Exploring census data

Shree Priya

NOTE: Answers are given in *italics*

Setup:

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

```
# Load standard libraries
library('dplyr')
library('censusr')
library('stringr')
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 ultimate 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.

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 (crime_data.csv). 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 csv file for consistency. We will henceforth call this dataset the "Crime Dataset." Perform a basic inspection of the Crime Dataset and discuss what you find.

```
#Reading the data
crd = read.csv("crime_data.csv.bz2")

#Checking the validity of the data
head(crd)
```

```
##   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
```

```
#Number of rows
nrow(crd)
```

```
## [1] 448821
```

```
#Number of columns
ncol(crd)
```

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

```
#Checking crime rate according to the neighbourhoods
n = crd %>% group_by(Neighborhood) %>% summarise(count=n())

#Arranging it highest to lowest
head(arrange(n, desc(count)))
```

```
## # A tibble: 6 x 2
##   Neighborhood        count
##   <fct>               <int>
## 1 DOWNTOWN COMMERCIAL 42077
## 2 NORTHGATE           26487
## 3 CAPITOL HILL        26421
## 4 QUEEN ANNE          23309
## 5 SLU/CASCADE         20232
## 6 UNIVERSITY          17804
```

Downtown commercial has more than 1.5 times of the other neighborhoods.

(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?

```
#Removing NA from occured date
crd = crd %>% filter(is.na(Occurred.Date)==FALSE)

#Checking the dimensions
dim(crd)
```

```
## [1] 448821     11
```

```r
#Checking the validity of the data
head(crd)
```

```
##   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
```

```r
#Making the column into the Date format for extracting year
crd$Occurred.Date = as.Date(crd$Occurred.Date, "%m/%d/%Y")

#Checking the validity of data
head(crd)
```

```
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 1     1.975e+12    1975-12-16           900    12/16/1975          1500
## 2     1.976e+12    1976-01-01             1    01/31/1976          2359
## 3     1.979e+12    1979-01-28          1600    02/09/1979          1430
## 4     1.981e+13    1981-08-22          2029    08/22/1981          2030
## 5     1.981e+12    1981-02-14          2000    02/15/1981           435
## 6     1.988e+13    1988-09-29           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
```

```r
#Extracting the year and ensuring its saved as a number
crd1 = crd %>% mutate(YEAR = as.numeric(format(crd$Occurred.Date, "%Y")))

