Data Analysis Assignment

Dhananjay Devraj

2024-04-22

INRODUCTION

The aim of this analysis is to delve into the datasets crime23.csv and temp2023.csv, extracting meaningful insights and patterns that shed light on policing and climate data in Colchester for the year 2023. These datasets provide a rich source of information regarding street-level crime incidents, daily climate measurements, and weather-related variables collected from a weather station near Colchester.

The crime23.csv dataset, sourced from the UK Police API, contains detailed information about street-level crime incidents in Colchester. It encompasses various attributes such as crime category, location details, outcome status, and more, offering a comprehensive view of the crime landscape in the region.

On the other hand, the temp2023.csv dataset captures daily climate data collected from a weather station in close proximity to Colchester. This dataset includes key meteorological variables such as temperature, wind speed, precipitation, humidity, and more, providing crucial insights into the climatic conditions experienced throughout the year.***

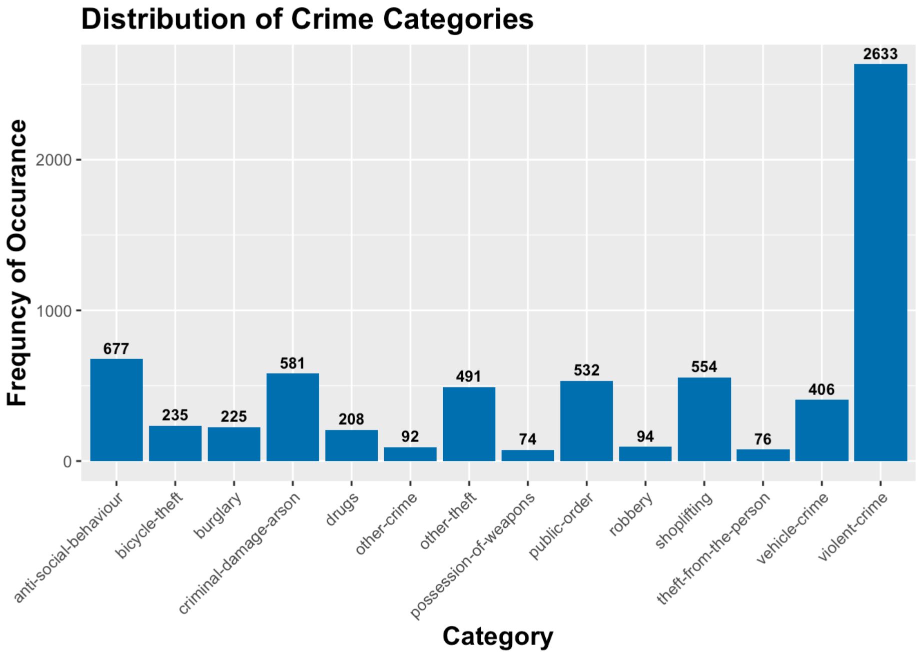
Through thorough analysis and visualization of these datasets using R and a range of data analysis tools such as ggplot2, leaflet, and plotly, this report aims to uncover patterns, trends, and correlations within the data. The findings will not only enhance our understanding of crime dynamics and weather patterns in Colchester but also pave the way for informed decision-making and policy interventions based on data-driven insights.

DATA ANALYSIS AND INTERPRETATION

Lets us begin with exploring the Crimedataset Provided to us.

```
# barchart
crime_df$category <- as.factor(crime_df$category)

ggplot(data = crime_df, aes(x = category)) +
  geom_bar(fill = "#0072B2", stat = "count") +
  geom_text(aes(label = stat(count)),stat = "count", vjust = -0.5,size = 3,color = "black",
            fontface = "bold") +
  labs(title = "Distribution of Crime Categories", x = "Category", y = "Frequncy of Occurance") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title = element_text(size = 14, face = "bold"),
        plot.title = element_text(size = 16, face = "bold"))
```



The bar chart depicting the distribution of crime categories is a crucial visual representation of the frequency of different types of crimes recorded in Colchester during the year 2023. These categories range from anti-social behavior, theft, burglary, and criminal damage to drug-related offenses, robbery, and violent crimes. Each bar in the chart represents a specific crime category, and the height of the bar corresponds to the frequency or count of that particular category.***

The bar chart showcases the frequency of each crime category. It's evident that certain types of crimes, such as violent crime, shoplifting, anti-social behavior, and criminal damage/arson, occur more frequently compared to others.

Key Observations:

1.Violent crime stands out as the most prevalent category with 2633 reported incidents. This highlights the need for measures to address violence in the community.
2.Shoplifting and anti-social behavior also have notably high frequencies, indicating potential challenges related to theft and public order.

