MILESTONE 2

Week 4

Abstract

This is the second milestone in this one. I will test a few characteristics of my data

Mohammad Hossein Movahedi

Movahedi.m@northeastern.edu

Contents

Milestone 1

Introduction

For this Milestone, I will use the data from Kaggle (Singh, 2014). The Robbery_2014_to_2019.csv dataset uses the APA system for the bibliography. Also, the R code that I used to process the data is available on my GitHub account, which address is mentioned in the bibliography part. I also put my code in the appendix part.

The objective is to get a sense of the data. Secondly, to find subsets of data, get descriptive statistics for each subgroup, and create visualizations for subset data.

The research question here is to determine the situation of rubbery crimes in Toronto. We will determine which part of Toronto is the most dangerous one.

Content and context of the dataset

This database contains 21543 rows of data about reported robberies in Toronto. Toronto is regarded as one of North America's safest cities. According to a study of more than 1,000 Canadians, the Economist Intelligence Unit's 2019 Safe Cities Index classified Toronto as the 6th safest city globally. (Patton, 2019)

With This dataset about Robbery, we can find out how the crime rate is going in Toronto, and we can also make subsets of locations to analyze them in detail.

This dataset has 27 columns, but most of them are about the time of the robbery and when it is reported (14 columns), and also four columns are identifiers, and seven columns are about the location of the theft, so this dataset needs a lot of cleaning before it can be used and it will be more effective if it is divided by subsets.

The content of the dataset is the following list

- Index: Record Unique Identifier
- eventuniqueid: Event Unique Identifier
- occurrencedate: Date of occurrence
- reporteddate: Date occurrence was reported
- premisetype: Premise where the occurrence took place
- ucrcode: URC Code ucrext: URC Code Extension
- offence : Offence related to the occurrence
- reportedyear : Year occurrence was reported
- reportedmonth : Month occurrence was reported

- reportedday : Day occurrence was reported
- reporteddayofyear : Day of week occurrence was reported
- reporteddayofweek : Day of year Occurrence was reported
- reportedhour : Hour occurrence was reported
- occurrenceyear : Occurrence year
- occurrencemonth : Occurrence month
- occurrenceday : Occurrence day
- occurrencedayofyear : Occurrence day of the year
- occurrencedayofweek : Occurrence day of week
- occurrencehour : Occurrence hour
- MCI : Major Crime Indicator related to the offence
- Division : Division where the event occurred
- Hood_ID : Neighbourhood Name
- Neighbourhood : Neighbourhood Identification
- Long : Longitude of point extracted after offsetting X and & Coordinates to nearest intersection node
- Lat : Latitude of point extracted after offsetting X and & Coordinates to nearest intersection node

Methodology

This milestone is divided into many parts, each dealing with one of the aspects of the project.

Part 1: Data cleaning

For this dataset, the data cleaning part is a challenge since there are many duplicated columns my goal here is to make three sub-tables one based on the type of the robbery, one based on the timing of the robbery and based on the location of the robberies

```
 # first of all I delete duplicate rows
rob <- data[!duplicated(data), ]
#Now I clean offence columns by deleting the "Robbery -" part
rob <- data %>%
  mutate_at("offence", str_replace, "Robbery - ", "")
#Now i delete the useless columns
nolist <- c("Index_","event_unique_id","occurrencedate","reporteddate",
            "ucr_code","ucr_ext","reportedyear","reportedmonth","reportedday",
            "reporteddayofyear","reporteddayofweek","reportedhour",
            "MCI","ObjectId")
rob <- rob[,!(names(rob) %in% nolist)]
#Now I combine time columns to make them one
rob$datatime <-paste(rob$occurrenceday," ",rob$occurrencemonth," ",
                      rob$occurrenceyear," ",rob$occurrencehour)

rob$datatime <- parse_date_time(rob$datatime, orders = "dmy_h")
#Now I can delete the rest of columns
notimelist <- c("occurrenceyear" ,"occurrencemonth","occurrenceday",
                "occurrencedayofyear","occurrencedayofweek","occurrencehour")
rob <- rob[,!(names(rob) %in% notimelist)]
rob1<- rob
```

```
#Now I set premise, Hood_Id,Neighbourhood  and Division  as factor

rob$premisetype<-as.factor(rob$premisetype)

rob$Division<-as.factor(rob$Division)

rob$Hood_ID<-as.factor(rob$Hood_ID)

rob$Neighbourhood<-as.factor(rob$Neighbourhood)

# Now I group by neighbourhood

Mode <- function(x) {

  ux <- unique(x)

  ux[which.max(tabulate(match(x, ux)))]

}

arrange(desc(number_player))

Hood<- rob %>%

  group_by(Neighbourhood,Hood_ID) %>%

  summarize(NumberOfRobbery = n(),MostOffence = Mode(offence),Division = Mode(Division),Datemean
= median(datatime,na.rm = T),

            Long = mean(Long),Lat = mean(Lat))

# Now I group by offence

Offence<- rob %>%

  group_by(offence) %>%

  summarize(NumberOfRobbery = n(),MostHood = Mode(Neighbourhood),MostHood_ID =
Mode(Hood_ID),Division = Mode(Division),Datemean = median(datatime,na.rm = T),

            Long = mean(Long),Lat = mean(Lat))

# Now I group by date

Date<- rob %>%

  group_by(year(datatime),month(datatime)) %>%

  summarize(NumberOfRobbery = n(),MostOffence = Mode(offence),MostHood =
Mode(Neighbourhood),MostHood_ID = Mode(Hood_ID),Division = Mode(Division),

            Long = mean(Long),Lat = mean(Lat))
```

