Homework 1

About The Data

We will be working with a simulated data set related to educational outcomes and technology. Students in an introductory Physics course were tracked throughout a semester long class. In addition to the lectures and textbook, the students also used a digital system to work practice problems. The system includes an algorithm that assesses the level of a student's knowledge in the topic. It also measures the amount of minutes spent on the subject during the relevant period of time. Prior to the midterm exam, the students were expected to use the system and reach a level of 2 on a number of topics. Prior to the final exam, the students were expected to reach a level of 5 on all of the topics. The students also completed homework assignments, a midterm, and a final, which were all scored on a scale from 0 to 100 points. Based on their performance in the class, the students received overall scores that would form the basis of their grades. After the completion of the class, the university wanted to study the performance of the students and the impact of the digital system. To incorporate prior levels of knowledge, the university gathered data about the performance of these students in their mathematics curriculum of trigonometry and calculus.

The data were recorded in the following files:

Prior Knowledge: ../Data/Prior Courses.csv

Digital System, Prior to the Midterm Examination: ../Data/Knowledge Check – Level 2.csv

Digital System, Prior to the Final Examination: ../Data/Knowledge Check – Level 5.csv

Scores on the Homework, Exams, and Overall: ../Data/Grades.csv

Delimiter: Each column of each file is separated with a comma **,** delimiter.

All of the data files include an identifier column **Student ID** so that the information from different files can be linked. These identifiers also link to the names of the student and other private information about them, which have been separately stored in a secure location.

In some places, the data may contain unusual values. Any value that does not match up with reasonable expectations for the measure should be converted to a missing value for the purpose of the analysis.

Completing the Assignment

Use the information in the files to answer the questions below. To receive full credit, you must provide the output along with the full set of code used to generate it.

This assignment includes a relatively small amount of information. It would be possible to open all of the files in spreadsheet programs, perform visual inspections, or even answer the questions using other tools. **However, you must fully complete this assignment using R.**

Tips: We recommend familiarizing yourself with the **data.table** package in R. This will enable you to work with large amounts of data. Many of the questions can be answered with a relatively small amount of code by making use of data.table's functionality. We also recommend organizing your code. Within a folder (e.g. Homework 1 for this class), create separate subfolders for the Data and the Analysis. Store this file in the Analysis folder. Then you can use relative directories to read in the data. The template for this assignment includes variables defined in the **constant** code chunk that refer to the names of all of the files.

Question 1: Preliminaries

One way to read data files is using the **fread** function. Read in the data and answer these questions:

Question 1a) Dimensions

How many rows and columns are there in each file? Use the **dim** command to display the dimensions.

Prior Knowledge

Knowledge Check 1

Knowledge Check 2

Grades

Question 1b) Subjects

How many unique students were in the class? Make sure this calculation includes information from all of the files.

Question 2: Multiple Records

Which files (if any) contain more than 1 row per student? Display the records from these files for any students with multiple rows. Write a function called **display.multiple.records** that will perform this work on each table of data. Use the **datatable** function in the **DT** package to display an HTML table of these results in sorted order of Student ID. (In spite of the similarity in their names, the **datatable** function in the **DT** library for displaying tables in HTML should not be confused with the **data.table package** for data processing.) If there are no students with multiple records in a given table, display an empty table as the result.

Prior Knowledge

Knowledge Check 1

Knowledge Check 2

Grades

Question 3: Reduction to a Single Record

To handle the students with multiple records, we decided to summarize their prior knowledge as follows:

- For each student, the highest score in a prior class will be used. If no numeric record is included, an NA value should be used. For reference, we have provided a function called **max.with.na** that can perform this calculation for a single student.
- We will also create an overall score called Prior Knowledge Level. For each student, this will be defined as the average of the student's highest score in Trigonometry and the student's highest score in Calculus. For students who did not take both of these classes, the overall score will be based on the measured values.

Based on these criteria, we will answer the following questions.

Question 3a) Summary Before the Reduction

Starting with the original table of Prior Knowledge scores, compute the following for each class: the number of measured records, the number of unique students with a measured record, the average score among all of the measured records, and the standard deviation of the scores among all of the measured records. Round all of the numeric values to 2 decimal places. Write a function called **summarize.pk.class** that will display the name of the prior class along with these figures using the **datatable** method from the **DT** package.

Trigonometry

Calculus

Question 3b) Reduction of Information

Now create a new table called **pk.reduced** that will contain 1 record per student according to the criteria specified above. For the students with multiple records in the original file, display their records in the **pk.reduced** table using the **datatable** function in the **DT** package.