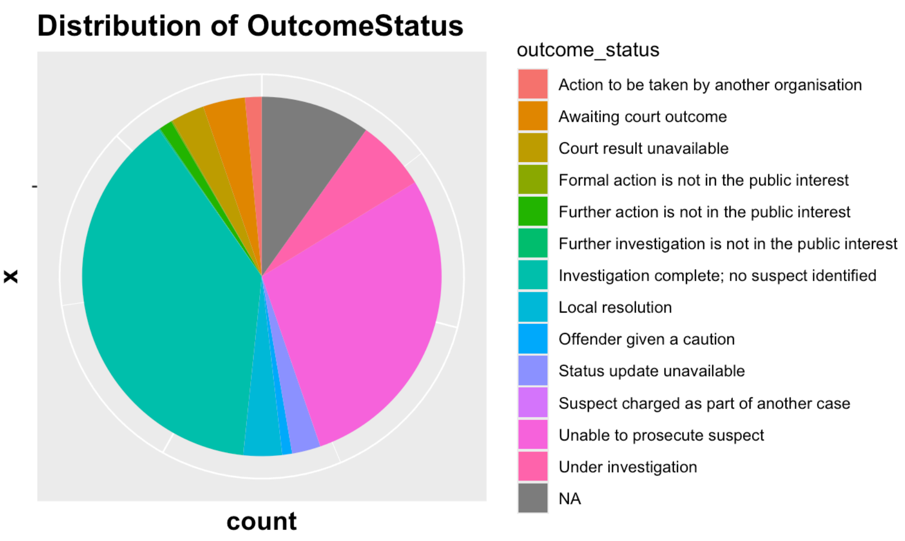
3.Other categories like possession of weapons and theft from the person show relatively lower but still significant numbers, indicating a diverse range of criminal activities in the area.

Understanding the distribution and frequency of different crime types is crucial for law enforcement agencies and policymakers. It helps in allocating resources effectively, devising targeted strategies for crime prevention, and enhancing public safety measures.

Overall, this bar chart serves as an essential visual tool in analyzing and presenting the distribution of crime categories in Colchester, providing valuable insights into the patterns and trends of street-level crimes in the region.

Now let us see how the outcome status of all of the crimes that has happend in colchester

```
ggplot(data = crime_df, aes(x = "", fill = outcome_status)) +
  geom_bar(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Distribution of OutcomeStatus") +
  theme(axis.text.x = element_blank(),
        axis.title = element_text(size = 14, face = "bold"),
        plot.title = element_text(size = 16, face = "bold"))
```



The distribution of outcome statuses provides valuable insights into the resolution of reported crimes in Colchester during 2023. Investigative outcomes varied widely, with “Investigation complete; no suspect identified” being the most frequent outcome (2,656 incidents), indicating challenges in identifying perpetrators or lack of evidence leading to closure without prosecution. This was followed by “Unable to prosecute suspect” (1,959 incidents), highlighting complexities or limitations in legal proceedings. On the other end, outcomes such as “Suspect charged as part of another case” and “Formal action is not in the public interest” occurred infrequently, suggesting unique circumstances or legal considerations in these cases. These outcomes are pivotal in understanding the overall efficacy of law enforcement efforts and the complexities involved in resolving different types of crimes.

The bar chart which displays the frequencies of various crime categories, shedding light on the prevalent issues within the community. For instance, violent crimes, totaling 2,633 incidents, highlight the urgency and significance of law enforcement efforts, encompassing offenses such as assault, robbery, and homicide. These incidents often lead to outcomes such as “Action to be taken by another organization,” “Awaiting court outcome,” or “Suspect charged as part of another case,” depending on the progress of investigations and legal proceedings. On the other hand, anti-social behavior, accounting for 677 incidents, represents disruptive behaviors impacting community well-being, with outcomes ranging from “Local resolution” to “Further action is not in the public interest.” Shoplifting, with 554 incidents, reflects property-related thefts, where outcomes vary from minor resolutions to challenges in prosecution. Additionally, criminal damage and arson, totaling 581 incidents, encompass vandalism and property damage issues, leading to outcomes like “Action to be taken by another organization” or “Unable to prosecute suspect” based on evidence and legal proceedings’ status. These connections underscore the complex landscape of crime management, necessitating tailored strategies for addressing diverse challenges effectively.

```
# crime occurences in each location type

two_way_table <- table(crime_df$location_type, crime_df$category)

kable(two_way_table, caption = "Crime Categories frequency vs location type", row.names = TRUE) %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE)
```

Crime Categories frequency vs location type

	anti-social-behaviour	bicycle-theft	burglary	criminal-damage-arson	drugs	other-crime	other-theft	possession-of-weapons	public-order	robbery	shoplifting	theft-from-the-person	vehicle-crime	v
BTP	0	4	0	1	0	0	4	0	6	0	0	0	1	

	anti-social-behaviour	bicycle-theft	burglary	criminal-damage-arson	drugs	other-crime	other-theft	possession-of-weapons	public-order	robbery	shoplifting	theft-from-the-person	vehicle-crime	v
Force	677	231	225	580	208	92	487	74	526	94	554	76	405	

