Analysis of Crime Clearance Rates in San Francisco

Final Project Submission for IDS 702, Fall 2021

Deekshita Saikia

12/12/2021

Summary

This analysis investigates factors that influence the clearance rates of crimes across the various Police Districts of San Francisco. The data analyzed contains reports of crime incidents filed daily with the San Francisco Police Department (SFPD) since January 1, 2018. The final valid model used is a hierarchical logistic regression model that models crime clearance rates as the response, and Police Districts as the random intercept. The results show that the time of day, if the crime occurred on a weekday or weekend, and the category of crime, have an impact on the crime clearance rates across the police districts. The dataset is updated daily, so the clearance rates used in this analysis might not be exactly accurate.

Introduction

San Francisco has long been an epicenter of crime incidents, with the most recent spurt of burglaries and homicides observed during the pandemic. The locals face a 1-in-16 chance each year of being a victim to a property or a violent crime, which makes the city more dangerous than around 98% of the cities in the US. While the overall crime rates have gone down since the peaks of the 1980s, the crime clearance rates have shown no improvement in recent years. The SFPD's clearance rates dashboard provides data on the number of crimes cleared, by type of crimes and police precincts, which shows a consistent decrease in the clearance rates across all types of crime categories since 2018.

In order to implement law and order effectively, crime rates, as well as crime clearance rates must be analyzed and steps taken to lower the number of unsolved crimes as much as possible. In this analysis, some factors that possibly influence the crime clearance rates are studied. In particular, we look at factors like season of the year, public holidays, crime categories, time of day and week, and how they affect the crime clearance rates, for the eleven police precincts of San Francisco.

Data

The dataset used for this analysis has been sourced from SFPD's Incident Report Database. The dataset compiles data from the department's Crime Data Warehouse (CDW) to provide information on incident reports filed daily by the SFPD with CDW, or filed by the public with the SFPD.

The incident reports data contains records of incidents filed daily since January 1, 2018 to November 11, 2021. It is the closest the department provides to raw counts of reported crimes across the city. All reports are approved by a supervising Sergeant or Lieutenant.

The dataset includes attributes like date and times of incidents, the street intersections where each incident occurred, whether multiple crimes occurred in the same incident, and the jurisdiction where the incident took place. We leverage these attributes for our analysis, to see which factors influence crime clearances, by the different police precincts of the city. In addition, public holidays ave also been merged into this dataset to examine if holidays had an effect on crime clearances.

A clearance rate describes the percentage of clearances reported to the number of crimes reported. The variable *Resolution* is a categorical variable with 4 classes: *Open or Active*, *Cite or Arrest Adult*, *Exceptional adult*, and *Unfounded*. We define a binary variable that signifies if the crime was resolved, classifying *Cite or Arrest Adult* and *Exceptional adult* as resolved, and the remaining classes as unresolved.

Before fitting a model, a series of pre-processing steps are carried out in the data. The redundant and irrelevant variables for our analysis are dropped. Some incident reports have multiple incident codes recorded against a single incident number. In these cases, a single incident code is retained against an individual incident for the purpose of this analysis. It is also worth noting that once a supervising officer provides approval for an incident report, no further changes can be made to it. If changes or additional information is required or discovered during an investigation, a supplemental report is filed to capture updates. Hence, for this analysis, the latest report corresponding to an incident is kept.

The *Incident Subcategory* column has 71 unique categories of incidents. Since this is very granular to model on, the categories corresponding this column are collapsed into four classes: *Property Crime*, *Violent Crime*, *Arson* and *Other*. This column also has a very minute proportion of missing values (~0.07%), which are dropped from the dataset. In addition, incidents that took place outside the jurisdiction of San Francisco have been dropped.

A few features are engineered out of existing features in the dataset. A *Season* variable is created out of the month of occurrence of the incident. The time of occurrence of incident is used to create a categorical variable $TimeofDay$, which consists of 6 classes, viz., *Late Night*, *Early Morning*, *Morning*, *Noon*, *Evening* and *Night*. An indicator for weekend is also created, and public holidays have been merged into this dataset from an external source. The dataset is then rolled up, so our unit of observation is *PoliceDistrict*, *IncCategory*, *Season*, *Month*, *isWeekend*, *Holiday* and *TimeOfDay*, and clearance rates are calculated at this level.

Exploratory analysis is carried out to examine how the distribution of clearance rates differ across each of the police districts. This can be seen in the graphs below.