Now that we cleaned the data and created the tables, we can continue to the next step.

Part 2: Data analysis

Now we run some descriptive analytics on data. the First thing that I want to see is whether are the numbers increasing each year or not? and what is the crime change rate each month

```
 #New we calculate crime rate

ggplot(Date, aes(x= ym , y =NumberOfRobbery)) + geom_point()+geom_smooth()+

  labs(title ="Number Of Robbery in time", x="", y = "Number of Robbery")

lead(Date$`month(datatime)`)



Date <- Date %>%

  mutate(crime_change = (NumberOfRobbery/lead(NumberOfRobbery) - 1) * 100)

# now I delete outliers

boxplot(Date$crime_change)$out

Date$crime_change <- rm.outlier(Date$crime_change, fill = TRUE, median = FALSE, opposite =
FALSE)

ggplot(Date, aes(x= ym , y =crime_change)) + geom_point()+geom_smooth()+

  labs(title ="crime change", x="",y="changes")
```

according to my calculations ang graph 1 there was no significant changes.

Now I use Hood table to see weather there is connection between crimes in neighborhood and divisions or not .

```
 # now testing connection between crimes and locations

p<-ggplot(data=Hood, aes(x=Division, y=NumberOfRobbery,fill=MostOffence)) +

  geom_bar(stat="identity")

p
```

after creating bar plot 1 I found out most crimes were mugging

Part 3: Discussion and conclusion

This database showed me that the crime rate isn’t changing a lot in Toronto and the most common type of robbery is mugging. After analyzing this database, I feel safer in Toronto now that I know the crime rate is so low.

Furthermore, we will test the neighbourhoods to decide which part of Toronto is better to live in and what is the relationship between locations and crimes.

Part4: graphs, Tables and charts

```
 > head (Hood)

# A Tibble: 6 × 8

# Groups:   Neighbourhood [6]

  Neighbourhood    Hood_ID NumberOfRobbery MostOffence Division Datemean             Long

  <fct>            <fct>             <int> <fct>       <fct>    <dttm>              <dbl>

1 Agincourt North… 129                 181 Mugging     D42      2017-08-07 21:00:00 -79.3

2 Agincourt South… 128                 164 Mugging     D42      2017-05-26 00:00:00 -79.3

3 Alderwood (20)   20                   41 Mugging     D22      2016-12-09 18:00:00 -79.5

4 Annex (95)       95                  245 Mugging     D53      2017-05-18 11:00:00 -79.4

5 Banbury-Don Mil… 42                   90 Mugging     D33      2016-04-27 12:30:00 -79.3

6 Bathurst Manor … 34                   56 Robbery Wi… D32      2017-05-06 12:00:00 -79.5

# … with 1 more variable: Lat <dbl>
```

