Names: Saee Risbud, Scout Farda, Kristi Pham, Krisha Chaudhari,

Shahar Rozenberg, Annie Yu, Seth Dayawansa

1. Introduction

The presence of crime has persisted in society and it is pivotal to understand not only the underlying factors that contribute to the ongoing act of crime, but also what aides towards their solve time to determine solutions that can reduce current crime rates and prevent them from occurring. Crime statistics play an important role in many aspects such as predictive policing to prevent crimes, improving community relations, appropriate budget formation and resource allocation, and more (Walden University). There are existing studies that delve into criminal case processing time (Ostram 2020) and police response time (Auten 1981) along with other reports that tackle different aspects of this issue that show the significance of this topic.

Of course, being in Austin, we decided to hone in our city's crime reports to see if we could identify patterns and curate our own findings. Upon discovering our dataset we observed that a majority of Austin crimes had not been cleared which leads us to question what impacts resolve time, also synonymous with clearance time. Based on what was available, we decided to see whether the council district the crime occurred in -or a different factor- had a stronger correlation.

2. Data

We found this dataset at austintexas.gov. This is a compilation of crimes reported and responded by the Austin Police Department that is regularly updated, beginning in 2003. There may be many offenses categorized to a single incident but this dataset only accounts for those that are of the highest degree. Thus, in this dataset, each incident has one offense attached to it. The dataset contained a number of variables, but the ones we were most interested in were the crime type (description of the highest degree of offense), occurrence date (reported date of the committed crime), clearance date (date that the crime was solved/cleared), clearance status (how/whether the crime was solved) and UCR category (code that details what the crime is). We chose to highlight these variables because a lot of our analysis was based on how long it took to get the crime cleared (or solved). Since the dataset was too huge to compute with, we decided to start our analysis using only the data from 2023. To find the clearance time to make the graphs, the difference of the occurrence date and clearance date was taken. Data about the mean family income corresponding to each district to analyze if affluence of each of the council districts influences the clearance time was obtained from Housingworks Austin.

To clean and process the data, we truncated all data not from 2023 and got rid of any NA values. Data points with a longitude or latitude of 0 were also ignored, because they may have been faulty. Rows with inconsistent labels and typos were also ignored.

3. Exploratory Analysis [30 points]

To start exploring the data, our team first decided to look at a map of Austin with specific data from our dataset to see if anything stood out. The first image we generated was a map showing every crime committed and its clearance status. This can be seen below in Figure 1.

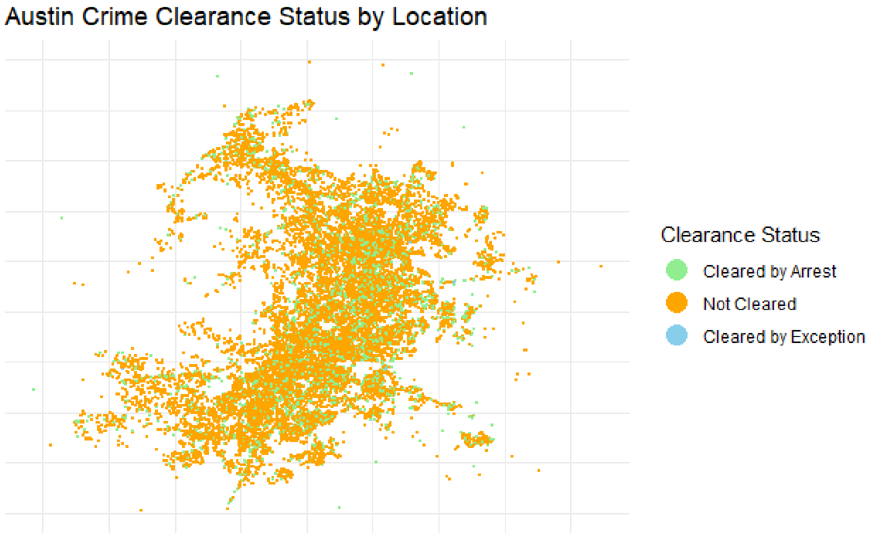

Figure 1 - Austin Crime Clearance Status by Location

Based on the above map, we can immediately see that the vast majority of Austin crimes committed in the last year remain unsolved. We can also see that the distribution of cleared cases (The green dots) seems clustered in certain locations.