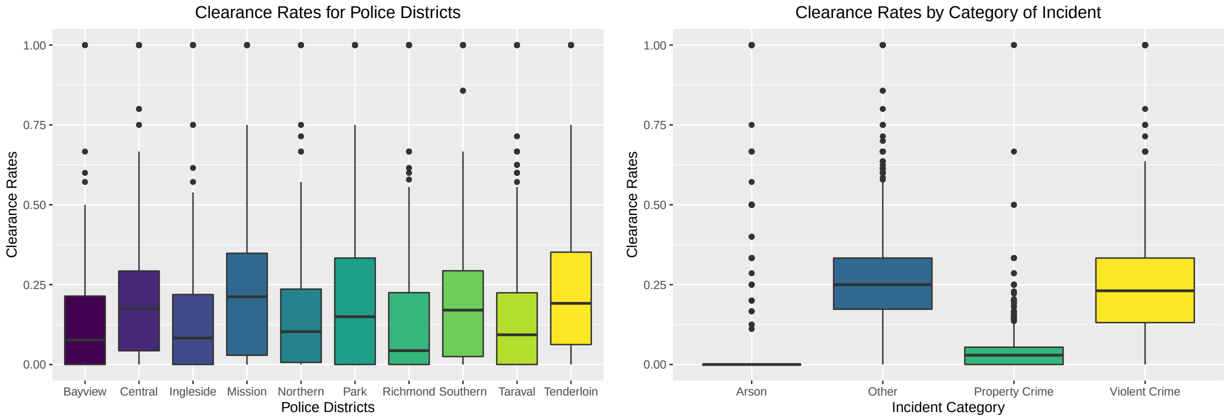


Figure 1: Left: Clearance rates across Police Districts; Right: Clearance rates by category of incident

The boxplot on the left shows that Mission and Tenderloin have the some of the highest average clearance rates amongst the ten police districts, while Richmond sees the lowest average clearance rates. This plot forms the motivation for fitting a hierarchical model, using separate intercepts for each group, i.e. Police District, to account for the different baseline clearance rates amongst the districts. The boxplot on the right displays how the average clearance rates vary for each category of incident. As can be seen from the plot, violent crimes have high clearance rates, while property crimes, larceny thefts in particular, have extremely low clearance rates.

Additional exploratory analysis focused on other predictors are also carried out. Interaction effects among the predictors are also explored, but there does not seem to be any significant trends to model on. Predictors like $Season$, $MonthOfYear$, $Holiday$ and $IncCategory$ do not significantly affect the clearance rates across the police districts. The graph below shows the clearance rates by time of day, across all police districts. It can be observed that the trend and magnitude of the rates vary across the districts. This suggests that random slopes could possibly be modeled for each group. Additional plots of the predictor variables and their interactions against clearance rates can be found in the appendix.

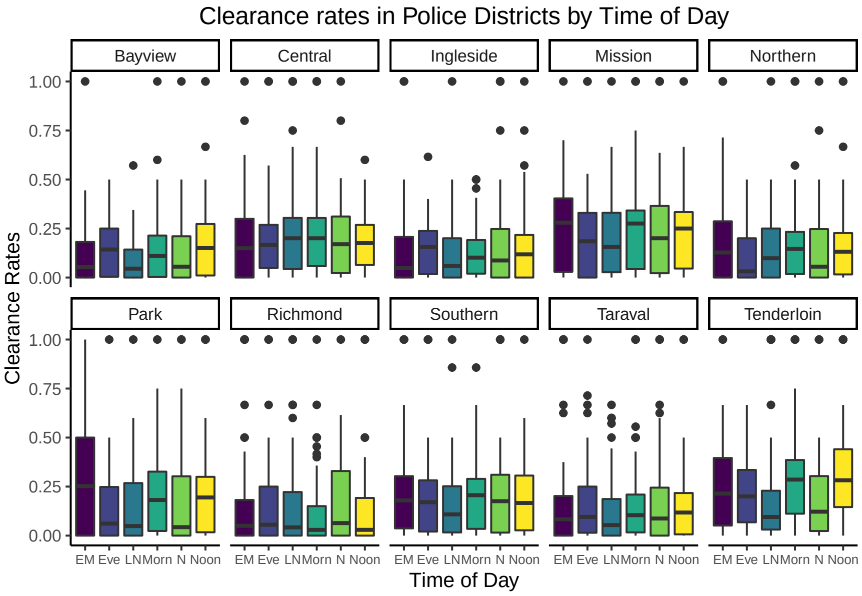


Figure 2: Clearance rates by Time of Day, by Police Districts, EM : Early Morning, Eve : Evening, LN : Late Night, Morn: Morning, N : Night, Noon:Noon

Model

Model selection was performed by accounting for different interactions and effects, which included both random and fixed effects. The baseline model that was fit for this analysis is a hierarchical logistic regression model, that models the clearance rates as the response variable, and $PoliceDistrict$ as the random intercept. This is carried out to account for the random effects that the different police districts might contribute to the model.

