Visualizing Crimes in Montgomery County

Alex Valentiner - alval20@student.sdu.dk

Magnus Kjær Sørensen - magns20@student.sdu.dk

Mikkel Priisholm Larsen - mikkl20@student.sdu.dk

Sebastian Christensen Mondrup - semon20@student.sdu.dk

Dashboard link:

Abstract

Background and Motivation

Our dataset consists of crimes committed in Montgomery County in the state of Maryland in the USA. Montgomery County is the most populous in Maryland and it borders Washington D.C. The population density of Montgomery is 2153.8 per square mile, compared to 632 per square mile in all of Maryland.
The dataset has not been chosen because of any specific interest in the subject of crime but for some of its technical properties. The dataset is quite large - more than 300,000 crimes - which is interesting, as one can be relatively confident that the trends found reflect reality. The large dataset also creates some interesting technical challenges, like making interactive visualizations performant.
Furthermore, the dataset was found interesting because of the data types involved. The date and time of each crime is recorded, which gives opportunity for a wide range of visualizations. The data also contains different categorical variables, which was found interesting. The type of place (e.g. street, hotel etc) where each crime happened is recorded as well as the type of the crime.

Project Objectives

The visualizations will answer questions about the crime in Montgomery County. The data will explain the different types of crime, describe where and when most crime occurs, and the number of victims involved in each crime. Additionally it will show how these have changed over time.
The aim of the project is to answer the questions in the list below, by visualizing the dataset.

- Which places have the highest number of victims?
    - Has the number of crimes changed over time?
- How have crime rates changed over the years?
    - Are there any specific days of the year where the high crime rate repeats?
- What type of crime is most prevalent over time?
- At what time of the day does crime occur?
    - How does the distribution look at different places?
    - How does the distribution look for different crime types
- How many victims are involved in crimes in different places?

Data

The data was retrieved from the Montgomery County (Maryland)'s own data website[1], where it is possible to find details about the dataset and even do basic visualization of the data. Montgomery County uses a law enforcement records-management system named "EJustice" that compiles this data for them. The data is updated daily, but at the time that the data was pulled from the site (November 8th 2022), the dataset contained 318.865 records, starting in late 2016 and onwards.

According to the Montgomery County data website, the data is "derived from reported crimes classified according to the National Incident-Based Reporting System (NIBRS) of the Criminal Justice Information Services (CJIS) Division Uniform Crime Reporting (UCR) Program and documented by approved police incident reports.". It should be noted that the data is based on preliminary reports, and therefore might not reflect the full/true picture e.g. due to lack of verification/follow-up on some reports.

While the dataset includes 30 different variables, only a select few were used in the making of this report. The relevant variables include:

- Start_Date_Time: The date and time the crime occurred from.
- Victims: The number of victims of the crime.
- Crime Name1: Crime against Society/Person/Property, Other or Not a Crime.
- Place: A categorisation of the place. Examples: Hotel/Motel/Etc., Liquor Store - County, Residence - Single Family, and many more.

Of the variables that were not used, most describe the physical location of the crime in a variation of ways (long and latitude, street type/name, police sector/patrol area, and more).

Data processing

This section describes some of the larger changes that were done to the dataset, and a description why it was done.

Cleaning

The dataset contains some rows which have some columns with the value *Not Available* (NA) or an empty string. This was not optimal to use for the diagrams, so some data cleanup was done. All rows which have an empty string in a used column were put into the category *Other* by mutating the row. The dataset was filtered by removing rows which have the value NA in a column used for a diagram.

[1] https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3

Time

The dataset provided the year, month, day, and time of day for a crime. An additional column was added which had the time omitted, so the value was only year, month, and day. Some diagrams only required the year, so another column was made where the values had the unit *year*. The unit *year* changed the month and day to 01. The reason behind this was to make grouping more easy for diagrams.

Place

The data contains a column called *Place* describing the place for a crime. A place can be locations such as a Library, University, and Gas Station. There are 99 unique places in the dataset. A decision was made to mutate some of these rows, so they could be grouped together into groups. For example a lot of rows had values with the keyword *Street* and then something else, these rows got their value in the column Place changed to *Street*.

Visualization/Dashboard

Which places have the highest number of victims?