From this point, we then looked into different ways we could stratify the data into different geographical regions. Examining the data we found the council district in which the crime occurred. Information about the citizens within each council district is readily available online, so we chose to further explore this. We next generated an image showing the bounds of the council districts in Austin in regards to crimes committed to double check the validity of the geographical data. This visualization can be seen in Figure 2. We also generated a bar chart showing clearance rate by council district to see if there is a notable difference. This can be seen in Figure 3.

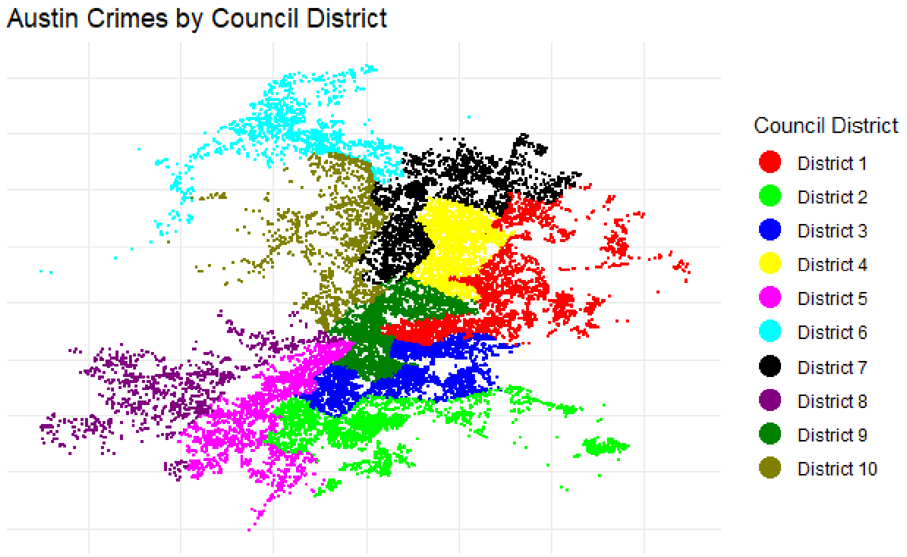

Figure 2 - Austin Crimes Sorted by Location and Council District

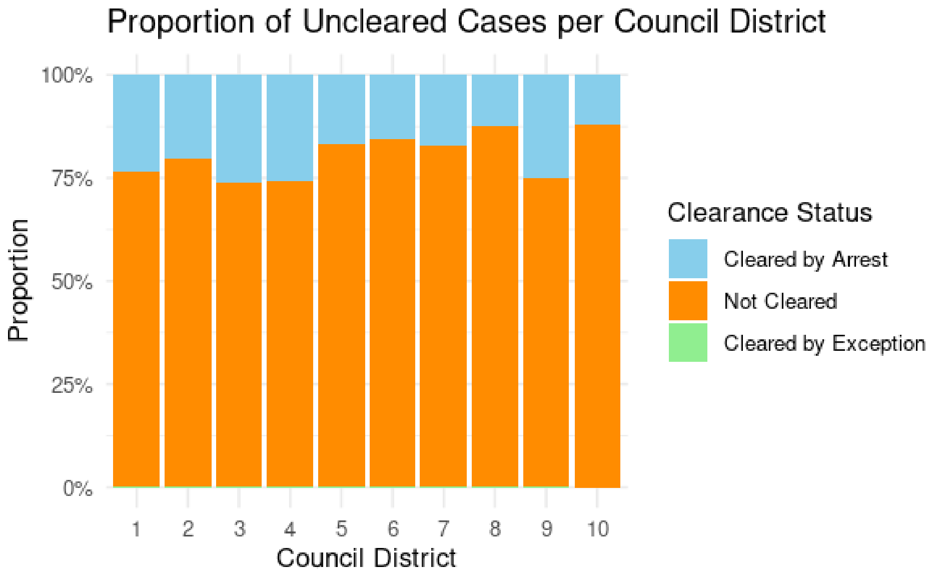

Figure 3 - Austin Crime Clearance Status by Location

From Figure 2, it can be seen that the number of crimes committed per council district varied significantly district to district, and from Figure 3 it can be seen that there is a very distinct difference between the clearance proportions of different council districts. This led us to explore a first hypothesis: *the clearance rate of a council district is highly dependent on the number of crimes committed within it*. And our second hypothesis: *the clearance rate of a council district is highly dependent on its affluence.*

We were also interested in the factors influencing how long it takes for crimes to be cleared. So we then went to graph the average clearance time per council district. The result of that exploration can be seen below in Figure 4.