The hierarchical model is fit with all the main effects, i.e. $Season$, $Month$, $isWeekend$, $Holiday$, $TimeOfDay$ and $IncCategory$. Both $Season$ and $Holiday$ were found to be insignificant predictors. To confirm this, ANOVA Chi-squared tests are performed, and the p-value was found to be significantly greater than 0.05, implying that these predictors do not add any additional information to the model.

In addition, random slopes for $TimeOfDay$ were also considered for the hierarchical model. Although the solution for this model does converge, the solution obtained has a singular fit, which is an indicator that the model might be too complex for the underlying data and hence, possibly over-fitting. The random intercept model is then retained as the final model. This model has an intercept standard deviation of 0.33, which is not very high. The model equation is as shown below:

$$y_i|x_i \sim Bernoulli(\pi_i);\ i = 1, 2, \dots, 6475;\ j = 1, 2, \dots, 10$$

$$log(\frac{\pi_i}{1-\pi_i}) = (\beta_0 + \gamma_{0j|i|}^{policedistrict}) + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \beta_3 * x_{i3} + \beta_4 * x_{i4};$$

$$\gamma_{0j} \sim N(0, \sigma^2_{policedistrict})$$

where, x_{i1} is $Month$, x_{i2} is $isWeekend$, x_{i3} is $TimeOfDay$, x_{i4} is $IncCategory$.

The coefficients of the model are as shown in the table below:

	Dependent variable:
	cbind(Resolution_resp, countIncidents - Resolution_resp)
Month2	0.024 (0.024)
Month3	−0.063*** (0.024)
Month4	−0.035 (0.024)
Month5	−0.083*** (0.024)
Month6	−0.202*** (0.025)
Month7	−0.137*** (0.024)
Month8	−0.128*** (0.024)
Month9	−0.120*** (0.024)
Month10	−0.183*** (0.024)
Month11	−0.086*** (0.025)
Month12	−0.074*** (0.026)
isWeekend1	−0.046*** (0.012)
TimeOfDayEvening	−0.162*** (0.019)
TimeOfDayLate Night	−0.199*** (0.021)
TimeOfDayMorning	−0.190*** (0.019)
TimeOfDayNight	−0.092*** (0.022)
TimeOfDayNoon	−0.020 (0.019)
IncCategoryOther	1.142*** (0.082)
IncCategoryProperty Crime	−1.018*** (0.083)
IncCategoryViolent Crime	0.952*** (0.083)
Constant	−2.003*** (0.134)
Observations	6,475
Log Likelihood	−13,021.150
Akaike Inf. Crit.	26,086.310
Bayesian Inf. Crit.	26,235.370
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: Coefficients of hierarchical logistic regression with random intercept

From the results in Table 1, it can be observed that the intercept indicates that across all police districts, the odds of a crime getting cleared are 0.13, when all the categorical predictors are at their baseline values. The coefficients are exponentiated for easier interpretation of the effects. For a fixed police district, day of the week, time of day and category of incident, the odds of a crime getting cleared is 18% less likely for the month of June, compared against the baseline month of January. For a police district, month, time of day and category of incident, the odds of a crime getting cleared on the weekend is 5% lesser than a weekday. For a fixed police district, month, day of week, and category of incident, the odds of a crime getting cleared is 18% less if it occurs late night, compared to if it occurs early morning. For a given police district, month, day of week and time of day, the odds of a property crime getting cleared is 64% less likely than the odds of an arson charge getting cleared. These results are consistent with what was observed in the exploratory analysis.

The dotplot below in figure 3 further highlights the variation in clearance rates across the police districts. As was observed from the exploratory analysis, Tenderloin and Mission have the greatest positive intercept, which are significantly different than zero.