#Checking the validity of data
head(crd1)
```

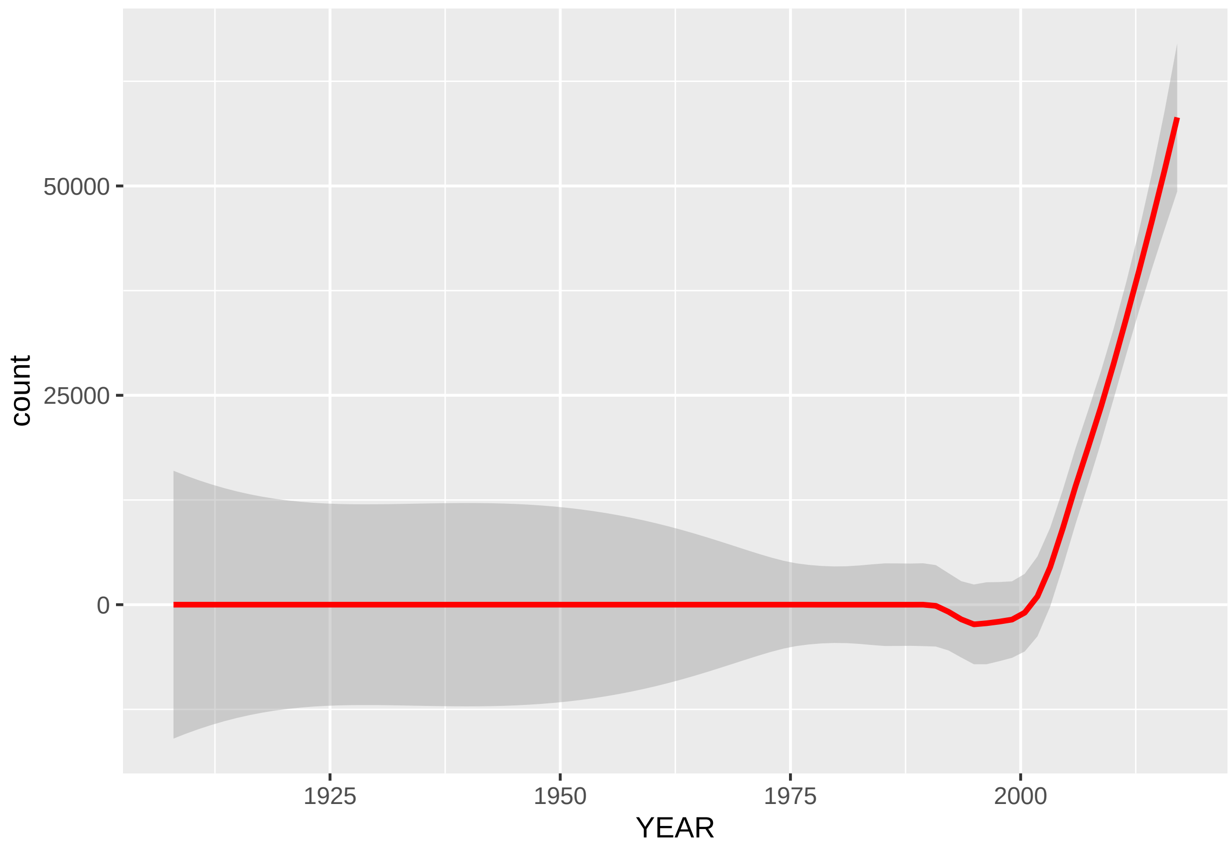
```
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 1     1.975e+12    1975-12-16           900    12/16/1975          1500
## 2     1.976e+12    1976-01-01             1    01/31/1976          2359
## 3     1.979e+12    1979-01-28          1600    02/09/1979          1430
## 4     1.981e+13    1981-08-22          2029    08/22/1981          2030
## 5     1.981e+12    1981-02-14          2000    02/15/1981           435
## 6     1.988e+13    1988-09-29           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 YEAR
## 1            LAKEWOOD/SEWARD PARK 1975
## 2                         UNKNOWN 1976
## 3        CENTRAL AREA/SQUIRE PARK 1979
## 4                 BRIGHTON/DUNLAP 1981
## 5 ROXHILL/WESTWOOD/ARBOR HEIGHTS 1981
## 6                     SLU/CASCADE 1988
```

```r
#To check the earliest year
head(arrange(crd1, YEAR),1) #1908
```

```
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 1     2.008e+13    1908-12-13          2114    12/13/2008          2114
##   Crime.Subcategory Primary.Offense.Description Precinct Sector Beat
## 1               DUI                  DUI-LIQUOR     EAST      G   G2
##               Neighborhood YEAR
## 1 CENTRAL AREA/SQUIRE PARK 1908
```

```r
#Annual number of crimes committed
crd2 = crd1 %>% filter(is.na(YEAR)==FALSE) %>% group_by(YEAR) %>% summarise(count=n())

#Plotting the yearly crime rate
ggplot(crd2, aes(x=YEAR, y=count)) + geom_smooth(color="Red") + scale_y_continuous()
```



As we can see from the above data, after the year 2000 the crime rate increases exponentially. This is one of the anamolies I found in the data

Let’s 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.