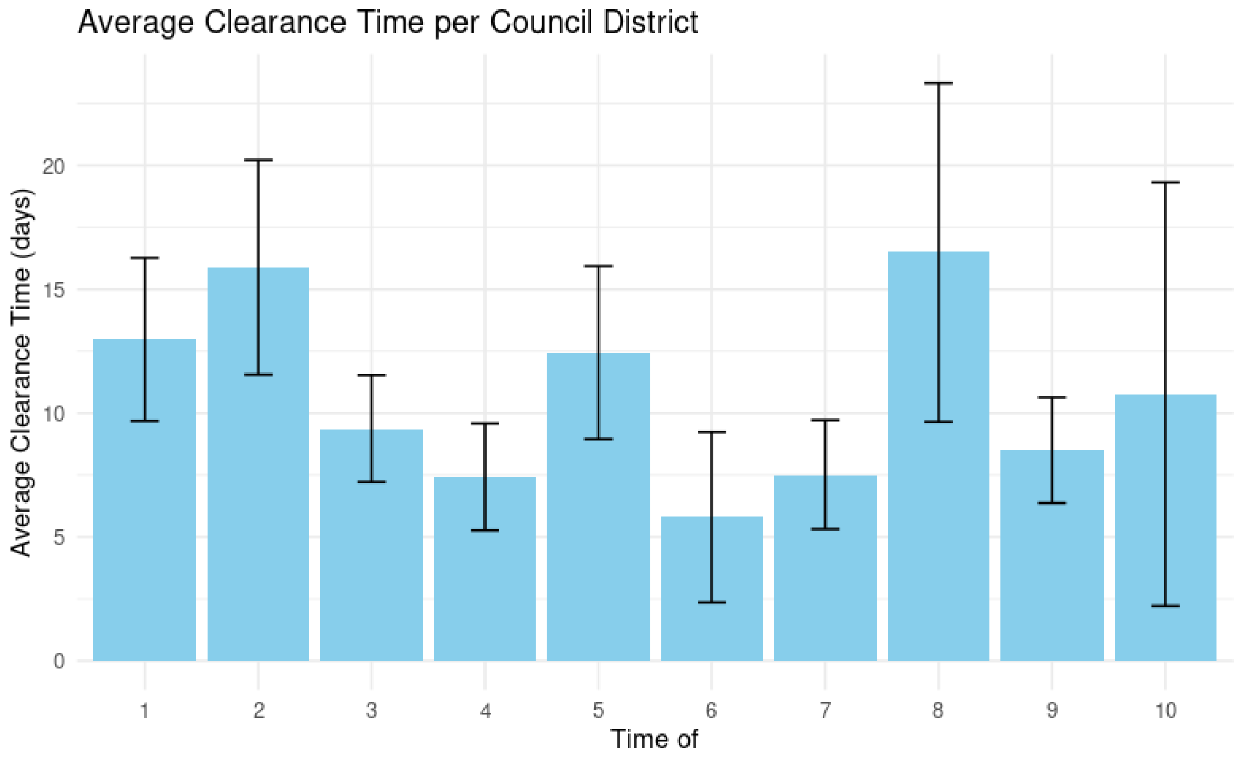

Figure 4 - Austin Crime Average Clearance Time in days by Council District

From the above figure, it can be seen that there is a significant variation between the clearance time between council districts, even though the error bars indicate a massive variance within each district. This led us to explore a third hypothesis: *the number of crimes committed per council district is highly dependent on its clearance rate*. And our fourth hypothesis was generated: *the affluence of each council district is highly dependent on its average clearance time.*

In order to test the hypotheses of whether council district affluence affected crime statistics, we needed data corresponding to the wealth of these districts. We chose median family income (MFI) to serve this metric, and we sourced the data from Housingworks Austin.

