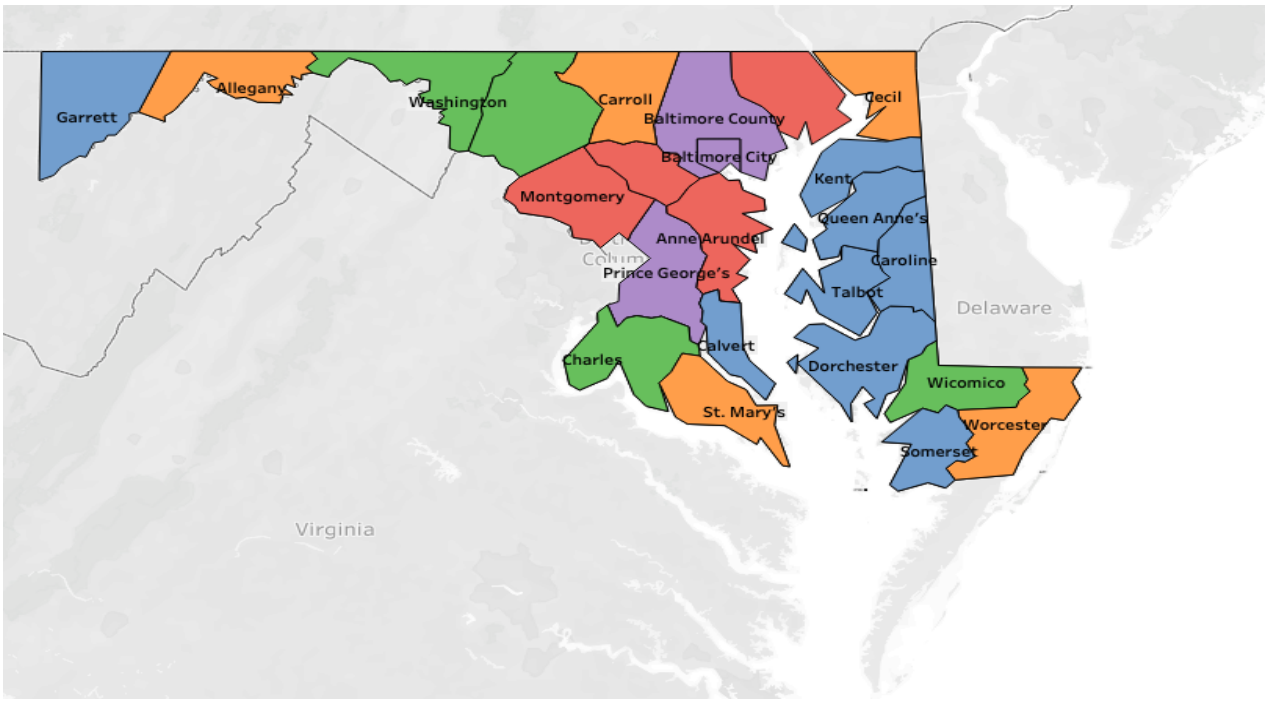
IS 733 Data Mining
A Project Report on
Maryland Crime Rate Trend Analysis



By
Akshay Negi
Ninad Sawant
Nishigandha Karle
Prerana Prabhakar
Vamshi Krishna Yenmangandla

Under the Guidance of Professor
Dr. James Foulds

Information Systems Department
University of Maryland, Baltimore County

Table of Contents

1. Abstract 2

2. Introduction 2

3. Background and Related Work 3

4. Methodology 3

4.1. Data Re-scaling 3

4.2. LMER 4

4.2.1. LMER used to predict crime rate 5

4.2.2. Using the values of LMER 6

5. Experimental Results 8

6. Conclusion 12

7. Future Scope 13

8. References 13

1. Abstract

Analysis of crime is a systematic approach for the identification and analysis of criminal patterns and trends. Crime data analysts will help Maryland law enforcement departments speed up the crime resolution process with the growing sources of computerized systems. Using the data mining concept, we can analyze previously unknown, useful information coming from unstructured data. Predictive policing involves detecting offenders using analytical and predictive techniques and it has been found that doing the same is pretty much successful. Owing to the rising crime rate in Maryland state over the years, we will have to tackle a large amount of crime data collected in warehouses that would be very difficult to manually examine, and even now a day, criminals are technologically advancing, so new technology need to be used to keep the police ahead of them.

The focus of this study is to compare various approaches to the issue of predicting the number of crimes in different areas of Maryland. In this study, we have used LMER (Linear Mixed Effect Regression). The predictive factors used in this project have been selected using feature selection technique. This approach allowed us to categorize the crime rate based on the locations and predicts which county in Maryland has the highest crime rate.

2. Introduction

Maryland is a state in the Mid-Atlantic region, with Baltimore as the largest city and capital being Annapolis. The current population of Maryland is estimated to be around 6 million, placing Maryland in the 19th position in terms of population. Crime in Maryland has been categorized into various types depending on the areas such as rape, robbery, theft, assault and so on. Even though the number of violent crimes has been declined, Baltimore city in Maryland remains higher than the national average. Under the U.S. News and World Report, Baltimore, MD is one of The United States most dangerous cities. Collecting, reviewing, and acting on in-city crime data has been a high priority over the past few years. If factors that lead to crime can be better understood, so anticipating or avoiding these incidents may be a long way off.