$PoliceDistrict

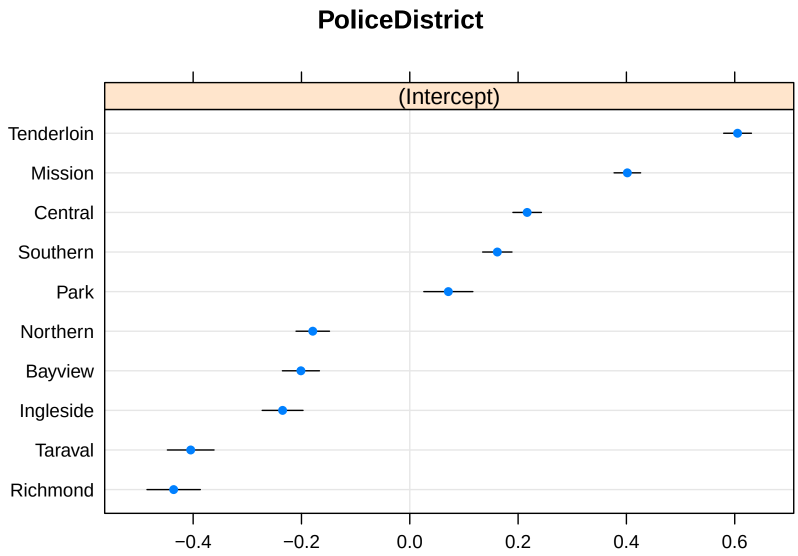


Figure 3: Dotplot of random effects

Conclusion

Overall, the model was valid and addressed questions about the factors that affect the clearance rates of crimes across the ten police jurisdictions of San Francisco. It was observed that *Season* and *Holiday* did not have a significant effect on the crime clearance rates. Some effects which tend to impact crime clearances are the time of day in which an incident occurs, if the crime takes place on a weekend, and the type of crime that occurs. Specifically, average clearance rates observed were lower on the weekends than on weekdays, and for incidents occurring late nights compared to early mornings. The odds of clearance rates for property crimes are also very low compared to other categories of crimes. This ties in with Proposition 47 that was passed in California, which updated the lower limit for thefts to be considered for prosecution to $950, and hence, a large number of these cases are unreported or uncleared.

The dataset is subject to change, and hence, some records may be modified or removed to comply with court orders to to keep certain records private during internal investigations. This dataset also does not necessarily capture all data surrounding policing and crime, and hence, conclusions drawn out of this analysis might not be exactly accurate. Since the data is grouped at a combination of different factors, which makes it very granular, we observe that not all groupings have enough observations recorded against them to fit a model on. Certain categories could be aggregated out so that there are sufficient observations against each grouping.

Appendix

Data Dictionary

A summary of all variables included in the dataset are as under:

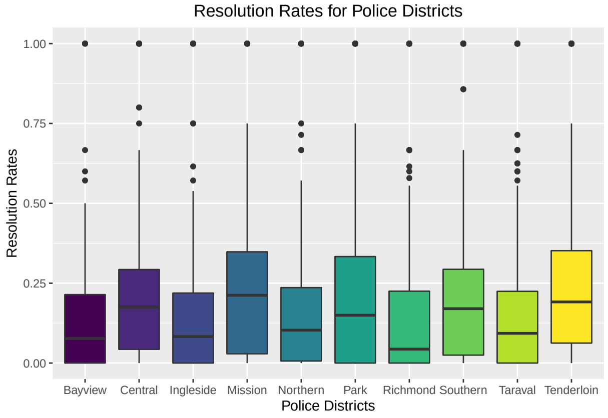
Variable	Description
Incident Datetime	The date and time when the incident occurred
Incident Date	The date the incident occurred
Incident Time	The time the incident occurred
Incident Year	The year the incident occurred, provided as a convenience for filtering
Incident Day of Week	The day of week the incident occurred
Report Datetime	Distinct from Incident Datetime, Report Datetime is when the report was filed
Row ID	A unique identifier for each row of data in the dataset
Incident ID	This is the system generated identifier for incident reports. Incident IDs and Incident Numbers both uniquely identify reports, but Incident Numbers are used when referencing cases and report documents
Incident Number	The number issued on the report, sometimes interchangeably referred to as the Case Number. This number is used to reference cases and report documents
CAD Number	The Computer Aided Dispatch (CAD) is the system used by the Department of Emergency Management (DEM) to dispatch officers and other public safety personnel. CAD Numbers are assigned by the DEM system and linked to relevant incident reports (Incident Number). Not all Incidents will have a CAD Number. Those filed online via Coplogic (refer to “Filed Online” field) and others not filed through the DEM system will not have CAD Numbers
Report Type Code	A system code for report types, these have corresponding descriptions within the dataset
Report Type Description	The description of the report type, can be one of: Initial; Initial Supplement; Vehicle Initial; Vehicle Supplement; Copologic Initial; Copologic Supplement
Filed Online	Non- emergency police reports can be filed online by members of the public using SFPD’s self-service reporting system called Coplogic Values in this field will be “TRUE” if Coplogic was used to file the report. Please reference the link below for additional info: (http://sanfranciscopolice.org/reports)