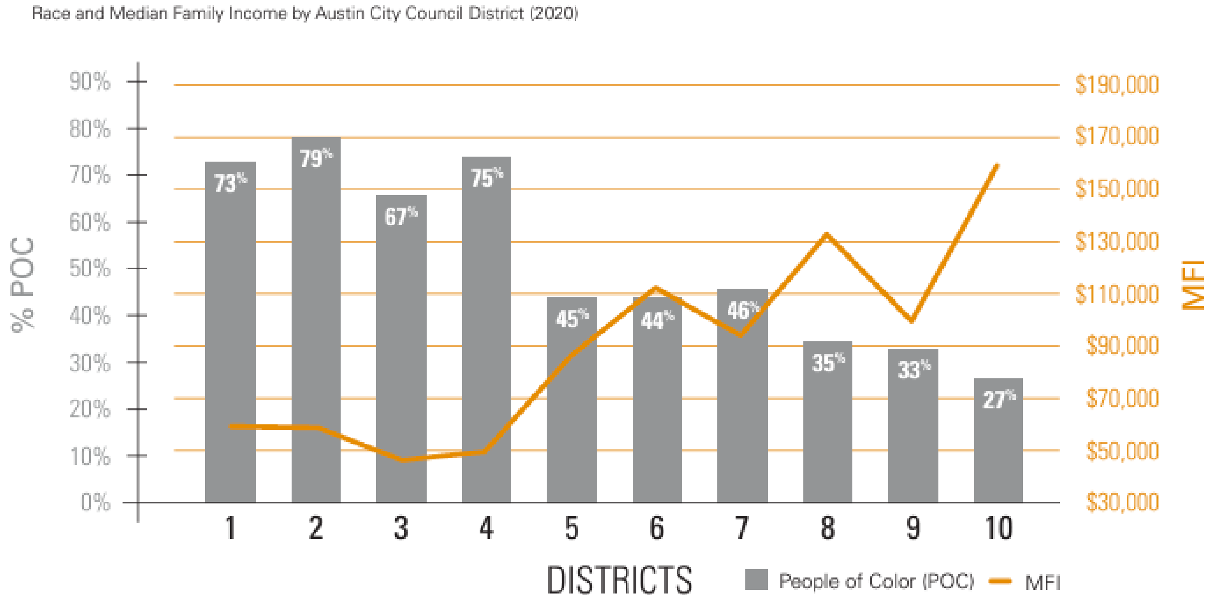

Figure 5 - Median Family Income by Council District

From the information in the figure above it is very apparent that income levels vary dramatically between districts, but it is not immediately apparent if there is any correlation.

4. **Modeling** [20 points]

Before examining the relationship between affluence and crime rates within council districts. We first tested whether the council district had any importance relative to these variables. This is because if the council district turns out to be unimportant, any analysis within the variable would be pointless.

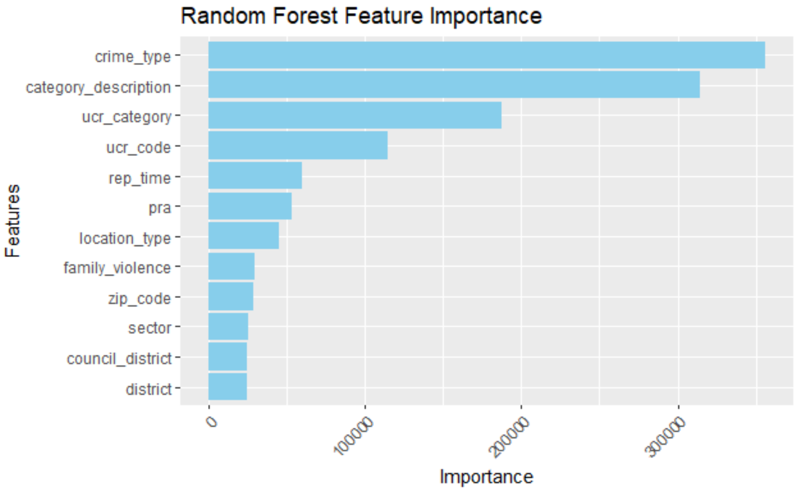

Figure 6 - Austin Crime Clearance Time Random Forest Feature Importance

Before starting our statistical analysis, we had some hypotheses about the data. Having those, we were still not sure what are the features that impact the clearance time the most. Does the council district affect the amount of time for a crime to be cleared? Does the clearance time depend on any other variables?

In order to figure this out, we used a *Random Forest* model. The model gets as input a *target variable* and *feature variables*, and outputs the "how much" the target variable depends on each feature variable. The data we get from the model is called the *importance* of each variable to 'predict' the target variable. Of course, the rows that have non-zero clearance time are the rows that represent cleared crimes. Therefore, we cleared the data from all the not-cleared crimes for this specific analysis. Furthermore, we separated our data to *training* and *test*, in order to make sure our model works.

