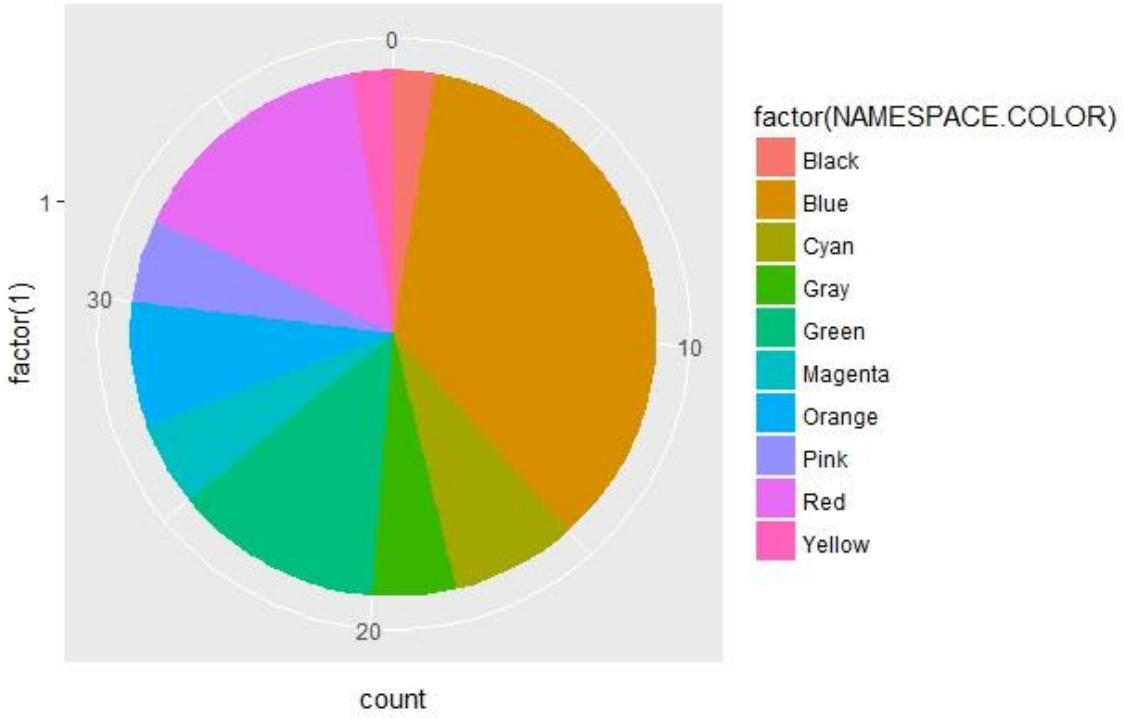
1. Considering the namespace colors, are all 12 namespace colors represented in the responses?
No, only 10 namespace colors were represented.

2. Build a 0R Rule for guessing someone's favorite namespace color. What is the 0R Rule? Use it to guess the Professor's favorite color.
0 R rule can be created for guessing someone's favorite color by considering majority. As seen in the following graph, 10 out of 39 users chose Blue as favorite color. Hence 0R rule gives Blue as Professor's favorite color. Following is the definition in R.
tail(names(sort(table(studentData$NAMESPACE.COLOR))), 1)



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? Why might the data be wrong?
Correct proportion of left-handed Vs right-handed is 1:9 i.e. 0.1111. In analysis of the quiz response data, left-handed percentage comes as 0.1142. So the results of the quiz do reflect valid data. However, some people might be ambidextrous which was not considered during the survey.

4. Considering the question about being polydactyl (having six fingers). I'm telling you that this is never supposed to happen. It would make it hard to use scissors. What is this question testing for? Why do you say this? Where any found, if so, which user-id(s)?
Wikipedia reports that 1 in 500 live births are polydactyl. Still there was 1 instance of polydactyly among 39 observations. ID was 799. May be it was testing for relation between typing speed and polydactyly. It can be observed that polydactyl has finished quiz in 9 minutes which is second shortest time. Also the polydactyl plays Clarinet which is easier to hold and play with more fingers.

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? Is there a pattern in the lower case responses?

Before testing if user gave correct answer for fiction question, data should be converted to lowercase or uppercase for comparing. It is observed that people who took more time to answer quiz have entered case sensitive responses. Of 25 people who spent more than 16 mins on quiz, 17 entered case sensitive responses. Which is 68%.

