DECEMBER 5, 2021

PREDICTION OF CRIME RATE

INSY 5339 - 001 PRINCIPLES OF DATA MINING

PROJECT REPORT

EXECUTIVE SUMMARY :

The Real Estate Company in Chicago wants to come up with a strategy that increases their profits. With the help of the Chicago Crime Data and the Chicago Population Data, the Crime Rate categorized by the community for the city is found. The Finding segregates the safe neighborhoods and unsafe neighborhoods of the city. This is used as a strategy to increase the selling prices and thereby improving their profits.

The 20 years crime data from 2001 - 2020 is taken and cleaned and aggregated according to the number of crimes by the Community Area. The Data Visualization provided insights about the distribution of the data. We calculated the crime rate distribution for the year 2020.

The K-means clustering gave us an overall idea about the data and the communities that has above average total number of crimes. In addition to that, statistical models such as Linear regression , Regression Tree and the KNN regression are done for predicting the future year's (2021) crime rate. Comparing all the models Linear Regression and Regression tree were the top two models having good accuracy, which are used for prediction.

The total crime and total population for the year 2021 is predicted and the crime rate is calculated and sorted. The list of the safe neighborhoods to invest is given to the real estate company. We found that there are changes top safe neighborhoods every year. With the help of data analytics , we were able to predict the safest community to invest which gives the client an edge over their competitors.

PROJECT MOTIVATION :

The Consultant Agency is approached by a Real Estate Company to come up with a way to improve their business. The Real Estate Company has to invest in houses, which they can sell to the people with better profits. We should come up with the strategy to increase their profits and recommend them the neighborhoods to invest.

STRATEGIC SOLUTION:

With the problem in hand, we had done some research and found that people are willing to pay more if the house they are buying is in a safe neighborhood. To address this, we took up the Chicago Crime data from 2001 - 2020 and the Chicago Population data into account. With the help of the total crime and total population we can calculate crime rate.

Crime Rate by Community = Total Crime / Total population

Crime rate gives us the probability a person living in the community be a victim to the crimes happening there. Safer neighborhoods are defined by the lower crime rate. Generally , the company takes up the current year data to make such investment or make investment in available properties. With the help of data analytics we can find the crime rate for the future year which helps in increasing the profits of the company.

DATA DESCRIPTION:

CRIME DATA :

The Data is from the official Chicago Data Portal and is provided by the Chicago police department. The data has details of the crimes reported from 2001 to present and has about 7.42M rows.

Variables	Description
ID	Unique identifier
Case Number	The Chicago Police Department RD Number (Records Division Number) unique to the incident
Date	Incident occurred date
Block	Street Address in which the incident occurred.
Primary Type	Type of Crime.
Arrest	Binary Variable indicating whether the arrest was made or not
Domestic	Binary Variable which indicates whether the crime comes under domestic violence
Community Area	Indicates the community area of the incident. Chicago has 77 community areas.
Year	Year the incident occurred.
Zip code	Zip code of the community Area.

POPULATION DATA:

The Dataset has the population counts of the Chicago city according to the age. The data is taken from the official portal as well and must be correlated with the crime data using the community area to find the crime rate.

Variable	Decription
Year	Indicates the year
Zipcode	Indicates the zip code of the area
Population Count	Total Population count
Population Age 0-17	Population count from the age 0 - 17
Population age 18+	Population count of people above age 18
Population Age 65+	Population count of people above age 65
Population Male	Male Population Count
Population Female	Female Population Count
Population Latinx	Population count of Latin people
Population Asian Non-Latinx	Population count of Asian non latin people
Population Black Non- Latinx	Population count of Black non latin people
Population White Non-Latinx	Population count of white non latin people

DATA CLEANING / PRE- PROCESSING:

• The Crime Data we have taken gives us the details of each crime happened in Chicago, which was taken and aggregated to get the total number of crimes for future prediction.

• We have merged the Community Area to the population data based on the zip code

• The Community Area is a categorical Variable, so one-hot encoding is done for the better prediction.

• Population of default values and elimination of missing values are done.

