```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
crime<- read.csv("F:/SEM1/Advance_Data_Analytics/Project/Project_2/crime1.csv", header=TRUE)
str(crime)
```

```
## 'data.frame':    1472853 obs. of  25 variables:
##  $ ID                 : int  10481979 10496416 10500341 10500433 10501254 10586318 10607316 1
0607777 10607895 10627410 ...
##  $ CASENUMBER         : chr  "HZ221661" "HZ237276" "HZ241461" "HZ240859" ...
##  $ DATE1              : chr  "01/01/2016 12:00:00 AM" "01/01/2016 12:00:00 AM" "01/01/2016 1
2:00:00 AM" "01/01/2016 12:00:00 AM" ...
##  $ BLOCK              : chr  "009XX N KEDZIE AVE" "018XX W EVERGREEN AVE" "012XX N SPRINGFIEL
D AVE" "049XX S DREXEL BLVD" ...
##  $ IUCR               : chr  "1154" "1154" "1154" "0810" ...
##  $ PRIMARYTYPE        : chr  "DECEPTIVE PRACTICE" "DECEPTIVE PRACTICE" "DECEPTIVE PRACTICE"
"THEFT" ...
##  $ DESCRIPTION        : chr  "FINANCIAL IDENTITY THEFT $300 AND UNDER" "FINANCIAL IDENTITY TH
EFT $300 AND UNDER" "FINANCIAL IDENTITY THEFT $300 AND UNDER" "OVER $500" ...
##  $ LOCATIONDESCRIPTION: chr  "RESIDENCE" "APARTMENT" "RESIDENCE" "RESIDENCE" ...
##  $ ARREST             : chr  "false" "false" "false" "false" ...
##  $ DOMESTIC           : chr  "false" "false" "false" "false" ...
##  $ BEAT               : int  1211 1424 2535 222 1523 533 1413 931 1115 835 ...
##  $ DISTRICT           : int  12 14 25 2 15 5 14 9 11 8 ...
##  $ WARD               : int  26 1 27 4 28 9 35 16 28 18 ...
##  $ COMMUNITYAREA      : int  23 24 23 39 25 54 22 61 26 70 ...
##  $ FBICODE            : chr  "11" "11" "11" "06" ...
##  $ XCOORDINATE        : int  1154860 1163545 1150155 1183147 1141611 1184011 1153320 1163858
1149451 1157423 ...
##  $ YCOORDINATE        : int  1906243 1909429 1908019 1872548 1902863 1818293 1915322 1871818
1899354 1851870 ...
##  $ YEAR               : int  2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...
##  $ UPDATEDON          : chr  "02/10/2018 03:50:01 PM" "02/10/2018 03:50:01 PM" "02/10/2018 0
3:50:01 PM" "02/10/2018 03:50:01 PM" ...
##  $ LATITUDE           : num  41.9 41.9 41.9 41.8 41.9 ...
##  $ LONGITUDE          : num  -87.7 -87.7 -87.7 -87.6 -87.8 ...
##  $ LOCATION           : chr  "(41.898545493, -87.706654407)" "(41.907109594, -87.674665093)"
"(41.903512024, -87.723889357)" "(41.805470526, -87.603810519)" ...
##  $ REGION             : chr  "North" "West" "North" "South" ...
##  $ ARREST_01          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ DOMESTIC_01        : int  0 0 0 0 0 0 1 1 0 0 ...
```

> The dataset consist of 25 variables with 1472853 observations.

```
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 = "great
er", 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 = "great
er", 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 of the lockdown people stayed mostly 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)), al
ternative = "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 and it makes sense because of the lockdown people stayed at home.

> 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)), altern
ative = "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
],all = 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),LocationDesc
2021[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 t.