6. Considering the Winnie-the-Pooh question, 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? (CS students, please help students from other disciplines that ask for help.)
Considering the bad responses for the Winnie the pooh question, following processing will help to test the answer:

1. Getting rid of leading and trailing whitespaces in the response. E.g. expected "pooh" appeared " pooh "
2. Converting response to lowercase. E.g. expected pooh appeared Pooh
3. Finding if the response contains correct answer. E.g. expected pooh, appeared Winnie the pooh
4. Finding the Edit distance between expected and actual response and accepting it as correct response if distance was less. E.g. expected pooh, appeared poop. Levenshtein distance and Oops distance. Levenshtein distance is measure of number of insertions deletions or substitutions required to convert one word to other.

7. Prof. Kinsman invented a new Boolean attribute. This attribute is set to true (or 1) if the person was involved in any team sports or group activities (such as orchestra). What is the word for an attribute that is derived from other attributes?
Attribute that is derived from other attributes is called feature of the dimension or attribute.

8. Considering the first question about the hours of sleep one gets, could this be taken the wrong way? 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?
The user might not have monitored the sleeps he or she takes to answer this question correctly. Some people might not want to tell about how much hours of sleep they get so it might be considered unethical question. Also from the following responses, it is obvious that users did not follow question correctly and hence intent of question was not fulfilled
For example:
Average – 7, Mode – 6, Median – 8 – Mode is less than median and average
Average – 7, Mode – 6, Median – 6.5 – Average is greater than mode and median
Average – 7.35, Mode – 7, Median – 8 – Mode is less than average and median
Average – 8, Mode – 7, Median – 7 – Mode and median less than average
Average – 7, Mode – 6.5, Median – 7.25 – Mode is less than average and median
Many others have responded by giving equal answers for mode, median and average

9. Considering all of the sleep questions, how do you think the order that the questions are asked changes the results?
People who paid attention to question, after answering the question about average sleep were tied to answer median and mode questions accordingly and 19 out of 39 did that by entering same answer for Average, median and mode i.e. symmetrical distribution which were the only people who followed the empirical relation between mean, mode and median as specified by Karl Pearson:
Mode = 3 median - 2 mean

10. Considering all of the questions about sleep, do you see any evidence of a mixture model?
It is observed that people get less than their average sleep most of the times during week as mode is less than average in most records.

11. I suspect that some students literally raced through the quiz. Can you find any evidence of racing? What evidence? Who are the suspects?
It is evident that some students raced through the questions, the time they took is really low.
Suspects are: 1392, 1043, 451, 1270, 767, 799, 1798

12. Do any attributes require deduplication? How would you propose this be performed?
Following questions need deduplication:

1. UNIQUE.ID field needs disambiguation. ID 1392 is repeated. Or due to some issue same user was allowed to give quiz twice.
2. First Name and Last Name might need both deduplication and disambiguation. Though such cases are not seen in provided data.
3. Swallow Speed – Converting m\s to mph
4. Fiction Question – Converting cases
5. Children's Story – Converting cases, extracting word
6. COMPUTER – converting case, extracting word