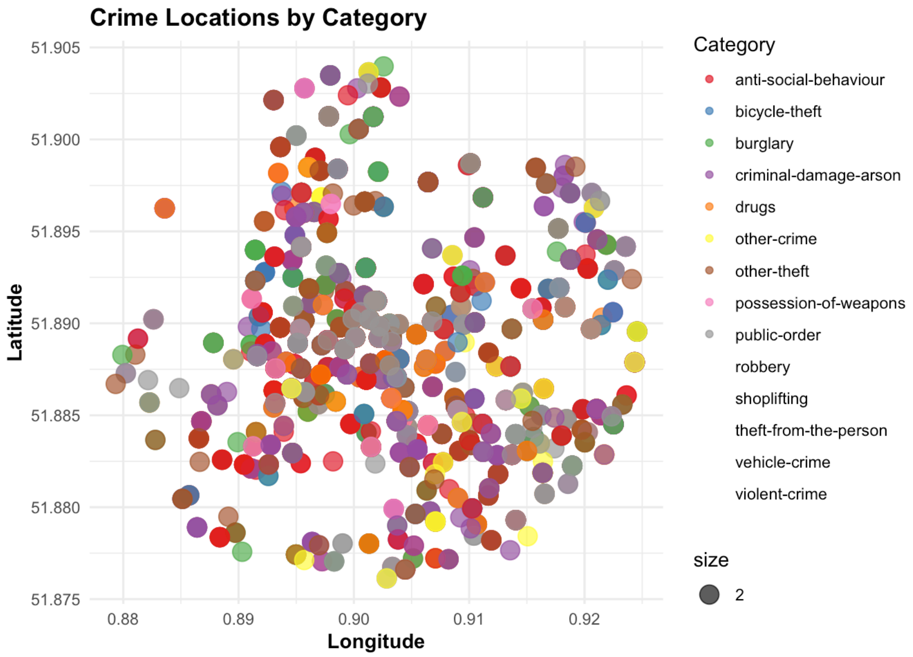
The table presents a comparison of crime categories across two distinct location types: Force and BTP (British Transport Police). Force denotes locations under a normal police force, while BTP indicates British Transport Police locations, which are within the boundaries of a regular police force.

Within the BTP category, the most prominent crime is “Theft from the person,” with eight reported incidents. This is followed by “Bicycle theft” and “Other theft,” each having four reported incidents. “Robbery” and “Possession of weapons” have six and one reported incidents, respectively, indicating a range of criminal activities within BTP locations.

Conversely, the Force category highlights significant concerns in terms of “Violent crime,” with 2,625 reported incidents, reflecting a substantial focus on public safety and law enforcement efforts. “Shoplifting” and “Criminal damage/arson” are also noteworthy, with 554 and 580 reported incidents, respectively, pointing to property-related crimes and vandalism issues. Additionally, “Anti-social behavior” and “Vehicle crime” have 677 and 405 reported incidents, respectively, showcasing a diverse range of criminal behaviors within regular police force locations.

We will see how the datapoints are scattered over different latitude and logitude values.

```
ggplot(crime_df, aes(x = long, y = lat, color = category, size = 2)) +
  geom_point(alpha = 0.7) +
  scale_color_brewer(palette = "Set1") +
  labs(title = "Crime Locations by Category", x = "Longitude", y = "Latitude", color = "Category") +
  theme_minimal() +
    theme(legend.position = "right",
        plot.title = element_text(face = "bold"),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"))
```