In Figure 6, we can see a bar graph of the feature variables' importance in predicting the target variable - *Clearance Time* in that case. We can observe the fact that the 2 most important feature variables are *crime_type* and *category_description*. Both of these variables are describing the same thing (which is the nature of the crime), while *category_description* is separated into only 7 common crime types. It allows us to understand that the nature of the crime has the most influence on the clearance time.

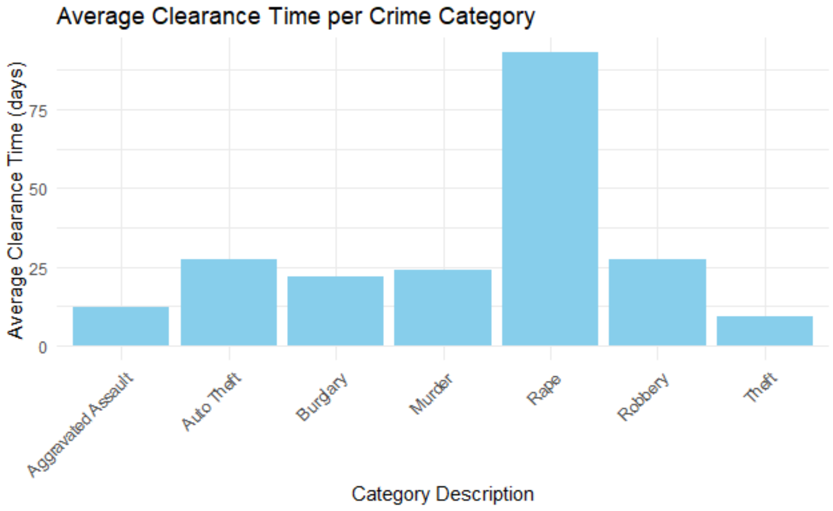

Figure 7 - Austin Crime Clearance Time per Crime Category

In the figure above , we can see a visualization of how the *category_descriptions* affects the clearance time. We can easily see that rape has the longest clearance time, while the others are quite similar.

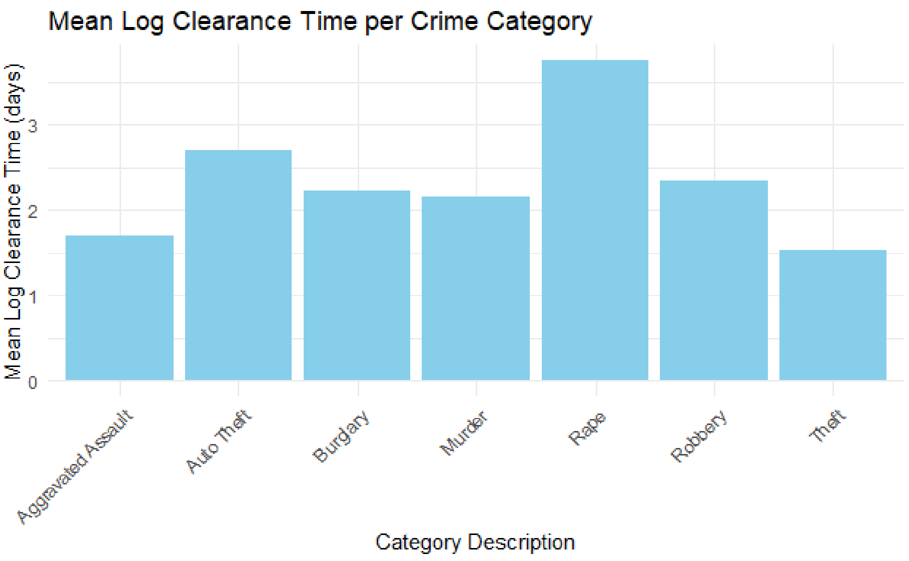

Figure 8 - Austin Crime Log Transformed Clearance Time per Crime Category

The bar graph above shows us that even after conducting a log transformation to the clearance time $(log(1 + clearancetime))$, the ratio stays the same. Now we can be sure that rape has the longest clearance time, without strong biases.