```
#Assuming after 2011 does NOT include 2011
crd_2011 = crd1 %>% filter(YEAR > 2011)

#Checking the dimensions of the data
dim(crd_2011)
```

```
## [1] 275320     12
```

```
#Checking the validity
head(crd_2011)
```

```
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 1     2.012e+13    2012-04-02          2040    04/03/2012            28
## 2     2.012e+13    2012-04-02          2100    04/02/2012          2103
## 3     2.012e+13    2012-04-02          1930    04/02/2012          2126
## 4     2.012e+13    2012-04-02          2144    04/02/2012          2144
## 5     2.012e+13    2012-04-02          2218    04/02/2012          2218
## 6     2.012e+13    2012-04-02          2229    04/02/2012          2229
##     Crime.Subcategory Primary.Offense.Description Precinct Sector Beat
## 1            NARCOTIC         NARC-POSSESS-MARIJU     WEST      K   K2
## 2  ROBBERY-COMMERCIAL        ROBBERY-BUSINESS-GUN    NORTH      B    B2
## 3 MOTOR VEHICLE THEFT              VEH-THEFT-AUTO    NORTH      J    J1
## 4                 DUI                  DUI-LIQUOR     EAST      E    E3
## 5 ROBBERY-RESIDENTIAL ROBBERY-RESIDENCE-BODYFORCE     EAST      C    C2
## 6           CAR PROWL              THEFT-CARPROWL    NORTH      U    U2
##     Neighborhood YEAR
## 1 PIONEER SQUARE 2012
## 2  BALLARD SOUTH 2012
## 3  BALLARD NORTH 2012
## 4   CAPITOL HILL 2012
## 5   MADISON PARK 2012
## 6     UNIVERSITY 2012
```

Here, I have assumed that year 2011 is NOT will not be present in the data. Getting 275320 rows and 12 columns after cleaning

(c) Looking at Frequency of Beats

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?

```
#Number of rows with Beat information
no_of_2011=nrow(crd_2011)

#Number of rows without Beat information
beat_2011 =nrow(crd_2011 %>% filter(Beat == ""))

#Frequency
no_of_2011/beat_2011
```

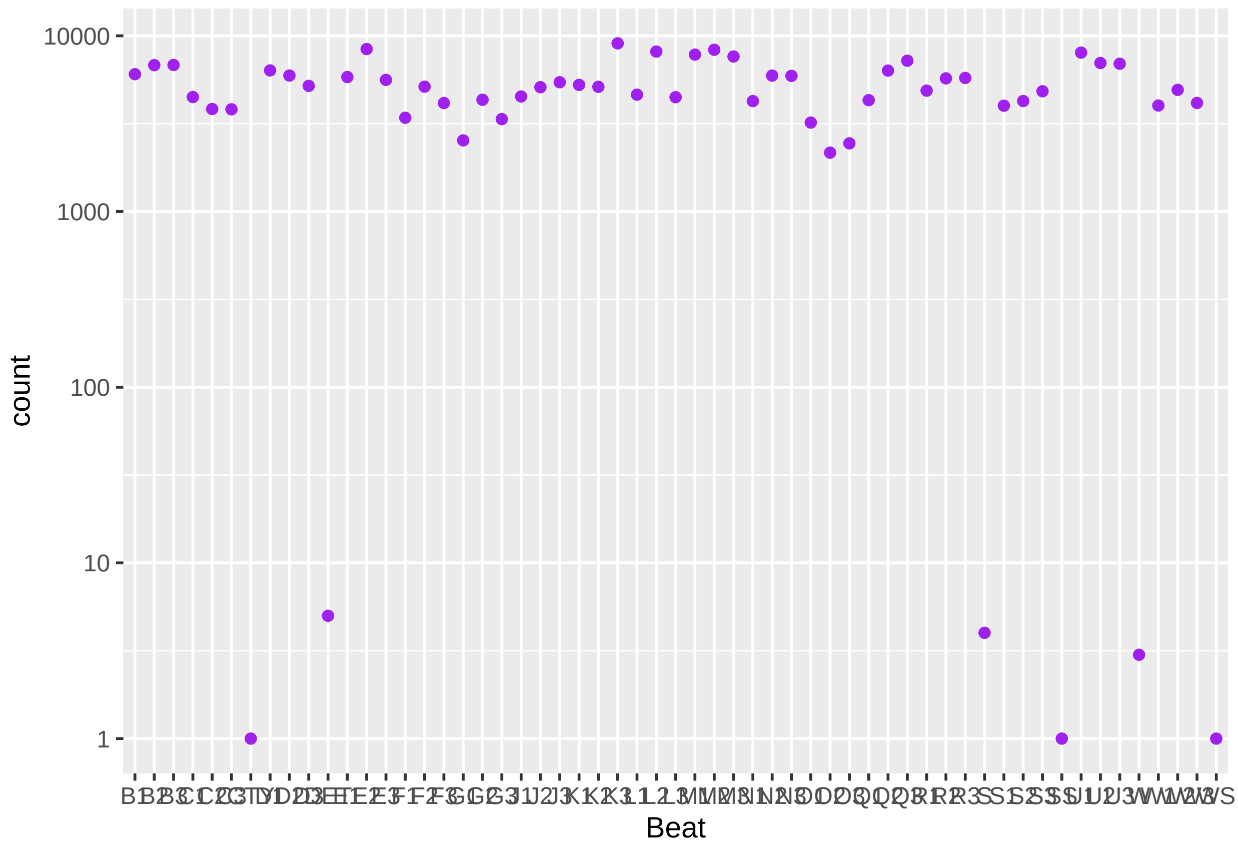
```
## [1] 182.452
```