Most states have shared in the decline in violent crime appeared in years. However, in studying crime patterns, it appears that there are states that have not engaged in the same reduction stages, or that still have exceptionally high rates of violent crime in which one of them is Maryland.

MARYLAND VIOLENT CRIMES
POPULATION: 6,042,718

Crimes	MURDER	RAPE	ROBBERY	ASSAULT
Report total	490	1979	9716	16,135
Report per 1000	0.08	0.33	1.61	2,67

We utilize data mining methodology to perform an exploratory data analysis on Maryland's crime rate, which has a higher violent crime rate (4.7) than the national average (3.7). The goal is to

identify 1032 rows and classify crime patterns, crime based on locations and figure out which of Maryland's counties fall into the highest crime rate categories. The first thing that matters most is the safety. Given that Maryland is known for its violence, predicting crime will help in demonstrating which county has the highest risk factor. Our motive is to raise consciousness among people about the factors that cause unsafe behavior. Analysis of crime can be provided to the Police department which will help them increase the security or avoid crimes from happening. By understanding the patterns of criminal behavior, many criminal investigations can be solved.

3. Background and Related Work

We surveyed work related to our own, both to point out the many contributions of previous researchers and to place our contributions in the proper context. Numerous techniques exist for analyzing crime data. In the past, representations were strictly based on the historic crime records in the database. Results show crimes like robbery, murder, highway robbery and burglary that are higher in regions which lacks proper security and also less inhabited.

One of the popular Visualization techniques is the use of heat maps. The crime prone areas can be graphically represented using a heat map which indicates level of activity, usually darker colors to indicate low activity and brighter colors to indicate high activity. However, in some graphical representations the majority of the projects took into consideration, Crime Rate vs Population in the respective areas or cities which seems little unfair. The underlying reason is because bigger the city, more would be the population and the analysis would not be precise.
In our Project we have Visualized “Crime Rate vs Population Density” instead of “Crime Rate vs Population”. We have performed a similar analysis in this project and used other concepts like LMER method.

4. Methodology

This project utilizes a dataset focused on Crime in Maryland State. The dataset was available in a comma-separated values (CSV) file, originally there were 1032 instances with 38 attributes, including Jurisdiction, Year ranging from 1975 to present, Population, Rape, Robbery, Assault, B&E Larceny, Theft, Grand Total, Percentage Change, Overall Crime rate per 100,000, Percent change. The raw data have been transformed in a useful and efficient format by doing Data cleaning using python. All the blank rows were taken care of and made data error free.

4.1. Data Re-scaling

To be compatible with analysis of the crime data set, an additional attribute was added to the crime data set. We rescaled the YEAR variable, before building a linear mixed-effects regression model. Regression models give best results when the intercept is near zero, but YEAR starts at 1975, in the raw data. The model will fail to converge, if we use YEAR without re-scaling. We created a new variable, YEAR_R that starts at zero rather than 1975.

```
# A tibble: 6 x 4
  JURISDICTION     YEAR POPULATION crime_rate
  <chr>           <dbl>      <dbl>      <dbl>
1 Allegany County  1975      79655       178.
2 Allegany County  1976      83923       104.
3 Allegany County  1977      82102       155.
4 Allegany County  1978      79966       128.
5 Allegany County  1979      79721       138
6 Allegany County  1980      80461       148.
>
```

Before Rescaling

```
> head(crime_use)
# A tibble: 6 x 5
  JURISDICTION     YEAR POPULATION crime_rate YEAR_R
  <chr>           <dbl>      <dbl>      <dbl>  <dbl>
1 Allegany County  1975      79655       178.      0
2 Allegany County  1976      83923       104.      1
3 Allegany County  1977      82102       155.      2
4 Allegany County  1978      79966       128.      3
5 Allegany County  1979      79721       138       4
6 Allegany County  1980      80461       148.      5
>
```

After Rescaling

4.2. LMER

LMER is a linear mixed model, also called a Multilevel model or hierarchical model depending upon the context. This is a type of regression model that considers both fixed (variation that is explained by the independent variables of interest) and random effect (variation that is not explained by the independent variables of interest). This model is also called a Mixed model because it includes a mixture of two effects. These random effects essentially give structure to the error term ϵ.

LMER stands for linear mixed effect regression model. This generic function fits a mixed-effects model with nested or crossed grouping factors for the random effects. The dataset is hierarchical in nature so LMER can be used. LMER deals with two types of parameters:

1. Fixed: This parameter never varies.
2. Random: These are parameter which are random by nature

This is similar to linear regression, where the data is assumed as random variables and the parameters are fixed effects.

