**DALLAS CRIME DATA**

APPLYING STASTICS TO YOUR CAPSTONE PROJECT

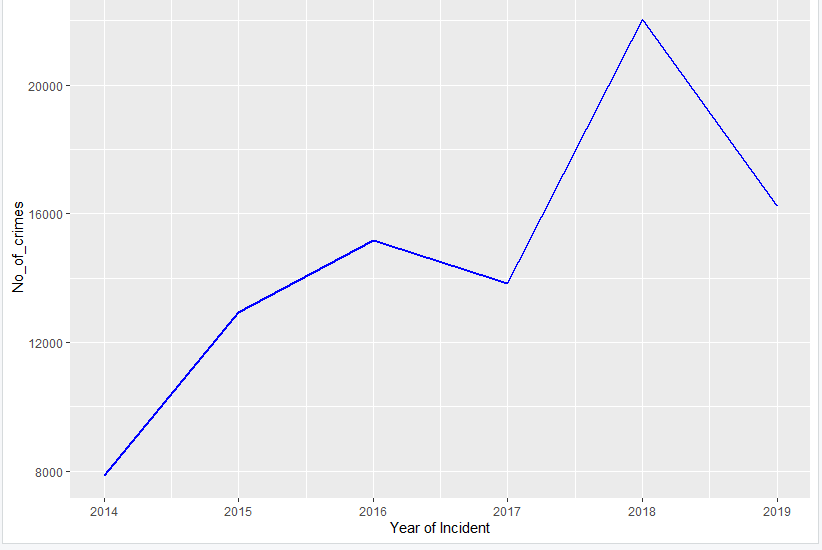
Having learnt the basics of probability, descriptive statistics and the fundamentals of Inferential statistics, Let’s apply some of its knowledge to the Dallas Crime data.

The dataset is cleaned up and ready for analysis. Let’s analyze by answering the following questions.

Can you count something interesting?

Let us count the number of incidents or crimes that happened in a particular DIVISION of Dallas say CENTRAL, over the years.

|  |
| --- |
| *crime\_rate <- crimedata %>% filter(Division == "CENTRAL") %>% group\_by(`Year of Incident`) %>% tally(name = "No\_of\_crimes")*  *ggplot(crime\_rate, aes(x=`Year of Incident`, y = `No\_of\_crimes`)) + geom\_line(col = "Blue", size = 1*) |

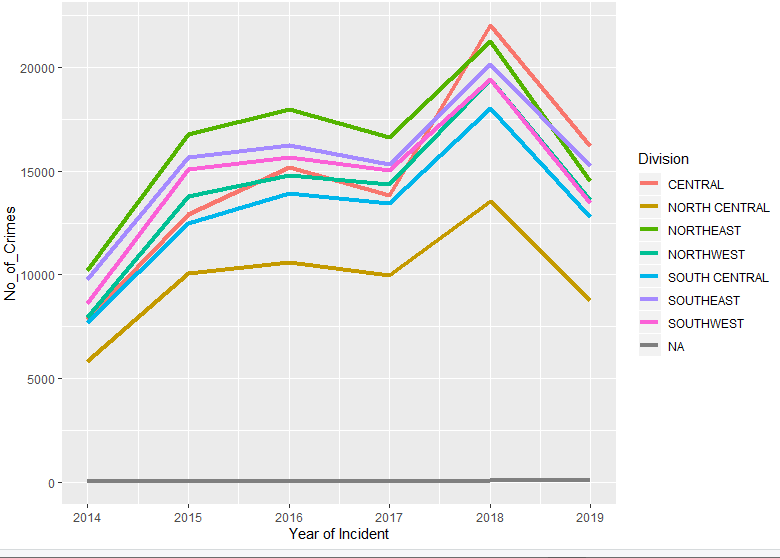


This plot clearly shows that the number of crimes increased in CENTRAL Dallas over the years. In the year 2014, there were approximately 8000 crimes reported. The trend seem to be increasing till the year 2015 and a slight drop in the number of crimes happened in the year 2017 and so on.

Can you find some trends?

Let us count the number of incidents that happened in every part of the city and also how it varied from the year 2014 to 2019.

|  |
| --- |
| *crime\_trends <- crimedata %>% group\_by(Division, `Year of Incident`) %>% tally(name = "No\_of\_Crimes")*  *ggplot(crime\_trends, aes(x=`Year of Incident`, y = No\_of\_Crimes , color = Division, size = 0)) + geom\_line(size = 1.5)* |

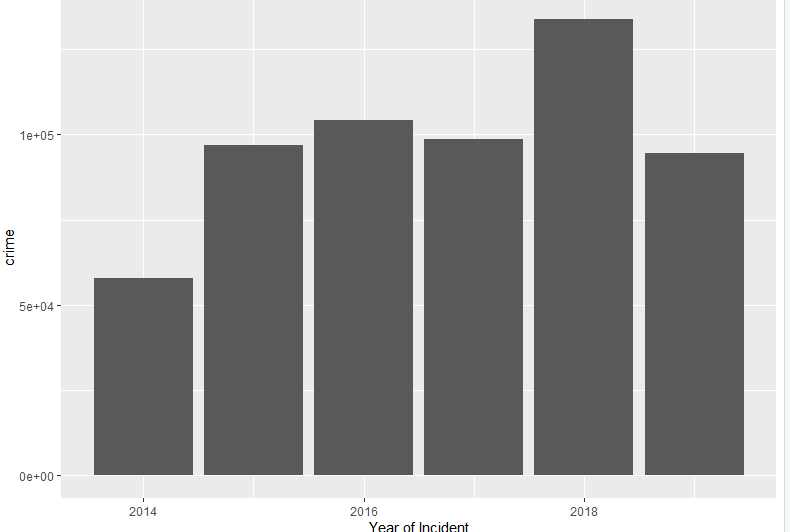


This plot shows the increase in the crime rate over years all over the city. There is huge increase in the crimes that happened in Dallas in the year 2015 compared to 2014. Crime rate decreased in the year 2019 when compared to 2018. However, we can see a Gray line in the plot which represents the anomalies. The dataset contains few data where there are no details about the Division where the crime took place, resulting in anomalies while analysis.