DATA VISUALIZATION:

1. BOX PLOT:

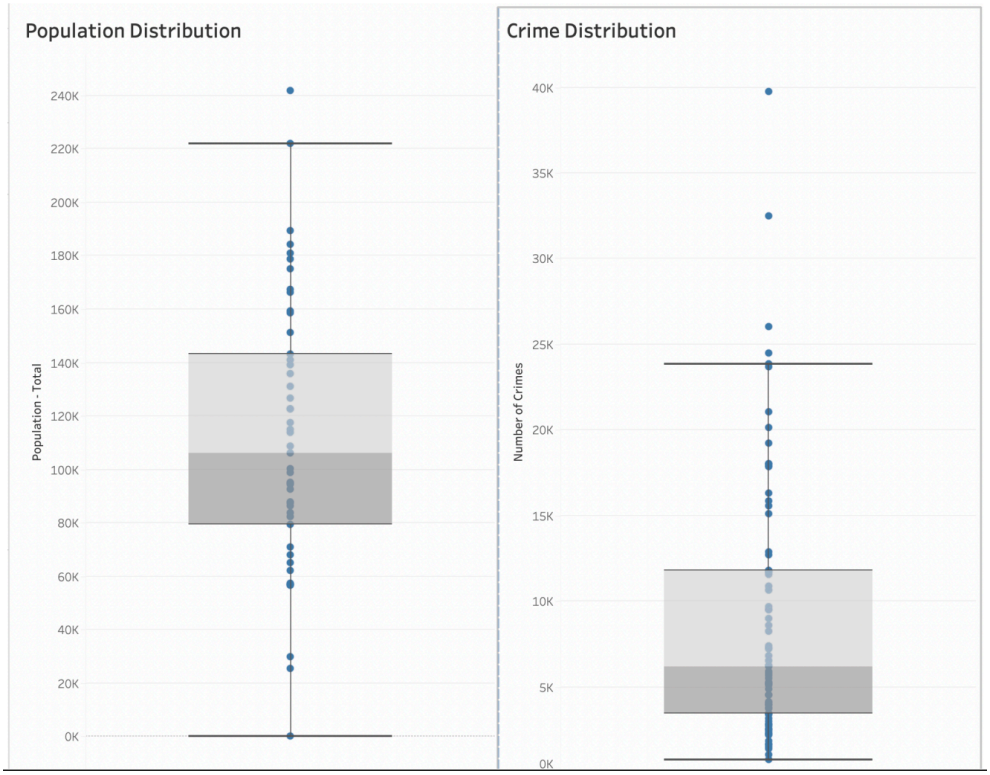
To Visualize the distribution of the population and crimes with respect to the community Area, box plot is used. The Data for the year 2020 is taken for the analysis.

POPULATION DISTRIBUTION:

The total population of Chicago for a year accounts for about 2705988. The Average population of the city is about 35142 when we divided them among the 77 community areas. The Community Area 6 has population of 241884 and tends to be an outlier. The Community Areas 55 and 32 has the lowest population of about 25000.

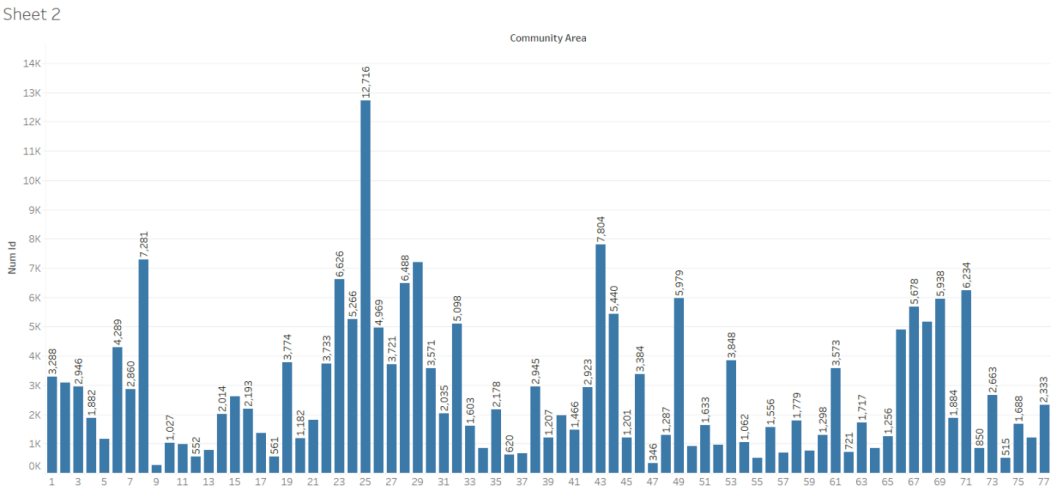
CRIME DISTRIBUTION:

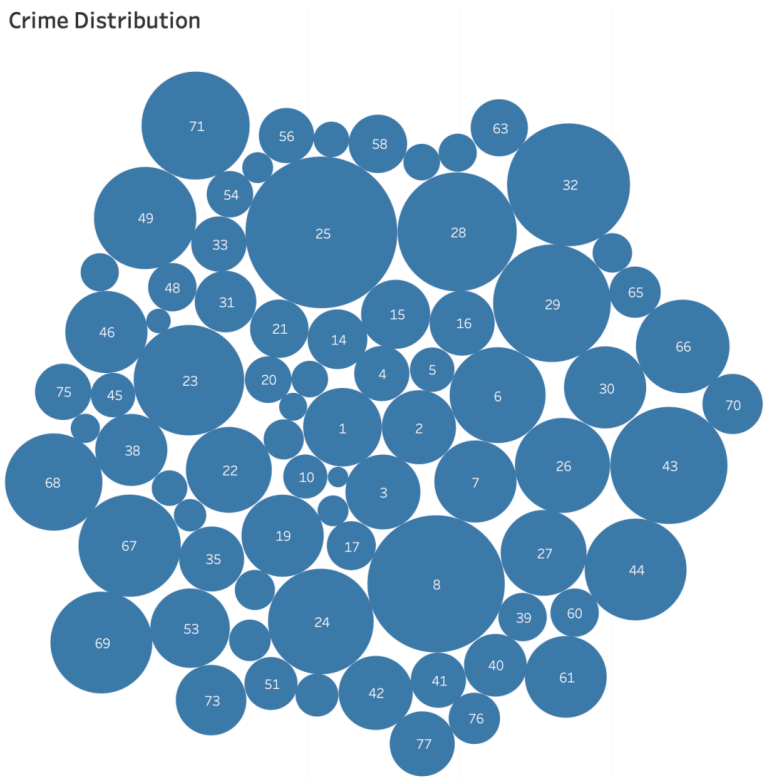
The Crime Distribution plot takes into account the number of crimes based on the community area. The Median Crime is about 6159. The most number of Crimes happen in community area 25 (39719) , 8(32468), 32(26025), 28(24455) and tend to be the outliers.The Lowest number of crimes of 720 happened in community area 9.



2. BAR CHART & BUBBLE CHART :

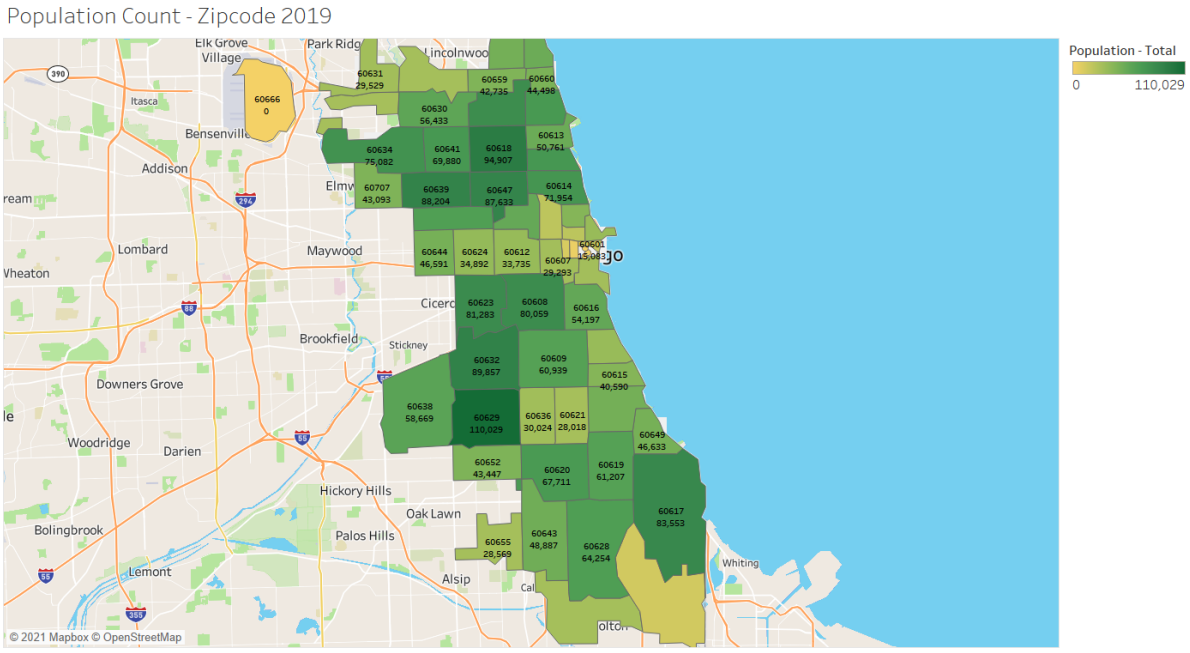
The bar chart represents the total number of crimes occurred in each community. With we we can find the top 10 /20 Community Areas with less number of crimes. The Bar Chart represents the year 2020. The total Crimes may change according to the future predictions.The Bubble Chart also represents the total number of crimes, but is more visually intuitive as the size differs based on the number of crimes. We can roughly find how many community areas should be avoided completely because of the higher crime rate.





