Basic Analysis For Crime Data And Arrest Data In LA

Author: Wan-Lun Tsai

BASIC INFORMATION

Data Description

- Data Set A : Crime_Date_from_2010_to_Present (LA)
  - include crime data from 2010 to 2017, and the date , time , district, crime category and so on
  - there are 1617731 rows and 26 columns
- Date Set B: Census_Data_by_Council_District(2010)(LA)
  - include gender, race, age of population in each council district
  - there are 15 rows and 41 columns
- Data Set C: Arrest_Data_from_2010_to_Present(LA)
  - include arrest day, time, district, charge group and so on
  - there are 1155631 rows and 17 columns

Data Source

https://catalog.data.gov/dataset?organization_type=City+Government&page=1

Question 1

Find out the crime type with the highest frequencies in LA in 2010, then find out the month with the highest rate of this crime type. Both results are needed to be shown in the form of graphs or tables. Then draw a map to show the differences of total population among different council districts in LA in 2010, and try to show the distribution of the crime type on the map.

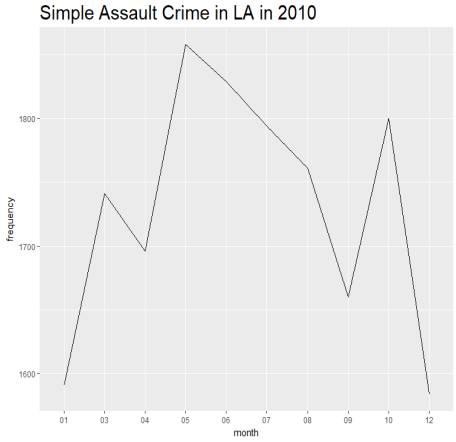
Question 2

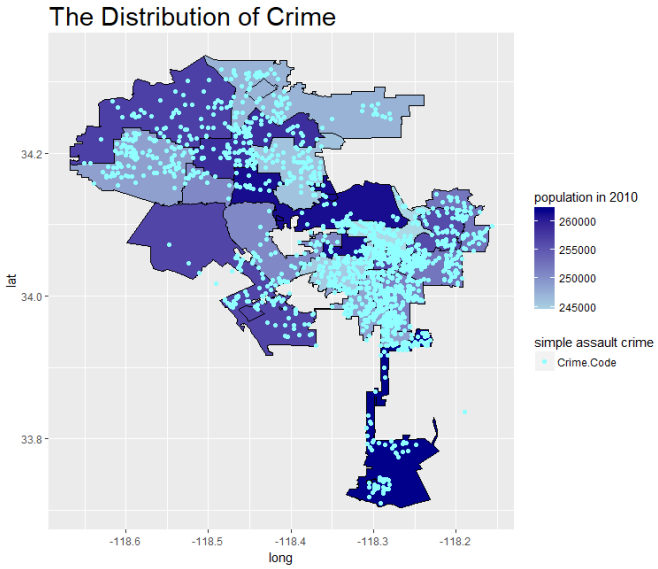
Compute the average time occurred (2010-2017) of every kind of crime and compute the pairwise distance between each kind of crime based on the average. Use MDS to create a 2-dimensional map. Repeat these steps by using Data Set C (based on the average time of each charge group).

QUESTION 1

Top Ten Crimes in LA in 2010

	Crime.Code	Crime.Code.Description	freq
1	624	BATTERY - SIMPLE ASSAULT	20331
2	330	BURGLARY FROM VEHICLE	17529
3	510	VEHICLE - STOLEN	16459
4	310	BURGLARY	15738
5	440	THEFT PLAIN - PETTY ($950 & UNDER)	10810
6	341	THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD0036	10482
7	626	INTIMATE PARTNER - SIMPLE ASSAULT	10345
8	354	THEFT OF IDENTITY	10101
9	740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 0114	9890
10	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	9836