To answer this question three variables are used. The first variable is a nominal categorical variable called “Crime.Name1”, and it is the type of crime. The second variable is a discrete numerical variable called “Victims”, and is the number of victims of a crime. The third variable is called “Place” and is a nominal categorical variable. As mentioned before, the variable “Place” has its values grouped into smaller categories. A downside with the grouping is that some places can dominate their category. These three variables were used to build two stacked bar charts. Both diagrams have the variable “Place” as their y-axis and the variable “Victims” as their x-axis. The bars are the type of crime and make it possible to see the largest type of crime in a place. The colors chosen are very different from each other. This choice was made so it is easier for people to distinguish the different types of crimes from each other because some of the bars were small. The first stacked bar chart shows the total number of victims on the x-axis, so it is possible to see the places with the highest number of victims. The second stacked bar chart has normalized data, and this is done so it is more clear which type of crimes are the most dominant for a place.

Has the number of crimes changed over time?

This question is referring to the number of crimes for each place over time.
The variables “Place” and “Crime.Name1” are also used to answer this question. A third variable called “year” is also used. The variable *year* is a categorical variable made from the variable called “Start_Date_Time”. The variable *year* has the

structure: year-month-day. The unit for the variable is declared as “year”, and that changes the month and day to 01-01, so the value will be year-01-01. The reason for using the unit *year* is that the months and days are not needed for this diagram, so giving all data the same day and month makes it easy to group the data by year. These three variables are used to make a line plot with multiple lines. The x-axis is the year range from 2017 to 2021. The dataset also contains data for some of 2016 and the current data in 2022, but this data is omitted. The reason is that it will show a lower amount of crimes for 2016 than it really was, and not show all the crimes there will be in 2022. The y-axis is the number of crimes. The number of crimes is calculated by counting the number of crimes for a place in a given year. The lines are representing the places, and they are shown as colored lines. Points are added to the lines to make it easier to see the number of crimes for a year.

At what time of the day are crimes committed?

To answer this question, the “start date time” variable is used, and from this, the time of the days is extracted. To visualize how the crimes are distributed over the day, a bar plot is used. The bar plot has the hour of the day on the x-axis and density on the y-axis. Thus, the channel used for time is position, and the channels used for density is length as well as area. The reason for visualizing the density within each hour instead of the count, is to make it natural to put a density curve on top of the diagram later on.

To visualize the circular nature of the data (the hour 0 comes after the hour 23) a similar diagram but with polar coordinates have been created. This diagram does a good job of giving an intuition about the data, but it also has its problems. First up, it is harder to read the density values on the round diagram. Secondly, length is now the only visual channel showing the density values, as the area of the bars doesn’t correlate with density on this plot. Only having one visual channel for a variable isn’t a problem in and of itself, but it might be confusing in this case as people could (consciously or unconsciously) mistake area for representing density. Lastly, it might be a little confusing for the reader to see a circle representing the 24 hours of the day, when we are used to seeing clocks, which only show 12 hours.

The original idea for a diagram showing the distribution of crimes over the day, was a kernel density plot or a violin plot. A bar plot was chosen though, but the option of overlaying a density curve was added. The added value of the density curve lies in it capturing the continuous nature of the variable. The bar plot will sometimes show a steep (vertical line) increase or decrease between two hours, when the change is actually more gradual, and this is captured in the density curve. It was somewhat of a technical challenge to create the density curve because of the circularity of the variable - the visualization software didn’t know that hour 0 came after hour 23, causing the density curve to not match up at the ends as it should.

In order to allow the user to answer more specific questions about when on the day crimes happen, two filtering options were added. It is possible to select one or more crime types to show data for, and to select one or more place types to show data for.

While experimenting with the diagram, it was discovered that seeing distributions for e.g. a specific crime type wasn't that informative if you couldn't compare it to the overall distribution. Therefore an option was added to show the overall distribution on top of the distribution with the selected filters. The overall distribution is shown with another color, and the bars are semi-transparent, allowing the user to see both distributions at the same time.
For this diagram, performance is somewhat of an issue. The filtering makes the diagram load really slowly if it is using the full dataset. Therefore a random sample of the full dataset is used for this plot. This is fine when showing the overall distribution, but if a very specific filter is chosen (e.g. only one place and one crime type) the diagram will sometimes show only a few records, making the diagram less usable. This is somewhat of a tradeoff between performance and correctness of the diagram.

Are there places where crimes with several victims are more prevalent?

