CSCI-420 – HW01
Due when the dropbox closes (see MyCourses)
(10 points total)
Prof. T. Kinsman

Look at the handout for the first quiz data and answer the following questions. As you answer the questions, realize that the data you see was put through a PERL (Practical Extraction and Reporting Language) parser and pre-processed. Then it was read into Excel for inspection.

The data was rendered anonymous by:

- To protect names and identities, the user-ids were run through a simple secure hashing algorithm to give each student a unique user id. We can then refer to each student using this UID.
- The student names were deleted or changed to random names.
- The data in each column was randomized, so the columns are not related to each other. The only two columns that were kept correlated were expected grade and weeknight homework should be due.
- I did some preliminary data cleaning to remove egregious mistakes.

Data Cleaning and Inspection:
Hang onto this data, we will do statistics on it later. There are patterns in the data. You will learn to find them.

The first problem you will have is reading in the data. It is best to read it into a spreadsheet so you can sort it by various columns. For example, for question 2 it is helpful to sort the data by the color.

Questions are generally worth a ½ point each unless otherwise stated. The math is not exact.

1. (3/2 pts) For each question (1 to N) is the recorded variable **N**ominal, **O**rdinal, **I**nterval, **R**atio, or **NA (***Not applicable)*? Make a vertical list numbered 1 to 24, and list your answer, with a sentence fragment explaining why you decided this.

2. (1 pt) Considering the namespace colors:
   a) Are all 12 namespace colors represented in the responses?
   b) Which was the most commonly selected namespace color?

3. Considering the question of left-handed – search online for the correct proportion of the population that is left vs. right-handed. Do you think that the results of our quiz reflect valid data?

(continued on next page.)

4. Considering the question about being polydactyl (having six fingers), what is this question testing for? Where any found, if so, which user-id(s)?

   An outlier is something that happens, and is valid data. Outliers are rate, but they do happen. You meet them walking around the street in your daily lives. An anomaly is something that is never expected to happen, and you are not prepared to deal with as a data analyst. For example, the CDC does not track diseases in anyone over age 85. They are poorly prepared, but these people become anomalies in the CDC data. They have to handle them on a case-by-case basis.

   Was this question testing for an outlier or an anomaly? Which do you think?

5. Considering the Star Wars Question, what kind of data cleaning should be done before testing to see if the user gave the correct answer? Should the machine test for case sensitive responses?

6. Considering Winnie-the-Pooh, there were bad responses. What processing would automatically help the analysis? Use your Computer Science skills here. How do you match a pattern? What pattern might you use?

7. What was the most popular team sport?

8. Considering the first question about the hours of sleep one gets, how could this be taken the wrong way? If the quiz was only taken once, and the person taking the quiz did not know what was coming, do you think the answers given correctly represented the ultimate intent of the question?

9. Considering the questions, about sleep, is there any missing data? Why do you think this happened?

10. Considering all of the sleep questions, do you think the order that the questions is asked matters?

11. Considering all of the questions about sleep, do you see any evidence that a single person is a mixture model?

12. I suspect that some students literally raced through the quiz. Such students are more likely to select the namespace color Orange, and cause incorrect data. I noticed this last semester. Can you find any evidence of racing? Why or why not?

13. Do any records contain attributes that are duplicate, but which should be unique. If so, suggest a deduplication method to correct it. (Deduplication is in the text book.)

14. Some surveys *require* the subject to provide a phone number. Sometimes 867-5309 comes up a lot. What message are these subjects trying to convey? How should the survey question be changed?

15. Considering the processing path that the data went through, why would someone report that they get "July 8th" worth of sleep a night? How might this have happened? (Honest, this really happened.)

16. By casually looking at the data, without doing any statistics, when should homework be due?

17. Are there any other data cleaning techniques in the book that I forgot to ask about?

18. Can you think of a good, intriguing, relevant or fun question we should ask next semester's class?