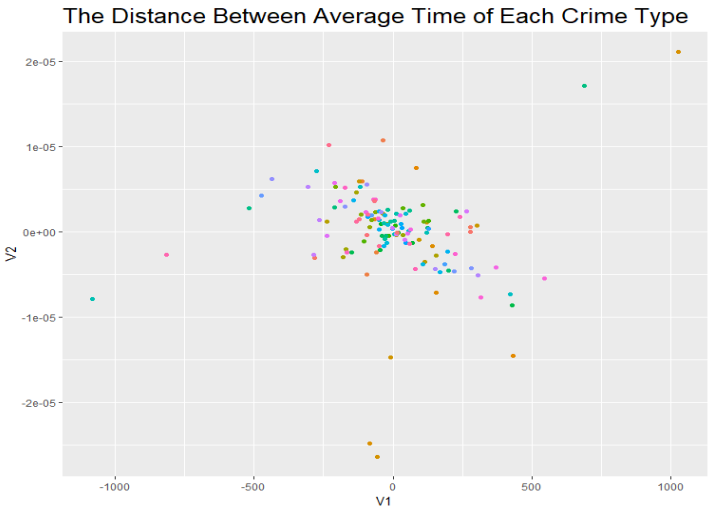
Methodology

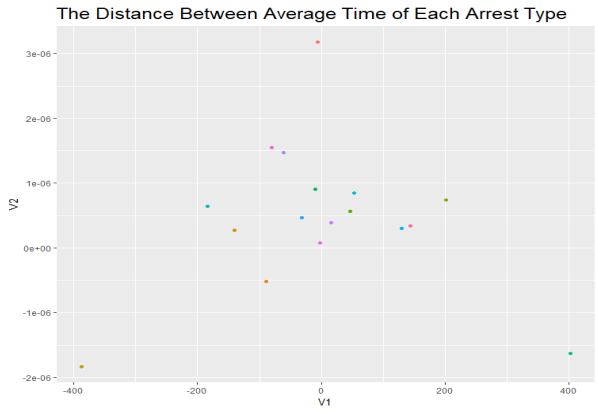
The purpose of this question is to find some relationship between population and crimes. Since the data is too large, it would be too complicated if using every data. Therefore, I find out the most frequent crime in 2010 first, and the result is shown by the table “Top Ten Crimes in LA in 2010”. And then I find out the most frequent month of this crime in 2010, and the result is shown by the plot “Simple Assault Crimes in LA in 2010”. Finally, I find out the location of each crime and draw them on the map of population.

Summary

Through the last graph, we could know in the districts with less population have more frequencies of “simple assault” crime. On the contrary, the districts with more population have relatively low frequencies of this crime type.

QUESTION 2





Methodology

First, I calculate the average time of each type of crime and arrest. Then, I compute their pairwise distances. Finally, I use MDS to create a 2-dimensional map for the crime type and arrest type, respectively.

Summary

From the first graph, we could find out most of the crime type happen almost at the same time, which means in a certain period of time, people have more chance to be in danger.

If comparing the two graphs above, we could also find out the average time among arrest type is less concentrated. Therefore, we may conclude that the average time of each arrest type is different, at least not as similar as the average time of each crime type.

