202120 – DATA-201-23387
Minh Tu Bui
December 16, 2020
Capstone Project - Final Report

Overview

The project focuses on the records of the police dispatched incidents in Montgomery County from 2017 to present. The main objectives are to conduct the forecast for the number of incidents using the time-series data, evaluate the responsiveness of the police stations in terms of travel distance to the incident locations, and recommend the optimal location of the current and new police stations.

The main dataset is the “Police Dispatched Incidents” from dataMontgomery that provides relevant information about crime types, incident locations and police district numbers. The analysis uses the additional “Police Station” dataset to generate significant findings that support the ultimate objectives of the project. The main programming language in the report is Python with multiple packages and libraries and Tableau Public for mapping visualization.

Data Cleansing

The analysis reads the data of the main dataset “Police Dispatched Incidents” directly from the dataMontgomery urls so that the report has the most updated data. The “Police Station” dataset has fixed information; therefore, the analysis uses the .csv file downloaded from the website.

“Police Dispatched Incidents” Dataset

The column names are adjusted from “name name” to “name_name” format so that all of the column names are consistent, which helps referencing the variables in the commands. The number of columns reduces from 26 to 17 as the report keeps the information needed for the analysis.

“Start_Time” Variable: The variable consists of the records of the time when the emergency center receives the calls. The report applies the date/time data type that is important in the time-series analysis. The recorded timeline is from July 2017 to present as the dataset is updated daily.

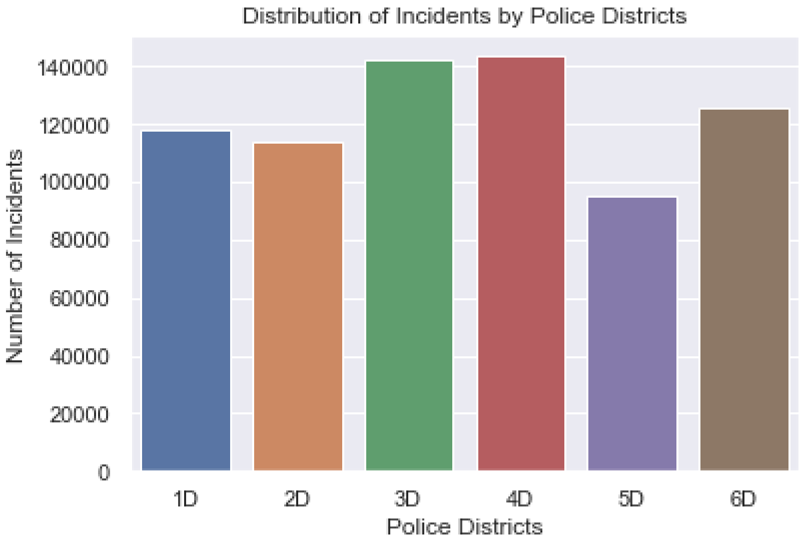
“City” Variable: The variable shows that the incident locations are not completely in Montgomery County and some locations belong to other counties in close proximity including Howard County, Prince George’s County, Anne Arundel County, and so forth. For the missing values in this variable, the Geopy library provides the implementation for third-party geocoders and other data sources to conduct the reverse geocoding that translates the coordinates of the incident locations to the actual addresses. The cities in the addresses are then retrieved for analyzing. The records with no information of City, Longitude and Latitude are removed.

“Distance” Variable: This new variable shows the travel distance between the police station and the incident location of each record. The report uses the “Police Station” dataset and the Vincenty formula to calculate the distance using the longitudes and latitudes in each record. The Vincenty formula calculates the geodesic distances between a pair of latitude/longitude points on an ellipsoidal model of the Earth. The five number summary shows that the majority of records are under 21 miles while there are eleven incidents with the distance over 200 miles. These observations need further discussions with the data owner and are removed from the dataset as they may affect the results of the analysis.

