**SMALS Research Project**

Eric Shin

University of Washington Bothell

December 5th, 2022

**Introduction**

The purpose of this section of the on-going SMALS research project was to clean up the Michigan State University IDs data set, msuids for short. The task challenge was to clean up this data set by extracting important variables from the original filenames column in the MSU data set and putting these important variables into separate columns in a newly extracted data set based on the file name category. The separate file name categories were Year, ExamType, Course, Instructor, QuestionNumber, Semester, StudentID, QuestionConcept, QuestionContext, as well as QuestionType. These categories were all in separate columns in the newly extracted data set. Another task challenge that was assigned was to visualize the frequency of each extracted variable based on the specific file name through bar graphs. The key purpose of this entire on-going SMALS project is to analyze student-constructed models of biological systems and statistical concepts through a network science approach for the purposes of exploring changes/growth in models as evidence of learning as well as comparing the growth of student models among beginning, middle, and end models through the growth of network graphs and plots. The reason why I am doing this project for credit is because I want to not only gain research experience as a student, but to also improve my coding skills/experience in RStudio by learning how to use several coding techniques to extract data from a data set in R through regular expressions, pulling individual strings using the msuids dataset, creating individual columns for a newly extracted data set, and eventually creating network plots with the dataset. The main goal for me was to clean up the MSU data set all the way until every key variable was extracted from the MSU data set into new columns in the extracted data set. We couldn’t quite get to the part where we visualize the frequencies of the variables through network graphs and plots, but this may be something that I can work on with Professor Trujillo next quarter for this project. This section of the SMALS project demonstrates my ability to perform data cleaning by using regular expressions and the extraction of data by using code from the stringr package, data visualization by using ggplot graphs to show the frequencies of the extracted variables, data interpretation by analyzing the points of interest and the frequencies of the extracted variables that really stand out, as well as technical reporting to showcase what I have learned through this research credit.

**Methods**

The original MSU data set contained separate columns of student IDs, file names, with courses, semesters, years, instructors, exam types, and question numbers within these file names, and parent names, which were the original file names that were assigned to specific variables. Again, as mentioned earlier in the Introduction, we wanted to extract the key variables from the MSU data set and put these variables into different file name categories which were Year, ExamType, Course, Instructor, QuestionNumber, Semester, StudentID, QuestionConcept, QuestionContext, as well as QuestionType. In terms of the very first approach to clean the data, after loading the MSU data set into R using the read.csv code, selecting the important columns from the MSU data set using the select code, as well as creating a new object and data set named filenames from the selected columns, we wanted to analyze the data set and identify as many key variables as possible. After identifying all of the key variables, we put these variables into different string lists based on the relevancy to the file name. The variables were each separated by OR functions. File name objects were created for each string list such as Specific\_Year, Exam\_Type, Instructor\_Initials, Question\_Number, Specific\_Semester, Question\_Concept, as well as Question\_Context for the use of regular expressions and extraction code. The next approach that we used was regular expressions along with extraction code from the stringr package. During this approach, we created a series of str\_extract functions that extracted every single variable that matched with either the variables from the string list or the regular expression, depending on the specific file name variable that was being extracted. We used case\_when functions to replace values that do not match with a specific variable. We were able to assign non-matching values to a specific variable using this function. We also used str\_c functions which combined two column names into one individual column name. We used this function to combine the Semester and Year columns into one single and separate column, named AcademicTerm, with the Semesters and Years separated by commas within the single column, as well as the QuestionConcept and QuestionContext columns into one single and separate column named QuestionType, with the Question Concepts and Question Contexts separated by slashes within the single column. These mentioned str\_extract, str\_c, and case\_when functions were inside of the mutate function which created variables and columns in a new data set based on existing variables from the original data set. A new object named Extracted\_Names was a data set created from the extraction of the filenames data set using the mutate function which contains the extracted data and columns. For the QuestionConcept and QuestionContext columns, we used case\_when functions to assign specific parent names to specific Question Concepts, which were Gene To Evolution, Carbon Cycle, as well as Gene To Phenotype, and Question Contexts, which were Isle Royale, DDT Resistance, Owl, as well as Frogs. Inside of these case\_when functions, we used vector lists to list all of the parent names that needed to be assigned to the specific Question Concept and Question Context in order to avoid creating separate case\_when functions for every single parent name. We also used %in% functions inside of these case\_when functions which identified any Question\_Concept and Question\_Context string list variable that matched with the parent names in the filenames data set. If there was a match, these string list variables were assigned to a specific Question Concept and Question Context using the ~ “(\*Question Concept or Question Context\*)”, TRUE ~ as.character() function where variables get assigned to a specific column in the Extracted\_Names data set. These mentioned case\_when functions were inside of the second mutate function which created other variables and columns in the Extracted\_Names data set based on existing variables from the original data set.

