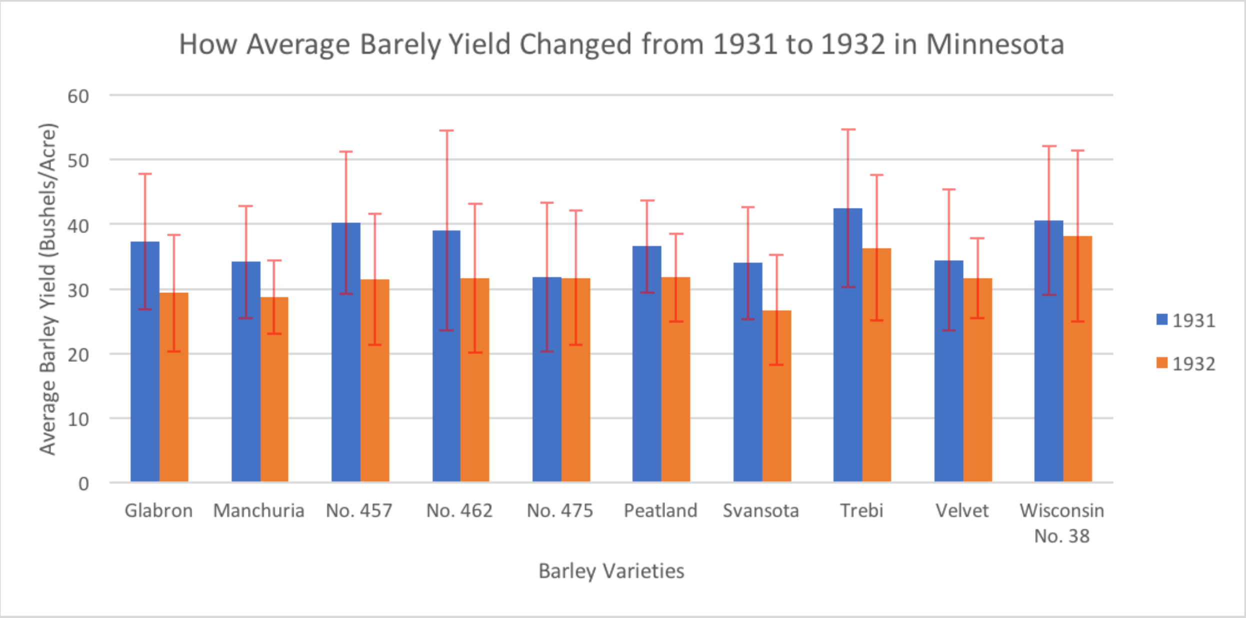
How Average Barely Yield Changed from 1931 to 1932 in Minnesota
Average Barley Yield (Bushels/Acre)
60
50
40
30
20
10
0
Glabron
Manchuria
No. 457
No. 462
No. 475
Peatland
Svansota
Trebi
Velvet
Wisconsin No. 38
Barley Varieties
1931
1932

The first thing I did to determine how to visualize this data was to figure out the meaning of each variable in the dataset. This dataset included the following: Site, Variety, Yield, and Year. The site at first glance is just nominal (N), but because locations can be interpreted as longitude/latitude points, it potentially has meaning to interpret them as quantitative (Q) values. I did not end up using this interpretation after playing around with the data. There is no special ordering or distance metric between varieties of wheat, so this must be encoded as nominal (N). Yield is quantitative, because it is meaningful to compare different values of yield. Year can also be quantitative, but since this dataset only contains two years, one could use Nominal.

After labeling each variable with the correct data type, I made the following observations. First, I have four variables (two of which are Nominal and have relatively high cardinality), and it would be much easier to only visualize two or three. The next step was which deciding which variables to combine into a smaller number of variables. This step involved me deciding on what the data was trying to show and to understand the rational of the experiment. My guess was that the important variable was the yield. This is due to knowing of the great depression and the dust bowl in the Midwest during the 1930s, and the need to grow more sustainable crops. The variable(s) under test seem to be the site and/or the variety. An educated guess was that they were testing which variety of barley would produce highest yield and for more accurate data, decided to test this at different locations. Using this insight, I decided to combine the yield results for each site while maintaining the variety variable. A natural way to combine these variables was to take the average and standard deviation of the results. Once I did this, I did notice an explicit trend of the yield for every single variety went down from 1931 to 1932, with some going down more than others. I decided to make this the main takeaway from my visualization.

Now that I know what data I wanted to show, the next step was materializing the visualization and choosing visual encodings for each of the variables. My variables were now Avg Yield/Standard Deviation, Variety, and year. I started with Yield and chose it to be a position-based encoding on the Y-axis of my chart. I chose this because yield is a quantitative variable and is the measured quantity which is natural for the Y-axis. Next, since Variety has cardinality of 11, I decided to encode this nominal value as positional on the x-axis. Any other choice (size, value, texture, color to some extent, orientation, and shape) would have been confusing and difficult to interpret due to the need to differentiate between 11 different types. For the Year, I chose to double encode this variable with both Color and with position (relative x-axis). I chose color since only two values could be interpreted as Nominal, it is easy for a user to just compare different varieties in a single year by looking at the color. As for the relative x-axis, I show how each variety's yield changes from 1931 to 1932 by putting the year next to each other on each axis. I think of this as projecting the year from the z-axis onto the x axis. Since I have two values (average and stddev) related to the yield, I decided to use a bar chart with an error bar since this is interpretable as an average and stddev.

The story I am trying to tell is the following. Yield for barley heavily depends on the variety; Changes in yield from 1931 to 1932 in general decrease, but change differently depending on the variety of barley. The error bars allow me to indicate to the reader that there is some variability in the yield data due to location differences and to not take the averages as absolute truths. I think that these decisions facilitate effective communication, because I am encoding each variable appropriately per its data-type, communicating trends in data, while still maintaining a notion of inexactness in the measurements.

I used mostly excel for this assignment. In addition, while attempting to understand the data, I used google maps to give me latitude and longitude for each Minnesota city in the dataset. This did not make it in the final visualization.