Chicago Crime Data Analysis

Pankaj, Sakshi and Prafful

12/19/2021

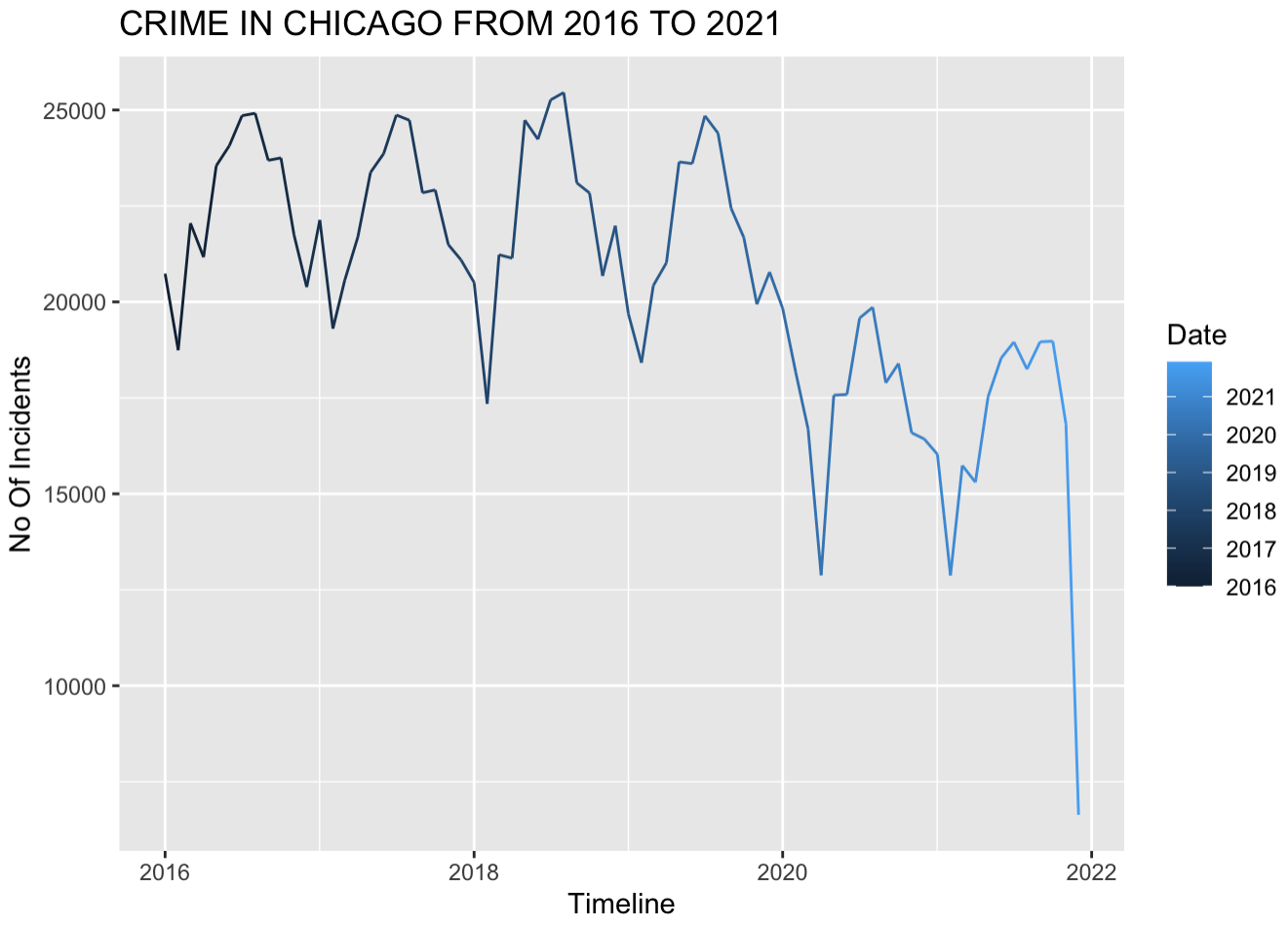
Importing Data and packages

```
# add packages you need for this assignment
library("xlsx")
library(tidyverse)  # includes tibbles, ggplot2, dyplr, and more.
```