4.2.1. LMER used to predict crime rate

Considering the Maryland data, LMER is used to predict the value of crime rate based on two factors, that is year and Jurisdiction. The equation is written as follows:

crime_rate ~ Year_R + (Year_R | Jurisdiction)

crime _rate which is the left part of ~ is the variable to be predicted. The right-hand side has two components: fixed effect and random effect. Year_R is the fixed effect and (Year_R|Jurisdiction) is the random effect. The fixed effect here is considered as Maryland state and the Random effect is considered as the 24 counties in the state of Maryland.

This LMER function is coded in R so the function. The below code snippet demonstrates the code.

```
#Building LMER- linear meixed effecr regression model
# load the lmerTest package
library(lmerTest)

# Build a lmer and save it as lmer_crime
lmer <- lmer(crime_rate ~ YEAR_R + (YEAR_R|JURISDICTION), data=crime_use)

# Print the model output
lmer


# Examine the model outputs using summary
summary(lmer)

coef(summary(lmer))
```

The output generated by the function is as follows:

```
> # Examine the model outputs using summary
> summary(lmer)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: crime_rate ~ YEAR_R + (YEAR_R | JURISDICTION)
   Data: crime_use

REML criterion at convergence: 13461

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.4418 -0.4827 -0.0538  0.4393  6.4565

Random effects:
 Groups       Name        Variance  Std.Dev. Corr
 JURISDICTION (Intercept) 179502.58 423.677
              YEAR_R          15.93   3.991  -0.68
 Residual                  23323.87 152.722
Number of obs: 1032, groups:  JURISDICTION, 24

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) 533.5250    86.9862  23.3456   6.133 2.77e-06 ***
YEAR_R       -1.7534     0.9003  23.1973  -1.948   0.0637 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
       (Intr)
YEAR_R -0.655
convergence code: 0
Model failed to converge with max|grad| = 0.122593 (tol = 0.002, component 1)
```

The model here gives the output of both fixed and random effects. The output is described as follows:

1. REML criterion at convergence is the fitting method of the model. REML stands for restricted or residual maximum likelihood. REML criterion at convergence is similar to deviance but dependent on the fixed-effects parameterization.
2. Summary of variance components:
    1) Scaled Residual: It is the summary of the distribution of the final residual error in the model.
    2) Random effects: It is the summary of the different variance components by group 24(Jurisdiction) on both Variance and standard deviation scales.
    3) Residuals are technically random effects; they are included in the random effects summary part.
    4) Number of observations = 1032 and number of groups=24 in the grouping term (Jurisdiction).
3. Summary of fixed effects:
    1) Comparing it to the traditional regression output: Parameter estimates 60 - Standard Error – t-values (ratio of estimate to error) – but no p-values, as it is not trivial to estimate the degrees of freedom in general.

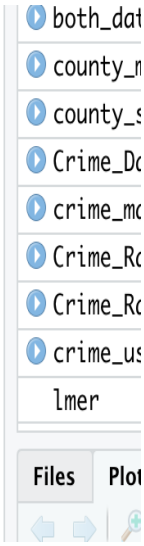
4.2.2. Using the values of LMER

The final summary looks as below:

	B	C	D	E	F	G	H	I
	fixef(lmer)							
		(Intercept)	YEAR_R		ranef(lmer)			
		533.524982	-1.753395		$JURISDICTION			
						Intercept	Year_R	
					Allegany County	-321.77111	6.29215891	
					Anne Arundel County	-121.33061	4.55757223	
					Baltimore City	1738.35225	-9.6580695	
					Baltimore County	373.72965	-3.7910075	
					Calvert County	-133.35748	-1.8437867	
					Caroline County	-138.10934	1.76573896	
					Carroll County	-299.78264	0.97997681	
					Cecil County	-64.23568	2.45909856	
					Charles County	-56.9524	2.43959271	
					Dorchester County	148.83885	-1.7813848	
					Frederick County	-69.97837	-1.2800768	
					Garrett County	-336.46011	2.30261214	
					Harford County	-180.64173	0.30882399	
					Howard County	-131.18086	-3.3070057	
					Kent County	-216.69378	2.19563676	
					Montgomery County	-256.22851	0.21898957	
					Prince George's County	446.59661	-3.8495166	
					Queen Anne's County	-182.13409	-0.7420452	
					Somerset County	-73.98743	0.77597996	
					St. Mary's County	-151.87616	0.09036298	
					Talbot County	-93.10705	-1.8311839	
					Washington County	-240.49311	2.65623111	
					Wicomico County	80.53461	6.28718998	
					Worcester County	280.26846	-5.245888	

The code in R:

```
62
63  # Add the fixed-effect to the random-effect and save as county_slopes
64  print(fixef(lmer)["YEAR_R"] )
65  print(ranef(lmer)$JURISDICTION["YEAR_R"])
66
67  county_slopes <- fixef(lmer)["YEAR_R"] + ranef(lmer)$JURISDICTION["YEAR_R"]
68
69
70  # Add a new column with county names
71  county_slopes <-
72    county_slopes %>%
73    rownames_to_column("county")
74
75  head(county_slopes)
```



By using LMER the values of intercept for fixed and random effects are generated. Now the next step was substituting these values in the equation:

crime_rate ~ Year_R + (Year_R | Jurisdiction)

