Chicago Crime Report

Dataset:
Chicago Crime Dataset 2020

Source:
https://data.cityofchicago.org/Public-Safety/Crimes-2020/qzdf-xmn8/data_preview

Research Question:
Find out where and when crime happens the most in Chicago.

Reason:
I have chosen the Chicago Crime Dataset 2020 for understanding the following:

1. *Strategic Resource Allocation:* Identifying peak times and crime hotspots enables targeted deployment of law enforcement resources, improving response and deterrence.
2. *Proactive Policing:* Focusing on high-crime areas allows for proactive patrols and operations, disrupting criminal activities and reducing overall crime rates.
3. *Community Involvement:* Using crime trend data, policymakers and community groups can implement tailored interventions, such as neighborhood watch programs, to address underlying causes of crime.
4. *Public Safety Improvement:* Addressing crime hotspots leads to safer communities, and fosters trust between residents and law enforcement.
5. *Informed Decision-Making:* Analysis of crime patterns will empower authorities to make data-driven decisions on resource allocation, policy, and community initiatives, enhancing effectiveness in crime prevention.

Goal:

1. Understand the factors that influence whether an arrest or not
2. Determine whether there is a correlation between crimes committed and locations.
3. Distinguish whether the type of crime at a location is arrestable.

Dataset Details:
The crime dataset includes 212,337 records.

1. *ID:* A unique identifier for each crime incident.
2. *Case Number:* The official case number associated with the incident.
3. *Month:* The month when the incident occurred.
4. *Date:* The specific date of the incident.
5. *Primary Type:* The primary classification of the crime.
6. *Description:* A more detailed description of the crime.
7. *Location Description:* Description of the location where the crime occurred.
8. *Arrest:* Indicates whether an arrest was made (True/False).
9. *Domestic:* Indicates whether the crime was domestic-related (True/False).
10. *District:* The police district where the incident occurred.
11. *Year:* The year when the incident occurred.

PLOTS:

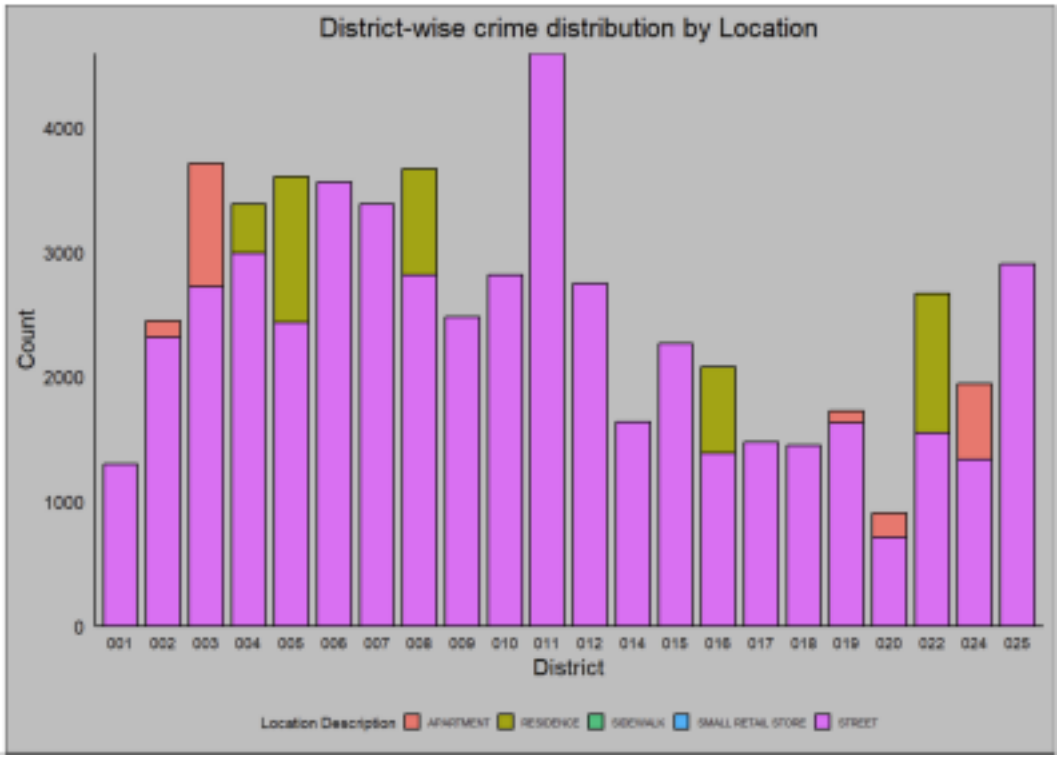
Plot 1:

I aim to understand the district-wise crime distribution by location through the first plot.

Code:

```
View(crime_data)
#Count occurrences of each location description
location_counts <- crime_data %>%
 count(crime_data$`Location Description`) %>%
 arrange(desc(n))

#Rename the location counts column
location_counts <- location_counts %>% rename(Location = 'crime_data$`Location Description`')

#Select the top 5 location descriptions
top_5_locations <- location_counts$Location[1:5]

#Categorize locations into top 5 or others
crime_data <- crime_data %>%
 mutate(`Location Description` = case_when(
 crime_data$`Location Description`
 %in% top_5_locations ~ crime_data$`Location Description`,
 TRUE ~ "OTHERS" ))

#Count occurrences of each location description
location_counts2 <- crime_data %>%
 count(crime_data$`Location Description`) %>%
 arrange(desc(n))

#Filter out "Others" category
filtered_data <- crime_data %>%
 filter(`Location Description` != "OTHERS")

#Plot 1 - District-wise crime distribution by Location
plot1 <- ggplot(filtered_data, aes(x = District, fill = `Location Description`)) +
 geom_bar(position = "identity",width=0.8,color='black') +
 scale_fill_discrete() +
 labs(title = "District-wise crime distribution by Location",
 x = "District",
 y = "Count") +
 scale_y_continuous(expand = expansion(mult = c(0, 0)), limits = c(0, NA)) +
 theme_minimal() +
 theme(legend.position = "bottom",legend.text = element_text(size = 5),
 legend.title = element_text(size = 7),
 legend.key.size = unit(0.3, "cm"),
 axis.text.x = element_text(size = 7),


 axis.text = element_text(color = "black"),
 plot.background = element_rect(fill = "grey"),
```

```
 plot.title = element_text(hjust = 0.5),
 axis.line = element_line(color = "black"),
 panel.grid.major = element_blank(),
 panel.grid.minor = element_blank())
plot1
```



Implications:

1. District 011 has the highest overall crime count, mainly on the street (pink).
2. Districts 003 to 008 also show high crime counts, with street and residence crimes predominant.
3. Apartment-related crimes (red) appear relatively low across all districts.
4. Residence-related crimes are less frequent but present in several districts.
5. Sidewalk and small retail store crimes are uncommon in these districts.
6. District 020 exhibits a lower crime distribution than others, suggesting it's relatively safer.

