UNIVERSITI TEKNOLOGI PETRONAS

DATA SCIENCE

(TFB2063)

SEMESTER SEPTEMBER 2024

GROUP PROJECT

LECTURER: TS. DR. KAMALUDDEEN USMAN DANYARO

No.	Student Name	ID	Programme
1	Mohamad Syazreeq Akmal bin Sarudin	22000454	Information Technology
2	Adam Shamel bin Mohad Shafie	22000766	Information Technology
3	Ahmad Faqrullah bin Hairol	22000450	Information Technology
4	Muhammad Aqfa bin Alias	21001180	Business Management

Table of Contents

1.0 Report Assessment

1.1 Introduction

The "Gun Violence in the United States" dataset provides a thorough and in-depth analysis of gun-related incidents nationwide, gathering a variety of data points that illustrate the complexity of this important topic. Gun violence has become one of the most important public safety issues in the US in recent years, sparking intense discussion among residents, community leaders, and legislators. By using this dataset to unearth insights that can drive preventative measures, improve public awareness, and inspire effective policy-making, this research seeks to explore the patterns, trends, and implications of gun violence.

Mass shootings, suicides, killings, and unintentional discharges are all considered forms of gun violence. Numerous research and publications indicate that the number of gun-related deaths has escalated to concerning proportions, with tens of thousands of deaths annually. Beyond the immediate victims, gun violence has an influence on communities, influencing social cohesiveness, mental health, and economic stability. Developing focused treatments that address the underlying causes of gun violence requires an understanding of its dynamics.

The dataset will be methodically examined in this research in order to pinpoint important trends and patterns related to gun violence in the US. We seek to provide a thorough picture of how gun violence appears in various contexts by looking at variables like location, demographics (such as age, gender, and race), incident types (such as domestic violence or gang-related shootings), and temporal factors (such as seasonality). Additionally, this analysis will look at relationships between the prevalence of gun violence and socioeconomic indicators like education levels, poverty rates, and access to mental health options.

We will use a range of analytical tools, such as descriptive statistics, data visualisation techniques (such charts and graphs), and possibly predictive modelling approaches, to accomplish these goals. By using these techniques, we will be able to simplify complicated data into insights that stakeholders of all levels can readily comprehend. With an emphasis on demographic characteristics, event kinds, and geographic distribution, this paper will use the information to examine trends and patterns in gun violence. By looking at these factors, we hope to find information that can help community leaders and legislators combat this widespread problem. Additionally, associations between rates of gun violence and socioeconomic variables including poverty and educational attainment will be investigated.

1.2 Problem Description

The growing number of mass shootings in the US has raised serious concerns about public safety and made the creation of prediction algorithms to project the number of victims in upcoming events necessary. Policymakers and emergency responders must comprehend trends in victim numbers because they can improve preventive efforts and allocate resources more effectively. This essay explains the reasoning behind creating a forecast model that uses historical data and statistical techniques to estimate the number of victims (both dead and injured) per month in 2024.

Mass shootings, defined as incidents where four or more individuals are killed or injured by gunfire in a single event, have surged in recent years. According to the Gun Violence Archive, there were 2,465 mass shootings reported in 2024 alone, with a total of 114,980 homicides and injuries related to gun violence. This alarming trend underscores the urgent need for effective analysis and intervention strategies. Research indicates that mass shootings occur approximately every 12.5 days in the U.S., with significant variations in victim numbers depending on various factors such as location, time of year, and underlying motivations of the shooters.

Predictive models that can calculate the number of victims both fatalities and injuries resulting from mass shootings are necessary in light of the startling increase in these instances in the United States. This predictive model's main goal is to use previous data to offer insights about upcoming mass shooting incidents in 2024. In order to create a solid foundation for comprehending and forecasting victim counts, this essay lists the main aspects that will be examined, such as the season, demographic characteristics, and incident data from the past.

The season is one of the most important factors in determining how many people will be killed in mass shootings. Researchers can find seasonal patterns that may affect the frequency and intensity of mass shootings by breaking down trends by month. For example, studies have indicated that a number of factors, including school schedules, community events, and holiday stress, may contribute to greater rates of gun violence during specific months. It is feasible to ascertain whether particular months routinely result in greater victim counts by looking at historical data on mass shootings across a number of years. During peak times, this knowledge can aid policymakers and emergency responders in more efficient resource allocation.

By analysing this historical data, researchers can identify trends over time and create statistical models that predict future outcomes based on established patterns. For instance, a study found that rampage mass shootings tend to result in significantly higher victim counts compared to non-rampage incidents. Understanding these distinctions allows for more nuanced predictions about potential future incidents.

1.3 Data Preparation

- Dataset

Initially, the dataset used contain data about mass shooting cases in US, 2024. The dataset contains 13 columns and 427 rows of data with features like Incident ID, Incident Date, State, City or County, Address, Victims Killed, Victims Injured, Suspects Killed, Suspects Injured, Suspects Arrested, Latitude, Longitude and "Coordinates_Found".

