

Group 2

Aishwarya Reddy Toom

Harshitha Suresh

Pavan Reddy Mamidi

Pavan Teja Nagiseti

Sasidhar Sirivella

BACKGROUND

Crime in the United States has been recorded since colonization. Crime rates have varied over time, with a sharp rise after 1963, reaching a broad bulging peak between the 1970s and early 1990s. Since then, crime has declined significantly, and current crime rates are approximately the same as those of the 1960s. (Source: wiki)

Overall, the number of crimes of all kinds reported in the United States remained fairly flat last year after a three-year increase, according to an annual FBI report. But while crimes against property were down, physical assaults against people were up, accounting for 61 percent of the 7,120 incidents classified as hate crimes by law enforcement officials nationwide. State and local police forces are not required to report hate crimes to the FBI but the bureau has made a significant effort in recent years to increase awareness and response rates. Still, many cities and some entire states failed to collect or report the data last year, limiting the conclusions that can be drawn from the FBI report.

Montgomery County is the most populous county in the U.S. state of Maryland, located adjacent to Washington, D.C. As of the 2010 census, the county's population was 971,777. As of July 1, 2018 The United States Census Bureau estimates the population of the county to be 1,052,567 residents with an increase of 8.3% since 2010.

As one of the most affluent counties in the United States, Montgomery County also has the highest percentage (29.2%) of residents over 25 years of age who hold post-graduate degrees. The county has been ranked as one of the wealthiest in the United States.

The Race of the county's residents is estimated to be 60.2% White (43.4% Non-Hispanic White), 19.9% African-American or Black, Hispanic or Latin, 15.6% Asian, 3.4% Two or more races, 0.7% American Indian or Alaskan Native, 0.1% Native Hawaiian or Pacific Islander

The Age and Sex is estimated to be :

- 6.3% Persons under 5 years
- 23.3% Persons under 18 years
- 15.5% Persons 65 years and over
- 51.6% Females

OBJECTIVE

Our study aims to find spatial and temporal criminal hotspots using a set of real-world datasets of crimes. We intend to perform extensive exploratory data analysis that will help us understand the factors affecting the crimes such as crime locations, timeshifts, response time etc., Based on the preliminary analysis and feature selection, we intend to predict the average response time

for a given crime and also predict crime rate based on temporal data. Finally, we provided an analysis/ suggestions by combining our findings that would help the officials to make decisions accordingly in order to improve the safety of Montgomery County.

DATA EXPLORATION

National crime data is an excellent tool in helping to improve relations with the community. Making crime data public increases transparency. While it can open criminal justice professionals to scrutiny, it also allows for a dialogue between law enforcement and the public they serve. As with any other government agency, the public deserves to be informed about how well police are protecting the community. Crime statistics are important in determining whether these initiatives are working, or if changes are needed. The data can show if crime is going up or down in the areas targeted.

Appendix – Overview of Data Included in Crime Dataset

Display Order	Column	Field Description
1	Incident ID	Police Incident Number
2	Offence Code	Code for an offense committed
3	CR Number	Police Report Number
4	Dispatch Date / Time	The actual date and time a Officer was dispatched
5	NIBRS Code	FBI NIBRS codes
6	Victims	Number of victims
7	Crime Name 1	Crime against society/person/property or others
8	Crime Name 2	Describes the NIBRS_CODE
9	Crime Name 3	Describes the Offence code
10	Police District Name	Name of District (Rockville, Wheaton etc.)
11	Block Address	Address in 100 block level
12	City	City
13	State	State
14	Zip Code	Zip code
15	Agency	Assigned Police Department
16	Place	Place description
17	Police Sector	Police Sector Name
18	Beat	Police patrol area subset within District
19	PRA	Police patrol are subset within Beat
20	Address Number	House or Business Number
21	Street Prefix	North, South, East, West
22	Street Name	Street Name
23	Street Suffix	Quadrant (NW, SW, etc.,)
24	Street Type	Ave, Drive, Road etc.,
25	Start Date / Time	Occurred from date/time
26	End Date / Time	Occurred to date/time
27	Latitude	Latitude
28	Longitude	Longitude
29	Police District Number	Major Police Boundary
30	Location	Location

Crime statistics Database including raw data and search functions of reported Country crime. The data presented 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.

Initially the dataset contained 197000 observations and 30 variables most of which are categorical and ordinal. Detailed explanation of variables is given in the table to the left. The data has crimes reported from 4 states that have Montgomery county present in those states, however we are considering only zip codes that are a part of Montgomery county belonging to Maryland. 51 zip codes were then considered for our further analysis. Few of the variables are unimportant and does not contribute much to

our analysis we intended to make, hence we added more data to our existing dataset from the latest census data with reference to zip codes such as population distribution, race distribution, unemployment details and median household income.

DATA PREPROCESSING

Feature Elimination

Initially we removed a few unnecessary variables based on general understanding, for example: CR_Number and NIBRS_Code which just give the police case number and FBI NIBRS Offence code respectively. We also removed a few variables with string values and no much significance like the block address and Address. All the pincodes belong to the state of Maryland, so we got rid of the state column as well. Agency (Assigned Police Department). Apart from that we also removed Address_Number, Street_Prefix, Street_Name, Street_Type and End_Date_Time, which we felt are insignificant to our analysis.

Data cleaning

Our data has crimes reported in 4 states. We are considering the zip codes which belong to Montgomery, MD. The data has variables which give the time related data for example, dispatch time which cannot be imputed and are irreplaceable. Imputing a dataset which consists of such variables will create noise. Apart from that, our dataset consists of around 160000 complete cases and only 15% missing values.

Data Integration

	count ▾	Zip_Code	Unemployment_rate	Median_Hous...	Population	Male	Female	White	Black
1	16813	20910	6.9	\$61723	42,868	20,746	22,122	21,454	13,506
2	15274	20902	2.4	\$75317	52,484	26,365	26,119	20,902	8,847
3	12528	20904	2.4	\$76928	57,035	26,306	26,306	15,588	26,246
4	12350	20906	2.4	\$65452	26,206	13,234	12,972	24,470	17,073
5	12323	20850	2.7	\$101621	51,568	23,986	27,582	28,000	6,640
6	12232	20874	2.4	\$80595	61,045	29,786	31,259	31,708	14,454
7	11387	20877	2.8	\$64770	38,885	19,378	19,507	18,302	7,210
8	10131	20878	2.8	\$109837	64,126	31,695	32,431	35,437	6,501
9	8265	20852	2.7	\$80493	46,904	21,570	25,334	28,913	4,766
10	7158	20901	2.4	\$77192	36,154	17,815	18,339	18,546	9,260

Here we would like to cite the Importance of a “Data Integration” Strategy:

1. Improve decision making.
2. Improve customer experience.
3. Streamline operations.
4. Increase productivity.
5. Predict the future.

By including an exhaustive set of features to our dataset we are enriching the available data set to gain more insights. We have included the above rows using public data from the data.gov website. In order to merge the data we explored all the variables and found that we can extract demographic data for all the zip codes available in the dataset. We have merged the new data with our initial data by performing a left join on the zip codes.

The included columns are:

1. **Unemployment rate** - According to a research paper, A one percent increase in the unemployment rate will increase the crime rate by 14.3 per 100,000 inhabitants. We have the unemployment for different zip codes in our dataset, so we would like to explore the relationship between `unemployment_rate` and crime rate in Montgomery county.
2. **Median Household Income** - According to aristotle, poverty is the parent of crime. So we would like to use this variable in our analysis. For example, in the zip code 20910, the unemployment rate is 6.9% and median household income is 61k. This tells us that the low household income is due to high unemployment rate. This can be the reason why 20910 has the most number of crimes.
3. **Population** - We included this to calculate the crime rate in a particular zip code. We would like to use this to discover patterns in the dataset.
4. **Male and Female** - In 2014, more than 73% of those arrested in the US were males. We want to see if the same applies to our dataset as well.
5. **Race** - We want to see which race group is committing the most number of crimes by taking the majority rate in a zip code into consideration.
6. **Location coordinates of police stations** - We added the location coordinates of police stations to get the distance between the crime scene and police stations. The distance can be used to see the relationship between distance and the response time.

