INFX 573: Problem Set 2 - Data Wrangling

Pamela Chakrabarty

Due: Monday, October 18, 2016

Collaborators: None but consulted on a question or two with Namrata Deshpande

Instructions:

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

1. Download the `problemset2.Rmd` file from Canvas. Open `problemset2.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset2.Rmd`.
2. Replace the "Insert 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.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
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 PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps2.Rmd`, knit a PDF and submit the PDF file on Canvas.

Setup:

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

```
# Load standard libraries
library(tidyverse)
library(nycflights13)
library(jsonlite)
```

Problem 1: Open Government Data

Use the following code to obtain data on the Seattle Police Department Police Report Incidents.

```
# importing and loading the dataset from the url data.seattle.gov
police_incidents <- fromJSON("https://data.seattle.gov/resource/7ais-f98f.json")
data(police_incidents)
```

```
## Warning in data(police_incidents): data set 'police_incidents' not found
```

(a) Describe, in detail, what the data represents.

Answer: The dataset "police_incidents" if based off of offense reports in the city of Seattle recorded by the police department. It is recorded on the Seattle Police Department Records Management System and stored on the website data.seattle.gov. The data set has total 19 variables and 1000 observations.

```
# Tells us the number of observations(rows) and variables(columns) of the current dataset
dim(police_incidents)
```

```
## [1] 1000   19
```

Observations: Observation 1: On applying the str function, we can see that the location variable has 3 sub variable under it. We do not need this location variable as we already have the variables "latitude" and "longitude" in out police_incidents dataset. So, we can write a script to discard this redundancy by:

```
# Cleaning the dataset by creating a subset of the orginal dataset to drop the location dataframe .
police_incidents = subset(police_incidents, select = -location)
```

(b) Describe each variable and what it measures. Be sure to note when data is missing. Confirm that each variable is appropriately cast - it has the correct data type. If any are incorrect, recast them to be in the appropriate format.

Answer: The following variables are : 1) offense_code_extension is a categorical variable that denotes the extension code for an offense 2) offense_type is a categorical variable denotes the type of offense 3) general_offense_number is a categorical variable that denotes a unique 10 digit number for each offense 4) offense_code is a categorical variable that denotes a 4 digit number code for a specfic ofense type 5) rms_cdw_id is a numerical variable that denotes a unqiue id for each incident report recored in the Record Management systems(rms) 6) census_tract_2000 is a categorical variable that denotes the census tract when the offense occured 7) zone_beat is a categorical variable denoting the zone of the incident's location 8) latitude denotes the latitude of the offense occurence 9) summarized_offense_description is a categorical variable that describes an offense 10) date_reported denotes the date when the incident was reported 11) occurred_date_or_date_range_start denotes the start date of an offense incident 12) summary_offense_code is a categorical variable that denotes the 13) year, month denotes the year and month when the offense occured 14) district_sector is a categorical variable that denotes the district sector of an offense 15) latitude, longitude are numercial variables denoting the 16)hundred_block_location denotes extent of distance of upto 100th block from incident location 17) occurred_date_or_date_range_start denotes the start date of a reported offense 18) occurred_date_range_end denotes the end date of a reported offense

Further Observations:

```
#Shows the data types of each variable of the dataset
str(police_incidents)
```

```
## 'data.frame':    1000 obs. of  18 variables:
##  $ offense_code_extension          : chr  "0" "0" "1" "0" ...
##  $ offense_type                    : chr  "EQUALS" "ASSLT-NONAGG" "VEH-THEFT-AUTO" "THEFT-SHOPLIFT"
##  $ general_offense_number          : chr  "2016239258" "2016340018" "2016340045" "2016339816" ...
##  $ offense_code                    : chr  "2903" "1313" "2404" "2303" ...
##  $ rms_cdw_id                      : chr  "949463" "1038931" "1038930" "1038854" ...
##  $ year                            : chr  NA "2016" "2016" "2016" ...
##  $ zone_beat                       : chr  NA "E2" "U1" "L3" ...
##  $ latitude                        : chr  NA "47.615837097" "47.667503357" "47.721984863" ...
##  $ summarized_offense_description  : chr  NA "ASSAULT" "VEHICLE THEFT" "SHOPLIFTING" ...
```

```
##  $ date_reported                    : chr  NA "2016-09-19T14:25:00" "2016-09-19T13:21:00" "2016-09-19
##  $ occurred_date_or_date_range_start: chr  NA "2016-09-19T13:00:00" "2016-09-18T15:00:00" "2016-09-19
##  $ summary_offense_code             : chr  NA "1300" "2400" "2300" ...
##  $ month                            : chr  NA "9" "9" "9" ...
##  $ census_tract_2000                : chr  NA "7500.5009" "4400.4003" "100.5005" ...
##  $ hundred_block_location           : chr  NA "16XX BLOCK OF 11 AV" "52XX BLOCK OF 12 AV NE" "127XX
##  $ district_sector                  : chr  NA "E" "U" "L" ...
##  $ longitude                        : chr  NA "-122.318168640" "-122.315200806" "-122.293640137" ...
##  $ occurred_date_range_end          : chr  NA NA "2016-09-19T13:00:00" NA ...
```

Observation 2: 2)Not all variables are having the correct data types, a couple of few that I noticed were year, month and latitude and I have converted them to relevant data types in my understanding. For instance, the variable latitude has decimal values so it should be a numeric value.

