CS 448 Data Visualization Homework 1

October 2, 2017

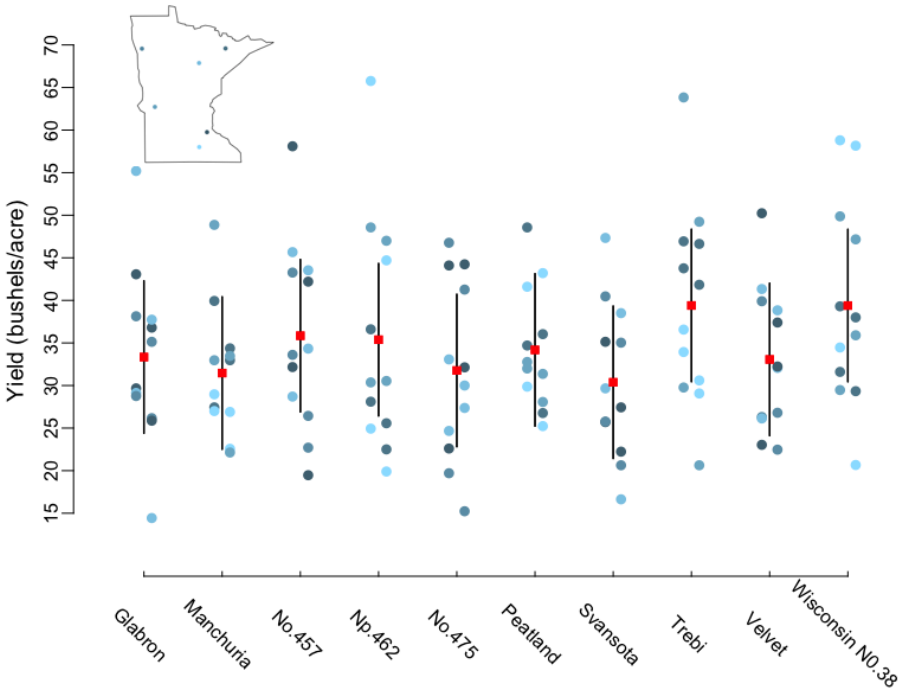

Figure Description x-axis shows variety, y-axis shows yield, each point represents yield in one year at one station, location of each station are shown in map in upper left corner. Station with higher average yield is represented by darker color. For each variety, year 1931 is plotted slightly on the left and 1932 on the right. Red points are estimated yield for each variety using model described below, arrows are a 95% confidence interval.

The model fitted is

$$y_{ijk} = \mu_i + \alpha_j + \beta_k + (\alpha\beta)_{jk},$$

for $y_{i,j,k}$ representing yield for variety i, year j and site k. Assume $\alpha_j \sim \mathrm{N}(0, \sigma_1^2)$,$\beta_k \sim \mathrm{N}(0, \sigma_2^2)$, and $(\alpha\beta)_{j,k} \sim \mathrm{N}(0, \sigma_3^2)$ all independent.

Result From the figure, there does not seem to be a particular variety that yields more, or a particular station that has higher yield in the two years. Year 1932 has less yield than 1931.

Graphical choice I want to answer the following question from data:

Is there a variety that has particularly high yield on average?

In the mean time, I also want to explore if there is a site that has high yield in general; if there is difference in yield between year 1931 and year 1932; if there is a site particularly good for a specific variety.

1. x-axis shows variety of each plant, as we want to compare varieties.
2. Axis. Shows relationship I am interested in. Varieties are order in alphabet.
3. Color. Use color to indicate station, order color according to yield at a station. A map is produced in top left to show position of each station so that one can explore connection between geography and yield. All colors are shown in light saturation. Notice it is harder to distinguish sites from color.
4. Statistics. There's a lot of noise in the data, use some statistics to combine information. Include a interaction term because there is a station that yields more in 1932 whereas most other stations yield less.

Technical detail

- Color. I used R to generate this graph. Color is set from ordering of average yield over two years and all varieties,

```
min<-0.4
col<-hsv(h=0.55,s=0.4,v=(1-min)*(order)/6+min)
```

- Model.

```
g<-lmer(Yield~Variety+(1|Site)+(1|Year)+(1|Site:Year)-1,barley)
```