For answering this question, two variables are needed: number of victims, and place. Number of victims is a discrete numerical variable and is very right skewed. The skew of the victims variable makes it hard to make a visualization which shows anything other than the fact that the data is skewed. One of the logical things to try when visualizing the distribution of records along a categorical and a discrete numerical variable would be a stacked bar chart (normalized or not). Attempts at this yielded diagrams where only the part showing the crimes with one victim were visible.
When dealing with very skewed data, one solution is a logarithmic scale. Applying a logarithmic scale to a bar chart seems like a bad idea though, as length as a visual channel is perceived linearly, and the risk of confusing the user is very high. It was decided that using color as the visual channel for the logarithmic scale would be less confusing, as change in color is not perceived as linearly as length. The resulting diagram is a heatmap, with the number of victims on the x-axis, place on the y-axis, and the logarithm of the count in each bin deciding the color. To decrease the chance of confusion because of the logarithmic scale, the number of records in each bin is displayed as a label on top of the bin. As the number of victims is so skewed, just a few crimes have more than 7 victims. To avoid having a lot of empty bins, the crimes with 7 or more victims have been grouped together.
For the color scheme, several things had to be taken into consideration. The colors had to be differentiable for people with color blindness and no bin could be too dark, as the text on top would be unreadable. The chosen colormap is a gradient between a semi-dark blue and a light sand color. These are different enough that small changes are visible, and at the same time, the scheme is color blind friendly, while the text can still be read on top of the darkest blue in the scheme.

One problem remaining, is that the different categories in the place variable contain vastly different numbers of records. This means that it is difficult to tell whether one place has a higher proportion of crimes with several victims. To make this easier, a checkbox is added allowing the user to normalize with respect to the place. With this option checked, each bin in the diagram shows what percentage of crimes in the given place has the given number of victims.

What type of crime is most prevalent over time?

To classify the categories each crime falls into, the nominal categorical variable “Crime Type 1” is used. This variable labels crimes into 5 categories: “Crime Against Person”, “Crime Against Property”, “Crime Against Society”, “Not a Crime” and “Other”. Although the dataset includes two additional variables that function as sub categories to the “Crime Type 1”, these are not used to answer the question, as they divide crimes into a much too large number of categories.
To map the distribution of crimes over time, the variable “Start_Date_Time” is used to add a label to each month that describes in which month they happened, creating a new categorical ordinal variable. As the question is only in regards to the distribution of crimes and not the precise number, and that the total number of crimes changes for each month, the data is normalized to be percentwise for each month. Each type of crime has a percentage tied to it, for each month in the dataset, that describes how many of the total crimes committed in that month were of that specific type. With all this information, what should actually be visualized is: The type of crime, the percentage that type took up, and the month.
An array of different types of graphs can be used to visualize these variables, such as: A stacked bar chart, line chart, multi-set bar chart, interactable/animated pie chart and more. While it is indeed possible to illustrate with a graph that shows the distribution for one month, and then either have the user interact to change between months, or animate the graph to switch between months, this does not give the best options for comparisons, which is needed to answer the question. Furthermore, knowing that one of, if not the, strongest visual channel is position, the catalog can be narrowed down quite a bit.
As the dataset includes data from 2016 till 2022, the amount of months is somewhere around 80 months, a stacked bar chart for each month is not suitable, but instead a line chart is chosen. The X-axis representing the months, one after the other, and the Y-axis showing the percentage. For each crime type a line on the graph is made The identity channels used to distinguish these lines from each other is color hue - while the worse option would be shape (dashed/dotted/etc. lines). As there are only five categories, choosing a set of colors that are easily distinguished from each other is not an issue.

All of this allows a clear view of changes between months and comparisons between the types of crimes for each month and over time. To further illustrate the time aspect of the graph, the graph is animated to reveal the lines month after month.

How have crime rates changed over the years?

To answer this question, the data needed are all the dates between a start date and an end date, and the amount of crime that occurred each day. This can be obtained with the variable “start date time”, where date occurrences can be counted. The best visualization for this would be a calendar heatmap. The calendar heatmap needs the categorical variables: month, day and year. This can also be obtained from the variable “start date time”.
The dataset includes data from 2016 to 2022 but it was decided that the calendar heatmap would only include data from 2017 to 2021, since the excluded years only contained data for half of the year. To visualize if there were any specific days of the year where the high crime rate repeats, the calendar heatmap includes the specific day of the week and the month. This should better visualize if a specific date repeated a high amount of crime each year. For some readers it might be confusing that the week starts with sunday but this is depending on the readers region. It was discussed to make the weeks start with monday but this would mess up the algorithm used to count the dates, so it was dismissed. Since the day of the week is not the same for each year, it was decided that each month containing the days should be clearly separated. This was done by drawing lines separating the months. The color scheme was decided to be a simple gradient from light yellow color to a dark red. To make the heatmap show the highlights of the years, the gradient contains more light colors before turning red. Which makes the difference between the low amount of crimes (under 150 occurrences) very small, but will clearly show a difference when going from 150 to 200 and from 200 to 250.

