IMT 573: Problem Set 3 - Working With Data II

Stephen V Tucker

October 28, 2022

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 Server.

1. Download the `problemset3.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problemset3.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset3.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. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
4. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. 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 dont' exist
# if you run this on its own it with give an error
```

6. 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 knitted PDF file to `ps3_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.
7. Collaboration is often fun and useful, 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.

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

```
# Load standard libraries
library('dplyr')
library('stringr')
library('httr')
library('jsonlite')
library("tidyverse")
```

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!).

```
crime_data_raw <- read.csv("~/Downloads/crime_data.csv")

# copy
crime_df <- crime_data_raw
```

(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 (note: the documentation for the provided dataset has been replaced by this newer documentation - most of the columns in the dataset map to columns in the documentation, but not the reverse). 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.

```
dim(crime_df)
```

```
## [1] 523591     11
```

```
str(crime_df)
```

```
## 'data.frame':    523591 obs. of  11 variables:
##  $ Report.Number              : num  1.98e+12 1.98e+12 1.98e+12 1.98e+13 1.98e+12 ...
##  $ Occurred.Date              : chr  "12/16/1975" "01/01/1976" "01/28/1979" "08/22/1981" ...
##  $ Occurred.Time              : int  900 1 1600 2029 2000 155 2213 0 1130 NA ...
##  $ Reported.Date              : chr  "12/16/1975" "01/31/1976" "02/09/1979" "08/22/1981" ...
##  $ Reported.Time              : int  1500 2359 1430 2030 435 155 2213 844 1700 NA ...
##  $ Crime.Subcategory          : chr  "BURGLARY-RESIDENTIAL" "SEX OFFENSE-OTHER" "CAR PROWL" "HOMICIDE
##  $ Primary.Offense.Description: chr  "BURGLARY-FORCE-RES" "SEXOFF-INDECENT LIBERTIES" "THEFT-CARPROWL
##  $ Precinct                   : chr  "SOUTH" "UNKNOWN" "EAST" "SOUTH" ...
##  $ Sector                     : chr  "R" "" "G" "S" ...
##  $ Beat                       : chr  "R3" "" "G2" "S2" ...
##  $ Neighborhood               : chr  "LAKEWOOD/SEWARD PARK" "UNKNOWN" "CENTRAL AREA/SQUIRE PARK" "BR
```

```
head(crime_df)
```

```
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 1     1.975e+12    12/16/1975           900    12/16/1975          1500
## 2     1.976e+12    01/01/1976             1    01/31/1976          2359
## 3     1.979e+12    01/28/1979          1600    02/09/1979          1430
## 4     1.981e+13    08/22/1981          2029    08/22/1981          2030
## 5     1.981e+12    02/14/1981          2000    02/15/1981           435
## 6     1.988e+13    09/29/1988           155    09/29/1988           155
##      Crime.Subcategory  Primary.Offense.Description  Precinct Sector Beat
## 1 BURGLARY-RESIDENTIAL           BURGLARY-FORCE-RES     SOUTH      R   R3
## 2    SEX OFFENSE-OTHER    SEXOFF-INDECENT LIBERTIES   UNKNOWN
## 3            CAR PROWL               THEFT-CARPROWL      EAST      G   G2
## 4             HOMICIDE HOMICIDE-PREMEDITATED-WEAPON     SOUTH      S   S2
## 5 BURGLARY-RESIDENTIAL           BURGLARY-FORCE-RES SOUTHWEST      W   W3
## 6  MOTOR VEHICLE THEFT               VEH-THEFT-AUTO      WEST      M   M2
##                    Neighborhood
```

```
## 1           LAKEWOOD/SEWARD PARK
## 2                        UNKNOWN
## 3       CENTRAL AREA/SQUIRE PARK
## 4                BRIGHTON/DUNLAP
## 5 ROXHILL/WESTWOOD/ARBOR HEIGHTS
## 6                    SLU/CASCADE
```

```
tail(crime_df)
```

```
##        Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 523586     2.019e+12    03/20/2019          1330    03/20/2019          1654
## 523587     2.019e+12    03/20/2019          1713    03/20/2019          1713
## 523588     2.019e+12    03/20/2019           730    03/20/2019          1721
## 523589     2.019e+12    03/20/2019          1724    03/20/2019          1724
## 523590     2.019e+12    03/20/2019          1750    03/20/2019          1904
## 523591     2.019e+12    03/19/2019          1800    03/20/2019          2237
##                Crime.Subcategory Primary.Offense.Description Precinct Sector
## 523586            THEFT-BUILDING              THEFT-BUILDING    NORTH      B
## 523587 FAMILY OFFENSE-NONVIOLENT                 CHILD-OTHER    SOUTH      O
## 523588      BURGLARY-RESIDENTIAL          BURGLARY-FORCE-RES     EAST      C
## 523589        ROBBERY-COMMERCIAL  ROBBERY-BUSINESS-BODYFORCE    SOUTH      S
## 523590            THEFT-SHOPLIFT              THEFT-SHOPLIFT    NORTH      L
## 523591           THEFT-ALL OTHER                   THEFT-OTH    NORTH      N
##        Beat         Neighborhood
## 523586   B2        PHINNEY RIDGE
## 523587   O3      MID BEACON HILL
## 523588   C2 MONTLAKE/PORTAGE BAY
## 523589   S2        RAINIER BEACH
## 523590   L2            NORTHGATE
## 523591   N1           BITTERLAKE
```

