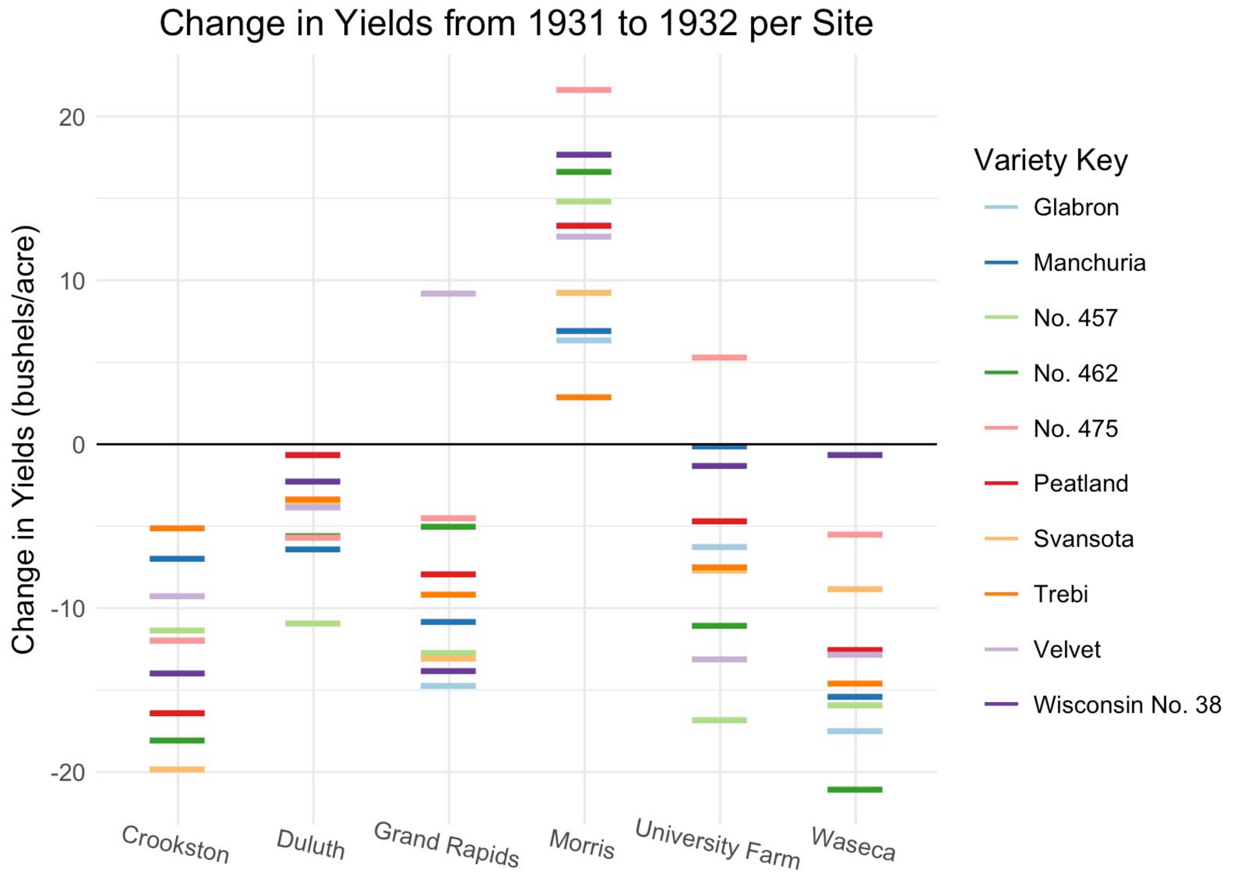
Noam Habot
CS 448B – Data Visualization
Assignment 1
2 October, 2017

An Effective Visualization of the Barley Dataset



<u>Short description of tools</u>: To generate this plot, I utilized *R*. Before plotting the data, I augmented it with a package called *dplyr* and then explored the data to find trends using various statistical tests and some of *R*'s summarization tools. Once I had the idea of the story that I wanted to portray in mind, I generated the plot and customized it using *ggplot2*. Furthermore, to ensure the colors would be distinct enough for the audience to tell the difference, I used a packaged called *RColorBrewer* with a palette called *Paired* such that the colors would be distinguishable, as opposed to a color gradient or rainbow.

I began this assignment by first analyzing the dataset to see which interesting trends I could find within the data. From the readings and from lectures, I learned that a data visualization ought not to be merely a display of the data at hand, but instead, a descriptive story that conveys new meaning behind the most important take-away of the dataset. After observing that for each pair of Site and Variety, there were exactly two rows of data (one for the yield in 1931 and one for the yield in 1932), I decided to meld the dataset such that instead of 120 total rows, there were now 60 rows with the features being: Site, Variety, Yield1931, and Yield1932. I then ran hypothesis tests to see if the means of the yields between 1931 and 1932 were different for each Site and for each Variety using *t.test*() in *R*. Right there, I learned a key fact: the means of the yields between the two years were not statistically significant when separating the groups by Variety, but were indeed statistically significant when separating them by Site.

This key observation led me to create a new column in the dataset for each Site and Variety pair – the difference between the yields in 1931 and 1932. Creating this variable allowed me to reduce the dimensionality of the variables that I believed were important to include in the graph from 4 (Site, Variety, Yield1931, Yield1932) to 3 (Site, Variety, Yield1932-Yield1931). I also noticed that for each Site, the differences seemed to cluster together amongst the various Varieties. This proved to be helpful in creating the layout of the final graph: the x-axis being the categorical variable Sites (sorted alphabetically), and the y-axis being the differences in the yields from 1931 to 1932.

The next decision was what to do regarding the separate Varieties within each Site. One option was to reduce the dimensionality of the visualization by not including the variation of Varieties within each site. I could have done this by taking the mean of the yield difference for each site and plotting that single point per site. Instead, I decided to keep the Varieties there and assigned “tick marks” to each of the observations, thereby having 60 total tick marks distributed between 6 different Sites. I used color to denote the 10 separate Varieties using a special color palette from an *R* package called *RColorBrewer*, which has predefined color palettes that are meant to be more easily distinguishable. Also, instead of using circles as points, I used tick marks to show more granularity such that nearby points won’t mask each other. The story I wanted to convey through this visualization had the following main components:

- Most sites except for “Morris” had reduced yields in 1932 than 1931, for almost all varieties. This is easier to tell using the horizontal line at y=0.
- “Morris” had all of its varieties yield more bushels/acre in 1932 than in 1931.
- The difference in yields from 1931 to 1932 within each site were roughly clustered together.
- The variance within each site is fairly consistent throughout all sites, and thus it is fairly simple to eyeball the mean yield difference within each site.

There is also an aspect of the data that is masked by this visualization – the baseline yields for each year. This visualization purely displays relative comparisons, and not absolute values of yields. However, this piece of information is not relevant to the story and the intent of the visualization, which is to convey differences in yields between the two years amongst various sites in the most intuitive way possible.