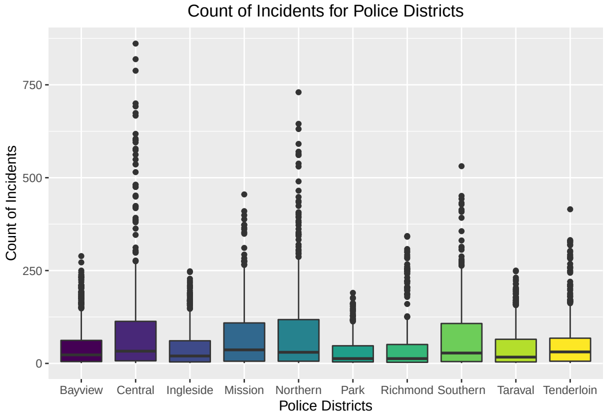
Variable	Description
Incident Code	Incident Codes are the system codes to describe a type of incident. A single incident report can have one or more incident types associated. In those cases you will see multiple rows representing a unique combination of the Incident ID and Incident Code
Incident Category	A category mapped on to the Incident Code used in statistics and reporting. Mappings provided by the Crime Analysis Unit of the Police Department
Incident Subcategory	A subcategory mapped to the Incident Code that is used for statistics and reporting. Mappings are provided by the Crime Analysis Unit of the Police Department
Incident Description	The description of the incident that corresponds with the Incident Code. These are generally self-explanatory
Resolution	The resolution of the incident at the time of the report. Can be one of: • Cite or Arrest Adult • Cite or Arrest Juvenile* • Exceptional Adult • Exceptional Juvenile* • Open or Active • Unfounded Note: once a report is filed, the Resolution will not change. Status changes and/or updates must be provided using a Supplemental Report * Incidents identifying juvenile information are not included in this dataset. Please see the Juvenile Data section for more information. *This is used to create the response variable.*
Intersection	The 2 or more street names that intersect closest to the original incident separated by a backward slash (). Note, the possible intersections will only include those that satisfy the privacy controls
CNN	The unique identifier of the intersection for reference back to other related basemap datasets. For more on the Centerline Node Network see https://datasf.gitbook.io/draft-publishing-standards/standard-reference-data/basemap/street-centerlines-nodes
Police District	The Police District where the incident occurred. District boundaries can be reviewed in the link below. Please note this field is entered by officers and not based on the point. Reference here: https://data.sfgov.org/d/wkhw-cjsf

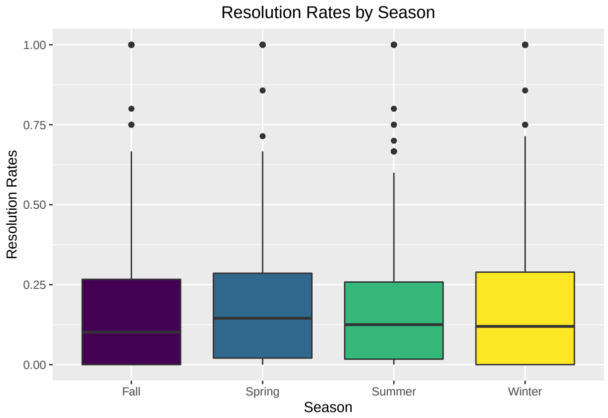
Variable	Description
Analysis Neighborhood	This field is used to identify the neighborhood where each incident occurs. Neighborhoods and boundaries are defined by the Department of Public Health and the Mayor's Office of Housing and Community Development. Please reference the link below for additional info: https://data.sfgov.org/d/p5b7-5n3h Please note this boundary is assigned based on the intersection, it may differ from the boundary the incident actually occurred within
Supervisor District	There are 11 members elected to the Board of Supervisors in San Francisco, each representing a geographic district. The Board of Supervisors is the legislative body for San Francisco. The districts are numbered 1 through 11. Please reference the link below for additional info: https://data.sfgov.org/d/8nkz-x4ny Please note this boundary is assigned based on the intersection, it may differ from the boundary the incident actually occurred within
Latitude	The latitude coordinate in WGS84, spatial reference is EPSG:4326
Longitude	The longitude coordinate in WGS84, spatial reference is EPSG:4326
Point	Geolocation in OGC WKT format (e.g, POINT(37.4,-122.3)