FEATURE ENGINEERING

Apart from the remaining variables in the dataset we have created new variables from the available ones to help our analysis.

1. **Crime rate**: We have created a new column `crime_rate` which is a count of `incident_id` grouped by zip code divided by population.
2. **Start date and time**: We broke down the Start date time into start date and start time using `datepart()` and `timepart()` functions.
3. **Dispatch date and time**: We broke down the dispatch date time variable into dispatch date and dispatch time using `datepart()` and `timepart()` functions.
4. **Response time**: Dispatch time - Start time. This tells us how quickly the crime was addressed by the police.
5. **Time Shift**: We split the start time of all the incidents into 4 categories, 0:00 to 6:00 is time shift 1, 6:01 to 12:00 is shift 1, 12:01 to 18:00 is shift 3 and 18:00 to 24:00 is shift 3.
6. **Distance_miles**: Distance between police station and crime scene using `geodist()` function.

7. **Quarters & Seasons:** We categorized the start date of the crime into different quarter and seasons based on the months in which they took place (For example: September to November is taken as fall quarter)
8. **Place:** We took substring of the place where the crime took place and categorised them into places like residence, street, parking etc.

DISCREPANCIES IN THE DATA

start_date	dispatch_date	start_time	dispatch_time	timeshift	Response_time_min ▲
06MAR2017	10MAR2010	13:00:00	17:31:19	12pm-6pm	-3676048.683
20DEC2019	08MAY2013	1:00:00	18:44:18	12am-6am	-3479415.7
11SEP2016	20SEP2010	22:00:00	15:38:23	6pm-12am	-3143901.617
31MAY2017	18SEP2011	15:15:00	20:48:21	12pm-6pm	-2997746.65
12OCT2018	05FEB2013	0:00:00	1:50:50	12am-6am	-2987889.167

1. One thing we noticed in the dataset is that the dispatch date time when the officer was dispatched is before the start date time of the crime. This kind of misentry is seen in 2200 records.
2. Apart from that we could also observe that for a few entries the dispatch date time is as long as 4-5 years which is again a misentry. For this reason we are considering the entries which have a difference of one day between start date time of the crime and dispatch date time.
3. The other discrepancy is that crimes against society and property like drugs/narcotics violation, robbery etc., which can affect a group of people but have been reported by one person have one victim in the entry, like discussed above.

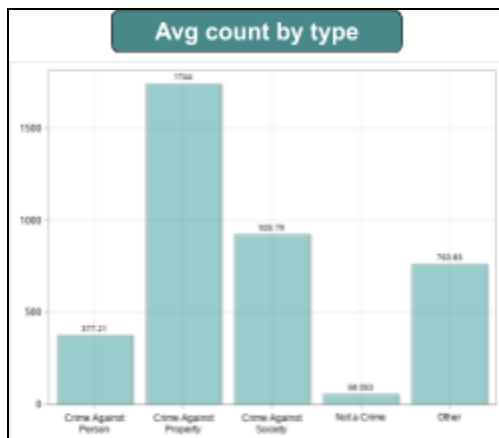
These come under the challenges we have faced while analyzing the data and made us filter our data by removing these discrepancies which could have affected the analysis.

TREATING MISSING VALUES

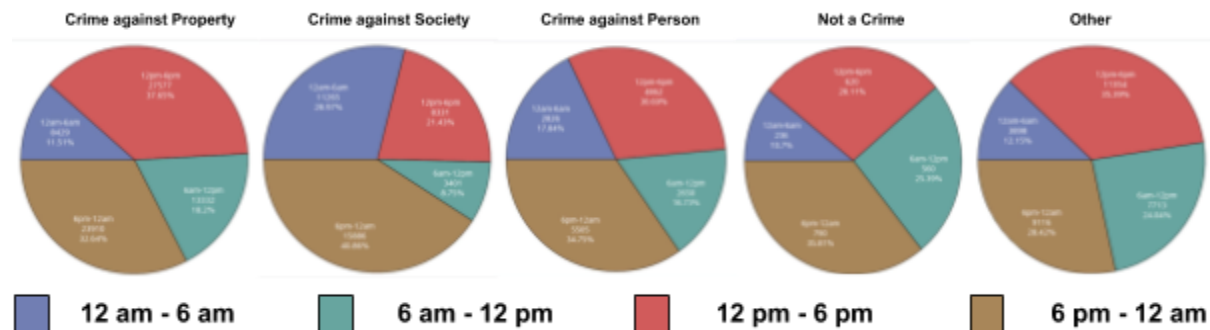
When we were calculating response time, which is a continuous variable in minutes from dispatch date and time start date time we came across wrong entries as discussed above in the discrepancies. As a result of these, the calculated response times were negative and we had to encode them as NAs. Whenever necessary, we imputed these response times, using mean imputation technique.

EXPLORATORY DATA ANALYSIS

Analysis by Type of Crime (CRIME 1)



mean_res_time	Crime_Name1
102.81108413	Crime Against Person
472.37000532	Crime Against Property
28.595625712	Crime Against Society
383.18311977	Not a Crime
174.66609273	Other



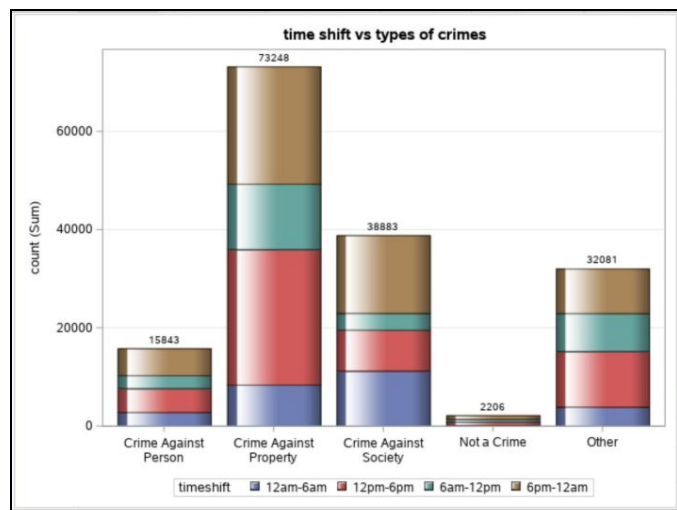
Crimes have been categorized into 4 types in our dataset- Crimes against Property that includes theft from buildings and vehicles, Crimes against Society such as identity fraud, drug violations, shoplifting, trespassing and misconduct, Crime against Person such as sexual assault, abuse, and rapes. Crimes reported such as lost property, mental illness are categorized under Others.

We observe that the majority are Crime against Property (73248). As you can see from the pie charts at the bottom of the dashboard, the number of crimes committed against property increases hourly from 12 pm and peaks through the afternoon and evening hours till 6 pm and then drops to a low point from 12 am to 6 am. Similarly, crimes against person and other crimes also increase from 12 pm - 6 pm on any given day. In contrast, the crimes against society drop to a low point during 6 am - 12 pm. We can see that crime against society mostly happens in the time frame 6pm to 12 am, this explains that crimes like drug narcotics offences, weapon law violations etc, happen usually in the night. According to our data from Montgomery, on average, 911 calls are responded to in about 28 minutes for the crimes against society, 472 minutes for the crimes against property and 102 minutes for the crimes against people.

The major takeaway here is that one would expect that, given the number of crimes against property (CAP), the response time would be least but on contrary to our expectation, the response time seems to be the highest for CAP. This could be because, on an average most crimes happen in the evenings (6pm to 12am). Even though the number of crimes committed during afternoons are more for CAP the average response time is high, which could probably be due to a significantly high response time for those crimes that are committed in the evening.

