Stat 503X
EXAM 1

Name:

Student Number:

There are 5 equally weighted questions, and you have 50 minutes to complete the exam. You should spend no more than 10 minutes on each question. This exam is also closed notes. You should only need a pen or pencil to complete it.

1. (10pts) Explain the method called “jittering”, and how it might be used for exploratory data analysis.
2. *(10pts)* What visual cues would you use to detect cluster structure if you were using a grand tour on high-dimensional data set?

3. *(10pts)* Which of these are valid interpretations of the mosaic plot of the tips data on the next page? Circle agree or disagree with the interpretation.

- More males pay the bill than females. (Agree Disagree)
- Relatively higher numbers of males pay the bill on Fridays, otherwise the numbers are similar on other days. (Agree Disagree)
- The distribution of female bill payers is different to males over days. (Agree Disagree)
- Sex and day of the week are independent variables. (Agree Disagree)
- Amongst males, relatively more smoking parties on Thursday and Friday with numbers dropping off to Sunday (Agree Disagree)
- In smoking parties where a female has paid the bill, there is a similar pattern to males from day to day, except that Friday has the larger number. (Agree Disagree)
- Thursday is a big day for males during the day, but there are almost no males on Friday. (Agree Disagree)
- The restaurant is closed Saturday and Sunday during the day. (Agree Disagree)
- At night there is a predominance of smokers. (Agree Disagree)
- No bill for a party of size 6 was paid by a female. (Agree Disagree)
4. (10pts) This data was collected on two species (Orange and Blue), both males and females of crabs in Australia. There are five measured variables: Frontal Lobe, Rear Width, Claw Length, Carapace Width, Body Depth. From the figures below, the primary question is to work out which variables contribute to the separation of species and sex. In the grand tour view all the variables were standardized first, and then projection pursuit was used to get this view.

(a) (3) Describe what you learn about the data from the grand tour plot.

(b) (3) Describe what you learn about the data from the dotplots and pairwise scatterplots combined.

(c) (4) Which variables contribute to the separation of species and sex?
5. (10pts) This data was collected on a field managed by Iowa State University. The primary question is how corn yield is related to the soil nutrients:

- **B** Boron (parts per million)
- **Ca** Calcium (parts per million)
- **Cu** Copper (parts per million)
- **Fe** Iron (parts per million)
- **K** Potassium (parts per million)
- **Mg** Magnesium (parts per million)
- **Mn** Manganese (parts per million)
- **Na** Sodium (parts per million)
- **P** Phosphorus (parts per million)
- **Zn** Zinc (parts per million)

In particular, how can we get higher values of the yield by controlling the values of the soil nutrients.

(a) (6) Briefly describe how you would approach to answering the primary question. (Like the Suggested approaches part of the assignments, give methods, and what parts of the question they might address.)
(b) (4) From the Figure below, describe the relationship between Yield and Boron. (High, medium and low values of Yield are brushed differently, but this is irrelevant for answering the question.) Explain why simple linear regression would have missed this important relationship.