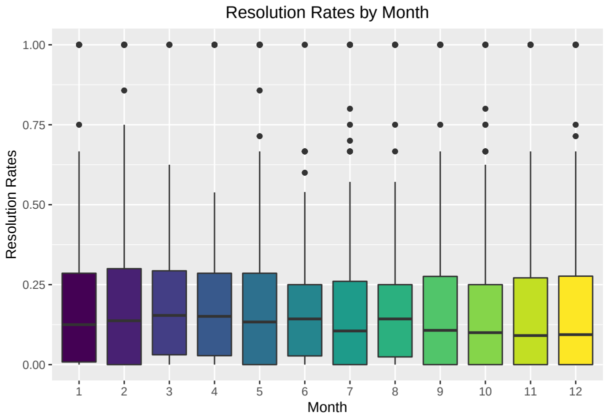
Exploratory Analysis

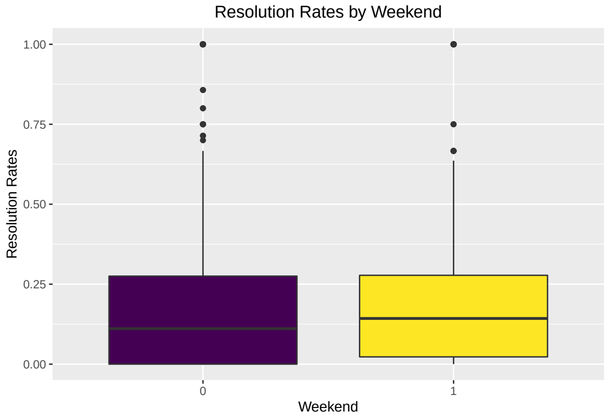
Distribution of response variable against predictors