Eg: Generating crime rate for Allegany County:

Crime_rate ~ -1.753395+6.29215891 = 4.53876391

Similar values for the crime_rate is generated for the remaining counties as shown in the following excel:

Calculating the county Slopes	
county_slopes <- fixef(lmer)["YEAR_R"] + ranef(lmer)$JURISDICTION["YEAR_R"]	
Jurisdiction	County_Slopes
Allegany County	4.53876391
Anne Arundel County	2.80417723
Baltimore City	-11.41146451
Baltimore County	-5.54440245
Calvert County	-3.59718174
Caroline County	0.01234396
Carroll County	0.97997681
Cecil County	0.70570356
Charles County	0.68619771
Dorchester County	-3.53477982
Frederick County	-3.03347178
Garrett County	0.54921714
Harford County	-1.44457101
Howard County	-5.0604007
Kent County	0.44224176
Montgomery County	-1.53440543
Prince George's County	-3.84951664
Queen Anne's County	-2.49544021
Somerset County	-0.97741504
St. Mary's County	-1.66303202
Talbot County	-1.83118389
Washington County	0.90283611
Wicomico County	4.53379498
Worcester County	-6.99928296

The next step was plotting these values on the map of Maryland. This is also done in R. Below is the code snippet: The usmap library is imported and the latitude and longitude is used for every county along with the crime_rate value to see the trend on the map.

```
# Plot the results
crime_map <-
  ggplot(data=both_data, aes(x=x, y=y, group=county, fill=YEAR_R)) +
  geom_text(aes(label = county), data = both_data,  size = 1, hjust = 1)+
  geom_polygon() +
  scale_fill_continuous(name = expression(atop("Change in crime rate","(Number year"^-1*")")),
                        low = "skyblue", high ="red")


# Look at the map
crime_map
```

```
# Load usmap package
library(usmap)

# load and filter map data
county_map <- us_map(regions = "counties", include = "MD")

# See which counties are not in both datasets
county_slopes %>% anti_join(county_map, by = "county")
county_map %>% anti_join(county_slopes, by = "county")

# Rename crime_names county
county_slopes  <- county_slopes  %>%
  mutate(county = ifelse(county == "Baltimore City", "Baltimore city", county))

# Merge the map and slope data frames
both_data <- county_slopes %>% full_join(county_map, by = "county")

# Peek at the data
head(both_data)

# Set the notebook's plot settings
options(repr.plot.width=10, repr.plot.height=5)
```

5. Experimental Results

The Linear Mixed Effect Regression (LMER) is shown in the figure 1. As the name suggests there suggest a linear trend to be happening but there is no linear trend which is seen in the map. The crime rate percent changes for every county in the state of Maryland. The scale on the right-hand side depicts the change in the crime rate. According to the color used in the below image (figure 1) 4 represents an increase in change in the crime rate and is represented by Dark red color. The 0% change is represented by the lighter version of red. The major decrease in the crime rate is represented by the light blue or the sky-blue color. The sky-blue color is represented by -8 on the graph. As we know that Baltimore city has the most crime rate but for the past 40 years the crime rate tends to be decreasing in Baltimore city and has majorly changed the crime trend.

Baltimore county has the light red color that means it has decreased by the slope of -4. To check the accuracy of the graph a second graph is plotted which is shown in figure 2. The figure 2 below shows the change in the crime rate. The graph below shows that Baltimore city has a reduced

crime rate of 20%. This means the Baltimore city has become safer over the span of 40 years. The countries like Caroline, Kent, Fredrik have a positive Increase in the crime rate. The same counties are described by the red color in the Heat map shown in figure 1. The left side in the graph in figure 2 shows the Highest crime rate that is happening in the counties. The graph depicts which county has the highest crime rate in the State of Maryland. From the graph we can say that Baltimore City has the highest crime rate in the state of Maryland, but on the other side the positive is that the trend over the span of 40 years is decreasing in Baltimore county.