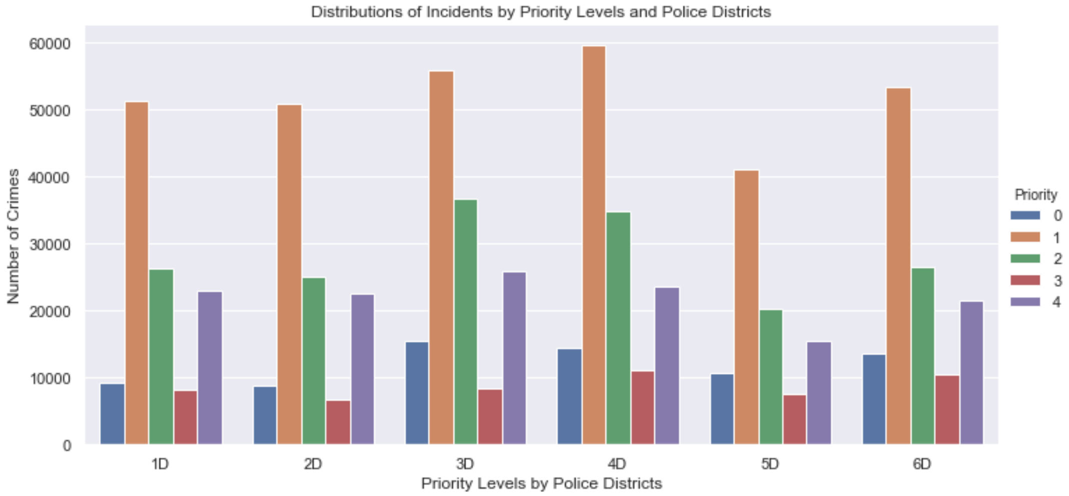
The data wrangling process also checks the “Priority”, “Police District Number” , and “CallRoute_Dispatch” variables. Based on the descriptions on the dataMontgomery website and the explanations from the data owners, the observations that do not match the predetermined categories are removed. The main purpose is to remove missing values that cannot be interpreted using the other variables in the dataset.