Importing required files from the path

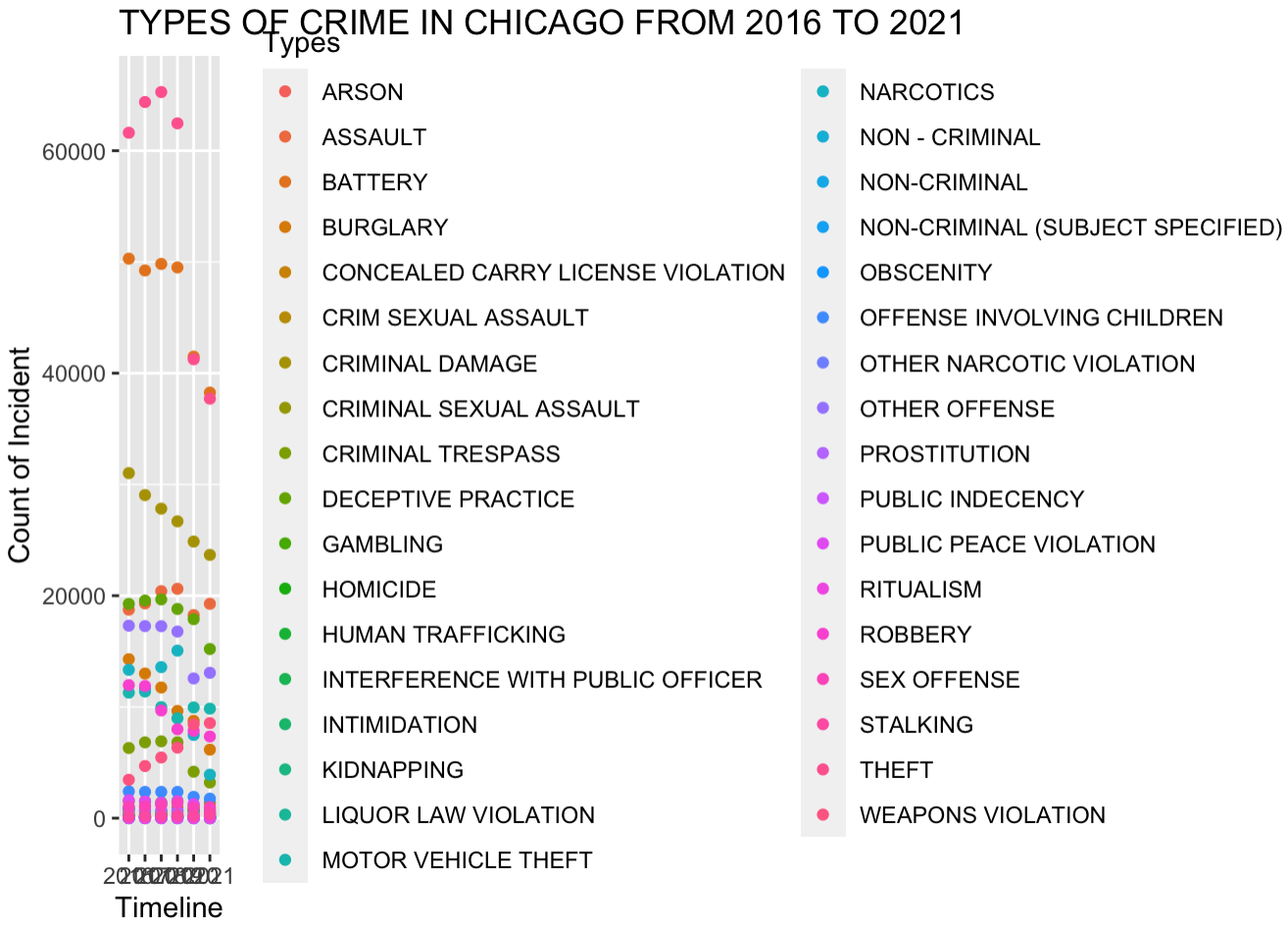
```
ccdata<- read.csv("~/Desktop/CC/ChicagoCrimeData20162021.csv")
crime<- read.csv("~/Desktop/CC/crime1.csv", header=TRUE)
yearcc <- read.xlsx("~/Desktop/CC/yeardata16-21.xlsx", 1, header=TRUE)
```

```
plot1 <-ggplot(yearcc, aes(x=Date)) +
        geom_line(aes(y = Year, color = Date))
print(plot1 +ggtitle("CRIME IN CHICAGO FROM 2016 TO 2021")+ labs(y="No Of Incidents", x = "Timelin
e"))
```



The above graph illustrates number of incidents occurred each year from 2016 to 2021. It clearly depicts that the highest number of incidents happened in 2018 and every year follows the same pattern. In the begning of every year there is a significant decline in number of incidents, as you can see in the Jan-Feb of 2016 the numbers are least and it gradually increases from March to August and achieve its peak during September to November and starts declining again from December. Suprisingly this pattern is identical for every year, we might notice overall decline in cases from 2020 but the pattern is exactly same.

```
CrimeType <- table(ccdata$Primary.Type, ccdata$Year) ## Making a table between Primary crime type
 and year in which it occurred.
typecc <- data.frame(CrimeType) # making it a data frame

plot2 <- ggplot(data = typecc) +
  geom_point(mapping = aes(x = Var2, y = Freq,colour = Var1))
print(plot2 +ggtitle("TYPES OF CRIME IN CHICAGO FROM 2016 TO 2021")+ labs(y="Count of Incident", x
= "Timeline", color = "Types"))
```