Analysis of Mean Response Time against different times of a day

mean_res_time	timeshift
171.4245304	12am-6am
258.8356155	12pm-6pm
259.91504938	6am-12pm
287.73731707	6pm-12am



As hypothesized before, the mean response time for Crime against Property is more because most of these crimes are committed between 12pm to 6pm and 6pm to 12am. As you can see above, the mean response time for these two time zones is significantly high.

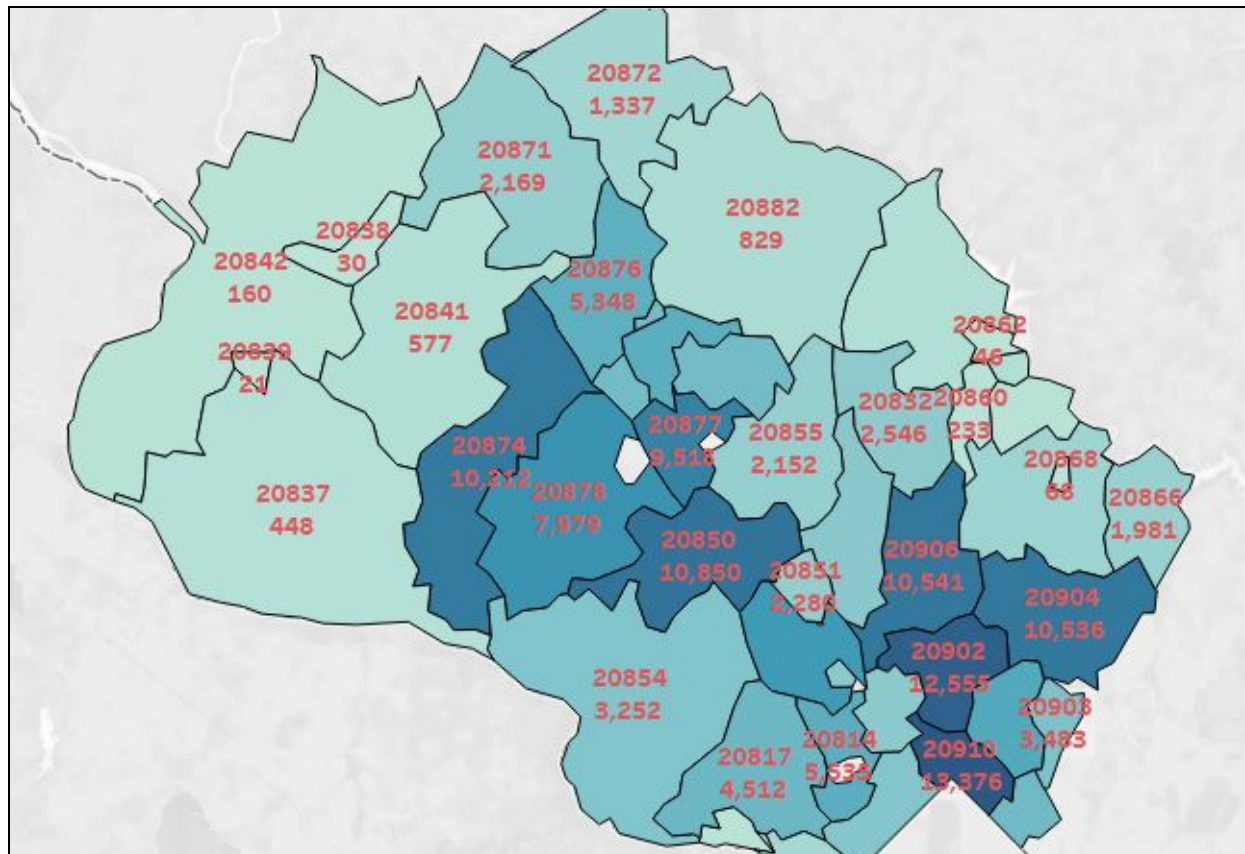
On an average, the standard response time for 911 calls is 171 minutes for the crimes reported between 12 am to 6 am. As we can see above the response time on an average for the time shift (12am-6am) is around the same. As seen in the clustered bar chart above, the number of crimes reported were significantly low during this time period. So, it is quite evident that the less number of reported crimes are the reason for the least response time during 12 am - 6am.

In summation, the most number of crimes are committed are of type “Crime against Property” which usually take place around the time shifts (12 pm-6pm & 6 pm to 12 am) and the average response is more for crime against property for these time shifts. As we can see in the pie charts, generally most crimes are committed in the time shift 6pm - 12am(shown in brown) across all types of crimes. **Hence, it would be suggested that Montgomery County Police should take initiatives that aim to reduce the response times for these time shift 6pm to 12 am.**

This analysis throws light on the issue of lack of quick response time(s) which helps us an understanding of why US citizens prefer to have some means of defense, like a firearm, more immediately to hand. According to the Small Arms Survey 2017, the United States has the highest estimate of firearms in civilian possession (393,347,000).

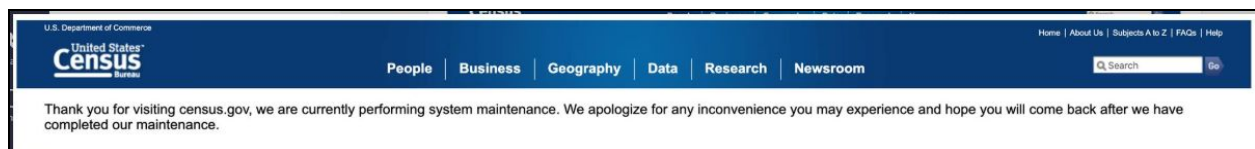
Analysis based on Zip Codes

Each county is an aggregation of different areas that are represented by unique ZIP Codes. An exploratory data analysis based on ZIP Codes will give us an understanding about which zipcodes are more crime prone regions across different geographical areas which fall under Montgomery County.



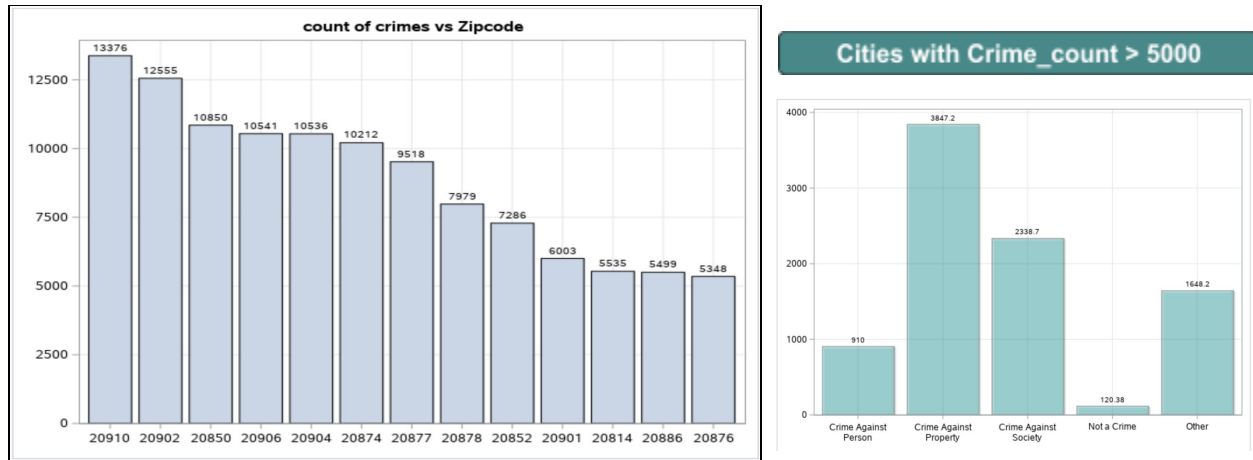
Variability of number of crimes at different ZipCodes across Montgomery County can be seen here. We tried to plot Choropleth map in SAS to represent the statistical crime data in Montgomery County through various shading patterns based on ZipCode, but we couldn't download the county maps (Server down in census website). So, we used Tableau to create this geographical map.