Exploratory Descriptive Analysis

Insight #1: Distribution of Incidents by Police Districts

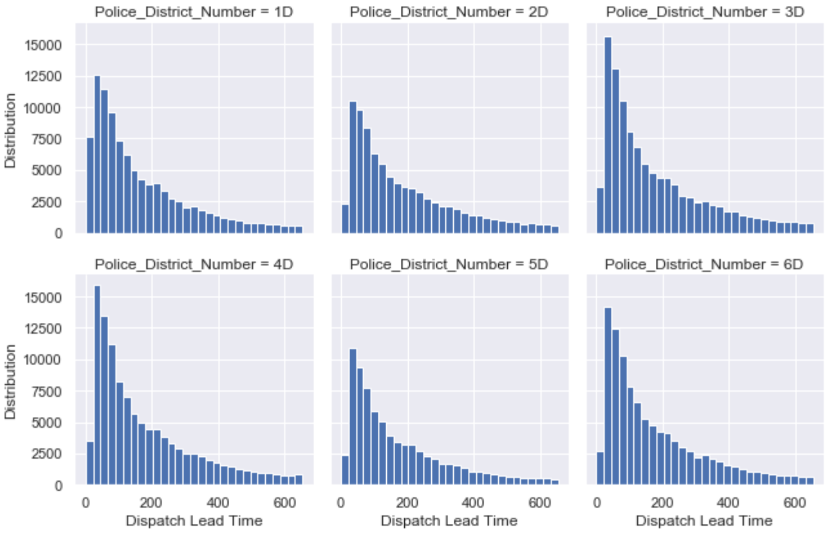


Insight #2: Distribution of Incidents by Priority Levels and Police Districts



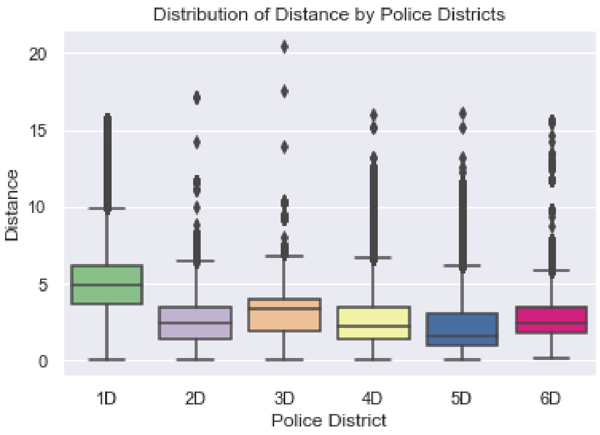
The bar charts show that the police districts in Silver Spring (3D) and Wheaton (4D) have the highest numbers of crime incidents with over 140,000 cases. Among the emergency calls assigned to the police stations, the majority of the cases have the priority level 1, followed by the priority level 2.

Insight #3: Relationship between Police District and Dispatch Lead Time



The “CallRoute_Dispatch” shows the time between the data entry to the first unit dispatched as the dispatch lead-time. The five number summary shows that the 75% quantile is 7 minutes. The 1.5 IQR concept helps identify the outliers. The one-way ANOVA test is not much affected if the data distribution is skewed unless the group sizes are small. Since the distributions of the groups are similar, we can still apply the ANOVA method despite the fact that they are not normal. Both ANOVA tests (with and without outliers) have the p-values lower than 0.05; therefore, we conclude that there is a relationship between police districts and dispatch lead-time while the expected result is that the dispatch lead-time should be equal among the police districts.

Insight #4: Relationship between Police District and Distance

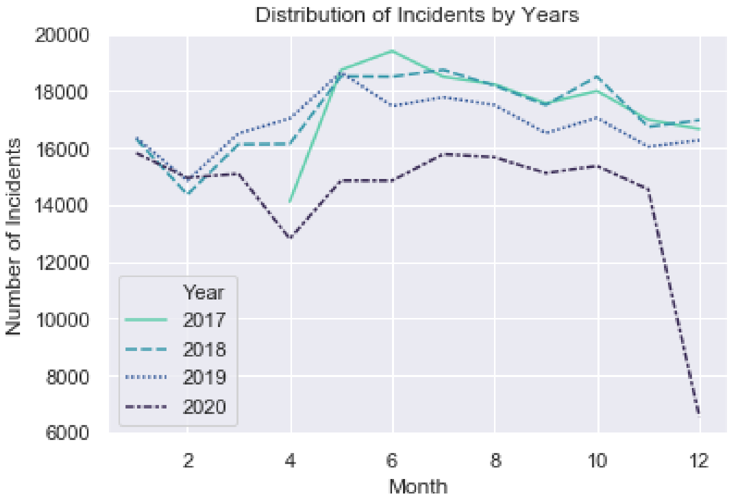


The report shows that the ANOVA test has the p-value significantly below 0.05; therefore, we conclude that there is a relationship between police districts and the distance between the stations and the incident locations. The result is the opposite to the expected result that the travel distances should be equal among police districts. The Rockville district (1D) has a higher median of 5 miles compared to the other districts.

Final Data Products

The first data product provides the time-series forecast for the number of calls received in 2021 using the SARIMAX method. The second data product presents the police dispatched incidents mapping on Tableau Public and the concept of the center of gravity to suggest the optimal location of the current and additional police stations.

Forecast of the number of calls received in 2021

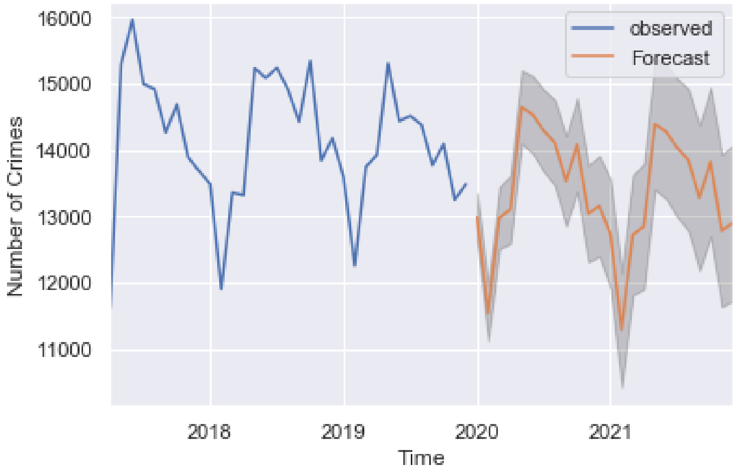


The line graph shows that the data from 2017 to 2019 have similar patterns while the number of calls in 2020 is significantly lower than the previous years. The dramatic reduction in April 2020 implies the impacts of the pandemic due to the stay-at-home order that limits gatherings and business operations in March 2020.

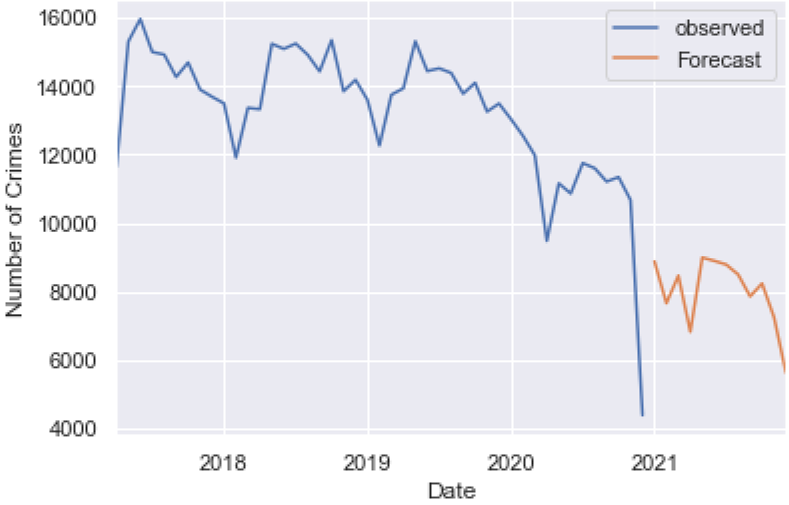
Originally, the analysis uses both SARIMAX method and Facebook Prophet library to conduct the time-series data. The report compares the root mean square errors of the forecasts to evaluate the performance of each method. The data product is the result of the SARIMAX model selection. Before conducting the forecast, the model check if the data is stationary. De-trending and de-seasonalizing applies to resolve the lack of stationarity.

Scenario 1: The forecast excludes the data of the year

The root mean square error of the first scenario is 567.55. While the number of incidents of each month is normally over 10,000 so the RMSE is acceptable.



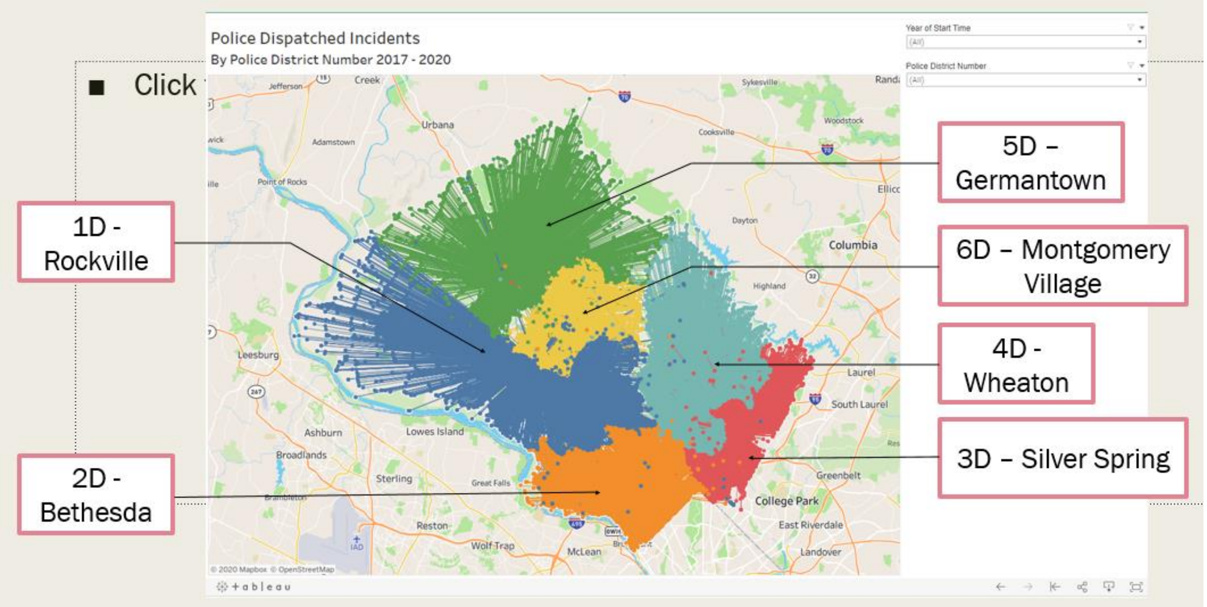
Scenario 2: The forecast includes the data of the year 2020.



The root mean square error of the first scenario is 1,973.35. The RMSE of the second scenario is higher than that of the first scenario due to the unique pattern of the data 2020. Furthermore, the cut-off time of the analysis is mid-December; therefore, the forecast may be affected by the insufficient data.

Optimal Location for Police Station

On Tableau Public, the latitude and longitude coordinates of the police stations and incident locations are used to create the origin-destination map showing the distribution of the dispatched incidents in the six police districts. The visualization supports the results of the one-way ANOVA test regarding the relationship between the police districts and travel distance. The Rockville station (1D) is responsible for the larger area compared to the other stations, and the descriptive statistics show that the Rockville station has longer travel distances. The suggestions are either relocating the 1D station or having an additional station near Dickerson and Poolesville cities.



To find the suggested location for the Rockville police district, the analysis uses the Shapely package to find the centroid of the incident locations. The package supports the analysis and manipulation of planar features using functions from the widely deployed GEOS library. The Vincenty formula helps recalculate the distances between the new police station and the incident locations for comparison.

Scenario #1: Re-locate the current Rockville police station

	Current Location	Include Dickerson	Exclude Dickerson
Count	117,754	117,754	117,064
Mean	4.967951	2.823915	2.724799
Std	1.940806	2.427731	2.223563
Min	0.02700000	0.03200000	0.05800000
25%	3.63200000	1.46800000	1.42700000
50%	4.95600000	2.20100000	2.15600000
75%	6.15300000	3.33100000	3.24200000
Max	15.62500000	19.24400000	16.31200000

In the first scenario, the suggested new location for the current police station is near Nelson St in Rockville. With this new location, the travel distances reduce significantly in general; however, there are outliers with longer distances compared to the current situation. These outliers are the incidents in Dickerson near the borderline between Rockville and Germantown districts.

Scenario #2: Add an additional station near Dickerson, Poolesville, and Beallsville

	Station 1 (Current Location)	**Station 2 (Near Dickerson)**
Count	114,131	3,623
Mean	4.810430	1.962484
Std	1.725272	1.804171
Min	0.02700000	0.008000
25%	3.584000	0.44500000
50%	4.890000	1.13700000
75%	6.057000	3.54900000
Max	15.62500000	6.55800000

The second scenario shows that the additional station helps solve the problem of long travel distances to those incidents in Dickerson, Poolesville and Beallsville. However, the travel distances have not been improved with the current police station. Therefore, re-locating the current police station is a relevant option in this scenario.

Limitations and Recommendations

The time-series forecasting applies SARIMAX and Facebook Prophet models and compares the root mean square errors to select the better option. There are other methods available for time-series forecasting that can be considered. Furthermore, the impacts of the pandemic and related events can affect the forecast. The predictions should be recalculated frequently due to the upcoming events such as the introduction of the vaccines.

The analysis uses a simplistic approach in finding the optimal location for the current police station with the assumption that the patrol police dispatch directly from the station. While the police are most likely on the road, the further analysis should consider the algorithm used to dispatch patrol police based on their current location. Furthermore, the analysis should check the financial requirements and resource capacity when proposing an additional station. The queueing theory is another relevant approach to check the performance of the police stations as the original dataset provides the time records of the call receipt, data entry, dispatch, and arrival of each incident.

Regarding the data quality, the analysis do not consider the extreme outliers in distances and dispatch lead-time that may cause the bias in the results. Therefore, future analyses need more discussion regarding these records to gain more insights.

DataMontgomery Experience

The search engine on the website is fast and precise. It allows filtering the options so the search process becomes more convenient. The visualization tools are easy to use with a variety of options to describe the data, specifically mapping. Compared to other data sources such as Kaggle, dataMontgomery provides more details about the datasets including data sources, update frequency, topic categories, and so forth. This advantage help students in finding suitable datasets for their research.

Some datasets have small numbers of variables and observations that can be less than 100 observations. The majority of the datasets focus on certain topics including spending disclosure, commercial impact tax, crimes, violations and so forth. Regarding the purpose of this course, it is difficult to find dataset with both categorical and numeric variables.

Bibliography

Demeter, T. (2020, September 14). *How to create an Origin-Destination Map in Tableau*. Retrieved from The Information Lab: https://www.theinformationlab.co.uk/2020/09/14/how-to-create-an-origin-destination-map-in-tableau/#:~:text=%20How%20to%20create%20an%20Origin-Destination%20Map%20in,we%20grab%20this%20calculation%20and%20bring...%20More%20

GeoPy Contributors Revision. (2018). *Welcome to GeoPy's documentation!* Retrieved from https://geopy.readthedocs.io/en/stable/search.html

Gillies, S. (2020, September 27). *The Shapely User Manual.* Retrieved from https://shapely.readthedocs.io/en/stable/manual.html

Hariharan, K. (2020, January 14). Retrieved from Time Series Forecasting — ARIMA vs Prophet: https://medium.com/analytics-vidhya/time-series-forecasting-arima-vs-prophet-5015928e402a#:~:text=From%20the%20experiment%2C%20we%20can,Model%20had%20RMSE%20of%2011.4%25.

Li, S. (2018, July 8). *An End-to-End Project on Time Series Analysis and Forecasting with Python*. Retrieved from towards data science: https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b

Prem. (n.d.). *iunera*. Retrieved from Top 5 Common Time Series Forecasting Algorithms: https://www.iunera.com/kraken/big-data-science-intelligence/time-series-and-analytics/top-5-common-time-series-forecasting-algorithms/

Rooy, N. (2016, December 14). *Calculate the Distance Between Two GPS Points with Python (Vincenty's Inverse Formula)*. Retrieved from https://nathanrooy.github.io/posts/2016-12-18/vincenty-formula-with-python/

Rooy, N. (2016, September 7). *Calculating the Distance Between Two GPS Coordinates with Python (Haversine Formula)*. Retrieved from https://nathanrooy.github.io/posts/2016-09-07/haversine-with-python/

Shao, V. (2020, September 15). Retrieved from Forecasting with a Time Series Model using Python: Part One: https://www.bounteous.com/insights/2020/09/15/forecasting-time-series-model-using-python-part-one/

Vincent, T. (2017, March 23). Retrieved from A Guide to Time Series Forecasting with ARIMA in Python 3: https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3

Vincent, T. (2017, April 4). *A Guide to Time Series Forecasting with Prophet in Python 3*. Retrieved from https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-prophet-in-python-3