```
#Finding the frequency of each beat in the Beats data set
freq_beats = crd_2011 %>% filter(Beat != "") %>% group_by(Beat) %>% summarise(count=n()) %>% arrange(de

#Checking the validity of the data
#We can see the anamolies in the data from the tail
tail(freq_beats,7)
```

```
## # A tibble: 7 x 2
##   Beat  count
##   <fct> <int>
## 1 O2     2163
## 2 DET       5
## 3 S         4
## 4 W         3
## 5 CTY       1
## 6 SS        1
## 7 WS        1
```

```
#Plotting the data
ggplot(freq_beats, aes(x=Beat, y=count)) + geom_point(color="purple") + scale_y_log10()
```



```r
#Filtering out missing beats
missing_beats = crd_2011 %>% filter(Beat == "")

#Number of rows of the missing beats
nrow(missing_beats)
```

```
## [1] 1509
```

The frequency of the occurence of each beat is presented above The anamolies with the frequency of the data- There are 6 beats with in single digits, and the rest all the beats are a 4 digit number. You can see this in the tail() function of the data above Yes, there are missing beats, The number of missing beats are 1509

(d) Importing Police Beat Data and Filtering on Frequency

Load the data on Seattle police beats (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 henceforth call this dataset the "Beats Dataset."

```r
#Reading the data
beats = read.csv("police_beat_and_precinct_centerpoints.csv")

#Checking the validity of the data
head(beats)
```

```
##   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
```

```
#Checking the dimensions of the data
dim(beats)
```

```
## [1] 57  4
```

We have now loaded the beats data set, it has 57 rows and 4 columns

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?

```
#Anti join to check if particular beats are not present in the Crime data set
an_crd = anti_join(crd_2011, beats, by=c("Beat" = "Name"))
```

```
## Warning: Column `Beat`/`Name` joining factors with different levels,
## coercing to character vector
```

```
#Checking the dimensions of the data
dim(an_crd)
```

```
## [1] 1521   12
```

```
#Checking the count of all the anamolies
an_crd %>% group_by(Beat) %>% summarise(count=n())
```

```
## # A tibble: 6 x 2
##   Beat  count
##   <fct> <int>
## 1 ""     1509
## 2 CTY       1
## 3 DET       5
## 4 S         4
## 5 SS        1
## 6 WS        1
```

There are 1509 beats missing, the rest of the beats above are in single digits which are not present in the Crime dataset. There are a total of 1521 values in the Crime data set not present in beats dataset

Would you say that these comprise a large number of the observations in the Crime Dataset or are they rather infrequent?

No, 1521 values are compared to 273796 values is very less. So I would say that it does not account to much part of the crime dataset

Do you think removing them would drastically alter the scope of the Crime Dataset?

No, I think that removing them would not affect the crime dataset a lot and will be beneficiary instead. The data should be cleaned.

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?