3. CHICAGO MAP:

The Chicago map is plotted with respect to the zip codes. A single zip code may comprise off one or more community areas.



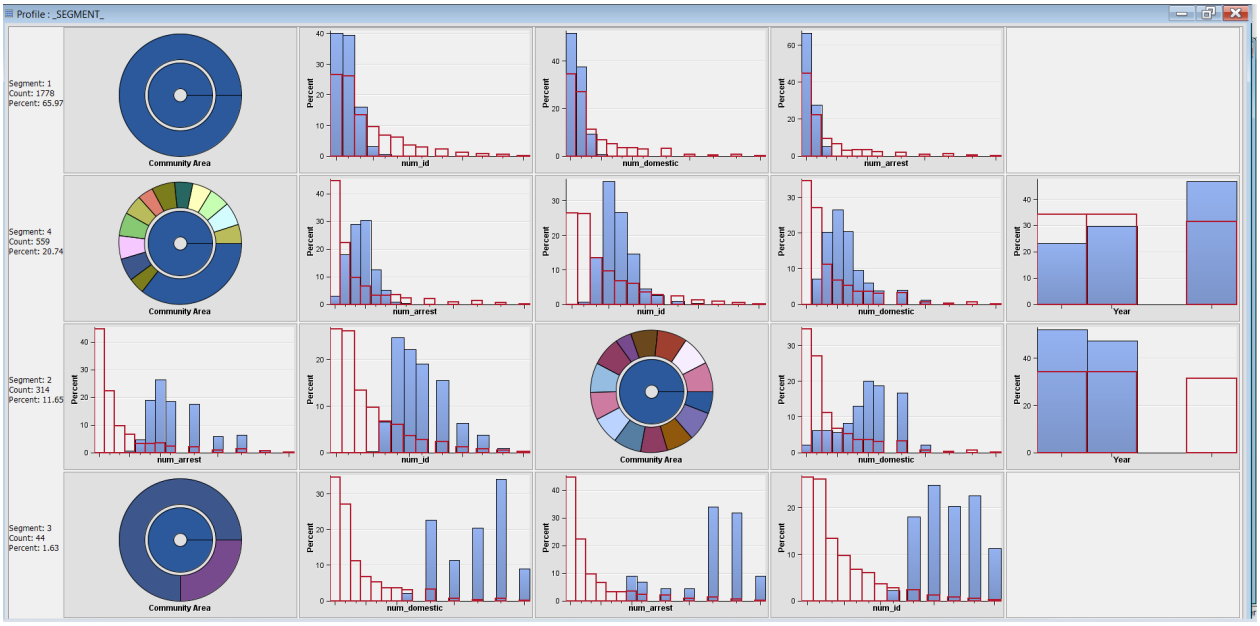
PREDICTION MODELS:

For the year 2020, the top 5 unsafe neighborhoods are given in the table.