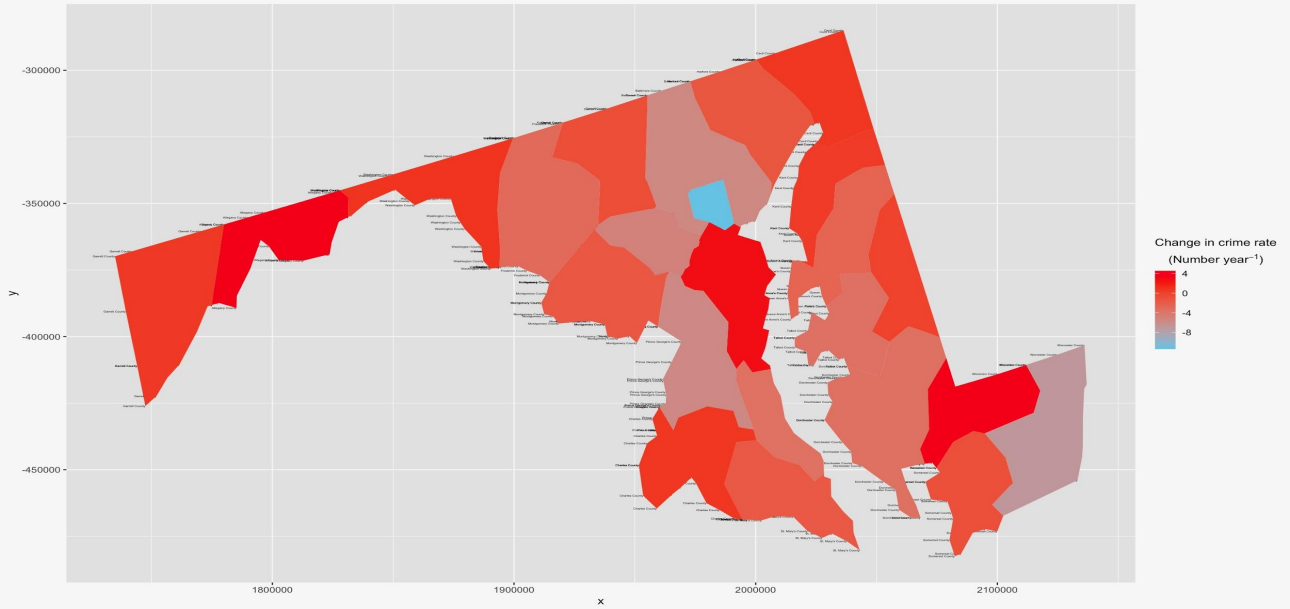


Figure 1: LMER heat map

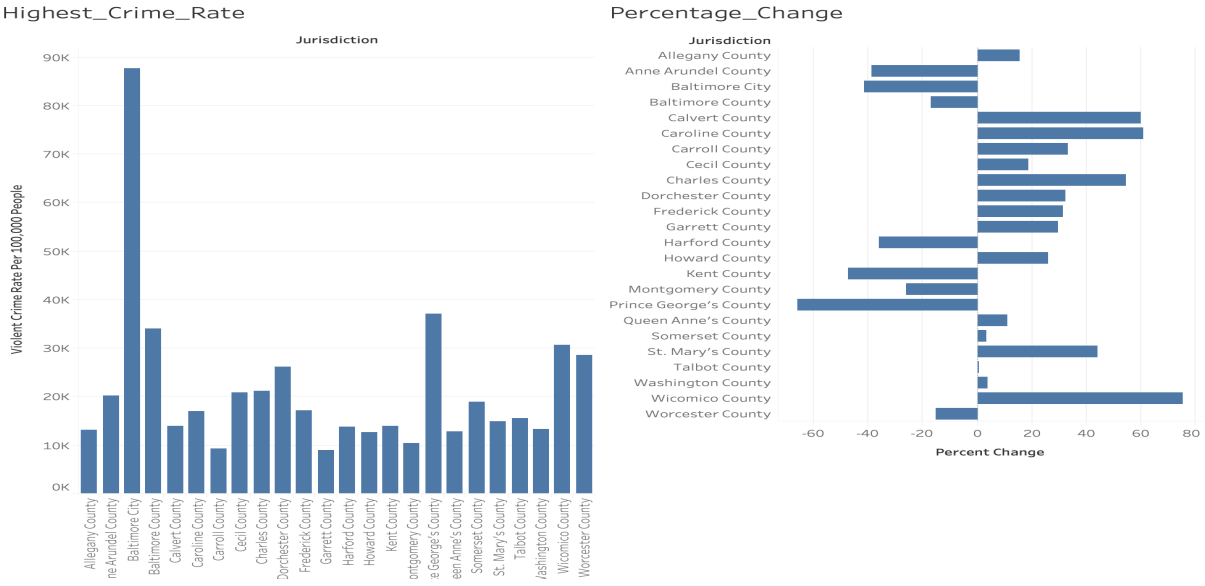


Figure 2: Percent change in Crime

The table in the Methodology is showing the various calculations done to calculate the slope of every county to be plotted on the map. The values are generated by using LMER and then another formula is applied to calculate the county slope and similarly the US map function is used for depicting this trend. (Figure 3) shows the values of LMER predicted using a scatter plot. Figure 3 has all the 24 Jurisdictions in Maryland and they are plotted as the crime rate against the year. It is shown in the form of Scatter Plot. Figure 4 is showing the various crimes happening in Maryland according to the year and their trends like for which year it is increasing and similarly for which year it is decreasing. There are crimes like Robbery, Rape, Murder, Larency, Motor Vehicle theft that is happening in various counties and the graph or the trend about which year does it takes place is shown in the figure.