```
summary(crime_df)
```

```
##  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
##
##
```

##
##

(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?

Answer

The earliest year was in 1908 and in regards of trends i do not see anything sticking out thats obvious except for how far back the data goes 1908 seems rather far compared to 1970's.

```
# Step one, this creates a substring of just the year column for further data manipulation.
crime_year_subset <- group_by(crime_df, year = substr(crime_df$Occurred.Date,7,10))

# earliest  year is 1908
earliest_year <- summarise(crime_year_subset, count= n())
print(earliest_year)
```

```
## # A tibble: 46 x 2
##    year   count
##    <chr>  <int>
##  1 ""         2
##  2 "1908"     1
##  3 "1964"     1
##  4 "1973"     1
##  5 "1974"     1
##  6 "1975"     2
##  7 "1976"     2
##  8 "1977"     1
##  9 "1978"     1
## 10 "1979"     2
## # ... with 36 more rows
```

```
earliest_year[earliest_year == ""] <- NA

min(earliest_year$year, na.rm = TRUE)
```

```
## [1] "1908"
```

```
# mutate
crime_year_subset_mutate <- mutate(crime_df, year =substr(crime_df$Occurred.Date,7,10))

## I wasnt sure which dplyr method was most optimal so i decided to keep both ways incase anything happ
```

Subset the Crime Data-set to only include crimes that were committed after 2011 (remember good practices of data provenance!). Going forward, we will use this data subset in the analysis.

```
## This creates the actual subset of 2011
crime_year_subset_filter <- filter(crime_year_subset, year >= 2011)
head(crime_year_subset_filter)
```

```
## # A tibble: 6 x 12
## # Groups:   year [1]
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
##           <dbl> <chr>                <int> <chr>                 <int>
```

```
## 1       2.01e13 03/27/2011             2100 03/28/2011             1624
## 2       2.01e13 03/25/2011             1600 03/28/2011             1736
## 3       2.01e13 03/17/2011             1220 03/28/2011             1641
## 4       2.01e13 03/22/2011             1607 03/28/2011             1645
## 5       2.01e13 03/28/2011              345 03/28/2011             1704
## 6       2.01e13 03/28/2011              700 03/28/2011             1943
## # ... with 7 more variables: Crime.Subcategory <chr>,
## #   Primary.Offense.Description <chr>, Precinct <chr>, Sector <chr>,
## #   Beat <chr>, Neighborhood <chr>, year <chr>
```

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

Answer

- a police beat is the geo-spacial are wherepolice patrol that are marked out by block numbers and streets.
- The total number of beats are 523591
- According to the table there are missing beats 3298, anomalies that exist are the ones that have a significanly lower number than the others by comparison.

```
# sumamry
police_beat_df <- summarise(crime_df, Name = Beat)
# Table
police_beat_df <- table(police_beat_df)
# data-frame
police_beat_df <- data.frame(police_beat_df) %>%
  rename(police_beat_df, Name = police_beat_df)



# total frequency of beats
police_beat_count <- sum(police_beat_df$Freq)

print(police_beat_count)
```

```
## [1] 523591
```

(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?

Answer

- Based on findings the set does not include beats that are not present it has more geospatial data Longitude and Latitude

- In comparison to our police beat dataframe we have 65 observations, one being an empty string in comparison to our police_beat_coords dataframe where we have 57 observations. It appears that we have fewer beats identifiers in the coords data-frame.

```
# Loaded data
# police_beat_and_precinct_centerpoints.csv
police_beat_coords <- read.csv("~/Downloads/Police_Beat_and_Precinct_Centerpoints.csv")

beat_join <-left_join(police_beat_coords,police_beat_df, by = "Name")

# NA's - CITYWIDE, E, SE, SW,
```

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?

Answer

I wanted to differentiate the beats identifiers and the frequency for clarification purposes. Furthermore, in this problem set the question of how many observation can be perceived

```
## in doing this we reduced our dataframe beat observations from 65 to 52
police_beat_df_update <- filter(police_beat_df, Freq > 10)