Community Area	Total Crimes	Population	Crime Rate in Percentage
32	10891	14675	74.21
8	13102	39019	33.57
28	9466	29591	31.99
25	15174	47712	31.80
67	7128	32203	22.13

K-MEANS CLUSTERING:

The k-means clustering is done to get an overview of the different segments of data available. By default, 2 segments are formed. The two segments were a mixture of all type of crimes like above average , average , below average etc. In an effort to get the top 25% percent of the dataset which has crimes below the average line, we tried to segment the data into four segments. By this type we were able to get some segments with the total crimes below average rate. This segmentation was useful in getting an idea to what types of community areas can be expected to have low number of crimes. Although when the population data is predicted and when the crime rate is calculated, these are expected to change to some extent.



LINEAR REGRESSION:

The Linear Regression is one of the widely used algorithms as part of the prediction models. At first the data is ran without pre-processing with the target variable as total crimes (num id). The Model has about 82% accuracy. Then the data is taken, aggregated based on the year and one-hot coding is done for the categorial variables, stat-explore is used as checkpoint to confirm about the missing variables. The data is again run to get an Adjusted R-Square value of about 0.9802.

The same is done for the population data. The population data has a few missing values in columns such as Age 0-17 , Age 18+ and Age 65+. All the missing values are imputed with the default value of 0. This is done because that category of people may not be there in the community. The Prediction model has the accuracy of 92%

KNN MODEL :

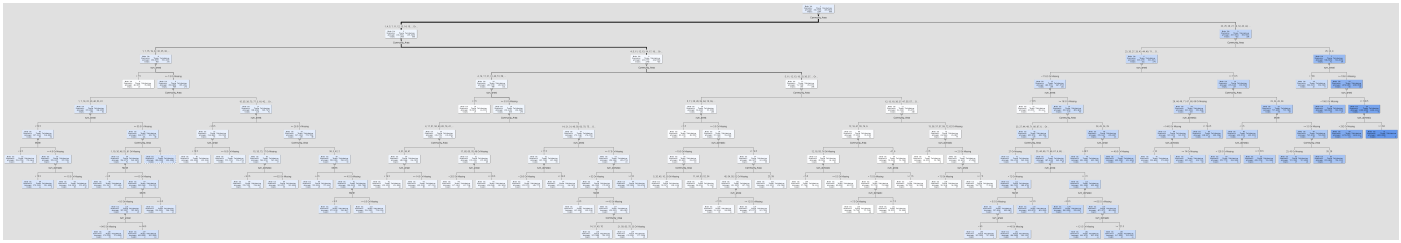
The K nearest neighbor regression model is done to predict the total crime based on the community. For our dataset, four near neighbor seemed appropriate as it has the lowest misclassification rate.

