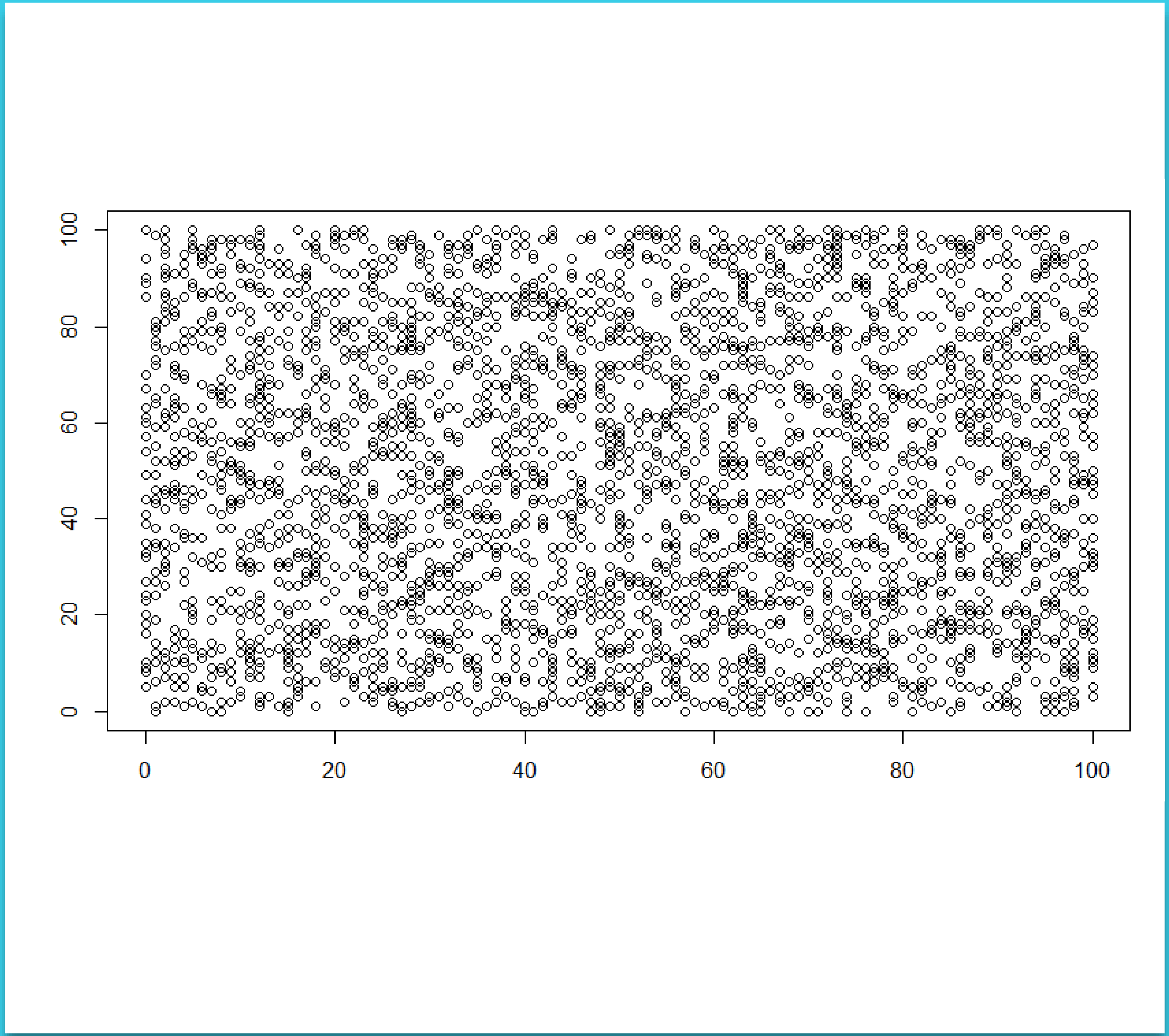
Mysterious Box Kaggle Challenge

Andrew Scovell

Homework 12

Looking at nonsense scatterplots

- I wanted to look at the relationships between inputs, so I plotted input 1 to input 2, 2 to 3, 3 to 4, 1 to 4, 2 to 4, etc.
- I did this with each individual sound and each switch type
- They pretty much all looked like this
- I decided to throw as much information at rpart to see if it could give a coherent response instead



New columns for the data

- I threw as much as I could think of at rpart
- This included columns for whether the ID is even or odd and what the final digit of the ID is
- The sum, product, mean, and standard deviation of the inputs for each row
- The differences between inputs 2 and 1, 3 and 2 & 4 and 3
- And finally, whether each input is even or odd

Finding a CP to use

- To avoid overfitting, I played around with various CP values and a minsplit of 20
- In the end, I decided to risk it with a CP of 0.00045
- It seemed low enough to grab at some of the new columns that I had added, but would probably still overfit some data
- It also never seemed able to tell when the sound would be a beep
- Cross Validation gave an error of 0.3109, an average accuracy subset of 0.6803, and an average accuracy for all of 0.6580
- This lined up with what the 10% Kaggle leaderboards were showing for what a lot of people had submitted, so I submitted mine fairly confident to have a good score