Aryan Sharma

Statistical Programming

Data Cleaning and Manipulation:

1. **Explore the variables included in the dataset.**

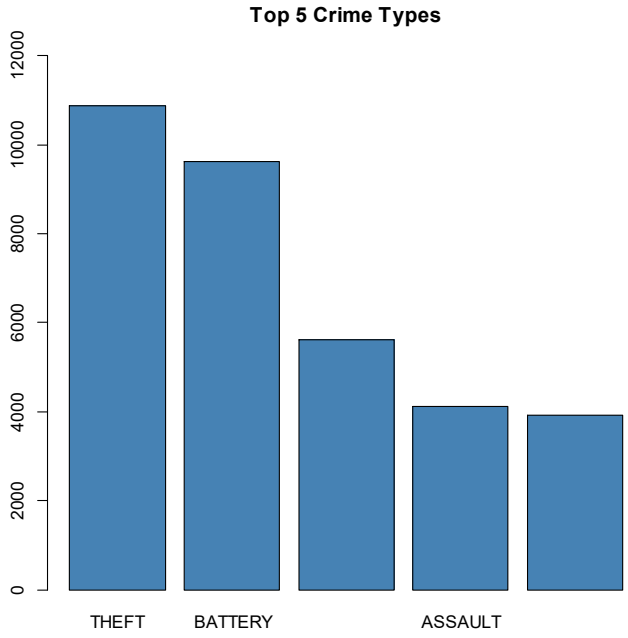
2. **Load the data into your environment and perform any necessary cleaning steps and any data preprocessing steps needed for your analysis. one important variable needs to be formatted.**

```
file_path <- ("C:/Users/91884/Desktop/BAIS/Advance data science/Assignment 3/chicago_crime-1.xlsx")
data <- read_excel(file_path, sheet = "data")
summary(data)
str(data)
```

	date	primary_type	arrest	year	date2
1	2019-01-01 01:00:00	THEFT	FALSE	2019	Jan 2019
2	2019-01-01 01:00:00	THEFT	FALSE	2019	Jan 2019
3	2019-01-01 01:08:00	WEAPONS VIOLATION	TRUE	2019	Jan 2019
4	2019-01-01 01:10:00	THEFT	FALSE	2019	Jan 2019
5	2019-01-01 01:14:00	BATTERY	TRUE	2019	Jan 2019

3. **Conduct a descriptive analysis of the key variables.**

```
top_crimes <- sort(table(data$primary_type), decreasing = TRUE)[1:5]
barplot(top_crimes, main = "Top 5 Crime Types", col = "steelblue", ylim = c(0, 12000))
mean(data$arrest, na.rm = TRUE)
```

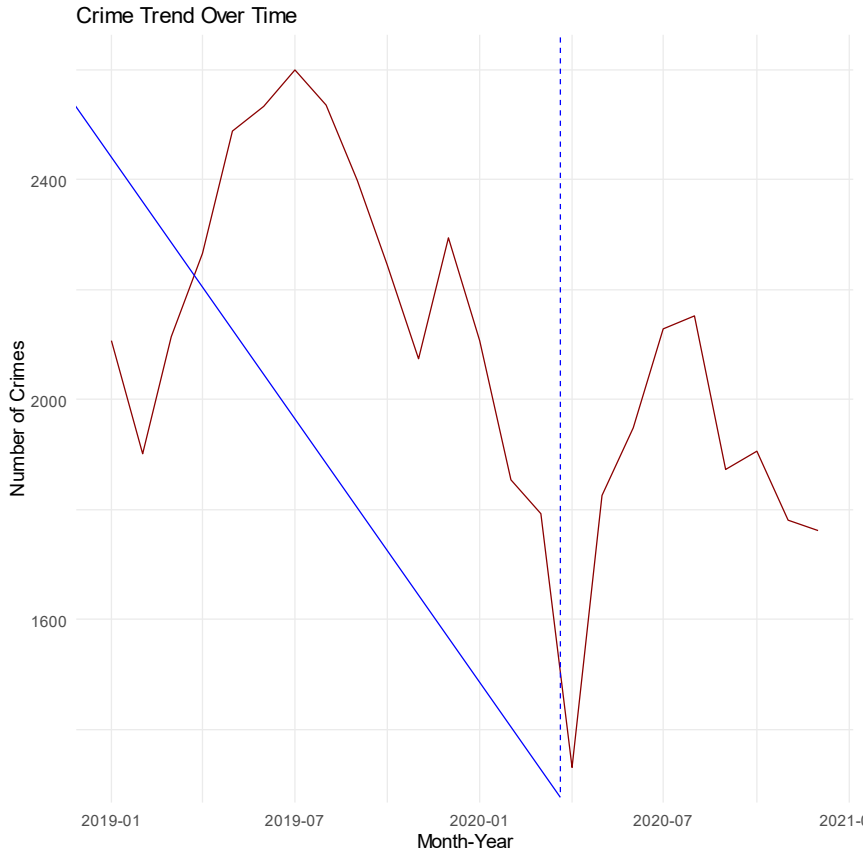



```
> mean(data$arrest, na.rm = TRUE)
[1] 0.18824
```