The above graph is to determine the type of incident that happend during the time period of 2016 - 2021. From the graph we can conclude that “Theft” as a crime type has highest occurence followed by “Battery”,“Criminal Damage” “Deceptive Practice” adn so on. While these were some top crimes commited there are some types which were less than 100. To better understand the figures following is the tabular view of the types of crime segregated by Year, sorted alphabatically.

```
CrimeType
```

```
##
##                                    2016  2017  2018  2019  2020  2021
##   ARSON                             516   444   373   375   588   496
##   ASSAULT                         18741 19306 20406 20617 18251 19269
##   BATTERY                         50297 49237 49822 49508 41490 38256
##   BURGLARY                        14289 13000 11747  9638  8750  6158
##   CONCEALED CARRY LICENSE VIOLATION  36    69   149   217   148   171
##   CRIM SEXUAL ASSAULT              1513  1529  1427   921    75     0
##   CRIMINAL DAMAGE                 31018 29044 27822 26681 24865 23667
##   CRIMINAL SEXUAL ASSAULT            97   138   266   707  1138  1368
##   CRIMINAL TRESPASS                6306  6815  6907  6818  4175  3207
##   DECEPTIVE PRACTICE              19260 19564 19667 18806 17888 15207
##   GAMBLING                          189   191   201   142    25    13
##   HOMICIDE                          790   676   601   506   791   770
##   HUMAN TRAFFICKING                  11    10    12    14     5    12
##   INTERFERENCE WITH PUBLIC OFFICER  936  1087  1306  1546   654   298
##   INTIMIDATION                      135   150   168   163   162   105
##   KIDNAPPING                        202   190   172   173   119    88
##   LIQUOR LAW VIOLATION              227   191   268   232   143   170
##   MOTOR VEHICLE THEFT             11286 11380  9983  8976  9952  9843
##   NARCOTICS                       13333 11674 13578 15060  7484  3907
##   NON - CRIMINAL                      5     0     0     0     0     0
##   NON-CRIMINAL                       49    37    36     4     1     4
##   NON-CRIMINAL (SUBJECT SPECIFIED)    1     2     3     0     0     0
##   OBSCENITY                          51    86    87    59    55    46
##   OFFENSE INVOLVING CHILDREN       2406  2360  2360  2359  1912  1752
##   OTHER NARCOTIC VIOLATION            4    11     1     8     6     2
##   OTHER OFFENSE                   17304 17263 17258 16775 12560 13066
##   PROSTITUTION                      800   735   718   680   277    73
##   PUBLIC INDECENCY                   10    10    14    11     9     3
##   PUBLIC PEACE VIOLATION           1607  1498  1372  1520  1270   574
##   RITUALISM                           0     0     0     0     1     0
##   ROBBERY                         11960 11880  9679  7994  7853  7342
##   SEX OFFENSE                      1028  1050  1168  1346   939  1002
##   STALKING                          176   191   207   226   200   340
##   THEFT                           61617 64377 65275 62461 41252 37710
##   WEAPONS VIOLATION                3450  4686  5456  6339  8429  8545
```

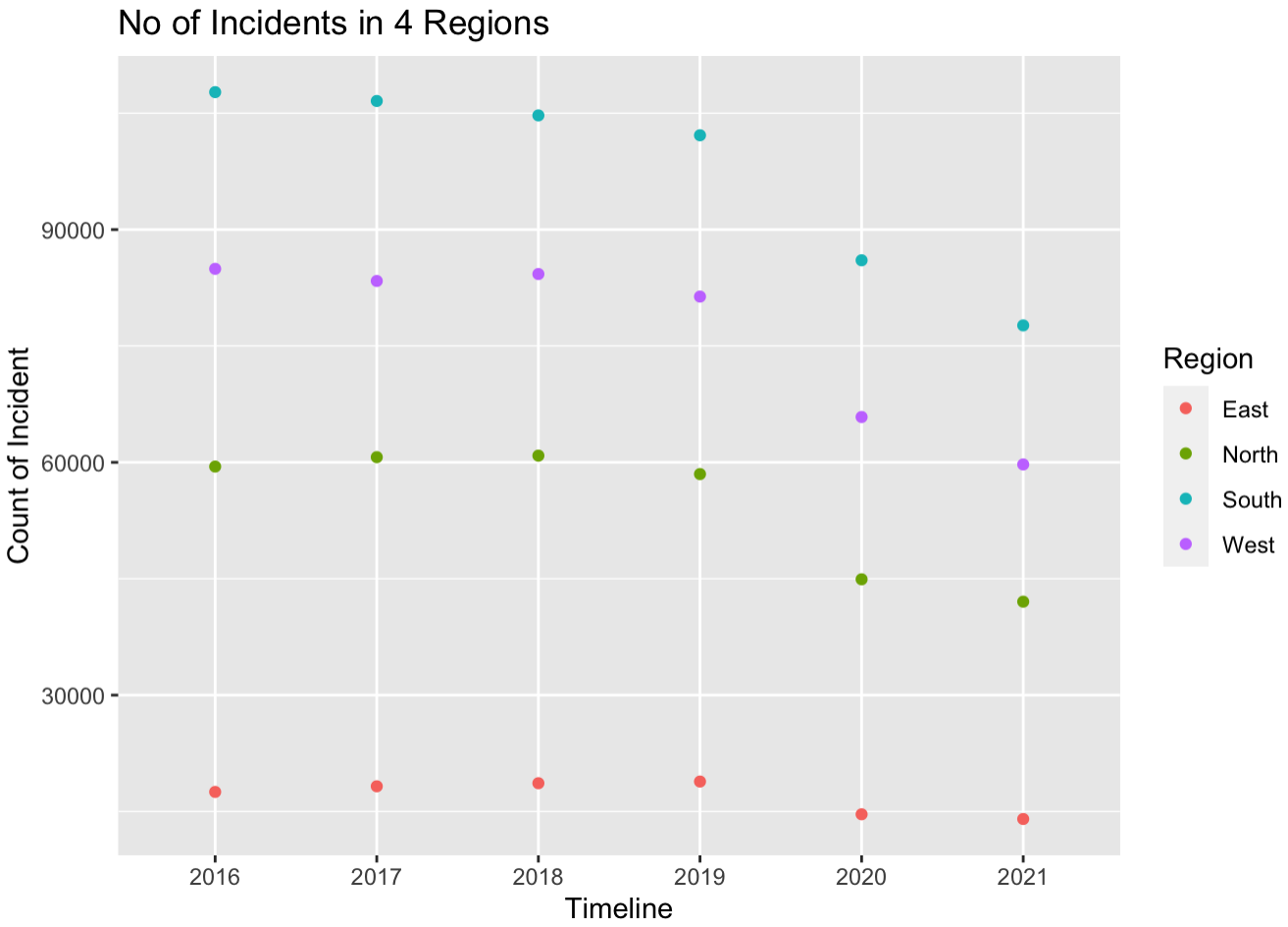
```
tt1 <- table(ccdata$YEAR,ccdata$REGION)
tt1
```

```
## < table of extent 0 x 0 >
```

```
head(ccdata$Block)
```

```
## [1] "028XX W 22ND PL"           "008XX S INDEPENDENCE BLVD"
## [3] "011XX N AVERS AVE"         "029XX N HOYNE AVE"
## [5] "002XX N DEARBORN ST"       "049XX N KILDARE AVE"
```

```
tt1 <- table(crime$YEAR,crime$REGION)

regioncc <- data.frame(tt1)

names(regioncc)[1] <- "Year"
names(regioncc)[2] <- "Region"
names(regioncc)[3] <- "Count"

regioncc$Region <- as.character(regioncc$Region)
regioncc$Year <- as.character.Date(regioncc$Year)

plot4 <- ggplot(data = regioncc) +
  geom_point(mapping = aes(x = Year, y = Count,colour = Region))
print(plot4 +ggtitle("No of Incidents in 4 Regions")+ labs(y="Count of Incident", x = "Timeline"))
```