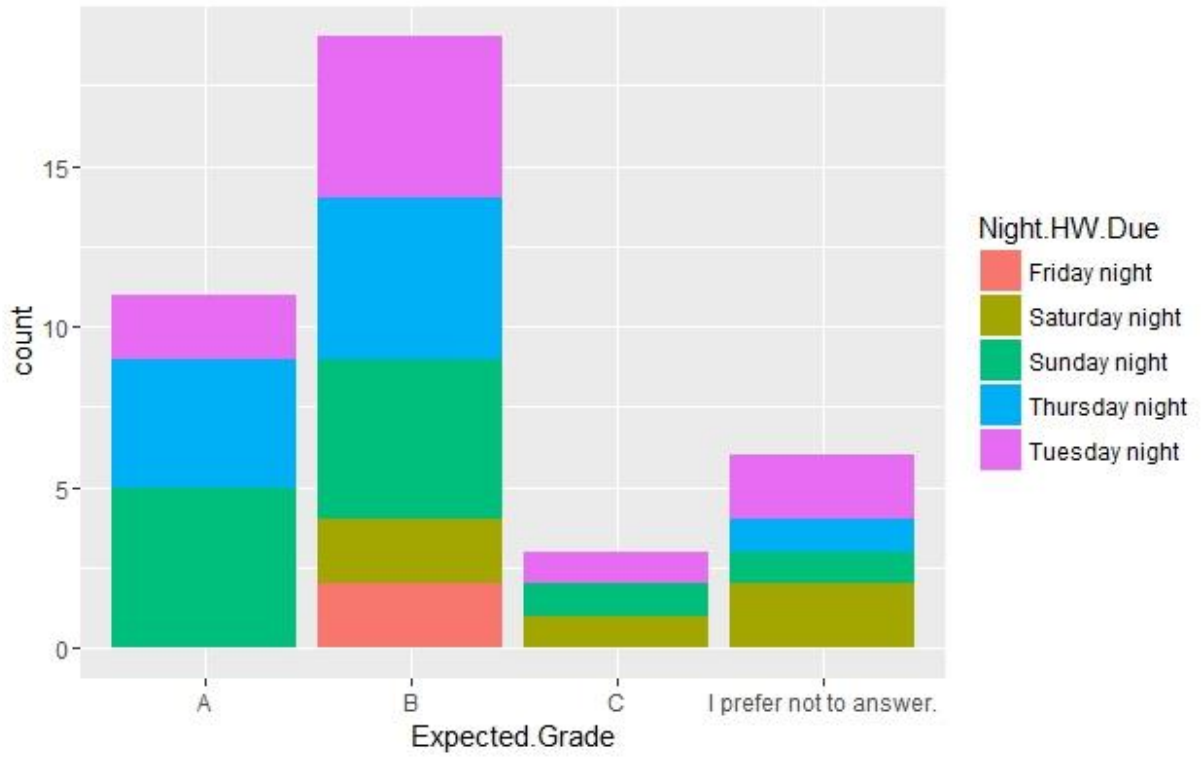
13. Does the answer to the question about the person who invented the first compiler help you differentiate between students at all?
No, All students have answered with same response: Grace Murray Hopper

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 such a survey question be changed?
867-5309 number was displayed in Tommy Tutone video as a phone number obtained from mens room wall. It ended up being famous fake number. Subjects who put this number are trying to convey they are not interested to give phone number. Survey question should be changed for this question to be optional.

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, and it was recently revealed as a problem invalidating up to 35% of all fMRI studies.)
People who wanted to enter multiple answers for sleep data entered slash or hyphen separated values which ended up as data in date format.

16. Of students that expect to get an A, what is the most common day that they want homework due?
Of 11 students who expect to get grade A, 5 want homework due on Sunday night.

17. Of students that expect to get a B, what is the most common day that they want homework due?
There is no one common homework due day among students who expect to get a B. Thursday night, Tuesday night and Sunday night have equal votes of 5 among 19 students which is seen in following plot.

