IMT 573: Problem Set 4

Working with Data: Part II

Jenny Skytta

Due: April 24, 2022

Collaborators: Independent work shared code with Jeff Ingham

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. Download the `04_ps_workingdatatwo.Rmd` file from Canvas or save a copy to your local directory on RStudio Cloud. Supply your solutions to the assignment by editing `04_ps_workingdatatwo.Rmd`.
2. 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 we 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. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object don't exist
# if you run this on its own it with give an error
```

7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, download and rename the knitted PDF file to `ps4_YourLastName_YourFirstName.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
install.packages("Rcpp", repos = "http://cran.us.r-project.org")
install.packages("tigris")
```

```
library(tidyverse)
library(lubridate)
library(tigris)
library(tidycensus)
library(stringr)
library(knitr) # this will keep code on the page!
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Problem 1: Joining Census Data to Police Reports In this problem set, we will be joining disparate sets of data - namely: Seattle police crime data, information on Seattle police beats, and education attainment from the US Census. Our goal is to build a dataset where we can examine questions around crimes in Seattle and the educational attainment of people living in the areas in which the crime occurred; this requires data to be combined from these two individual sources.

As a general rule, be sure to keep copies of the original dataset(s) as you work through cleaning (remember data provenance!).

(a) Importing and Inspecting Crime Data Load the Seattle crime data from the provided `crime_data.csv` data file. You can find more information on the data here: https://data.seattle.gov/Public-Safety/Crime-Data/4fs7-3vj5. This dataset is constantly refreshed online so we will be using the provided csv file for consistency. We will call this dataset the "Crime Dataset." Perform a basic inspection of the Crime Dataset and discuss what you find.

```
# loading the crime_data csv into global environment with
# variable name
crime_data <- read.csv("data/crime_data.csv", stringsAsFactors = FALSE,
    na.strings = "")
# View(crime_data) # inspect the table
ls(crime_data)  # listing the variables of the table
```

```
##  [1] "Beat"                          "Crime.Subcategory"
##  [3] "Neighborhood"                  "Occurred.Date"
##  [5] "Occurred.Time"                 "Precinct"
##  [7] "Primary.Offense.Description" "Report.Number"
##  [9] "Reported.Date"                 "Reported.Time"
## [11] "Sector"
```

```
summary(crime_data)
```

```
##  Report.Number       Occurred.Date      Occurred.Time  Reported.Date
##  Min.   :2.008e+08   Length:523591      Min.   :   0   Length:523591
##  1st Qu.:2.008e+13   Class :character   1st Qu.: 900   Class :character
##  Median :2.012e+13   Mode  :character   Median :1500   Mode  :character
##  Mean   :1.635e+13                      Mean   :1359
##  3rd Qu.:2.016e+13                      3rd Qu.:1920
##  Max.   :2.019e+13                      Max.   :2359
##                                         NA's   :2
##  Reported.Time  Crime.Subcategory  Primary.Offense.Description
##  Min.   :   0   Length:523591      Length:523591
##  1st Qu.: 950   Class :character   Class :character
##  Median :1407   Mode  :character   Mode  :character
##  Mean   :1353
```

```
##  3rd Qu.:1817
##  Max.   :2359
##  NA's   :2
##    Precinct            Sector              Beat           Neighborhood
##  Length:523591      Length:523591      Length:523591      Length:523591
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

This dataset appears to be a dataset of crimes in the Seattle areas from 1973 to 2018 capturing event and report times and dates, crime categories, neighborhoods, report numbers, police precinct and beat. I'm unfamiliar with the term beat in this context but in reviewing online, it appears to be another characterization of geographic territory.

(b) Looking at Years That Crimes Were Committed Let's start by looking at the years in which crimes were committed. What is the earliest year in the dataset? Are there any distinct trends with the annual number of crimes committed in the dataset?

```
time_crime <- crime_data %>%
  mutate(Occurred.Date=as.Date(Occurred.Date, format = "%m/%d/%Y")) %>% #set observations to date class
  drop_na(Occurred.Date) %>% # drop the multiple NAs from the column
  summarise(min(Occurred.Date)) #pull the earliest date

crime_occurred <- crime_data %>%
  mutate(Occurred.Date=as.Date(Occurred.Date, format = "%m/%d/%Y")) %>% #set observations to date class
  group_by(year = lubridate::floor_date(Occurred.Date, "year")) %>% #group by year
  tally() #total

crime_reports <- crime_data %>%
  mutate(Reported.Date=as.Date(Reported.Date, format = "%m/%d/%Y")) %>% #set observations to date class
  group_by(year = lubridate::floor_date(Reported.Date, "year")) %>% #group by year
  tally() #total

annual_post08 <- crime_data %>%
  mutate(Occurred.Date=as.Date(Occurred.Date, format = "%m/%d/%Y")) %>% #set observations to date class
  group_by(year = lubridate::floor_date(Occurred.Date, "year")) %>% #group by year
  tally() %>% #total
  filter(n>400) %>% # filter greater than 400
  drop_na(year) %>% # remove the NA values in year column
  summarise(average = mean(n)) #average of tally

annual_pre08 <- crime_data %>%
  mutate(Occurred.Date=as.Date(Occurred.Date, format = "%m/%d/%Y")) %>% #set observations to date class
  group_by(year = lubridate::floor_date(Occurred.Date, "year")) %>% #group by year
  tally() %>% # total
  filter(n<400) %>% # filter less than 400
  drop_na(year) %>% # remove NA values in year column
  summarise(average = mean(n)) #average of tally

#View(annual_pre08)
```