The above graph is between number of incidents recorded and region in which it took place. The scatter plot clearly illustrates that highest number of incidents were in the South region followed by West, North and East. However` we can notice decline in the recorded cases as we progress in the year but the sequence of region is identical every year.

```
tt2 <- table(ccdata$Year,ccdata$Location.Description)
locationcc <- data.frame(tt2)

names(locationcc)[1] <- "Year"
names(locationcc)[2] <- "Location"
names(locationcc)[3] <- "Count"

locationcc$Year <- as.character.Date(locationcc$Year)
locationcc$Location <- as.character(locationcc$Location)
locationcc$Count <- as.numeric(locationcc$Count)

loc <- subset(locationcc,locationcc$Location == "APARTMENT" | locationcc$Location == "RESIDENCE" |
locationcc$Location == "STREET" | locationcc$Location == "SIDEWALK"  )
loc
```

```
##      Year  Location Count
## 109  2016 APARTMENT 34474
## 110  2017 APARTMENT 33591
## 111  2018 APARTMENT 34800
## 112  2019 APARTMENT 34948
## 113  2020 APARTMENT 36004
## 114  2021 APARTMENT 41268
## 847  2016 RESIDENCE 46200
## 848  2017 RESIDENCE 46098
## 849  2018 RESIDENCE 45170
## 850  2019 RESIDENCE 43252
## 851  2020 RESIDENCE 38671
## 852  2021 RESIDENCE 29767
## 973  2016  SIDEWALK 23498
## 974  2017  SIDEWALK 21011
## 975  2018  SIDEWALK 21168
## 976  2019  SIDEWALK 20344
## 977  2020  SIDEWALK 13410
## 978  2021  SIDEWALK 11248
## 1003 2016    STREET 60943
## 1004 2017    STREET 59977
## 1005 2018    STREET 59060
## 1006 2019    STREET 56490
## 1007 2020    STREET 50469
## 1008 2021    STREET 49157
```

```
data2016<-subset(crime,crime$YEAR == 2016)
data2017<-subset(crime,crime$YEAR == 2017)
data2018<-subset(crime,crime$YEAR == 2018)
data2019<-subset(crime,crime$YEAR == 2019)
data2020<-subset(crime,crime$YEAR == 2020)
data2021<-subset(crime,crime$YEAR == 2021)
```

```
crime2016<-data.frame( table(data2016$PRIMARYTYPE))
names(crime2016)[1]<-'CrimeType'


crime2017<-data.frame( table(data2017$PRIMARYTYPE))
names(crime2017)[1]<-'CrimeType'


crime2018<-data.frame( table(data2018$PRIMARYTYPE))
names(crime2018)[1]<-'CrimeType'

crime2019<-data.frame( table(data2019$PRIMARYTYPE))
names(crime2019)[1]<-'CrimeType'


crime2020<-data.frame( table(data2020$PRIMARYTYPE))
names(crime2020)[1]<-'CrimeType'


crime2021<-data.frame( table(data2021$PRIMARYTYPE))
names(crime2021)[1]<-'CrimeType'
```

