Final Report

Lindsay Knupp

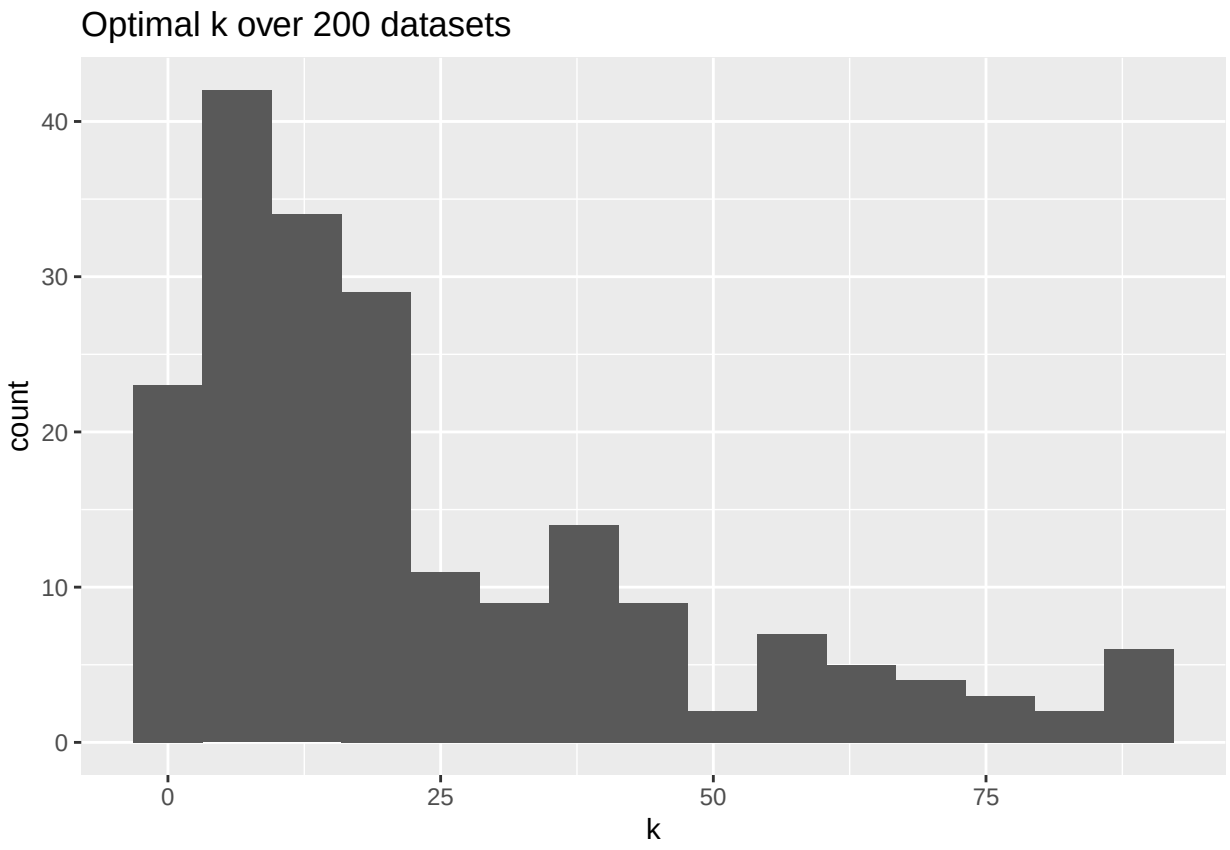
2024-05-14

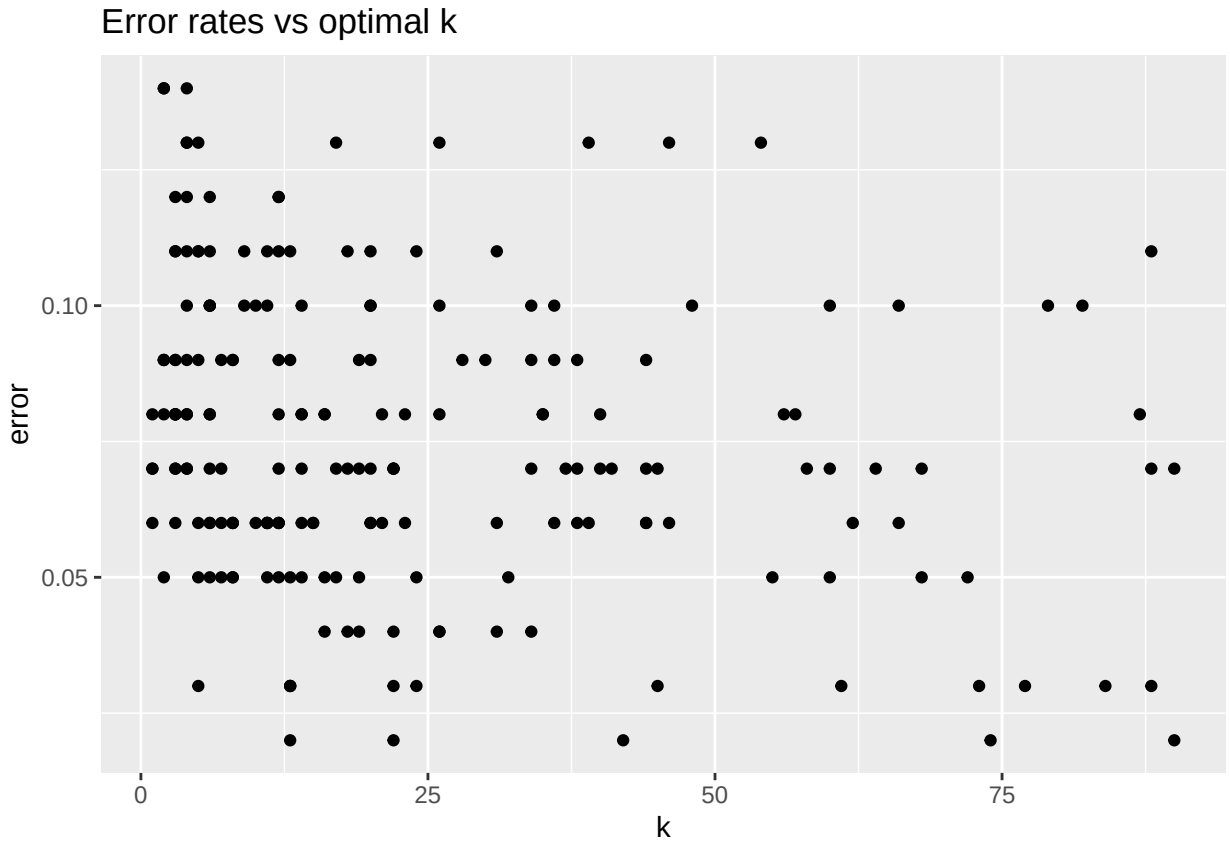
Simulation Study

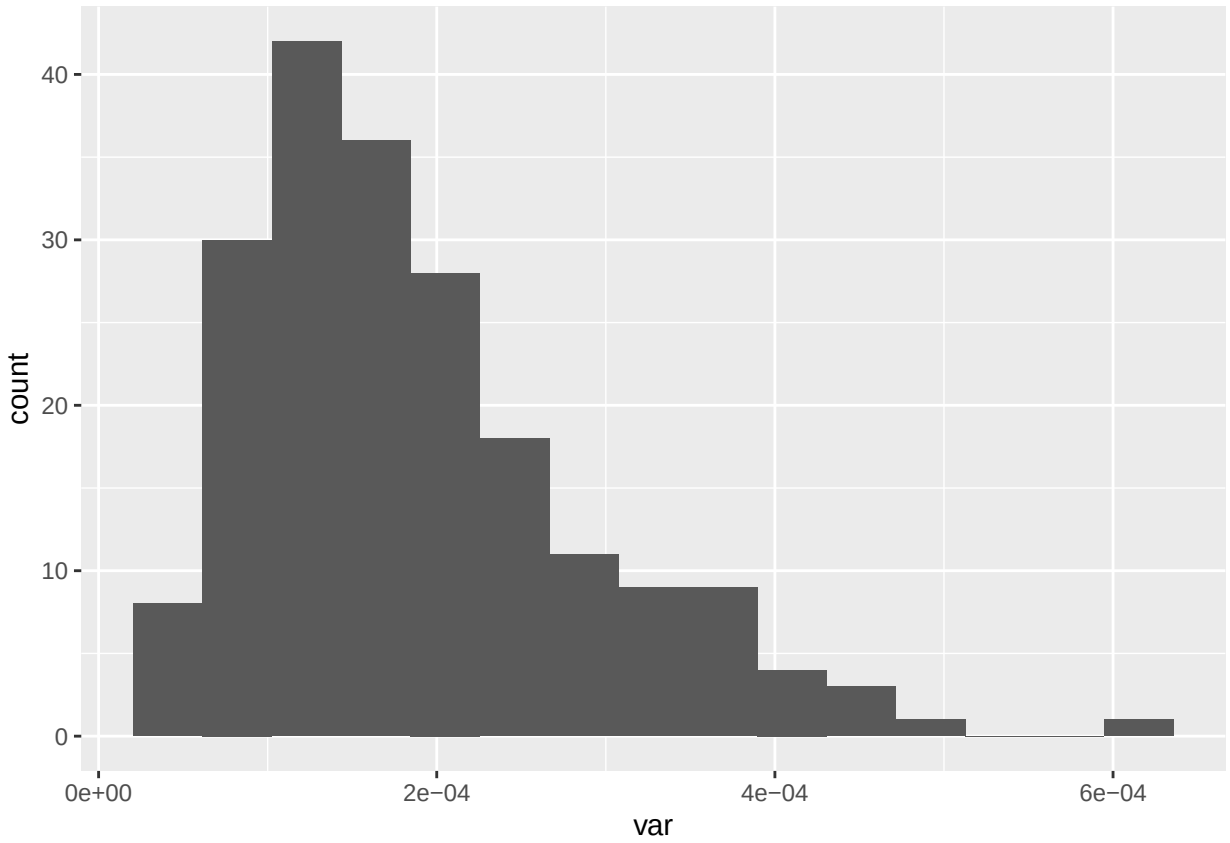
We were interested in understanding how our data affected the optimal choice of k in the k -nearest neighbors algorithm. We already experienced some variability when running our model through on different computers. Therefore, we wanted to see if more simulated datasets would produce the same variablility.