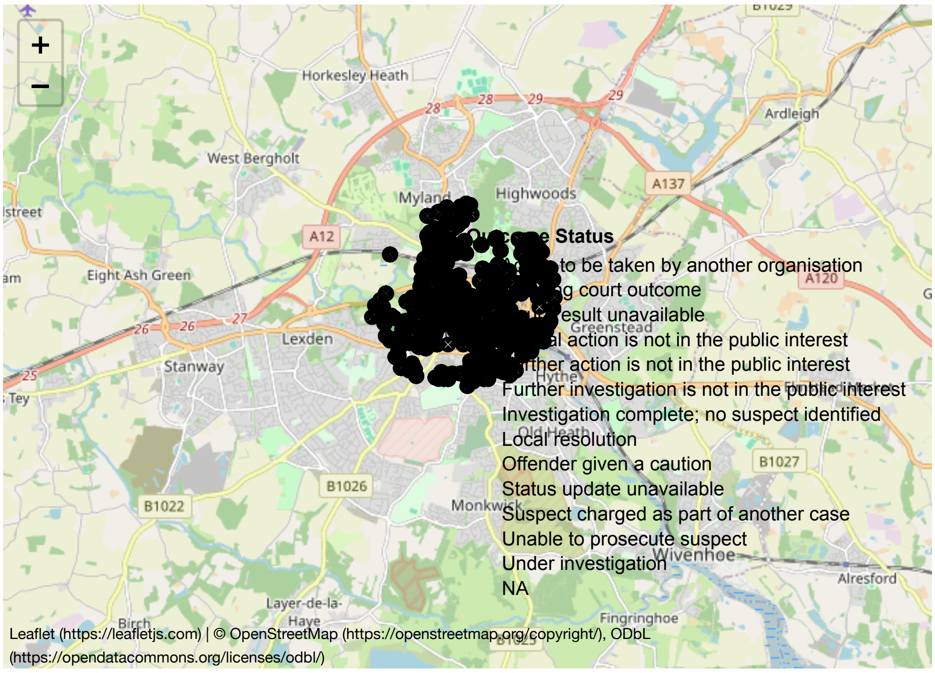


The scatter plot offers a visual representation of the correlation between different crime categories and their respective geographical coordinates, delineated by latitude and longitude. An observation from the plot indicates a notable concentration of criminal incidents occurring at longitude values exceeding 0.89, suggesting a higher occurrence rate in regions characterized by these longitudes. Conversely, the lower portion of the graph, particularly in terms of latitude, exhibits a denser clustering of data points, indicative of increased criminal activity in areas characterized by lower latitudes within the region.

We can Enhance the above visualization using map/leaflet visuals.

```
crime_map <- leaflet(data = crime_df) %>%
  # Add tiles for the base map
  addTiles() %>%
  # Add markers for crime incidents with custom icons and colors based on outcome status
  addAwesomeMarkers(~long, ~lat, icon = customIcons, popup = ~paste("Category:", category, "Street:", street_nam
e)) %>%
  # Set initial map view
  setView(lng = mean(crime_df$long), lat = mean(crime_df$lat), zoom = 12)%>%
  # Add legend for outcome status colors
  addLegend("bottomright", title = "Outcome Status",
            colors = unique(colors), labels = legend_labels,
            opacity = 1)

# Display the map
crime_map
```


+
−
Horkesley Heath
West Bergholt
Myland
Highwoods
Ardleigh
A137
A12
A120
Eight Ash Green
Lexden
Stanway
Greenstead
Hythe
Old Heath
B1027
B1026
B1022
Monkwick
Wivenhoe
Alresford
Layer-de-la-Haye
Birch
Fingringhoe
e Status
to be taken by another organisation
g court outcome
esult unavailable
al action is not in the public interest
rther action is not in the public interest
Further investigation is not in the public interest
Investigation complete; no suspect identified
Local resolution
Offender given a caution
Status update unavailable
Suspect charged as part of another case
Unable to prosecute suspect
Under investigation
NA
Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org/copyright/), ODbL (https://opendatacommons.org/licenses/odbl/)

The leaflet/map plot created for the data in crime23.csv provides a visual representation of anti-social behavior incidents based on their geographical locations. Each marker on the map corresponds to a specific incident, with the latitude and longitude coordinates indicating where each event occurred.

The map allows us to observe the distribution of anti-social behavior incidents across different areas, as indicated by the markers. For example, incidents are clustered around Military Road, Culver Street West, Ryegate Road, Market Close, and Lisle Road, among other locations. You may learn more about each occurrence, by hovering your cursor over the markers. The fact that incidences involving different types of crimes in the same location and resulting in different outcomes is an intriguing finding. The observation implies that the outcomes of incidents, influenced by contextual factors and law enforcement responses, reflect the complexity and interconnectedness of crime dynamics and community dynamics.

Additionally, the color or styling of markers on a map can provide valuable insights into the data. In this case, the use of different marker colors based on the outcome status of each incident can help viewers quickly identify and understand the nature or severity of each event. However, it’s crucial to note that the outcome_status column in the provided dataset contains missing values (“”). These missing values can pose a challenge when trying to visualize outcomes based on color or style.

Overall, the crime23.csv dataset’s data can be visualised using the map made in R using the “leaflet” package.Hovering over the markers makes it simple to explore the data and gives a generalised perspective of where occurrences of crime over different regions. The leaflet/map plot offers a spatial perspective on the occurrence of anti-social behavior incidents, highlighting hotspots and patterns that may be of interest for further analysis or investigation.

