An Empirical Review of Arrest Rates in the DMV area

Lizzie Healy
Rachna Rawalpally
Trey Roark
Sophia Rutman
Zoo Un Park

DSAN 5100
Georgetown University
December 10th, 2024

Table of Contents

Section I: Introduction

The Washington District of Columbia (D.C.) area is a dynamic social environment known for its historical, cultural, and political significance. However, like many metro urban areas, it faces difficulties related to public safety and law enforcement. It becomes increasingly difficult to measure these fluid concepts to provide ideas of reform and improvement. Arrest rates, for example, can serve as a critical metric to understand police activity, crime rates, and broader sociality within the region. They can also be a massive indication of public patterns when looking at the data through a sociological and behavioral perspective. Ultimately, the crimes being committed have this ulterior motive behind it, rather than the binary classification in the data, and thus, it's important to recognize the nuance of these arrests as well. By analyzing the trends and anomalies that we find within the District of Columbia, Maryland, and Virginia (DMV) area, we'll gain a better understanding of the challenges that are critical for maintaining public safety, ensuring equitable treatment, and fostering trust through transparency between law enforcement and the communities they serve.

The importance of understanding these trends is not limited to academia or policy makers. Our topic directly impacts the lives of residents and the effectiveness of the community-law enforcement relationship. Insights in criminal justice reform data can help uncover disparities, highlight systemic challenges, and provide actionable insights to preserve this relationship. Initial questions still linger: What types of offenses dominate arrest records, How do different jurisdictions within the DMV compare, or how have external shocks reshaped the landscape of law enforcement targeting? Through volatile relationships like these, it's required to ask these questions of our society, and build a reputable, empirical foundation to begin answering these questions.

In this report, we delve deep into statistical analyses of arrest rates across the D.C. area, with a keen focus on identifying patterns over time across various jurisdictions. Our analysis seeks to answer the question: *How have arrest rates in the DMV area changed over the past 10 years, and what key trends or patterns can be identified?* Addressing this question requires a multi-faceted approach, by taking a focus area approach when sectioning our research question into different parts. To do this, we split this research question into four key focus areas, each with their respective research questions. These topics can be mapped here:

- Temporal: *What trends and anomalies in adult arrest rates have emerged over the past decade? How do seasonal and annual variations impact arrest trends?*
- Categorical: *How do arrest categories (e.g., violent crime, drug-related offenses) differ in their distribution and frequency?*
- Geospatial: *How do police district boundaries influence arrest counts, and are there noticeable regional trends in enforcement or crime?*
- Generational: *What distinctions exist between juvenile and adult arrest rates, and do they reveal unique generational dynamics?*

To conduct this analysis, we utilized publicly available arrest data from local law enforcement agencies, applying statistical and geospatial methods to garner actionable and significant insights. These findings could help a plethora of stakeholders, ranging from policymakers and law enforcement officials to community leaders, understand the systemic challenges ongoing in the region more effectively. Policymakers can use the results to inform legislation regarding criminal justice, law enforcement officials can adapt their strategies, and community leaders can better advocate for their peers. Overall this report hopes to garner some empirical review of the activity that happens around us, and do so with complete transparency.

Through a thorough examination of these research questions, we hope to provide a better view of arrests rates and law enforcement activity in the D.C. region, adding another level of transparency to residents of the area. We hope to contribute to the ongoing efforts with data visibility, especially pertaining to social justice. By fostering these informed data-driven insights, our group aspires to help all levels of society, from residents to leaders to government officials to work together towards safer, more just communities. Ultimately, this report seeks to emphasize informed discussions and support the development of societal accountability that prioritize safety and equity in such a historically significant region.

Section II: Data Collection and Processing

Section II.I: Adult Data

The raw data was collected directly from the Metropolitan Police Department (MPD) of Washington D.C. The extensively documented site allows for public viewing of arrest reports by arrest category and for each calendar year, arrest summary statistics, interactive crime reports, and more. For the purpose of this paper, the MPD adult arrests by year reports were downloaded including 11 reports from 2013-2023.

With these particular reports, we were able to read in each CSV file and append them together utilizing Python's row-binding function. For the reports from 2013 to 2017, the row-binding was immediately successful, however, there was a slight discrepancy in the variable naming in all the reports beginning in 2018, thus, these names were rectified before row binding these sets. For note, there was no difference in the data and what is being reported; it was simply a change in the naming convention of a few variables. The data was otherwise well reported and maintained, thus, our only other data-cleaning task was standardizing the date variable. The arrest date variable contained a minor discrepancy in later years, with some entries formatted as MM/DD/YYYY. These were standardized to YYYY-MM-DD to ensure consistency for the time series analysis.

There are 26 variables included that provide information about the arrest itself, the arrestee, and the relevant location information. For the arrest information, the dataset includes year, date, and hour. In addition, it provides a Criminal Complaint Number (CNN), which is an anonymized number not unique to the arrestee, and an arrest number, which is also anonymized but is unique to each arrestee. The arrestee age variable is calculated by the date of the arrest subtracted by the arrestee's reported date of birth (DOB). The arrestee's race and ethnicity are included, which are denoted based on the arresting officer's observation with ethnicity being categorized as Hispanic, not Hispanic, or unknown. The arrestee's sex is also based on the observation of the arresting officer and includes male, female, and unknown. Further, the police service area of the home address of the arrestee is given in the defendant PSA variable as well as the district they preside in with the defendant district variable. The arrest category variable provides one of the 29 distinct types of arrests and the charge description further categorizes the arrest into one of 5,341 D.C. arrest code charges. The remaining 12 variables give a precise and in-depth idea about the location of the arrest. These include the arrest location police service area (PSA), the offense police service area (PSA), the police district where the arrest took place, and the police district where the offense occurred. The difference between the two occurs if the arrest took place at a separate location from where the alleged crime took place. The latitude and longitude of the arrest and the offense are also available. Finally, the dataset provided the variables on the location of arrest and offense utilizing the Maryland State Plane Coordinate System (SPSC) at the block level. The SPSC is a more precise mapping system that uses a Cartesian coordinate grid and maps the state of Maryland in detail.

The final dataset included a total of 275,416 observations with a unit of observation of individual arrest, where each row of the data is a reported arrest, not necessarily a crime or charge. The number of arrests peaked at 32,513, began to dip in 2020, and saw the lowest incidence in 2022 (Table 2.1). In Table 2.2, we see the majority of arrests over the 10-year reporting period involved individuals identified as Black, accounting for 86.14% of arrests. This was followed by individuals identified as White, who represented 9.09% of arrests. All other racial groups collectively made up less than 5% of total arrests. Individuals identified as male made up 77.39% percent of arrests, compared to individuals identified as female who made up 22.59% of arrests (Table 2.3). The average age of the arrested individual was 35, with a minimum of 18 (due to this report only including adult arrests) and a maximum age of 102. (Table 2.4). The five most common types of arrests are simple assault (61,440 incidents), traffic violations (32,447), release violations/fugitives (31,814), narcotics (24,957), and theft (24,940). The five least common arrest types are kidnapping (217), fraud and financial crimes: fraud (62), arson (47), fraud and financial crimes: forgery (4), and fraud and financial crimes: counterfeit (1).

(1) Year	(2) Arrests
2013	32,513
2014	32,313
2015	27,672
2016	29,980
2017	31,209
2018	29,115
2019	27,938
2020	18,491
2021	15,653
2022	14,991
2023	15,541

Table 2.1: Number of arrests each year.

Defendant Race	(1) Frequency	(2) Percentage
1. Asian	1,610	0.58
2. Black	237,253	86.14
3. Multiple	83	0.03
4. Other	559	0.20
5. Unknown	10,889	3.95
6. White	25,022	9.09

Table 2.2: Frequency and percentage of arrests by arrestee race (as reported by the arresting officer).

Defendant Sex	(1) Frequency	(2) Percentage
1. Female	62,224	22.59
2. Male	212,912	77.31
3. Unknown	280	0.10

Table 2.3: Frequency and percentage of arrests by arrestee sex (as reported by the arresting officer).

	(1) Minimum	(2) Maximum	(5) Average
1. Age	18	102	35.07

Table 2.4: Summary statistics for the arrestee.

Section II.II: Juvenile Data

The raw data pertaining to juvenile arrests was also collected directly from the Metropolitan Police Department (MPD) of Washington D.C. The data was not as comprehensive as that of the adult arrest CSVs, having only four features: date of arrest (titled ARREST_DATE), charge description (TOP_CHARGE_DESCRIPTION), home PSA of the arrestee (HOME_PSA), and the PSA where the arrest took place (CRIME_PSA). The MPD website provides biannual reports from January 2016 to June 2024 in CSV form, each containing these four features. Therefore, similar to the Adult arrest data, combining the CSVs into a singular dataframe was as simple as row-binding the data from each file in Python. Processing this data entailed converting the Arrest Dates into datetime objects in R.

In total, the final juvenile arrest datatable contained 15,123 arrests. 2016 demonstrated the highest number of arrests with 3278, and 2020 the lowest with 693 (Table 2.5). Simple assault accumulated the highest incidence of arrests with 2033 occurrences over the database,

followed by Absconder (Custody Order) with a similar rate of 2023. There were 275 categories of arrest over the eight year interval. In terms of Arrest Location PSA, PSA 602 has the highest incidents of arrests with 649, followed by PSA 703 with 578. The range of arrests over PSAs was from 649 in PSA 602 to 12 in PSA 201.

(1) Year	(2) Arrests
2016	3,278
2017	1420
2018	2720
2019	1434
2020	693
2021	741
2022	1682
2023	2188
2024	15,653
Undisclosed	19

Table 2.5: Number of arrests each year.

Section III: Time Series/Outlier Detection

Section III.I: Introduction

The dynamic process of arrest is inherently influenced by a multitude of indicators, including seasonal patterns, societal events, and many more outside factors. These arrest rates can fluctuate over time, driven by community behaviors, law enforcement priorities, and external disruptions, such as a public health crisis or protest. A Time series analysis permits us to explore these patterns in more depth, identifying significant trends and anomalies that reflect the underlying dynamics of crime and enforcement in the DMV area. This section (Section III) investigates the temporal trends in arrest counts over the past decade (2013-2023), and takes note of any existing outliers that correspond to key events, impacting the overall arrest count or law enforcement strategy. Through visualizations, statistical decomposition, and a two-sample t-test, Section III aims to identify both long-term and short-term trends.

Section III.II: Exploratory Data Analysis (EDA)

To gain a robust foundation of our research inquiry, we can look at the exploratory data analysis of daily arrest counts to reveal distinct temporal patterns and outliers from the data encompassing 2013 to 2024 statistics.

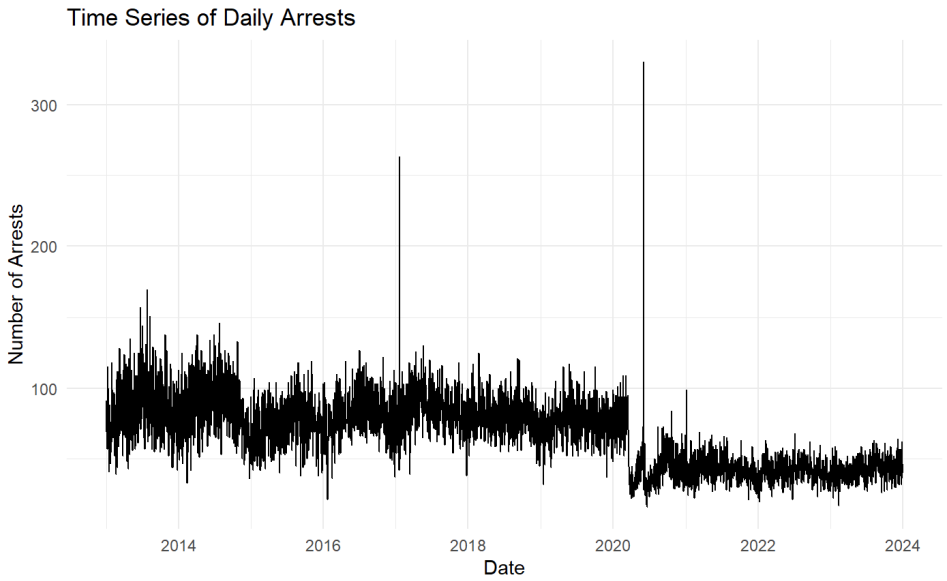


Figure 3.1: Time Series of Daily Arrests

The time series plot of daily arrests (Figure 3.1) highlights key changes in arrest rates over time. From 2013 to early 2023, the data show a seemingly stable and almost visually stationary pattern, with daily arrests averaging approximately 86 per day. As shown in the time series graph above, however, there's two massive spikes indicating outliers within the data. These noticeable peaks in arrest counts occur sporadically, with some days exceeding 250

arrests. There were two instances of this, once on January 20th, 2017, and the latter increase of involvement corresponds to June 1st, 2020. These anomalies are likely driven by special events or targeted law enforcement operations. When conducting further research into this question, there were two police reports that were written in response to a significant amount of arrests in the D.C. area for protests. The first on January 20th was the Metropolitan Police Department (MPD) responding to a "peaceful-turned-violent" protest from the 2017 Presidential Inauguration, evident from the police report made by the MPD themselves (Tobin 2018). The second outlier, conducted on June 1st, 2020 was directly correlated to crackdown on peaceful protesters in Lafayette Square from a George Floyd protest during that year. The Natural Resources Committee released a report directly correlating a significant increase in police activity and arrest rates during this day due to the law enforcement involvement in this protest (Grijalva 2023). These sources allow us to attribute most of the explanatory power in the outliers to these significant events, so if further analysis was needed for a time series model, then exogenous variables such as these protests could be made.

Furthermore, there was a significant decrease in arrest rates that continued from 2020 to 2024. Initial signs point to the onset of the COVID-19 pandemic in early 2020 marking the stark reduction in arrest counts, with daily averages dropping to approximately 46 arrests. However, there are many possibilities to the exact cause of this idea. We'll be conducting significance tests to help support this point. The pre-COVID period also displays a slight downward trend starting in 2016, which could reflect changes in crime rates, decriminalization efforts, or shifts in policing strategies. These preliminary observations set the stage for a more detailed analysis of the factors driving these trends.

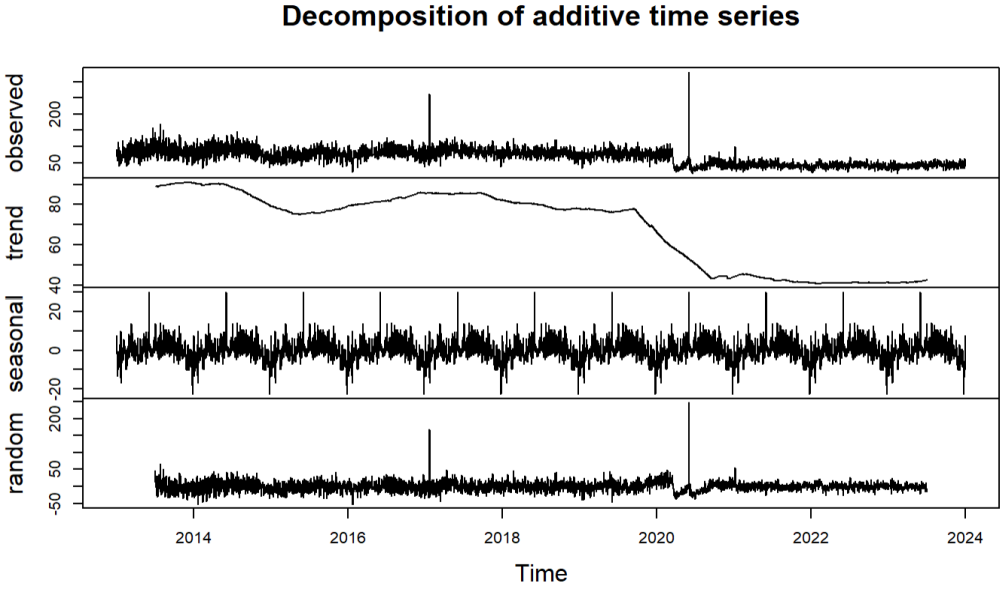


Figure 3.2: Decomposition of Additive Time Series

The decomposition of the time series into trend, seasonal, and residual components can be seen above in Figure 3.2. The trend component also illustrates a long-term decrease with a start difference happening in 2020 during the pandemic. The seasonal component captures recurring patterns, such as increased arrests during certain months or events. What we gain from that particular portion of the graph is the indication of a semi-seasonal trend when it comes to the arrest rates themselves. This particular question of seasonality will be further explored in Section IV.

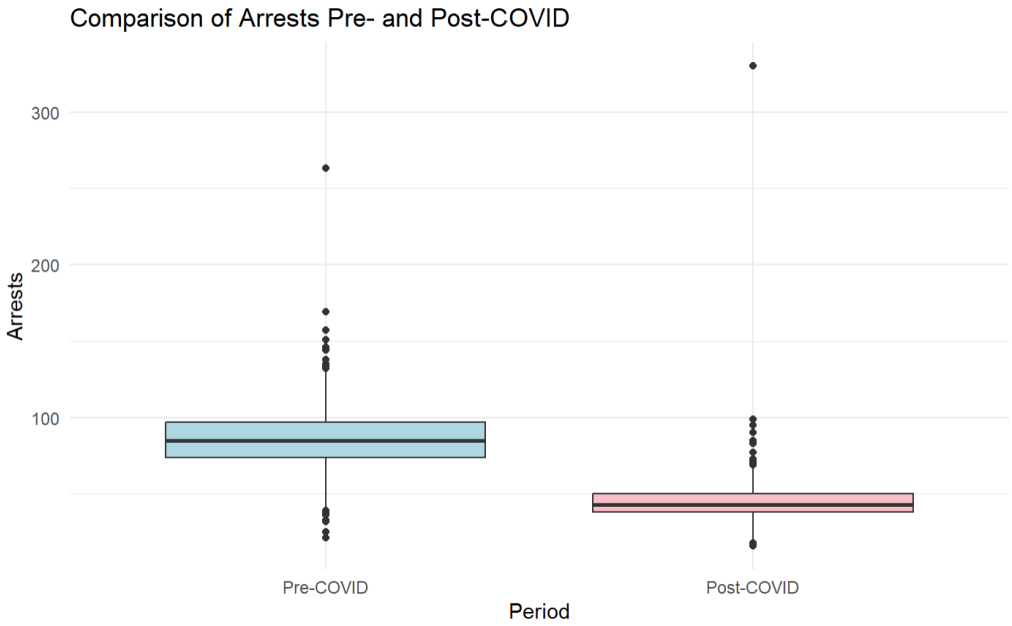


Figure 3.3: Comparison of Pre- and Post-COVID Arrest Distributions

Before conducting our t-test in the next section, we can visualize the arrest rate distributions directly using the side-by-side boxplot displayed in Figure 3.3. These boxplots compare the distribution of arrests specifically before and after the COVID-19 pandemic. The pre-COVID distribution is characterized by slightly higher variability, with a wider interquartile range and more frequent outliers representing exceptionally high arrest counts. In contrast, the post-COVID distribution shows a narrower spread, reflecting reduced variability in daily arrests. The median arrests post-COVID are significantly lower, underscoring the pandemic's impact on policing and public behavior. Looking into further research on this question, we can also find reports such as Hou et al. that correlate COVID-19 and its direct impact on arrest rates within major cities like Washington D.C., adding more evidence to support our claim to the inquiry (Hou et al. 2022). This EDA is promising for statistically significant results, as the data does support the claim that the average arrest rates can be initially attributed to COVID-19 given its prevalence in society around the time it emerged.

Section III.III: Methodologies

For our statistical significance test to gain insight into the impacts of the COVID period on arrest rates, we separated the data into two subsets like mentioned above in Section III.II. Arrest data were segmented into pre- and post-COVID periods (before and after March 11th, 2020, or when Washington D.C. was declaring emergency lockdown). The Welch Two-Sample t-test with a significance level of 0.5 was used to determine whether the mean daily arrests differed significantly between these periods. The t-test is a statistical method used to determine whether there is a significant difference between the means of two independent groups. It accounts for differences in group variances and sample sizes, making it suitable for this analysis. The null hypothesis (H_0) assumed no difference in mean daily arrests between the two periods, while the alternative hypothesis (H_A) posited a significant difference. The test statistics, *t*, calculated during this test can be seen in Appendix E. As a result, this analysis provided will function as a statistical basis for understanding the pandemic's impact on law enforcement activity.

Section III.IV: Results/Findings

Firstly, it was found through literature review and empirical research that specific days with arrest counts exceeding 250 arrests were identified as outliers and attributed to major public protests in the D.C. region. This significant increase in arrests was tracked through different perspectives and continued to contribute to some of the other spikes within our data.

Next, The comparative analysis of pre- and post-COVID data revealed a statistically significant reduction in daily arrests. The Welch Two-Sample t-test confirmed this finding, with a mean difference of approximately 40 arrests per day, and a p-value less than .05 ($p < 0.001$). It showed us a pre-COVID mean of approximately 87 arrests per day, with post-COVID's amount dropping to approximately 47 arrests per day. Our t-statistic of 383.03 indicated a statistically significant p-value, allowing us to reject the null hypothesis, confirming a significant reduction in daily arrests after the onset of the pandemic. This result reflects both behavioral changes (e.g., reduced public activity during lockdowns) and deliberate shifts in law enforcement practices (e.g., fewer arrests for non-violent offenses). Although we can't necessarily identify the specific portions of COVID-19 that affected arrests directly, we do now know that some of these arrest trends were made in initial reduction due to the pandemic itself.

Section III.V: Discussion/Conclusion

The findings from the time series analysis underscore this dynamic nature of arrest trends in the DMV area, and our results had several key themes emerge.

Firstly, the decline in arrest rates in 2020 reflect systemic changes in how law enforcement agencies can approach public safety. These changes could include the decriminalization of certain offenses, community-based policing strategies, or reductions in crime rates. The sharp drop in arrests post-COVID further highlights the influence of external

events on policing practices and how external factors such as pandemics can inflict changes in all social spheres of society.

Secondly, looking at the decomposition of our time series data, seasonal fluctuations in arrest rates demonstrate the importance of considering the contextual factors when analyzing this kind of data. For instance, further research should be conducted in how these frequent spikes during holidays coordinate with different granularities of time, which will be done in the following section.

Third, the pandemic represents a clear inflection point in our time series data, with a dramatic and later sustained reduction in arrests around the D.C. area. This result highlights the adaptability of law enforcement during crises and emphasizes the potential for alternative approaches to public safety that don't solely rely on high arrest rates as an indication of police involvement.

Additionally, the identification of extreme outliers offered us an opportunity for a deeper investigation into specific events that drive short-term spikes in arrest counts. These findings indicate a significant increase in arrest rates can be attributed to protest frequency. Significant social events such as this confirm a change in police activity. By examining these anomalies directly, we better understood the interaction between public events, law enforcement responses, and community dynamics.

Lastly, the results of the Welch t-test provided solid statistical evidence of significant changes in arrest practices post-COVID in D.C. These findings have important implications for our stakeholders noted in the introduction. This finding indicates the need for adaptive, data-driven approaches to public safety that consider both long-term trends and the impact of external disruptions.

The results of this section contribute to the ongoing discussions of equity, accountability, and the role of law enforcement over time. Future research can build on these results by exploring the causal factors behind observed trends and evaluating the broader societal impacts of changes in arrest practices. However, we can continue with our analysis on our other research questions.

Section IV: Trend Analysis of Seasonality

Section IV.I: Introduction

As discussed in Figure 3.1, the time series decomposition reveals a distinct seasonality trend. This section will further investigate the temporal trends in the arrest data, with a particular focus on the factors driving these seasonal variations.

Section IV.II: Exploratory Data Analysis (EDA)

Initially, we categorize the arrests over the seasons of the year: winter, spring, summer, and fall based on the arrest date given for each observation. Figure 4.1 depicts the frequency counts of crime for each of the four seasonal bins. Visually, it appears the arrests are at their peak in the summer months and dip significantly in the winter months, with spring and fall sitting at 2nd and 3rd highest incidence, respectively. Next, the arrests are binned into the 12 months of the year. In Figure 4.2, we see that each month falls somewhere between 20,000 and 25,000 arrests reported with July and August having the highest arrest frequency. November, December, and February have the smallest arrest frequencies according to the visual plot. Following this, the observations are split by the day of the week they occur, again leveraging the arrest date for this process. Visually, Figure 4.3 shows an upward trend in arrests toward the end of the week, peaking on Friday, with the lowest occurrence on Monday. Figure 4.4 splits the arrests by the weekdays and weekends (Friday through Sunday, inclusive), which shows that arrests are much more prevalent on the weekdays.

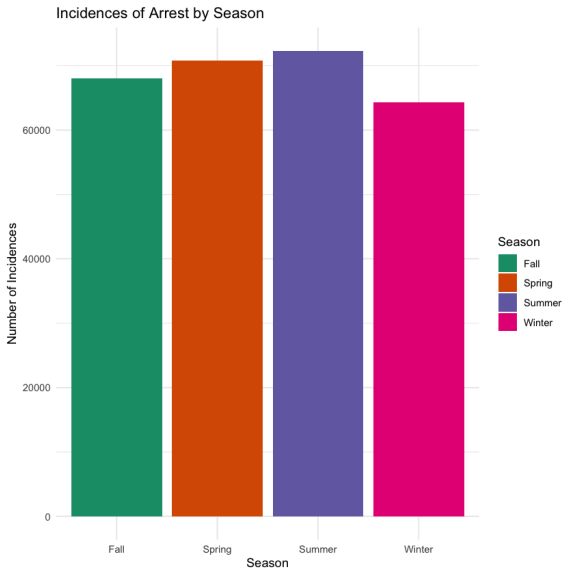


Figure 4.1: Arrest Counts by Season.

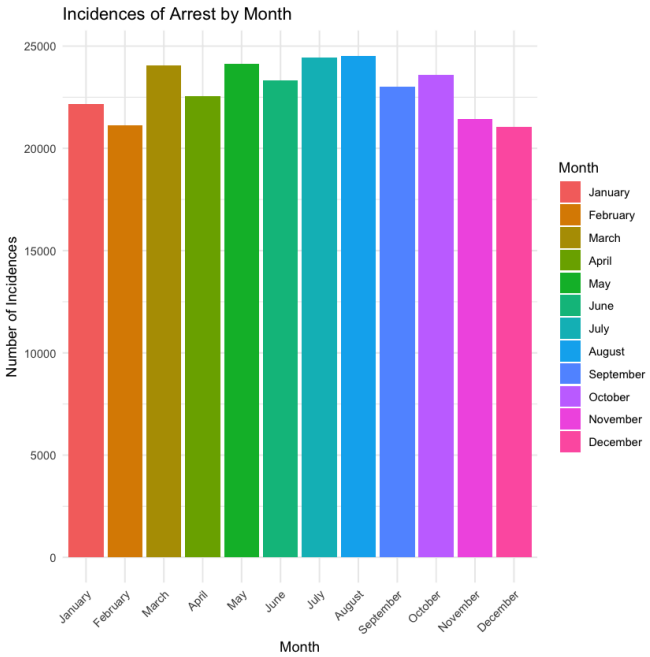


Figure 4.2: Arrest Counts by Month.

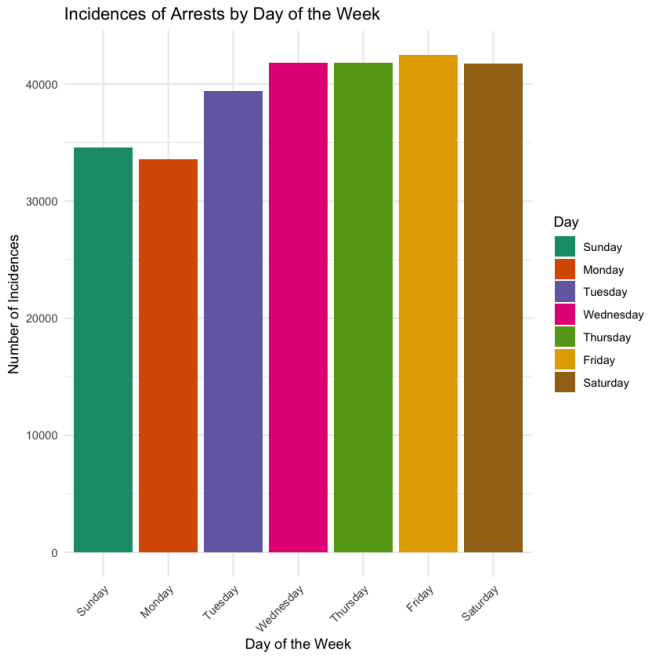


Figure 4.3: Arrest Counts by Days of the Week.

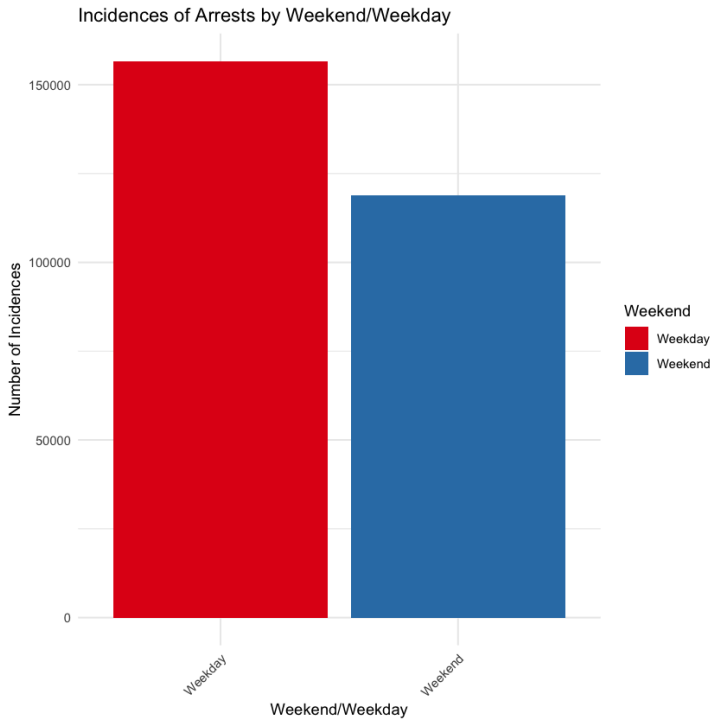


Figure 4.4: Arrest Counts on the Weekdays versus Weekend (Fri-Sun).

Section IV.III: Methodologies

In order to statistically verify the observations made through exploratory data analysis and to achieve the goal of investigating arrest seasonality, two hypothesis tests were utilized for each of the following categories: seasons, months, days of the week, and weekend/weekday split. The first test was a chi-squared test, which was used to statistically test whether the arrest counts were evenly distributed across the aforementioned groupings. The null hypothesis in each case was that crime counts were evenly distributed across the seasons, months, days, or weekends/weekdays, depending on the context. Following this, a pairwise proportion test with Bonferroni p-value adjustment was applied to determine which groupings were driving the results of the chi-square test, if a significant result was found. The null hypothesis in each case was that there was no difference in the arrest proportion for each season, month, day, and weekend/weekday. A two-sample t-test was also used to test for statistical significance between the weekend and weekday arrest rates (see Appendix for t-statistic calculation). The null hypothesis in this case was that the true difference in means between the Weekdays and the Weekends is equal to zero.

Section IV.IV: Results/Findings

When looking at the hypothesis testing for seasonality, we found a p-value of 2.2e-16 and, thus, we can reject the null hypothesis that the arrest counts are evenly distributed across the seasons, at the 5% significance level. In addition, we can reject the null hypothesis that there is

no difference in the arrest proportion between seasons due to a p-value of less than 0.05 for each seasonal pairing in the pairwise testing (see Appendix A.1).

For the monthly testing, the chi-square test resulted in a p-value of 2.2e-16, meaning we can reject the null hypothesis that the arrest counts are evenly distributed across the months. When looking at the pairwise results we cannot reject the null hypothesis that there is no difference in the arrest proportion between the months for the following monthly pairings: April and January, April and September, March and May, August and July, August and March, August and May, December and February, December and November, February and November, July and March, July and May, June and October, June and September, March and May, and March and October. All other monthly pairings had a p-value of less than 0.05 (see Appendix A.2).

The chi-square tests for days of the week yielded a p-value of 2.2e-16, thus we can reject the null hypothesis that the arrest counts are evenly distributed across the days. The pairwise testing results allowed us to reject the null hypothesis that there was no difference in the arrest proportions between days for all pairing except Sunday and Monday, Wednesday and Thursday, Wednesday and Friday, Wednesday and Saturday, Thursday and Friday, Thursday and Saturday, and Friday and Saturday (see Appendix A.3).

The p-value for the weekday versus weekend chi-square testing was 2.2e-16, meaning we can reject the null hypothesis that the arrest counts are evenly distributed across the weekend and weekdays. The t-statistic from the t-test was 0.8905, meaning we cannot reject the null hypothesis that the true difference in means between weekdays and weekends is equal to zero.

Section IV.V: Discussion/Conclusion

With the hypothesis testing results, we can conclude that arrests are variable based on the season of the year, are most prevalent in the summer, and are least prevalent in the winter. Further, arrest counts do differ across the months of the year for certain monthly pairings. July and August have significantly higher arrest rates than all other months March and May. December and February have a lower incidence of arrest rates than all other months except November. These results back up the seasonal findings, showing the arrest rates in summer are particularly driven by July and August. We cannot fully conclude that Friday has a significantly higher arrest rate than all other days, however, it is significantly more than the early weekdays (Sunday, Monday, and Tuesday). This trend is true for Thursday and Saturday as well, indicating a significant upward trend in arrests towards the latter part of the week. We can conclude that Monday has a lower incidence of arrests than all other days of the week except for Sunday, again supporting the upward trend in arrests. In terms of whether the weekdays or weekends see more arrests being made, we have conflicting results. On one hand, the chi-square test leads us to believe that weekdays have a significantly higher amount of arrests, however, the t-test results lead us to believe there is no significant difference in the mean number of arrests. Based on the difference in how these two tests are constructed, we can reasonably conclude that the significant result in the first instance is driven simply by the fact that there are more days in the week than

the weekend and not a true difference. Thus, while there is an upward trend in arrests throughout the week there is not a significantly higher number of arrests on the weekends or weekdays.

Section V: Arrest Categorization

Section V.I: Introduction

Much like the time and seasonality analysis, the reasons for arrest and categories of crimes provide valuable insights into the nature of law enforcement activity and societal behaviors. Understanding the distribution and trends across categories such as theft, assault, and traffic violations helps identify shifts in enforcement priorities and societal trends in criminal activity. However, this is *not* a holistic view at crime rates, as the arrest rates made for certain crimes and the crime rates of a particular area are not interchangeable or synonymous (Ghandnoosh & Budd 2024). Comparing these categories can give a semblance of areas with the highest need of resources, which enables policymakers and law enforcement officials to analyze these areas of interest. It also acts as an accountability mechanism, to view some of the most popular arrest categorizations and a breakdown of the makeup of those arrests. This section focuses on analyzing the monthly trends and variations in arrest counts across major categories over the past decade.

Section V.II: Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) of this section focused on identifying trends and differences across arrest categories.

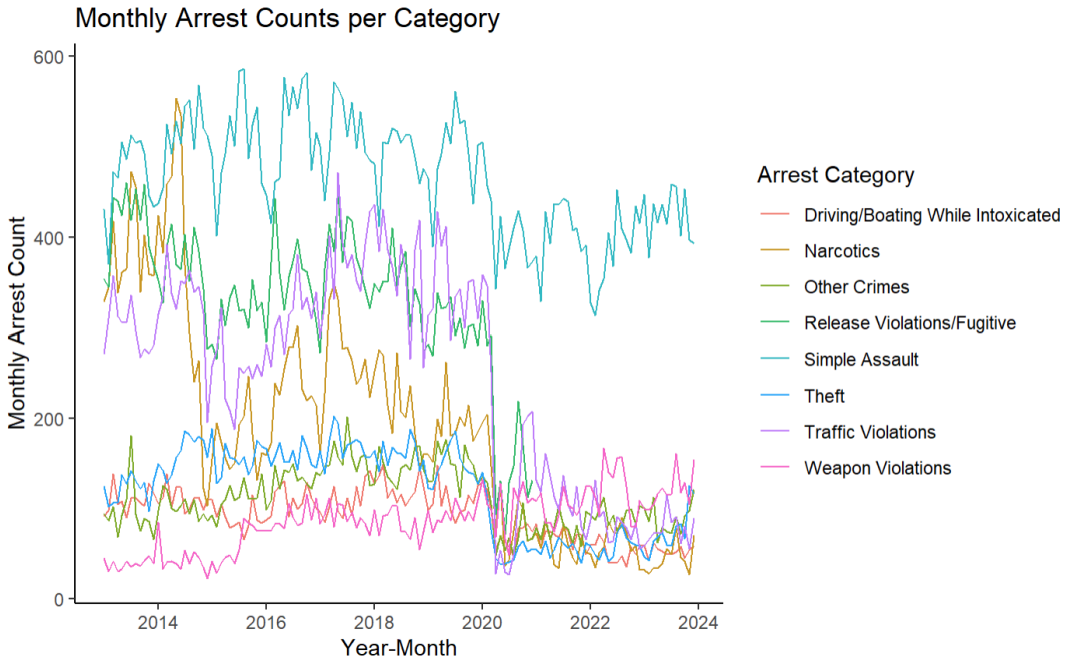


Figure 5.1: Time Series of Monthly Arrest Counts by Arrest Category

Above in Figure 5.1 we made a time series graph sub sectioned by the specific arrest category to get a glimpse at the monthly arrest counts in each of the major crime categories from

2013 to 2024. We see the top 7 arrest categories with the others that don't show as much frequency categorized as "Other Crimes" to preserve readability. There were 35 different arrest categories in total, but we looked at those most frequent in the dataset. As seen in the graph, narcotics-related arrests dominated the early part of the decade, peaking in 2014 before experiencing a steady decline. Theft and traffic violations also showed significant activity, with notable peaks. Categories such as simple assault and release violations displayed more consistent trends over time, while some categories like weapon violations exhibited sharp fluctuations likely tied to specific enforcement actions or policy changes. The pandemic period (indicated by March 2020 onward) is marked by a universal decline across all categories, reflecting the broader drop in arrest rates during that time. Although, one of the crimes that didn't necessarily take as much of a hit when it came to the differences between pre and post-COVID was simple assaults. Looking at the graph, it seems that simple assault arrests were the only category not to suffer a massive reduction in monthly arrest counts. This could be due to COVID-19 potentially limiting resources, decrease in major criminal activity, or even decriminalization. Either way, this provides an empirical review of these specific crimes overtime in the area.

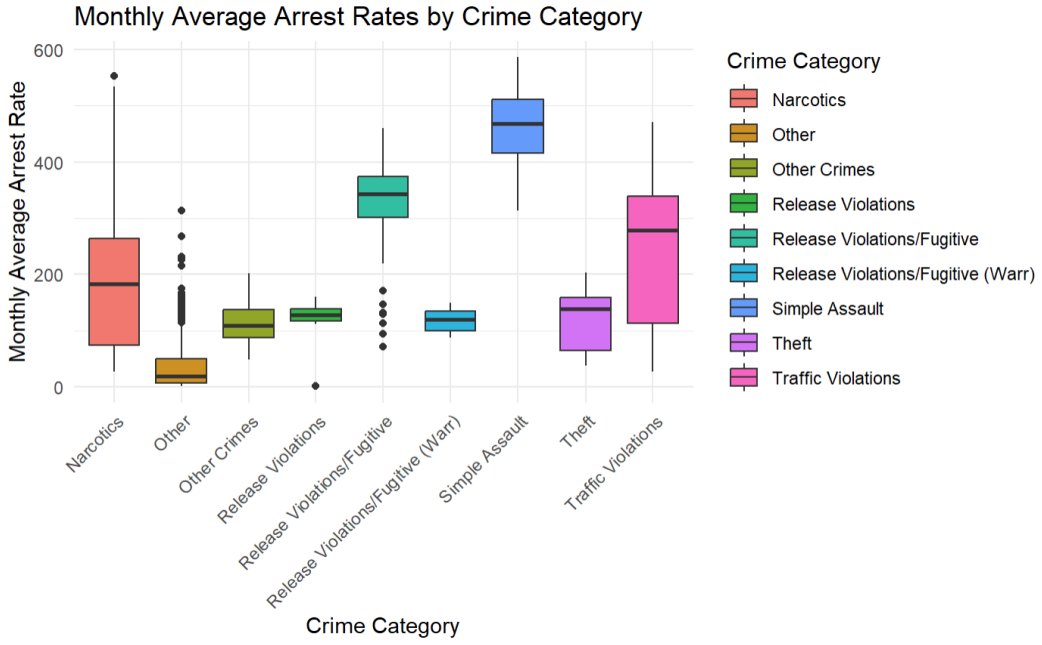


Figure 5.2: Monthly Average Arrests by Crime

The side-by-side boxplot in Figure 5.2 highlights the distribution of monthly average arrest rates for each category, allowing us to compare them directly. Looking at the above figure, we see that arrests pertaining to narcotics show the highest variability, with a wide range of monthly counts and numerous outliers. Theft and traffic violations also exhibit significant variability, while other categories, such as release violations and simple assault, demonstrate relatively stable distributions. The clear differences in variability and median values across categories underscore the distinct nature of enforcement and societal factors affecting each crime

type. As seen similarly to our time series portion of EDA, we see that simple assault arrests have a significantly higher distribution than the rest of our other categories when looking at the monthly average.

Section V.III: Methodologies

To analyze the distribution and trends across arrest categories, we'll utilize an ANOVA test. This ANOVA test will help determine whether the mean monthly arrest counts across our categories differed significantly. The null hypothesis (H_0) posited no significant differences between categories, while the alternative hypothesis (H_A) suggested that at least one category has a significantly different mean. The F-statistic splits the total variation into two components, the between-group variability (or Mean Square of Treatment) and the within-group (or Mean Square Error) variability. The ratio of these two components is computed as the F-statistic. These F-statistic equations and how they're calculated can be seen in the Appendix. After conducting the test, we achieved an F-statistic of approximately 1914 and a p-value of approximately zero, which strongly suggests rejecting the null hypothesis, indicating the difference in means monthly arrest counts across categories is statistically significant.

After performing an ANOVA test, we'll use a post-hoc Tukey test to identify specific pairs of categories with the most significant differences, allowing us to view the interactions between these categories. As a general overview, this Tukey test calculates pairwise differences and compares them to a critical value. It accounts for multiple comparisons that uses a studentized range statistic, q, to control for the increased risk of Type I errors due to the increase in tests. This q-statistic equation can be found in the Appendix. Next, we'll conduct these tests and analyze their results.

Section V.IV: Results/Findings

After gaining insight into the distributions of crime categories and arrest rates having to do with such, we've garnered a lot of actionable insights.

Firstly, we concluded there are trends in several of the arrest categories through our EDA, and that shifts have significantly affected some of these categories, reflecting possible policy changes such as decriminalization or shifts in enforcement priorities. Some of these categories, on the other hand, like simple assault, stayed more stable over time, suggesting a consistent enforcement pattern for these offenses. More research into this needs to be conducted in order to come to a more specific conclusion. Through our visualizations, we saw the distributions of these arrest categories as well, which indicated an initial difference between the priorities of law enforcement activities and the crime rates that correlate with arrest rates in the city. However, these were semi-volatile, with some of the categories exhibiting higher variability than others with some outliers. Furthermore, the ANOVA test revealed statistically significant results confirming this average difference, meaning there are certain crimes that people in the DMV are arrested for more often than usual, such as simple arrests. Lastly, The Tukey Post-hoc test indicated significant pairwise differences between the arrest categories. For example, the

pairwise differences between a variety of the arrest categories and simple assault yielded the higher test statistics, indicating that simple assault is one of the most statistically significant categories compared to the other categories. Also, theft vs. simple assault showed a significantly higher difference and test statistic than most other categorical combinations as well. A table of the full list of pairwise differences can be found in the appendix. It also revealed that some of the categories don't have a ton of differences, showing similar distributions in their average arrest rates. For example, theft vs. release violations did not reveal a statistically significant pairwise difference, indicating their distribution of average arrest counts is very similar.

Section V.V: Discussion/Conclusion

The categorization analysis in this section allows for a glimpse into the nature of arrest activities across the DMV area. We saw similar trends to our overall time series and seasonality analyses when it comes to the specific arrest categories. Although, it doesn't necessarily line up with the trend of broader national efforts to reduce arrests for non-violent or simple offenses to focus on serious crimes, as our data indicated the only arrest that didn't take a massive reduction after COVID-19 was simple assault. Also, simple assault, theft, and traffic violations consistently ranked among the top categories in arrest counts. These seasonal spikes may reflect the social behaviors of the public, such as increased travel during the holidays or even higher theft rates during busy retail seasons. Furthermore, the significant results from our ANOVA and Tukey test confirmed that arrest counts vary significantly between these categories. This creates the foundation of importance for analyzing each category separately to understand the unique factors driving arrest trends.

Overall, by analyzing the arrest trend categories in our dataset, this section highlights the complexities in the interplay between societal patterns, enforcement trends, and even external factors. These findings can potentially inform resource allocation decisions, policy development, and steps to ultimately address systemic inequities in law enforcement practices. Future research regarding arrest categorization could also explore some of the more causal factors behind these trends, and take a more policy approach towards analyzing the trends of these specific arrest types.

Section VI: Geospatial Analysis

Section VI.I: Introduction

This section examines the geospatial characteristics of police districts. We analyze police districts rather than police wards due to the broader geographic coverage and more extensive data on police activity at the district level, which facilitates a more comprehensive analysis. The focus is identifying trends in arrest counts across the police districts within the Districts of Columbia, Maryland, and Virginia (DMV) region.

Section VI.II: Exploratory Data Analysis (EDA)

The DMV region is divided into seven police districts, and our analysis focuses on police arrest data spanning 2013 to 2023. A key variable in this dataset is the '*Offense. District'* column, which identifies the district where the alleged crime occurred. To gain an initial understanding of arrest rates, we generated a heat map to visualize the geographical distribution of arrests. The heat map is centered around the mean latitude and longitude of the reported crimes, using the columns '*Offense. latitude'* and '*Offense.longitude.'* This visualization highlights areas with varying levels of arrest activity. The heat map reveals notable patterns in the DMV region. District 1 stands out with a significant red hotspot, indicating a high concentration of arrests in this area. This observation raises questions about the factors contributing to this elevated activity. In contrast, the northwest region of the map predominantly shows green areas, signifying lower arrest rates. However, some smaller hotspots, such as around Foggy Bottom, are also evident. Figure 6.2 presents a bar plot summarizing total arrest counts across the seven districts from 2013 to 2023. The data reveals that Districts 1, 6, and 7 have the highest arrest counts, while District 2 has the lowest. Building on this, Figure 6.3 provides a stacked bar plot that categorizes the top eight crimes by district. The results show District 1 has disproportionately high arrests for release violations and fugitives. Similarly, simple assault emerges as a prevalent offense in Districts 6 and 7, reflecting significant trends in those areas. These visualizations provide a foundation for deeper analysis into the distribution and nature of arrests across the DMV's police districts.

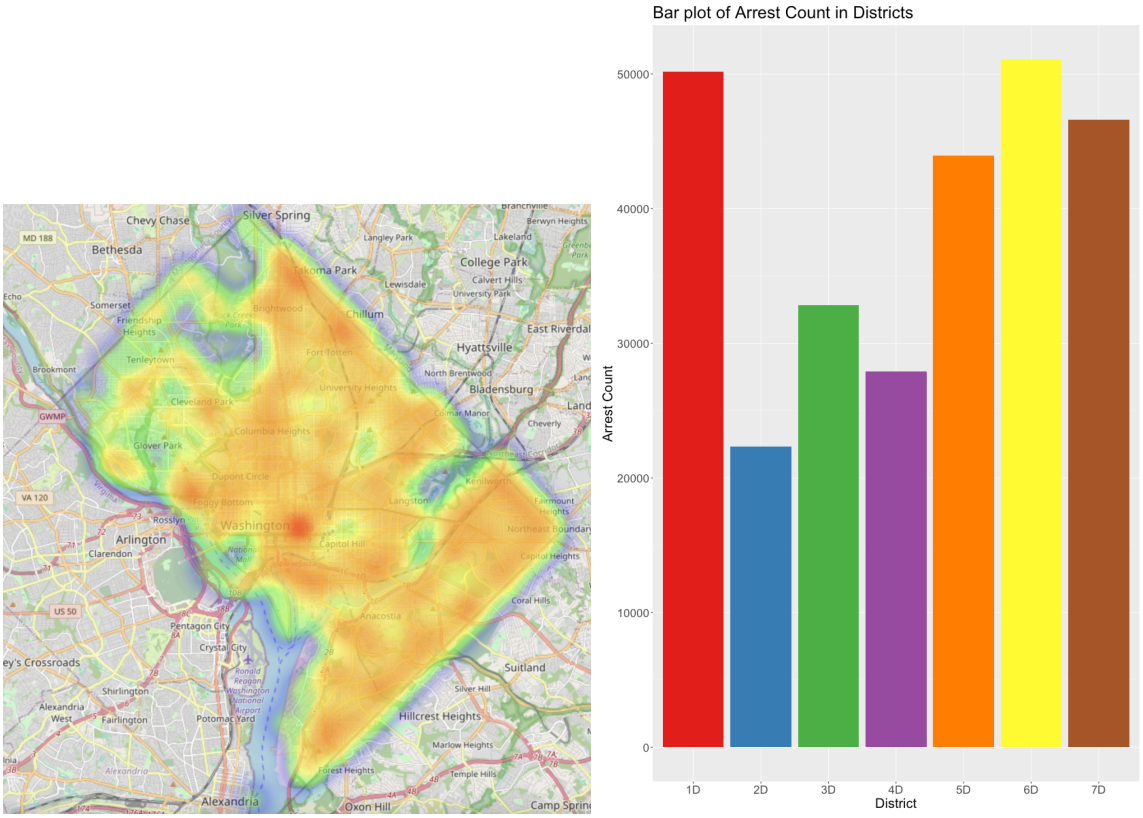


Figure 6.1 & 6.2: Heat Map of the DMV arrest counts & Bar Plot of Arrest in broken up by Districts

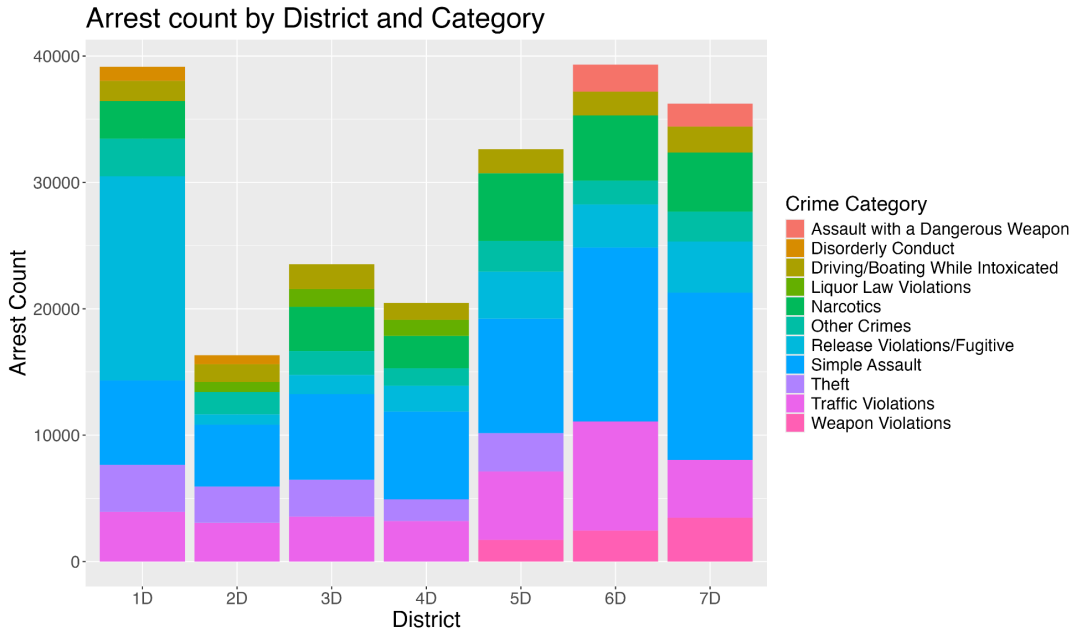


Figure 6.3: Stacked Bar Plot of Arrest Count broken up by Districts and Crime Category

Section VI.III: Methodologies

Based on the information provided by the heatmap, we first aim to test whether arrest counts differ across districts. To do this, we will perform an ANOVA test. The null hypothesis is that arrest counts are the same across all districts, while the alternative hypothesis is that there is a difference in arrest counts between at least one district. If we reject the null hypothesis, the ANOVA test will indicate a difference but will not specify which district or districts differ. We will conduct a Tukey's test to determine which specific districts differ, allowing pairwise comparisons between districts. Next, we will examine the types of crimes occurring in each district to determine if certain crimes are more prevalent in specific areas. Are certain crimes more likely to happen in a particular district? Are any crimes district-specific? To investigate this, we will perform a chi-square test. The null hypothesis is that crime category and district are independent, while the alternative hypothesis suggests that crime category and district are dependent. After performing the chi-square test, we will further analyze the results by examining the chi-square residuals to identify which crimes are most strongly associated with specific districts.

Section VI.IV Results/Findings

After running the ANOVA test in R, where we grouped total arrests by district, we obtained a very small p-value, leading us to reject the null hypothesis. This finding indicates that there are differences in arrest counts across districts. In the appendix is a plot of Tukey's test. To interpret this plot, if the horizontal line does not fall between the dotted zero lines, it indicates a significant difference between the two districts. The longer the difference in the means line, the more significant the difference. According to the plot, the following district pairs show significant differences in arrest counts: 1D-2D, 1D-3D, 1D-4D, 2D-5D, 2D-6D, 2D-7D, 3D-6D, 4D-6D, and 4D-7D. The most notable differences occur between District 2 and Districts 1, 5, 6, and 7. These results are also reflected in the bar plot, which shows that District 2 has the lowest arrest count, while Districts 1, 5, 6, and 7 have the highest. Next, we ran a chi-squared test, which returned a p-value of less than 2.2e-16. This extremely small p-value makes us reject the null hypothesis, indicating a dependent relationship between the crime category and district. In other words, certain crimes are significantly more likely to occur in specific districts. To dive deeper, we examined the chi-square residuals. The most significant positive residuals, indicating higher-than-expected arrest counts, were observed in the following districts and crime categories:

- 1D for release violations/fugitive
- 7D for simple assault
- 2D for theft

These results show that these crimes occur at significantly higher rates in these districts than in others, making these crimes district-specific hotspots.

Section VI:V: Discussion/Conclusion

This analysis highlights how arrest rates across districts, particularly for specific crimes, follow distinct patterns. With this information, law enforcement can develop more targeted interventions based on the crime categories prevalent in each district.

Let us revisit the initial questions we had when examining the heatmap. In District 1, we identified a hotspot of release violations/fugitive arrests, notably higher than in other districts. This hotspot was located near the courthouse, which makes sense, as individuals who violate bail or parole conditions are often arrested by judges or law enforcement in these areas. Conversely, we observed fewer arrests in District 2, which aligns with both the bar plots and Tukey's test, where District 2 significantly differed from Districts 1, 5, 6, and 7. District 2 is the wealthiest in Washington, D.C., encompassing Georgetown, Dupont Circle, Logan Circle, and Kalorama neighborhoods. Historically, wealthier neighborhoods tend to have lower crime rates. However, our chi-square residuals indicated that theft is a significant hotspot in District 2, which aligns with the district's socioeconomic makeup. When comparing District 2 (the wealthiest) with District 7 (the poorest), the arrest counts reveal a stark contrast. This suggests that various factors, including wealth and socioeconomic status, influence crime rates in these districts. The patterns seen in the heatmap and chi-square residuals underscore the complex relationship between crime and district characteristics. The story told by the arrest data in the DMV area is fascinating, and further analysis could explore how factors like socioeconomic status and race influence arrest rates. Understanding these dynamics could help shape more effective policing strategies across different neighborhoods.

Section VII: Juvenile vs. Adult Research Question

Section VII.I: Introduction

Many of the questions we ask here about geospatial data, seasonality, and trends over time can also be asked of the juvenile data. This section delves into the differences in adult and juvenile data over time, location, and arrest category, with the aim of discovering significantly significant trends in occurrence numbers and categorical splits. The analysis utilizes the adult arrest and juvenile arrest data described in Section II.

Section VII.II: Exploratory Data Analysis (EDA)

Meaningly differences emerge between the juvenile and adult arrests through completing simple aggregations over arrest date, arrest category, and arrest PSA of each dataset.

To begin however, we completed a decomposition, similar to the decomposition over the adult arrest DataFrame, to find overall trends in juvenile data over time. This graph demonstrates a dip after the primary COVID-19 years over the trend analysis, as seen in the adult arrest count over time. It occurs slightly earlier than that of adult arrests, and increases dramatically as the years approach 2024. By visually observing this trend, we can see that the arrests grow closer to pre-covid rates as time goes on. The graph also demonstrates a clear seasonality in arrests.

To further compare the adult and juvenile arrests since 2016, we compiled Figure 7.1, which shows the proportion of arrests completed each particular day. The combination of each day of data provides a distribution over time, allowing analysis of spikes in arrests mentioned in Section VI.V. By simply observing these graphs, we see that juvenile arrest numbers dip after 2020, but not to the extent of adult data.These graphs visually demonstrate significant differences over time between adult and juvenile arrests, which leads to the question: are these differences also statistically significant? We explore this question in subsection III.

We also made a bar chart showing the top ten arrest categories for juvenile and adult arrests, and their respective proportions (Figure 7.2). The top category for each is Simple Assault, and both include theft/robbery and assault with a dangerous weapon. However, they are not of the same prevalence in each data set, and the other seven top categories differ. The differences pictured show another major split in arrest data: reason for arrest.

Our final EDA was related to the arrest PSA. In other words, in which PSA are most adults/juveniles arrested? The distributions are shown in Figure 7.3. They demonstrate similar dips and spikes. The next steps for this figure, as for others, is discovering whether any of the above analyses produce significantly significant results.

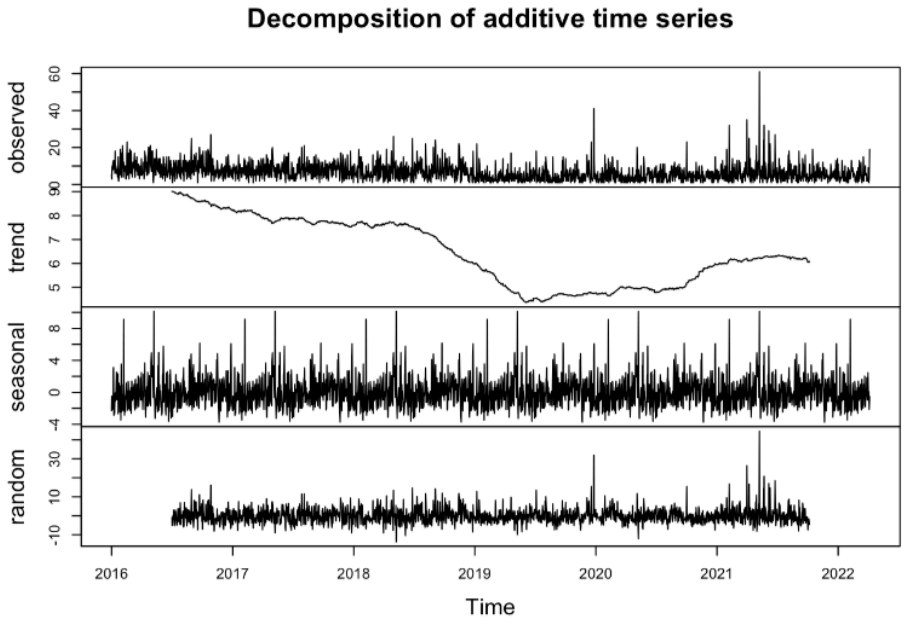


Figure 7.1: Decomposition of Additive Time Series

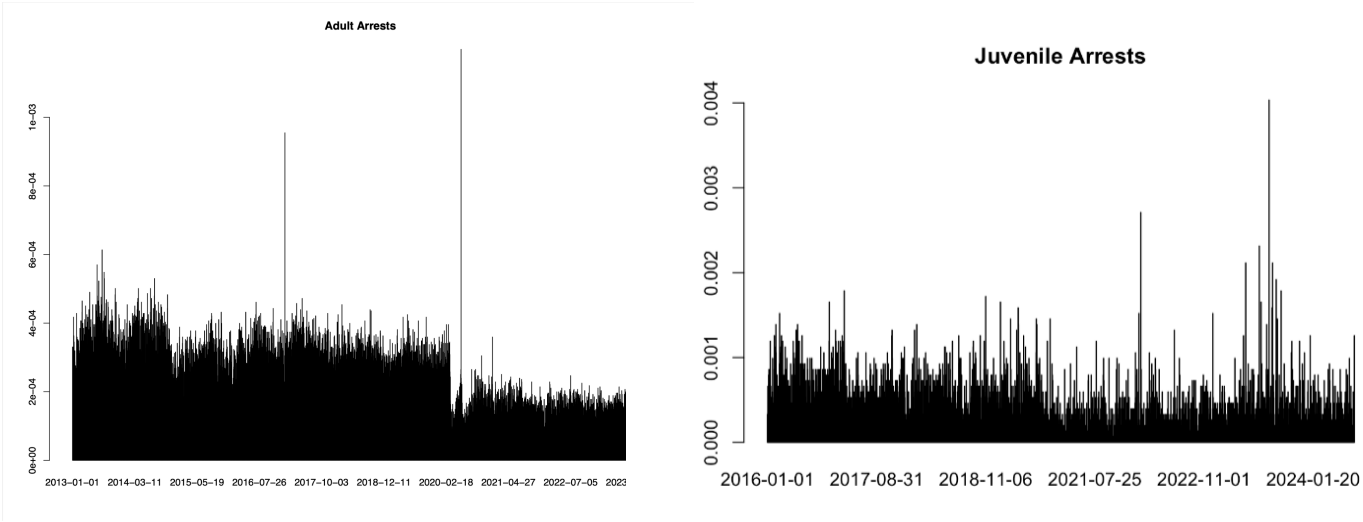


Figure 7.2: Distribution of Arrests over 2016-2024 for Adults and Juveniles

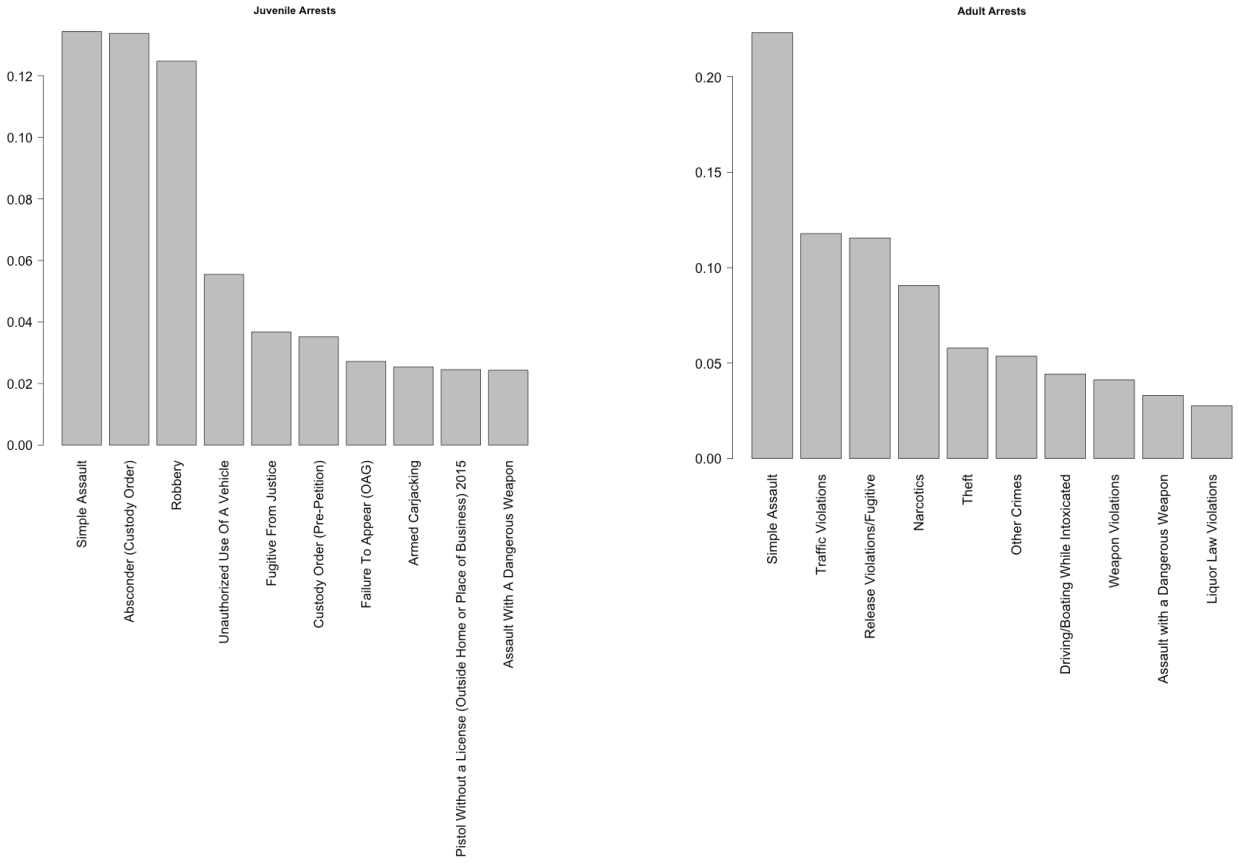


Figure 7.3: Distribution of Adult and Juvenile Arrests Across Categories

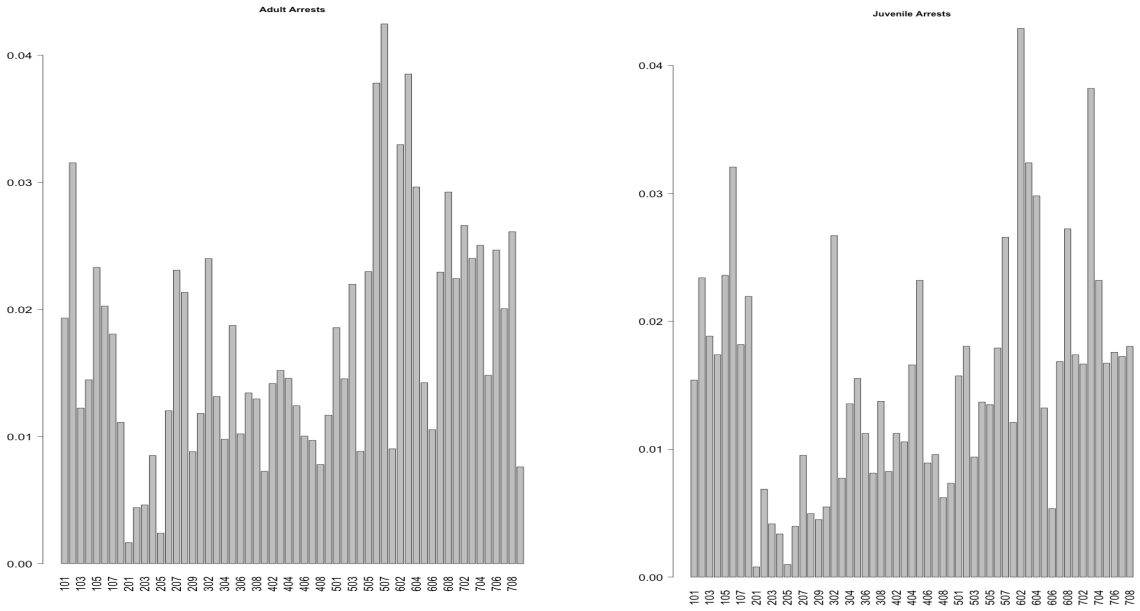


Figure 7.4: Distribution of Adult and Juvenile Arrests across PSAs

Section VII.III: Methodologies

The methodologies for Section VI include two Kolmogorov-Smirnov (KS) tests and a G-test. The first KS test is used to discover the truth behind the null hypothesis: both the Adult and Juvenile adult distributions over the eight years of data gathered are sampled from the same continuous distribution. The arrests normalized by the total number of arrests were used in the KS test in order to avoid a skewed result due to the great difference in number of arrests over time. Testing for statistical significance here tells us whether there is enough evidence to say that the arrest rates truly differ over time.

The G-test was used on the normalized proportion of arrests over arrest categories. We used a G-test as opposed to a Chi-Squared test because there were many small expected frequencies over various categories, as juvenile arrests contained over 200 more arrests categories than adult arrests. We wanted to see whether juvenile and adult arrest categories were independent or dependent variables, which is the result of the G-test.

Finally, another KS test was used to determine whether the arrests over PSA are similar enough to have been sampled from the same distribution. Just like the former two hypothesis tests, we used the normalized proportion of arrests in the test to account for the largely different number of arrests.

Section VII.IV Results/Findings

From these three hypothesis tests, we found two statistically significant results. The first relates to the KS test completed on the two arrest distributions over time. With a p-value of 2e-16 and D statistic of 0.59, we can reject the null hypothesis and claim that these two datasets are sampled from different continuous distributions. In other words, adults and juveniles were arrested at different rates over the span of 2016 to 2024.

The second statistically significant result was between the categories of arrest between age groups. The G-test yielded a G-statistic of 83058 and a p-value of 1. This means that we cannot reject the null hypothesis that the two categorical arrests are independent variables. We concluded from this test that adults and juveniles were arrested at different rates over different categories.

The test that did not provide statistically significant results was the final KS test inspecting the distribution of arrest proportions over PSA. The test yielded a p-value of 0.2238 and a D statistic of 0.18542. We cannot reject the null hypothesis that both adult and juvenile arrest proportions over arrest PSAs were sampled from the same continuous distribution. There is no statistically significant difference over the two distributions illustrated in Figure 7.4.

Section VII:V Discussion/Conclusion

There were a few anomalies in juvenile arrest rates over time, including two large spikes in arrests on 2022-02-10 and 2023-07-28. After combing through the MPD database and other

related literature, we could not find specific events that may have led to these elevated numbers. There were many youth-led abortion protests between 2022 and 2023, perhaps leading to these rates, but no information was released about mass arrests specifically on these dates as there was for spikes in the adult arrests mentioned in Section III.

A large consideration for juvenile arrests is how arrests impact their schooling. For a majority of youth in D.C., a large portion of their lives is spent in school buildings. Education is immensely important for youth well-being and development. Without implementing measures and targeted interventions for crimes that occur in specific school districts, students will not be able to learn in a physical and psychologically safe environment.

Although the arrest categories were largely different between juvenile and adult arrests, the category that remained the highest in proportion for both was Simple Assault. As shown in Figure 7.5, a large percentage of the arrests related to these events take place in schools. These assaults make for a dangerous school environment for youth in schools around the District of Columbia. A robust study of simple assault across school districts as future work can help inform schools and police about how to prevent these offenses.

In addition, as mentioned in Section V, many arrests across the DMV are concentrated in neighborhoods with certain characteristics, such as low socioeconomic status. School based arrests also follow this trend, with males and black students being most commonly arrested (Figures 7.6 AND 7.7). Further analysis of juvenile arrests and school based arrests can further combat any systemic racism experienced by black students in the D.C. area, and can provide targeted interventions for schools experiencing high amounts of arrests relating to all or certain categories.

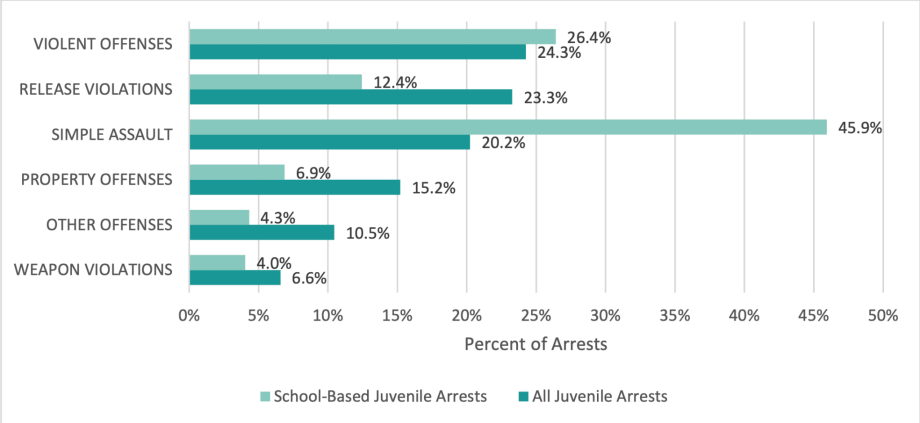


Figure 7.5: Distribution of Juvenile Arrests Across Broad Categories and In vs. Out of School

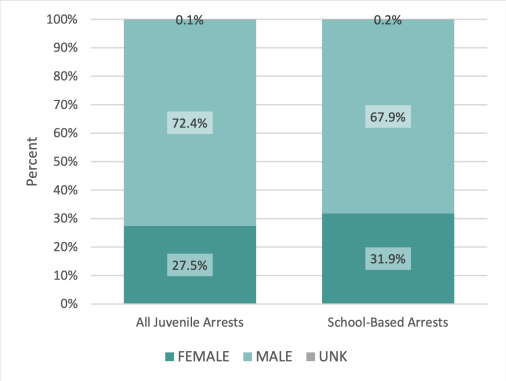


Figure 7.6: Juvenile Arrests by Gender and In/Out of School

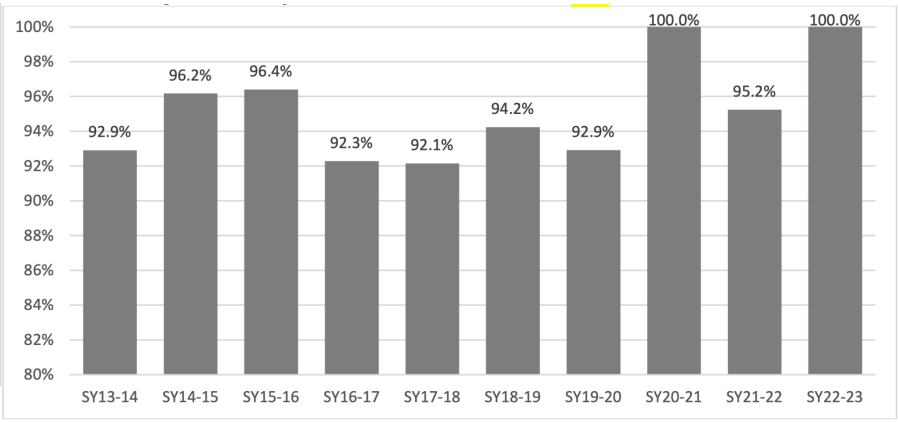


Figure 7.7: In School Juvenile Arrests by Percentage of Total Arrests that were of Black Students

Section VIII: Age Group Significance Tests and Transition Matrices

Section VIII.I: Introduction

This section aims to focus more on crimes committed by adults(age 18 or over). It has two objectives: first, observing whether crimes have increased by each age group in comparison with pre-COVID (defined before 2020) and post-COVID (after 2021, excluding 2020 and 2021 since these two years are a peak of COVID, not after COVID); and second, getting a transition matrix that is a probability that will lead to the next crime for each age group.

Section VIII.II: Exploratory Data Analysis (EDA)

Before analysis, the age group was divided into six age groups; 18-25,26-30, 31-35, 36-45,46-55,56 and more, considering the frequency of each age as inside the data young age, for example, less than 30 years old have more frequency than ones with more than 56 and over. Also, crimes are categorized into eight groups based on their relevance: family, sex-related crime, general law violation, release violation, simple assault, substance-related crime, theft crime, traffic violation, and violent crime.

Figure 8.1 below is a histogram showing the frequency of crimes by category considering age group and year from 2013 to 2023. Simple assault takes the largest portion for all age groups. Also, the 18-25 age groups have the largest frequency, meaning the number of crimes committed is higher in young people.

Figure 8.2 below is a histogram that compares pre-covid and post-COVID's proportion of crimes committed by age group. For the age group 18-25 and 46-55, pre-covid has a higher proportion compared to post covid, whereas in age groups, it is the opposite. Also, one thing to note is that for age 56 and over two periods seems to have very small differences compared to other age groups.

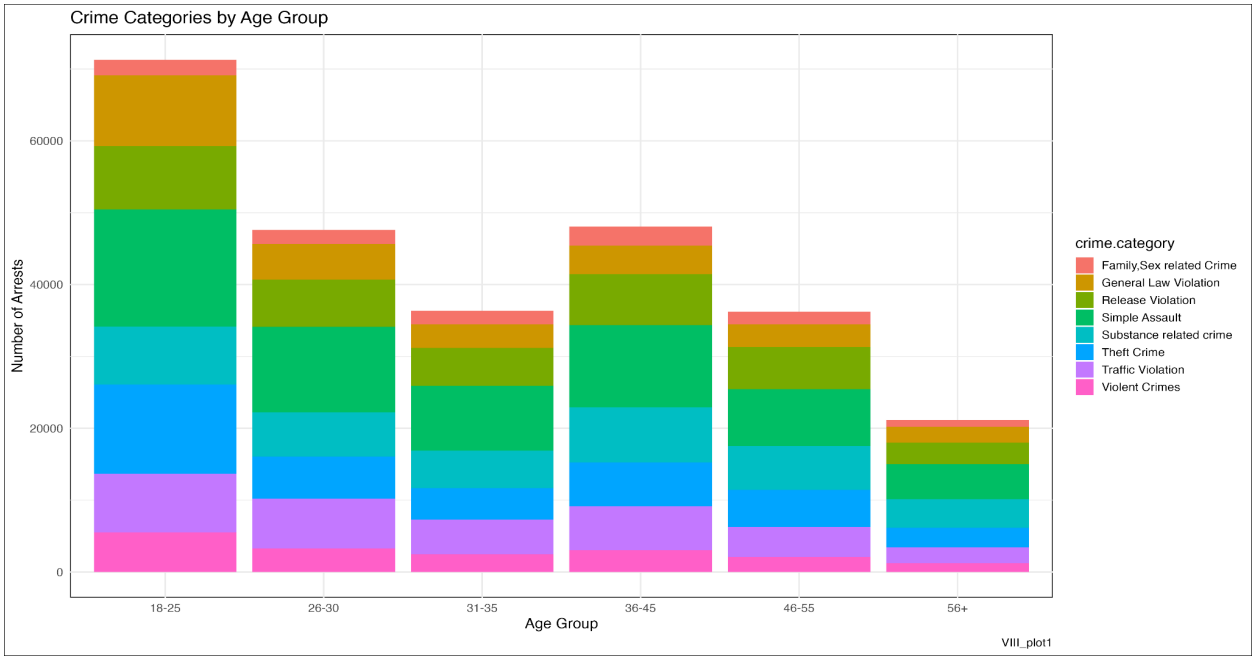


Figure 8.1: Crime Categories by Age Group

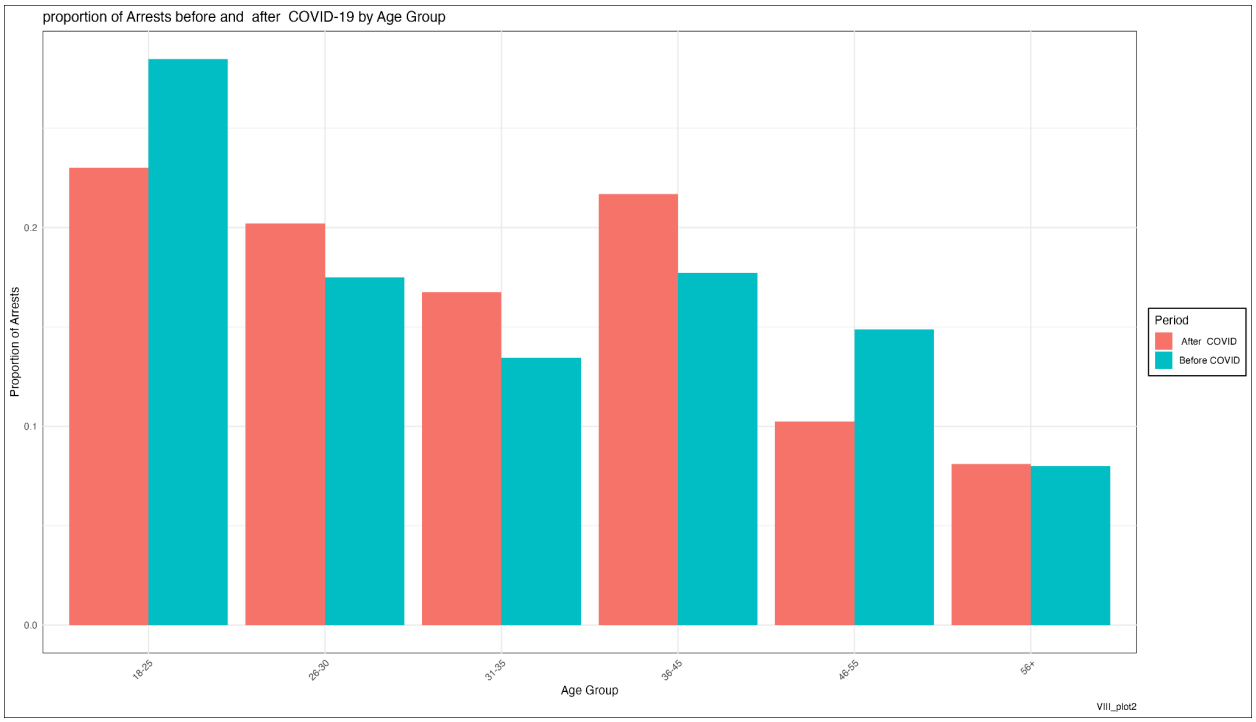


Figure 8.2: Proportion of Arrests before and after COVID-19 by Age Group

Section VIII.III: Methodologies

The methodology for section VIII includes bootstrapping and transition matrix. First, pre-covid and post-covid data are imbalanced data; there is a significant difference in the number of observations. So to get robust results, bootstrap sampling was used and with this, a hypothesis test was applied, with the null hypothesis, there is no significant difference in means for pre and post covid, and the alternative hypothesis: there is an increase in mean post-covid (crime has increased in post-covid). Second method is the transition matrix. The second method is the transition matrix. The purpose of using the transition matrix is to examine the transition probability between crime categories for each group. The question to solve by using this method is what will be the next crime, what is the probability that one crime category transitions to another crime category.

Section VIII.IV: Results/finding

Table 8.1: Bootstrap Sampling Results

Age_Group	*Observed_Difference*	*P_Value*	*Bias*
18-25	-0.07992012	0.5000	3.322196e-05
26-30	0.187424033	0.5060	1.798778e-04
31-35	-012175393	0.5126	3.924899e-04
36-45	0.01741627	0.5082	2.138117e-04
46-55	-0.17950426	0.4958	-2.870193e-04
56+	-0.82913793	0.4930	-2.714498e-04

Bootstrap was repeated 5000 times, with two data separated as before 2020 as pre-covid data, and after 2021 as pre-covid data. Table 8.1 is a result of bootstrap sampling with a hypothesis test. For all age groups, the p-value is larger than 0.05 significance level, meaning for all age groups, results fail to reject the null hypothesis, there is no significant evidence that crime has increased after COVID-19. This implies that after the DMV area went through COVID-19, for all age groups, there is less evidence to say that crime has increased. Also, for all the age groups, bias is very low, it can be interpreted that the accuracy for bootstrap is accurate.
The transition matrix is formulated with some rules. First, after a day, it is treated as a different person committing a crime. Second, if a crime has been committed by multiple criminals, then it is considered not the same case. Lastly, crimes are categorized as eight crimes since there are closely related crimes.

Figure 8.3 is a transition matrix for the age group 18-25 and the plots in the appendix are the matrix plots for All groups. Transition on Simple assault for all crime categories is observed as the highest transition probability among all the crime categories. This is due to the frequency of simple assault, as the histogram in EDA section VIII shows, simple assault has the highest frequency. Also as seen in the EDA section, the 18-25 age group has the highest frequency, so the 18-25 group's transition matrix is used first, to see how the transition matrix works and check overall trends for all groups.

There are some notable findings in the 18-25 age group transition matrix. The first transition probability from simple assault to simple assault is 0.32, which means that the probability of the next crime if committed after committing simple assault is 0.32. Also, the highest probability is 0.33, the transition probability to simple assault when a general law violation happened. This might assume a case when an arrestee argues with a police officer after a simple law violation and while resisting there was a simple assault.

For all other age groups, there is a similar pattern in the transition matrix. Except related to simple assault, there are interesting findings related to substance-related crimes. Substance-related crimes include drug usage and liquor-related violations, and in all age groups except age group 18-25, transition to release violation when someone commits this type of crime has a high probability. Also, the transition probability from substance-related crime to substance-related crime has a high probability for all age groups, which means that the probability of the next crime, if committed after a substance-related crime, has a high probability of violating another substance-related crime. From this finding, it is possible to infer that arrestees who are charged related substance-related crimes are quite related to release violations, which violate the terms of the condition for the arrestee's release by committing another crime. And committing other substance-related crimes probability is high in all age groups except 18-25, there is a possibility that violation is related to other substance-related crimes.

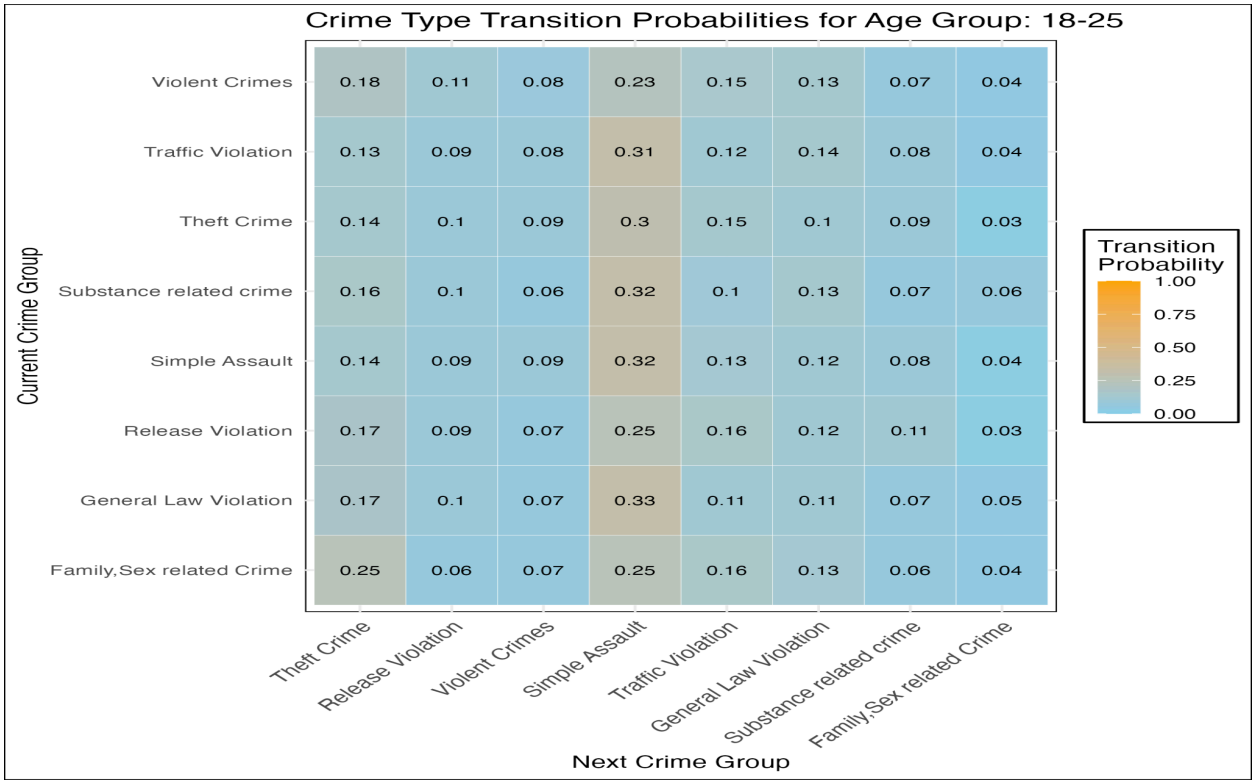


Figure 8.3: Transition Matrix of Crimes for Age Grouping 18-25

Section VIII.V: Discussion/Conclusion

Bootstrap test shows that in all age groups, there is no significant evidence that crime has increased post-covid. From the time series analysis, we know that there was a large increase in crime committed during COVID-19. However, after performing the test, it is hard to say that crime has steadily increased even after COVID-19, it can be interpreted as people tend to calm down after COVID-19.

The purpose of this test is to observe whether there was an increase in age groups between pre and post-COVID-19. Related to this topic, further tests can be suggested on the increase before and post-COVID in felony-level crimes to observe the difference in crime severity.

Markov chain transition matrix shows after the crime is committed, what is the probability that leads to other category crimes by each age group. There are many crime categories, but the highlighted part is on simple assault and substance-related problems. In all age groups, simple assault took over the highest probability for transition. This is due to the high frequency of simple assaults in all age groups. Also for substance-related crimes, there is a high probability that the next crime will be a substance-related crime and also a high probability of release violation. Narcotic issues are the ones that have high repetition chances and compared to assault or violent crime, it tends to be free in the short term. If considering that this can be interpreted as a danger of substance-related crime since it has a high chance of repeating the

crime. To extend this idea, checking by individual on what crime is likely to be committed can be conducted by the markov chain, by checking an individual's crime record.

Section IX: Conclusion

We explored many different facets of arrest data in the District of Columbia in this report, each having drastic impacts on daily life of inhabitants[1]. From discovering trends before and after COVID to creating transition matrices indicating which secondary crimes are most likely after an initial crime, this analysis has the potential to inform police and D.C. administrators on which areas and crimes need targeted interventions in order to support their citizens and prevent further growth of crime rates. Although a broad overview, the conclusions of this paper provide dozens of avenues for future work with the potential to produce reform and inform policy moving forward.

For example, observances of the arrest rates before and after COVID, and the policies that influenced this change, show a definitive shift in policing as a result of legislation. Further analysis is needed to discover whether this change was overall helpful or harmful to the community, but there is definitive change in arrest rates.

Looking into the future, how can we implement similar policies to enact positive change? We discovered that there was a statistically significant increase in crime rates over the summer and on weekends. Inspecting programs and outreach over these days and months can help to decrease these numbers and instances of crime that do not result in arrest. A similar example is the transition matrix of likely secondary crimes after an initial crime. What educational tools can D.C. create to inform previous arrestees and lower rates of repeat arrests? D.C. can tailor education tools towards primary offenses with knowledge of the most likely next offense based on historical data. A study conducted by Arizona State University states strategies such as Deterrence, where punishment escalates as offenders continue to commit crimes, and Informal Social Control, where the community around an offender produces negative sentiment on crime (Tilley 2024). By educating communities and creating policies that make progressively worse punishments as crimes increase, D.C. can reduce the differences in crime throughout the seasons and week, and prevent repeat offenses.

Other conclusions we drew from hypothesis tests demonstrated different levels of crime between adults, juveniles, and specific age-groups, as well as between different police wards in D.C. Finding these differences can inform targeted interventions for different demographics and areas. These interventions can positively influence schools, shared spaces such as parks and biking paths, and the overall physical and psychological safety of different areas in D.C. Implementing policy and community strategies to educate youth adults can take place in a multitude of settings to target those most at risk, such as schools, businesses, and, again, shared spaces. Cities such as Baltimore, Detroit, and San Antonio have invested in social services and interventions by local communities to lower their crime rates over the years since COVID. They have seen promising results using these methods, which can be used to inform reforms across at-risk areas and age-groups in the District of Columbia. Again, despite the broad nature of this report, the results can be used to create change in the lives of D.C. residents across the city.

[1] For reproducibility and validity, the code for our statistical tests and visualizations can be found here: https://github.com/sophiarutman/DC-Crime

Section X: References

Andone, Dakin, and Emma Tucker. "'This Is Not Luck. This Is a Systemic Approach': These Major US Cities Are Trying to Curb Violent Crime - and It's Working." *CNN*, Cable News Network, 29 Sept. 2024, www.cnn.com/2024/09/29/us/us-violent-crime-rates-down-dg/index.html#:~:text=With%20similar%20trends%20emerging%20across,intervention%20by%20trusted%20community%20members. Accessed 10 Dec. 2024.

Bowser, Muriel. "Stay Home DC." *Coronavirus*, 20 Mar. 2020, coronavirus.dc.gov/stayhome.

"G-TEST." *Wikipedia*, Wikimedia Foundation, 2 Feb. 2024, en.wikipedia.org/wiki/G-test#:~:text=for%20some%20cell%20case%20the,sense%20of%20Hodges%20and%20Lehmann. Accessed 03 Dec. 2024.

Ghandnoosh, Nazgol and Budd, Kristen M. . "Incarceration and Crime: A Weak Relationship." The Sentencing Project, 13 June 2024, www.sentencingproject.org/reports/incarceration-and-crime-a-weak-relationship/.

Grijalva, Raul M. "On Anniversary of June 1 Crackdown on Peaceful Protesters in Lafayette Square, Ranking Member Grijalva Releases Report with New Evidence on the Trump Administration's Involvement | the House Committee on Natural Resources." House.gov, June 2023, democrats-naturalresources.house.gov/media/press-releases/on-anniversary-of-june-1-crackdown-on-peaceful-protesters-in-lafayette-square-ranking-member-grijalva-releases-report-with-new-evidence-on-the-trump-administrations-involvement. Accessed 11 Dec. 2024.

Hou et al., Miaomiao. Investigating the impact of the COVID-19 pandemic on crime incidents number in different cities. Journal of Safety Science and Resilience. 2022 Dec;3(4):340–52. doi: 10.1016/j.jnlssr.2021.10.008. Epub 2022 Feb 17. PMCID: PMC8849845.

How. (2021, October 4). *Congress Heights on the Rise*. Congress Heights on the Rise. https://www.congressheightsontherise.com/blog/how-to-find-your-dc-police-district-and-police-service-area

Metropolitan Police Department. "Data and Statistics." *Mpdc*, mpdc.dc.gov/node/1379551. Accessed 20 Nov. 2024.

"Pairwise.t.Test: Pairwise T Tests." RDocumentation, www.rdocumentation.org/packages/stats/versions/3.6.2/topics/pairwise.t.test. Accessed 3 Dec. 2024.

Tilley, Nick. “Analyzing and Responding to Repeat Offending: Page 2.” *ASU Center for Problem-Oriented Policing*, Arizona State University, 2 Dec. 2022, popcenter.asu.edu/content/analyzing-and-responding-repeat-offending-page-2. Accessed 10 Dec. 2024.

Tobin, Michael B. “Police Foundation Releases Report on 2017 Presidential Inauguration.” *Police Foundation Releases Report on 2017 Presidential Inauguration*, 10 July 2018, policecomplaints.dc.gov/release/police-foundation-releases-report-2017-presidential-inauguration.

Yulia, et al. “Pairwise T-Test : Excellent Reference You Will Love.” Datanovia, www.datanovia.com/en/lessons/pairwise-t-test/#google_vignette. Accessed 3 Dec. 2024.

Appendix A: Pairwise Test Results (Tukey)

	1	2	2
2	2.7e-16	-	-
3	< 2e-16	1.0e-05	-
4	< 2e-16	< 2e-16	< 2e-16

Table A.1: Pairwise proportion test results for seasons.

	1	2	3	4	5	6	7	8	9	10	11
2	1.6e-05	-	-	-	-	-	-	-	-	-	-
3	< 2e-16	< 2e-16	-	-	-	-	-	-	-	-	-
4	1.0000	4.9e-11	2.2e-11	-	-	-	-	-	-	-	-
5	< 2e-16	< 2e-16	1.0000	2.6e-12	-	-	-	-	-	-	-
6	8.2e-07	< 2e-16	0.0215	0.0153	0.0070	-	-	-	-	-	-
7	< 2e-16	< 2e-16	1.0000	< 2e-16	1.0000	3.1e-06	-	-	-	-	-
8	< 2e-16	< 2e-16	1.0000	< 2e-16	1.0000	4.1e-07	1.0000	-	-	-	-
9	0.0011	< 2e-16	4.1e-05	1.0000	9.3e-06	1.0000	4.9e-10	4.0e-11	-	-	-
10	2.5e-10	< 2e-16	1.0000	5.4e-05	0.5713	1.0000	0.0017	0.0003	0.5574	-	-
11	0.2413	1.0000	< 2e-16	1.63-06	< 2e-16	< 2e-16	< 2e-16	< 2e-16	2.3e-13	< 2e-16	-
12	1.4e-06	1.0000	< 2e-16	1.8e-12	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	1.0000

Table A.2: Pairwise proportion test results for months.

	1	2	3	4	5	6
2	0.0083	-	-	-	-	-
3	< 2e-16	< 2e-16	-	-	-	-
4	< 2e-16	< 2e-16	< 2e-16	-	-	-
5	< 2e-16	< 2e-16	< 2e-16	1.0000	-	-

6	< 2e-16	< 2e-16	< 2e-16	0.3380	0.2893	-
7	< 2e-16	< 2e-16	< 2e-16	1.0000	1.0000	0.2101

Table A.3: Pairwise proportion test results for days of the week.

Tukey's Honest Significant Difference Test Results

	Mean Difference (Diff)	Lower Bound (Lwr)	Upper Bound (Upr)	Adjusted p-Value (p adj)
Other-Narcotics	-157.324556	-171.05552	-143.59359	0.0000000
Other Crimes-Narcotics	-77.242424	-96.22242	-58.26243	0.0000000
Release Violations-Narcotics	-75.496753	-118.83722	-32.15629	0.0000024
Release Violations/Fugitive-Narcotics	142.327652	121.64468	163.01062	0.0000000
Release Violations/Fugitive (Warr)-Narcotics	-71.443182	-105.65986	-37.22650	0.0000000
Simple Assault-Narcotics	276.386364	257.40637	295.36636	0.0000000
Theft-Narcotics	-68.310606	-87.29060	-49.33061	0.0000000
Traffic Violations-Narcotics	56.742424	37.76243	75.72242	0.0000000
Other Crimes-Other	80.082132	66.35117	93.81310	0.0000000
Release Violations-Other	81.827803	40.51564	123.13997	0.0000000
Release Violations/Fugitive-Other	299.652207	283.64957	315.65485	0.0000000
Release Violations/Fugitive (Warr)-Other	85.881374	54.27315	117.48960	0.0000000
Simple Assault-Other	433.710919	419.97996	447.44188	0.0000000
Theft-Other	89.013950	75.28299	102.74491	0.0000000
Traffic Violations-Other	214.066980	200.33602	227.79794	0.0000000
Release Violations-Other Crimes	1.745671	-41.59479	45.08614	1.0000000
Release Violations/Fugitive-Other Crimes	219.570076	198.88710	240.25305	0.0000000
Release Violations/Fugitive (Warr)-Other Crimes	5.799242	-28.41744	40.01592	0.9998552
Simple Assault-Other Crimes	353.628788	334.64879	372.60879	0.0000000
Theft-Other Crimes	8.931818	-10.04818	27.91182	0.8736452
Traffic Violations-Other Crimes	133.984848	115.00485	152.96485	0.0000000
Release Violations/Fugitive-Release Violations	217.824405	173.71158	261.93723	0.0000000
Release Violations/Fugitive (Warr)-Release Violations	4.053571	-47.80139	55.90853	0.9999996
Simple Assault-Release Violations	351.883117	308.54265	395.22358	0.0000000
Theft-Release Violations	7.186147	-36.15432	50.52661	0.9998774
Traffic Violations-Release Violations	132.239178	88.89871	175.57964	0.0000000
Release Violations/Fugitive (Warr)-Release Violations/Fugitive	-213.770833	-248.96069	-178.58098	0.0000000
Simple Assault-Release Violations/Fugitive	134.058712	113.37574	154.74169	0.0000000
Theft-Release Violations/Fugitive	-210.638258	-231.32123	-189.95528	0.0000000
Traffic Violations-Release Violations/Fugitive	-85.585227	-106.26820	-64.90225	0.0000000
Simple Assault-Release Violations/Fugitive (Warr)	347.829545	313.61287	382.04622	0.0000000
Theft-Release Violations/Fugitive (Warr)	3.132576	-31.08410	37.34925	0.9999988
Traffic Violations-Release Violations/Fugitive (Warr)	128.185606	93.96893	162.40228	0.0000000
Theft-Simple Assault	-344.696970	-363.67697	-325.71697	0.0000000
Traffic Violations-Simple Assault	-219.643939	-238.62394	-200.66394	0.0000000
Traffic Violations-Theft	125.053030	106.07303	144.03303	0.0000000