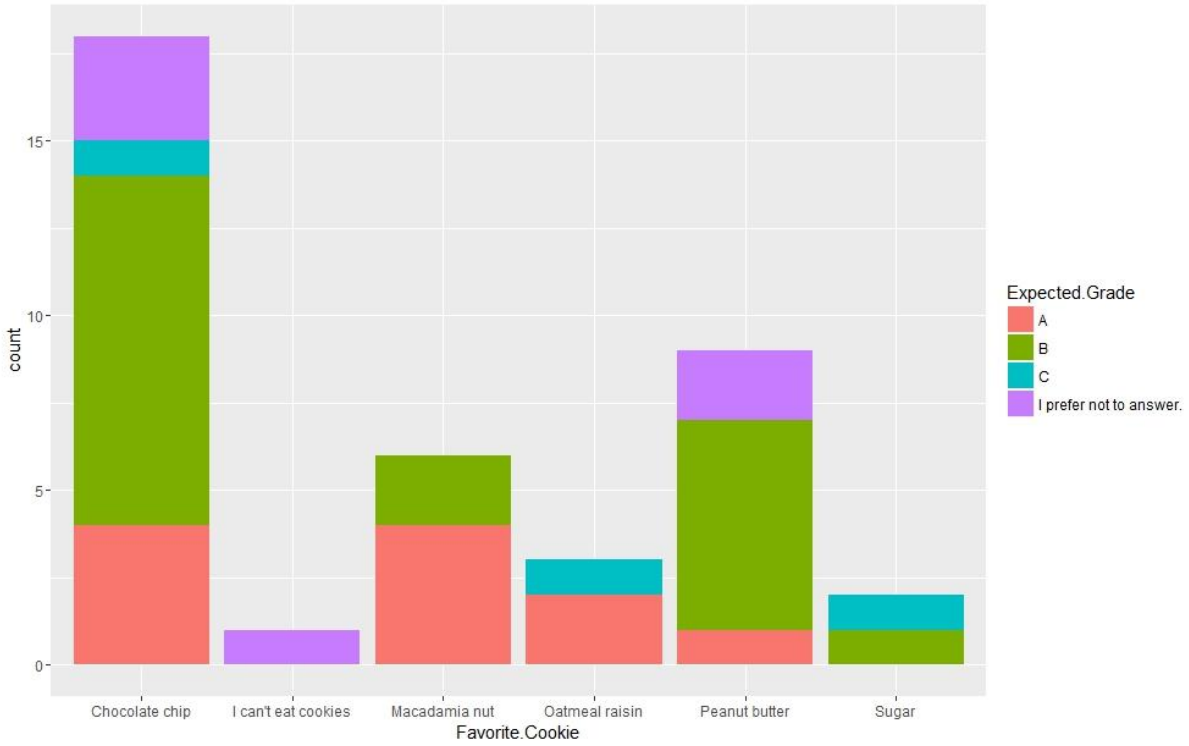


Sort the data by favorite kind of cookie:

18. For students whose favorite cookie is chocolate chip, what grade are they most likely to expect?
It is evident from the graph below that students whose favorite cookie is chocolate chip, expect to get grade B.

19. For students whose favorite cookie is peanut butter, what grade are they most likely to expect?
It is evident from the graph below that students whose favorite cookie is peanut butter, expect to get grade B.

20. For students whose favorite cookie is macadamia nut, what grade are they most likely to expect?
It is evident from the graph below that students whose favorite cookie is macadamia nut, expect to get grade A.



21. The survey lets the user refuse to answer the question about their expected grade. This generates missing data. Is this better or worse than the alternatives? Why do you think that? Support your answer with reasoning.

Survey lets user to refuse to answer the question which has following advantages over its alternatives:

1. Responses will be close to correct value: Since user have not analyzed the question and does not know answer, instead of giving fake answer he can choose not to answer.
2. User will not skip the question: If user skips the question because he/she does not know the answer, it will lead to empty values or NA values in the extracted data. These NA values need to be dealt with separately. Providing this option can provide substitute for NA values.

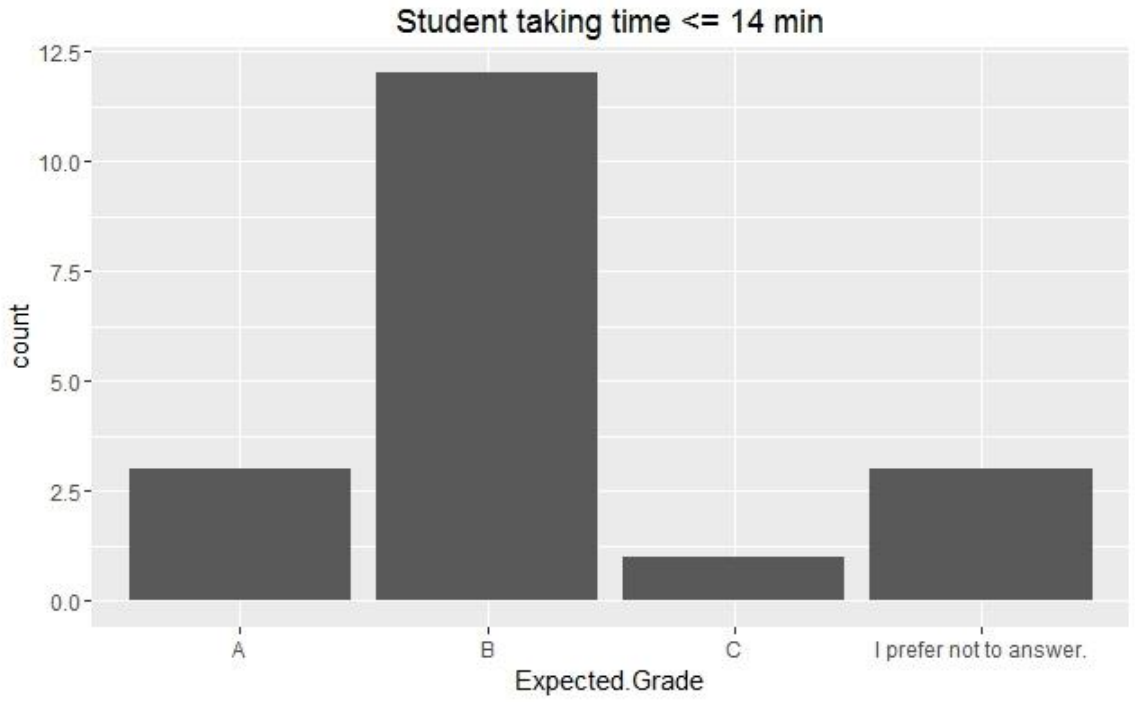
22. For students that refused to answer the question about their expected grade, could we guess the grade they expected to get using statistical analysis? What would you do to guess their expected grades?