Incident ID	Incident Date	State	City Or County	Address	Victims Killed	Victims Injured	Suspects Killec	Suspects Injured	Suspects Arre	Latitude	Longitude	Coordinates_Found
3052758	October 21, 2024	Washington	Fall City	7700 block of Lake Alice Rd	5	1	0	0	1	47.56812	-121.891	Yes
3052028	October 20, 2024		Jackson	2310 N Highland Ave	1	8	0	0	0	35.6139	-88.8194	Yes
3051984	October 20, 2024	Louisiana	Baton Rouge	9700 block of Greenwell Sp	0	5	0	0	0	30.44335	-91.1866	Yes
3051041	October 19, 2024	Pennsylvania	Philadelphia	2517 N Jessup St	0	7	0	0	0	39.95222	-75.1622	Yes
3050940	October 19, 2024	Mississippi	Lexington	24904 MS-17	3	8	0	0	0	33.11464	-90.0528	Yes
3051717	October 19, 2024	Georgia	Albany	504 College Dr	1	4	0	0	0	31.5814	-84.156	Yes
3051689	October 19, 2024	Indiana	Fort Wayne	4911 Manistee Dr	1	10	0	0	0	41.08026	-85.1383	Yes
3047186	October 14, 2024	Michigan	Detroit	Haverhill St and Southampt	1	3	0	0	0	42.33168	-83.048	Yes
3045998	October 12, 2024	Tennessee	Nashville	Jefferson St and 27th Ave	1	9	0	0	2	36.16784	-86.7782	Yes
3045675	October 12, 2024	Oklahoma	Oklahoma City	5878 S Agnew Ave	1	14	0	0	3	35.47203	-97.5211	Yes
3047310	October 12, 2024	Georgia	Elberton	Evergreen Dr	1	7	0	0	1	34.10988	-82.8641	Yes
3046124	October 12, 2024		Cleveland (East Cleveland)	15449 Euclid Ave	1	4	0	0	0	41.50904	-81.6117	Yes
3046477	October 11, 2024	South Carolina	Newberry	1400 block of Martin St	0	4	0	0	0	34.27455	-81.6192	Yes
3044413	October 9, 2024	Missouri	Kansas City	300 block of Westport Rd	0	5	0	0	0	39.10344	-94.5831	Yes
3041753	October 6, 2024	Texas	Fort Worth	100 block of NW 24th St	1	5	0	0	1	32.75095	-97.3309	Yes
3041259	October 5, 2024	Oklahoma	Oklahoma City	208 Johnny Bench Dr	0	4	0	0	3	35.47203	-97.5211	Yes
3040192	October 4, 2024	California	Redding	6438 Westside Rd	2	3	0	0	1	40.58762	-122.392	Yes
3037386	September 29, 2024	Georgia	Atlanta	2500 block of Gresham Rd	1	3	0	0	0	33.74831	-84.3911	Yes
3036427	September 29, 2024	Virginia	Danville	510 Spring St	1	3	0	0	1	36.58598	-79.4172	Yes
3036055	September 28, 2024	District of Columbia	Washington	2700 block of Langston Pl S	1	3	0	0	0	38.89037	-77.032	Yes
3037002	September 28, 2024	North Carolina	Henderson	1110 S Garnett St	0	5	0	0	0	36.33006	-78.3985	Yes
3037040	September 27, 2024	South Carolina	Woodruff	115 Griffin Rd	3	1	0	0	0	34.74011	-82.036	Yes
3034398	September 26, 2024	Illinois	Chicago	4255 N Knox Ave	0	4	0	0	0	41.88425	-87.6325	Yes
3033981	September 25, 2024	California	Los Angeles	449 W 74th St	0	6	0	0	4	34.05357	-118.245	

us_gun_violence_data_2024.csv

- Data Cleaning

From the dataset, there are missing values inside "State" and "Coordinate_Found" column. The missing values inside "State" column is handled by searching the state location according to the City or County data of the rows that have missing state data. After all missing states data are gathered, the missing states data stored inside a list and appended into the dataset.

```
# Column State & Coordinates_Found containing missing values therefore data cleaning is required
    # Handling missing values in State column
missing_state <- which(is.na(dataset_gun$State))
correct_state <- c("Tennessee", "Ohio", "New York", "Alabama", "Ohio", "California", "Texas", "California")
dataset_gun$State[missing_state] <- correct_state
```

Apart from that, the missing values inside "Coordinate_Found" column are handled by automatically insert "Yes" or "No" value inside the missing values depending on the availability of latitude and longitude of the row.

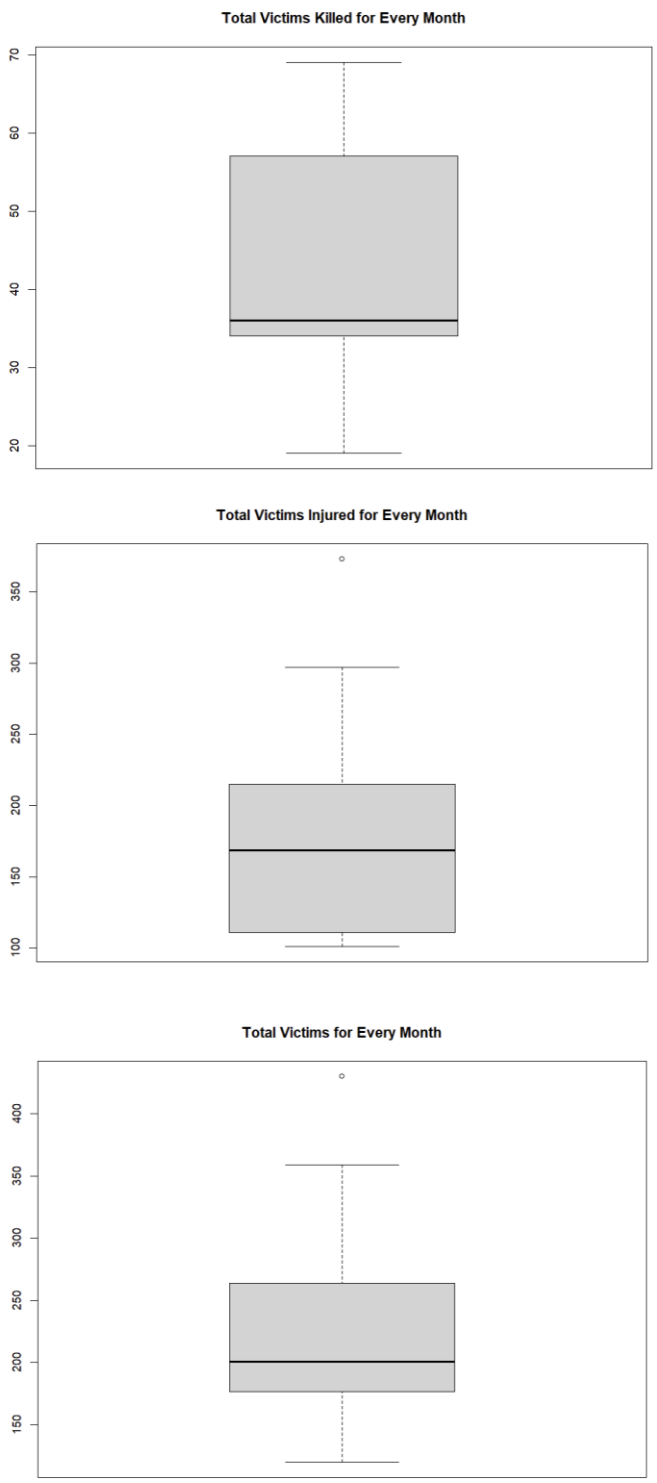
```r
# Handling missing values in Coordinates_Found column
for (i in 1:nrow(dataset_gun)){
    if (is.numeric(dataset_gun$Latitude[i]) && is.numeric(dataset_gun$Longitude[i])){
        dataset_gun$Coordinates_Found[i] <- "Yes"       # Value Yes if got coordinate on latitude and longitude
    }else if (!is.numeric(dataset_gun$Latitude[i]) || !is.numeric(dataset_gun$Longitude[i])){
        dataset_gun$Coordinates_Found[i] <- "No"        # Value No if no coordinate on latitude or longitude
    }
}
```

When the dataset is complete and having no missing values, new features added in the dataset which are "Total Victim" and "Month". The "Total Victim" feature is derived from the addition of "Victims Killed" and "Victims Injured" data. On the other hand, "Month" feature is derived from "Incident Date" data by extracting the month data from the date. Both features were added for model building and data prediction purpose. A data frame was created from this dataset named "monthly_victims" that contain "Month", "Victims Killed", "Victim Injured" and "Total_Victim" attributes. The data is sorted according to the sum of victim count monthly.