```
#Grouping by Beat and counting the number of occurences
crd_new = crd_2011 %>% filter(Beat != "") %>% group_by(Beat) %>% summarise(count=n()) %>% arrange(desc(

#Filtering out less than 10 occurences
crd_new = crd_new %>% filter(count>=10)

#Checking the validity of data
head(crd_new)
```

```
## # A tibble: 6 x 2
##   Beat  count
##   <fct> <int>
## 1 K3     9056
## 2 E2     8408
## 3 M2     8328
## 4 L2     8136
## 5 U1     8023
## 6 M1     7813
```

```
#Removing the occurences from the beats data set
crd_2011_new = inner_join(crd_2011, crd_new, by="Beat")

#Number of rows of new data set
nrow(crd_2011_new)
```

```
## [1] 273796
```

```
#Checking the validity of the data
head(crd_2011_new)
```

```
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 1     2.012e+13    2012-04-02          2040    04/03/2012            28
## 2     2.012e+13    2012-04-02          2100    04/02/2012          2103
## 3     2.012e+13    2012-04-02          1930    04/02/2012          2126
## 4     2.012e+13    2012-04-02          2144    04/02/2012          2144
## 5     2.012e+13    2012-04-02          2218    04/02/2012          2218
## 6     2.012e+13    2012-04-02          2229    04/02/2012          2229
##     Crime.Subcategory Primary.Offense.Description Precinct Sector Beat
## 1            NARCOTIC        NARC-POSSESS-MARIJU     WEST      K   K2
## 2  ROBBERY-COMMERCIAL        ROBBERY-BUSINESS-GUN    NORTH      B   B2
## 3 MOTOR VEHICLE THEFT               VEH-THEFT-AUTO    NORTH      J   J1
## 4                 DUI                  DUI-LIQUOR     EAST      E   E3
## 5 ROBBERY-RESIDENTIAL ROBBERY-RESIDENCE-BODYFORCE     EAST      C   C2
## 6           CAR PROWL               THEFT-CARPROWL    NORTH      U   U2
##     Neighborhood YEAR count
## 1 PIONEER SQUARE 2012  5125
## 2  BALLARD SOUTH 2012  6807
## 3  BALLARD NORTH 2012  4516
## 4   CAPITOL HILL 2012  5611
## 5   MADISON PARK 2012  3830
## 6     UNIVERSITY 2012  7001
```

```
#Removing the last row containing the count
crd_2011_new = crd_2011_new[1:(length(crd_2011_new)-1)]

#Checking the validity of the data
head(crd_2011_new)
```

```
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 1     2.012e+13    2012-04-02          2040    04/03/2012            28
## 2     2.012e+13    2012-04-02          2100    04/02/2012          2103
## 3     2.012e+13    2012-04-02          1930    04/02/2012          2126
## 4     2.012e+13    2012-04-02          2144    04/02/2012          2144
## 5     2.012e+13    2012-04-02          2218    04/02/2012          2218
## 6     2.012e+13    2012-04-02          2229    04/02/2012          2229
##     Crime.Subcategory Primary.Offense.Description Precinct Sector Beat
## 1             NARCOTIC         NARC-POSSESS-MARIJU     WEST      K   K2
## 2  ROBBERY-COMMERCIAL        ROBBERY-BUSINESS-GUN    NORTH      B   B2
## 3 MOTOR VEHICLE THEFT              VEH-THEFT-AUTO    NORTH      J   J1
## 4                 DUI                  DUI-LIQUOR     EAST      E   E3
## 5 ROBBERY-RESIDENTIAL ROBBERY-RESIDENCE-BODYFORCE     EAST      C   C2
## 6           CAR PROWL              THEFT-CARPROWL    NORTH      U   U2
##     Neighborhood YEAR
## 1 PIONEER SQUARE 2012
## 2  BALLARD SOUTH 2012
## 3  BALLARD NORTH 2012
## 4   CAPITOL HILL 2012
## 5   MADISON PARK 2012
## 6     UNIVERSITY 2012
```