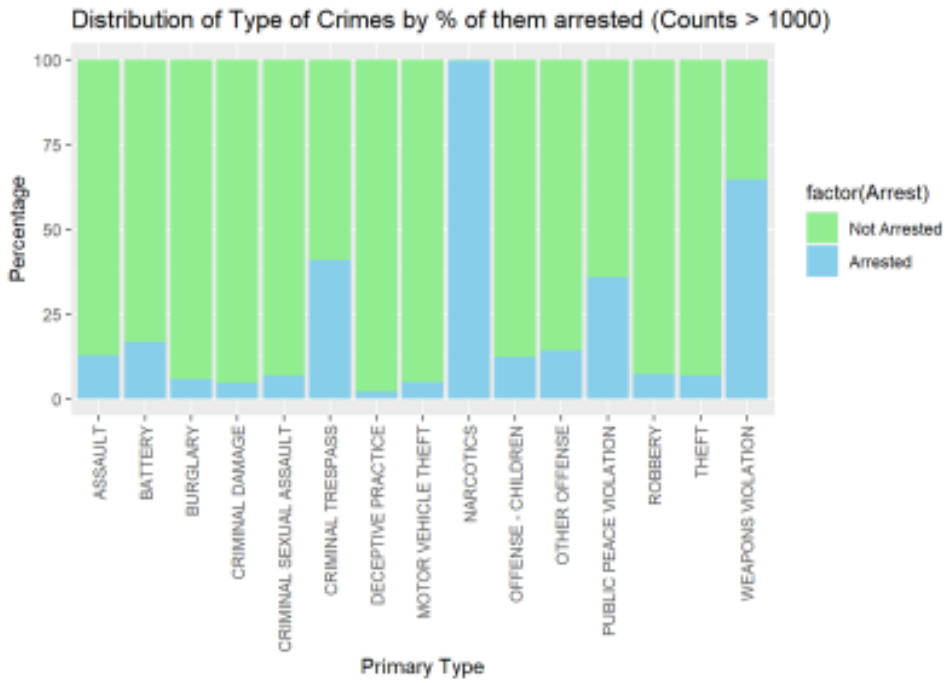
Plot 2:

In plot 2, I try to understand the distribution of types of crimes by the percentage of them arrested.

Code:

```
#Create a count of each crime type
crime_data_counts <- crime_data %>%
 group_by(`Primary Type`) %>%
 summarise(Count = n()) %>%
```

```
#Ungroup the data for plotting
ungroup()

#Filter for primary types with counts greater than 1000
filtered_crime_data <- crime_data %>%
 filter(`Primary Type` %in% crime_data_counts$`Primary Type`[crime_data_counts$Count > 1000])

#Calculate the percentage of arrests and non-arrests for each primary type arrest_percentages <- filtered_crime_data %>%
 group_by(`Primary Type`, Arrest) %>%
 summarise(count = n()) %>%
 mutate(percentage = count / sum(count) * 100)

# Plot the distribution of domestic crimes by primary type
ggplot(arrest_percentages, aes(x = `Primary Type`, y = percentage, fill = factor(Arrest))) +
  geom_bar(stat = "identity") +
 labs(title = "Distribution of Type of Crimes by % of them arrested (Counts > 1000)",
x = "Primary Type",
 y = "Percentage") +
 scale_fill_manual(values = c("lightgreen", "skyblue"), labels = c("Not Arrested", "Arrested")) +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Implications:

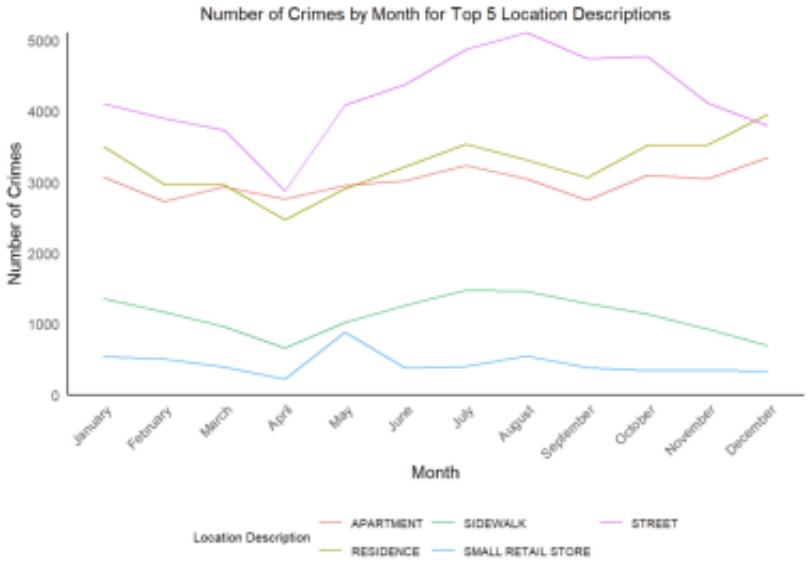
1. Narcotics, the crime type, have been arrested the most.
2. Over half of the Weapons Violation crimes have been arrested.
3. Deceptive Practice, Motor Vehicle Theft, and Criminal Damage are getting the least arrested.

Plot 3:

Plot 3 indicates the Number of Crimes by Month for Top 5 Location Descriptions.

Code:

```
#aggregate data
aggregated_data <- filtered_data %>%
 group_by(Month, `Location Description`) %>%
 summarise(Count = n())

aggregated_data$Month <- factor(aggregated_data$Month, levels = month.name)

#Plotting the data
plot2 <- ggplot(aggregated_data, aes(x = Month, y = Count, color = `Location Description`,
group = `Location Description`)) +
 geom_line() +
 labs(title = "Number of Crimes by Month for Top 5 Location Descriptions",
x = "Month",
y = "Number of Crimes") +
 scale_y_continuous(expand = expansion(mult = c(0, 0)), limits = c(0, NA)) +
scale_color_discrete() +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1),
 legend.position = "bottom",
 legend.text = element_text(size = 7),
 legend.title = element_text(size = 8),
 plot.title = element_text(hjust = 0.5,size = 11),
 axis.line = element_line(color = "black"),
 panel.grid.major = element_blank(),
 panel.grid.minor = element_blank())+
 guides(color = guide_legend(nrow = 2))
plot2
```



Implications:

1. The number of crimes that happen the most on the street, which is high throughout the year. There is a rise in crimes on the street between June - December.
2. A rise in crimes is noticed later in the year.
3. Small retail store crime happens more in April, May, and June.

Machine Learning Analysis 1: Linear Regression

The linear regression model will analyze how the number of reported crimes varies by month across different districts.

Code:

```
> library(readr)
> crime_data <- read.csv("Downloads/Crime_edited data.csv")
> str(crime_data)
> crime_data$Month <- factor(crime_data$Month, levels = c("January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November", "December"))
> crime_data$crime_count <- 1
> crime_lm <- lm(crime_count ~ District + Month, data = crime_data)
> summary(crime_lm)
```

Summary of the model:

Call:

lm(formula = crime_count ~ District + Month, data = crime_data)

Residuals:

Min 1Q Median 3Q Max

-1.058e-09 0.000e+00 0.000e+00 1.000e-14 7.000e-14

Coefficients:

Estimate Std. Error t value Pr(>|t|)

*(Intercept) 1.000e+00 1.812e-14 5.519e+13 <2e-16 ****

District 9.240e-16 7.182e-16 1.286e+00 0.1983

MonthFebruary -1.022e-17 2.353e-14 0.000e+00 0.9997

MonthMarch -6.423e-17 2.405e-14 -3.000e-03 0.9979

MonthApril 1.715e-17 2.592e-14 1.000e-03 0.9995

*MonthMay -5.986e-14 2.374e-14 -2.522e+00 0.0117 **

MonthJune 6.395e-18 2.371e-14 0.000e+00 0.9998

MonthJuly 6.008e-17 2.307e-14 3.000e-03 0.9979

MonthAugust -1.650e-16 2.299e-14 -7.000e-03 0.9943

MonthSeptember -1.352e-16 2.362e-14 -6.000e-03 0.9954

MonthOctober -1.157e-16 2.345e-14 -5.000e-03 0.9961

MonthNovember -1.761e-16 2.410e-14 -7.000e-03 0.9942

MonthDecember -2.440e-16 2.415e-14 -1.000e-02 0.9919

---

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 2.297e-12 on 212324 degrees of freedom

Multiple R-squared: 0.5, Adjusted R-squared: 0.5

F-statistic: 1.769e+04 on 12 and 212324 DF, p-value: < 2.2e-16

Implications:

Intercept: The intercept coefficient indicates that when all other predictors are zero (District and Month), the estimated average crime count is approximately 1. This intercept is statistically significant.

District: The coefficient for District (9.240e-16) suggests a very small, practically insignificant change in crime count per unit change in the district, with a p-value (0.1983) indicating that the effect of the district is not statistically significant at conventional levels (e.g., 0.05).

Months: Except for May, which shows a statistically significant decrease in crime count (-5.986e 14) with a p-value of 0.0117, the coefficients for all other months are not statistically significant, with p-values much greater than 0.05. The significance of May suggests that crime counts in May are significantly lower than the baseline month (not specified, but typically the one not included in the summary, often January if you encoded months categorically from January to December).

F-statistic and p-value: The F-statistic is very high (1.769e+04) with an extremely low p-value (< 2.2e-16), indicating that the overall model is statistically significant.

After conducting the Linear Regression analysis, it became evident that the model's performance was unsatisfactory due to the insignificance of the coefficients associated with features like month and district number. This suggests the selected features may not adequately capture the relationship between crime occurrences and predictors. Given the limitations in obtaining desired variables directly related to criminal counts, I've opted to pivot our approach. I've chosen "arrest" as the deciding variable for subsequent analysis.

Machine Learning Analysis 2: Random Forest

The random forest model examines the relationship between Crime Types and Location and their impact on Arrest decisions.

Code:

```
#Machine Learning:
#random forest: To understand how Crime Type and Location affect the Arrest.

#Library
library(randomForest)
#to convert everything to a category
crime_data[,5:8]<-lapply(crime_data[,5:8],factor)

#Set seed
set.seed(50)

#Run randomForest model to Predict Arrest based on 'Primary Type' and 'Location
Description' #Use na.omit to handle any NA values in the data
ran <- randomForest(Arrest ~ crime_data$`Primary Type` + crime_data$`Location
```

```
Description`,  data = crime_data,
 na.action = na.omit,
 ntree = 500,
 importance = T,
 mtry = 2)

#Print the randomForest model object
ran
```

Summary:

```
##
## Call:
## randomForest(formula = Arrest ~ crime_data$`Primary Type` + crime_data$`Location Description`, data = crime_data,
ntree = 500, importance = T, mtry = 2, na.action = na.omit)
## Type of random forest: classification
## Number of trees: 500
## No. of variables tried at each split: 2
##
## OOB estimate of error rate: 10.65%
## Confusion matrix:
##            FALSE    TRUE    class.error
## FALSE  173962    4249     0.02384252
## TRUE  18362     15764     0.53806482
```

Implications:

With 500 trees, the model achieves an Out-of-Bag (OOB) error rate of 10.65%, indicating an accuracy of 89.35% in predicting arrest outcomes.

1. *High Predictive Accuracy:* The model achieves 89.35% accuracy in predicting arrest outcomes based on Crime Types and Location features.
2. *Variable Importance:* Crime Types and Location are significant predictors, emphasizing their role in determining arrest decisions.
3. *Error Rate:* The model's 10.65% Out-of-Bag error rate indicates areas for potential improvement.
4. *Comprehensive Evaluation:* Ensemble decision trees allow for a thorough assessment of crime scenarios and their impact on arrests.
5. *Actionable Insights:* The model provides actionable insights for resource allocation and targeted interventions in high-crime areas.

Key Findings and Conclusion:

The analysis of the Chicago Crime Dataset 2020 aimed to understand where and when crime occurs most in the city. Through visualizations and machine learning, I provided insights for strategic resource allocation, proactive policing, community involvement, and public safety improvement. Initial visualizations highlighted district-wise crime distributions, revealing hotspots and trends. Machine learning analyses, including Linear Regression and Random

Forest models, explored factors influencing crime occurrences and arrest decisions. Findings emphasize the importance of data-driven approaches in addressing public safety challenges and informing decision-making processes for safer communities.