	Warnings	Predicted: num_id	Residual: num_id	Nearest neighbor	Nearest neighbor	Nearest neighbor	Nearest neighbor	Observation Number	A	Year	Month	num_id	num_arrest	num_domestic	Community_Area
1		193.0	96.0	1494.0	1360.0	1290.0	217.0	2.0	1	2018	1	289	40	33	2.0
2		484.75	-178.75	1464.0	1618.0	1179.0	717.0	3.0	2	2018	1	306	62	37	3.0
3		379.75	35.25	1079.0	1618.0	1464.0	2087.0	6.0	5	2018	1	415	67	39	6.0
4		606.75	440.25	1105.0	27.0	1645.0	1568.0	8.0	7	2018	1	1047	142	58	8.0
5		14.0	21.0	2623.0	394.0	2652.0	1857.0	9.0	8	2018	1	35	0	0	9.0
6		74.75	25.25	1125.0	2012.0	1173.0	118.0	10.0	9	2018	1	100	7	8	10.0
7		191.0	50.0	2065.0	93.0	224.0	2409.0	16.0	15	2018	1	241	22	35	16.0
8		416.0	102.0	871.0	1538.0	1177.0	237.0	24.0	23	2018	1	518	56	30	24.0
9		742.0	-66.0	1800.0	1569.0	1332.0	562.0	29.0	28	2018	1	676	291	131	29.0
10		136.75	23.25	245.0	1297.0	1406.0	1438.0	33.0	32	2018	1	160	25	21	33.0
11		145.25	-72.25	496.0	1063.0	2394.0	1005.0	34.0	33	2018	1	73	15	12	34.0
12		109.5	8.5	2289.0	1166.0	380.0	1219.0	39.0	38	2018	1	118	12	16	39.0
13		160.25	-3.25	1727.0	1682.0	1309.0	1066.0	41.0	40	2018	1	157	21	19	41.0
14		292.75	-11.75	1097.0	1116.0	908.0	107.0	42.0	41	2018	1	281	72	57	42.0
15		139.25	-53.25	1427.0	1348.0	1869.0	835.0	48.0	47	2018	1	86	15	22	48.0
16		188.5	-108.5	1561.0	1745.0	1001.0	1540.0	50.0	49	2018	1	80	30	16	50.0
17		173.75	-63.75	1851.0	670.0	1132.0	1479.0	52.0	51	2018	1	110	22	28	52.0
18		482.0	-169.0	1139.0	46.0	754.0	2549.0	61.0	60	2018	1	313	82	66	61.0
19		67.75	9.25	1672.0	996.0	942.0	603.0	62.0	61	2018	1	77	8	13	62.0
20		334.75	92.25	1439.0	1501.0	746.0	1362.0	66.0	65	2018	1	427	72	85	66.0
21		541.0	-17.0	71.0	995.0	915.0	760.0	67.0	66	2018	1	524	135	132	67.0
22		474.5	-16.5	148.0	222.0	1954.0	2102.0	68.0	67	2018	1	458	117	118	68.0
23		517.75	-10.75	275.0	148.0	665.0	1684.0	69.0	68	2018	1	507	124	117	69.0
24		251.5	-10.5	156.0	1129.0	1498.0	309.0	79.0	78	2018	2	241	40	37	2.0
25		237.75	120.25	724.0	35.0	1700.0	281.0	83.0	82	2018	2	358	47	21	6.0
26		1080.75	-274.75	624.0	571.0	879.0	1625.0	85.0	84	2018	2	806	154	37	8.0
27		121.0	-48.0	1474.0	2166.0	11.0	1925.0	88.0	87	2018	2	73	11	14	11.0
28		39.5	-7.5	1972.0	2559.0	2536.0	2578.0	89.0	88	2018	2	32	4	6	12.0
29		280.5	41.5	1039.0	138.0	273.0	173.0	96.0	95	2018	2	322	61	58	19.0
30		221.5	-113.5	1470.0	1078.0	328.0	1573.0	97.0	96	2018	2	108	28	20	20.0
31		114.25	77.75	2253.0	1956.0	199.0	591.0	98.0	97	2018	2	192	23	23	21.0
32		480.75	-165.75	1832.0	1755.0	754.0	2549.0	104.0	103	2018	2	315	82	60	27.0
33		453.0	153.0	1951.0	1567.0	1856.0	1029.0	105.0	104	2018	2	606	109	52	28.0
34		99.25	-54.25	2086.0	1946.0	2061.0	1237.0	111.0	110	2018	2	45	10	6	34.0
35		140.5	35.5	219.0	974.0	262.0	287.0	112.0	111	2018	2	176	35	19	35.0
36		259.0	-70.0	477.0	862.0	1657.0	1575.0	117.0	116	2018	2	189	37	46	40.0
37		143.25	-59.25	251.0	1496.0	1049.0	1386.0	122.0	121	2018	2	84	20	16	45.0
38		303.0	-12.0	1097.0	107.0	1116.0	1424.0	123.0	122	2018	2	291	68	56	46.0
39		562.5	-138.5	490.0	819.0	121.0	144.0	126.0	125	2018	2	424	137	104	49.0
40		202.75	-82.75	1615.0	462.0	1631.0	1372.0	128.0	127	2018	2	120	35	24	51.0
41		71.75	-22.75	1577.0	213.0	1599.0	37.0	132.0	131	2018	2	49	17	9	55.0
42		82.25	-18.25	2012.0	118.0	2211.0	72.0	134.0	133	2018	2	64	6	9	57.0

So, whenever a new input comes with the input similar to the given observation number, the prediction gives us the total crimes for the year 2021. In Comparison, the KNN regression seemed to be the least performing model.

