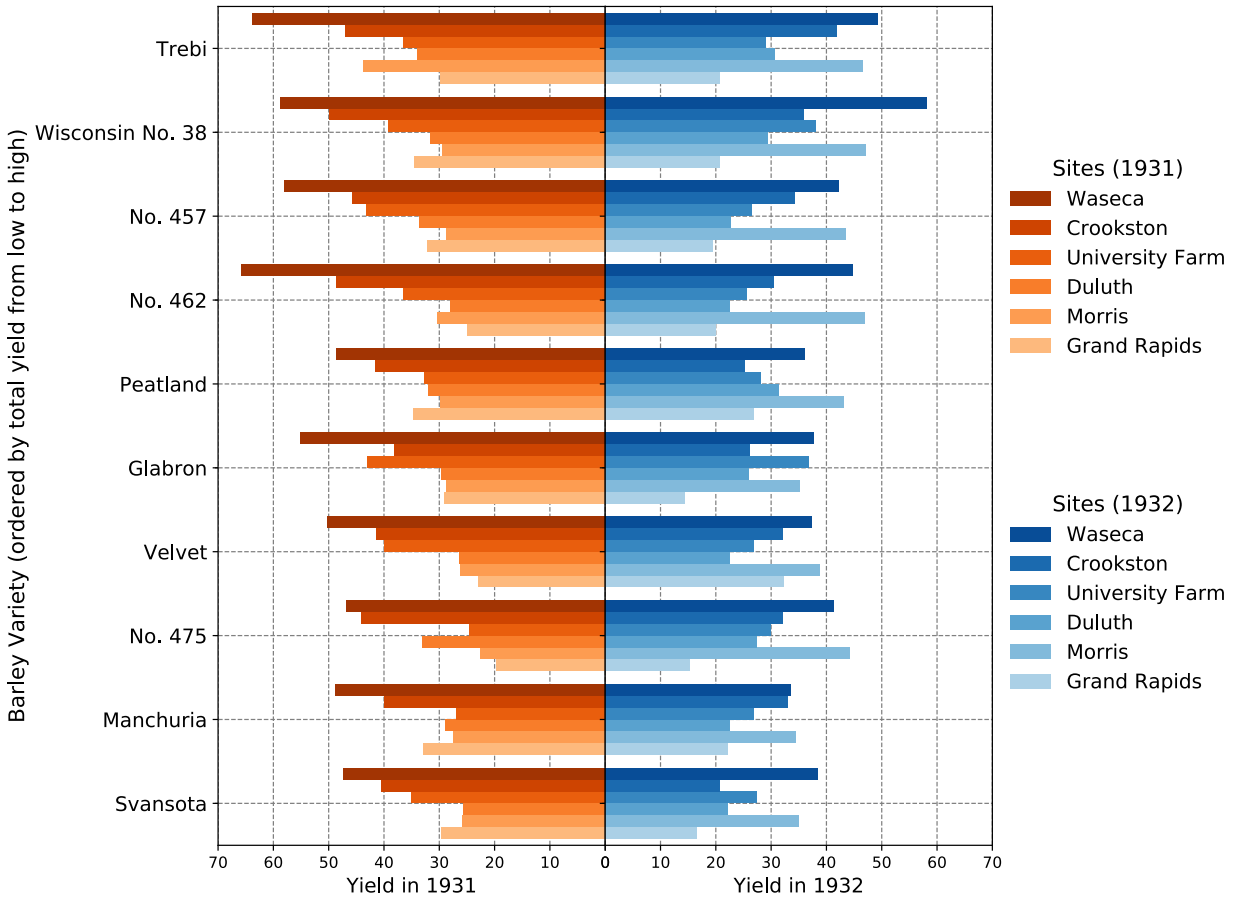
Trebi
Wisconsin No. 38
No. 457
No. 462
Peatland
Glabron
Velvet
No. 475
Manchuria
Svansota
Barley Variety (ordered by total yield from low to high)
70 60 50 40 30 20 10 0 10 20 30 40 50 60 70
Yield in 1931
Yield in 1932
Sites (1931)
Waseca
Crookston
University Farm
Duluth
Morris
Grand Rapids
Sites (1932)
Waseca
Crookston
University Farm
Duluth
Morris
Grand Rapids

Rationale: the aspect of the data set I wanted to high light is the discrepancy between yield across different sites for each type of variety, and back-to-back comparisons between the yields in 1931 and 1932. The story I want to focus on is to impose ordering of the yield from high to low for total yield across all sites across both years for each site and for each barley type.

Design: due to this rationale, I prioritized the variables as follows: (1) Yield, (2) Variety, (3) Year (4) Site. Further, I prioritized the ordering of the yields per variety per size over total yield per variety. Based on the priorities, my x-axis was first set to encode the yield, but I also used horizontal bar plots and the left-right orientation to display year-to-year discrepancies. This choice could allow me to reserve other visual variables for the remaining data variables. Next, I set the y-axis to be the Variety. The ticks were ordered based on the total yield over two years over all sites. Hence, the total mass of the grouped bars should go from low to high from bottom to top. Finally, for each type, I creates groups of bar plots that would indicate the Sites, where I used colors to encode the Site as well as the Year. The ordering of the total yield per site is encoded in the shade (with darker indicating higher yield), and I made an arbitrary decision to use orange and blue for the two colors (accounting for color blindness). Overall, one could follow the visualization to understand the individual Varieties and Sites as well as the ordering based on the tick marks and the figure legend.

Limitations: despite the design rationale above, there are aspects of the data that are down-played as a trade-off. For instance, the back-to-back comparison by comparing the length of a left-oriented bar to a right-oriented bar is as straightforward as putting them in the same orientation in a grouped fashion. One would need to use a sense of symmetry to see the discrepancy between yields. In addition, the comparison between total yields per site across all varieties per Site is not easy to see, because one would need to add up the bars of the same color and then compare.

Implementation: the data was analyzed and transformed in python (edited in a Jupyter notebook) using pandas and matplotlib. The sorting operations and grouping were performed using the built-in functions in the pandas package. The horizontal bar plots are a default option for plotting pandas data frames, and customization was made for flipping axes for one subplot and positioning the ticks and legends. The colors were extracted from the colormaps “Blues” and “Oranges” in matplotlib, and I left out shades that were too light with the given background.