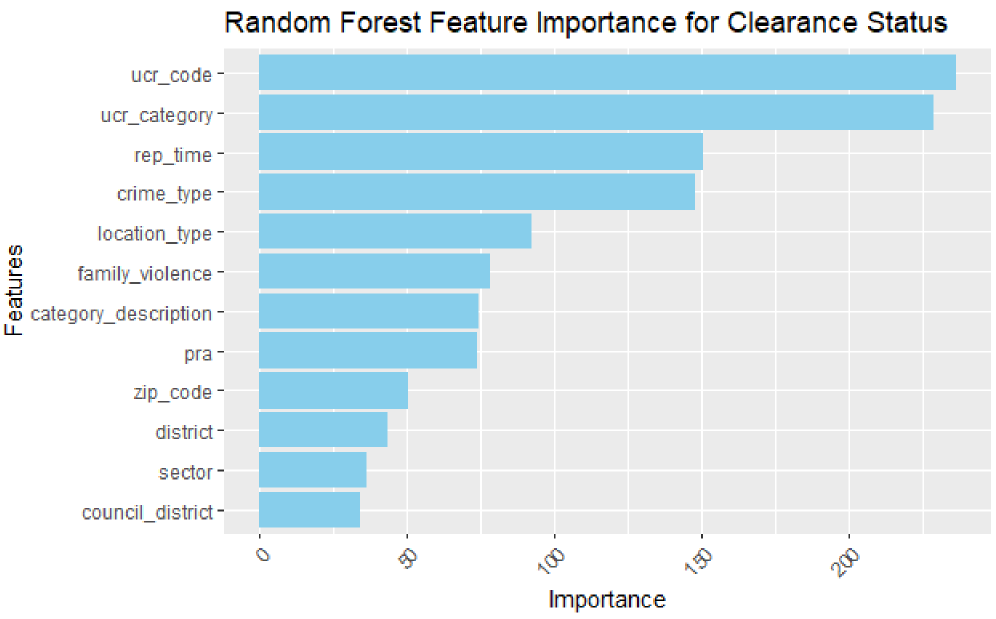

Figure 9 - Austin Crime Clearance Status Random Forest Feature Importance

After applying the random forest model in order to find the most important features in predicting *clearance time*, we decided to do the same, but this time the target variable is *clearance status*. We wanted to check which variables affect the clearance itself.

In Figure 9, we can see a bar graph of the feature variables' importance in predicting the target variable - *Clearance Status* in that case. This time, the most important variables are the *ucr_code* and the *ucr_category* (Categories - groups of codes). The UCR (Uniform Crime Reporting) categories are the FBI's way to categorize crimes in their statistical data. The visualization helps us understand the UCR category affects the clearance status.

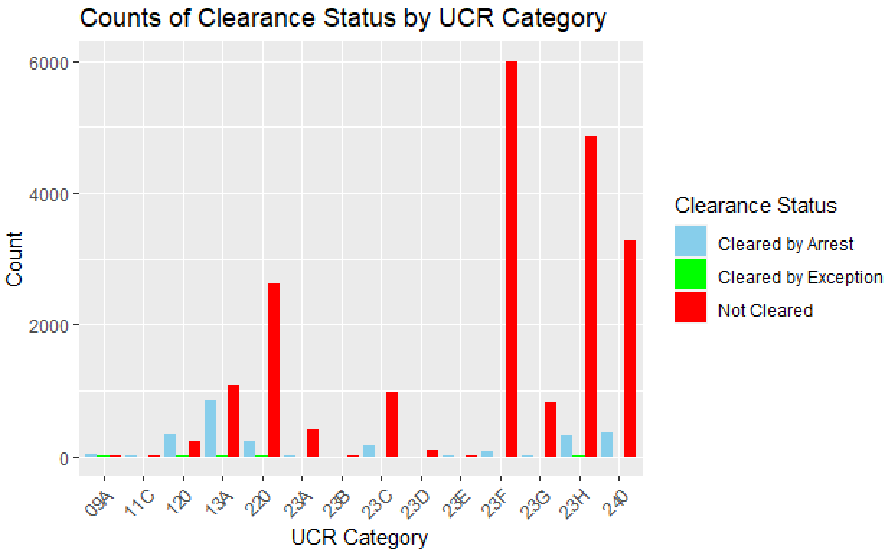

Figure 10 - Austin Crime Clearance Status by UCR Category

In the figure above , we can see a visualization of how the *ucr_category* affects the clearance status. Each category has 3 bars - cleared by arrest/exception and not cleared. We can easily notice that there are some crimes with more not-cleared crimes than cleared ones. For example, 23F ('Theft From Motor Vehicle ') has thousands of crimes not cleared, while there are barely any crimes that have been cleared.