HYPOTHESIS 1

To check whether the crime type committed most in year 2016 is the same type of crime committed in year 2017 using Hypothesis Testing. Here we are applying prop-test through which we will get the p-value. And on the basis of p-value we can come to a conclusion. We are using prop-test because to get accurate data with respect to the total number of crime.

$$H_0 : \mu 1 = \mu 2,$$
$$H_1 : \mu 1 \neq \mu 2.$$

H_0 is Null Hypothesis and H_1 is Alternative Hypothesis.

```
#Hypothesis 1

crime2016$CrimeType[which.max(crime2016$Freq)] # Primary type in 2016
```

```
## [1] THEFT
## 34 Levels: ARSON ASSAULT BATTERY BURGLARY ... WEAPONS VIOLATION
```

```
theft2016<-subset(crime2016$Freq,crime2016$CrimeType == 'THEFT')
theft2016
```

```
## [1] 61617
```

```
theft2017<-subset(crime2017$Freq,crime2017$CrimeType == 'THEFT')
theft2017
```

```
## [1] 64377
```

```
prop.test(x = c(theft2016,theft2017), n = c(nrow(data2016),nrow(data2017)), alternative = "greate
r", conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(theft2016, theft2017) out of c(nrow(data2016), nrow(data2017))
## X-squared = 89.49, df = 1, p-value = 1
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.01281968  1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.2285073 0.2394256
```

For the year 2016 THEFT is the most committed crime. The p-value of this hypothesis is greater than alpha i.e., 0.05. So we can accept the null hypothesis H_0 and agree that crime type committed most in 2016 is the same crime type committed in 2017.

HYPOTHESIS 2

To check whether the crime committed most in a Region in year 2016 is the same Region for the year 2017 using Hypothesis Testing. Here we are applying prop-test through which we will get the p-value. And on the basis of p-value we can come to a conclusion.

$$H_0 : \mu 1 = \mu 2,$$
$$H_1 : \mu 1 \neq \mu 2.$$

H_0 is Null Hypothesis and H_1 is Alternative Hypothesis.

```
Region2016<-data.frame(table(data2016$REGION))
names(Region2016)[1]<-'Region'
Region2016
```

```
##   Region   Freq
## 1   East  17524
## 2  North  59461
## 3  South 107720
## 4   West  84945
```

```
Region2017<-data.frame(table(data2017$REGION))
names(Region2017)[1]<-'Region'
Region2017
```

```
##   Region   Freq
## 1   East  18250
## 2  North  60671
## 3  South 106581
## 4   West  83379
```

```
Region2016$Region[which.max(Region2016$Freq)] # region in 2016
```

```
## [1] South
## Levels: East North South West
```

```
South2016<-subset(Region2016$Freq,Region2016$Region == 'South')
South2016
```

```
## [1] 107720
```

```
South2017<-subset(Region2017$Freq,Region2017$Region == 'South')
South2017
```

```
## [1] 106581
```

```
prop.test(x = c(South2016,South2017), n = c(nrow(data2016),nrow(data2017)), alternative = "greate
r", conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(South2016, South2017) out of c(nrow(data2016), nrow(data2017))
## X-squared = 5.365, df = 1, p-value = 0.01027
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.0008956395 1.0000000000
## sample estimates:
##    prop 1    prop 2
## 0.3994808 0.3963872
```

The p-value of this hypothesis is smaller than alpha i.e., 0.05 but mu1 is equal to mu2.So, we can accept the null hypothesis H_0 and agree that the Region in which crime committed most in 2016 is the same Region for year 2017 where crime were committed most i.e., Region South.

HYPOTHESIS 3

To check whether the domestic crime rate in 2020 increased or not in comparison to 2019 we are using Hypothesis Testing. As per my assumption domestic cases should increase because during lockdown people mostly stayed at home. Here we are applying prop-test through which we will get the p-value. And on the basis of p-value we can come to a conclusion.

$$H_0 : \mu 1 = \mu 2,$$
$$H_1 : \mu 1 \neq \mu 2.$$

H_0 is Null Hypothesis and H_1 is Alternative Hypothesis.

```
#Hypothesis 3
Domestic2019<-subset(crime,crime$YEAR == 2019 & crime$DOMESTIC_01 == 1)
nrow(Domestic2019)
```

```
## [1] 43249
```

```
Domestic2020<-subset(crime,crime$YEAR == 2020 & crime$DOMESTIC_01 == 1)
nrow(Domestic2020)
```