Table 1 : Hood table

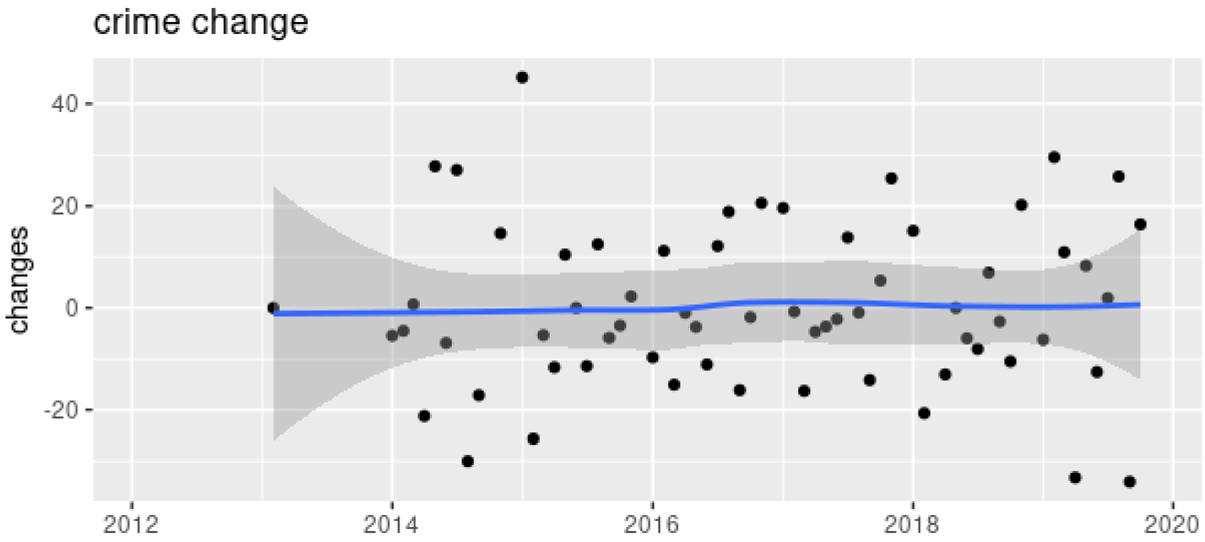
```
 > head (Offence)

# A tibble: 6 × 8

  offence    NumberOfRobbery MostHood MostHood_ID Division Datemean             Long   Lat

  <chr>                <int> <fct>    <fct>       <fct>    <dttm>              <dbl> <dbl>

1 Armoured…               33 Bedford… 39          D32      2016-04-22 04:00:00 -79.4  43.7

2 Atm                     76 Church-… 75          D51      2017-05-20 02:00:00 -79.4  43.7

3 Business              2434 Church-… 75          D51      2017-05-14 03:00:00 -79.4  43.7

4 Delivery…              215 York Un… 27          D31      2017-11-26 02:00:00 -79.4  43.7
```

```
5 Financia…              644 Bay Str… 76           D22       2017-01-24 10:00:00 -79.4  43.7

6 Home Inv…              830 Waterfr… 77           D43       2016-11-02 02:00:00 -79.4  43.7
```

