IMT 573: Problem Set 3 - Working With Data II

LEE CHEN HSIN

Due: Tuesday, October 26, 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 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.


```r
# Load standard libraries
library('dplyr')
library('censusr')
````
````

```
library('stringr')
library(tidyverse)
library(magrittr)
```

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.

```
CrimeDataset<-read.csv(file = '/Users/leechenhsin/Desktop/study@USA/07_UW_School/IMT573/crime_data.csv')

summary(CrimeDataset)
```

```
##  Report.Number          Occurred.Date    Occurred.Time     Reported.Date
##  Min.   :2.008e+08   07/01/2017:   199   Min.   :   0   12/31/2008:   238
##  1st Qu.:2.008e+13   05/26/2017:   193   1st Qu.: 900   03/31/2014:   196
##  Median :2.012e+13   01/20/2016:   186   Median :1500   10/10/2018:   196
##  Mean   :1.635e+13   12/01/2015:   184   Mean   :1359   06/18/2018:   195
##  3rd Qu.:2.016e+13   07/19/2018:   183   3rd Qu.:1920   05/12/2014:   193
##  Max.   :2.019e+13   11/25/2015:   182   Max.   :2359   07/05/2016:   193
##                      (Other)   :522464   NA's   :2      (Other)   :522380
##  Reported.Time            Crime.Subcategory      Primary.Offense.Description
##  Min.   :   0   CAR PROWL           :148263   THEFT-CARPROWL    :131297
##  1st Qu.: 950   THEFT-ALL OTHER     : 54420   THEFT-SHOPLIFT    : 48638
##  Median :1407   THEFT-SHOPLIFT      : 48638   THEFT-OTH         : 47276
##  Mean   :1353   BURGLARY-RESIDENTIAL: 46843   VEH-THEFT-AUTO    : 37840
##  3rd Qu.:1817   MOTOR VEHICLE THEFT : 43529   BURGLARY-FORCE-RES: 27984
##  Max.   :2359   BURGLARY-COMMERCIAL : 23531   THEFT-BUILDING    : 21438
##  NA's   :2      (Other)             :158367   (Other)           :209118
##       Precinct          Sector            Beat
##           :     6   M      : 42976   K3     : 16939
##  EAST     : 77475   U      : 40699   U1     : 14989
##  NORTH    :168392   K      : 38022   M1     : 14547
##  SOUTH    : 74426   B      : 37984   L2     : 14532
##  SOUTHWEST: 49332   D      : 35435   Q3     : 14329
##  UNKNOWN  :  3346   E      : 35038   M2     : 14238
##  WEST     :150614   (Other):293437   (Other):434017
##                Neighborhood
##  DOWNTOWN COMMERCIAL: 48942
##  NORTHGATE          : 30820
##  CAPITOL HILL       : 30735
##  QUEEN ANNE         : 27402
##  SLU/CASCADE        : 23343
##  UNIVERSITY         : 20868
##  (Other)            :341481
```

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

Occurredtime<-getmode(CrimeDataset$Occurred.Time)
Occurredtime
```

```
## [1] 2200
```

```
#For my observation, I found that the most time that crime event happened is at
#night like 22:00. Besides, the most happened crime subcategory is car prowl.
#In beat area, the most common area that beat happened is K3 and the most common
#neighborhood that crime event happened is downtown commercial.
```

(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?

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

```
Occurred.Year<- substr(CrimeDataset$Occurred.Date,7,10)
Occurred.Year=as.integer(Occurred.Year)
Occurred.Year=na.omit(Occurred.Year)
min(Occurred.Year)
```