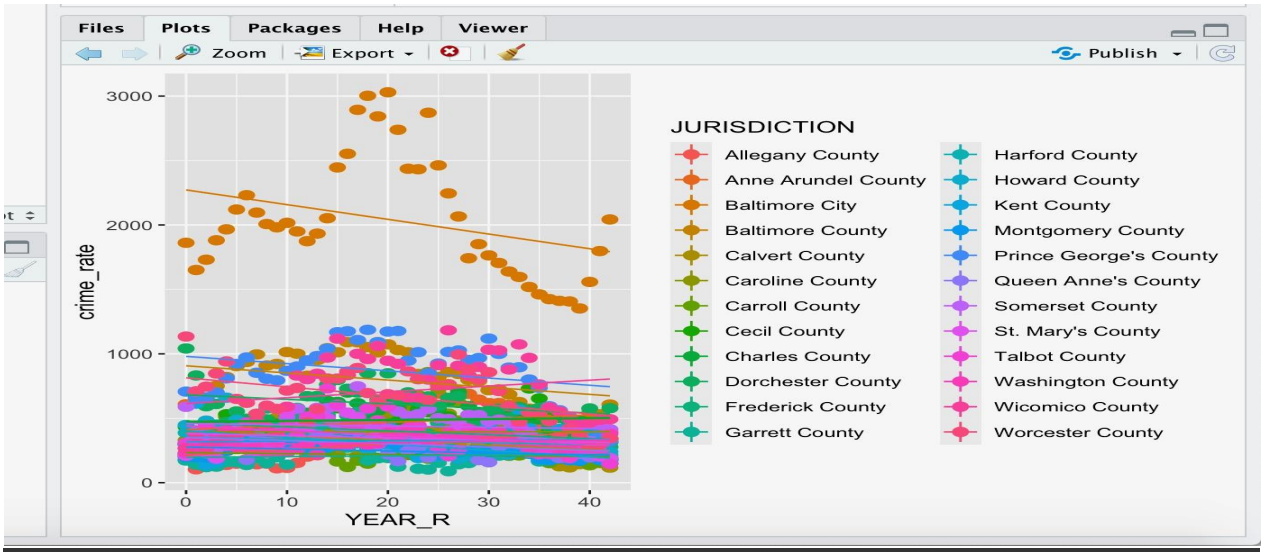


Figure 3: Scatter plot for the LMER values. Both fixed and random effects

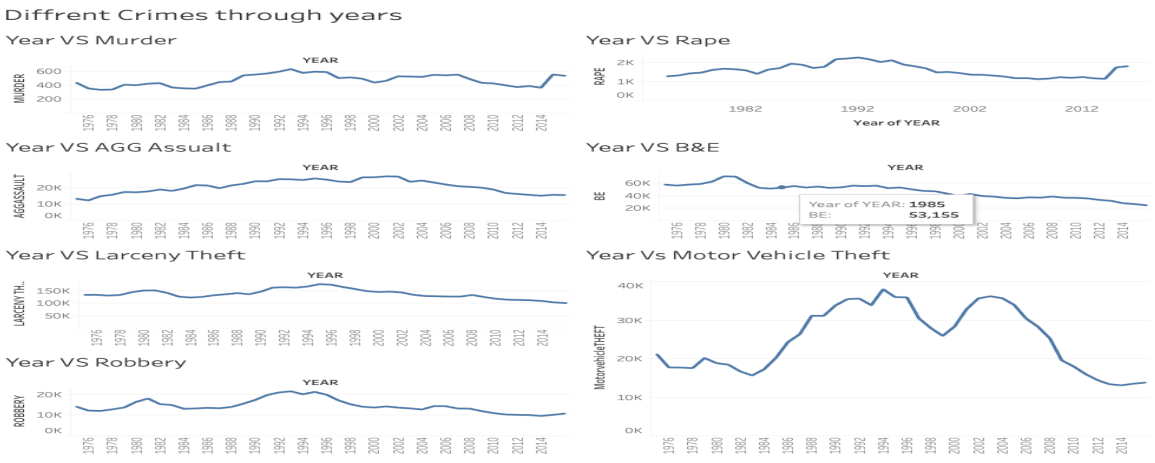


Figure 4: Crime Rates against years in Various Counties

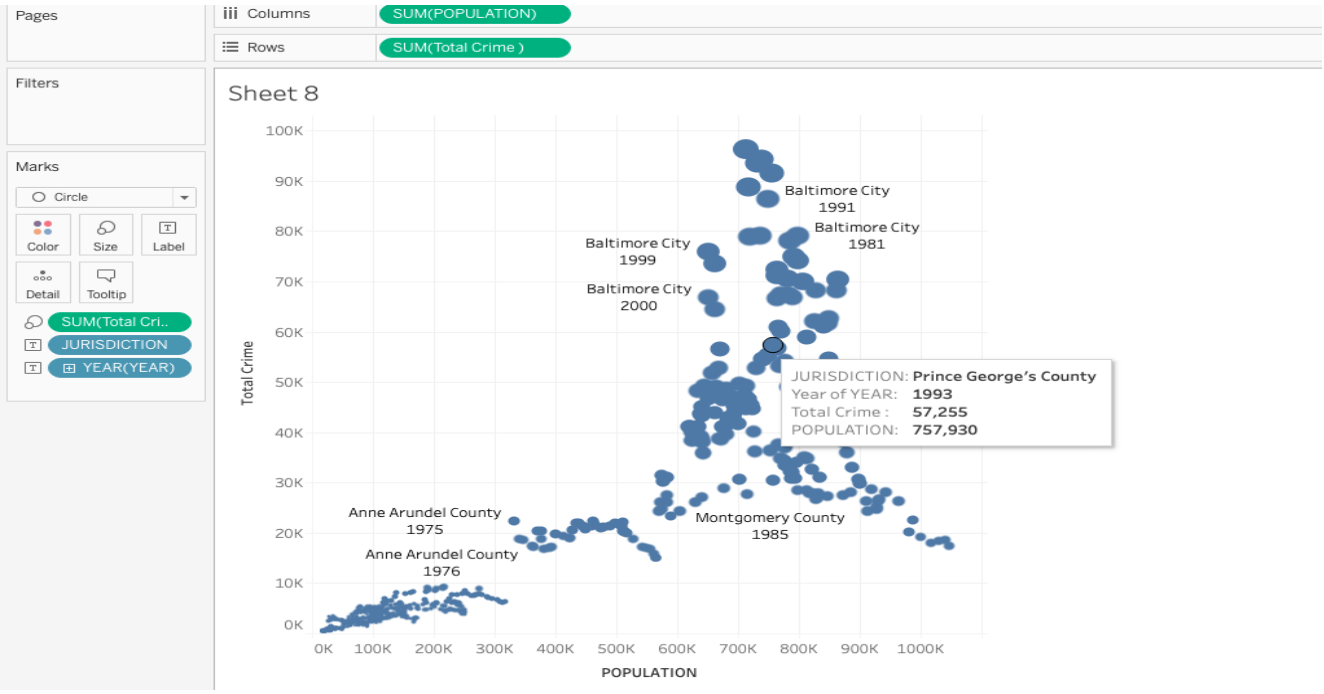


Figure 5: Population VS Total crime

After analyzing the data there were few questions like whether the population influences crime rate, this question can be answered through Tableau Visualization as shown in (Figure 5) and it can be found that it has a negative correlation and thus population does not matter for crime to take place. In figure 6 we have described the population density vs the total crime in the state of Maryland. Baltimore county is one of the counties with the highest rating of crime per population density. In (Figure 6) we can see the trend of crime happening in the state of Maryland, for the span of 40 years. In (Figure 7), the implementation of crime rate per population density using the density heat map.