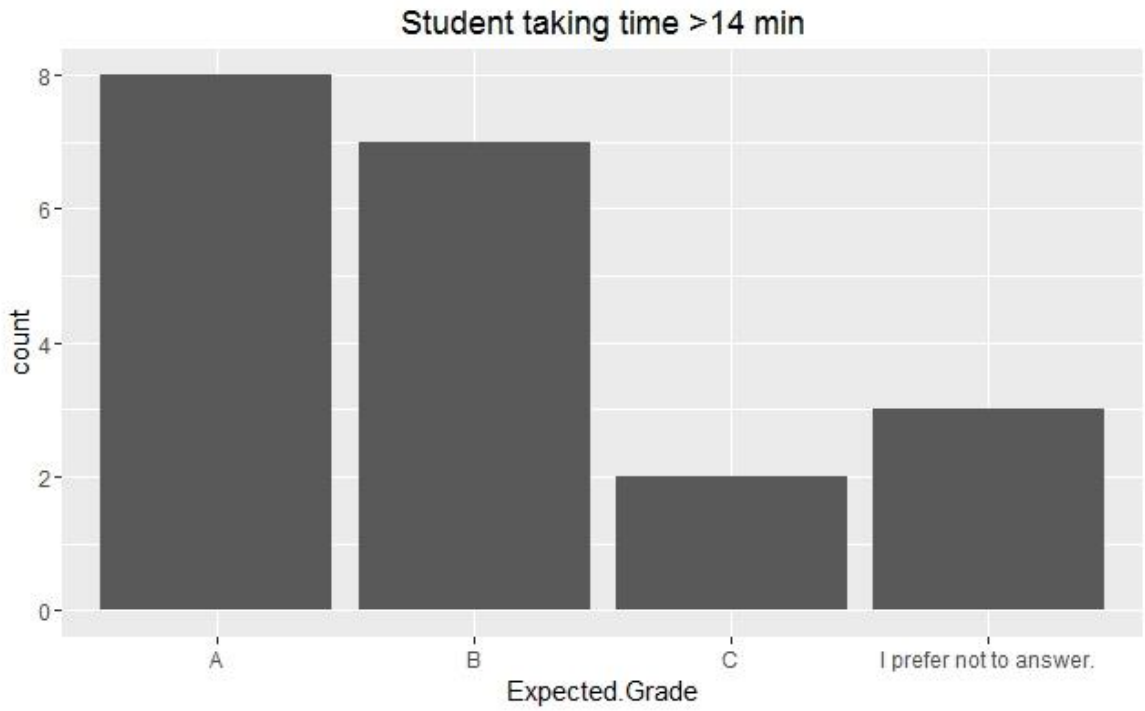
The 0R rule for grade attribute gives B as answer for students who refused to answer expected grade.

Expected Grade of student who refused to answer can also be guessed by using association of subset of students who refused to answer with other attributes like NameSpace.Color, Favorite.Cookie and Night.HW.Due to relate to the grade attribute. Attribute which gives best miss-classification rate can be chosen for 1R rule. Hence I would use 0R rule with their Cookie choice. 6 of 7 students who prefer not to answer, eat Chocolate chip or Peanut Butter cookie. Students eating Chocolate chip or Peanut Butter cookie expect to get grade B.