Lets us explore the Temp dataset provided to us, which contains daily climate data collected from a weather station close to Colchester

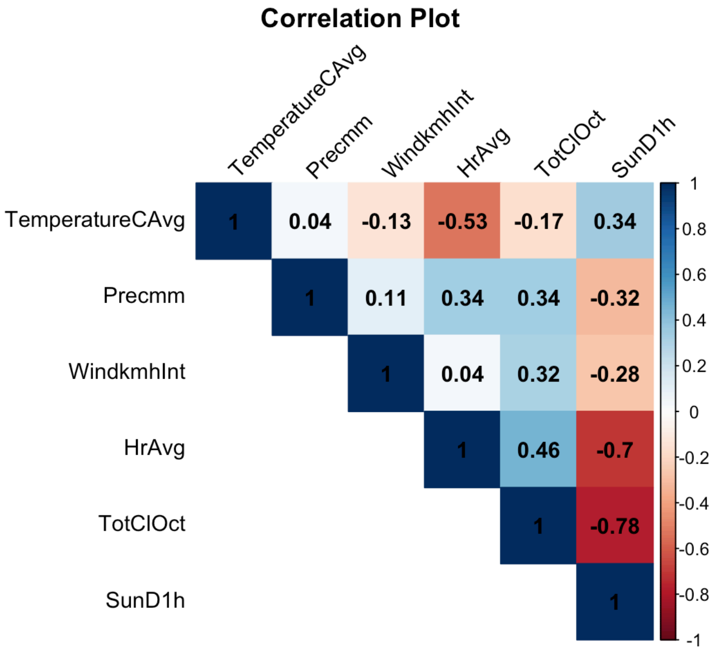
```
summary(temp_df)

head(temp_df)

# Remove NA values and calculate correlation matrix
correlation_matrix <- cor(temp_df[, c("TemperatureCAvg", "Precmm", "WindkmhInt", "HrAvg", "TotClOct", "SunD1h")],
use = "complete.obs")

# Print the correlation matrix
print(correlation_matrix)

# Create a correlation plot with customized aesthetics
corrplot(correlation_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45, tl.pos = "lt",
         addCoef.col = "black", number.cex = 1,
         title = "Correlation Plot", mar = c(0,0,2,0))
```



The correlation matrix shows the relationships between various weather variables, revealingh the correlation coefficients among the weather variables TemperatureCAvg, Precmm, WindkmhInt, HrAvg, TotClOct, and SunD1h.

Each variable's correlation with itself is shown in the matrix's diagonal elements and is always 1, representing perfect correlation.

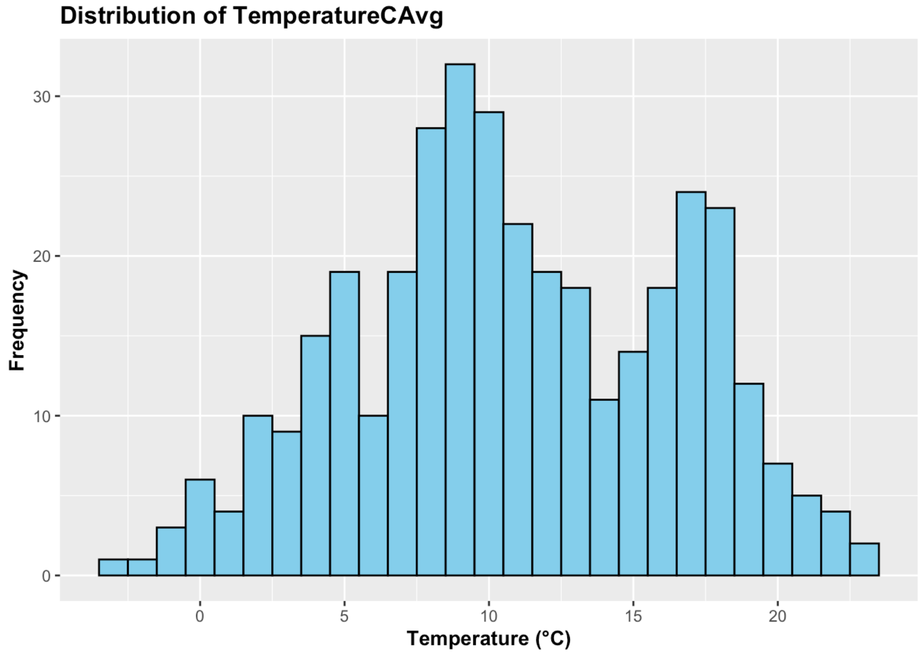
Among the variables, the strongest negative correlation (-0.53) is observed between TemperatureCAvg and HrAvg, indicating that as temperature rises, the hourly average tends to decrease. Conversely, there is a moderate positive correlation (0.34) between SunD1h and TemperatureCAvg, implying that higher temperatures are associated with increased sunlight duration.

A strong negative correlation (-0.78) is found between TotClOct and SunD1h, indicating that higher total cloud cover is linked to decreased sunlight duration. Additionally, there is a moderate positive correlation (0.46) between HrAvg and TotClOct, suggesting that as the hourly average increases, total cloud cover also tends to increase.

Overall, these correlations provide insights into how the different weather variables interact with each other, which can be valuable for understanding weather patterns and making informed decisions based on weather forecasts.