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

```
subsetcrime<- na.omit(CrimeDataset)

subsetcrime$Year <- Occurred.Year

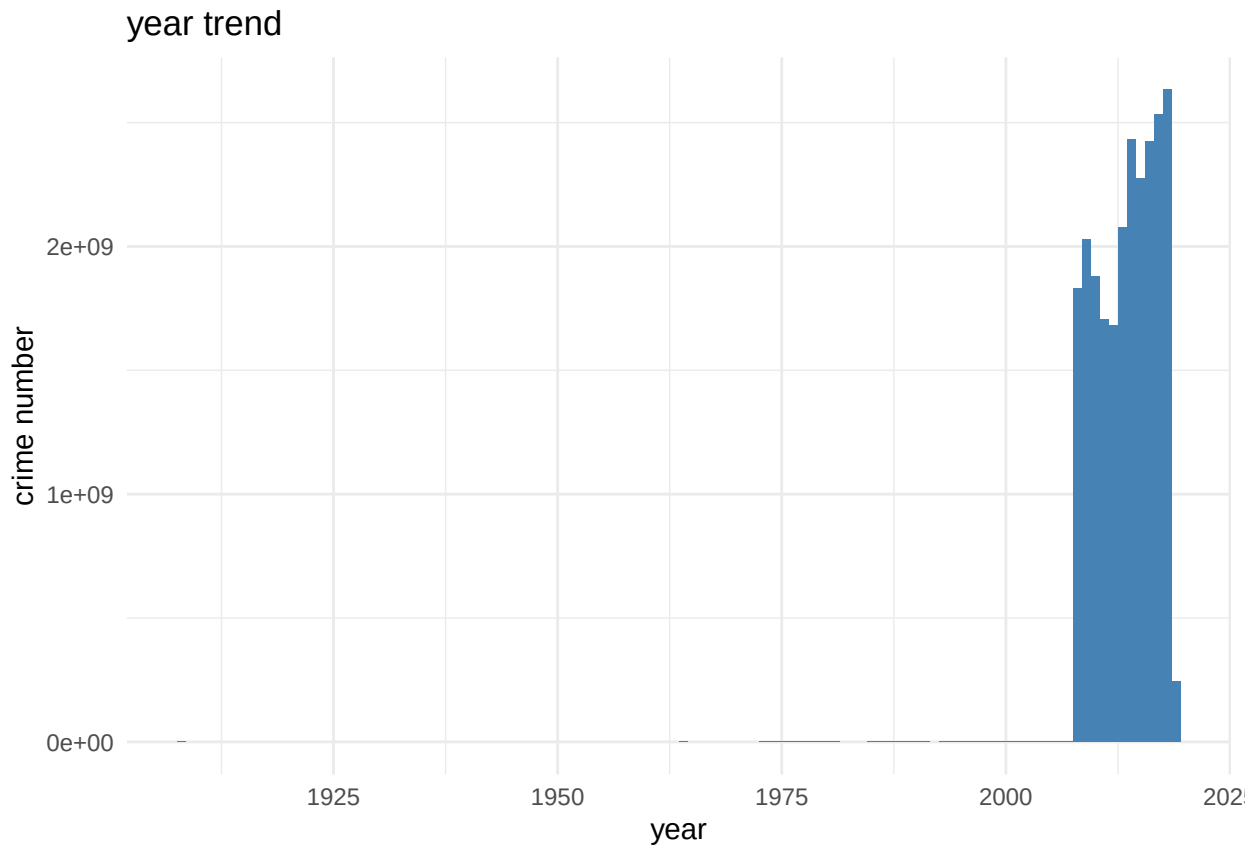
subsetcrime %>%
  group_by(Year)%>%
      summarise(count=n())%>%
        arrange(-count)
```

```
## # A tibble: 45 x 2
##     Year count
##    <int> <int>
##  1  2018 51302
##  2  2017 50334
##  3  2014 49322
##  4  2016 49220
##  5  2015 47693
##  6  2013 45551
##  7  2009 45056
##  8  2010 43353
##  9  2008 42793
## 10  2011 41298
## # ... with 35 more rows
```

```
number<-group_by(subsetcrime, Year)
yearnumber<-summarise(number,count=n())

subsetcrime2<- merge(subsetcrime,yearnumber,by='Year')
```

```
ggplot(data=subsetcrime2, aes(x=Year, y=count)) +
  geom_bar(stat="identity", fill="steelblue")+
  theme_minimal()+ ggtitle("year trend")+xlab("year") + ylab("crime number")
```