Error message when downloading the county maps:

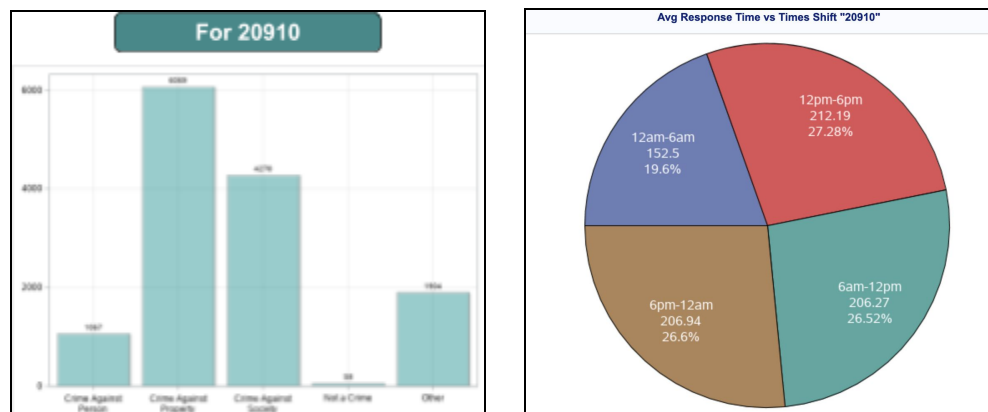


The graphs to the below represents zip codes that have more than 5000 crimes reported. From the graph we can see that 20910 has the highest crimes reported estimating to 13000. The second highest number of crimes reported is for zip code 20906 and 20904 estimating to 10500.

Going furthermore deep, we can analyze the type of crimes that are committed more in these zip codes.

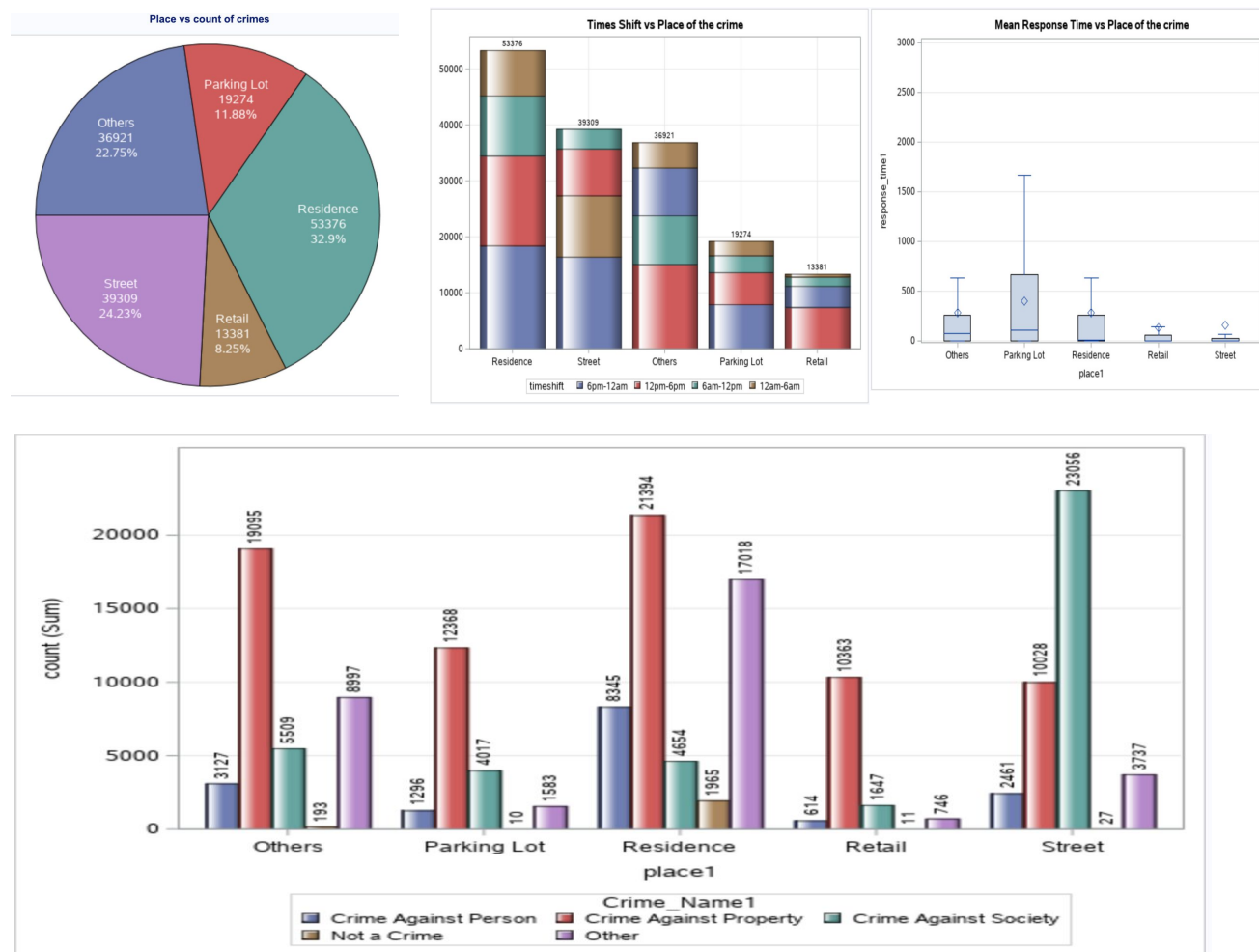


From the below graph, we can see that the majority of crimes committed in 20910 is categorized under crimes against property and society. According to recent updates, job growth in 20190 is positive resulting in a better standard of living. Based on this we can analyse that people who are well off are victimized more often and from our dataset, we see that unemployment rate is also high with 6.9% and the median household income is not significantly high from which we can infer that, unemployment and education are significant factors for crimes to happen.



20910 has the highest number of crimes reported among all the zip codes in our dataset. But the average response time of 199 minutes is comparatively less when compared to the other zip codes. Number of crimes reported are fairly the same in all the time shifts and the response time for the time shifts are also fairly the same. From this also we can say that **this is a well managed zip code because even though the frequency of crimes is high, response time seems fairly good.**

Analysis based on Crime Location



The above graphs explain the location where crimes have been committed. Violent crime is geographically concentrated in particular neighborhoods and in more localized areas known as hot spots. Analyzing the location of the crime will give us more insights about the locations that are not often safe in Montgomery County. We have categorized place variables into 5 subcategories based on where the crimes are reported. For example, Residence - Apartment/Condo, Residence-Yard, Residence- Shed etc is been categorized as Residence. Parking Lot- Garage, Parking Lot- County, Parking Garage- Commercial etc are categorized as Parking and Others include places like Restaurants, theatres, Library, Construction sites etc. We can see that the majority of the crimes reported belong to residences, streets and others. The next graph explains the type of crimes committed at these locations. Around 32% of the crimes happen at residences. Around 70% of the crimes that were reported at residences were committed during the time period of 12 pm -12 am.

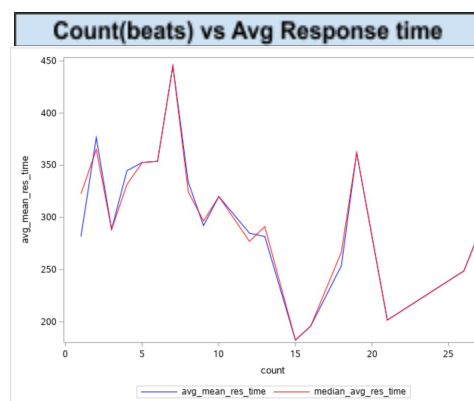
Residence, Street and Parking Lots are the places where most number of crimes have been committed during the time shift 6 pm - 12 am. The mean response time of police for the crimes that were committed at Parking Lots is the highest among all other crimes and this could be because of the reason that the number of crimes committed at Parking Lots being the second least. Crime against property is the most common type of crime here. It includes theft, destruction of property etc. The police response time is the least in case of Retail crimes, even though the number of crimes committed there was the least with the majority of them being committed during 12 pm - 6 pm. This could be because of the increased security and patrolling at retail environments like shopping malls etc.

From graph 3 we see that most of the crimes belong to Crimes Against Property (CAP) from which we infer that crimes like theft, burglary, breaking usually happen either at homes or from vehicles. So it makes sense that crimes reported at residence have category CAP in majority. Similarly around 24.23% of crimes happen on the streets. Furthermore, for streets, majority of the crimes committed belong to Crimes against Society which makes sense because on the streets, crimes like misconduct, drug violations, prostituion, trespassing happen more often. The mean response time of police for the crimes in Streets were the lowest among all and it might be because of high police patrolling on the streets. Around 22.75% of crimes is categorized as Others with type of crime to be Crime against property (CAP). Crimes like traffic offenses, animal bite, suicide attempts, credit card frauds comes under Other crimes. If we see locations grouped under others such as Library, Restaurant, theatres- all these are public places where people often visit. Hence, it supports the analysis why CAP is majority for Others. Parking Lot and Retail also have CAP as the majority type of crime from which we can draw the same inference as Others.

Secluded places like Residence and Parking Lot exhibit the same behaviour in terms of the “crimes committed” and “time shifts”. The crimes committed in these areas constitute 45% of the total crimes. The mean response times for these places are significantly higher than crimes that involve public places. Public places are more unsafe between 6pm-12am because the majority of crimes committed on streets are of type Crime against Society and the average response time is less because of effective police patrolling on streets. We can infer that **there is need to increase for security measures to be taken in secluded areas around the time 12 PM to 12 AM.**

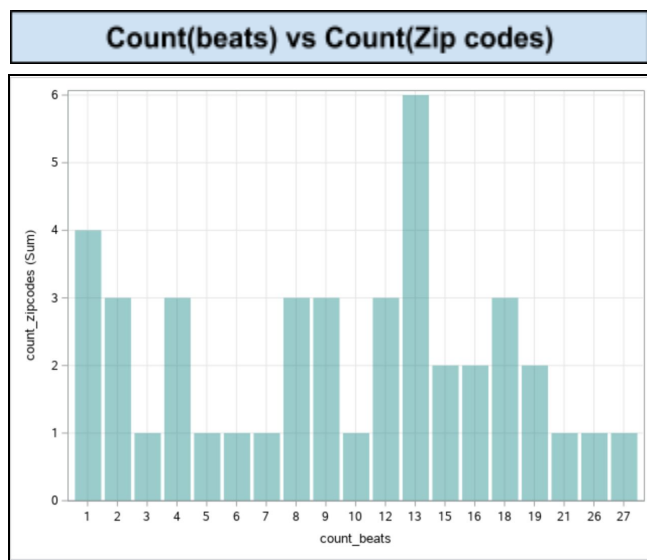
Analysis based on Beat(s)

In police terminology, a beat is the territory and time that a police officer patrols. Beat police typically patrol on foot or bicycle which provides more interaction between police and community members. So the number of beats in a given zip code are synonymous to the size of the police force for that particular zip code. As in higher the count of beats in a zip code higher the police force. In the analysis that follows, we

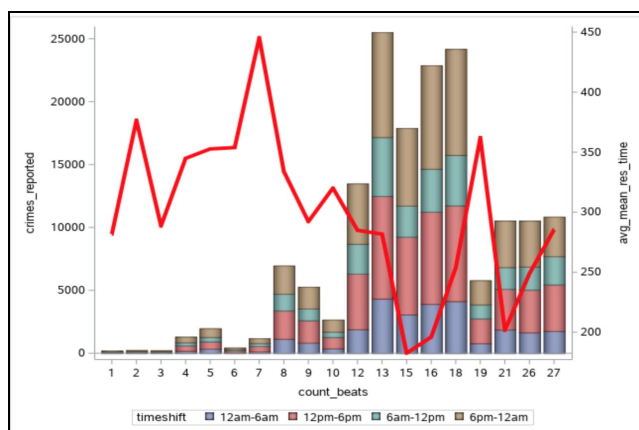


seek to understand the effect of the number of beats on the crime patterns in Montgomery county.

The general expectation would be that the response time should be low wherever the number of beats is high. If we observe the graph to the right that represents the average mean and median response time for the number of beats. It seems so that the general trend of the graph represents the same hypothesis but there are a few anomalies (around count_of_beats = 20) such as response rate is high even where the count of beats is high whereas in a few cases the response time is very high where the count of beats is very low (around count_of_beats = 7). To understand the number of zip codes under each bucket, we plotted a graph that represents (the graph to the right) the number Zip Codes bucked by the number of beats within the zip code. As you can see in the plot, there are around 20 Zip codes that have more than 10 beats in summation, which represents a significantly large police force. As we can see the peak in the graph, there are 6 zip codes with 13 beats. When we see the graph to the right, there are two ZIP Codes with the number of beats equal to 15, these ZIP Codes have a very less response time. Even though at 18 on xaxis, There are three ZIP Codes that have a high response time. Also, there are some zip codes which have almost 27 beats. So we wanted to understand what is the crime rate in these zip codes that have a high number of beats and how quick the crimes are addressed. Furthermore, we planned on analyzing it by the number of crimes reported for all zip codes.



In order to answer this we took the number of crimes into account. The graph to the right shows the number of crimes happening in areas with respect to the number of beats in different time zones. The red line represents the response time across different zip codes bucketed by number of beats. Here we can see that the zip codes with number beats between 13 to 17 are functioning efficiently because the response time is less though the number of crimes is more whereas if we see for the zip code with number of beats in between 4 to 7 have significantly high response rate though the number of crimes are very less.

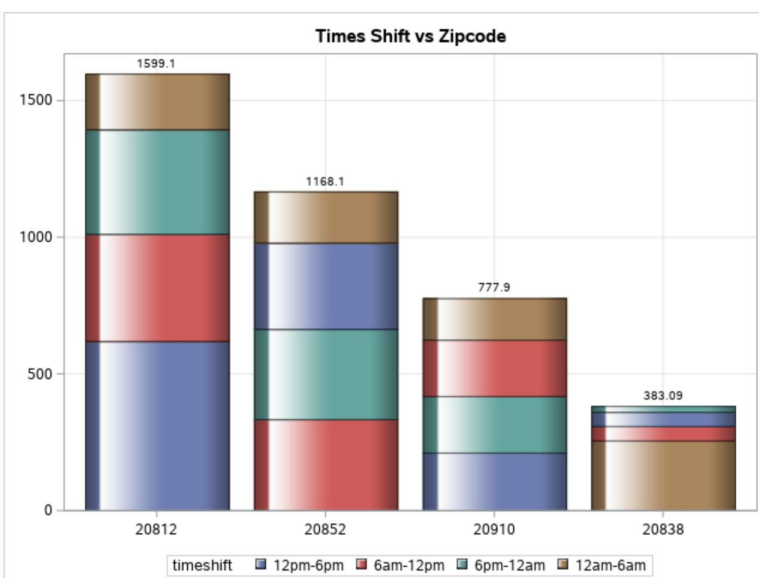
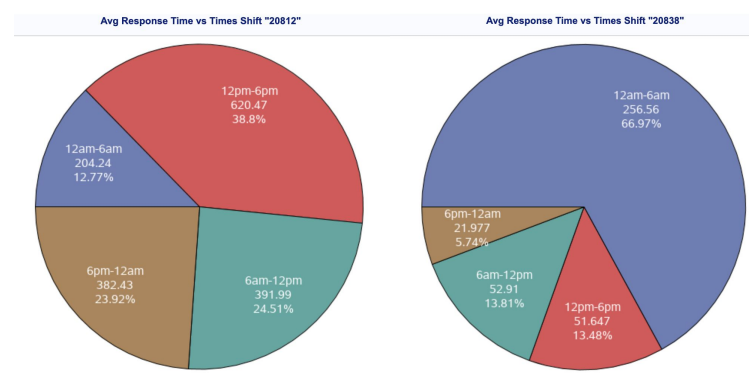


Hence one can infer that **Montgomery County has made efforts in the past to reduce response time in areas with high crime rates** but it is also evident that **areas with low crime rates have a significantly high response time, which needs to be worked on**. Also we can see that for the bucket with 19 beats, the response time is significantly large. Therefore, it is suggested to Montgomery county to emphasize on the beats responsible in the zip codes 20832 and 20854 to function promptly where the number of crimes is very high.

Analysis based on response time (low crime count regions)

Response time indicates the time taken for the police to attend the crime location. The assessment and coding of an incoming call, directing the call to the appropriate agency, assignment of the call to the specific units and that unit's arrival at the scene (even assuming the responder has an accurate location) are all the factors why the response time varies. We have considered zip codes 20812, 20838, 20910 with total number of crimes reported, average

response time, timeshifts for count of crimes as well as for the response time for our analysis.



Zip code 20812 and 20838 have the same number of crimes(30) reported. However, the response time is significantly different from each other. We wanted to analyse why the response time varies even though the same number of crimes are reported. Average response time for 20812 is 431 minutes and when we look at that bar graph that represents response time for each time shift for the respective zip codes, we can see that the response time is more between 12:00pm- 6:00pm. From the pie chart plotted for 20812, we can see that the time in which more numbers of crimes are reported is in the same time shift. We can interpret that the response time is more because this particular time shift is at the peak time of the day. High

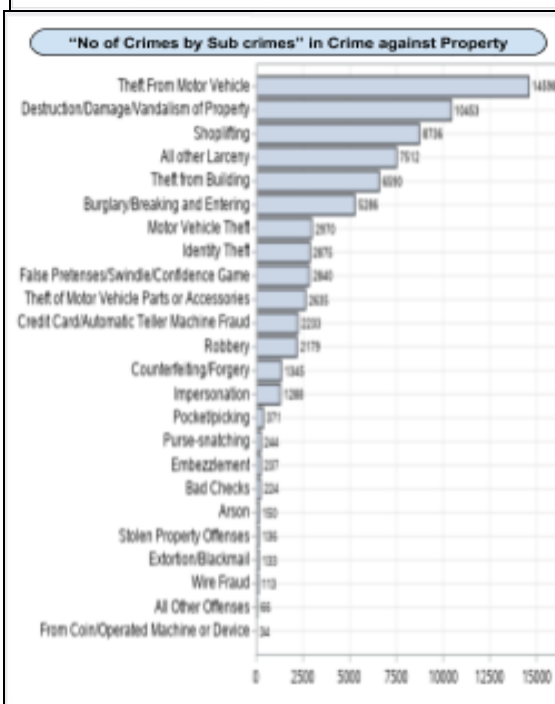
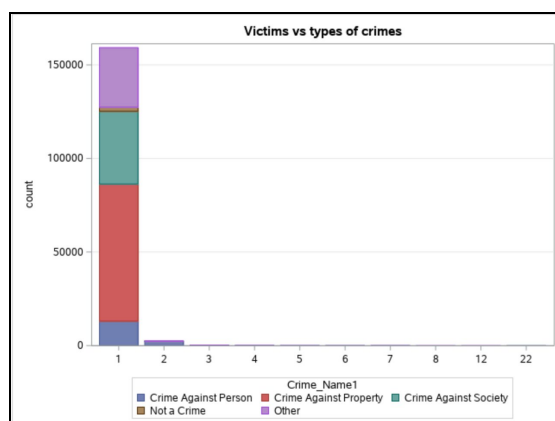
response time may be due to the multiple cases that have to be addressed by the officers and also can be due to traffic which is the reason why the police officer reaches the location late. This is something the police department of 20812 has to work on so that they resolve the crimes

at the earliest. On the other hand, for 20838 which also has the same number of crimes reported has a very less average response time of 94 minutes. From the bar graph we see that between 6:00pm:12:00am is the timeshift in which more crimes are committed and the time shift in which the response time is least is between 6:00pm:12:00am. This justifies why the average response time is less for this Zip code. We can say that this zip code is well managed by its police department in responding to the crimes reported.

Based on the above analysis we can see that **external factors play a role in the variation of response time in regards to particular time shifts**. As we've said before, there is a need for initiatives that need to be taken by Montgomery county in the areas with less crimes rates as the response times were notably large. Our analysis suggests that 20838 is a well managed zip code which has a low crime count and can be considered as a **reference for other zip codes in the low crime count zone to follow the same measures as taken in 20838**.

Analysis based on number of crimes

We did an analysis to see what kind of crime is happening when people are going alone and when they are in a group. When there is one victim, we can see that the most number of crimes are crime against property - which includes crime like theft from vehicles, apartments, burglary etc. We can even see that the crime against society and other crimes are also high. From the above graph, we can see that Crime against Property and society is committed on a larger scale in the zipcodes that we have considered.

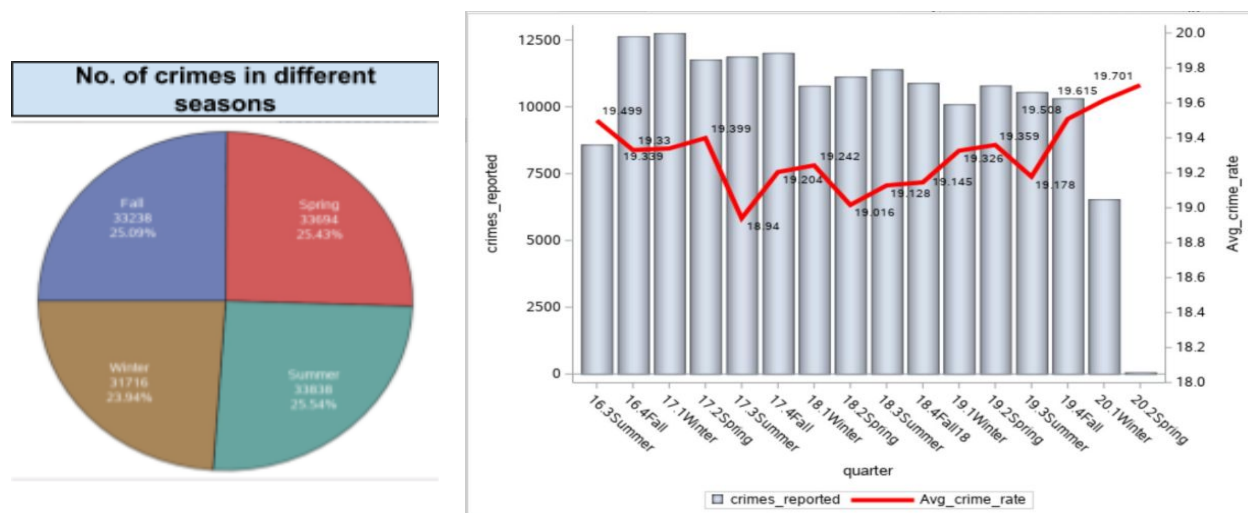


Over the years, as per the report produced by the Montgomery Police Department, Crimes against persons have decreased over the years and Crimes against Property and Society has increased overtime justifying the reason for most number of crimes committed belonging to these two categories. We can see from the earlier graph that the crimes happened mostly when there was one person. So we did an analysis on what kind of crimes happened when there were more than 2 people involved. We can see that the majority of them were Crime against person like intimidation, sexual assault, rapes etc.

If we see the breakdown of crime against property which is happening the most, we can see that theft from motor vehicles and destruction of property is the highest. Apart from that theft of motor vehicle parts and shoplifting is also significant. This tells us that the people in Montgomery should make sure to keep themselves safe while in vehicles and have enough security for their property. Here if we think about the kinds of crimes in crime against property we can notice that even though the crime happened against an entire household or shop, only one person reports it and only that person will be considered as the victim and not all the people who live in that particular house. The same thing applies to crimes like drug/narcotics violations, driving against the influence of alcohol where victims can be more but only one person reports it or the police files the case.

The key takeaways(as represented by the given data) is that most crimes that are committed involve a single victim and a majority of which are reported as crime against property which seem to have a very high response time in general. These crimes mostly happen during the times 12pm-12am. A key inference is that **security measures in regards to properties like house and motor vehicles need to be improved in this county** to assure safety of its citizens and a decreased crime rate.

Analysis based on seasons

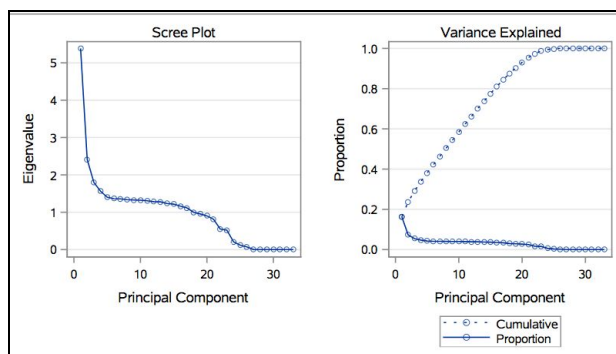


The temporal analysis of crime data produces analytics that describe patterns in criminal activity of Montgomery County based on time.(From July 2016 to March 2020). It allows us to provide decision-support to law enforcement agencies that want to optimize their tactical crime fighting. We tried to use temporal data analysis to make inferences and draw correlations that we can use to monitor and predict what crimes are happening when and why. Assuming that the second half of 2016 has the same number of crimes as in the first half of the year, we observe that the number of crimes reported in all the years (2016-2019) were consistent and are around 50000. Over the years in our data, the number of crimes that have been reported are almost equal in all four seasons as can be seen in the below pie chart. We plotted the number of

100

[illegible]


reason for the high correlation among them. It can be seen that all the crime types are negatively correlated with each other. Also, the crimes reported at different places were also negatively correlated. Similarly, the number crimes reported in different time shifts and in different seasons were also negatively correlated among themselves. Shown to the right, is the results of PCA in SAS and interpretation of Principal Components that explain the contribution of variables. From the below graph we see a deviation in the elbow of the plot at 20 which implies that A total of 20 components together explain 93% of the variance in our dataset.



	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
Avg_distance	-0.14077	0.00272	-0.0088	0.00270	0.01917	0.02701	0.07209	0.00101	-0.0109	0.02404	0.04010	0.00362
Avg_Response_time	-0.02005	-0.04047	0.00025	-0.00054	0.01070	-0.00023	0.01700	0.01404	-0.0109	-0.00037	-0.02430	0.07304
Unemployment_rate	-0.04054	0.02708	-0.00549	0.00270	0.00005	-0.01010	0.11000	0.14017	0.13000	-0.02710	0.00012	0.00010
Race_Percentage	0.00000	-0.14000	0.01000	0.02000	0.01000	0.04000	-0.02000	-0.01000	-0.04000	0.00000	-0.00000	0.00000
White	0.01000	-0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Black	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Hispanic_or_Latino	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Population	0.01000	-0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Male	0.01000	-0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Female	0.01000	-0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
count_beds	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Crime_Name1_Crime_Against_Person	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Crime_Name1_Crime_Against_Property	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Crime_Name1_Crime_Against_Society	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Crime_Name1_Not_a_Crime	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Crime_Name1_Other	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
place1_Others	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
place1_Parking_Lot	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
place1_Residence	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
place1_Retail	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
place1_Street	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
timeShift_12am_6am	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
timeShift_12pm_6pm	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
timeShift_6am_12pm	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
timeShift_6pm_12am	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
season_Fall	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
season_Spring	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
season_Summer	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
season_Winter	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Race_Majority_African_American	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Race_Majority_Black	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Race_Majority_Hispanic	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000
Race_Majority_White	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.01000

Principal Component 1: Coefficient of Population distribution contributes more to the first principal component that explains around 41% of the variance. We can say that **PC1 primarily measures the general population index** in our data. So one can say that the population of a given zipcode is a major indicator of the crime_rate. The higher the population the higher the crime rate.

Principal Component 2: Absolute values for the coefficients of Racial distribution of white and blacks contribute more to the second principal component that explains around 60% of the variance. From this we can infer that **PC2 primarily measures a contrast between the races because of the presence of negative value for whites.**



Principal Component 3: Coefficients for the type of crimes namely Crime against Property contribute more to PC3 that explains around 62% of the variance. This PC **primarily measures the type of crime for which more number of crimes are committed.**

Principal Component 4: Coefficients for the place where the crime is committed contribute more to PC4 that explains around 50% of the variance. This PC **primarily measures the location where the crimes are committed.**

Principal Component 5 and 6 : Coefficients for timeshifts namely between 6:00am-6:00pm contribute more to PC5 and PC6. Hence we can say that these two PCs **primarily measures timeshifts in which the number of crimes committed are more.**

Principal Component 7: Coefficients for season fall and winter contributes more to PC7. It can be inferred that PC7 **measures crimes happening in fall season in the timeshift 6pm-12am.**

Principal Component 8: The Coefficient of the season-summer contributes more to the Principal Component 8. We can say that **Principal Component 8 primarily measures the crimes committed in the summer season.** The inference is that a certain amount of crimes have a pattern of being committed in the summer season.

Principal Component 9 : The coefficient of the season-winter contributes more to the Principal Component 9. We can say that **Principal Component 9 primarily measures the crimes committed in the winter season.** The inference is that a certain amount of crimes have a pattern of being committed in the winter season.

Principal Component 10 : The coefficient of the season-spring contributes more to the Principal Component 9. We can say that **Principal Component 10 primarily measures the crimes committed in the Spring season.** The inference is that a certain amount of crimes have a pattern of being committed in the spring season.

Principal Component 11 - Coefficients for unemployment rate and race majority of hispanics contribute more principal Component 10. We can say that **Principal Component 11 primarily measures the unemployment rate where the race majority is Hispanics.**

Principal Component 12 : Coefficients for type of crime namely Others and the place- others such as library, restaurant, construction site etc contribute more to Principal Component 12. From this we can infer that **PC12 primarily measures the location of the crimes that belong to the category type Others.**

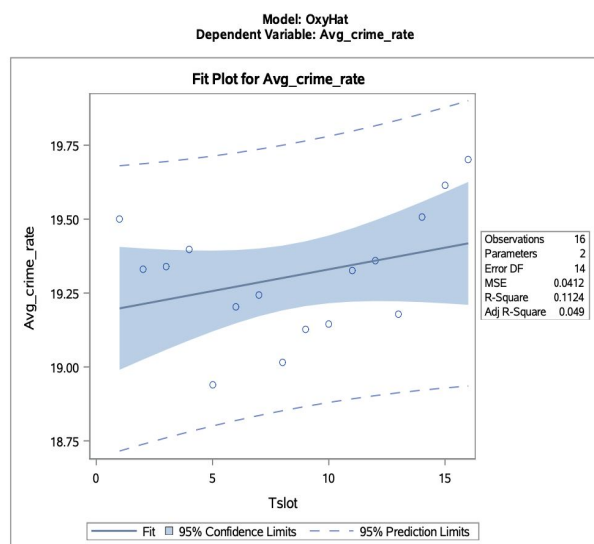
As we have seen from the analysis we've done so far, the principal components are to be considered as criterions of analysis in order of importance from top to bottom i.e... Principal component one is more important than principal component 2 which in turn is more important than principle component 3 and so on. Principal Component Analysis helped us to emphasize the variation and in bringing out strong patterns in our dataset and made data easy to explore

and visualize. Hence, we have built a linear regression model that will take into account all the above said factors grouped by the order of priority with average response time as the target variable.

