Missing Data

In this assignment you will complete a variety of tasks related to working with missing data.

Deliverable: All of your work for this assignment should be done in an R Markdown document. Knit your document into a Word file and submit the Word file as the deliverable for this assignment.

Libraries: For this assignment you will need the following libraries: tidyverse, VIM, and mice.

Before beginning the assignment tasks, you should read-in the data for the assignment into a data frame called grades. This data contains grade information from an engineering course. The dataset was originally used to investigate how student performance in the course would be predictive of student grades on the "Final" exam. The "Prefix" column is a surrogate for enrollment year in the engineering program. Smaller values imply older (more mature?) students.

Task 1: How much data is missing and in what variables?

Task 2: Use the VIM package to visualize missingness. Does there appear to be systematic missingness? In other words, are there students that are mising multiple pieces of data?

Task 3: Use row-wise deletion of missing values to create a **new data frame**. How many rows remain in this data frame?

Task 4: Use column-wise deletion of missing values to create a **new data frame** (from the original data frame not from the data frame created in Task 3). How many columns remain in this data frame?

Task 5: Which approach (Task 3 or Task 4) seems preferable for this dataset? Briefly discuss your answer.

Task 6 Use the code below to impute the missing values in the dataset using the mice package.

```r
grades_imp = mice(grades, m=1, method = "pmm", seed = 12345)
#in line above: m=1 -> runs one imputation, seed sets the random number seed to get repeatable results
summary(grades_imp)
densityplot(grades_imp)
#red imputed, blue original, only shows density plots when more than 1 value the variable was imputed
#note that the density plots are fairly uninteresting given the small amount of missing data
grades_complete = complete(grades_imp)
summary(grades_complete)
```

Task 7: Briefly discuss potential issues that could be encountered when working with missing data. Describe situations where imputation may not be advisable.