The earliest date appears to be December 13th 1908. There is a larger record of crimes that occurred versus reported with 2008 representing a giant uptick in both records of occurrence and reporting. Before 2008, we see and average of just 13 crimes occurring in the dataset. Beginning in 2008, there is an average of 40244 crimes occurring (excluding NA values where no year is recorded).

(c) Looking at Frequency of Beats What is a Police Beat? How frequently are the beats in the Crime Dataset listed? Are there any anomolies with how frequently some of the beats are listed? Are there missing beats?

```
# View(cd_beats)

cd_beats <- crime_data %>%
    group_by(Beat) %>%
    arrange(desc(Beat)) %>%
    tally()
```

From the crime_data dataset, there appear to be 64 named Beats and one NA included. Beats represent a geographic area within a neighborhood. Some of the Beats appear to be underrepresented in the data with fewer than 10 incidents associated. Its difficult to ascertain if there are missing Beats as I'm not finding an exact number of expected Beats.

(d) Importing Police Beat Data and Filtering on Frequency Load the data on Seattle police beats provided in `police_beat_and_precinct_centerpoints.csv`. You can find additional information on the data here: (https://data.seattle.gov/Land-Base/Police-Beat-and-Precinct-Centerpoints/4khs-fz35). We will call this dataset the "Beats Dataset."

```
Beats_DS <- read.csv("data/police_beat_and_precinct_centerpoints.csv",
    stringsAsFactors = FALSE, na.strings = "")
# loading Beats Dataset

# View(Beats_DS) #Viewing Beats
```

Does the Crime Dataset include police beats that are not present in the Beats Dataset? If so, how many and with what frequency do they occur? Would you say that these comprise a large number of the observations in the Crime Dataset or are they rather infrequent? Do you think removing them would drastically alter the scope of the Crime Dataset?

There are 64 Beats included in the crime_data dataset and 57 listed in the Beats dataset, accounting for a difference of 7 identified extra Beats in the crime data and 1 NA.

Let's remove all instances in the Crime Dataset that have beats which occur fewer than 10 times across the Crime Dataset. Also remove any observations with missing beats. After only keeping years of interest and filtering based on frequency of the beat, how many observations do we now have in the Crime Dataset?

```
Beats_filter <- crime_data %>% # creating variable
  group_by(Beat) %>% # grouping by Beats
  tally() %>%  # tallying Beats
  filter(n > 10) %>%  #filtering over 10 Beats tally
  drop_na() # removing the NA (missing Beats)

#View(Beats_filter)
```

After filtering to remove the NA missing Beats, and those with less than 10 across the dataset, we are left with 51 Beats within the crime_data dataset.

(e) Importing and Inspecting Police Beat Data To join the Beat Dataset to census data, we must have census tract information. Use the `censusr` package to extract the 15-digit census tract for each police beat using the corresponding latitude and longitude. Do this using each of the police beats listed in the Beats Dataset. Do not use a for-loop for this but instead rely on R functions (e.g. the 'apply' family of functions). Add a column to the Beat Dataset that contains the 15-digit census tract for the each beat. (HINT: you may find `censusr`'s `call_geolocator_latlon` function useful)

```
# load census data into census_data variable
Census_data <- read.csv("data/census_edu_data.csv", stringsAsFactors = FALSE,
    na.strings = "")

# add column using apply with the call_geolocator_latlon
# function to calculate census tract
Beats_DS$census_code <- apply(Beats_DS, 1, function(row) call_geolocator_latlon(row["Latitude"],
    row["Longitude"]))

# View(Beats_DS)
```

We will eventually join the Beats Dataset to the Crime Dataset. We could have joined the two and then found the census tracts for each beat. Would there have been a particular advantage/disadvantage to doing this join first and then finding census tracts? If so, what is it? (NOTE: you do not need to write any code to answer this)

(f) Extracting FIPS Codes Once we have the 15-digit census codes, we will break down the code based on information of interest. You can find more information on what these 15 digits represent here: https://transition.fcc.gov/form477/Geo/more_about_census_blocks.pdf.