4. **Using the ggplot, create a line plot of crime incidents over time, with a focus on the impact of a specific change. The visualization should highlight the cutoff point of this change. Draw the time plots with month-year as x-axis.**

```
data$date2 <- as.Date(paste0("01 ", data$date2), format = "%d %b %Y")
monthly_crime <- data %>%
  group_by(date2) %>%
  summarise(total_crimes = n())

ggplot(monthly_crime, aes(x = date2, y = total_crimes)) +
  geom_line(color = "darkred") +
  geom_vline(xintercept = as.Date("2020-03-21"), linetype = "dashed", color = "blue") +
  labs(title = "Crime Trend Over Time",
       x = "Month-Year", y = "Number of Crimes") +
  theme_minimal()
```



5. **Do you see any change in crimes after the cutoff point?**

```
cutoff_date <- as.Date("2020-03-21")
pre_covid <- monthly_crime %>% filter(date2 < cutoff_date)
post_covid <- monthly_crime %>% filter(date2 >= cutoff_date)

cat("Mean crime incidents before COVID:", mean(pre_covid$total_crimes), "\n")
cat("Mean crime incidents after COVID:", mean(post_covid$total_crimes), "\n")
```

```
> cat("Mean crime incidents before COVID:", mean(pre_covid$total_crimes), "\n")
Mean crime incidents before COVID: 2220
> cat("Mean crime incidents after COVID:", mean(post_covid$total_crimes), "\n")
Mean crime incidents after COVID: 1855.556
```

Yes, there is a noticeable decline in crime incidents after the COVID-19 cutoff point of March 21, 2020. The mean number of monthly crime incidents dropped from 2220 before COVID to 1855.56 after COVID, indicating a reduction in criminal activity during the post-pandemic period. This suggests that the onset of the pandemic, possibly due to lockdowns and reduced public activity, may have contributed to the decrease in crime rates.

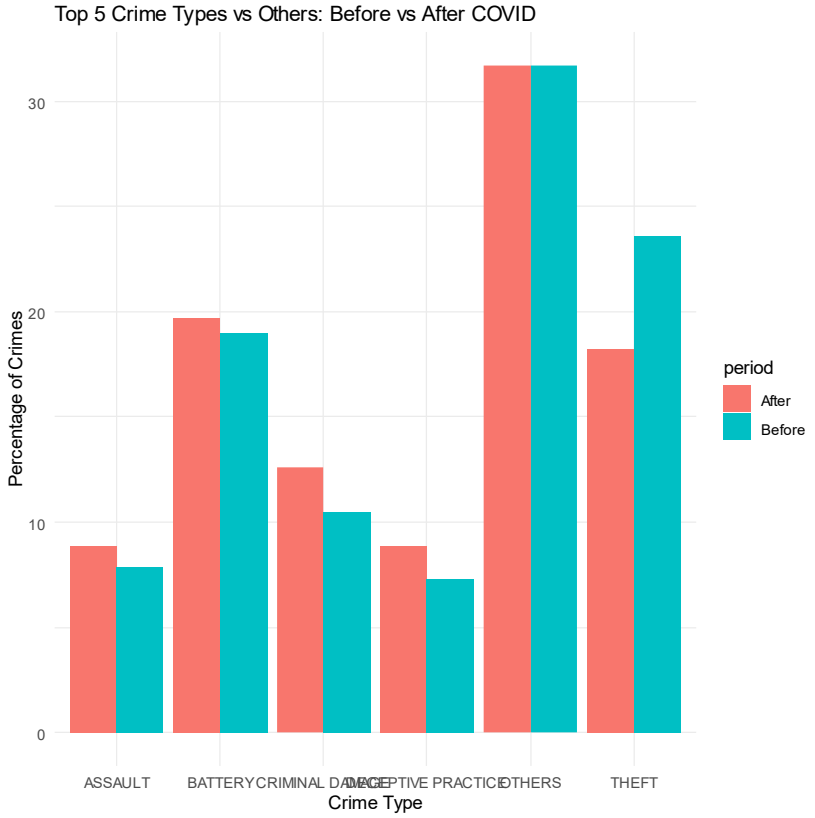
6. **How has the distribution of the top 5 types of crimes and other crimes (variable 'others') changed before and after the onset of the COVID-19 pandemic in Chicago?**

```
top5_types <- names(top_crimes)
data$crime_group <- ifelse(data$primary_type %in% top5_types, data$primary_type, "OTHERS")
data$period <- ifelse(data$date2 < cutoff_date, "Before", "After")

crime_distribution <- data %>%
  group_by(period, crime_group) %>%
  summarise(count = n()) %>%
  group_by(period) %>%
  mutate(percent = round(100 * count / sum(count), 1))

ggplot(crime_distribution, aes(x = crime_group, y = percent, fill = period)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Top 5 Crime Types vs Others: Before vs After COVID",
       x = "Crime Type", y = "Percentage of Crimes") +
  theme_minimal()
```