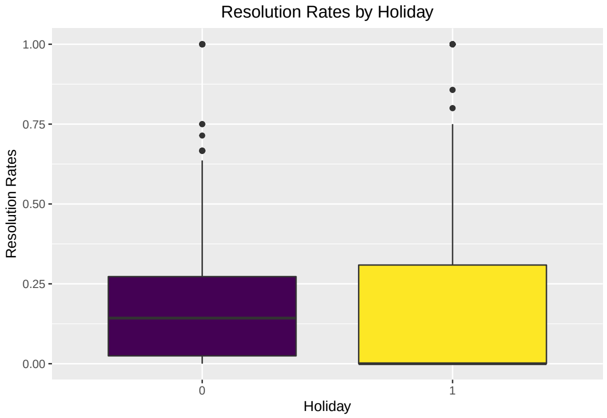


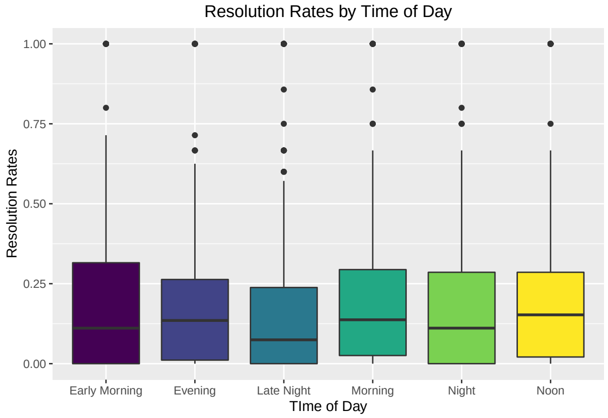


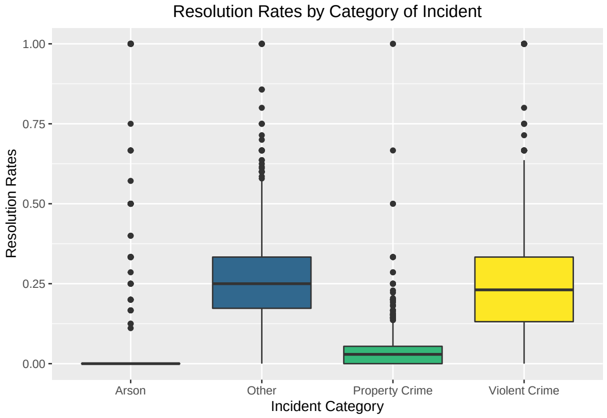

Resolution Rates by Season


Resolution Rates by Month

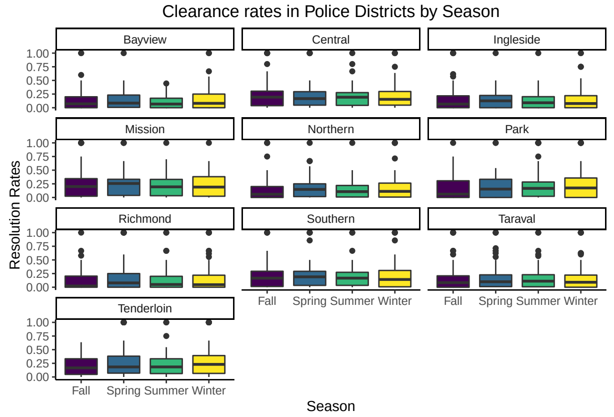

Resolution Rates by Weekend

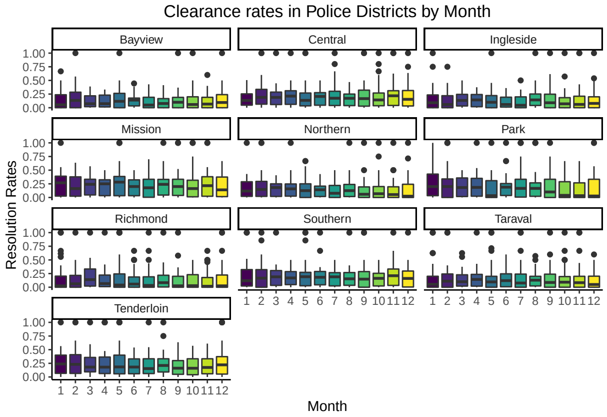

Resolution Rates by Holiday

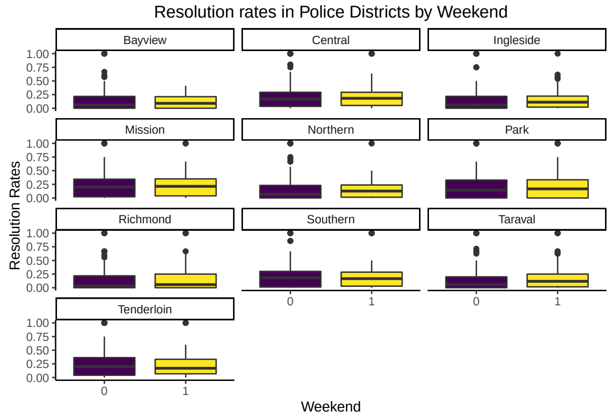

Resolution Rates by Time of Day

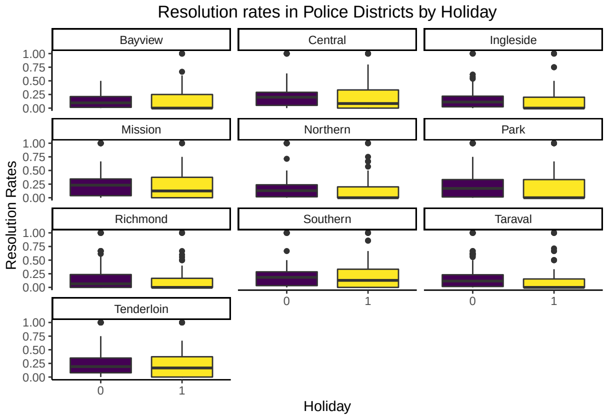

Resolution Rates by Category of Incident

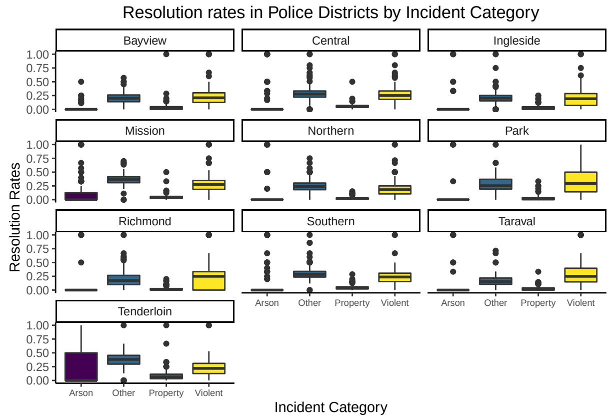
Distribution of response variable against predictors by Police Districts

