IMT 573 Lab: Data Wrangling

Naga Soundari Balamurugan

October 9th, 2018

Don't forget to list the full names of your collaborators!

Collaborators: Jayashree Raman

Instructions:

Before beginning this assignment, please ensure you have access to R and/or RStudio.

1. Download the `week3a_lab.Rmd` file from Canvas. Open `week3a_lab.Rmd` in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week3a_lab.Rmd`. You will also want to download the `weather.txt` data file, containing a dataset capturing daily temperatures in Cuernavaca, Mexico during 2010.
2. Run `install.packages('babynames')` in your "console" in R Studio to install the library.
3. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
4. Be sure to include code chucks, figures and written explanations as necessary. Any collaborators must be listed on the top of your assignment. Any figures should be clearly labeled and appropriately referenced within the text.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, rename the R Markdown file to `YourLastName_YourFirstName_lab3a.Rmd`, and knit it into a PDF. Submit the compiled PDF on Canvas.

In this lab, you will need access to the following R packages:

```r
# Load some helpful libraries
library(dplyr)
library(reshape2)
library(tidyverse)
library(babynames)
```

Problem 1: Data Cleaning

In this problem we will use the `weather.txt` data. Import the data in **R** and answer the following questions.

Hint: You might find the function `read.table()` useful here.

```r
#Read the weather dataset from text file
weatherData <- read.table("weather.txt")
```

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

Hint: There are five variables of interest here.

```
# EDIT ME
```

The five variables in the weather dataset are id, year, month, element and day(d1 to d31). The ID could be a location ID but from the data, there is only a single ID which is MX000017004. The variable year represents the year in which the temperature is recorded. It is 2010 throughout the dataset. The variable month represents the month in which the temperature is recorded. The dataset contains temperature from all month except september. Each month is represented twice for TMin and TMax. The variable element represents of the temperature is minimum(TMIN) or maximum(TMAX). The elements d1 to d31 represents the days of a month. A major observation is most of the data remains NA.

(b) Tidy up the weather data such that each observation forms a row and each variable forms a column. You might find the following functions helpful:

- `melt`
- `mutate`
- `dcast`

```
# EDIT ME
```

Problem 2: Data Manipulation

In this problem we will use the `babynames` data. Use the data to answer the following questions. Start by familiarizing yourself with the data.

```
#Load the babynames data
ls("package:babynames")
```

```
## [1] "applicants" "babynames"  "births"     "lifetables"
```

The babynames package has four datasets: applicants, babynames, births, lifetables.

(a) What name/s has/have been used for the most number of years (when used for a single gender)?

```
# EDIT ME
babynamesData <- babynames::babynames
femaleBabyNames <- babynamesData %>% filter(sex == 'F') %>% group_by(name) %>% dplyr::summarize(count =
maleBabyNames <- babynamesData %>% filter(sex == 'M') %>% group_by(name) %>% dplyr::summarize(count = n


frequentFemaleBabyNames <- femaleBabyNames %>% filter(count == max(femaleBabyNames$count)) %>% select(na
frequentMaleBabyNames <- maleBabyNames %>% filter(count == max(maleBabyNames$count)) %>% select(name)

freqMaleNameCount <- nrow(frequentMaleBabyNames)
freqFemaleNameCount <- nrow(frequentFemaleBabyNames)

freqMaleNameCount
```

```
## [1] 481
```

```
freqFemaleNameCount
```

```
## [1] 452
```

The list *frequentFemaleBabyNames* and *frequentMaleBabyNames* has the list of most frequent female and male baby names respectively. There are 481 male names and 452 female names that have been used for most number of years(136 years)

(b) Which name received the largest percentage of any name for any year for any gender?

```
# EDIT ME
mostFameName <- babynamesData %>% filter(prop == max(babynamesData$prop))
mostFameName
```

```
## # A tibble: 1 x 5
##    year sex   name      n   prop
##   <dbl> <chr> <chr> <int>  <dbl>
## 1 1880. M     John   9655 0.0815
```

John is the most famous name for any gender/year and had a largest percentage of 0.0815.

(c) Which name recorded in the data set has been out of use for the longest time?

```
# EDIT ME
longestTimeDiff <-  babynamesData %>% group_by(name) %>%
  dplyr::summarize(diffInUse = (max(year) - min(year)))

namesNotInUseLong <- longestTimeDiff %>% filter(diffInUse == max(longestTimeDiff$diffInUse)) %>% select
```

> There are 1511 names which were not in use for longest time(135 years) and they are listed in *namesNotInUseLong* variable.

(d) For each year, what is the total number of names that were recorded? Treat boy and girl versions of the same name as two separate names. Did you need to look at the data to answer this question?

```
# EDIT ME
namesCount <- babynamesData %>% group_by(year, sex) %>%
  dplyr::summarise(count = n())

namesCountByYear <- namesCount %>% group_by(year) %>% dplyr::summarize(total = sum(count))
```

> Yes, for any calculation we need to look at the data first. And after each step the structure of the data is to be analysed for proceeding further.