Story/Results

Which places have the highest number of victims?

The stacked bar chart shows that the place with the most victims is “Residence”, and “Residence” has over 105000 victims. “Residence” accounts for places such as “Residence - Apartment/Condo”, “Residence - Yard”, and more. The place with the second most victims is “Street”, and “Street” has just below 75000 victims. “Street” is places like “Street - In vehicle”, “Street - Bus Stop”, and more. The place with the third highest number of victims is “Other”, and “Other” has around 40000 victims. “Other” is the category with the most places because the dataset had a lot of places that did not fit into our categories and they had relatively few victims, so they were put into “Other”. The normalized stacked bar chart can then be used to see the most

common type of crime for a place. The diagram shows that the most common type of crime for “Residence” is “Crime Against Property”, but “Residence” also has quite a high percentage of crimes in the category “Crime Against Person” compared to most of the other places. The most common type of crime for “Street” is “Crime Against Society”, and that can be things such as drugs or gun violations.

Has the number of crimes changed over time?

This diagram can be used to see if a place experiences a decrease in the number of crimes, and therefore the number of victims. The place “Residence” has experienced a decrease in the number of crimes. The decrease is especially between 2018 to 2019 and then again from 2020 to 2021. The place “Street” has experienced a massive decrease in the number of crimes. In 2017 the place “Street” had shy of 15000 crimes, and that was dropped to 7500 in 2021. It is possible to see that in general the total amount of crimes has dropped since 2017, since almost all places have experienced a decrease.

At what time of the day are crimes committed?

This part will look at the distribution of crimes over the day, first generally and then for some specific crime types and places.

Overall distribution

The diagram shows that most crimes are committed in the afternoon and the evening - between 12 O’clock and midnight. The peak in this period is in the hour between 15 and 16 in the bar plot and between 16 and 17 in the density curve. Less crimes are committed between 24 and 9, where a distinctive “U” shape can be seen, with the least crimes being committed between 5 and 6 in the morning. The hours between 9 and 12 are somewhat in-between.
It is worth noting the very large peak at midnight. This seems unnatural, and a possible explanation might be that the officers sometimes just enter the day where a crime happened, with the system then defaulting to 00:00 for the time of the day. A similar, although not as prominent, peak can be seen between 12 and 13. A theory might be that crimes that happened “around noon” might often be registered at the time 12:00. Something similar may explain the peak at midnight. Maybe crimes that happened “in the night” are often registered at 00:00. All this is just guessing though.

Applying some specific filters

Here, some specific filters have been chosen, which reveal clear trends. This is by no means a comprehensive list of trends that can be found with these diagrams. Choosing all crime types but only crimes committed on streets shows a distribution with way fewer crimes during the day (6:00-18:00) and more crimes during the night. This is what could be expected, as it is easier to get away with doing something

criminal on a street during the night, compared to during the day when a lot of people are on the streets.
The distribution seen if only the “school/university/college” place is chosen, is also very distinct. Almost no crimes are committed here between 24:00 and 6:00, only a few are committed after 18:00, and by far the most are committed between 9:00 and 15:00, where people are at school. So this is also not that much of a surprise.
When choosing specific crime types, some distinct trends can also be seen. Crimes against society are committed relatively more during the night, while crimes in the “other” category are committed more between 6:00 and 15:00. These categories are so broad that it is hard to say what was expected, as well as give an explanation for these trends.
Using these diagrams to answer the research questions worked very well. The option to overlay the general distribution was especially helpful. The diagram also had some superfluous elements though. The diagram with polar coordinates wasn’t as helpful at answering questions, but it was good to glance at it at times, especially when trying to compare the bars before and after midnight, which is hard on the standard bar plot. The density curves also weren’t as helpful and could perhaps be omitted. Making them optional with a checkbox was at least a good idea.

Are there places where crimes with several victims are more prevalent?

From the heatmap of victims, it seems that banks, stores and government buildings are places where crimes with several victims are rare. On the other hand, the place “residence” has by far the largest bins of 2 and 3 victims, but this also seems to be the place with the most crimes overall. When the normalization option is enabled, other interesting trends can be seen. Residences still relatively have the most crimes with two victims, but it can now be seen that for every number of victims from three and up, the school/university/college category has the highest prevalence - a scary conclusion.
This heatmap isn’t the easiest to read, but the skewed data was hard to visualize, and this seems like a decent attempt. The colors help the user get a quick overview, but the number of records / percentages on the bins were indispensable when trying to draw any conclusions.

