# IMT 573: Module 4 Lab

Data Integration

Jenny Skytta

Due: April 24, 2022

**Collaborators: Independent work** List collaborators here.

### Objectives

In this lab exercise you will practice data cleaning and integration skills.

### Instructions

Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Open the `04_lab_dataintegration.Rmd` and save a copy to your local directory. Supply your solutions to the assignment by editing `04_lab_dataintegration.Rmd`.
2. First, replace the "YOUR NAME HERE" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and I encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`. When the PDF report is generated rename the knitted PDF file to `lab4_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

In this lab you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(nycflights13)
library(knitr) # this will keep code on the page!
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

**Problem 1: Data Cleaning**

In this problem we will use data found in the file `weather.txt`. Import the data into **R** and answer the following questions. This is challenging! I have given you no other information other than the file name. See what you can come up with for these questions.

```
Weather <- read.delim("data/weather.txt")

summary(Weather)  #Summary that includes variables, descriptive stats
```

```
##       id                 year          month          element
##  Length:22          Min.   :2010   Min.   : 1.000   Length:22
##  Class :character   1st Qu.:2010   1st Qu.: 3.250   Class :character
##  Mode  :character   Median :2010   Median : 6.000   Mode  :character
##                     Mean   :2010   Mean   : 6.273
##                     3rd Qu.:2010   3rd Qu.: 9.500
##                     Max.   :2010   Max.   :12.000
##
##        d1              d2              d3              d4            d5
##  Min.   :138.0   Min.   :144.0   Min.   :144.0   Min.   :120   Min.   : 79.0
##  1st Qu.:178.2   1st Qu.:158.2   1st Qu.:167.2   1st Qu.:158   1st Qu.:141.5
##  Median :218.5   Median :218.0   Median :208.0   Median :196   Median :210.5
##  Mean   :218.5   Mean   :223.2   Mean   :211.5   Mean   :196   Mean   :208.6
##  3rd Qu.:258.8   3rd Qu.:283.0   3rd Qu.:252.2   3rd Qu.:234   3rd Qu.:276.5
##  Max.   :299.0   Max.   :313.0   Max.   :286.0   Max.   :272   Max.   :321.0
##  NA's   :20      NA's   :18      NA's   :18      NA's   :20    NA's   :14
##        d6              d7             d8            d9               d10
##  Min.   :105.0   Min.   :129   Min.   :173.0   Mode:logical   Min.   :168.0
##  1st Qu.:148.2   1st Qu.:167   1st Qu.:202.2   NA's:22        1st Qu.:212.2
##  Median :191.5   Median :205   Median :231.5                  Median :256.5
##  Mean   :191.5   Mean   :205   Mean   :231.5                  Mean   :256.5
##  3rd Qu.:234.8   3rd Qu.:243   3rd Qu.:260.8                  3rd Qu.:300.8
##  Max.   :278.0   Max.   :281   Max.   :290.0                  Max.   :345.0
##  NA's   :20      NA's   :20    NA's   :20                     NA's   :20
##       d11           d12              d13             d14             d15
##  Min.   :134.0   Mode:logical   Min.   :165.0   Min.   :130.0   Min.   :105.0
##  1st Qu.:174.8   NA's:22        1st Qu.:198.2   1st Qu.:156.2   1st Qu.:150.5
##  Median :215.5                  Median :231.5   Median :230.0   Median :196.0
##  Mean   :215.5                  Mean   :231.5   Mean   :222.2   Mean   :196.0
##  3rd Qu.:256.2                  3rd Qu.:264.8   3rd Qu.:296.0   3rd Qu.:241.5
##  Max.   :297.0                  Max.   :298.0   Max.   :299.0   Max.   :287.0
##  NA's   :20                     NA's   :20      NA's   :18      NA's   :20
##       d16             d17           d18             d19             d20
##  Min.   :176.0   Min.   :175.0   Mode:logical   Mode:logical   Mode:logical
##  1st Qu.:209.8   1st Qu.:201.2   NA's:22        NA's:22        NA's:22
##  Median :243.5   Median :227.5
##  Mean   :243.5   Mean   :227.5
##  3rd Qu.:277.2   3rd Qu.:253.8
##  Max.   :311.0   Max.   :280.0
##  NA's   :20      NA's   :20
##    d21             d22             d23             d24             d25
##  Mode:logical   Mode:logical   Min.   :107.0   Mode:logical   Min.   :156.0
##  NA's:22        NA's:22        1st Qu.:139.2   NA's:22        1st Qu.:191.2
##                                Median :207.0                  Median :226.5
```

```
##                                 Mean   :205.0                 Mean   :226.5
##                                 3rd Qu.:272.8                 3rd Qu.:261.8
##                                 Max.   :299.0                 Max.   :297.0
##                                 NA's   :18                    NA's   :20
##       d26             d27             d28             d29             d30
##  Min.   :121   Min.   :142.0   Min.   :150.0   Min.   :153.0   Min.   :145.0
##  1st Qu.:161   1st Qu.:170.8   1st Qu.:190.5   1st Qu.:173.2   1st Qu.:178.2
##  Median :201   Median :229.5   Median :231.0   Median :230.0   Median :211.5
##  Mean   :201   Mean   :243.8   Mean   :231.0   Mean   :228.5   Mean   :211.5
##  3rd Qu.:241   3rd Qu.:318.2   3rd Qu.:271.5   3rd Qu.:285.2   3rd Qu.:244.8
##  Max.   :281   Max.   :363.0   Max.   :312.0   Max.   :301.0   Max.   :278.0
##  NA's   :20    NA's   :16      NA's   :20      NA's   :18      NA's   :20
##       d31
##  Min.   :154
##  1st Qu.:179
##  Median :204
##  Mean   :204
##  3rd Qu.:229
##  Max.   :254
##  NA's   :20
```

```
# View(Weather)
```

