IMT 573: Module 4 Lab

Data Integration

Tommy Huynh

Due: July 16, 2021

Collaborators: List collaborators here.

Objectives

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)
```

```
## Warning: package 'tidyr' was built under R version 4.1.1
## Warning: package 'readr' was built under R version 4.1.1
```

```
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 4.1.1
```

```
weather_data <- read.table("weather.txt")
```

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.

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

```
head(weather_data, 4)
```

```
##            V1   V2    V3      V4   V5   V6   V7   V8   V9  V10  V11  V12  V13
## 1          id year month element   d1   d2   d3   d4   d5   d6   d7   d8   d9
## 2 MX000017004 2010     1    TMAX <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 3 MX000017004 2010     1    TMIN <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 4 MX000017004 2010     2    TMAX <NA>  273  241 <NA> <NA> <NA> <NA> <NA> <NA>
##    V14  V15  V16  V17  V18  V19  V20  V21  V22  V23  V24  V25  V26  V27  V28
## 1  d10  d11  d12  d13  d14  d15  d16  d17  d18  d19  d20  d21  d22  d23  d24
## 2 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 3 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>
## 4 <NA>  297 <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA> <NA>  299 <NA>
##    V29  V30  V31  V32  V33  V34  V35
## 1  d25  d26  d27  d28  d29  d30  d31
## 2 <NA> <NA> <NA> <NA> <NA>  278 <NA>
## 3 <NA> <NA> <NA> <NA> <NA>  145 <NA>
## 4 <NA> <NA> <NA> <NA> <NA> <NA> <NA>
```

The variables in the data set are the ID, Year, Month, Measurement Element, and Day of the Month

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

```
weather_summary  <- weather_data %>%
  group_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.

I think that weather will have a correlation with flight delays. The lower the tmin or the higher tmax is, the more likely it is for flights to be delayed due to weather related issues.