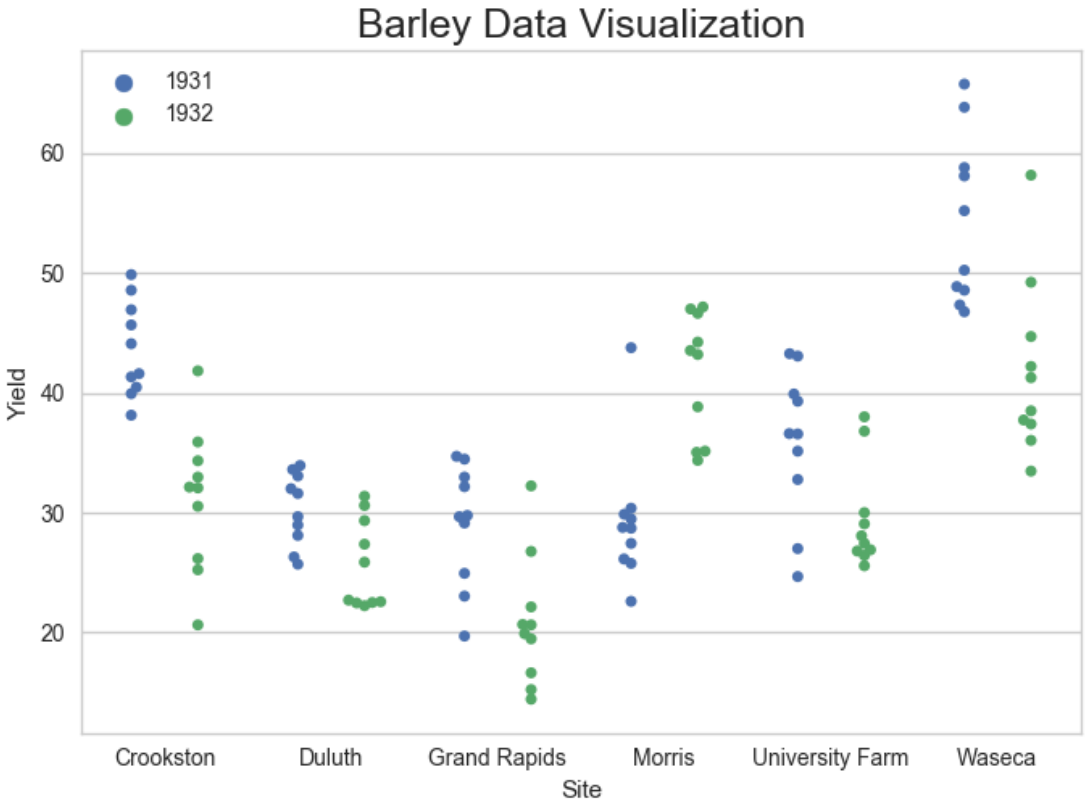
CS448B - Assignment 1

Vighnesh Sachidananda

Visualization



Analysis

Ranking Data Variables

In this assignment, I aimed at coming up with an automated way to rank data attributes. From the analysis of Bertin and Mackinlay, we are exposed to the idea that visual variables are not equal. Some are better for communicating data than others. The same is true for data variables, some attributes contain relationships of interest and others are more noisy (encoding less information).

Concretely, I came up with a ranking algorithm for data attributes that works by aiming to find variables that **manifest correlations**. This concept is abstract but mathematically we can use the notion of mutual information to understand how "correlated" or mutually dependent variables are.

For example for variables Y and X:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

We can compute the above for every pair of variables and find the pair of variables with the highest mutual information. This pair we will choose to allot the two position visual variables (x, y axis). Then we find the data variable that is most correlated with either of the two data variables assigned the position variables (site and yield in this case). This process can be continued until we have a complete ranked list of our data variables.

Feature Pair	Mutual Information
'Site', 'Yield'	0.456
'Year', 'Yield'	0.123
'Year', 'Site'	0.019
'Yield', 'Variety'	0.005
'Year', 'Variety'	0.003
'Site', 'Variety'	0.000

The ranked data variables were in the following order as a result:

1. Site
2. Yield
3. Year
4. Variety

Now that we have our ranked list of data variables, we can apply them to the 3 visual variables we are using for our visualization. Site and Yield were ranked 1st and 2nd by the algorithm and we give them the x and y axis as a result. Year is ranked 3rd and was given the hue visual variable. Variety was omitted from the visualization for simplicity. Note that in our mutual information table the pair 'Site' and 'Variety' have a mutual information of 0, meaning that knowing the Site gives no information about the Variety (which is true since all Sites grow the same Variety of barley).

Visualization Insights

From our visualization we can gain insights into the barley dataset:

- Firstly, we can see that indeed Site and Yield are correlated variables.
- Waseca, Crookston and Morris seem to have the highest yields.
- Grand Rapids and Duluth seem to have the lowest yields.
- For most sites, the yield in 1932 were less than or equal to the 1931 yields with the exception of the Morris site.
- In the Duluth and University Farm sites, the yield for the two years is closest and Morris, Crookston, and Waseca have the highest difference between the two years.

Tools Used

All code is written in python.

The following libraries (imports) were used:

```
import pandas as pd
import numpy as np
from sklearn import feature_selection
import scipy
import operator
import matplotlib.pyplot as plt
import seaborn as sns
```