```
#Checking the dimensions of the data
dim(crd_2011_new)
```

```
## [1] 273796     12
```

After removing the beats with less than 10 occurences, we have 273796 rows and 12 columns .

(e) Importing and Inspecting Police Beat Data

To join the Beat Dataset to census data, we must have census tract information.

First, let's remove the beats in the Beats Dataset that are not listed in the (cleaned) Crime Dataset.

Then, let's 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.

```
#Removing the beats that are not present in the crime data set
beats_new = anti_join(beats, crd_2011_new, by=c("Name"="Beat"))
```

```
## Warning: Column `Name`/`Beat` joining factors with different levels,
## coercing to character vector
```

```
#checking the dimensions of the data
dim(beats_new)
```

```
## [1] 6 4
```

```
#Anti join the removed rows with the beats data set
beats_new_2 = anti_join(beats, beats_new, by=c("Name"))

#checking the dimensions of the data
dim(beats_new_2)
```

```
## [1] 51  4
```

```
#Applying the geolocator function to get the census code
beats_new_2$census_data = apply(beats_new_2, 1, function(row) call_geolocator_latlon(row['Latitude'], ro

#Checking the validity of the data
head(beats_new_2)
```

```
##   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
##       census_data
## 1 530330014004000
## 2 530330032001003
## 3 530330029003016
## 4 530330065001015
## 5 530330075002001
## 6 530330063004005
```

Removed the beats that are not present in the Cleaned crime data set and added the census code. It can be seen in the head() function above.

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)

No, it would be very inefficient to join and then find the census code because there are more number of rows when it's joined. 273796 rows as compared to the 57 rows in the beats data set. The function runs of the server, so it would be very slow if it ran on 273796 rows

(f) Extracting FIPS Codes

Once we have the 15-digit census codes, we will break down the code based on information of interest.

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?

```
#Checking the validity of the data
head(beats_new_2)
```

```
##   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
##       census_data
## 1 530330014004000
## 2 530330032001003
## 3 530330029003016
## 4 530330065001015
## 5 530330075002001
## 6 530330063004005
```

```
#Making a state code and the county code in the beats data set
beats_new_2 = beats_new_2 %>% mutate(state_code = substr(census_data, 0, 2), county_code = substr(censu

#Checking the validity of the data
head(beats_new_2)
```

```
##   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
##       census_data state_code county_code
## 1 530330014004000         53         033
## 2 530330032001003         53         033
## 3 530330029003016         53         033
## 4 530330065001015         53         033
## 5 530330075002001         53         033
## 6 530330063004005         53         033
```

```
#Checking the dimensions of the data
dim(beats_new_2)
```

```
## [1] 51  7
```

Yes, it is the same as what is expected on the internet. The WA state code is 53 and King County's county code is 033. We can see the same thing in the data as

(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
#Extracting and Adding the 11 digit code to the beats data set
beats_new_2 = beats_new_2 %>% mutate(code_11 = substr(census_data, 0, 11))

#Checking the validity of data
head(beats_new_2)
```

```
##   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
##       census_data state_code county_code     code_11
## 1 530330014004000         53         033 53033001400
## 2 530330032001003         53         033 53033003200
## 3 530330029003016         53         033 53033002900
## 4 530330065001015         53         033 53033006500
## 5 530330075002001         53         033 53033007500
## 6 530330063004005         53         033 53033006300
```

```r
#Checking the dimensions of the data
dim(beats_new_2)
```

```
## [1] 51  8
```

Added the 11 digit code to the beats data set

(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
#Reading the census data
cen = read.csv("census_edu_data.csv.bz2")