REGRESSION ANALYSIS

Linear Regression : Crime rate estimation for Spring'20.

To start off with, we built a basic model to explore the linear regression method in SAS. Variables considered - Seasons (Independent variable) and Crime rate (Dependent variable)



We expect the crime rates to increase during warm seasons which is associated with the average crime rate. The above regression analysis result represents the crime rate values that lie on the linear regression line which is fit for the average crime rate values of the entire dataset. Based on this we predicted the crime rate value for the spring 2020 which is the 16th entry in the result. From this we can infer that crime rates have increased from winter to spring. This corroborates our results from temporal analysis from which we drew an inference that crime rate increase in spring season.

Regression Scoring Example Predicted Scores for Regression

Obs	Avg_crime_rate	quarter	Tslot	OxyHat
1	19.4994	16.3Summer	1	19.1982
2	19.3301	16.4Fall	2	19.2128
3	19.3391	17.1Winter	3	19.2275
4	19.3987	17.2Spring	4	19.2421
5	18.9397	17.3Summer	5	19.2568
6	19.2042	17.4Fall	6	19.2715
7	19.2424	18.1Winter	7	19.2861
8	19.0157	18.2Spring	8	19.3008
9	19.1276	18.3Summer	9	19.3154
10	19.1452	18.4Fall18	10	19.3301
11	19.3262	19.1Winter	11	19.3448
12	19.3594	19.2Spring	12	19.3594
13	19.1778	19.3Summer	13	19.3741
14	19.5081	19.4Fall	14	19.3887
15	19.6147	20.1Winter	15	19.4034
16	19.7014	20.2Spring	16	19.4181

Linear Regression : Response time estimation

In our second model, we want to predict the average response time based on a few predictor variables. We



grouped the data according to the order shown in the figure to the left where average response time is evaluated for a given particular season, at a certain timeshift, at a certain distance from the police station, at a given place and for a type of crime. In order to perform the data aggregation for input features, we have also taken the population estimates, the number of beats for every zip code and performed a left join with the table containing the data shown in the funnel to the left.

Number of Observations Read	9347
Number of Observations Used	9347

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	316871986	13777043	155.11	<.0001
Error	9323	828062474	88819		
Corrected Total	9346	1144934459			

Root MSE	298.02570	R-Square	0.2768
Dependent Mean	242.58221	Adj R-Sq	0.2750
Coeff Var	122.85555		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	B	203.18215	16.63686	12.21	<.0001
Avg_distance	1	-0.01275	0.02090	-0.61	0.5420
Unemployment_rate	1	-0.19483	2.57641	-0.08	0.9397
Race_Percentage	1	0.24327	0.19177	1.27	0.2046
Black	1	-0.00458	0.00128	-3.58	0.0003
Hispanic_or_Latino	1	-0.00112	0.00106	-1.05	0.2918
Population	1	0.01507	0.00722	2.09	0.0369
Male	1	-0.00645	0.01000	-0.64	0.5192
Female	1	-0.02369	0.00681	-3.48	0.0005
count_beats	1	1.21770	1.17865	1.03	0.3016
Crime_Name1_Crime_Against_Person	B	-91.78583	9.37317	-9.79	<.0001
Crime_Name1_Crime_Against_Proper	B	308.04838	8.68190	35.48	<.0001
Crime_Name1_Crime_Against_Societ	B	-133.29352	9.09783	-14.65	<.0001
Crime_Name1_Not_a_Crime	B	131.14661	14.50436	9.04	<.0001
Crime_Name1_Other	0	0	.	.	.
place1_Others	B	77.48678	9.38621	8.26	<.0001
place1_Parking_Lot	B	63.37761	10.08030	6.29	<.0001
place1_Residence	B	51.86532	9.07280	5.72	<.0001
place1_Retail	B	-63.37651	11.46526	-5.53	<.0001
place1_Street	0	0	.	.	.
timeshift_12am_6am	B	-12.67100	8.99020	-1.41	0.1587
timeshift_12pm_6pm	B	-4.75357	8.33987	-0.57	0.5687
timeshift_6am_12pm	B	-19.70900	8.66091	-2.28	0.0229
timeshift_6pm_12am	0	0	.	.	.
season_Fall	B	-7.95396	8.69557	-0.91	0.3604
season_Spring	B	-9.40195	8.82081	-1.07	0.2865
season_Summer	B	-9.79007	8.73153	-1.12	0.2622
season_Winter	0	0	.	.	.

The number of observations used are 9347. If we can see the analysis of variance, the p-value is less than 0.05 so we can successfully reject the null hypothesis (means are equal) and say that there is a relationship between the independent variables and the dependent variable (response time).

R-Square is the proportion of variance in the dependent variable (response time) which can be predicted from the independent variables (type of crime, place, distance, timeshift and season), which is 27% in our case.

The p-value for a few independent variables is less than 0.05 implies that they are statistically significant towards our dependent variable. We have also considered the other variables based on our analysis which we found to be important.

The intercept represents the minimum of average response time when there are zero values for all other variables. If we consider a statistically significant variable like crime against property, for every incident of crime reported the response time will increase by a minimum value of 308 minutes i.e... for a crime against property the minimum response time will come out to be (Intercept + Coefficient) ~500 minutes, given that all other variables account to no change in the average response time.

In our earlier analysis, the average response time for streets is found to be the least. This inference is supported by the regression model to the left as the coefficient of linear regression is 0. Also we can see that Crime against society has a negative

value for its coefficient. This means that the average response time is very less for the crimes committed against society which tend to happen mostly on the street as we've seen before.

Test case

If we see the table, the predicted values are the response times and there is a 95% probability that the response times fall between the given confidence interval range. Standard error of the mean tells you how accurate your estimate of the mean is likely to be.

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	8.67E-01	152.6044	12.9768	127.1670	178.0419	-151.7378
2	9.78E+00	83.9474	14.0353	56.4351	111.4597	-74.1641
3	7.48E+02	582.8124	12.7874	557.7462	607.8786	165.5542
4	8.11E+01	581.3647	12.9591	555.9621	606.7674	-500.2647

CONCLUSIONS

The following are the key takeaways from our analysis and models:

- ❑ It would be suggested that Montgomery County Police should take initiatives that aim to reduce the response times for the crimes that take place during the time shift 6pm to 12 am.
- ❑ It is suggested to Montgomery county to emphasize on the beats responsible in the zip codes 20832 and 20854 to function promptly where the number of crimes is very high.
- ❑ Secluded - Residence and Parking Lot exhibit the same behaviour; constitute 45% of the crimes- secluded; public- street > crime against society.
- ❑ We see a significant dip in crime rate during summers and this is because the census data gets updated every June resulting in change in the numbers of the total population.
- ❑ According to our data, most crimes that are committed involve a single victim.
- ❑ To assure security measures in regards to properties like house and motor vehicles need to be improved in this county
- ❑ Our suggestion to the citizens of Montgomery county is to stay vigilant during the months of June-August.

LIMITATIONS

- ❑ We added data regarding races to analyze how crime and race are related but the population for different races can vary over the years. But since we have considered the latest update, race majority is similar across different zip codes. Hence, we won't be able to get significant insights.
- ❑ This analysis is confined to Montgomery county and the results may vary for different cities in the united states
- ❑ We did not consider the imputed data in a few cases because this might bias our analysis.
- ❑ Intensive feature selection was not done as some of the coefficients of predictor variables in the model that estimates the average response time were zero.



REFERENCES

- <https://smartech.gatech.edu/bitstream/handle/1853/53294/theeffectsofunemploymentoncrime-rates.pdf>
- https://en.wikipedia.org/wiki/Estimated_number_of_civilian_guns_per_capita_by_country
- <https://www.policeone.com/community-relations/articles/do-police-response-times-matter-8pDUo0L5OxmerFAb/>
- <https://www.zmescience.com/science/weather-crime-connection-04234/>
- <https://support.sas.com/en/documentation.html>