**Results**

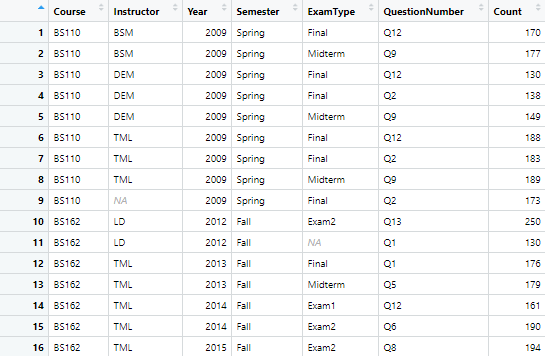


Table 1 shows the frequencies of each unique combination of variables from the Course, Instructor, Year, Semester, ExamType, and QuestionNumber file names.

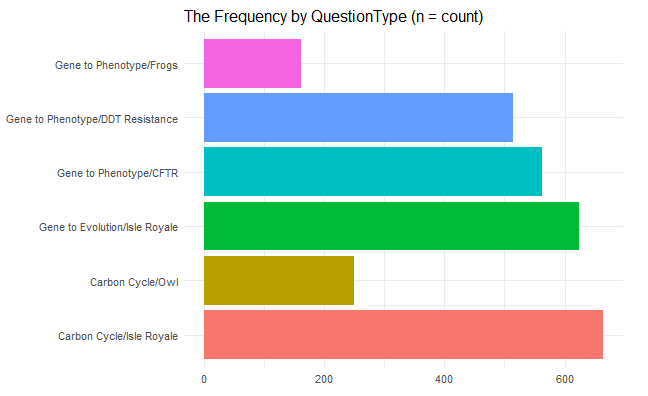


Figure 1 visualizes the frequencies of each QuestionType variable.

From analyzing the results in the table, we can see that the frequency of the transcribed models from each unique variable in the Extracted\_Names data set, which is measured by count, is spread evenly, hovering around the low to high 100s for each unique variable count with very few outliers. The only outlier that can be recognized is the unique variable of BS162 LD 2012 Fall Exam2 Q13 with a count of 250 transcribed models, the unique variable with the highest frequency in the table. This finding suggests that the type of course, instructor, year, and semester may have an influence on the student’s interest and needs in terms of their growth in learning and what they want to learn. From analyzing the results in the figure, we can see that the frequency of the transcribed models from each QuestionType variable in the Extracted\_Names data set, which is measured by count, is spread evenly again for the most part, with four variables hovering around the range of the 500s and 600s along with two noticeable variable outliers that are hovering around the range of the 100s and 200s. This shows that most students are mainly learning the Isle Royale Question Context, with a frequency of more than 1,200 transcribed models. This also shows that students are mainly learning the Gene To Phenotype Question Concept, with a frequency of more than 1,200 transcribed models. These findings suggest what types of Question Concepts and Question Contexts that students are understanding the best.

**Discussion**

Overall, the experience of contributing to an on-going research project was both very enjoyable and challenging at the same time. From having to research on my own and learn from Professor Trujillo on how to write the correct code to create variables in a string list and extract variables from a data set to assigning variables to a specific column in a data set, as well as visualizing and analyzing the frequencies of each extracted variable, I have gained valuable experience and knowledge from this research credit. Being able to clean up data for better analysis, which is a process that is part of data wrangling, as well as being able to visualize extracted data through graphs and tables are two main practices that I have learned during this research credit. The understanding of these practices is important because the data science field will need people who can perform these types of practices, especially for companies that utilize both data cleanup and data visualization. As someone who is trying to pursue a career in this field, this research credit has not only helped me gain valuable data research experience, but this has also helped me prepare for what type of work that I need to emphasize in order to be successful in this field.