## This steps coereces a empty string into a NA value
police_beat_df_update[police_beat_df_update == ""] <- NA

### This gets rid of NA value
police_beat_df_update <- police_beat_df_update %>%
  drop_na(Name)

## This gets the sum of observation of frequencies of the total beat identifiers.

## Sum of total beat identifiers / observations
observations <- length(police_beat_df_update$Name)

## sum of number of frequencies
frequencies <- sum(police_beat_df_update$Freq) # number of observations is 523550


## research: https://statisticsglobe.com/replace-blank-by-na-in-r
```

(e) Adding Census Codes to Police Beat Data To join the Beat Dataset to census data, we must have census tract information. Visit this page to the FCC's API that allows us to extract census tracts based on coordinates. We can use this API to provide a 15-digit census tract for each police beat using the corresponding latitude and longitude. Use the provided function (`get_census_code()`) to do this for 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). You are welcome to modify the `get_census_code` function to allow you to work with the appropriate `apply` function, but it is not necessary to do so. Add a column to the Beat Dataset that contains the 15-digit census tract for the each beat.

```
get_census_code <- function(lat, long){
  req <- paste0('https://geo.fcc.gov/api/census/area?lat=',
                lat,
                '&lon=',
```

```
                long,
                '&format=json')
  res <- GET(req)
  output <- fromJSON(rawToChar(res$content))
  output <- output$results
  output <- output$block_fips[1]
  return(output)
}

##
beat_tract <- mutate(police_beat_coords, census_tract = mapply(get_census_code, police_beat_coords$Lati

head(beat_tract )
```

```
##   Name                           Location.1 Latitude Longitude    census_tract
## 1   B1 (47.7097756394592, -122.370990523069) 47.70978 -122.3710 530330014004000
## 2   B2 (47.6790521901374, -122.391748391741) 47.67905 -122.3918 530330032021003
## 3   B3 (47.6812920482227, -122.364236159741) 47.68129 -122.3642 530330029003016
## 4   C1 (47.6342500180223, -122.315684762418) 47.63425 -122.3157 530330065001015
## 5   C2 (47.6192385752996, -122.313557430551) 47.61924 -122.3136 530330075022001
## 6   C3 (47.6300792887474, -122.292087128251) 47.63008 -122.2921 530330063002008
```

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)

Answer

The one thing im following along with is that the tract information has all beats and the updated on has been altered and I want to ensure my final dataset has the correct amount 51

- beats with missing values are below
- NA's - CITYWIDE, E, SE, SW,

(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?

```
## Created Column
beats_code_11_code <- mutate(beat_tract, state_code = substr(beat_tract$census_tract,1,2)) %>%
  mutate(beat_tract, county_code = substr(beat_tract$census_tract,3,5)) %>%   mutate(beat_tract, digita

head(beats_code_11_code)
```

```
##   Name                           Location.1 Latitude Longitude    census_tract
## 1   B1 (47.7097756394592, -122.370990523069) 47.70978 -122.3710 530330014004000
## 2   B2 (47.6790521901374, -122.391748391741) 47.67905 -122.3918 530330032021003
```

```
## 3   B3 (47.6812920482227, -122.364236159741) 47.68129 -122.3642 530330029003016
## 4   C1 (47.6342500180223, -122.315684762418) 47.63425 -122.3157 530330065001015
## 5   C2 (47.6192385752996, -122.313557430551) 47.61924 -122.3136 530330075022001
## 6   C3 (47.6300792887474, -122.292087128251) 47.63008 -122.2921 530330063002008
##   state_code county_code digital_code_11
## 1         53         033     53033001400
## 2         53         033     53033003202
## 3         53         033     53033002900
## 4         53         033     53033006500
## 5         53         033     53033007502
## 6         53         033     53033006300
```

```
## This question requested two seprate code
## FIPS - code King county: 53033 033
```

(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.

```
beats_code_11 <- mutate(beat_tract, state_code= substr(beat_tract$census_tract,1,2)) %>%
  mutate(beat_tract, county_code = substr(beat_tract$census_tract,3,5)) %>%     mutate(beat_tract, digi

# To keep code effecient i uses another pip to add the 11 digit code.
```

(h) Extracting 11-digit Codes From Census Now, we will examine census data provided in `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.

```
census_edu <- read.csv("~/Downloads/census_data_2020_edu_attainment.csv")
edu_code_11 <- mutate(census_edu, digital_code_11 = substr(census_edu$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?

```
## review
joined_data <- left_join(beats_code_11, edu_code_11, by = "digital_code_11")
```

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 are the final dimensions of the joined dataset?

```
joined_beat_census <- left_join(police_beat_df_update, joined_data, by = "Name")
```

Seeing what the final dataset looks like, what is an interesting question you could ask of this data and how would you go about answering it?

Interesting Question

An interesting question I have pertaining to the data Is how do we leverage this data-set to gain insights. What ideas could we come up with to optimize the data as well.

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