AVG. TEMPERATURE DISRIBUTION OVER THE ENTIRE PERIOD

```
ggplot(temp_df, aes(x = TemperatureCAvg)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Distribution of TemperatureCAvg", x = "Temperature (°C)", y = "Frequency")+
  theme(legend.position = "right",
        plot.title = element_text(face = "bold"),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"))
```



The histogram reflects the distribution of average temperatures and provides insights into the shape of this distribution, whether it is symmetric, skewed, or exhibits unusual patterns. The x-axis denotes temperature in degrees Celsius, while the y-axis indicates the frequency of days falling within specific temperature ranges. Each bar in the histogram represents a temperature range, with its height corresponding to the number of days falling within that range.

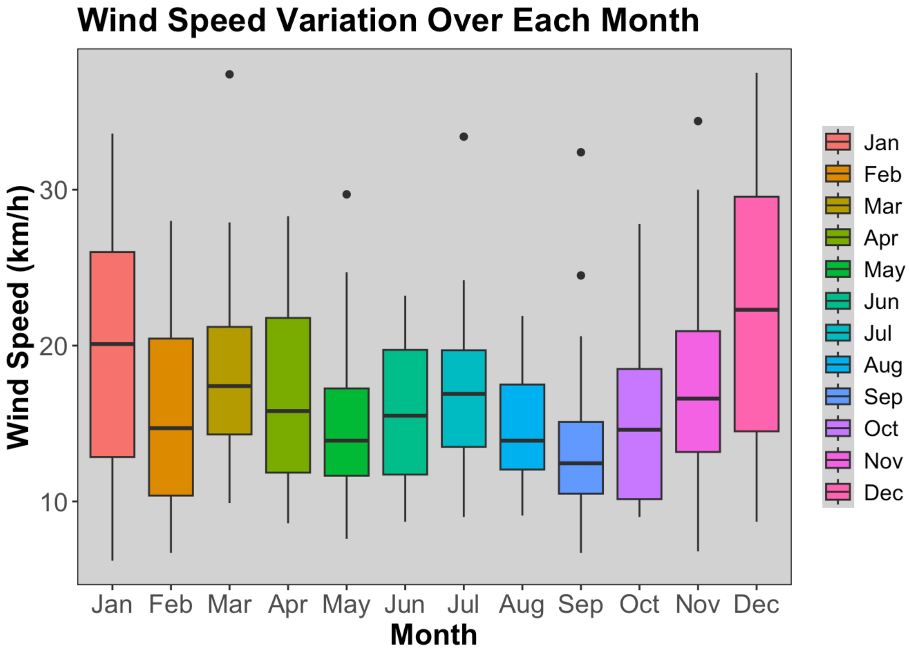
By analyzing the histogram, one can identify temperature clusters and observe where the most frequent temperatures occur. The normally distributed histogram of average temperatures suggests that the majority of days throughout the analyzed period experienced average temperatures clustered around a central value. This central value is represented by the peak or highest point of the histogram, indicating the most common or typical temperature observed. The even distribution of temperatures in a normal histogram suggests that weather conditions in the region were consistent and stable over the analyzed period which ranges from 5°C to 15°C. It also implies that extreme weather events, such as heatwaves or cold snaps, were relatively rare during this time frame.

BOXPLOT TO DEPICT THE WIND SPEED OVER EACH MONTH

```
temp_df$Date <- as.Date(temp_df$Date)

# Extract month from Date column
temp_df$Month <- month(temp_df$Date, label = TRUE)

ggplot(temp_df, aes(x = Month, y = WindkmhInt, fill = Month)) +
  geom_boxplot() +
  labs(
    title = "Wind Speed Variation Over Each Month",
    x = "Month",
    y = "Wind Speed (km/h)",
    fill = "Month"
  ) +
  scale_fill_discrete(name = "Month",
                      labels = month.abb) +  # Use abbreviated month names in legend
  theme(plot.title = element_text(size = 18, face = "bold"),
        axis.title = element_text(size = 16, face = "bold"),
        axis.text = element_text(size = 14),
        legend.title = element_blank(),
        legend.text = element_text(size = 12),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_rect(color = "black", fill = NA),
        panel.background = element_rect(fill = "lightgray"))
```



The boxplot visually represents the variation in wind speed across different months. Each box in the plot represents the distribution of wind speeds for a specific month, allowing for a quick comparison of wind speed characteristics over time.

The height of the box indicates the interquartile range (IQR), which represents the middle 50% of the data. The thicker line within the box represents the median wind speed for that month, providing a central measure of tendency. The whiskers extending from the box show the range of wind speeds within 1.5 times the IQR above and below the upper and lower quartiles, respectively. Any data points beyond this range are considered outliers and are plotted individually as points.