Recasting some of the variables of the dataset:

```
# Converting the variables year and month each from character to date data type
police_incidents$year <- as.Date(police_incidents$year, format="%Y")
police_incidents$month <- as.Date(police_incidents$month, format="%m")
#Converting the variable latitude into a numeric datatype
police_incidents$latitude <- as.numeric(police_incidents$latitude)
```

(c) Produce a clean dataset, according to the rules of tidy data discussed in class. Export the data for future analysis using the Rdata format.

Answer:

```
# Exporting the data for future use means that we need to save our cleaned dataset in RData format for
cleaned_policeincidents <- police_incidents
save(cleaned_policeincidents, file="cleaned_policeincidents.RData")
```

(d) Describe any concerns you might have about this data. This may include biases, missing data, or ethical concerns.

Answer: Some of the possible concerns that comes to mind could be: The data might be biased, meaning might be a possibility of minor offenses not being recorded or might be missing out on any offense incidents that were not reported to the police or not reported on the system. The first obervation of the dataset are having multiple NA values and this might lead someone to discard this row in their data analysis. Also, many of the values are marked as "X". But we dont know what X means, is it same as NA or is something different so that could lead us to make wrong assumptions about the data. I also see many of the values are missing for variables like occured_date and offense_code (does it mean that all there might be more offenses that occurred but were not reported as they did not suit the previous codes?).

Problem 2: Wrangling the NYC Flights Data

In this problem set we will use the data on all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013. You can find this data in the `nycflights13` R package.

(a) Importing Data:

Load the data.

```
library(tidyverse)
#Loading the nycflights dataset
library(nycflights13)
```

(b) Data Manipulation:

Use the flights data to answer each of the following questions. Be sure to answer each question with a written response and supporting analysis.

- How many flights were there from NYC airports to Seattle in 2013?

Answer: There are total 3923 flights from NYC airports to Seattle in 2013.

To answer this question, I have filtered the nycflights13 dataset to narrow down the dataset with only observations recorded for the variable(column) "dest" where dest is Seattle (the value of which is referred by the string "SEA"). Next, I have used the piping function to combine the filtered dataset with the count of the total number of flights

```
# Filtering the flights with Seattle as destination and then summarizing the dataset by counting the to
filter(flights, dest=="SEA") %>%
summarize(number_of_flights = n())
```

```
## # A tibble: 1 × 1
##   number_of_flights
##               <int>
## 1              3923
```

- How many airlines fly from NYC to Seattle?

Answer: There were total 5 airlines flying from NYC to Seatlle

```
#
filter(flights, dest=="SEA") %>%
group_by(carrier) %>%
summarize(number_of_airlines = n())
```

```
## # A tibble: 5 × 2
##   carrier number_of_airlines
##     <chr>              <int>
## 1      AA                365
## 2      AS                714
## 3      B6                514
## 4      DL               1213
## 5      UA               1117
```

- How many unique air planes fly from NYC to Seattle?

Answer: There were 936 unique airplaines flting from NYC to Seattle

```
# Filtering the dataset to flights with Seattle as destination and piping the result with using the dis
filter(flights, dest=="SEA")  %>%
distinct(tailnum)
```

```
## # A tibble: 936 × 1
##    tailnum
##      <chr>
## 1   N594AS
## 2   N3760C
## 3   N45440
## 4   N37464
## 5   N503JB
## 6   N77296
## 7   N553AS
## 8   N3ETAA
## 9   N3772H
## 10  N76523
## # ... with 926 more rows
```

- What is the average arrival delay for flights from NYC to Seattle?

Answer: The average arrival delay for flights flying from NYC to Seattle in 2013 was -1.099.

```
# Filtering dataset to flights flying from Seattle to NYC in 2013 and then summarizing the result based
flights %>%
filter(dest=="SEA")  %>%
summarise(avg_delay = mean(arr_delay, na.rm=TRUE))
```

```
## # A tibble: 1 × 1
##   avg_delay
##       <dbl>
## 1 -1.099099
```

- What proportion of flights to Seattle come from each NYC airport?

Answer: Around 46.67% of Seattle flights flew from the EWR airport and 53.32% of Seattle flights flew from JFK airport.

```
# Filtering dataset to flights flying from Seattle to NYC in 2013, grouping them by origin, then calcul
filter(flights, dest=="SEA") %>%
group_by(origin) %>%
summarise(total_flights = n()) %>%
mutate(prop_of_NYCflights= total_flights/sum(total_flights))
```

```
## # A tibble: 2 × 3
##   origin total_flights prop_of_NYCflights
##    <chr>         <int>              <dbl>
## 1    EWR          1831          0.4667346
## 2    JFK          2092          0.5332654
```