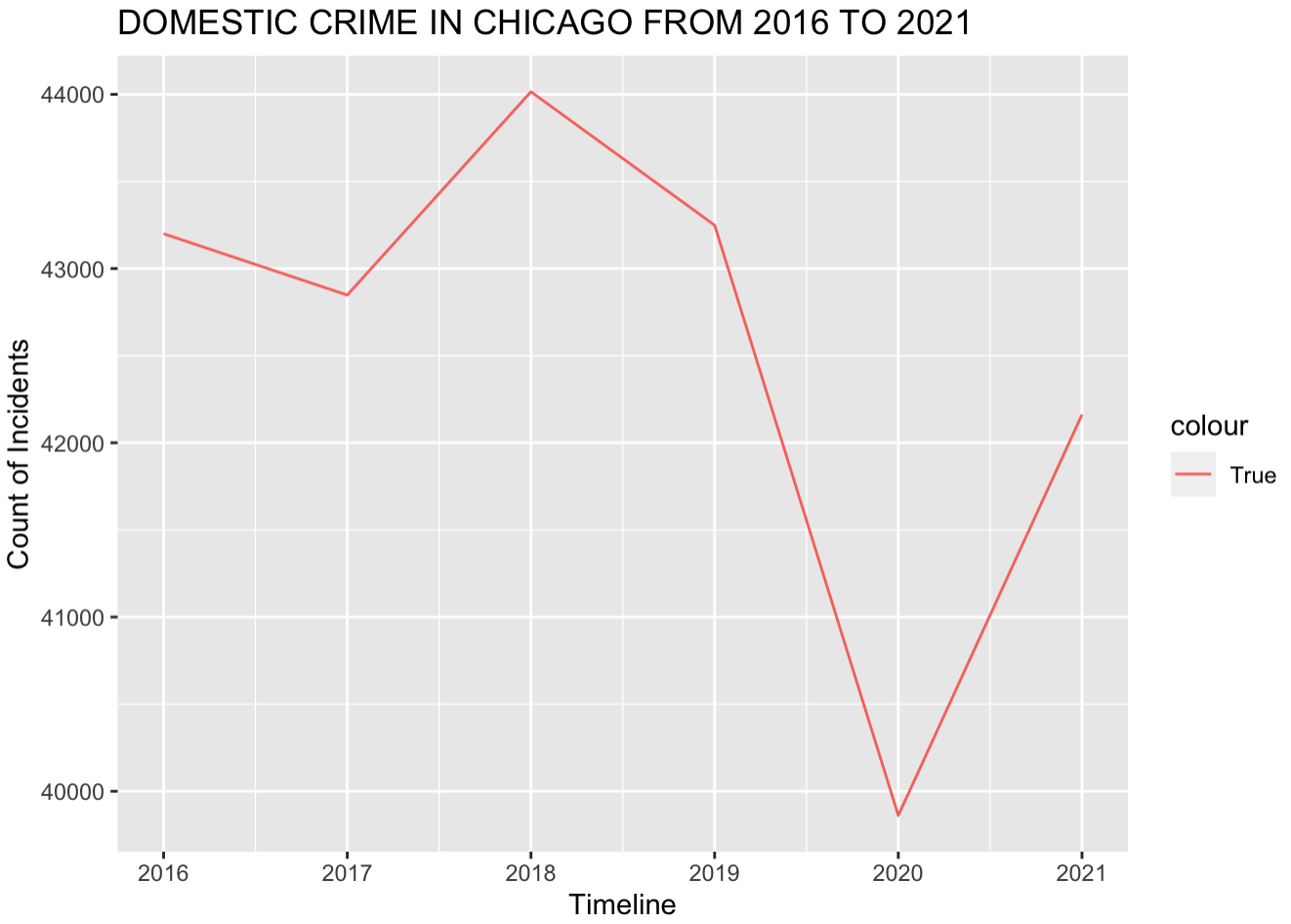
```
## [1] 39861
```

```
prop.test(x = c(nrow(Domestic2019),nrow(Domestic2020)), n = c(nrow(data2019),nrow(data2020)), alte
rnative = "greater", conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(nrow(Domestic2019), nrow(Domestic2020)) out of c(nrow(data2019), nrow(data2020))
## X-squared = 415.57, df = 1, p-value = 1
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.0245634  1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.1657799 0.1884975
```

The p-value of this hypothesis is greater than alpha i.e., 0.05 so, we can accept the NULL hypothesis H_0 and agree that the Domestic Violence cases increased in year 2020 in comparison to the year of 2019. Additonally I am using line graph to visualize the pattern of domestic cases from 2016 - 2021.

```
domesticcc <- read.xlsx("~/Desktop/CC/Domesticcc.xlsx", 1, header=TRUE)

plot3 <-ggplot(domesticcc, aes(x=Year)) +
        geom_line(aes(y = TRUE., color = "True"))
print(plot3 +ggtitle("DOMESTIC CRIME IN CHICAGO FROM 2016 TO 2021")+ labs(y="Count of Incidents",
 x = "Timeline"))
```



From the graph we can see that the cases around 43000 in 2016 and went upto 44000 in 2018 but we saw a sudden decline in cases in 2019 achieving its all time low of 40000. However, the number started increasing from the begning of 2020 and incrased by 2000 in 2021 which directly relate to lockdown amid pandemic.