	Month	Total_Victims	Victims Killed	Victims Injured	Month_Num
1	January	180	69	111	1
2	February	145	36	109	2
3	March	177	36	141	3
4	April	206	29	177	4
5	May	264	49	215	5
6	June	382	57	357	6
7	July	359	62	297	7
8	August	245	34	211	8
9	September	196	36	160	9
10	October	120	19	101	10

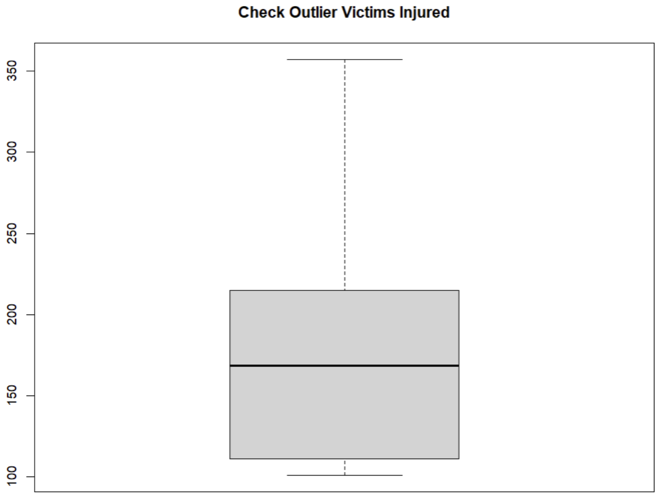
monthly_victims data frame

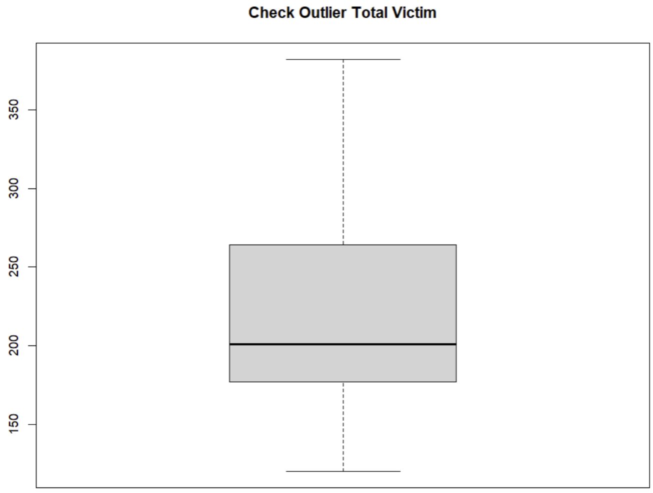
This data frame used to undergo data visualization, model building and data prediction process but prior these processes, it is important to find any extreme values (outliers) to elevate any errors or unreliable output happen. Outliers can significantly impact the reliability of visualizations, skew model performance, and lead to unreliable predictions. By identifying and addressing these anomalies, we can ensure the integrity and accuracy of the analysis, leading to more firm and reliable output. Detecting and handling outliers is an essential preparatory step to minimize errors and enhance the quality of the final outcomes derived from the monthly_victims data frame.



Total victim killed, total victim injured and total victim by month represented in Boxplot

These boxplot diagrams are produced to check any outliers exist inside the data frame. The data for “Total Victims Injured” and “Total Victims” are both having an outlier outside the upper fence of the boxplot. These outliers identified using the interquartile range (IQR) method are handled by replacing the outliers with the value of upper fence value for their respective boxplot graph. These are the boxplot graph after the outliers were eliminated:

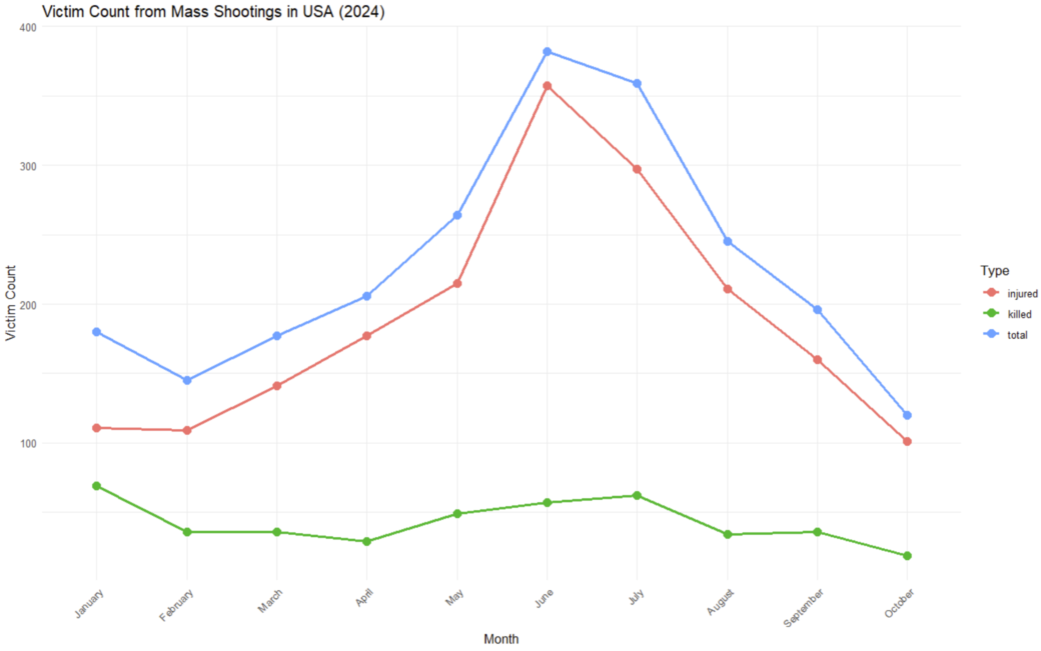




Total victim injured and total victim Boxplot after cleansing

- Data Visualization

The data frame monthly_victims is used for visualizing data of victim killed, injured and total by month from January until October. We are using line graph to visualize the distribution of the data.