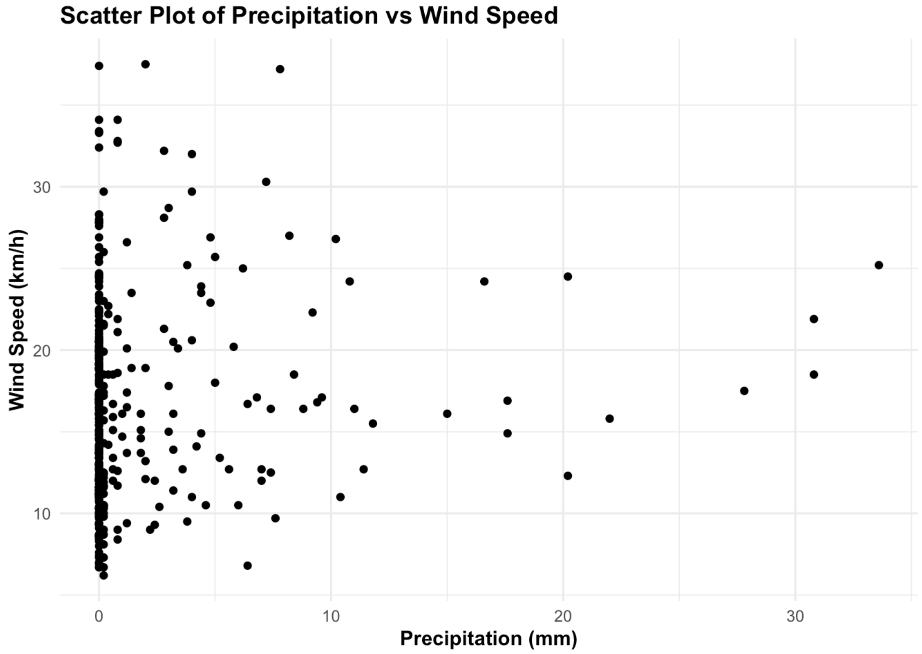
The analysis of the boxplot data unveils distinct seasonal patterns in wind speed fluctuations. Notably, months characterized by elevated median wind speeds and wider box distributions, such as December and January, signify periods of heightened wind activity. Conversely, a decline in wind speeds is observed in the subsequent months of February, March, and April, indicated by narrower box widths and lower median values. The remaining months exhibit relatively consistent wind speed levels, with August and September recording the least wind activity. This pattern elucidates the seasonal variability in wind speeds and provides valuable insights into meteorological trends throughout the year.

Overall, the boxplot provides a clear and concise way to visualize the variation in wind speed across different months, allowing for easy comparison and identification of trends or outliers.

Let us analyse by comparing how precipitation (in millimeters) is affected by wind speed (in kilometers per hour) by analyzing a scatter plot and calculating the correlation coefficient between these two variables.

DOES WINDSPEED AFFECT THE PRECIPITATION OR HAS CORRELATION WITH IT? LET US CHECK.

```
ggplot(temp_df, aes(x = Precmm, y = WindkmhInt)) +
  geom_point() +
  labs(title = "Scatter Plot of Precipitation vs Wind Speed",
       x = "Precipitation (mm)",
       y = "Wind Speed (km/h)") +
  theme_minimal()+
  theme(legend.position = "right",
        plot.title = element_text(face = "bold"),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"))
```



```
# Calculate correlation coefficient
correlation_coefficient_wind <- cor(temp_df$Precmm, temp_df$WindkmhInt)
print(paste("Correlation Coefficient:", correlation_coefficient_wind))
```

In the scatter plot depicting precipitation versus wind speed, a notable observation is the distribution of data points. The majority of points cluster densely along the y-axis, indicating consistent wind speeds across various precipitation levels. However, as precipitation levels increase beyond 10 millimeters, the number of data points sharply declines, with very few instances of high precipitation levels accompanied by high wind speeds.

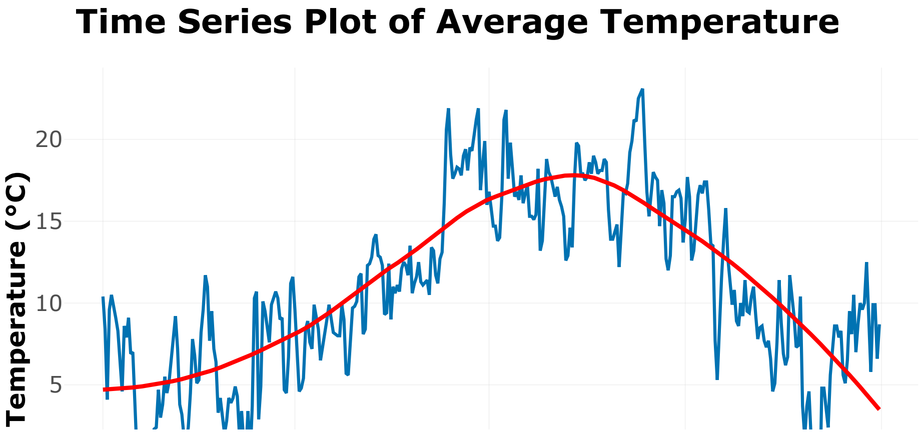
Despite the visually apparent pattern in the plot, the correlation coefficient calculation reveals that there is no significant linear correlation between precipitation and wind speed. The correlation coefficient value, which is close to zero or within a negligible range, suggests that changes in precipitation levels do not have a linear impact on wind speed variations. This finding is in line with the observation from the plot, where the spread of data points does not exhibit a clear trend or pattern along a straight line.