HYPOTHESIS 4

To check whether the number of arrest in year 2017 increased or decreased in comparsion to year 2016. We will check this by using Hypothesis Testing. Here we are applying prop-test through which we will get the p-value. And on the basis of p-value we can come to a conclusion.

$$H_0 : \mu 1 = \mu 2,$$
$$H_1 : \mu 1 \neq \mu 2.$$

H_0 is Null Hypothesis and H_1 is Alternative Hypothesis.

```
#Hypothesis 4
Arrest2016<-subset(crime,crime$YEAR == 2016 & crime$ARREST_01 == 1)
nrow(Arrest2016)
```

```
## [1] 52995
```

```
Arrest2017<-subset(crime,crime$YEAR == 2017 & crime$ARREST_01 == 1)
nrow(Arrest2017)
```

```
## [1] 52597
```

```
prop.test(x = c(nrow(Arrest2016),nrow(Arrest2017)), n = c(nrow(data2016),nrow(data2017)), alternat
ive = "greater", conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(nrow(Arrest2016), nrow(Arrest2017)) out of c(nrow(data2016), nrow(data2017))
## X-squared = 0.71416, df = 1, p-value = 0.199
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.0008653798  1.0000000000
## sample estimates:
##    prop 1    prop 2
## 0.1965325 0.1956144
```