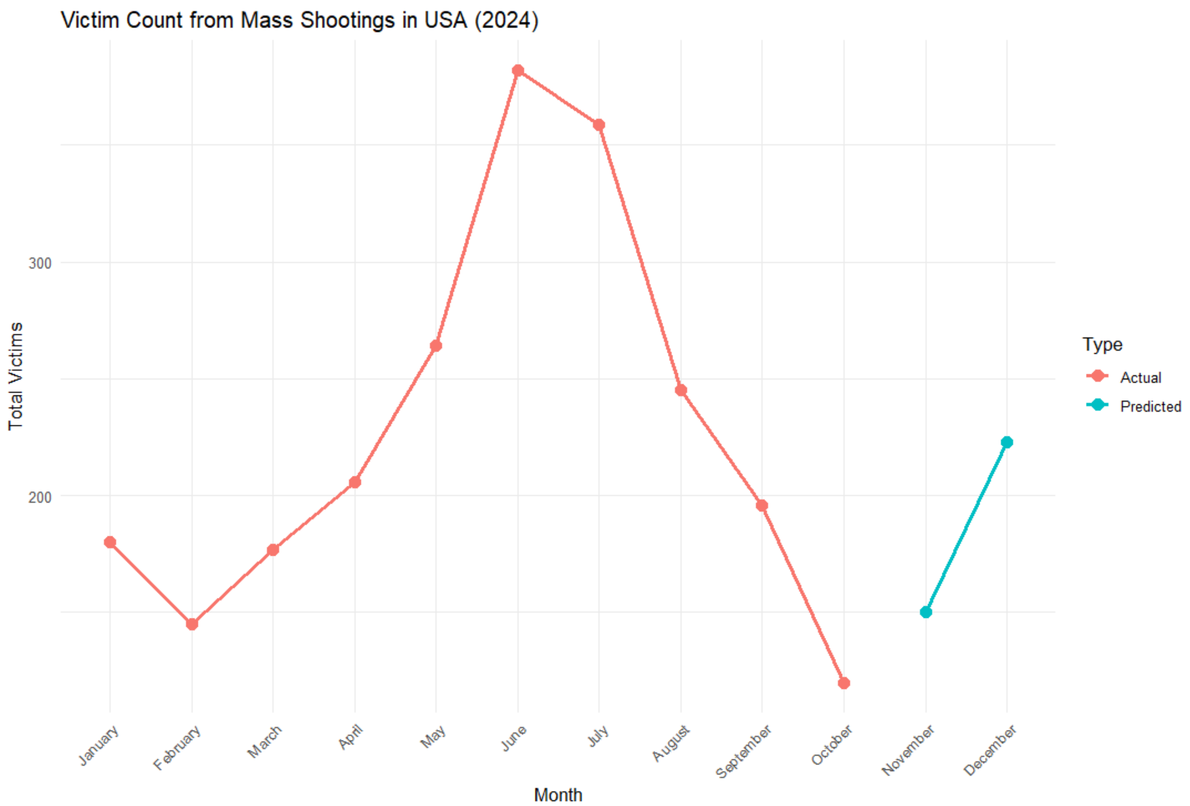
Victim Count from Mass Shootings in USA (2024) line chart

This graph shows the monthly number of victims from mass shootings in the US for 2024, divided into three groups: injured, killed and total (injured + killed). The number of victims increased steadily from January, reaching a peak in June, before decreasing by October. The number of people killed follows a similar pattern, with a sharp increase during the summer, especially in June. Meanwhile, the number of injured victims remained more consistent throughout the year, with only minor changes. The summer months, especially June and July, show the highest figures, which could mean that certain factors make catches more common at this time. This graph highlights the need to study these trends to understand the causes and find ways to prevent such incidents during this high-risk month.

1.4 Solution

The dataset provides information on gun violence incidents in the United States, specifically mass shootings, from January to October 2024. The goal is to predict the number of victims in the remaining months, November and December, using the data from the first ten months. The dependent variable in this analysis is the total number of victims, which is the sum of those killed and injured in these incidents, while the independent variable is the month of the year. The reason why is to lower the risk and number of total victims for all the months.

There are two different data type used to calculate the prediction of total victims, which are dependent data which is the total victims itself and the independent variable which is the month. To improve the prediction accuracy, outliers in the dataset are identified and replaced with upper or lower extreme values, ensuring the model is based on more reliable data. For the prediction, a polynomial regression model is used because of its ability to handle non-linear relationships between the variables and since the data is still numerical. Unlike linear regression, which assumes a constant rate of change, polynomial regression can accommodate the curved, fluctuating patterns often observed in mass shooting data over the course of the year. This flexibility allows the model to better capture trends and seasonal variations. However, choosing the right degree of freedom for the polynomial regression is essential to avoid overfitting, where the model would be too complex and inaccurate, or underfitting, where it would not capture important patterns. The ideal degree of freedom is found through trial and error to balance accuracy and model complexity.



Predicted total victims count from mass shooting in USA for month November and December (2024)

The analysis is conducted using the R programming language, with several libraries aiding the process. The code uses several different libraries in order to access installed packages, dplyr package is used for data manipulation, ggplot2 is employed for visualizing the data, and readr simplifies the reading and handling of the dataset. After applying the polynomial regression model, predictions are made for the number of victims in November and December 2024. The model's accuracy is assessed using the Mean Absolute Percentage Error (MAPE), a metric that compares the predicted values to the actual values, providing insight into the prediction's reliability. With a MAPE of 7.37%, the model shows a relatively low error rate, indicating that the predictions for the last two months of the year are likely to be accurate.

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

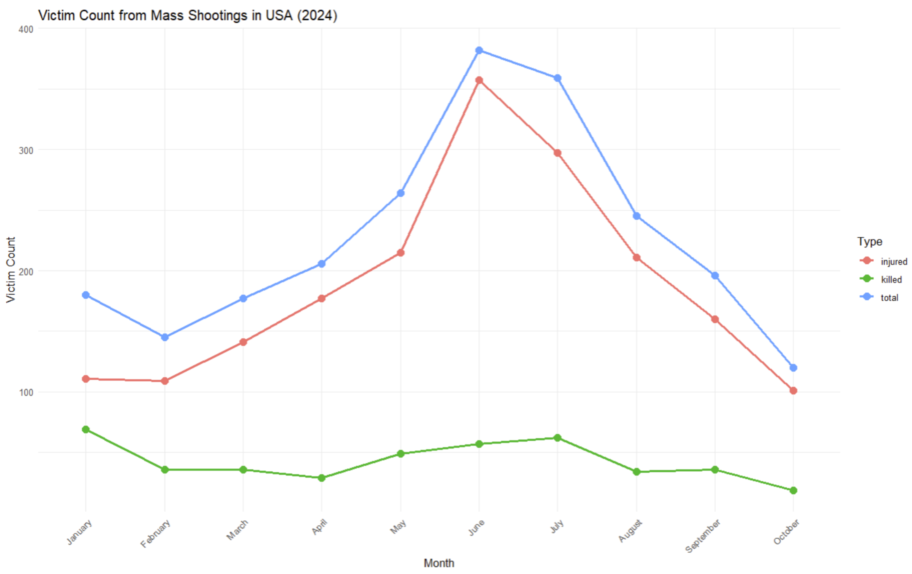
MAPE Formula

The results suggest a significant spike in victims during the summer months which is on June, with a slight decrease in subsequent months. Predictions for November and December are 150 and 223 victims, respectively, based on the predicted values from the R code, indicating an increase in gun violence towards the year's end. These forecasts are critical for policymakers and emergency responders in terms of resource allocation, planning, and preventive strategies. While the model is based on the available data, it does not account for external factors such as socio-economic conditions or significant policy changes, which could influence mass shooting incidents. Nevertheless, the model provides valuable insights into the trends and can be a useful tool for enhancing public safety and reducing gun violence in the future.

1.5 Discussion on Findings

The dataset comprises a total of number of gun violence victims that is combination of victims killed and injured across 1 to 10 months, with 11 and 12 months predicted using a polynomial regression model. The increasing a total of victims was observed in 6 months to indicate a specific incident that driving the growth of incidents. With this, it underscores the importance of identifying regular patterns for resource allocation and preventing an action.

Findings on Victim Count from Mass Shootings in USA (2024) line chart:



Victim Count from Mass Shootings in USA (2024) line chart

The Victim Count from Mass Shootings in USA (2024) line chart shows the victim count (killed, injured, and total) from mass shootings in the USA in 2024, with the following trends observed:

1. **Overall Trend (Total Victim – Blue Line)**

The total victim counts increase gradually from January through June to reach its peak, indicating that mass shooting incidents might have been on the rise during this period. The number of victims then begins to slide downward after June, with a sharp drop seen in the months from July to October. It could be that there was indeed a decrease in either the frequency or intensity of mass shootings as the year unfolded.

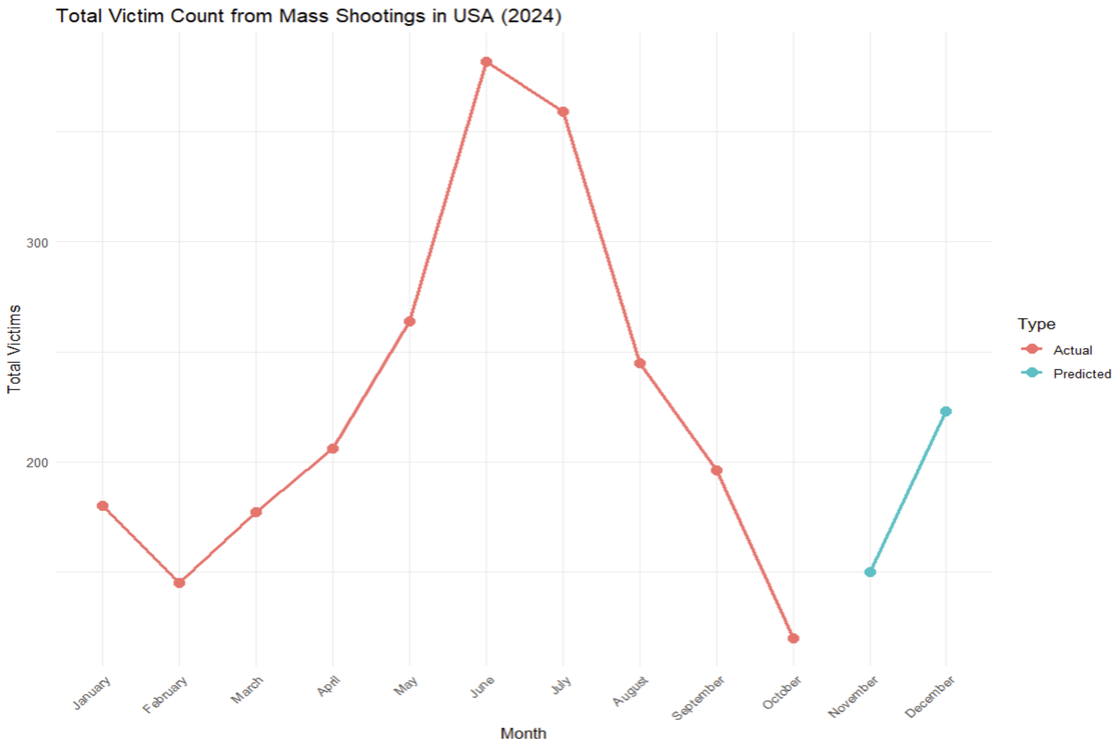
2. **Victims Killed (Green Line)**

The number of victims killed stays relatively even month after month, with a slight increase from May through June. In contrast to the previous categories, the change in casualty count is not as significant, which could indicate that patterns of casualty in mass shootings are much more stable or less influenced by monthly difference.

3. **Victims Injured (Red Line)**

The number of victims Injured also reveals a similar pattern to the total victim count, peaking in June. However, the difference between the total victims and those who are injured will suggest that most incidents of mass shootings cause more injuries compared to the number of those dead. This could point to either the severity of the incidents or how the responses by the emergency services turn out.

Findings on Victim Count from Mass Shootings in USA (2024) line chart:



Total Victims count from Mass Shooting in USA (2024) line chart

The Total Victim Count from Mass Shootings in USA (2024) line chart shows actual total victims count (January – October) and predicted total victims count (November and December) from mass shootings in the USA in 2024, with the following trends observed:

1. **Trends of Actual Data of Total Victim Count (January - October):**

The number of victims from US mass shootings has seen a significant uptick from January to June 2024, reflecting that both frequency and severity have risen at an alarming rate in the first half of the year. This upward trajectory peaked in June 2024, when the total victim count reached its highest value recorded. Beginning with this peak, however, the victim count started to decline. This shows a seeming decrease in mass shooting incidents from July to October.

2. **Trends of Predicted Data of Total Victim Count (November – December):**

A noticeable peak in victims is observed in November to December 2024, indicating a potential rise in mass shooting incidents.

Findings on the prediction model:

A polynomial regression model with a degree of freedom (df) of 6 was used to predict total victims for November and December to the nonlinear nature of data. This model captures a relationship more effectively than linear regression while maintaining the numerical data. The degree of freedom was optimized through trial and the error of balance the trade-off between underfitting and overlifting. Excessively high degrees were avoided as they would lead to overlifting, capturing noise rather than true trend. Prediction accuracy was evaluated using MAPE (Mean Absolute Percentage Error).

1. **Data Prediction:**

The model used total victim data from monthly_victims data frame to predict the value of total victims in November and December.

	Month	Total_Victims	Victims Killed	Victims Injured	Month_Num
1	January	180	69	111	1
2	February	145	36	109	2
3	March	177	36	141	3
4	April	206	29	177	4
5	May	264	49	215	5
6	June	382	57	357	6
7	July	359	62	297	7
8	August	245	34	211	8
9	September	196	36	160	9
10	October	120	19	101	10

monthly_victims data frame

Predicted total victim in November and December:

November: 150 victims

December: 223 victims

2. **Mean Absolute Percentage Error (MAPE) Evaluation**

The model evaluation using MAPE evaluation method is consist of using actual and predicted data of total victims in January to October 2024.

	Month	Total_Victims
1	January	180
2	February	145
3	March	177
4	April	206
5	May	264
6	June	382
7	July	359
8	August	245
9	September	196
10	October	120

	predicted_values
1	123
2	153
3	160
4	177
5	179
6	210
7	275
8	296
9	347
10	354

Actual data vs Predicted data of Total Victim (January – October 2024)

Result of Evaluation:

```
Mean Absolute Percentage Error percentage of the model: 7.37 %
```

The model's Mean Absolute Percentage Error (MAPE) was calculated to be 7.37%. This means that the model has an average error of about 7.37% with respect to the actual values when making the forecast of the total victim count. In general, within predictive modelling, a MAPE less than 10% is usually regarded as indicating a good model fit. This entails that the model does quite a reasonable job in predicting the total victim count from mass shootings in the USA.

The dataset range is limited to 10 months of data, and the predictions for month of data, and the predictions for November and December rely heavily on the established trends and patterns. These limitations could affect the model's ability to capture long-term seasonality. The predictions do not account for external factors like socioeconomic conditions, legislative changes, or major events that would influence gun violence trends. The predictions can guide authority by identifying months with higher predicted victim counts. This enables better allocation of resources, targeted prevention strategies, and timely inventions.

Implication:

The implication from this analysis should provide clear warning for USA country to take serious action to avoid the mass shootings victims arise as what have been predicted from the data analysis. Some of the action that important to be taken to tackle the issue of mass shootings:

Some of the following steps could be taken or initiated by the USA, in view of the trend shown in the chart above, to tackle the issue of mass shootings:

- Stricter Implementation of Gun Control Laws
- Lengthening the waiting period for purchasing guns.
- Enhancing Mental Health Services.
- Improve security measures in schools, places of worship, and other public places.
- Invest in technology that will make it easier to detect and prevent threats.
- Share research findings with policymakers and the public.

1.6 Conclusion

In conclusion, this analysis followed a data-driven approach to identifying high-risk areas for gun violence by utilizing incident records and machine learning techniques. From systematic data cleaning, feature engineering, and application of the Random Forest model, we were able to classify locations into high-risk and low-risk categories based on incident frequency and geographic information.

Key Findings:

1. Risk Stratification
2. Model Performance
3. Feature Insights

Implications and Applications:

The practical implications of this study are seen in the realms of law enforcement, policymaking, and community organizations. Conclusively, through the identification of high-risk areas, patrols, community outreach, and other preventive measures can be channelled to where they are most needed. In essence, the application of machine learning in this context makes the solution scalable and automated enough to solve a societal challenge.

Limitations and Future Directions:

This analysis does provide a number of insights, but there is still room for improving it. In the binary classification of risk, for example, multi-level risk classes can be considered, or other sources of data can be used, such as demographic, socioeconomic, and temporal data. Refining high-risk thresholding based on domain expertise or advanced statistical method needs to be done for enhanced precision. Further optimization of hyperparameters may be required, and other machine learning algorithms need to be explored.

In general, this study developed evidence for the benefit of combining data analysis with predictive modelling in dealing with complex social problems. The insights gained here can serve as the base in informed decision-making and further research aimed at trying to reduce gun violence and making communities safe.

