Exercise Sheet 1

For the ncyflights13 data set:

1. Find all flights that:
   1. Had an arrival delay of two or more hours
   2. Flew to Houston (IAH or HOU)
   3. Departed in summer (July, August, and September)
   4. Were delayed by at least an hour, but made up over 30 minutes in flight
   5. Departed between midnight and 6am (inclusive)

Can you use the between() function to simplify the code needed to answer the previous questions?

1. How many flights have a missing dep\_time? What other variables are missing?
2. Currently dep\_time and sched\_dep\_time are convenient to look at, but hard to compute with because they’re not really continuous numbers. Convert them to a more convenient representation of number of minutes since midnight.

e.g. the value 517, represents five hours and 17 minutes past midnight. To calculate the number of minutes past midnight we can write 5 x 60 +17 = 317.

1. Find the 10 most delayed flights using a ranking function. Carefully read the documentation for min\_rank().
2. Download the Pb\_messy.csv and Pb\_tidy.csv data sets from Blackboard (In the folder for Exercise Sheet 1). The data represents the levels of lead (Pb) measured in soil samples taken from different treatments. For each site (1 and 2), there were two plots (A and B) where measurements were taken for a control sample and a treatment sample. Can you create an R script that will transform the messy data set into the tidy data set?
3. A census collecting household data included information on an individual’s age and income. The variables were checked for missing data and the plots that were generated are shown in Figures 1-3 below.

**Figure 1**



**Figure 2**



**Figure 3**



* 1. Comment on the extent of the missing data for the variable ‘Income’ and the variable ‘Age’.
  2. Is there a missing data pattern evident for either of the variables? If so what is the pattern?

1. Download the cholesterol.csv data set from blackboard. Note that to use the VIM package, the data will need to be in the form of a dataframe rather than a tibble. Perform some exploratory data analysis on the cholesterol data set and present the results in a short report including:
   1. A summary table.
   2. Variable distributions.
   3. Scatterplots.
   4. Comment on the relationships between the continuous variables.
   5. Comment on the extent of the missing data and identify whether there are any missing data patterns.
   6. Identify whether there are any outliers in the data set.