Due to the results of this analysis it became apparent that further analysis into the relationship between council district and our crime statistics of interest was pointless. We were now able to respond to our initial hypotheses.

5. Discussion

We used feature importance analysis within the random forest learning algorithm to test our hypothesis and form our conclusions about the importance of district council on crime clearance time and clearance status. We generated 2 feature importance analyses to determine which category had the greatest importance on crime clearance time and clearance status to explore our hypothesis.

From the feature importance analysis of the categories' importance on crime clearance time, we found that crime type had the highest importance score meaning that it played the biggest role in determining clearance time. To dig deeper into the specific crime categories, we created a bar graph of all the average clearance times to each crime type. We found that rape had a significant longer clearance time at more than 75 days , compared to the other types of crime that were all cleared within 30 days. Additionally, to our surprise, we found that the council district had the second lowest importance score meaning that it did not play an important factor in determining clearance time.

In the feature importance analysis showing the significance of each category on crime clearance status (not cleared, cleared by arrest, cleared by exemption). We found that UCR (uniform crime report) code, which shows the type of nature of crime, had the greatest importance in determining clearance status while council districts had the lowest importance score showing little significance on clearance status. We did another bar graph to look into the clearance statuses organized by UCR code and found that the code 23F, which represents petty crime like theft, had the highest number of uncleared cases.

The models showed us that our original thought about council districts playing an important role in determining crime clearance time and clearance status was proven wrong. The difference in resources, wealth, and police department organization didn't play a huge significance in determining crime rate. In contrast, we found that type of crime was a bigger determinant for both clearance time and clearance status. This would lead to further investigations of the reasons behind this; however, a few hypotheses could be the difference in evidence availability, resource allocation, and complexity of the investigation.

Limitations to our analyses is that our data set was too big to run in R, so we only used data from 2023, meaning that it could potentially be an outlier and not a holistic representation of crime in Austin throughout these past years if we wanted to see an overall trend. This would lead to further investigations of a comparison randomly selected dataset. However, we did choose the current year, 2023 as we wanted to see the current trends about crime.

Furthermore, the dataset did not include crime that was not reported which would lead to skews in the analyses or misinterpretation as many crimes like rape and assault can be mentally hard for the victim to report. This would lead to missed data which can influence our conclusions.

Additionally, there could be potential missed outliers within the dataset when we formulated our analysis of the average clearance time for each crime type which could be a limitation to our conclusion about the clearance time for the types of crime.

Moreover, 'district' serves as an indicator of location and in our analysis, we included all of the major variables of the dataset which included many other location-related variables. This overlap could lead to information leakage across variables and potentially limit our conclusion.

Lastly, a challenge for our future investigation is to quantify the impact of district council on our 2 main variables (clearance time and clearance status). We only found that the variable ‘council district’ had a low importance score; however, we weren’t able to quantify the days or number of cases for each status it affected which would be a challenging, but informative future investigation.

6. Ethics

As with every project, it is important to discuss the ethical considerations regarding the findings and nature of the topic at hand. The AREA Plus (4p) Framework describes 16 principles in which 3 will be further described within the context of this project.

The first principle asks about future uses and impacts and whether they are sustainable and socially desirable. Based on the findings, though location does have some correlation with clearance time – as represented by variables such as council district, zip code, sector, and district – the analysis showed that crime type, specifically rape, is the most influential factor towards clearance time with instances of rape taking the longest to be resolved. Using this information, an effective measure could include thoughtfully considered budget and resource allocation towards districts to ensure they have the means and tools to ensure justice. Further research needs to be conducted to determine the specifities regarding long clearance time for rape cases, but it is imperative that police departments within these districts allocate more resources for incidents of this type as they require a heavy workload relative to other crime types. It is challenging to determine the sustainability and social desirability of these measures without knowing the current circumstances regarding resource distribution within departments, however it is reasonable to infer that a majority of society can come to a common understanding that rape is a heinous act that can not go unpunished, therefore it should have increased dedication. That is not to undermine the atrociousness of other crime types but considering the longer clearance time, increased efforts are needed.