**(a) What are the variables in this dataset? Describe what each variable measures.**

Hint: There are five variables of interest here.

This appears to be a tibble with 35 variables; 5 of which include ID which represents the Meteorological Station ID for Cuernavaca Mexico, a Year (2010), all 12 months as integers, a thermal max and thermal min of temperature within each moth of that year. I searched for the id in google which brought up a page https://geographic.org/global_weather/mexico/cuernavaca_004.html and just by knowing that this represents weather and a location, I was able to infer the rest of the information. The d columnns logically must represent the day within the month where there was a max and min. This is inferred from the basis that they are d1 through d31 which is the maximum number of days within the longest month.

**(b) Tidy up the weather data such that each observation forms a row and each variable forms a column.**

```
weather_long <- Weather %>% #load weather tibble
  gather(key = "Days",  #gather to truncate on days
         value = Temps, #set column name to Temps for the temp values
         c(-id, -year, -month, -element),na.rm = TRUE) %>%
  # use all observations except year, month, and element, and remove na values
  arrange(desc(month)) #arrange descending by month
```

### Problem 2: Data Integration

Flight delays are often linked to weather conditions. How does weather impact flights from NYC? We utilize both the `flights` and `weather` datasets from the `nycflights13` package to explore this question.

First consider conducting a brief exploratory analysis of the weather data. In your EDA you might want to consider which weather variables are associated with impact on flights. Explain your choices in how you are measuring or evaluating impact on flights. You will likely need to integrate the flights and weather datasets in your analysis.

```
data("weather")  #load weather and flights tibbles into global environment
data("flights")

# View(weather) # View the tabular tibble to explore
ls(weather)  #display of variables included
```

```
##  [1] "day"        "dewp"       "hour"       "humid"      "month"
##  [6] "origin"     "precip"     "pressure"   "temp"       "time_hour"
## [11] "visib"      "wind_dir"   "wind_gust"  "wind_speed" "year"
```

```
dim(weather)  # dimensions of weather tibble
```

```
## [1] 26115    15
```

```
#View(averageflights)

flightweather <- left_join(weather, flights, by = c("time_hour" = "time_hour", "origin" = "origin", "day
  drop_na() #remove the NA values

flightweather_delays <- flightweather %>%
  filter(dep_delay > 1 & arr_delay > 1) %>% #filter for delays
  select(wind_gust, wind_speed, pressure, dep_delay, arr_delay, origin, visib) %>%  #select columnn
  summarise(visib = mean(visib), delay = mean(dep_delay - arr_delay), wind_speed = mean(wind_speed), wi


flightweather_ontime <- flightweather %>%
  filter(dep_delay < 1 & arr_delay < 1) %>% #filter for on time flights
   select(wind_gust, wind_speed, pressure, dep_delay, arr_delay, origin, visib) %>%
     summarise(visib = mean(visib), delay = mean(dep_delay - arr_delay), wind_speed = mean(wind_speed),
 knitr::kable(flightweather_delays, align = "cccc", caption = "Delayed Flight's Average Weather Metrics
```

Table 1: Delayed Flight's Average Weather Metrics.

| visib | delay | wind_speed | wind_gust |
|---|---|---|---|
| 9.64904 | -0.7803335 | 16.78011 | 25.19768 |

```
knitr::kable(flightweather_ontime, align = "cccc", caption = "On Time Flight's Average Weather Metrics
```

Table 2: On Time Flight's Average Weather Metrics.

| visib | delay | wind_speed | wind_gust |
|---|---|---|---|
| 9.898951 | 11.37362 | 16.34368 | 24.71535 |

In the EDA of the weather tibble, I see that there are some affecting measures related to flights; specifically visibility, and wind variables which could affect a flights speed if they're either flying against the wind or with the wind. It appears that higher wind gust, higher wind speed and lower visibility impact flight average delays.

**Citations**

Reading Data From TXT|CSV Files: R Base Functions http://www.sthda.com/english/wiki/reading-data-from-txt-csv-files-r-base-functions

Code written above is from the previous course IMT 511 which used the below text to support class scripts. https://www.google.com/books/edition/Programming_Skills_for_Data_Science/BnB6DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover