Nevra Diker

DATA 205

32118

Reported Crimes Within Montgomery County

For the past few months, I’ve been analyzing reported crime data within Montgomery County. This has allowed me to understand trends in public safety within my own community. Through my project, I was able to work with real world data and draw predictions that could potentially help the public safety workers and locals of Montgomery County.

The dataset I will be working with is the ‘Crimes’ dataset from dataMontgomery which is compiled by the Montgomery County Police Department. It has 440,000 rows of data on reported crimes within Montgomery County. It consists of 30 columns such as the type of crime that was reported, the category the crime fits into, where the crime reportedly happened, and the agency the crime was reported to. It is updated daily and includes crimes dating back to July 1st, 2016.

The main focus of my project is to analyze patterns and trends of reported crime in Montgomery County to help enhance public safety and support crime prevention efforts. My goal is important because using and exploring this data to understand what kind of crimes are reported and where they take place could help police agencies distribute their resources more efficiently and prevent crime before it even occurs.

The main tools I used during the duration of my project are R Studio, Excel, and dataMontgomery. I did all of my data cleaning, exploratory data analysis, and visualization in R Studio. I used Excel to have a copy of the untouched dataset; I found that having the original dataset on hand was useful while I was conducting my exploratory data analysis and changing certain aspects of the dataset.

To begin my data cleaning and pre-processing journey, I checked the amount of NA (Not Available) values within each column of the dataset. None of the columns that I worked with contained NA values, so I left them alone. For the few columns that I worked with, I made sure to check the data for any inconsistencies. Things like capitalization differences or misspellings would cause issues during the exploratory data analysis phase of my project. I was able to use the ‘unique’ function in R Studio to check each variable present and confirm that these inconsistencies were not present.

Going into this project, I knew that I would be working with yearly data for a good portion of my analysis. Since the dataset didn’t have a column that only showed the year a crime was reported, I had to create one. Using the ‘Start\_Date\_Time’ column which gives us information on the time and day the crime was reported, I extracted the year values and inputted them in the new column, ‘Year’. Since the dataset doesn’t have a full year of data for 2016 and 2025, I knew that I would have to exclude both of these years in some of my visualizations and analysis, such as the yearly average of reported crime and future predictions of crime reports, in order to prevent any sort of skewing or bias.

There were a few columns I worked with that had vague names. I believe that they didn’t accurately portray the variables within them. The first column was originally named ‘Crime Name1’, which was then changed to ‘Category’. This column contains information about the category of crime that was reported. These categories can be one of four: ‘Crime Against Person’, ‘Crime Against Society’, ‘Crime Against Property’, or ‘Crime Against Not a Crime’.

The second column name I changed was ‘Crime Name2’. This column lists the actual crime that was reported. It contains anything from ‘Simple Assault’ and ‘Embezzlement’ to ‘Runaway’ and ‘Littering’. This column became ‘Crime’.

The third and final column name I changed was ‘Crime Name3’, which was changed to ‘Description’. The ‘Description’ column has additional information about the crime. For example, if a crime is reported as ‘Destruction/Damage/Vandalism of Property’ in the ‘Crime’ column, the ‘Description’ column would have a data entry of ‘Damaged Property – Private’ or ‘Damaged Property – Public’. It’s important to note that not all reported crimes are actually crimes; the dataset also contains variables such as ‘Runaway’ or ‘Sudden Death’.

The ‘Crime’ column had a crime named ‘All Other Offenses’ which was an umbrella term for over 90 other crimes. These crimes were described in the ‘Description’ column, and they ranged from ‘Police Information’ to ‘Lost Property’. I was able to extract the values from the ‘Description’ column and replace every ‘All Other Offenses’ value for the actual crime.

The ‘Crimes’ dataset has a large number of interesting columns and variables. However, although the dataset contains 30 columns, most of them are relatively useless for the purpose of my project. I focused on the ‘Category, ‘Crime’, ‘Description’, and ‘Year’ columns. I also used the ‘Agency’ column, which tells us which public safety agency the crime was reported to. These agencies include the Montgomery County Police Department (MCPD), Gaithersburg Police Department (GPD), Montgomery County Fire Marshall (MCFM), Montgomery County Sheriff’s Office (MCSO), Rockville City Police Department (RCPD), and Takoma Park Police Department (TPPD). Through exploratory data analysis, these five variables provided insight into the deeper layers of the dataset.

I was able to find out that the average number of reported crimes per year is about 50,000 and the Montgomery County Police Department sees the highest number of reported crimes per year. There is a noticeable decrease in reported crime during 2020 and 2021. This is most likely due to the COVID-19 pandemic and the quarantine that everyone was under. Overall, there has been a decrease in the number of reported crimes since 2017.

I was also able to figure out that the most reported crime is ‘Theft From Motor Vehicle’ and the most reported crime type is ‘Crime Against Property’. However, before I extracted the values from ‘All Other Offenses’, it used to be the most reported ‘crime’. The place where the most crimes are reportedly committed turned out to be ‘Residence – Single Family’. The crimes that are reported as taking place in ‘Residence - Single Family’ tend to be ‘Simple Assault’ and ‘Burglary’.

For my final data product, I wanted to create a 5-year predictive model of reported crimes by agency and the places in which they were reportedly committed. To begin this process, I had to filter the data to only include crime reports from 2017 to 2024 since 2016 and 2025 don’t have a full year’s worth of data. Then, I grouped together and counted the number of crime reports by ‘Year’ and ‘Agency’. Using the filtered data, I created a visualization of the total crime reports by agency per year, so that I could compare it to the prediction model. Looking at the graph, it’s obvious that MCPD has the highest number of crime reports. It’s important to note that there is a decrease in the amount of reported crime at every agency over the past few years. Although COVID-19 is the main reason for this, we can see a decrease in reported crime even after the pandemic happened. We should expect to see something similar to this in the prediction model: MCPD having the highest amount of crime reports but also seeing a slight decrease.

**A graph with numbers and lines

AI-generated content may be incorrect.**

To build the prediction model, I used a linear regression approach to forecast crime reports for each agency from 2026 through 2030. I first grouped the cleaned data by agency and fit a simple linear model to each one, using the year as the independent variable and the number of reported crimes as the dependent variable. This allowed me to visualize the trend of reported crimes over time for each individual agency. Then, I generated predictions for the next five years and created a line chart to visualize the predictions.

**A graph with numbers and lines

AI-generated content may be incorrect.**

We can see that the Montgomery County Police Department is predicted to continue holding the spot of the agency with the highest amount of crime reports. We can also see that the number of crimes reported to MCPD is predicted to continue decreasing. The other agencies seem to be constant with the amount of reported crime that they are predicted to see over the next five years.

In addition to creating a 5-year prediction model for crime reports by agency, I also wanted to create a prediction model that shows the percentage of crime reports by place. To start, I once again filtered out the years 2016 and 2025 since they don’t contain a full year’s worth of data, ensuring the results would be consistent and reliable. I then aggregated the number of reported crimes by place and year, and identified the top five places with the highest overall crime counts between 2017 and 2024. For each of these top places, I built a linear regression model using the year as the predictor and the number of reported crimes as the outcome, in order to generate crime forecasts from 2026 to 2030. After creating the predictions, I calculated what percentage of the total predicted crime each place represented for each year. I then created a bar chart showing how crime might be distributed across the top places over the next five years, and whether certain places are expected to continue contributing a large share of reported crimes in the future.

**A graph of numbers and colors

AI-generated content may be incorrect.**

To display my prediction models and crime data in a more interactive and user-friendly way, I built a Shiny dashboard in R Studio that allows users to explore various aspects of the data. The interface includes a sidebar where users can select a specific agency or view all agencies, as well as adjust the year range for predictions from 2026 to 2030. The main panel of the app contains multiple tabs with which you can toggle to the different visualizations that I have created. The ‘Predicted Crime Plot’ tab shows the line graph of predicted yearly crime reports by agency. In association, the ‘Predicted Crime Table’ provides a detailed table of those predictions, including upper and lower confidence intervals. The ‘Total Crimes by Agency’ tab presents an overview of total reported crimes by agency per year, along with the top place where each agency reported the most crimes. To observe the total crimes that are reported to each agency without MCPD minimizing the lines of the other agencies, the ‘Total Crimes (Excl. MCPD)’ tab excludes MCPD from the plot. The ‘Agency Totals Table’ summarizes overall crime report counts per agency. For location-based insights, the ‘City-Agency Table’ shows the relationship between cities and their reporting agencies. Finally, the ‘Place Predictions Plot’ visualizes the predicted percentage distribution of crime reports across the top five places from 2026 to 2030, while the ‘Place Predictions Table’ gives the raw values used in that visualization. To conclude, the dashboard I’ve created allows users to look at trends that are present in the dataset and future predictions.

Acknowledgements

I would like to thank Professor Jane Valentine for helping me throughout the Capstone experience and providing me with valuable advice when necessary. Thank you to my DATA 205 classmates that went through the struggle of creating a fulfilling project with me. I would also like to thank Professor Mais Alraee for helping me troubleshoot my code and showing me a different way of looking at things. Finally, I would like to thank Professor Rachel Saidi for teaching me the foundations of Data Science.

References

Montgomery County, MD. “Crime: Open Data Portal.” *Crime | Open Data Portal*, Montgomery County, MD, 6 May 2025, data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3/about\_data.

“Montgomery County, MD.” *Montgomery County Maryland*, montgomerycountymd.gov/. Accessed 6 May 2025.

“Learn R.” *Codecademy*, Codecademy, www.codecademy.com/learn/learn-r. Accessed 7 May 2025.

“Forecast.” *RDocumentation*, www.rdocumentation.org/packages/forecast/versions/8.23.0. Accessed 7 May 2025.

*Shiny for R :: Cheatsheet*, rstudio.github.io/cheatsheets/html/shiny.html?\_gl=1%2A13qc3lq%2A\_ga%2AMTg5MzQyMzY0Ny4xNzQ0MzI1MTE1%2A\_ga\_2C0WZ1JHG0%2AczE3NDY2MDExNDAkbzIkZzAkdDE3NDY2MDExNDAkajAkbDAkaDA. Accessed 7 May 2025.

“Learn Linear Regression with r: Linear Regression in R Cheatsheet.” *Codecademy*, www.codecademy.com/learn/learn-linear-regression-in-r/modules/linear-regression-in-r/cheatsheet. Accessed 9 May 2025.

Thulin, Måns. “Modern Statistics with R.” *11 Predictive Modelling and Machine Learning*, modernstatisticswithr.com/mlchapter.html. Accessed 9 May 2025.