IMT 573: Problem Set 4

Working with Data: Part II

Tommy Huynh

Due: July 11, 2021

Collaborators:

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
library(tidyverse)
```

```
## Warning: package ’tidyr’ was built under R version 4.1.1

## Warning: package ’readr’ was built under R version 4.1.1
```

```
library(censusr)
```

```
## Warning: package ’censusr’ was built under R version 4.1.1
```

```
library(stringr)
library(tidycensus)
```

```
## Warning: package ’tidycensus’ was built under R version 4.1.1
```

```
library(tigris)
```

```
## Warning: package ’tigris’ was built under R version 4.1.1
```

```
library(dplyr)

options(tigris_use_cache = 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.

```
crime_dataset <- read_csv("crime_data.csv")
```

```
## Rows: 523591 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): Occurred Date, Reported Date, Crime Subcategory, Primary Offense De...
## dbl (3): Report Number, Occurred Time, Reported Time

##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

(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?

```
#This code shows the first couple rows of the dataset
head(crime_dataset)
```

```
## # A tibble: 6 x 11
##   ‘Report Number‘ ‘Occurred Date‘ ‘Occurred Time‘ ‘Reported Date‘
##             <dbl> <chr>                     <dbl> <chr>
## 1         1.98e12 12/16/1975                  900 12/16/1975
## 2         1.98e12 01/01/1976                    1 01/31/1976
## 3         1.98e12 01/28/1979                 1600 02/09/1979
## 4         1.98e13 08/22/1981                 2029 08/22/1981
## 5         1.98e12 02/14/1981                 2000 02/15/1981
## 6         1.99e13 09/29/1988                  155 09/29/1988
## # ... with 7 more variables: Reported Time <dbl>, Crime Subcategory <chr>,
## #   Primary Offense Description <chr>, Precinct <chr>, Sector <chr>,
## #   Beat <chr>, Neighborhood <chr>
```

According to the dataframe, the earliest year in the dataset is 1975.It appears that as the years go by, the more case reports there are.

(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?

A police beat is a territory that police officers patrol. It appears that there are almost always a police beat listed with each case. There are some beats that are missing.

(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."

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?

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?

```r
police_beats <- read_csv("police_beat_and_precinct_centerpoints.csv")
```

```
## Rows: 57 Columns: 4

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Name, Location 1
## dbl (2): Latitude, Longitude

##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
crime_dataset[crime_dataset$Beat %in% names(which(table(crime_dataset$Beat) > 10)), ]
```

```
## # A tibble: 520,252 x 11
##    ‘Report Number‘ ‘Occurred Date‘ ‘Occurred Time‘ ‘Reported Date‘
##              <dbl> <chr>                     <dbl> <chr>
##  1         1.98e12 12/16/1975                  900 12/16/1975
##  2         1.98e12 01/28/1979                 1600 02/09/1979
##  3         1.98e13 08/22/1981                 2029 08/22/1981
##  4         1.98e12 02/14/1981                 2000 02/15/1981
##  5         1.99e13 09/29/1988                  155 09/29/1988
##  6         1.99e13 10/08/1993                 2213 10/08/1993
##  7         1.99e13 06/08/1994                    0 06/12/1994
##  8         2.00e13 12/08/1996                 1130 12/08/1996
##  9         2.00e13 05/12/2000                 2330 05/14/2000
## 10         2.00e12 01/15/2001                 2310 01/15/2001
## # ... with 520,242 more rows, and 7 more variables: Reported Time <dbl>,
## #   Crime Subcategory <chr>, Primary Offense Description <chr>, Precinct <chr>,
## #   Sector <chr>, Beat <chr>, Neighborhood <chr>
```

```r
crime_dataset[!is.na(crime_dataset$Beat), ]
```

```
## # A tibble: 520,293 x 11
##    ‘Report Number‘ ‘Occurred Date‘ ‘Occurred Time‘ ‘Reported Date‘
##              <dbl> <chr>                     <dbl> <chr>
##  1         1.98e12 12/16/1975                  900 12/16/1975
##  2         1.98e12 01/28/1979                 1600 02/09/1979
##  3         1.98e13 08/22/1981                 2029 08/22/1981
##  4         1.98e12 02/14/1981                 2000 02/15/1981
##  5         1.99e13 09/29/1988                  155 09/29/1988
##  6         1.99e13 10/08/1993                 2213 10/08/1993
##  7         1.99e13 06/08/1994                    0 06/12/1994
##  8         2.00e13 12/08/1996                 1130 12/08/1996
##  9         2.00e13 05/12/2000                 2330 05/14/2000
## 10         2.00e12 01/15/2001                 2310 01/15/2001
## # ... with 520,283 more rows, and 7 more variables: Reported Time <dbl>,
## #   Crime Subcategory <chr>, Primary Offense Description <chr>, Precinct <chr>,
## #   Sector <chr>, Beat <chr>, Neighborhood <chr>
```

There are now 520,252 observations

(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)

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)

```r
replacement_function <- function (lat, lon, benchmark, vintage)
{
  if (missing(benchmark)) {
    benchmark <- "Public_AR_Census2020"
  }
  else {
    benchmark <- benchmark
  }
  if (missing(vintage)) {
    vintage <- "Census2020_Census2020"
  }
  else {
    vintage <- vintage
  }
  call_start <- "https://geocoding.geo.census.gov/geocoder/geographies/coordinates?"
  url <- paste0("x=", lon, "&y=", lat)
  benchmark0 <- paste0("&benchmark=", benchmark)
  vintage0 <- paste0("&vintage=", vintage, "&format=json")
  url_full <- paste0(call_start, url, benchmark0, vintage0)
  r <- httr::GET(url_full)
  httr::stop_for_status(r)
  response <- httr::content(r)
  return(response$result$geographies$`Census Blocks`[[1]]$GEOID)
  if (length(response$result$geographies$`2020 Census Blocks`[[1]]$GEOID) ==
      0) {
    message(paste0("Lat/lon (", lat, ", ", lon, ") returned no geocodes. An NA was returned."))
    return(NA_character_)
  }
  else {
    if (length(response$result$geographies$`2020 Census Blocks`[[1]]$GEOID) >
        1) {
      message(paste0("Lat/lon (", lat, ", ", lon, ") returned more than geocode. The first match was re
    }
    return(response$result$geographies$`2020 Census Blocks`[[1]]$GEOID)
  }
}

police_beats$census_id <- with(police_beats, replacement_function(police_beats$Latitude[1], police_beat
```

Could not get call_geolocator_latlon to work. Keeps return 400 error. ##### (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?

```r
police_beats$state_code <- substr(police_beats$census_id, 1, 2)

police_beats$county_code <- substr(police_beats$census_id, 3, 5)
```

(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.

```r
police_beats$eleven_code <- substr(police_beats$census_id, 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.

```r
edu_data <- read.csv("census_edu_data.csv")

edu_data$eleven_code <- substr(edu_data$GEO.id, 10, 21)
```

(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?

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?

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

```r
merged_beats <- merge(edu_data, police_beats)
merged_census <- merge(merged_beats, crime_dataset)
```

There are 29,844,687 observations and 47 variables in the new dataseet.