Tyler Paulsen
CSCI420-01
9/11/15

1) (3/2 pts) For each question (1 to N) is the recorded variable Nominal, Ordinal, Interval, Ratio, 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.

1. nominal - there is a clear choice between the data submitted - true or false.
2. nominal - same data as question 1. There is a clear choice between the data submitted.
3. nominal - was a radio box selection of the colors. Clear choice between data submitted.
4. nominal - there is a clear separation between what is selected, and the order of the answer does not matter to the question.
5. nominal - Many different answers were presented, has no order, no interval, and no absolute zero
6. nominal - clear choice between what data was submitted (6 different types of answers)
7. nominal - same type of data as question 6. Clear choice between data submitted.
8. nominal - almost all the same answer (34 of 34) clear choice between data.
9. nominal - only two choices of data available true/false via radio button selection.
10. nominal - four data choices given via radio buttons. Clear choice between data points.
11. nominal - can clearly separate based on type of sport given.
12. nominal - same type of question as 11. Can separate based on the type of sport
13. nominal - same type of question as 11 and 12. Can separate based on the instrument played
14. nominal - same type of question as 11,12, and 13. Can separate based on club type.
15. nominal - majority of data has the same pattern. The digits 1,2,3,4, and 5 are present in most of the data submitted.
16. ordinal - can order the grades from highest (A) to lowest (D) with I prefer not to answer as a special case.
17. ordinal - can order based on the number of classes selected.
18. ordinal - can order data based on how many hours on average a student sleeps.
19. ordinal - can order data based on how many hours a student typically sleeps .
20. ordinal - can order data based on median of the hours slept by the student.
21. nominal - data given can be clearly be split apart from other data points
22. ordinal - can order the data based on how many years the student has been at RIT
23. nominal - there is a clear answer to the question asked.
24. nominal - has a clear choice in what is selected for the weekday.

2) (1 pt) Considering the namespace colors:

A. Are all 12 namespace colors represented in the responses?
no, pink, brown, and white were all not selected from the namespace colors given.

B. Which was the most commonly selected namespace color?
blue was the most commonly selected.

Tyler Paulsen
CSCI420-01
9/11/15

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?

> Approximitly 10% of the population is left handed (https://en.m.wikipedia.org/wiki/Handedness). The data that we had ~68% right handers and ~31% left hander. I do not think the quiz reflected valid data, the question was asked in an odd way, and some people may have not answered correctly.

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)?

> This question was to test for outliers; it is possible for someone to have six fingers, so the answer is valid and should be handled.
>
> There was one UID that had six fingers: 2875

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?

> The data should all be reduced to lower case and trimmed of any whitespace.No, the machine should not 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?

> The data should be made all the same case. This should eliminate a lot of the sorting problem that excel has. Once this is done, the data can be looked at and counted correctly by the excel count algorithm. The other bad responses look to be either legitimate responses or just typos.

7) What was the most popular team sport?

> The most popular answer was: none. The most popular sport was a tie between: swimming, track and field, golf, and cross country.

Tyler Paulsen
CSCI420-01
9/11/15

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?

> This question could be answered as giving the average number of hours slept each night, when the question was asking for the mode. This question would not be answered correctly based on the intent of the question. The questions need to be more descriptive on the type of answer needed.

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

> There was no indication of only submitting a number response. Some of the data is given in sentence form explaining why the hours of sleep is the way it is -- this is not needed, and it makes it hard to parse out the correct answer from the response.

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

> Yes, I think they should be presented all at once rather than three separate questions. When the questions are asked separated like that, it is hard not to think that it is just the same question asked three separate ways unless it says explicitly that it wanted the mode, median or mean.

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

> In the data given, the answers that were valid all were within an hour or two from each other, so there was no mixture model for a given person. There was also only 3 possible answers that could happen, and this would be near to impossible to claim to be a mixture model with such little data.

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?

> The evidence i would use for racing would be to look at the questions that required and keyboard input. If the user had 2 or more outliers, I would consider them racing. UID 2442 gave 2 different answer that had an outlier. I would suspect this UID of racing.

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.)

> There are duplicate UIDs in the data -- 2512. A method to deduplicate the data would be to give all unique UIDs.

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?

> The subjects gave a phone number, it may not be the one the survey question asked for. The survey question should be changed to ask for the individuals number, not just a phone number.

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.)

> Auto correct could have corrected 8 to July 8th.

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

> Sunday night because it was picked the most just at a glance.

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

What to do with missing values. There was no missing values at a glance, and there was a question about missing values, but there was no method asked on how to solve it.

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

shoe size.