INTERACTIVE TIMESERIES PLOT FOR AVG. TEMPERATURE THROUGH DIFFERENT MONTHS

```
start_date <- as.Date("2023-01-01")
end_date <- as.Date("2023-12-31")
p <- ggplot(data = temp_df, aes(x = Date, y = TemperatureCAvg)) +
  geom_line(color = "#0072B2", size = 0.6) +  # Custom line color and size
  geom_smooth(method = "loess", se = FALSE, color = "red", size = 0.8) +  # Add LOESS smooth line
  labs(title = "Time Series Plot of Average Temperature",
       x = "Date",
       y = "Temperature (°C)") +
  xlim(start_date, end_date) +
  theme_minimal() +  # Use a minimalistic theme
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.grid.major = element_line(color = "lightgray"),
        plot.title = element_text(size = 18, face = "bold"),
        axis.title = element_text(size = 14, face = "bold"),
        axis.text = element_text(size = 12),
        legend.position = "none")

# Convert the ggplot object to an interactive plotly object
p <- ggplotly(p)

# Print the interactive plot
p
```



The time series plot visualizes the variation in average temperature over different months within a one-year period. The plot includes a line graph showing the actual average temperatures (in degrees Celsius) and a smoothed line using the LOESS method, providing a trend representation of temperature changes over time.

From the plot, we can observe fluctuations in temperature throughout the year, with distinct patterns such as seasonal shifts and potential trends. The line graph helps identify temperature trends, while the smoothed line offers a more generalized view of temperature changes, highlighting overall patterns or tendencies.

The graph displaying the variation in average temperature over different months exhibits a normal distribution with a slight left skew. This indicates that while the majority of temperature values are centered around the mean, there is a slightly higher frequency of lower temperatures compared to higher temperatures with majority of them falling below 15 degree celcius.

The left skewness suggests that there are relatively fewer occurrences of exceptionally high temperatures compared to lower temperatures throughout the year. This observation aligns with typical seasonal temperature patterns, where colder temperatures prevail during January to April, resulting in the leftward skew in the distribution. The avg Temperature seems to increase over the next few months and remain moderately fluctuaing at higher temperatures upto the month of september. after which the temperature seems to decreases gradually.

Additionally, the presence of a normal distribution implies that temperature variations over the course of the year follow a relatively consistent pattern, with deviations from the mean temperature occurring within expected ranges.

The interactive nature of the plot adds an extra dimension to the analysis, allowing users to delve deeper into the data by interacting directly with the graph. With interactive features like zooming, panning, and hovering over data points, users can explore specific temperature values for each date more precisely.

CONCLUSION

In conclusion, the analysis of the crime23.csv and temp2023.csv datasets has provided valuable insights into policing dynamics and climate patterns in Colchester for the year 2023. Through various visualizations such as bar charts, scatter plots, maps, correlation matrices, histograms, and boxplots, significant trends and correlations within the data have been uncovered.

The distribution of crime categories revealed prevalent issues such as violent crime, shoplifting, anti-social behavior, and criminal damage/arson, underscoring the need for targeted strategies for crime prevention and enhancing public safety measures. Outcome statuses of reported crimes also shed light on the complexities in legal proceedings and law enforcement efforts.

The correlation matrix for weather variables provided insights into relationships between temperature, precipitation, wind speed, cloud cover, and sunlight duration, offering a deeper understanding of weather patterns and interactions among meteorological factors.

Overall, this analysis serves as a foundation for informed decision-making, resource allocation, policy interventions, and further research in areas such as crime prevention, climate resilience, and public safety strategies. Leveraging data-driven insights from such analyses can significantly contribute to enhancing community well-being and addressing societal challenges more effectively in Colchester and similar regions.

REFERENCES

[1] rbokeh: How To Create Interactive Plots In R. towardsdatascience 2022. [Online]. Available: https://towardsdatascience.com/rbokeh-how-to-create-interactive-plots-in-r-cf8fd528b3d5 (https://towardsdatascience.com/rbokeh-how-to-create-interactive-plots-in-r-cf8fd528b3d5)