APPENDIX

```
library(tidyr)
library(stringr)
library(ggmap)
library(mapdata)
library(tidyverse)
library(ggplot2)
library(rgeos)
library(maptools)
library(dplyr)
library(tmap)
library(sp)
library(grid)
library(gridExtra)

##load data
crime=read.csv("C:/Users/zubad/Downloads/Crime_Data_from_2010_to_Present.csv")
census=read.csv("C:/Users/zubad/Downloads/Census_Data_by_Council_District.csv")
arrest=read.csv("C:/Users/zubad/Downloads/Arrest_Data_from_2010_to_Present.csv")
district_map=rgdal::readOGR(dsn="C:/Users/zubad/Desktop/council_district.shp",layer=
"council_district")

##separate some columns
newcrime=crime%>%separate(Date.Reported,into=c("month","date","year"),sep="/")%>
%extract(Location,into=c("long","lat"),"([0-9]*\\.*[0-9]*),(.*[0-9]*\\.*[0-9]*).")
newarrest=arrest%>%separate(Arrest.Date,into=c("month","date","year"),sep="/")%>%e
xtract(Location,into=c("long","lat"),"([0-9]*\\.*[0-9]*),(.*[0-9]*\\.*[0-9]*).")
newcensus=census%>%separate(Council.District,into=c("District.ID","NAME"),sep=" - ")
newcensus=as.data.table(newcensus)[order(as.numeric(District.ID))]

#1
##find out the top ten crimes
crime_freq_topten=as.data.table(newcrime)[year==2010,.SD,.SDcols=c(2,10,11)][,.(freq=
.N),by=.(Crime.Code,Crime.Code.Description)][order(-freq)][c(1:10)]
table=tableGrob(crime_freq_topten)
```

```
grid.arrange(table,top=textGrob("Top Ten Crimes in LA in 2010", vjust=9, hjust=1.1,
gp=gpar(fontsize=18)), nrow=1)

##find out the month with the most many frequencies
crime_freq_624=as.data.table(newcrime)[year==2010&Crime.Code==624,.SD,.SDcols=c(
2,10,11)][,.(freq=.N),by=.(month)][order(-freq)][c(1:10)]
ggplot(data=crime_freq_624,aes(x=month,y=freq,group="")) +
   geom_line()+
   labs(title='Simple Assault Crime in LA in 2010',y="frequency")+
   theme(plot.title = element_text(size=22))

##draw map
`2010crime`=as.data.table(newcrime)[year==2010,.SD,.SDcols=c(2,10,11,28,29)]
May_624crime=`2010crime`[Crime.Code==624&month=='05',]
numberr=nrow(district_map@data)
district_map@data$District.ID=1:numberr
total_map=merge(district_map,newcensus)
fort_map=fortify(district_map,region="District.ID")
new=merge(fort_map,newcensus,by.x='id', by.y='District.ID', all.x=TRUE)
pop_map=ggplot(new,aes(long,lat,group=group))+
   geom_polygon(color='black',fill='white')+
   geom_polygon(aes(x=long,y=lat, group=group, fill=Pop2010), data=new,
color='black')+
   scale_fill_gradient(low='lightblue',high='darkblue',name="population in 2010")
pop_map+
   geom_point(aes(x=as.numeric(lat),y=as.numeric(long),color='Crime.Code'),data=May
_624crime,inherit.aes=F)+
   scale_color_manual(values =c('Crime.Code'='#90FFFF'),name='simple assault crime')+
   labs(title="The Distribution of Crime")+
   theme(plot.title = element_text(size=22))

#2
##compute average times
avergt.crime=newcrime%>%select(Crime.Code.Description,Time.Occurred)%>%group_b
y(Crime.Code.Description)%>%summarise(averg=mean(Time.Occurred))
avergt.arrest=newarrest%>%select(Charge.Group.Code,Charge.Group.Description,Time)
```

```
%>%group_by(Charge.Group.Description)%>%summarise(averg=mean(Time))%>%filter(
averg!='NA')

##calculate the distance and do the visualization
as.matrix(avergt.crime)
dist_crime=dist(avergt.crime,diag=T,upper=T)
cmd_crime=cmdscale(dist_crime)
total_crime=cbind(as.tibble(cmd_crime),avergt.crime$Crime.Code.Description)
ggplot(total_crime,aes(x=V1,y=V2,color=avergt.crime$Crime.Code.Description))+
   geom_point()+
   theme(legend.position="none")+
   labs(title='The Distance Between Average Time of Each Crime Type')+
   theme(plot.title = element_text(size=19.5))
as.matrix(avergt.arrest)
dist_arrest=dist(avergt.arrest,diag=T,upper=T)
cmd_arrest=cmdscale(dist_arrest)
total_arrest=cbind(as.tibble(cmd_arrest),avergt.arrest$Charge.Group.Description)
ggplot(total_arrest,aes(x=V1,y=V2,color=avergt.arrest$Charge.Group.Description))+
   geom_point()+
   labs(title='The Distance Between Average Time of Each Arrest Type')+
   guides(fill=guide_legend(title="crime type"))+
   theme(plot.title = element_text(size=19), legend.position="none")
```