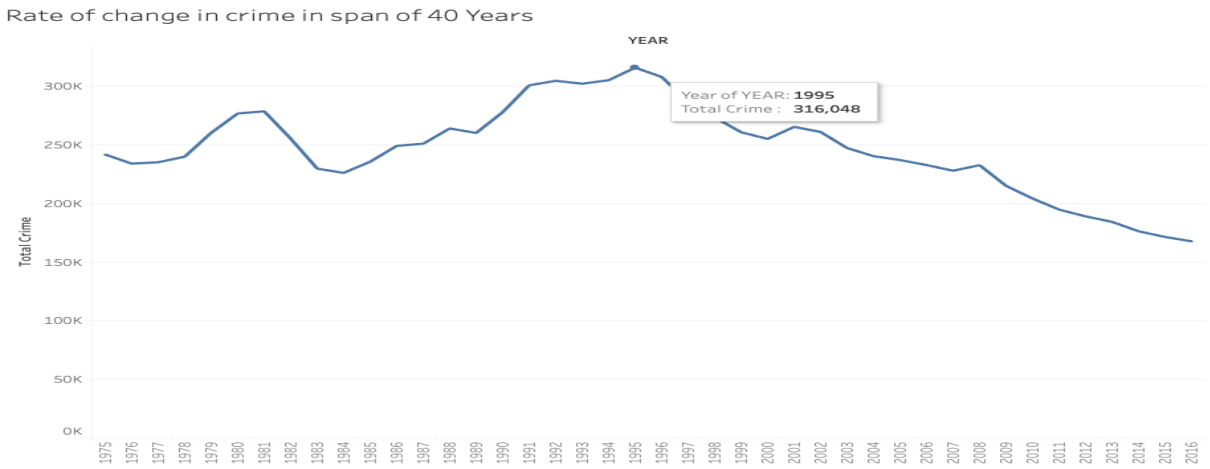


Figure 6: Rate of change in crime over the span of 40 years

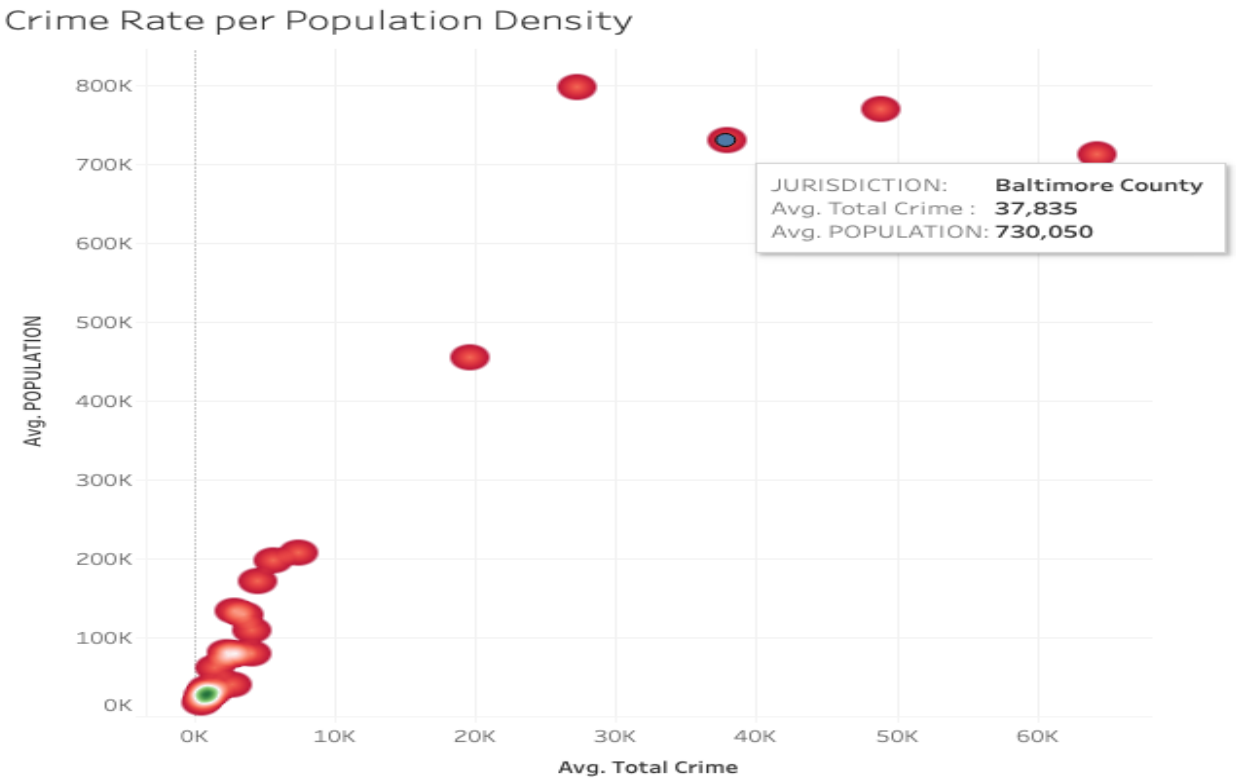


Figure 7: Population Density Vs Crime

6. Conclusion

Crime rates are varying across the entire world, so it is very difficult to detect. By visualizing the trends in the crime rate, the policy makers can decide on different types of plans based on the trends or the happenings of crime in that county. The data which we are dealing with has been collected from Maryland Statistical Analysis Center and Open Maryland Site. The data is collected for a 40-year time span and this data is enough to visualize and detect the trends to help the government to increase the safety measures in that particular county. As the concern was to show trends, a regression model was used. Post research Linear Mixed Effect Regression (LMER) was used to show the trend. To cross verify the result of the LMER, we plotted the same trend using Tableau to show the percent change in the crime and the result matches with the LMER output. After analyzing the data there were few questions like whether the population influences crime rate, this question can be answered through Tableau Visualization as shown in (Figure 5) and it can be found that it has a negative correlation and thus population does not matter for crime to take place .Hence we conclude that there is no linear trend in state of Maryland, and it differs from county to county. This crime data and analysis will help us make Maryland a better place to live in.

7. Future Scope