#Checking the validity of the data
head(cen,2)
```

```
##                   GEO.id    GEO.id2                       GEO.display.label
## 1 1400000US53033000100 5.3033e+10 Census Tract 1, King County, Washington
## 2 1400000US53033000200 5.3033e+10 Census Tract 2, King County, Washington
##   total no_schooling nursery_school kindergarten X1st_grade X2nd_grade
## 1  5708           82              0            0          0          0
## 2  6079          115              0            0          0          0
##   X3rd_grade X4th_grade X5th_grade X6th_grade X7th_grade X8th_grade
## 1         59         59          0         44          0        110
## 2          0          0          0         66          3          0
##   X9th_grade X10th_grade X11th_grade X12th_grade_no_diploma
## 1          0          28          27                    112
## 2         41          17          42                    125
##   high_school_diploma ged_or_alternative_credential
## 1                 833                           239
## 2                 614                           169
##   some_college_less_than_1_year some_college_1_or_more_years_no_degree
## 1                           259                                    669
## 2                           229                                    739
##   associates_degree bachelors_degree masters_degree
## 1               470             1600            584
## 2               458             2105           1045
##   professional_school_degree doctorate_degree
## 1                        319              214
## 2                         77              234
```

```
#Extracting the 11 digit code and making a column
cen2 = cen %>% mutate(code_11 = substr(GEO.id, 10, 21))

#Checking the validity of the data
head(cen2,2)
```

```
##                 GEO.id    GEO.id2                      GEO.display.label
## 1 1400000US53033000100 5.3033e+10 Census Tract 1, King County, Washington
## 2 1400000US53033000200 5.3033e+10 Census Tract 2, King County, Washington
##   total no_schooling nursery_school kindergarten X1st_grade X2nd_grade
## 1  5708           82              0            0          0          0
## 2  6079          115              0            0          0          0
##   X3rd_grade X4th_grade X5th_grade X6th_grade X7th_grade X8th_grade
## 1         59         59          0         44          0        110
## 2          0          0          0         66          3          0
##   X9th_grade X10th_grade X11th_grade X12th_grade_no_diploma
## 1          0          28          27                    112
## 2         41          17          42                    125
##   high_school_diploma ged_or_alternative_credential
## 1                 833                           239
## 2                 614                           169
##   some_college_less_than_1_year some_college_1_or_more_years_no_degree
## 1                           259                                    669
## 2                           229                                    739
##   associates_degree bachelors_degree masters_degree
## 1               470             1600            584
## 2               458             2105           1045
##   professional_school_degree doctorate_degree     code_11
## 1                        319              214 53033000100
## 2                         77              234 53033000200
```

```
#Selecting only the GEO id and code for display purposes
head(cen2 %>% select(GEO.id, code_11), 2)
```

```
##                 GEO.id     code_11
## 1 1400000US53033000100 53033000100
## 2 1400000US53033000200 53033000200
```

Added the 11-digit code to the census data

(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?

```
#merging the beats and the census data
mer_b_c = left_join(beats_new_2, cen2, by="code_11")

#Checking the validity of the data
head(mer_b_c,2)
```

```
##   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
##       census_data state_code county_code     code_11               GEO.id
## 1 530330014004000         53         033 53033001400 1400000US53033001400
## 2 530330032001003         53         033 53033003200 1400000US53033003200
##      GEO.id2                        GEO.display.label total no_schooling
## 1 5.3033e+10 Census Tract 14, King County, Washington  4155            0
## 2 5.3033e+10 Census Tract 32, King County, Washington  6896           26
##   nursery_school kindergarten X1st_grade X2nd_grade X3rd_grade X4th_grade
## 1              0            0          0          0         15          0
## 2              0            0          0          0          0          0
##   X5th_grade X6th_grade X7th_grade X8th_grade X9th_grade X10th_grade
## 1          0          0          0         33         18         110
## 2          0          0          0         15          0           0
##   X11th_grade X12th_grade_no_diploma high_school_diploma
## 1          20                     34                 472
## 2          15                      0                 348
##   ged_or_alternative_credential some_college_less_than_1_year
## 1                           100                           245
## 2                           102                           205
##   some_college_1_or_more_years_no_degree associates_degree
## 1                                    536               310
## 2                                    776               444
##   bachelors_degree masters_degree professional_school_degree
## 1             1301            760                         64
## 2             3000           1433                        374
##   doctorate_degree
## 1              137
## 2              158
```