The second principle asks about possible consequences, what is not known, and how could social desirability be ensured. This analysis did not identify the reasons for longer clearance time of rape cases, therefore there is the possibility that increased resource and budget allocation by police departments might not actually be the ideal action. More effective measures may have to be enforced by families or witnessing bystanders, school systems, the overall public, or other departments. However, if redistribution of a budget and resources was undertaken, a potential consequence could be that other types of crime would experience the ramifications of decreased tools. This leads into how social desirability could be ensured which would be to find a sort of balanced base line or compromise that takes into consideration the complexities of each crime type with each getting the respective means needed to solve them.

The third principle asks how to engage a wide group of stakeholders. To initiate this process, it is essential to foster an inclusive and collaborative environment that encourages open dialogue. Establishing a task force comprising representatives from law enforcement, legal professionals, victim advocacy groups, healthcare providers, and community leaders can ensure diverse perspectives are considered. Conducting regular town hall meetings, workshops, and focus groups can provide a platform for stakeholders to voice their concerns, share insights, and propose solutions. Furthermore, leveraging technology through online forums and surveys can facilitate broader participation. It is imperative to emphasize the importance of sensitivity and empathy when discussing such sensitive matters and to prioritize survivor-centered approaches. By fostering a collective commitment to addressing the complexities surrounding rape cases, we can collaboratively work towards expediting the resolution process, supporting survivors, and promoting justice within our communities.

7. Conclusion

In the city of Austin, we see that a majority of the crimes are uncleared and the clearance time for each of the council districts varies. We hypothesized that the affluence of districts might indicate the difference in clearance time across districts and thus made a graph that indicated the affluence of each district. This graph also showed differences across the districts however there was no direct correlation between the affluence and the clearance time. Since we could not find any significant relationship between the affluence and clearance time, we conducted a Random Forest Feature Importance and found out that crime type was the most significant variable in defining the clearance time while council district was one of the least significant ones. Based on this information, we made a graph for finding the clearance time for each crime type and concluded that rape had the highest average clearance time. Next we conducted a random forest feature importance for finding the clearance status and found that the UCR code (which defines the type of crime) was the most determinant in finding the clearance status.When the UCR code and the number of reported crimes along with their clearance status was graphed, we observed that UCR codes for crimes such as theft (such as motor theft).

8. Acknowledgment [1 point]

Team Member Name	**Percentage Contribution**
Kristi Pham	100%
Saee Risbud	100%
Scout Farda	100%
Krisha Chaudhari	100%
Shahar Rozenberg	100%
Annie Yu	100%
Seth Dayawansa	100%

9. Bibliography

Sources

Austin City Council District-by-District Analysis 2020, housingworksaustin.org/wp-content/uploads/2021/06/2020_ExecSumm_Methodology_FNL.pdf. Accessed 4 Dec. 2023.

Auten, J. “Response Time - What’s the Rush.” *Law and Order*, vol. 29, no. 11, November. 1981, pp 24-27, U.S. Department of Justice, https://www.ojp.gov/ncjrs/virtual-library/abstracts/response-time-whats-rush

City of Austin, Texas - data.austintexas.gov. “Crime Reports: Open Data: City of Austin Texas.” *Data.AustinTexas.Gov - The Official City of Austin Open Data Portal*, 4 Dec. 2023, data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu.

Ostram, Brian, et al. “Timely Justice in Criminal Cases: What the Data Tells Us.” *National Center for State Courts*, Effective Criminal Case Management, August. 2020, https://www.ncsc.org/__data/assets/pdf_file/0019/53218/Timely-Justice-in-Criminal-Cases-What-the-Data-Tells-Us.pdf

“Why National Crime Statistics Are Important”, *Walden University*, https://www.waldenu.edu/online-bachelors-programs/bs-in-criminal-justice/resource/why-national-crime-statistics-are-important

Libraries

Hadley Wickham developed the "dplyr" package for data manipulation in R (Version 1.0.0) [Computer software]. (2019). Retrieved from https://CRAN.R-project.org/package=dplyr

Hadley Wickham created the "ggplot2" package for data visualization in R (Version 3.3.2) [Computer software]. (2016). Retrieved from https://CRAN.R-project.org/package=ggplot2

Garrett Grolemund and Hadley Wickham developed the "lubridate" package for handling date-time data in R (Version 1.7.9) [Computer software]. (2018). Retrieved from https://CRAN.R-project.org/package=lubridate

Andy Liaw and Matthew Wiener developed the "randomForest" package for building random forests in R (Version 4.6-14) [Computer software]. (2018). Retrieved from https://CRAN.R-project.org/package=randomForest