REGRESSION TREE :

The Regression tree seemed to perform well with the data, the first split of the tree occurs with the community area. The Model kind of splits up the similar types of community areas and then continued splitting by other criterions such as the arrest numbers, domestic violence etc. The Month is also included since the data might get split based on the patterns. A regression tree of depth 9 is modeled. The tree has a misclassification rate of about 0.21



MODEL COMPARISON :

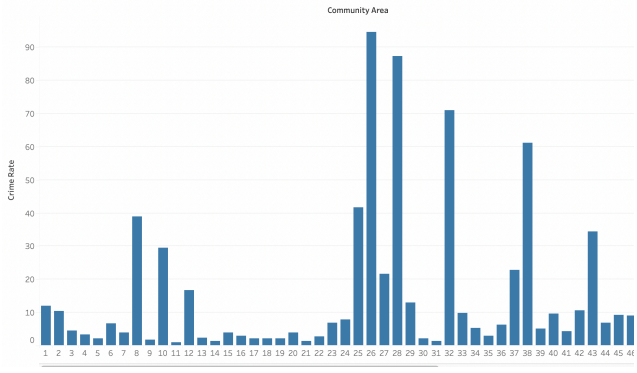
The Models Linear Regression and Regression Tree is selected for the prediction as it has good predictive model with less misclassification rate. The Prediction of the total crimes and total population are done for all the 77 communities and are taken for further crime rate calculation. The Crime rate for each community is calculate by dividing the total crimes by the total population. The Sorted list of the crime rate is generated.

INTERPRETATION:

We found that the top five unsafe community changes for the predicted year 2021. Which is the same in case of safer neighborhoods. The top five unsafe neighborhoods for the predicted year.

Community Area	Crime Rate in Percentage
26	94.58
28	87.30
67	77.19
32	70.95
68	63.62

The Tree map and the bar graph represents the calculated crime rate for the predicted year.



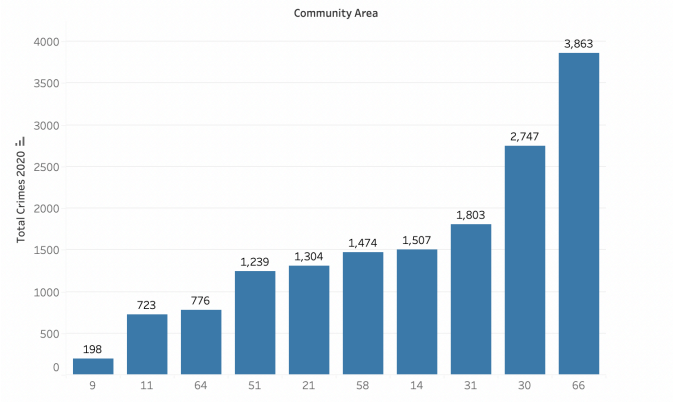
As we can see from the table and visualization that the community area 26 has the highest crime rate.The community Area 32, 67 and 28 were in the top 5 for the year 2020 as well, but the percentage of crime rate changes as the population and the total crimes increases year by year.

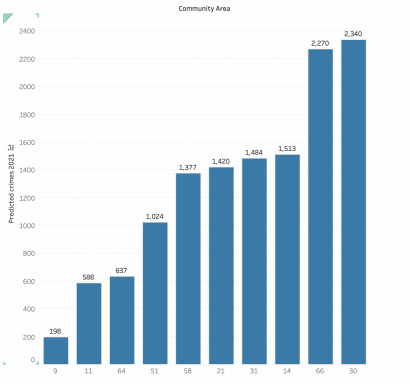
INSIGHTS:

The top ten safest neighborhood for the year 2021 is given to the real estate company. The Real estate company goes and invest in the given order. The Real estate company can also use this strategy for branding. The Company can increase their prices in accordance with the safety of the community.

Community Area	Area	Crime Rate
51	South Deering	0.93
11	Jefferson Park	1.04
64	Clearing	1.12
58	Brighton Park	1.25
21	Avondale	1.29
31	Lower west side	1.34
14	Albany Park	1.37
9	Edison Park	1.71
66	Chicago Lawn	2.06
30	South Lawndale	2.12

This is a tight packed game plan for the company because it uses the future data to invest and it takes into account two dimensions such as Crime and Population.





The two bar graphs represents the total crimes of the top ten community for the year 2020 and 2021, as we can find that the community area 9 has lowest total number of crimes in both the years, but when population is taken into account and crime rate is calculated we can find that the community area 9 has been placed as top 8th position. Such type of calculations and analysis gives our client a very big competitive edge over others. The People moving in to the properties will also be at peace because they have a better odds of not being susceptible to the crimes in their city which gives us a win-win situation.

CONCLUSION:

Thus, with the help of the visual analysis and the statistical models, we were able to fulfill the requirement of the client by coming up with a different strategy that helps them to develop their branding as they care about their customers and also increase their business profits.