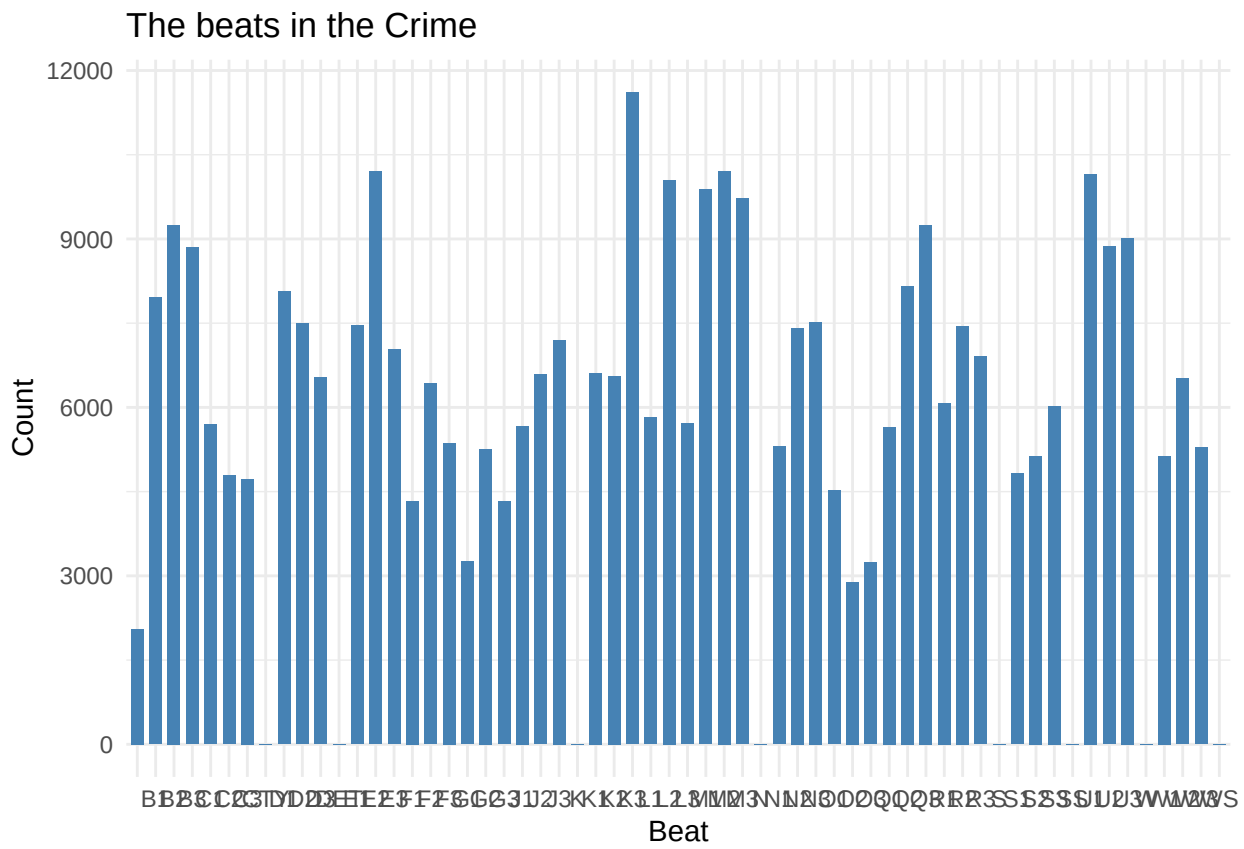

```
#the annual number of crimes committed in the dataset is located mostly
#during 2000-2025

crime.after.2012=filter(subsetcrime2,Year>=2012)
```

(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 means the police department and the territory that a police
#officer patrols.
#And in this dataset beat means designated police sector boundary where
#offense(s) occurred.