Algorithms such as KNN and decision trees can be applied in future to train the model. All the categorical variables when converted to numerical variables give unique ids. All the counties with crime will have different ids, Example: vehicle theft is given as crime is 10 and so on. With this approach, it is not clear how to assign numerical values to categorical values in a meaningful way. Hence this would likely cause difficulties for the learning algorithms, since assigning different numbers would lead to different results.

Therefore, treatment of categorical variables can be done using "one-hot encoding" approach where categorical variables are converted into binary variables with a single "1" and "0"s.

Also, decision trees can deal better with large datasets that have many layers with different nodes. Decision-trees provide better balance of flexibility and accuracy while limiting the number of possible decision points. It would be good to visualize the internal representations of the model, e.g. show the structure of a decision tree, and discuss any insights from the data retrieved from these representations.

8. References

1. Maryland Crime Retrieved from https://www.neighborhoodscout.com/md/crime

2. Crime Statistics Retrieved from http://goccp.maryland.gov/crime-statistics/

3. Crime Dataset Retrieved from https://opendata.maryland.gov/Public-Safety/Violent-Crime-Property-Crime-by-County-1975-to-Pre/jwfa-fdxs

4. https://web.stanford.edu/class/psych252/section/Mixed_models_tutorial.html

5. https://www.biorxiv.org/content/biorxiv/suppl/2016/07/06/062299.DC1/062299-1.pdf

6. https://www.kaggle.com/kevin2k1503/maryland-crime-data-by-county-from-19752016

7. https://www.neigh

8. https://www.ipl.org/div/stateknow/popchart.html

9. https://www.sciencedirect.com/science/article/pii/S1877050918315667

10.https://www.researchgate.net/publication/280722606_Crime_Analysis_and_Prediction_Using_Data_Mining