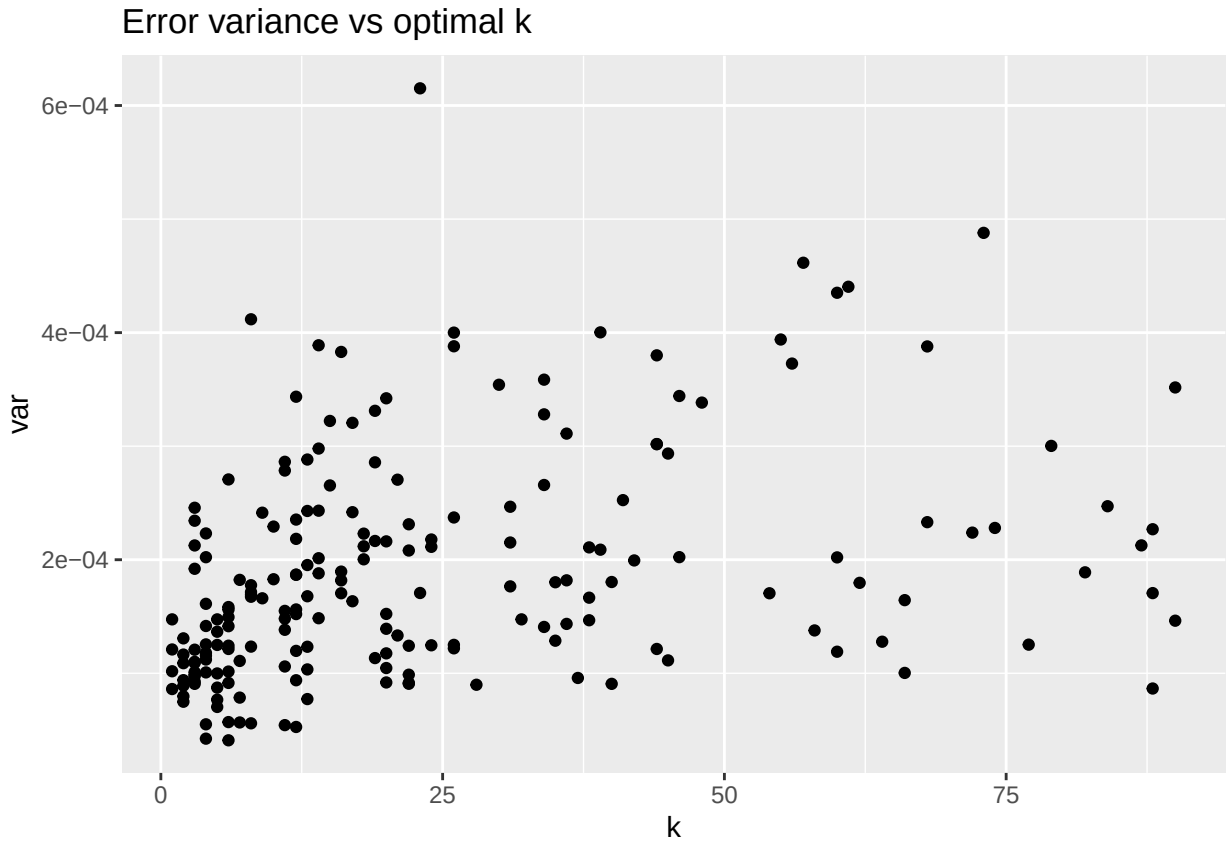
To replicate our 13 predictors and 2 response variables, we used a package called `faux` to simulate our numerical predictors from a multivariate normal distribution.

We used a standard normal distribution to replicate our 13 predictors.




Error rates vs optimal k
error
0.10
0.05
0
25
50
75
k


40
30
20
10
0
count
0e+00
2e−04
4e−04
6e−04
var


Error variance vs optimal k
6e−04
4e−04
2e−04
var
0
25
50
75
k