```
#checking the number of rows of the merge data
dim(mer_b_c)
```

```
## [1] 51 36
```

```
#Checking the number of rows in beats
dim(beats_new_2)
```

```
## [1] 51  8
```

```
#Same number of rows

#To check if there are any beats data that do not have any associated census data

check = anti_join(beats_new_2, cen2, by="code_11")

check
```

```
## [1] Name        Location.1  Latitude    Longitude   census_data state_code
## [7] county_code code_11
## <0 rows> (or 0-length row.names)
```

```
#No rows so all the police beats have census data.
```

In the above chunk of the code, we check the number of rows in beats and the number of rows of the merged data set. Both are the same. Also we check if the beats table contains any value that is not present in the census data. The number of rows of check is 0 confirming that there are no rows in the beats data set that is there in the census data set.

Now 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?

```
#Final merging of the crime data and the merge data
final_merge = left_join(crd_2011_new, mer_b_c, by=c("Beat"="Name"))
```

```
## Warning: Column `Beat`/`Name` joining factors with different levels,
## coercing to character vector
```

```
#Checking the validity of the data
head(final_merge,2)
```

```
##   Report.Number Occurred.Date Occurred.Time Reported.Date Reported.Time
## 1     2.012e+13    2012-04-02          2040    04/03/2012            28
## 2     2.012e+13    2012-04-02          2100    04/02/2012          2103
##    Crime.Subcategory Primary.Offense.Description Precinct Sector Beat
## 1           NARCOTIC         NARC-POSSESS-MARIJU     WEST      K   K2
## 2 ROBBERY-COMMERCIAL        ROBBERY-BUSINESS-GUN    NORTH      B   B2
##     Neighborhood YEAR                            Location.1 Latitude
## 1 PIONEER SQUARE 2012 (47.5998930290529, -122.326813620856) 47.59989
## 2  BALLARD SOUTH 2012 (47.6790521901374, -122.391748391741) 47.67905
##   Longitude     census_data state_code county_code     code_11
## 1 -122.3268 530330092001012         53         033 53033009200
## 2 -122.3918 530330032001003         53         033 53033003200
##                 GEO.id     GEO.id2
## 1 1400000US53033009200 53033009200
## 2 1400000US53033003200 53033003200
##                           GEO.display.label total no_schooling
## 1 Census Tract 92, King County, Washington  2529           56
## 2 Census Tract 32, King County, Washington  6896           26
##   nursery_school kindergarten X1st_grade X2nd_grade X3rd_grade X4th_grade
## 1              0            0          0          0         37          5
## 2              0            0          0          0          0          0
##   X5th_grade X6th_grade X7th_grade X8th_grade X9th_grade X10th_grade
## 1         17        156          4        100         49          19
## 2          0          0          0         15          0           0
##   X11th_grade X12th_grade_no_diploma high_school_diploma
## 1          14                     63                 354
## 2          15                      0                 348
##   ged_or_alternative_credential some_college_less_than_1_year
## 1                            88                           134
## 2                           102                           205
##   some_college_1_or_more_years_no_degree associates_degree
## 1                                    503               114
## 2                                    776               444
##   bachelors_degree masters_degree professional_school_degree
## 1              536            172                         71
## 2             3000           1433                        374
##   doctorate_degree
## 1               37
## 2              158
```

```
#Checking the dimensions of the data
dim(final_merge)
```

```
## [1] 273796     47
```

After merging the data sets, the final dimensions of the data set are 273796 rows and 47 columns

Once everything is joined, save the final dataset for future use. We'll revisit it in future problem sets!