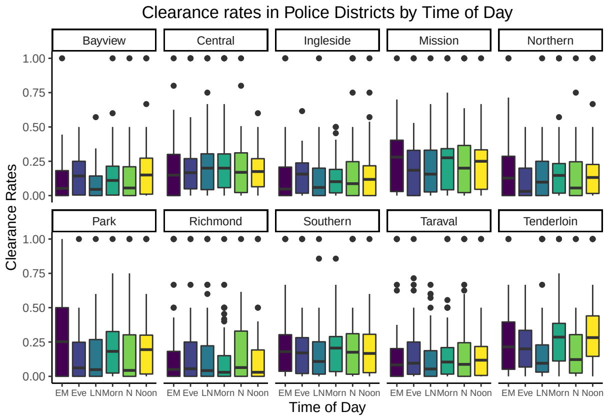


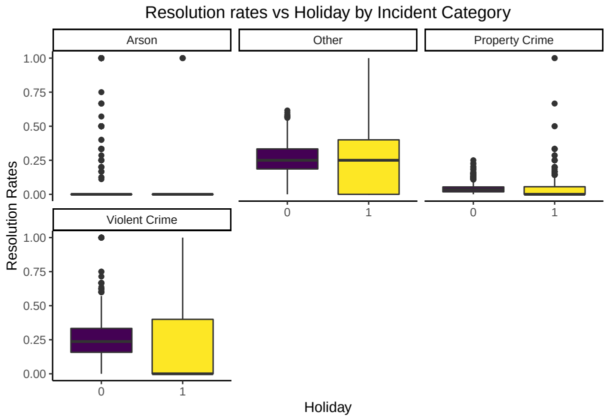


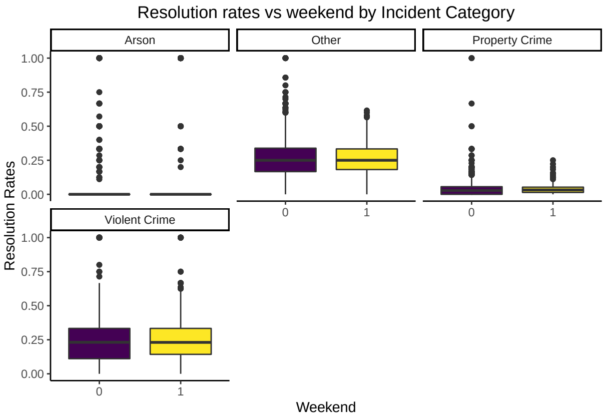


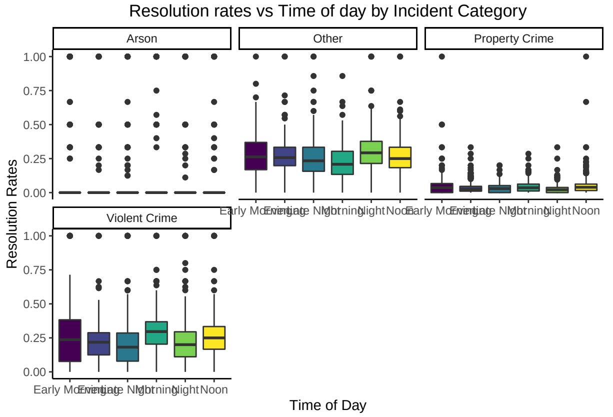


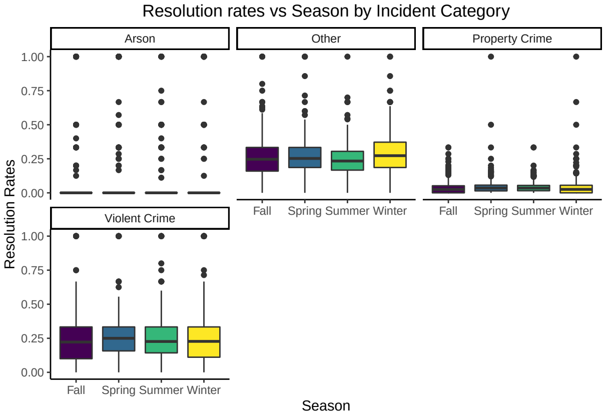


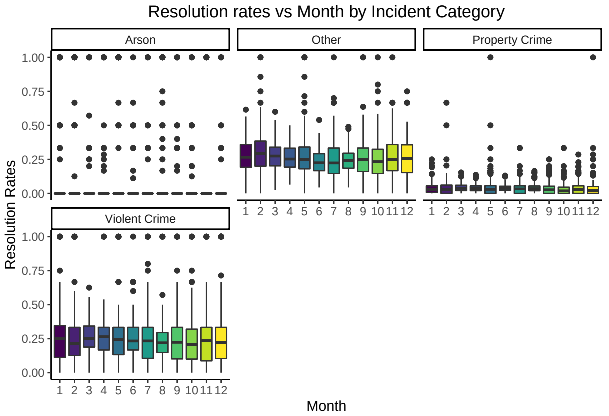












Code

The code can be found in the Github repository here.