# Bar
ggplot(crime.after.2012, aes(x=factor(Beat)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal()+ ggtitle("The beats in the Crime")+xlab("Beat") + ylab("Count")
```




```
#There are more beats in K and M area compared with other beats.

summary(crime.after.2012$Beat)
```

```
##          B1    B2    B3    C1    C2    C3    CS   CTY    D1    D2    D3   DET
##  2054  7954  9253  8846  5694  4789  4726     0     1  8066  7491  6530     7
##    E1    E2    E3    F1    F2    F3    G1    G2    G3    H1   INV    J1    J2
##  7459 10200  7032  4332  6429  5361  3257  5259  4327     0     0  5668  6585
##    J3     K    K1    K2    K3    L1    L2    L3  LAPT    M1    M2    M3     N
##  7203     1  6611  6560 11611  5823 10049  5710     0  9883 10210  9723     1
##    N1    N2    N3    O1    O2    O3    Q1    Q2    Q3    R1    R2    R3     S
##  5303  7409  7517  4523  2894  3239  5647  8159  9249  6080  7448  6909     4
##    S1    S2    S3    SS    U1    U2    U3     W    W1    W2    W3    WS    X9
##  4819  5139  6027     1 10157  8866  9019     3  5135  6514  5286     1     0
```

```
#Yes,there are 2054 missing 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."