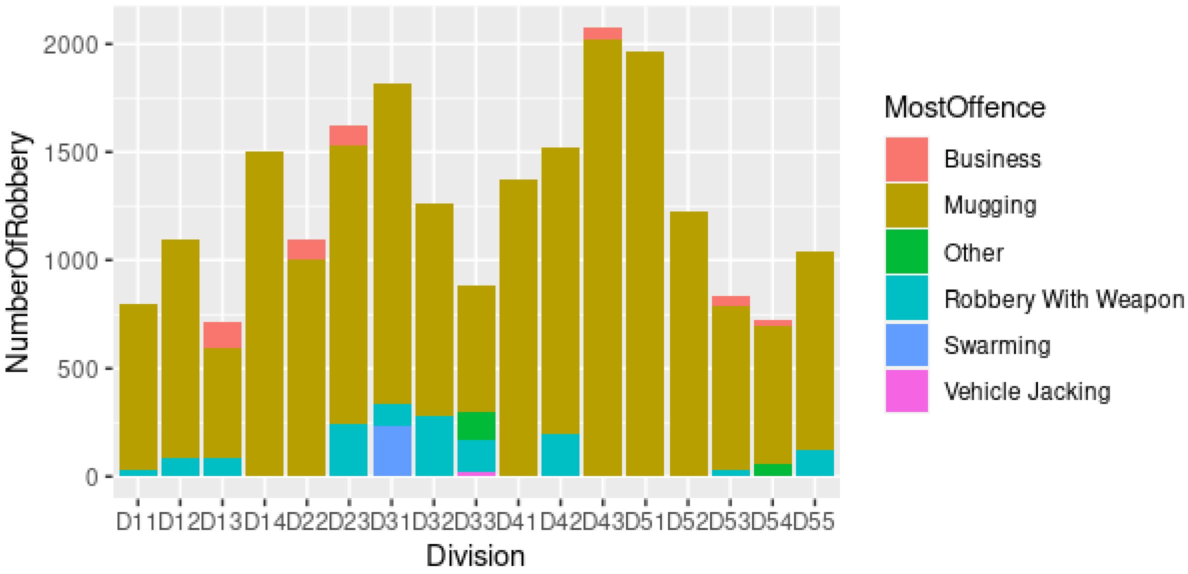
Table 2 : Offences

```
 > head (Date)

# A tibble: 6 × 11

# Groups:   year(datatime) [2]

  `year(datatime)` `month(datatime)` NumberOfRobbery MostOffence     MostHood MostHood_ID

             <dbl>             <dbl>           <int> <chr>           <fct>    <fct>

1             2012                 2               1 Mugging         Keelesd… 110

2             2013                 1               2 Mugging         Moss Pa… 73

3             2013                 2               1 Mugging         West Hi… 136

4             2013                 6               1 Other           Church-… 75

5             2013                 8               2 Other           Moss Pa… 73

6             2013                 9              18 Robbery With W… Waterfr… 77

# … with 5 more variables: Division <fct>, Long <dbl>, Lat <dbl>, ym <dttm>,
```

Table 3 : Offence



Graph 1 Crime Changes



Graph 2 Crime vs Division

Milestone 2

introduction

According to the findings of milestone one, the most common kind of robbery in Toronto is mugging, and also the most dangerous neighbourhood in Toronto is Church-Yonge Corridor. However, there are a few questions that are still unanswered.

1- Is there a difference between the number of crimes in neighbourhoods in the east of Toronto and west of Toronto?
2- Is there a difference between the number of crimes in the district in the north of Toronto and south of Toronto?
3- Is time a factor in the number of crimes and do as common belief more crime happens at night compared to the day in Toronto's neighbourhoods?

Methodology

First of all, we need to define which neighbourhoods are on each side of Toronto. We calculate the mean of all latitudes and then compare them to each area's mean and group the neighbourhoods.

```
 clong <- mean(rob$Long)

clat<- mean(rob$Lat)

ehoods<-filter(Hood,Long >= clong)

whoods<-filter(Hood,Long < clong)

nhoods<-filter(Hood,Lat >= clat)

shoods<-filter(Hood,Lat < clat)
```

now we can run a t-test for question one

```
 > t.test(ehoods$NumberOfRobbery,whoods$NumberOfRobbery)



         Welch Two Sample t-test



data:  ehoods$NumberOfRobbery and whoods$NumberOfRobbery

t = 1.6073, df = 99.659, p-value = 0.1112

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -9.259877 88.258434

sample estimates:

mean of x mean of y

 175.6032  136.1039
```

As seen in the results of the t-test, the p-value is 11% which is more than 5%, so we fail to reject the null hypothesis. There is no substantial evidence suggesting that neighbourhoods in the east of Toronto are safer or more dangerous than those in the west.

Now we test if there is a difference in the crimes in the north of Toronto compering to the south of Toronto.

```
 > t.test(nhoods$NumberOfRobbery,shoods$NumberOfRobbery)

         Welch Two Sample t-test

data:  nhoods$NumberOfRobbery and shoods$NumberOfRobbery

t = 0.74196, df = 137.3, p-value = 0.4594

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -28.80301  63.39937

sample estimates:

mean of x mean of y

 163.5161  146.2179
```

The p-value of this test is 45%, which is considered a significant p-value suggesting that we fail to reject the null hypothesis with more excellent confidence than the previous test.

To answer the third question, we will test whether or not the common belief is that the streets are more dangerous at night than in the day.