First, create a column that contains the state code for each beat in the Beats Dataset. Then create a column that contains the county code for each beat. Find the FIPS codes for WA State and King County (the county of Seattle) online. Are the extracted state and county codes what you would expect them to be? Why or why not?

```
# create a column in the Beats dataset with the state
# census code
Beats_DS$state_census_code <- substr(Beats_DS$census_code, 1,
    2)
# create a column in the Beats dataset with the county
# census code
Beats_DS$county_census_code <- substr(Beats_DS$census_code, 3,
    5)
# adding table to Rmd report
kable(head(Beats_DS), caption = "Beats dataset with Census Codes")
```

Table 1: Beats dataset with Census Codes

Name	Location.1	Latitude	Longitude	census_code	state_census_code	county_census_code
B1	(47.7097756394592, -122.370990523069)	47.70978	-122.3710	530330014004003	53	033
B2	(47.6790521901374, -122.391748391741)	47.67905	-122.3918	530330032021003	53	033
B3	(47.6812920482227, -122.364236159741)	47.68129	-122.3642	530330029003015	53	033
C1	(47.6342500180223, -122.315684762418)	47.63425	-122.3157	530330065001015	53	033

Name	Location.1	Latitude	Longitude	census_code	state_census_code	county_census_code
C2	(47.6192385752996, -122.313557430551)	47.61924	-122.3136	53033007502200	53	033
C3	(47.6300792887474, -122.292087128251)	47.63008	-122.2921	53033006300200	53	033

The FIPS code for Washington state is 53, while the King County FIPS code is 53033. The WA state census county code is 033 with the state census code being 53. These appear to match the FIPS codes.

(g) Extracting 11-digit Codes The census data uses an 11-digit code that consists of the state, county, and tract code. It does not include the block code. To join the census data to the Beats Dataset, we must have this code for each of the beats. Extract the 11-digit code for each of the beats in the Beats Dataset. The 11 digits consist of the 2 state digits, 3 county digits, and 6 tract digits. Add a column with the 11-digit code for each beat.

```
# adding the tract code to the Beats dataset
Beats_DS$tract_census_code <- substr(Beats_DS$census_code, 6,
    11)
# adding the eleven digit codes to the Beats dataset
Beats_DS$eleven_dig <- substr(Beats_DS$census_code, 1, 11)
```

(h) Extracting 11-digit Codes From Census Now, we will examine census data (`census_edu_data.csv`). The data includes counts of education attainment across different census tracts. Note how this data is in a 'wide' format and how it can be converted to a 'long' format. For now, we will work with it as is.

The census data contains a "GEO.id" column. Among other things, this variable encodes the 11-digit code that we had extracted above for each of the police beats. Specifically, when we look at the characters after the characters "US" for values of GEO.id, we see encodings for state, county, and tract, which should align with the beats we had above. Extract the 11-digit code from the GEO.id column. Add a column to the census data with the 11-digit code for each census observation.

```
# added the eleven digit column to the Census dataframe
Census_data$eleven_dig <- substr(Census_data$GEO.id, 10, 21)
# this data appears to already be captured in GEO.id2
```

(i) Join Datasets Join the census data with the Beat Dataset using the 11-digit codes as keys. Be sure that you do not lose any of the police beats when doing this join (i.e. your output dataframe should have the same number of rows as the cleaned Beats Dataset - use the correct join). Are there any police beats that do not have any associated census data? If so, how many?

```
# left join the Beats and Census datasets on the eleven digit column
Beats_Census <- left_join(Beats_DS, Census_data, by = "eleven_dig")

Missing_census <- Beats_Census %>%
  replace(is.na(.), 0) %>%  # replaced NA with 0
  select(Name, GEO.display.label, GEO.id) %>%
  arrange(-desc(GEO.display.label)) #count the observations without census data

#View(Missing_census)
```

There are 24 police beats that do not have any associated census data.

Then, join the Crime Dataset to our joined beat/census data. We can do this using the police beat name. Again, be sure you do not lose any observations from the Crime Dataset. What is the final dimensions of the joined dataset?

```
Beats_Census <- Beats_Census %>%
    rename(Beat = Name)  #renamed column for join

# right joined the Beats_Census and crime_data datasets by
# Beat.
Beat_Census_Crime <- right_join(Beats_Census, crime_data, by = "Beat")
```

The final dimensions are 523581 observations with 45 variables.

Once everything is joined, save the final dataset for future use.

Citations

Retrieve Census tract from Coordinates [closed] https://stackoverflow.com/questions/51499410/retrieve-census-tract-from-coordinates

More About Census Blocks https://transition.fcc.gov/form477/Geo/more_about_census_blocks.pdf

Code written above is from the previous course IMT 511 which used the below text to support class scripts. https://www.google.com/books/edition/Programming_Skills_for_Data_Science/BnB6DwAAQBAJ?hl=en&gbpv=1&printsec=frontcover