What type of crime is most prevalent over time?

Over the first 3 years, from 2016 to 2019, the distribution remains relatively stable with some slight fluctuations. Crimes classified as Crime Against Property take up the vast majority, fluctuating between about 38% and 48%.

Between 2019 and 2020 there is a small decrease in Crimes Against Society with a small general increase in Crimes Against Property, but in early 2020 this trend witnesses a sudden spike. As Crimes Against Society sharply decreases, Crimes

Against Property does the exact opposite in sharp increase. After this point, Crimes Against Society remains at around 15-18% where it used to be around 28%-32%, and Crimes Against Property stays around 48%-55%, with a temporary exception. This sudden change could be a sign that some very common crimes that were previously categorized as Crime Against Society would after 2020 be classified as Crime Against Property. Similarly, in mid 2021, crimes classified as Other experienced a sudden and very temporary drop, only to go back to its usual average at around the 20%-24% mark. This decrease matches an increase in Crimes Against Property, and could again be a sign of categorisation changing, but this time temporarily.

While it is possible that these sudden changes could occur due to other events (Covid, protests, etc.), it seems unlikely that criminals suddenly, almost overnight, abandoned one type of crime to switch over to another type, without it affecting the rest of the categories.

According to the National Incident-Based Reporting System[2], Crimes Against Property's goal is to obtain money, property or other benefit - herein lies stealing, burglary and such. Crimes Against Society are crimes such as drug violations, gambling, prostitution, etc. Lastly Crimes Against Persons are crimes where "victims are always individuals" - murder, rape, assault, etc. With these distinctions, and looking at the graph as a whole, it is clear that crimes in search of monetary or other sorts of gain are much more prevalent than the rest. The "Other" category does contain a large share of the total crimes committed, which brings in a lot of uncertainty, considering that the category is in this visualization a sort of "unknown". To further visualize and investigate this unknown, the sub categories can be explored to learn what this "Other" includes.

How have crime rates changed over the years?

And are there any specific days of the year where the high crime rate repeats?

Over the years 2017 to 2021 there is a clear decrease in crime rates. The years 2017 and 2018 show the highest amount of crimes, which decreases in 2019. In March 2020 there is a big change in crime rates. This could be caused by the police changing what kinds of crime they document. If the crime is small and occurs a lot, it could be that they stopped documenting these to save time spent at the desk. It could also be speculated if this is due to the quarantine caused by Corona, as it shows a cutoff in crime at the start of 2020 and it slowly increases at the end of 2020. Another reason why it could be caused by the quarantine is that new year

---

2 https://ucr.fbi.gov/nibrs/2018/resource-pages/crimes_against_persons_property_and_society-2018.pdf

2021 has the lowest amount of crime of all the 5 years. This could be explained by a big decrease in new year parties, caused by Corona. This way less people would be out to get robbed or drinking and driving. Year 2021 ends with a clear increase in crime rates, therefore reinforces that it could have been caused by the quarantine and Corona.
Significant days of the year where there is a pattern in the crime rates, are days like new years. New year is consistently a day with some of the highest amount of crime throughout the year. This could be caused by a lot of things, like drunk people getting robbed, drinking and driving, people misusing fireworks. The first day of most months also has a high crime rate, which could be explained by, if the police don't know when the crime occurred they could be documenting it as the first of the current month.
An interesting observation is that there is not any significant amount of crime occuring christmas, the 24th and 25th. This is interesting since most people are out of their house visiting family and friends, which thieves take advantage of. National Crime Victimization Survey (NCVS) says that robbery increases by 20 percent during December.[3] But the calendar heatmap shows no significant difference in crime rates in December. The explanation for this could be that while robberies increased by 20 percent, other crimes may be decreasing in December, which may be leveling out the crime rates in December.

Conclusion/Discussion

- Some of the most interesting trends found
    - High prevalence of crimes with several victims at schools.
    - Kriminalitet på gaden er halveret
        - Covid???

- What was hard to visualize - what would we have liked to include
    - The "tree" of crime categories
    - Victims - right skewed
    - Calendar - Mikkel winger noget
        - Each year was different
        - Weeks start with Sunday
        - Required the whole dataset
    - Interaction with the line plot (and maybe more interaction in general
        - Tried with plotly, but it took too long to load
            - Alternative should be found
            - Too much data
                - Maybe sample?

[3] https://www.hg.org/legal-articles/understanding-the-reality-of-holiday-related-crimes-49947

finally you conclude the report by a summary of what you achieved, how you achieved, what were the challenges for you and how the course can be improved.