IMT 573: Module 4 Lab

Data Integration

Ali Qazi

Due: July 17th, 2022

Collaborators: List collaborators here. Akeel Qazi, Anthony Mercado

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)
library(nycflights13)
library(data.table)
library(magrittr)
library(tidyr)
```

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.

Hint: There are five variables of interest here.

The variables in this data set are id, year, month, element, day, and temp. The variables consist of the date (month, day, year) of the weather condition, the temperature min and max, and the id associated with the information.

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

```
data <- fread("weather.txt")
head(data)
```

```
##             id year month element d1  d2  d3 d4  d5 d6 d7 d8 d9 d10 d11 d12 d13
## 1: MX000017004 2010     1    TMAX NA  NA  NA NA  NA NA NA NA NA  NA  NA  NA  NA
## 2: MX000017004 2010     1    TMIN NA  NA  NA NA  NA NA NA NA NA  NA  NA  NA  NA
## 3: MX000017004 2010     2    TMAX NA 273 241 NA  NA NA NA NA NA  NA 297  NA  NA
## 4: MX000017004 2010     2    TMIN NA 144 144 NA  NA NA NA NA NA  NA 134  NA  NA
## 5: MX000017004 2010     3    TMAX NA  NA  NA NA 321 NA NA NA NA 345  NA  NA  NA
## 6: MX000017004 2010     3    TMIN NA  NA  NA NA 142 NA NA NA NA 168  NA  NA  NA
##    d14 d15 d16 d17 d18 d19 d20 d21 d22 d23 d24 d25 d26 d27 d28 d29 d30 d31
## 1:  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA 278  NA
## 2:  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA 145  NA
## 3:  NA  NA  NA  NA  NA  NA  NA  NA  NA 299  NA  NA  NA  NA  NA  NA  NA  NA
## 4:  NA  NA  NA  NA  NA  NA  NA  NA  NA 107  NA  NA  NA  NA  NA  NA  NA  NA
## 5:  NA  NA 311  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 6:  NA  NA 176  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

```
?weather
```

```
# making columns wider
# ignore warning
datatable <- melt(data, id.vars = c("id", "year", "month", "element"),
          variable.name = "day",
          value.name = 'temp')
```

```
## Warning in melt.data.table(data, id.vars = c("id", "year", "month",
## "element"), : 'measure.vars' [d1, d2, d3, d4, ...] are not all of the same type.
## By order of hierarchy, the molten data value column will be of type 'integer'.
## All measure variables not of type 'integer' will be coerced too. Check DETAILS
## in ?melt.data.table for more on coercion.
```

```
datatable[, day := as.integer(gsub("d", "", day))]
view(datatable)
```

```
datatable[, Year := paste(year)]
datatable[, Month := paste( month)]
datatable[, Day := paste(day)]

# remove columns not needed
datatable[, c("year", "month", "day") := NULL]

datatable[, element := tolower(element)]

# wider
datatable <- dcast(datatable, ... ~ element, value.var = "temp")

# remove entries with NA values
datatable <- datatable[!(is.na(tmax) & is.na(tmin))]

view(datatable)
```

Problem 2: Data Integration

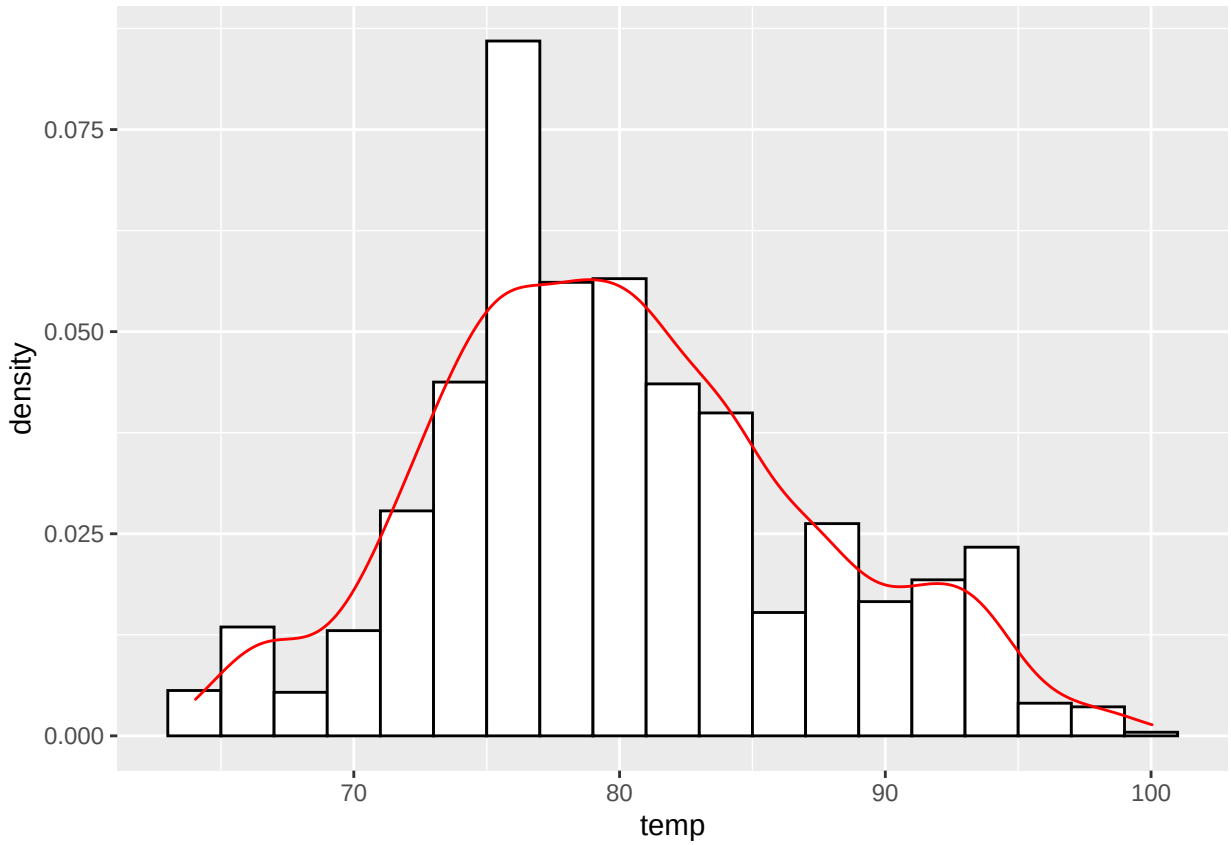
Flight delays are often linked to weather conditions. How does weather impact flights from NYC? We utilize both the `flights` and `weather` datasets from the `nycflights13` package to explore this question.

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.

```
library(tidyverse)
library(nycflights13)

# looking at temperature from month 7
# month with the most delays
weather %>% filter(month ==7) %>% ggplot(aes(temp, ..density..)) +

geom_histogram(binwidth = 2, colour = "black", fill = "white") + geom_density(colour = "red")
```



The histogram shows that the distribution of temperature in month 7 is skewed to the right. From last week's assignment we know that month 7 is the month with the most delays. We can see that temperature plays a big factor in this.

Something to note is that the weather.txt data does not show weather from the year 2013. So it would not make sense to compare the 2010 data with the 2013 NYC data. However,by looking at the data set weather in the nycflights13, we can see the correlation of weather and flights delays.