Can you make a bar plot or a histogram?

Let us find the number of incidents happened all over the city from the year 2014 till 2019

|  |
| --- |
| crimes\_per\_year <- crimedata %>% group\_by(`Year of Incident`) %>% summarize(crime = n())  ggplot(crimes\_per\_year, aes(x= `Year of Incident`,y=`crime` )) + geom\_bar(stat = "identity") |



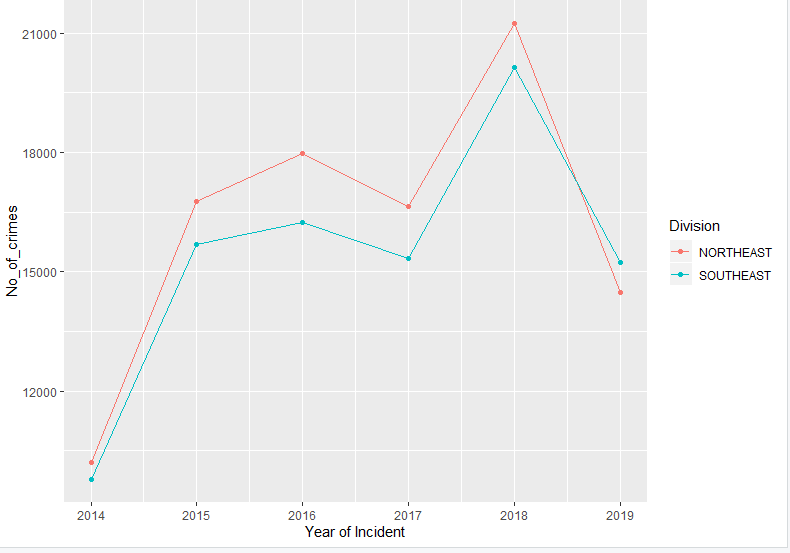
The above bar plot shows the overall crime rate in the city from 2014 to 2019.

Can you compare two related quantities?

Let us compare the crime rates in two different divisions of the city.

|  |
| --- |
| crime\_rate <- crimedata %>% filter(Division == "NORTHEAST" | Division == "SOUTHEAST") %>% group\_by(`Year of Incident`, Division) %>% tally(name = "No\_of\_crimes")  ggplot(crime\_rate, aes(x=`Year of Incident`, y = `No\_of\_crimes`, color = Division)) + geom\_point() + geom\_line() |

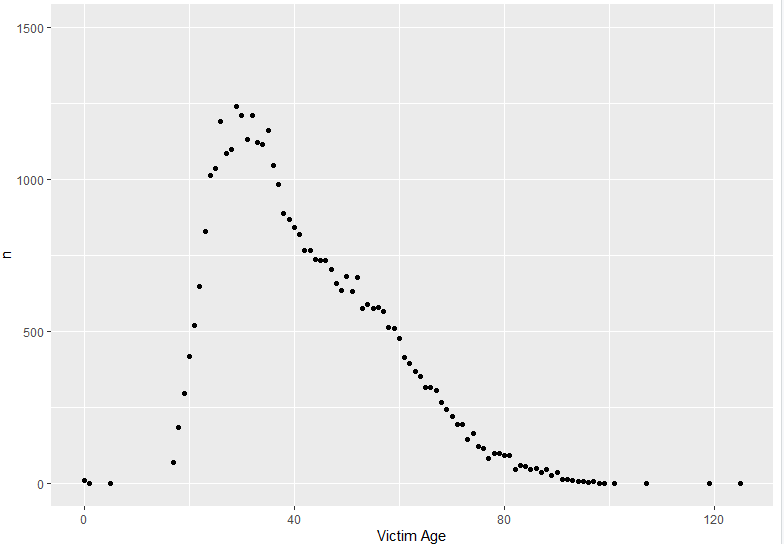
From the plot below we can easily compare the crime rates of the two divisions of the city. The NORTHEAST part of Dallas always has the increased crimes all these years.



Can you make a scatter plot?

Let’s plot a scatter plot to see people between ages 20 and 40 are most affected by BURGLARY.

|  |
| --- |
| victim\_affected <- crimedata %>% filter(grepl("^BURGLARY", `Type of Incident`)) %>% select(`Type of Incident`,`Victim Age`)%>% group\_by(`Victim Age`) %>% summarize(n = n())  victim\_affected <- na.omit(victim\_affected)  ggplot(victim\_affected, aes(x=`Victim Age`, y = n)) + geom\_point() + ylim(0,1500) |



Can you make a time series plot?

The time series plot below shows the number of crimes with its offense status plotted against time. It’s evident from the plot that a large number of were suspended.

|  |
| --- |
| offense\_trends <- crimedata %>% group\_by(`Offense Status`,`Year of Incident`) %>% summarize(count = n())  ggplot(offense\_trends, aes(x=`Year of Incident`, y = count, color = `Offense Status`)) + geom\_point() + geom\_line() |