1.7 Appendix

```
# This is the R script file for the analysis to make prediction of total victim count

# Library imported

library(readr)

library(ggplot2)

library(dplyr)

# General dataset

dataset_gun <- read_csv("C:/Users/user/OneDrive/Desktop/Syaz/2nd Year/2nd sem/DS/Project/us_gun_violence_data_2024.csv")

# Data Cleaning

# Column State & Coordinates_Found containing missing values therefore data cleaning is required

    # Handling missing values in State column

missing_state <- which(is.na(dataset_gun$State))

correct_state <- c("Tennessee", "Ohio", "New York", "Alabama", "Ohio", "California", "Texas", "California")

dataset_gun$State[missing_state] <- correct_state

    # Handling missing values in Coordinates_Found column

for (i in 1:nrow(dataset_gun)){

    if (is.numeric(dataset_gun$Latitude[i]) && is.numeric(dataset_gun$Longitude[i])){

        dataset_gun$Coordinates_Found[i] <- "Yes"

# Value Yes if got coordinate on latitude and longitude
```

```
    }else if (!is.numeric(dataset_gun$Latitude[i]) ||
!is.numeric(dataset_gun$Longitude[i])){

        dataset_gun$Coordinates_Found[i] <- "No"

# Value No if no coordinate on latitude or longitude

    }

}

    # Create new features: 'Total Victim' and 'Month'

dataset_gun$'Total Victim' <- dataset_gun$'Victims Killed' +
dataset_gun$'Victims Injured'

dataset_gun <- dataset_gun[, c(1:7, ncol(dataset_gun),
9:ncol(dataset_gun)-1)]

# Position 'Total Victim' column next to 'Victims Injured' column


dataset_gun$`Incident Date` <- as.Date(dataset_gun$`Incident Date`,
format="%B %d, %Y")

dataset_gun <- dataset_gun %>% arrange(`Incident Date`)

dataset_gun$Month <- format(dataset_gun$`Incident Date`, "%B")

dataset_gun <- dataset_gun[, c(1:2, ncol(dataset_gun),
4:ncol(dataset_gun)-1)]


    # Create new data frame: Monthly victim

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

monthly_victims <- dataset_gun %>% group_by(Month) %>%

summarise(

    Total_Victims = sum(`Total Victim`, na.rm = TRUE),

    'Victims Killed' = sum(`Victims Killed`, na.rm = TRUE),
```

```
    'Victims Injured' = sum(`Victims Injured`, na.rm = TRUE)

    )

View(monthly_victims)

    # Using boxplot to check any outliers in 'Total Victim' column in
monthly_victim data frame

boxplot(monthly_victims$'Victims Killed', main = "Total Victims Killed for
Every Month") # No outliers found

boxplot(monthly_victims$'Victims Injured', main = "Total Victims Injured
for Every Month") # 1 outlier found

boxplot(monthly_victims$Total_Victims, main = "Total Victims for Every
Month") # 1 outlier found

    # handling outliers for Victims Injured

first_q<-quantile(monthly_victims$'Victims Injured', 0.25)

third_q<-quantile(monthly_victims$'Victims Injured', 0.75)

iqr<-IQR(monthly_victims$'Victims Injured')

le<-round(first_q- 1.5 * iqr, digits = 0)

ue1<-round(third_q + 1.5 * iqr, digits = 0)

#avg <- round(mean(monthly_victims$'Victims Injured'), digits = 0)

monthly_victims$'Victims Injured'[monthly_victims$'Victims Injured' > ue1]
<- ue1 # replace the outliers with ue1

boxplot(monthly_victims$'Victims Injured', main = "Check Outlier Victims
Injured")

    # handling outliers for Total Victim

first_q<-quantile(monthly_victims$Total_Victims, 0.25)

third_q<-quantile(monthly_victims$Total_Victims, 0.75)

iqr<-IQR(monthly_victims$Total_Victims)
```

```
le<-round(first_q- 1.5 * iqr, digits = 0)

ue2<-round(third_q + 1.5 * iqr, digits = 0)

#avg <- round(mean(monthly_victims$Total_Victims), digits = 0)

monthly_victims$Total_Victims[monthly_victims$Total_Victims>ue2] <- ue2 #
replace the outliers with ue2

boxplot(monthly_victims$Total_Victims, main = "Check Outlier Total
Victim")


# Data Visualization using line graph

killed <- data.frame(Month = monthly_victims$Month,Victim =
monthly_victims$`Victims Killed`,Type = "killed")

injured <- data.frame(Month = monthly_victims$Month,Victim =
monthly_victims$`Victims Injured`,Type = "injured")

total <- data.frame(Month = monthly_victims$Month,Victim =
monthly_victims$Total_Victims, Type = "total")

plot_data <- rbind(killed, injured, total)


ggplot(plot_data, aes(x = Month, y = Victim, group = Type, color = Type))
+

  geom_line(linewidth = 1) +  # Draw lines

  geom_point(size = 3) +  # Highlight data points

  labs(

    title = "Victim Count from Mass Shootings in USA (2024)",

    x = "Month",

    y = "Victim Count"

  ) +
```

```
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


# Building Predictive Model using polynomial regression modell
    # building the prediction model
monthly_victims$Month_Num <- match(monthly_victims$Month, month.name)
poly_model <- lm(Total_Victims ~ poly(Month_Num, df = 6), data =
monthly_victims)
future_months <- data.frame(Month_Num = c(11, 12))
predict_victimCount <- round(predict(poly_model, newdata = future_months),
digits = 0)


    # update new data frame with predicted values
future_months$Total_Victims <- predict_victimCount
future_months$Month <- c("November", "December")
future_months <- future_months[, c("Month", "Total_Victims", "Month_Num")]


# Prediction model evaluation using MAPE calculation
actual_values <- monthly_victims$Total_Victims
predicted_values <- round(predict(poly_model, newdata = monthly_victims),
digits = 0)
MAPE <- mean(abs((actual_values - predicted_values) / actual_values) *
100)


# Visualize the actual and predicted values
```

```
    # Combine actual and predicted data for plotting
plot_data <- rbind(
  data.frame(Month = monthly_victims$Month, Total_Victims = actual_values,
Type = "Actual"),
  data.frame(Month = future_months$Month, Total_Victims =
future_months$Total_Victims, Type = "Predicted")
)
    # Plot the line chart
ggplot(plot_data, aes(x = Month, y = Total_Victims, group = Type, color =
Type)) +
  geom_line(linewidth = 1) +  # Draw lines
  geom_point(size = 3) +  # Highlight data points
  labs(
    title = "Total Victim Count from Mass Shootings in USA (2024)",
    x = "Month",
    y = "Total Victims"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


cat("Predicted total victim in November and December:\n")
print(predict_victimCount)
cat("Mean Absolute Percentage Error percentage of the model:", round(MAPE,
2), "%\n")
```

2.0 Critical Thinking on Generative AI

Generative AI is an essential tool that is widely used for its ability to reduce workload or turns complex work into something simpler, in short, it is able to make work easier. Not only does it act as a tool to help with work, but it can also do many things such as generate images, videos, or even voice. This technology if used correctly can create many wonders beyond imagination, although it can still be considered a new technology, its impact towards society is very large. However, this piece of technology can pose a threat to many if it is misused as there is not many rules and regulation against the usage of generative AI technology and sometimes it is hard to find any distinction between something real or AI generated. For example, from observational studies, it can be concluded that many students have used a generative AI known as ChatGPT, the most commonly used AI, especially for their studies or assignments, which is proven to be very effective. However, in a different light, using ChatGPT too much can lead to the student being dependent towards the AI and their original thoughts began to wither since most of it are generated by the AI which can be a very serious problem as it affects the mind of the younger generation to think less and depends more on AI. In order to reduce the usage of generative AI amongst students, some institutes have taken the measure to detect any work that is produced entirely by AI so that students can at least try to do the whole work themselves. Some scammers even resorted to using voice changer in order to trick their victims into thinking that they are a trusted personnel member, which can pose as a danger towards those who are unfamiliar with the scam tactics. Although generative AI can be seen as something that is used mostly in a bad way, it is also proven to provide solution to problems and act as a guide or assistant for a much higher quality work. For example, when making an essay, a student can use generative AI in order to make it become a much more professional written essay or if a student is having problem understanding a subject, the student can ask for guidance from the generative AI regarding the subject and it will provide notes or even the full walkthrough of the preferred subject. In conclusion, generative AI can be helpful or it can be harmful, it mostly depends on how the technology is used and who is using it.