Table A.4: Tukey Post-Hoc Results - Full Pairwise Table

Appendix B: Geospatial Graphs and Transition Matrices

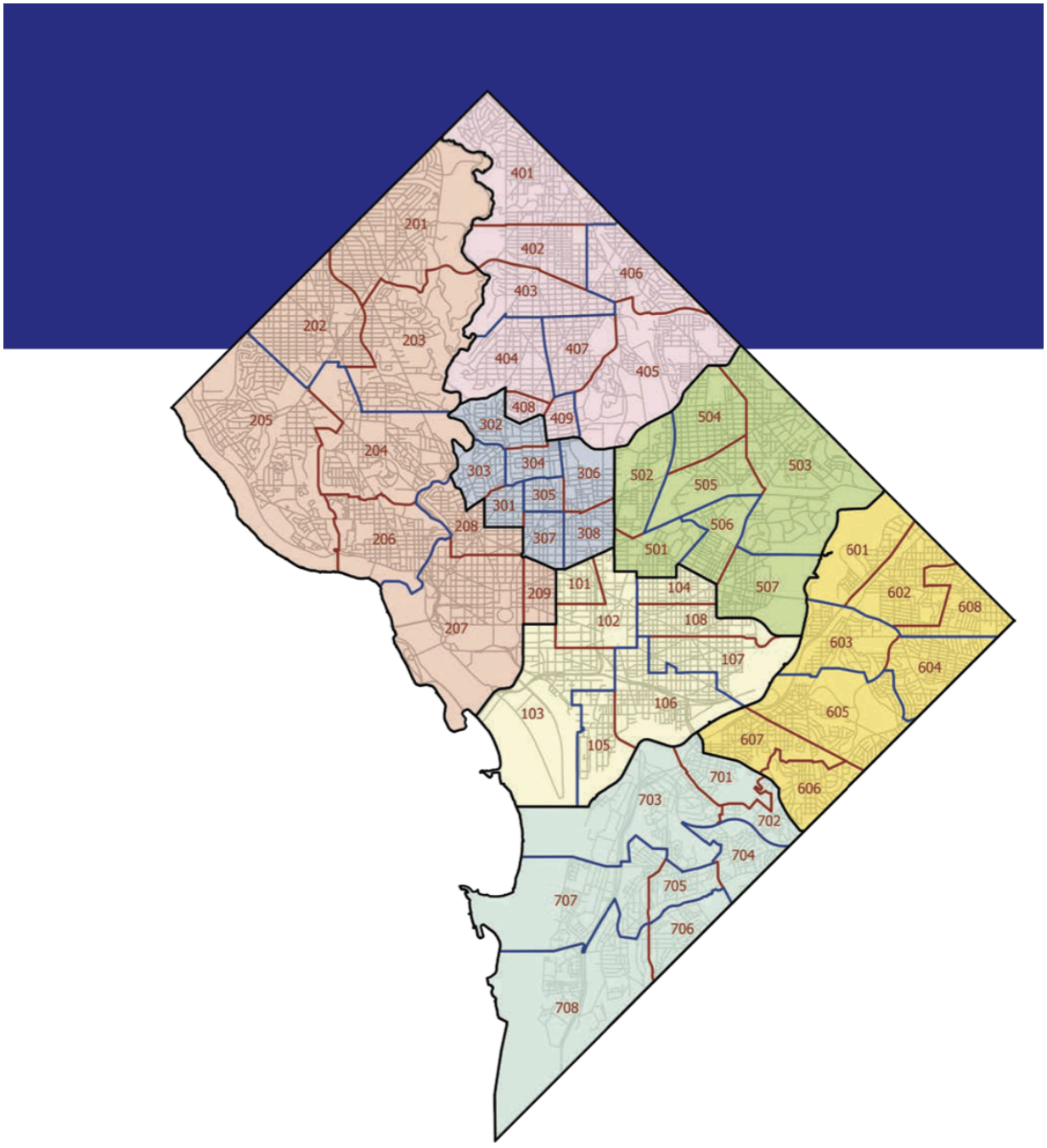


Figure B.1: Map of the all PSAs across DC

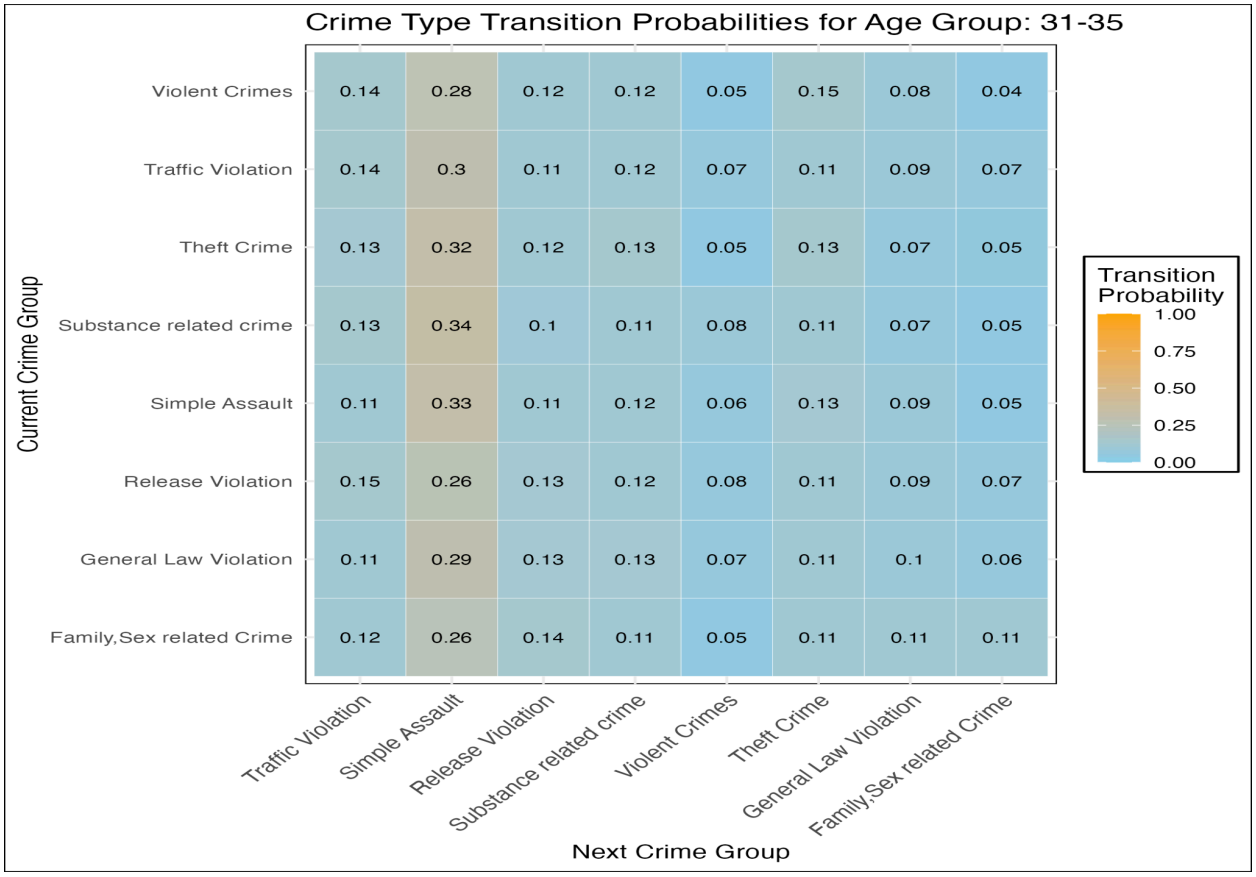


Figure B.2: Transition Matrix for Age Group 31-35

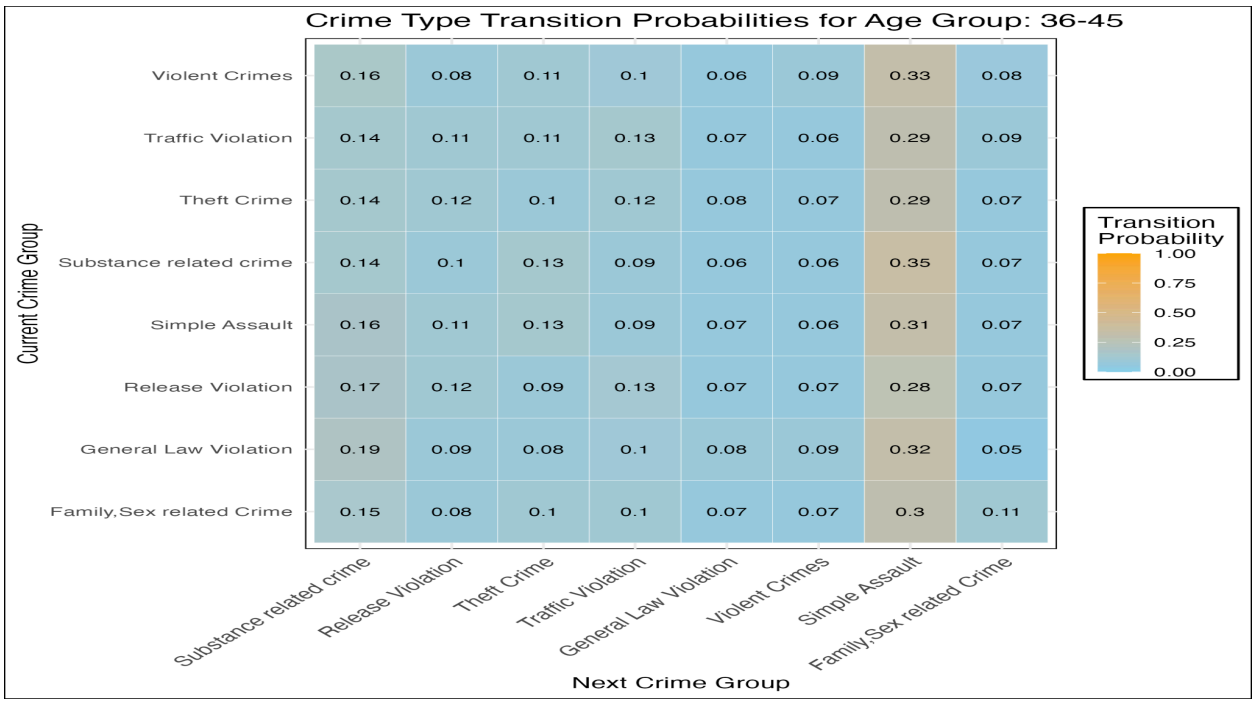


Figure B.3: Transition Matrix for Age Group 36-45

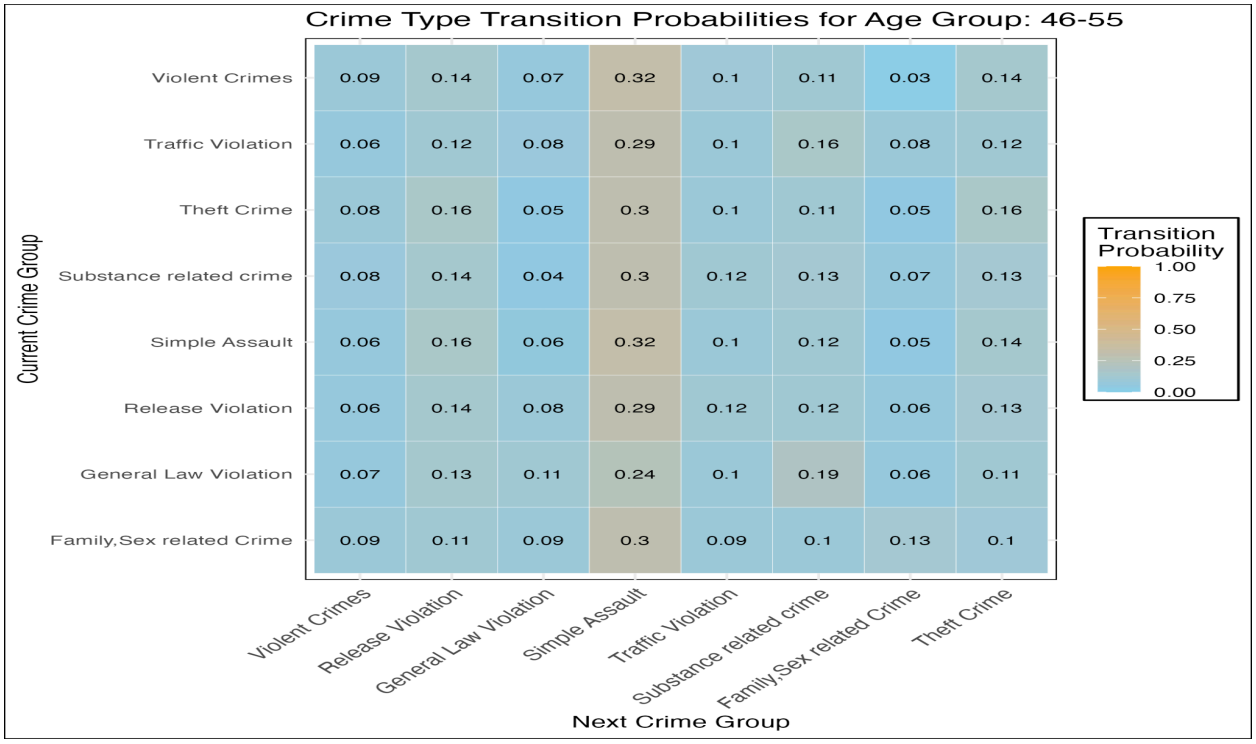


Figure B.4: Transition Matrix for Age Group 46-55

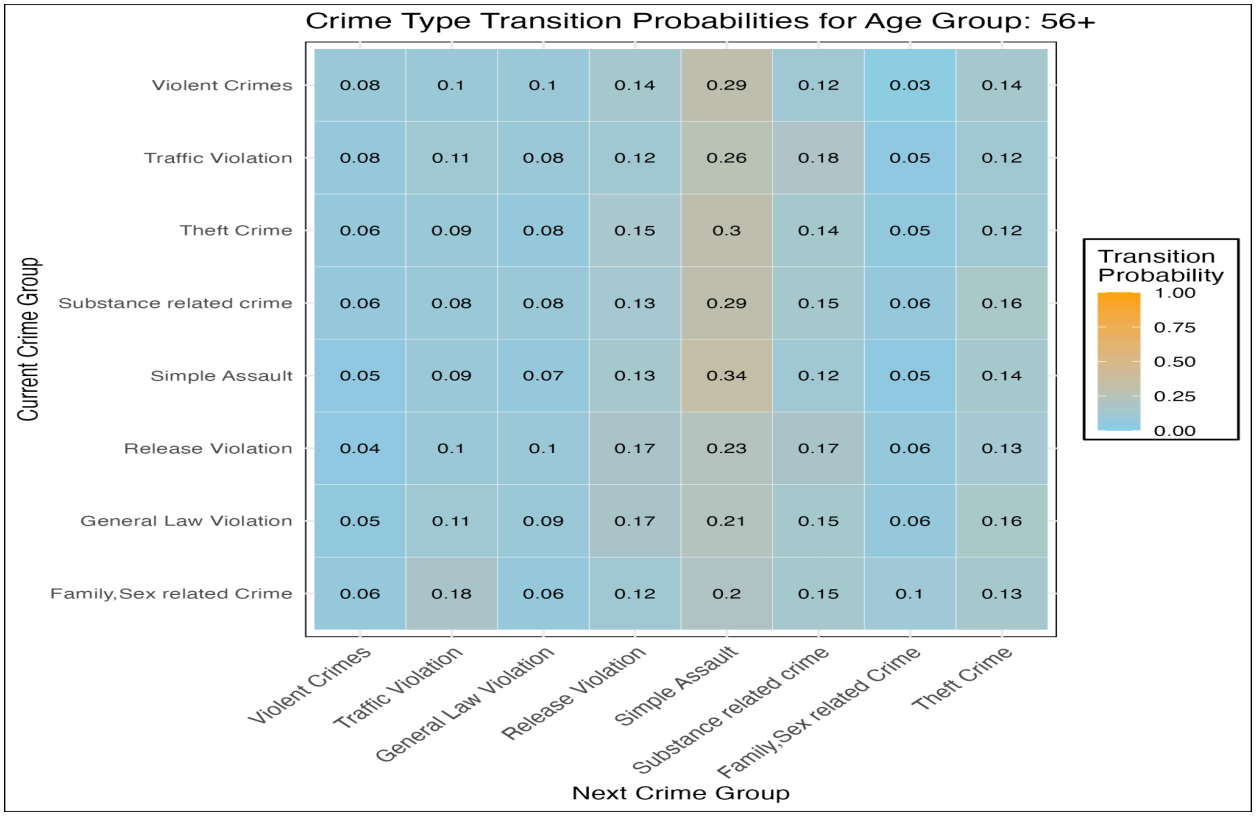


Figure B.5: Transition Matrix for Age Group 56+

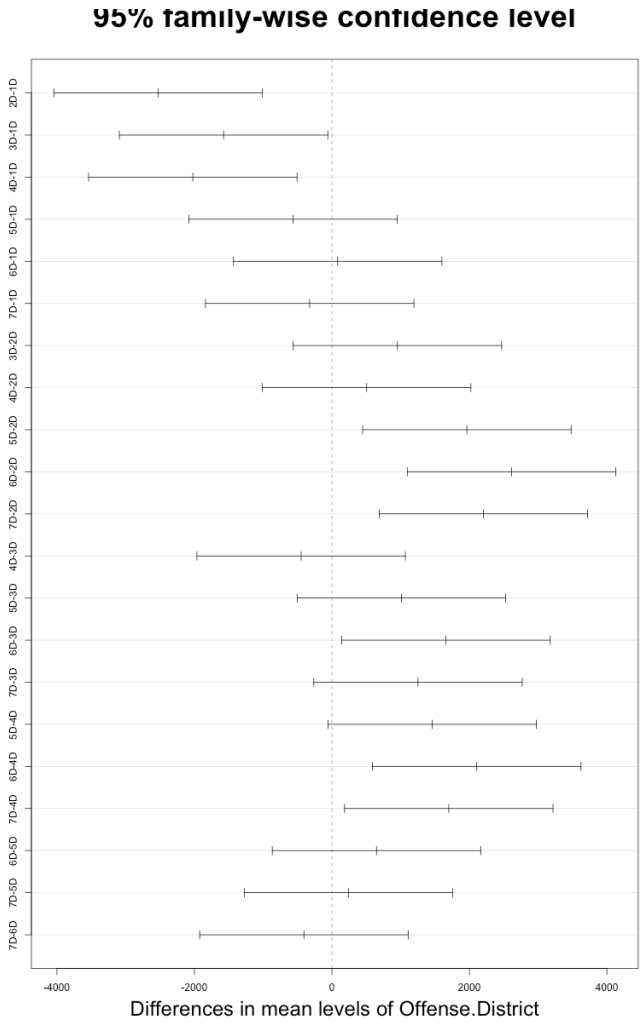
Appendix C: Test Statistics and Equations

$$D_n = \sup_x \left| F_n(x) - F(x) \right|$$

C.1: D Statistic

$$G = -2\sum_{i=1}^{m} O_i \ln\left(\frac{E_i}{O_i}\right)$$
$$= 2\sum_{i=1}^{m} O_i \ln\left(\frac{O_i}{E_i}\right)$$

C.2: G-statistic



C.3: Tukey's Test

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where:

- $\bar{x}$ = the sample mean
- μ_0 = the hypothesized population mean
- s = the sample standard deviation
- n = the sample size

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

C.4: T-statistic

$$F = \frac{MST}{MSE}$$

$$MST = \frac{\sum_{i=1}^{k}(T_i^2/n_i) - G^2/n}{k-1}$$

$$MSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k}(T_i^2/n_i)}{n-k}$$

C.5: ANOVA F Statistic, Mean Square of Treatment (MST), and Mean Square Error (MSE)