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?

```
BeatsDataset<-read.csv(file = '/Users/leechenhsin/Desktop/study@USA/07_UW_School/IMT573/police_beat_and_
BeatsDataset
```

```
##        Name                          Location.1 Latitude Longitude
## 1        B1 (47.7097756394592, -122.370990523069) 47.70978 -122.3710
## 2        B2 (47.6790521901374, -122.391748391741) 47.67905 -122.3918
## 3        B3 (47.6812920482227, -122.364236159741) 47.68129 -122.3642
## 4        C1 (47.6342500180223, -122.315684762418) 47.63425 -122.3157
## 5        C2 (47.6192385752996, -122.313557430551) 47.61924 -122.3136
## 6        C3 (47.6300792887474, -122.292087128251) 47.63008 -122.2921
## 7  CITYWIDE (47.6210041048652, -122.332993498998) 47.62100 -122.3330
## 8        D1 (47.6274421308028, -122.345705781837) 47.62744 -122.3457
## 9        D2 (47.6256548876049, -122.331370005506) 47.62565 -122.3314
## 10       D3 (47.6103493249325, -122.328653706199) 47.61035 -122.3286
## 11        E (47.6201542748144, -122.304782602556) 47.62015 -122.3048
## 12       E1 (47.6203486882073, -122.324419823241) 47.62035 -122.3244
## 13       E2  (47.6118432671102, -122.32016086571) 47.61184 -122.3202
## 14       E3  (47.603162336406, -122.319319689671) 47.60316 -122.3193
## 15       F1 (47.5484146593035, -122.354809670155) 47.54841 -122.3548
## 16       F2 (47.5254502461741, -122.365817548329) 47.52545 -122.3658
## 17       F3 (47.5261052985115, -122.336388313318) 47.52611 -122.3364
## 18       G1 (47.6091373306494, -122.307899616793) 47.60914 -122.3079
## 19       G2 (47.5958952989518, -122.306633195511) 47.59590 -122.3066
## 20       G3 (47.6031821881675, -122.292398835358) 47.60318 -122.2924
## 21       J1  (47.676809900774, -122.337899655521) 47.67681 -122.3379
## 22       J2 (47.6613374516723, -122.363818988307) 47.66134 -122.3638
## 23       J3 (47.6563781774877, -122.336468775341) 47.65638 -122.3365
## 24       K1 (47.6077552981764, -122.334107460638) 47.60776 -122.3341
## 25       K2 (47.5998930290529, -122.326813620856) 47.59989 -122.3268
## 26       K3 (47.5903972078525, -122.333545010682) 47.59040 -122.3336
## 27       L1 (47.7265488817709, -122.302631931191) 47.72655 -122.3026
## 28       L2 (47.7095588837442, -122.303661007867) 47.70956 -122.3037
## 29       L3 (47.6808531540255, -122.277032733938) 47.68085 -122.2770
## 30       M1 (47.6157584422587, -122.350867935301) 47.61576 -122.3509
## 31       M2 (47.6146150193586, -122.340275405136) 47.61462 -122.3403
## 32       M3 (47.6077571617787, -122.340896390036) 47.60776 -122.3409
## 33        N (47.6902980120839, -122.328757390104) 47.69030 -122.3288
## 34       N1 (47.7226875390406, -122.340459039106) 47.72269 -122.3405
## 35       N2  (47.698470493249, -122.351867710243) 47.69847 -122.3519
## 36       N3 (47.7045005246442, -122.329961214037) 47.70450 -122.3300
## 37       O1 (47.5822859359213, -122.311799603309) 47.58229 -122.3118
## 38       O2 (47.5656855826482, -122.330941962362) 47.56569 -122.3309
## 39       O3 (47.5345836385751, -122.303020266287) 47.53458 -122.3030
## 40       Q1  (47.650261230265, -122.400003042555) 47.65026 -122.4000
## 41       Q2 (47.6428529450151, -122.362673076853) 47.64285 -122.3627
## 42       Q3 (47.6269804063179, -122.362807276708) 47.62698 -122.3628
## 43       R1 (47.5758114569194, -122.288707022144) 47.57581 -122.2887
## 44       R2  (47.562285343514, -122.304240734006) 47.56229 -122.3042
## 45       R3 (47.5527951110333, -122.268210782218) 47.55280 -122.2682
## 46       S1 (47.5439339496481, -122.286476209963) 47.54393 -122.2865
## 47       S2 (47.5263519484816, -122.274095175041) 47.52635 -122.2741
## 48       S3 (47.5093533353672, -122.259542630385) 47.50935 -122.2595
```

```
## 49       SE (47.5476766838051, -122.284789228904) 47.54768 -122.2848
## 50       SW (47.5478566154038, -122.361787408364) 47.54786 -122.3618
## 51       U1 (47.6848677676269, -122.309913082907) 47.68487 -122.3099
## 52       U2  (47.6585545300635, -122.30659481859) 47.65855 -122.3066
## 53       U3 (47.6660083487855, -122.312204733721) 47.66601 -122.3122
## 54        W (47.6300237833357, -122.368053164444) 47.63002 -122.3680
## 55       W1 (47.5788164080083, -122.378814011668) 47.57882 -122.3788
## 56       W2 (47.5607068301888, -122.386946475037) 47.56071 -122.3869
## 57       W3 (47.5255479889804, -122.384581696918) 47.52555 -122.3846
```

```
summary(BeatsDataset$Name)
```

```
##       B1       B2       B3       C1       C2       C3 CITYWIDE       D1
##        1        1        1        1        1        1        1        1
##       D2       D3        E       E1       E2       E3       F1       F2
##        1        1        1        1        1        1        1        1
##       F3       G1       G2       G3       J1       J2       J3       K1
##        1        1        1        1        1        1        1        1
##       K2       K3       L1       L2       L3       M1       M2       M3
##        1        1        1        1        1        1        1        1
##        N       N1       N2       N3       O1       O2       O3       Q1
##        1        1        1        1        1        1        1        1
##       Q2       Q3       R1       R2       R3       S1       S2       S3
##        1        1        1        1        1        1        1        1
##       SE       SW       U1       U2       U3        W       W1       W2
##        1        1        1        1        1        1        1        1
##       W3
##        1
```

```
summary(crime.after.2012$Beat)
```

```
##          B1    B2    B3    C1    C2    C3    CS   CTY    D1    D2    D3   DET
##  2054  7954  9253  8846  5694  4789  4726     0     1  8066  7491  6530     7
##    E1    E2    E3    F1    F2    F3    G1    G2    G3    H1   INV    J1    J2
##  7459 10200  7032  4332  6429  5361  3257  5259  4327     0     0  5668  6585
##    J3     K    K1    K2    K3    L1    L2    L3  LAPT    M1    M2    M3     N
##  7203     1  6611  6560 11611  5823 10049  5710     0  9883 10210  9723     1
##    N1    N2    N3    O1    O2    O3    Q1    Q2    Q3    R1    R2    R3     S
##  5303  7409  7517  4523  2894  3239  5647  8159  9249  6080  7448  6909     4
##    S1    S2    S3    SS    U1    U2    U3     W    W1    W2    W3    WS    X9
##  4819  5139  6027     1 10157  8866  9019     3  5135  6514  5286     1     0
```

```
#dplyr summarize (setdf)

#Yes,Crime Dataset include 10  police beats that are not present in the Beats
#Dataset. The frequency of them are CTY:1, DET:7,H1:0, INV:0, K:1,
#LAPT:0, S:4, SS:1, WS:1,X9:0


#They are rather infrequent in the dataset.

#I think remove them will not drastically alter the scope of the Crime Dataset
#since the number of them does't account big enough.

summary(crime.after.2012$Beat)
```

```
##          B1    B2    B3    C1    C2    C3    CS   CTY    D1    D2    D3   DET
##  2054  7954  9253  8846  5694  4789  4726     0     1  8066  7491  6530     7
##    E1    E2    E3    F1    F2    F3    G1    G2    G3    H1   INV    J1    J2
##  7459 10200  7032  4332  6429  5361  3257  5259  4327     0     0  5668  6585
##    J3     K    K1    K2    K3    L1    L2    L3  LAPT    M1    M2    M3     N
##  7203     1  6611  6560 11611  5823 10049  5710     0  9883 10210  9723     1
##    N1    N2    N3    O1    O2    O3    Q1    Q2    Q3    R1    R2    R3     S
##  5303  7409  7517  4523  2894  3239  5647  8159  9249  6080  7448  6909     4
##    S1    S2    S3    SS    U1    U2    U3     W    W1    W2    W3    WS    X9
##  4819  5139  6027     1 10157  8866  9019     3  5135  6514  5286     1     0
```

```
#crime.after.2012$Beat=toString(crime.after.2012$Beat)

new_crime.after.2012=filter(crime.after.2012,crime.after.2012$Beat!="")


beat_group=group_by(new_crime.after.2012,new_crime.after.2012$Beat)

beat_summary=summarise(beat_group,count=n())

beat_greater_than_10=filter(beat_summary,count>=10)

sum(beat_greater_than_10$count)
```

```
## [1] 347980
```