To do this, first, we subset all robberies that happened outside, then divide them into two groups by the time they happened and then we do a two-sample t-test on the results.

```
 #testing based on time

robout<- filter(rob,rob$premisetype =="Outside")

robday<- filter(robout,(hour(datatime)>=6 & hour(datatime) < 18))

robnight<- filter(robout,(hour(datatime)>18 | hour(datatime) < 6))

#grouping

Hoodday<- robday %>%

  group_by(Neighbourhood,Hood_ID) %>%

  summarize(NumberOfRobbery = n(),MostOffence = Mode(offence),Division = Mode(Division),Datemean
= Mode(datatime),

            Long = mean(Long),Lat = mean(Lat))

Hoodnight<- robnight %>%

  group_by(Neighbourhood,Hood_ID) %>%

  summarize(NumberOfRobbery = n(),MostOffence = Mode(offence),Division = Mode(Division),Datemean
= Mode(datatime),

            Long = mean(Long),Lat = mean(Lat))
```

Because these data are for the same neighbourhoods, we can use paired t-test to calculate the numbers.

```
 > t.test(Hoodday$NumberOfRobbery,Hoodnight$NumberOfRobbery,paired = T)



         Paired t-test



data:  Hoodday$NumberOfRobbery and Hoodnight$NumberOfRobbery

t = -7.5127, df = 139, p-value = 6.399e-12
```

```
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -28.54783 -16.65217

sample estimates:

mean of the differences

                  -22.6
```

So, according to the results of the t-test, there is a difference between the number of crimes, and we reject the null hypothesis that there is no difference. As the results show, nights are more dangerous than days.

Summery

In summery, There is no indication that the neighbourhoods to the east of Toronto are safer or more dangerous than those to the west.

Also the test comparing north and south has a p-value of 45 percent, which is considered a significant p-value, indicating that we are unable to reject the null hypothesis with greater confidence than the prior test.

Finally, the t-test findings show that there is a difference in the number of offences, rejecting the null hypothesis that there is no difference. The data reveal that nights are riskier than days.

Bibliography

Patton, J. (2019). Toronto ranked among the safest cities in the world by Economist Intelligence Unit. [online] Global News. Available at: https://globalnews.ca/news/5829962/toronto-safest-cities-index-2019/ [Accessed 9 Mar. 2022].

Singh, A. (2014). Toronto Robbery 2014-2019. [online] Kaggle.com. Available at: https://www.kaggle.com/cosmicakshh/toronto-robbery-20142019 [Accessed 9 Mar. 2022].

momova97 (2022). momova97/ALY6010_Movahedi: This is the place that I will keep my projects R code. [online] GitHub. Available at: https://github.com/momova97/ALY6010_Movahedi [Accessed 8 Mar. 2022].