While most crime categories slightly decreased post-COVID, theft stood out as an exception with a significant rise, altering the crime distribution landscape in Chicago during the pandemic era.



7. **Suppose you decide to apply the RD design to draw causal inference. What would be your dependent variable? Why? How would you create the DV?**

In a Regression Discontinuity (RD) design applied to study the impact of the COVID-19 pandemic on crime, the appropriate dependent variable (DV) would be the *total number of crimes per month*. This variable captures the overall level of criminal activity and serves as a measurable outcome that may have been influenced by the pandemic. The rationale for choosing this as the DV is that the onset of COVID-19, marked by the cutoff date (March 21, 2020), represents a natural intervention point that may have caused a structural break in crime trends due to lockdowns, mobility restrictions, and economic disruptions. To create this DV, crime data should be aggregated monthly, summing all incidents recorded within each month. This monthly total crimes then becomes the outcome variable whose behavior is examined around the cutoff point to assess whether there is a statistically significant discontinuity, thereby allowing for a causal inference about the effect of the pandemic on crime rates.

8. **For RDD analysis, how would you choose/create an independent variable and identify a specific cutoff date? Discuss the importance of the cutoff date and the creation of the independent variable.**

In a Regression Discontinuity Design (RDD) analysis, the independent variable—often referred to as the *running variable* or *forcing variable*—must be a continuous variable that determines treatment assignment based on whether it crosses a specific threshold or cutoff. In the context of analyzing the impact of COVID-19 restrictions on crime, the independent variable would be a time-based variable, such as a sequential monthly or daily index, representing the date of each observation. For example, one could create a numeric variable where values increase with time (e.g., number of days or months from a fixed starting point), and the cutoff date would be coded as March 21, 2020—the day COVID-19 restrictions were officially implemented in Chicago.

The importance of the cutoff date lies in its role as the point of policy intervention; observations before this date are considered untreated (pre-COVID) while those on or after are considered treated (post-COVID). It represents the threshold at which potential changes in the dependent variable—here, monthly crime totals—can be attributed to the onset of the pandemic and associated policies. The independent variable must be continuous and precisely measured around the cutoff to validly estimate causal effects. Creating this variable with fine temporal granularity (e.g., daily or monthly) ensures that the RD design can detect any discontinuity in the outcome right at the intervention point, isolating the impact of COVID-19 from other time-related trends.

9. **Set up the model suitable for RDD Analysis and conduct the analysis. Interpret the results.**

```
rdd_data <- data[!is.na(data$dv_battery), ]

rdd_result <- rdrobust(y = rdd_data$dv_battery, x = rdd_data$running_var, c = 0)
summary(rdd_result)

#Optional: Visualize Discontinuity
ggplot(rdd_data, aes(x = running_var, y = dv_battery)) +
  geom_point(alpha = 0.3, position = position_jitter(width = 0.3, height = 0.05)) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
```

labs(title = "RDD: Effect of COVID Cutoff (March 21, 2020) on Battery Arrests", x = "Months from Cutoff", y = "Battery Arrest (1 = Yes)")

```
Sharp RD estimates using local polynomial regression.

Number of Obs.                50000
BW type                       mserd
Kernel                    Triangular
VCE method                        NN

Number of Obs.                31509        18491
Eff. Number of Obs.            6254         6894
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   3.144        3.144
BW bias (b)                   5.377        5.377
rho (h/b)                     0.585        0.585
Unique Obs.                      14           10

===============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
===============================================================================
  Conventional     0.005     0.009     0.571     0.568    [-0.013 , 0.023]
        Robust         -         -     0.589     0.556    [-0.016 , 0.031]
===============================================================================
```

The RD analysis suggests that there is no significant causal impact of the COVID-19 lockdown (as implemented on March 21, 2020) on the overall level of monthly crime incidents. The small, positive coefficient implies a slight increase in crime at the cutoff, but this is not statistically meaningful due to the wide confidence intervals and high p-values. Thus, the observed changes in crime before and after the cutoff could be attributed to normal variation rather than a true causaleffect of the pandemic restrictions.

Aryan Sharma