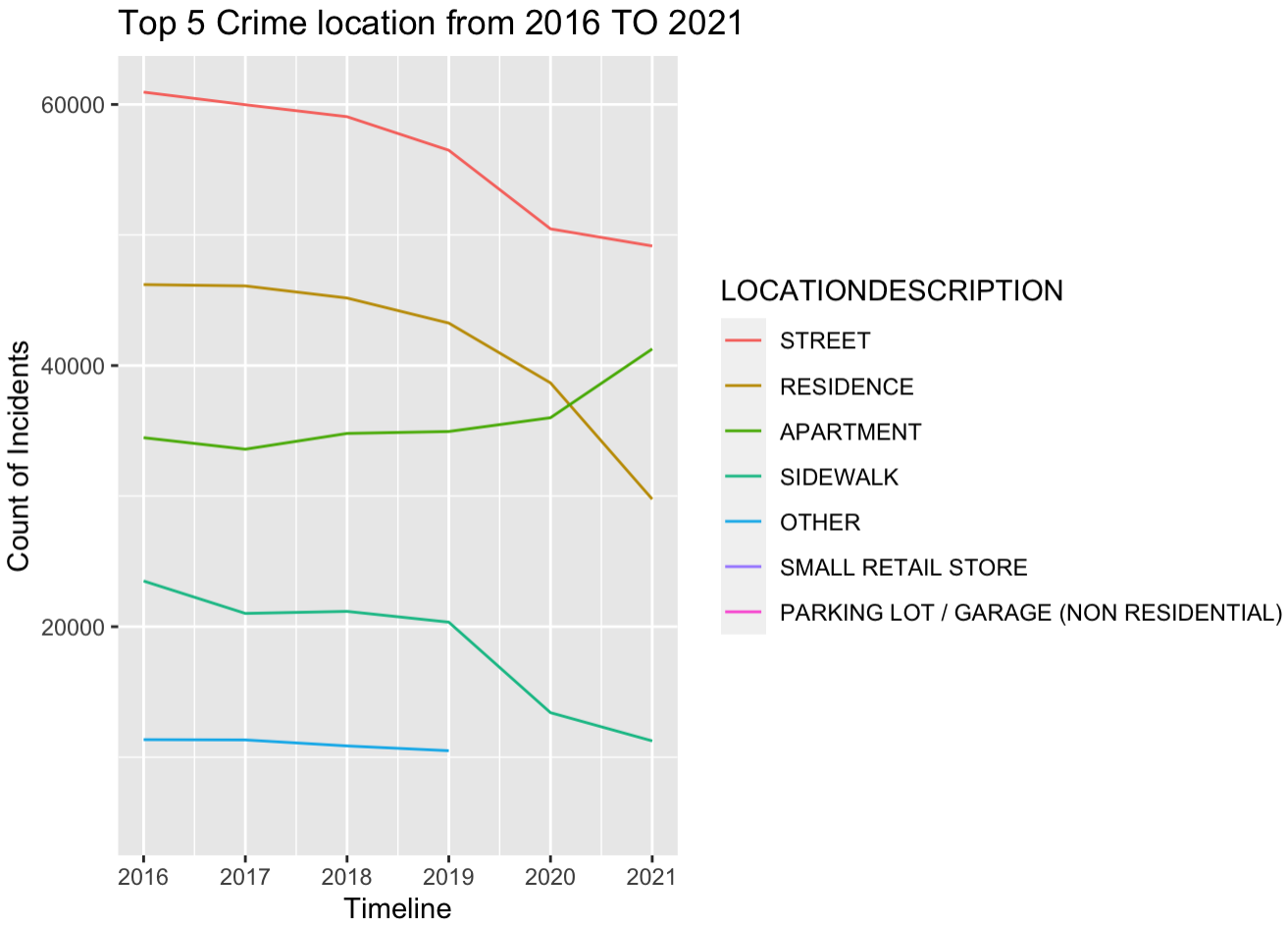
The p-value of this hypothesis is smaller than alpha i.e., 0.05 and the mu1 is also not equal to mu2. So, the proportion is also not equal. So we can accept the Alternate hypothesis H_1 and agree that the less criminal got arrested in year 2017 incomparison to the year 2016.

```
LocationDesc2016<-data.frame(sort(table(data2016$LOCATIONDESCRIPTION),decreasing = TRUE))
names(LocationDesc2016)[1]<- 'LOCATIONDESCRIPTION'
LocationDesc2016<- cbind(LocationDesc2016,Year=c(2016))


LocationDesc2017<-data.frame(sort(table(data2017$LOCATIONDESCRIPTION),decreasing = TRUE))
names(LocationDesc2017)[1]<- 'LOCATIONDESCRIPTION'
LocationDesc2017<- cbind(LocationDesc2017,Year=c(2017))



LocationDesc2018<-data.frame(sort(table(data2018$LOCATIONDESCRIPTION),decreasing = TRUE))
names(LocationDesc2018)[1]<- 'LOCATIONDESCRIPTION'
LocationDesc2018<- cbind(LocationDesc2018,Year=c(2018))



LocationDesc2019<-data.frame(sort(table(data2019$LOCATIONDESCRIPTION),decreasing = TRUE))
names(LocationDesc2019)[1]<- 'LOCATIONDESCRIPTION'
LocationDesc2019<- cbind(LocationDesc2019,Year=c(2019))


LocationDesc2020<-data.frame(sort(table(data2020$LOCATIONDESCRIPTION),decreasing = TRUE))
names(LocationDesc2020)[1]<- 'LOCATIONDESCRIPTION'
LocationDesc2020<- cbind(LocationDesc2020,Year=c(2020))


LocationDesc2021<-data.frame(sort(table(data2021$LOCATIONDESCRIPTION),decreasing = TRUE))
names(LocationDesc2021)[1]<- 'LOCATIONDESCRIPTION'
LocationDesc2021<- cbind(LocationDesc2021,Year=c(2021))


Top5allyears<- merge(merge(merge(merge(merge(LocationDesc2016[1:5,1:3],LocationDesc2017[1:5,1:3],a
ll = TRUE,sort = FALSE),LocationDesc2018[1:5,1:3],all = TRUE,sort = FALSE),LocationDesc2019[1:5,1:
3],all = TRUE,sort = FALSE),LocationDesc2020[1:5,1:3],all = TRUE,sort = FALSE),LocationDesc2021[1:
5,1:3],all = TRUE, sort = FALSE)

Top5allyears
```

```
##                         LOCATIONDESCRIPTION  Freq Year
## 1                                    STREET 60943 2016
## 2                                 RESIDENCE 46200 2016
## 3                                 APARTMENT 34474 2016
## 4                                  SIDEWALK 23498 2016
## 5                                     OTHER 11345 2016
## 6                                    STREET 59977 2017
## 7                                 RESIDENCE 46098 2017
## 8                                 APARTMENT 33591 2017
## 9                                  SIDEWALK 21011 2017
## 10                                    OTHER 11324 2017
## 11                                   STREET 59060 2018
## 12                                RESIDENCE 45170 2018
## 13                                APARTMENT 34800 2018
## 14                                 SIDEWALK 21168 2018
## 15                                    OTHER 10864 2018
## 16                                   STREET 56490 2019
## 17                                RESIDENCE 43252 2019
## 18                                APARTMENT 34948 2019
## 19                                 SIDEWALK 20344 2019
## 20                                    OTHER 10497 2019
## 21                                   STREET 50469 2020
## 22                                RESIDENCE 38671 2020
## 23                                APARTMENT 36004 2020
## 24                                 SIDEWALK 13410 2020
## 25                     SMALL RETAIL STORE  5264 2020
## 26                                   STREET 49157 2021
## 27                                APARTMENT 41268 2021
## 28                                RESIDENCE 29767 2021
## 29                                 SIDEWALK 11248 2021
## 30 PARKING LOT / GARAGE (NON RESIDENTIAL)  6035 2021
```

From the above result we can see the top 5 crime location for each year. From 2016 to 2019 top 5 crime location are same for all the 4 years i.e., Street, Residence, Apartment, Sidewalk, Other. For the year 2020 Other crime location is replaced by Small Retail Shop crime location in the top 5 list and for year 2021 Small Retail Shop crime location is replaced by Parking Lot/ Garage crime location in top 5 crime location.

```
plot6 <-ggplot(Top5allyears, aes(x=Year)) +
        geom_line(aes(y = Freq, color = LOCATIONDESCRIPTION))
print(plot6 +ggtitle("Top 5 Crime location from 2016 TO 2021")+ labs(y="Count of Incidents", x =
"Timeline"))
```



The above graph illustrate top 5 loation in the city of Chicago where crimes took place. From the year 2016 ‘Street’ is dominating the graph with 60k cases followed by ‘Residence’, ‘Apartment’and ’Sidewalk’. We can also see the decline in the cases from 2019 and increase in ‘Apartment’ cases from 2020.

This project idetifies the pattern in crimes commited by identifying the month it is commited in, the location where it was commited and the region where it was committed.