INFX 573 Lab: Data Wrangling

Charudatta Deshpande

Collaborators: Manjiri Kharkar, Ram Ganesan

(don't forget to list the names of your collaborators!)

Instructions:

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

1. Download the `lab04b-explore-clean.rmd` file from Canvas. Open it in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week3a_lab.Rmd`. Download also the `weather.txt` data file.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
3. 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.
4. 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:

```
# Load some helpful libraries
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.

```
# Following code will read the weather.txt dataset. I prefer to use
#data.table package for all table operations.
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

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

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

Hint: These figures originate from Cuernavaca, Mexico, 2010. You may want to consult tutiempo website.

Hint 2: There are five variables of interest here.

```
# Description of variables -
# id - This is likely the identifier of the weather station that supplied the reading.
# year - year the temperature was noted. Value '2010' for this dataset.
# month - month the temperature was noted. value ranges from 1 to 12, except 9.
# the records for month 9 (september) are missing.
# element - This field identifies the type of data represented by this record.
# For this dataset, the values are TMAX (maximum temperature of the day) and
# TMIN (minimum temperature of the day). Here, the decimal point is missing which
# indicates a problem with the data and needs to be corrected.
# d1 through d31 - Day of the month specified by the 'month' field. Each day is
# a separate field on the dataset.
```

(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:

- mutate()
- gather()
- spread()

```
# The below code will gather the weather dataset by converting d1 thru d31 variables
# in a single variable 'day' and values into single variable 'temp'.
weather <- weather %>%
           gather(d1:d31, key="day", value="temp")
#
# The below code will spread the weather table so that TMIN and TMAX are new fields.
#
weather <- weather %>%
          spread(key="element", value="temp")
#
# The below code will mutate the TMIN and TMAX fields to divide them by 10.
# This will show appropriate value for the temperature.
# The calculation for NA values will be ignored. Those values will stay NA.
weather <- weather %>%
           mutate (TMIN = (TMIN/10), TMAX = (TMAX/10))
#
# The above process creates a tidy version of the dataset.
# There are still rows with TMIN and TMAX as NA, but they can be easily removed if needed.
#
```

Problem 2: Data Manipulation

In this problem we will use the `babynames` data. Use the data to answer the following questions.

(a) What name has been used for the most number of years

(consider boy/girl names as distinct names)

```
# The below code will load the dataset babynames.
data(babynames)
#
# The below code will create a new dataset t which will include sex, name and count of
# 'sex, name' combination. Only 1 row will be kept for each 'name, sex' combo.
t <- babynames %>%
    select(sex, name) %>%
    group_by(name, sex) %>%
    mutate(count=n()) %>%
    filter(row_number(count) == 1) %>%
    arrange(desc(count))
# Below code will store rows with maximum number of counts
answer1 <- t %>% filter(count == 136)
head(answer1)
```

```
## # A tibble: 6 x 3
## # Groups:   name, sex [6]
##     sex      name count
##   <chr>     <chr> <int>
## 1     F      Mary   136
## 2     F      Anna   136
## 3     F      Emma   136
## 4     F Elizabeth   136
## 5     F    Minnie   136
## 6     F  Margaret   136
```

```
#
# 136 years is the longest time any name can be used. The question asks what names
# have been in use for most number of years, i,e, 136 years.
# Viewing answer1 dataset, there are 933 names that have been used for 136 years.
# Male and Female names have been treated as unique.
# Some of the names are - Mary, Anna, Emma, John, William etc. It is not
# possible to list 933 names.
```

(b) For each name, what year was it most popular (measured as the percentage of names for a given year)?

(consider boy and girl names as distinct)?

```
# Below code will select the record for the largest 'prop' for a given
# name and sex combination.
t1 <- babynames %>%
    group_by(name, sex) %>%
    filter(prop == max(prop))
# Answer 2- Below code will only select name sex, name and year from the t1
# dataset.
answer2 <- t1 %>%
           select (year, sex, name)
# Viewing answer2 dataset provides the answer for the question.
head(answer2)
```

```
## # A tibble: 6 x 3
## # Groups:   name, sex [6]
##    year   sex      name
##   <dbl> <chr>     <chr>
## 1  1880     F      Mary
## 2  1880     F Elizabeth
## 3  1880     F    Minnie
## 4  1880     F       Ida
## 5  1880     F     Alice
## 6  1880     F      Ella
```

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

```
# Below code will create a dataset t2, with largest year for a name/sex
# combination. And dataset will be ordered by year.
t2 <- babynames %>%
    group_by(name, sex) %>%
    filter(year == max(year)) %>%
    arrange(year)
head(t2)
```

```
## # A tibble: 6 x 5
## # Groups:   name, sex [6]
##    year   sex   name      n         prop
##   <dbl> <chr>  <chr> <int>        <dbl>
## 1  1880     M Merida     5 4.223009e-05
## 2  1881     F Zilpah     9 9.104244e-05
## 3  1881     M   Roll     5 4.617573e-05
## 4  1882     F  Crete     8 6.914673e-05
## 5  1883     F  Franc     5 4.164619e-05
## 6  1885     F  Lelie     5 3.522392e-05
```

```
#
# Answer - Looking at first 2 rows of dataset, following names have been
# out of use for the most time.
# Male name Merida - last used in 1880
# Female name - Zilpah - last used in 1881
#
```

(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?

```
# Below code will group the dataset by year and sex and count the
# number of names for each sex for each year.
t3 <- babynames %>%
    group_by(year, sex) %>%
    mutate(count=n()) %>%
    filter(row_number(count) == 1) %>%
    select(year, sex, count) %>%
```

```
    arrange(year)
head(t3)
```

```
## # A tibble: 6 x 3
## # Groups:   year, sex [6]
##    year   sex count
##   <dbl> <chr> <int>
## 1  1880     F   942
## 2  1880     M  1058
## 3  1881     F   938
## 4  1881     M   997
## 5  1882     F  1028
## 6  1882     M  1099
```

```
# Viewing dataset t3 gives the details of each year and each sex, and the count
# of names for each year/sex combination.
```