Appendix

```
 print("Mohammad Hossein Movahedi")
print("Milestone 2")
#importing and instaling libraries
install.packages('FSA')
install.packages('FSAdata')
install.packages('magrittr')
install.packages('dplyr')
install.packages('tidyr')
install.packages('plyr')
install.packages('tidyverse')
install.packages('outliers')
install.packages('ggplot2')
install.packages('lubridate')

library(ggplot2)
library(outliers)
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(tidyr)
library(dplyr)
library(tidyverse)
library(scales)
library(lubridate)
#importing dataset
```

```
data <-
read.csv("https://raw.githubusercontent.com/momova97/ALY6010_Movahedi/main/Robbery_2014_to_
2019.csv")

# first of all I delete duplicate rows

rob <- data[!duplicated(data), ]

#Now I clean offence columns by deleting the "Robbery -" part

rob <- data %>%

  mutate_at("offence", str_replace, "Robbery - ", "")

#Now i delete the useless columns

nolist <- c("Index_","event_unique_id","occurrencedate","reporteddate",

            "ucr_code","ucr_ext","reportedyear","reportedmonth","reportedday",

            "reporteddayofyear","reporteddayofweek","reportedhour",

            "MCI","ObjectId")

rob <- rob[,!(names(rob) %in% nolist)]

#Now I combine time columns to make them one

rob$datatime <-paste(rob$occurrenceday," ",rob$occurrencemonth," ",

                      rob$occurrenceyear," ",rob$occurrencehour)



rob$datatime <- parse_date_time(rob$datatime, orders = "dmy_h")

#Now I can delete the rest of columns

notimelist <- c("occurrenceyear" ,"occurrencemonth","occurrenceday",

                "occurrencedayofyear","occurrencedayofweek","occurrencehour")

rob <- rob[,!(names(rob) %in% notimelist)]

rob1<- rob

#Now I set premise, Hood_Id,Neighbourhood  and Division  as factor

rob$premisetype<-as.factor(rob$premisetype)

rob$Division<-as.factor(rob$Division)

rob$Hood_ID<-as.factor(rob$Hood_ID)

rob$Neighbourhood<-as.factor(rob$Neighbourhood)

# Now I group by neighbourhood
```

```
Mode <- function(x) {

  ux <- unique(x)

  ux[which.max(tabulate(match(x, ux)))]

}



Hood<- rob %>%

  group_by(Neighbourhood,Hood_ID) %>%

  summarize(NumberOfRobbery = n(),MostOffence = Mode(offence),Division =
Mode(Division),Datemean = Mode(datatime),

            Long = mean(Long),Lat = mean(Lat))

head (Hood)

# Now I group by offence

Offence<- rob %>%

  group_by(offence) %>%

  summarize(NumberOfRobbery = n(),MostHood = Mode(Neighbourhood),MostHood_ID =
Mode(Hood_ID),Division = Mode(Division),Datemean = Mode(datatime),

            Long = mean(Long),Lat = mean(Lat))

head (Offence)

# Now I group by date

Date<- rob %>%

  group_by(year(datatime),month(datatime)) %>%

  summarize(NumberOfRobbery = n(),MostOffence = Mode(offence),MostHood =
Mode(Neighbourhood),MostHood_ID = Mode(Hood_ID),Division = Mode(Division),

            Long = mean(Long),Lat = mean(Lat))

Date <- Date[-c(1,82),]

Date$ym <- paste(Date$`year(datatime)`,"-",Date$`month(datatime)`)

Date$ym <- parse_date_time(Date$ym,order = "ym")

head (Date)

#New we calculate crime rate

ggplot(Date, aes(x= ym , y =NumberOfRobbery)) + geom_point()+geom_smooth()+
  labs(title ="Number Of Robbery in time", x="", y = "Number of Robbery")

lead(Date$`month(datatime)`)



Date <- Date %>%

  mutate(crime_change = (NumberOfRobbery/lead(NumberOfRobbery) - 1) * 100)

# now I delete outliers

boxplot(Date$crime_change)$out

Date$crime_change <- rm.outlier(Date$crime_change, fill = TRUE, median = FALSE, opposite =
FALSE)

ggplot(Date, aes(x= ym , y =crime_change)) + geom_point()+geom_smooth()+

  labs(title ="crime change", x="",y="changes")

# now testing connection between crimes and locations

p<-ggplot(data=Hood, aes(x=Division, y=NumberOfRobbery,fill=MostOffence)) +

  geom_bar(stat="identity")

p

#milestone 2

print("milestone 2")

#calculate the mean of all latitudes and then compare them to each area’s mean and group
the neighbourhoods.

clong <- mean(rob$Long)

clat<- mean(rob$Lat)

ehoods<-filter(Hood,Long >= clong)

whoods<-filter(Hood,Long < clong)

nhoods<-filter(Hood,Lat >= clat)

shoods<-filter(Hood,Lat < clat)

#testing part

t.test(ehoods$NumberOfRobbery,whoods$NumberOfRobbery)

t.test(nhoods$NumberOfRobbery,shoods$NumberOfRobbery)

#testing based on time

robout<- filter(rob,rob$premisetype =="Outside")
```

```
robday<- filter(robout,(hour(datatime)>=6 & hour(datatime) < 18))

robnight<- filter(robout,(hour(datatime)>18 | hour(datatime) < 6))

#grouping

Hoodday<- robday %>%

  group_by(Neighbourhood,Hood_ID) %>%

  summarize(NumberOfRobbery = n(),MostOffence = Mode(offence),Division =
Mode(Division),Datemean = Mode(datatime),

            Long = mean(Long),Lat = mean(Lat))

Hoodnight<- robnight %>%

  group_by(Neighbourhood,Hood_ID) %>%

  summarize(NumberOfRobbery = n(),MostOffence = Mode(offence),Division =
Mode(Division),Datemean = Mode(datatime),

            Long = mean(Long),Lat = mean(Lat))

#testing

t.test(Hoodday$NumberOfRobbery,Hoodnight$NumberOfRobbery,paired = T)
```