Hence we can say that students who refused to answer expect to get grade B.

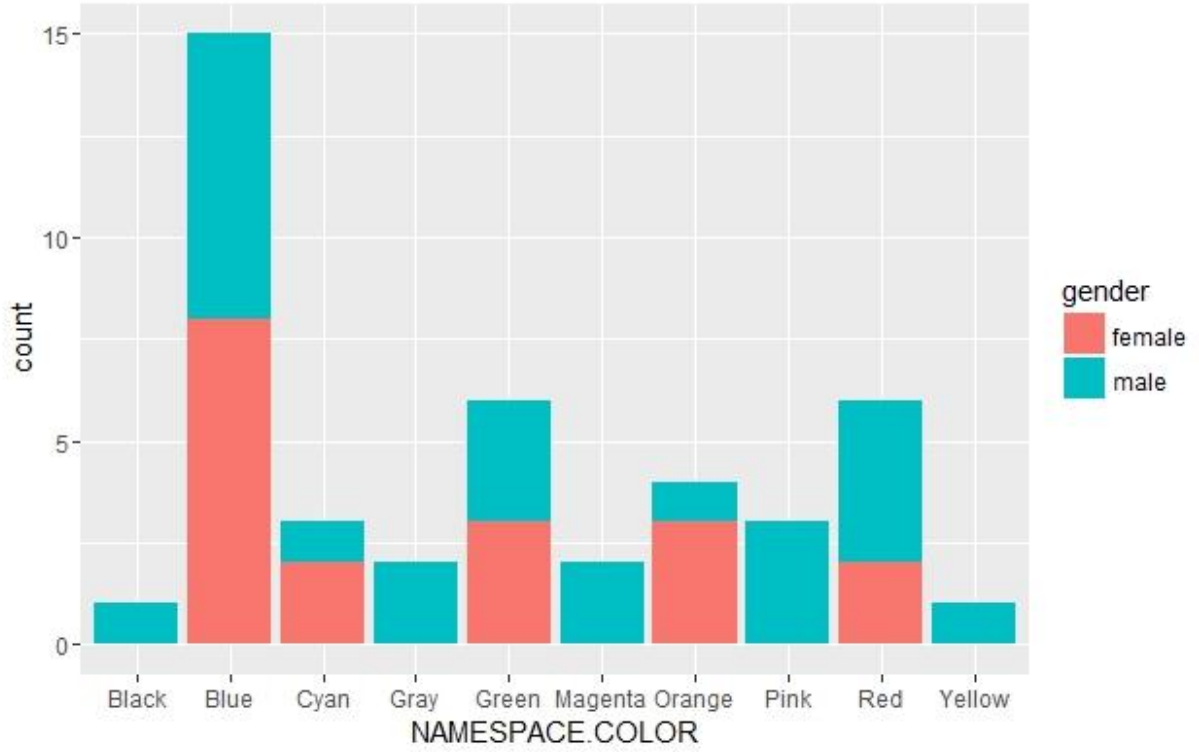
23. Sort the data by the amount of time taken on the quiz. Do you notice any patterns for respondents that took less than 0 to 14 minutes, compared to those that took 15 to 30 minutes?
It is observed that 19 out of 39 student took less than 15 minutes to respond quiz. It is noticed that students who took less time tend to like Chocolate Chip cookie and Peanut butter Cookie than other cookies. Whereas students who took more time liked Chocolate chip and Macadamia nuts cookies. It is also observed that majority of students who took less time, expect grade B and students who take more time expect grade A.





24. Do you notice any other patterns in the data, as presented?

I tried to guess gender of user based on their first name using the gender library in R. After plotting Namespace.Color against gender it was observed that only males liked gray, magenta, Pink colors.



25. Think of a good, intriguing, relevant or fun question we should ask next semester's class? What should we ask?

Which sandwich dressing you like the most?

Pesto Mayo, Chipotle Mayo, Horse-radish Mayo, Yellow Mustard, Honey Mustard, Brown Mustard, Mayonnaise, Thousand Island Dressing, Sriracha Mayo, Italian Dressing, None.