(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)

```
library(tigris)
```

```
## To enable
## caching of data, set `options(tigris_use_cache = TRUE)` in your R script or .Rprofile.
```

```
##
## Attaching package: 'tigris'
```

```
## The following objects are masked from 'package:censusr':
##
##     append_geoid, call_geolocator, call_geolocator_latlon
```

```
call_geolocator_latlon(40.61847, -74.02123)
```

```
## [1] "360470152003001"
```

```
geolocate <-mapply(call_geolocator_latlon, lat = BeatsDataset$Latitude, lon = BeatsDataset$Longitude)


BeatsDataset$geoloaction=geolocate
```

```
#the advantage of join first and then finding census tracts is to have completed
#data and information.
#the disadcantage of join first and then finding cansus tracts is that database
#will be too large and needs to take time to load it.
```

(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?

```
Statecode<- substr(BeatsDataset$geoloaction,1,2)
Countrycode<- substr(BeatsDataset$geoloaction,3,5)

BeatsDataset$Statecode=Statecode
BeatsDataset$Countrycode=Countrycode


#yes, the extracted state and country codes are what I am expect to be.
#since the state code of WA State is 53 and country code is 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.

```
digitcode<- substr(BeatsDataset$geoloaction,1,11)
BeatsDataset$digitcode=digitcode
```

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

```
censusdata<-read.csv(file = '/Users/leechenhsin/Desktop/study@USA/07_UW_School/IMT573/census_edu_data.cs

digitcode<- substr(censusdata$GEO.id,10,21)
censusdata$digitcode=digitcode
```

(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.

```
beat_census= censusdata %>% inner_join(BeatsDataset,by="digitcode")

save(beat_census,file='/Users/leechenhsin/Desktop/study@USA/07_UW_School/IMT573/beat_census.RData')
```