- *Adam Shamel, 22000766*

In my point of view on generative ai is tool that often been used by people in public, although it is still new in this decade, there are already many users from all around the world. From my perspective generative ai has a lot of tools that can help people in generate their daily life as for example generative ai can generate text, speech, or visuals to assist people with disabilities, such as creating audio descriptions for the visually impaired or translating text into multiple languages. This is example is used to enhanced their accessibility in order to live life as a normal person. The generative ai also been used in education to help both teachers and students enhanced their idea generating for their studies. For education field generative ai can produce adjusted educational content, simulations, and interactive learning tools, making education more engaging and accessible. This will help both teacher and student understanding what they teaching and learning in the education. Beside enhanced education capability, generative ai also empowers entertainment industries. By enhanced individuals' creativity such as arts, music, literature, sculpture, and photographic, it can it can also improve skills in their respective expertise. Generative ai also can help people public communication by improving enabling governments and organizations to present complex information in a simplified, accessible manner. It can generate summaries of detailed reports, policies, or datasets, making them easier for the public to understand. It also can be used to craft public service announcements adapted to specific audiences, ensuring important messages are effectively conveyed. With all this, it shown that generative ai is really helpful to the society by enhancing all types of fields that can improving every single people in their daily life.

- *Ahmad Faqrullah, 22000450*

Generative artificial intelligence (AI) tools are really giving such a big impact to the world, as for me too. Generative AI starts to bloom and getting everywhere with many types and different functionality all because from the beginning in November 2022 ChatGPT was introduced. It became so popular and really influence the AI booming the internet because it is open source. Now there are tons of generative AI we can simply use through the internet or application in mobile devices. The generative AI that I regularly use is Consensus. Consensus is a generative AI tool designed to simplify research by analysing and summarizing academic literature. It purpose is to provide answers or insight for the user's prompt with evidence-based answer by searching through scientific papers, thesis, articles, journal and many more to generate a concise answers for the user. This AI can be simply known as an AI tool that came from a combination between ChatGPT and Google Scholar. For a university student like me, Consensus gives valuable tools for me doing any research work or finding the best literature material that match on what I am looking for. This AI enables me to have faster access to any credible and relevant sources of information.

The use of Consensus is very simple yet useful. For instance, instead of manually scrolling through multiple academic papers over some topic on Google Scholar, Consensus has the capability to summarize key findings, outline opposing viewpoints, and identify trends within the research field. This functionality provides such a big value in university tasks such as assignments, literature reviews or analysing articles. Consensus saves me time identifying reliable sources and helps in maintaining the information credibility even on very tight deadlines. However, Consensus is not a substitute for applying critical thinking analysis on the academic papers in significantly improving research efficiency. This is because I have to assess its outputs, verify the suggested sources, and put into proper context the AI-generated insights. Generative AI tools are only as effective as the user's ability to utilize the information, use it wisely by understanding the output from the AI prior integrating it into my work. Therefore, Consensus gives a great help in locating and analysing academic papers, but the comprehension and execution that might ensure that my work is accurate or reliable are left for me.

- *Mohamad Syazreeq Akmal, 22000454*

In the rapidly advancing field of artificial intelligence, generative AI has emerged as a powerful tool capable of creating text, images, music, and other forms of content. While its potential for innovation is immense, critical thinking is essential to effectively and responsibly engage with this technology. This involves understanding how generative AI works, evaluating its outputs, addressing ethical considerations, and ensuring its appropriate application in diverse contexts. Critical thinking begins with comprehending the fundamental principles of generative AI. These systems, often powered by machine learning models like GPT or diffusion networks, are trained on massive datasets to produce content that mimics human creativity. However, despite their impressive outputs, these models have limitations. They lack true understanding, operate based on patterns in their training data, and can generate content that is nonsensical, biased, or factually incorrect. Recognizing these constraints is the first step toward informed and cautious use. One of the key aspects of critical thinking is evaluating the quality of generative AI's outputs. Users must assess the accuracy of the content, particularly when it is used in sensitive fields such as education, healthcare, or journalism. For instance, while AI-generated text can be persuasive, it may inadvertently include factual errors or reflect biases present in the training data. Similarly, AI-generated images or videos must be scrutinized for authenticity, as they can be used to spread misinformation or deceive audiences. Evaluating appropriateness also involves ensuring the content aligns with ethical and cultural norms. Generative AI's applications vary widely, from education and art to business and decision-making. Critical thinking ensures its ethical and effective use in each context. In education, for instance, AI tools should foster learning rather than encourage academic dishonesty. Students and educators must balance using AI for assistance with preserving the integrity of the learning process. Similarly, in the creative industries, generative AI raises questions about originality, copyright, and the balance between human and machine creativity.

- *Muhammad Aqfa Bin Alias,21001180*

Group opinion on generative AI

Generative AI, especially ChatGPT, has quickly become a powerful tool in today's world, offering immense potential to simplify challenging tasks, boost creativity, and improve productivity in various fields. It is widely seen as an innovative technology that has brought significant benefits to education, research, communication, and accessibility. However, while its advantages are noteworthy, it's important to approach its use with care and responsibility to address some of the challenges it presents.

In education, ChatGPT has proven to be an incredibly effective tool for both students and teachers. Students can rely on it to clarify difficult topics, brainstorm ideas, and enhance the quality of their writing. It provides tailored assistance, simplifies complex concepts, and serves as a helpful guide for academic research. Educators, on the other hand, can use it to create engaging learning materials and streamline lesson planning, making teaching more effective and interactive. However, there is a risk of over-reliance, which could affect students' ability to think critically and creatively. To address this, many institutions have introduced systems to detect AI-generated work, encouraging students to develop their own ideas and strengthen their reasoning skills.

Outside the classroom, ChatGPT also shines in improving accessibility and communication. It can assist individuals with disabilities by converting text to speech, translating languages, or generating audio descriptions, helping them engage more fully with the world around them. In public communication, ChatGPT plays a valuable role in breaking down complicated information, such as policies or reports, so it's easier for everyone to understand.

In professional and research settings, ChatGPT is a game-changer for analyzing and summarizing large amounts of information. It can highlight key points from academic papers or technical documents, saving time and making it easier to make well-informed decisions. Tools like Consensus, which combine ChatGPT's capabilities with research databases, provide evidence-based summaries that are especially helpful for academic and professional work. Still, users must remain critical of the AI's output, verifying its accuracy and ensuring it fits the context to maintain credibility and reliability.

While its advantages are significant, ChatGPT and other generative AI tools also come with risks when used irresponsibly. For instance, voice-generating AI can be misused by scammers to impersonate trusted individuals, and biased or inaccurate AI-generated content can spread misinformation. In creative industries, there are growing concerns about originality, copyright, and the balance between human and AI contributions. These challenges highlight the need for ethical guidelines and regulations to prevent misuse and promote the responsible use of generative AI.

To sum up, ChatGPT is a transformative tool that can enhance productivity, drive innovation, and close gaps in education, communication, and accessibility. Its benefits are undeniable, but to unlock its full potential for the greater good, users must approach it thoughtfully, balancing its capabilities with ethical considerations, critical thinking, and proper oversight.