Question 3c) Summary After the Reduction

Now, using the **pk.reduced** table, compute the following for each class and the Prior Knowledge Level: the number of measured records, the number of unique students with a measured record, the average score among all of the measured records, and the standard deviation of the scores among all of the measured records. Round all of the numeric values to 2 decimal places.

Trigonometry

Calculus

Prior Knowledge Level

Question 4: Combining the Tables

Now we want to create one comprehensive table called **dat**. It should be constructed according to the following criteria:

- Each student has 1 row of information (1 record).
- The Student IDs are sorted in increasing order (1, 2, 3, etc.)
- The columns first include the Prior Knowledge, then the Knowledge Check 1, then the Knowledge Check 2, and then the Grades.
- Every column has a unique and meaningful name.

After creating this table, display it using the **datatable** function in the **DT** package. Round all of the numeric measures to 2 decimal places.

Hints: There are a number of ways to combine multiple tables. The **data.table(a, b, c, ...)** function will bind the columns of multiple objects. The **merge(x, y, by, all.x, all.y)** function will combine (join) two tables **x** and **y** according to a character vector of column names **by** (or alternatively **by.x* and** by.y**). Specifying** all.x** and **all.y** as TRUE or FALSE (in different combinations) will give different options for combining tables with different records or elements. By comparison, **data.table** and **merge** have different advantages and drawbacks. Using **data.table** is more straightforward, but it assumes more about the structure of the tables. Using **merge** is more flexible in terms of the ordering and differences in the tables, but it requires more code and complexity for joining more than two tables together. Meanwhile, combining the tables based on the **Student ID** can encounter ordering problems as a character vector. It may be easier to extract a numeric version of the **Student ID** for sorting purposes. However, the final version of the table should only include the original identifier.

Question 5: Knowledge Check 1

How did the students do on the first knowledge check? Create a table with the following columns:

- Topic

- Number of Students (with measured scores)
- Mean Score
- Standard Deviation of Scores
- Percentage Reaching Threshold 2 or Higher (on a scale from 0 to 100).
- Mean Minutes
- Standard Deviation of Minutes

The table should have one row for each topic in the first Knowledge Check. Round all numeric values to 2 decimal places.

Question 6: Knowledge Check 2

How did the students do on the second knowledge check? Create a table with the following columns:

- Topic
- Number of Students (with measured scores)
- Mean Score
- Standard Deviation of Scores
- Percentage Reaching Threshold 2.0 or Higher (on a scale from 0 to 100).
- Mean Minutes
- Standard Deviation of Minutes

The table should have one row for each topic in the first Knowledge Check. Round all numeric values to 2 decimal places.

Question 7: Is Time Spent Time Well Used?

For each knowledge check topic at each level, compute the correlation between the students' scores and their minutes spent using the digital system. Display the results using the **datatable** function in the **DT** package. Round the numeric values to 2 decimal places.

Then comment on the findings. Do you think spending more time on the digital system is beneficial? Is your notion confirmed by the data? Whatever you believe, why might the opposite be true?

Hint: Reshaping the data to place all of the knowledge check scores in one column and all of the time spent in another may simplify the calculation. To do this, consider using the **melt** or **melt.data.table** function.

Question 8: Summary of Scores

For each homework assignment, the student's average homework score across all assignments, the midterm exam, the final exam, and the overall score, compute the following quantities:

- The number of students with a measured value.
- The mean score.
- The standard deviation of the score.

Display these results using the **datatable** function in the **DT** package. Round all of the numeric values to 2 decimal places.

Question 9: Correlations with Outcomes

For the purpose of evaluating the class, consider the following outcomes:

- Homework Average
- Midterm Score
- Final Exam Score
- Total Score

For each of these outcomes, compute their correlation with each of the following predictors:

- The Prior Knowledge Level
- Total Minutes spent on the knowledge check activities.
- The average score on the Level 2 knowledge checks (with a 5 threshold).

For the purpose of the time calculations, consider any missing value as a zero in computing each student's total minutes using the system.

Display the results using the **datatable** function in the **DT** package. Round all of the numeric values to 2 decimal places.

Question 10: Qualitative and Quantitative Conclusions

Based on the results that you have seen, what can you conclude about the digital system that the students used? Does spending more time on the system seem to help improve the scores on the homwork, midterm, final exam, and overall grade? How do higher scores on the system impact these outcomes? And how does this compare to the impact of the Prior Knowledge Level? Write a few sentences to outline your conclusions and recommendations.