Exploratory Data Analysis

- StudentID: 22000176

- Name: JiMin Kim

- 1st Major: Economics

- 2nd Major: AI convergence

This EDA aims to approach the raw_Dallas.dta data to reveal the trends, patterns, or check assumptions using visual techniques. Further more, it seeks to identify key elements for predicting different types of crime. The model created particularly specifies the hour variable and attempts to forecast what kind of crime will occur at each hour, as it has one of the greatest weighted scores among the feature variables that have been selected.

1. Data overview

The data considering the crimesi of Dallas contains 663249 crimes with the following 107 variables.
(['incidentnum', 'UCR_ctype', 'cnt', 'ctype', 'year', 'servicenumberid', 'watch', 'call911problem', 'typeofincident', 'typelocation', 'typeofproperty', 'incidentaddress', 'apartmentnumber', 'reportingarea', 'beat', 'division', 'sector', 'councildistrict', 'targetareaactiongrids', 'community', 'date1ofoccurrence', 'year1ofoccurrence', 'month1ofoccurence', 'day1oftheweek', 'time1ofoccurrence', 'day1oftheyear', 'date2ofoccurrence', 'year2ofoccurrence', 'month2ofoccurence', 'day2oftheweek', 'time2ofoccurrence', 'day2oftheyear', 'dateofreport', 'dateincidentcreated', 'offenseenteredyear', 'offenseenteredmonth', 'offenseentereddayoftheweek', 'offenseenteredtime', 'offenseentereddatetime', 'cfsnumber', 'callreceiveddatetime', 'calldatetime', 'callcleareddatetime', 'calldispatchdatetime', 'specialreportprerms', 'personinvolvementtype', 'victimtype', 'victimname', 'victimrace', 'victimethnicity', 'victimgender', 'victimage', 'victimageatoffense', 'victimhomeaddress', 'victimapartment', 'victimzipcode', 'victimcity', 'victimstate', 'victimbusinessname', 'victimbusinessaddress', 'victimbusinessphone', 'respondingofficer1badgeno', 'respondingofficer1name', 'respondingofficer2badgeno', 'respondingofficer2name', 'reportingofficerbadgeno', 'assistingofficerbadgeno', 'reviewingofficerbadgeno', 'elementnumberassigned', 'investigatingunit1', 'investigatingunit2', 'offensestatus', 'ucrdisposition', 'victiminjurydescription', 'victimcondition', 'modusoperandimo', 'familyoffense', 'hatecrime', 'hatecrimedescription', 'weaponused', 'gangrelatedoffense', 'victimpackage', 'drugrelatedistevencident', 'rmscode', 'criminaljusticeinformationservic', 'penalcode',

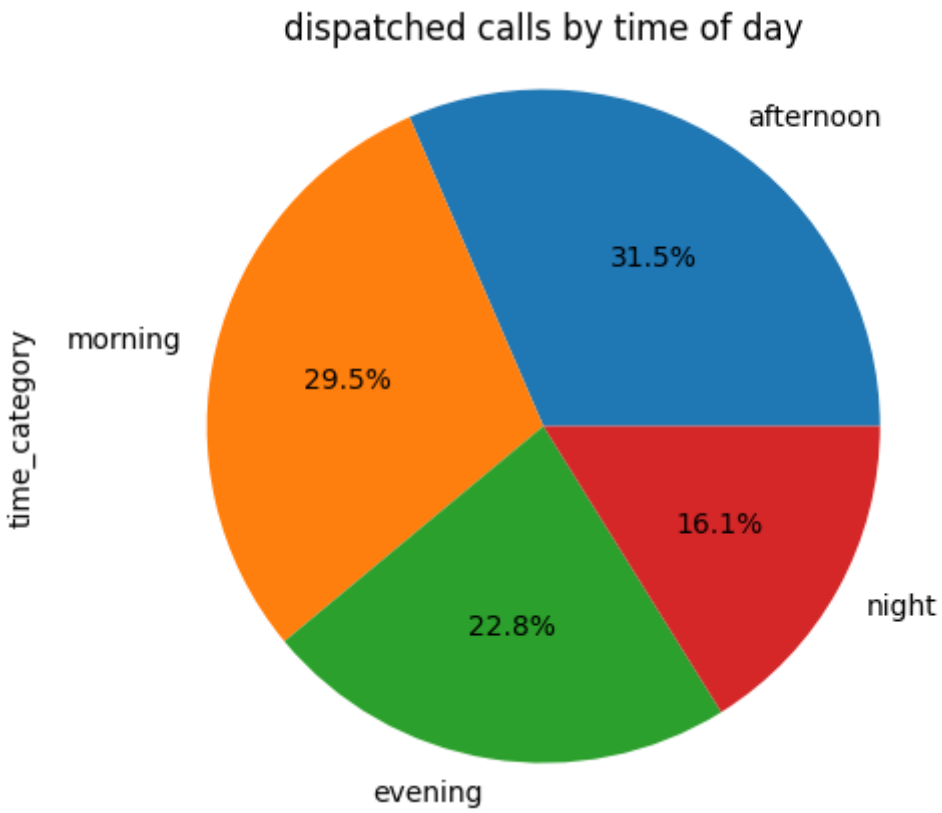
'ucroffensedescription', 'ucrcode', 'offensetype', 'nibrscrime', 'nibrscrimecategory', 'nibrscrimeagainst', 'nibrscode', 'nibrsgroup', 'nibrstype', 'updatedate', 'xcoordinate', 'ycordinate', 'zipcode', 'city', 'state', 'location1', 'address', 'geo_lat', 'geo_long', 'blkidfp00', 'blkidfp10']).

These variables primarily consider five elements of information about the crime in Dallas; crime type, geographic distribution, temporal patterns, clearance status, and demgraphic information. The dataset includes information on the crime type for each occurrence, which may be used to determine the most prevalent forms of crime in Dallas throughout the time period covered by the dataset. The geographic distribution is the second type of information.The dataset contains details on each incident's location, including latitude and longitude coordinates. The dataset also contains information about on each incident's date and time, which may be used to spot trends over time in criminal incidences. For instance, there can be particular times of day or days of the week when crime is more likely to happen. The dataset also contains data on each incident's clearance status, which shows whether the case has been resolved or not. This could be able to report trends in the efficiency of law enforcement activities by looking at clearance rates for certain crimes or in various parts of Dallas.The demographic data is the 5th category. The dataset contains details on the victim's ethnicity and gender. By examining this data, it could be possible to recognize trends in the characteristics of crime suspects and victims, which might assist guide the development of policies that reduce crime and advance equity.

2. Univariate analysis

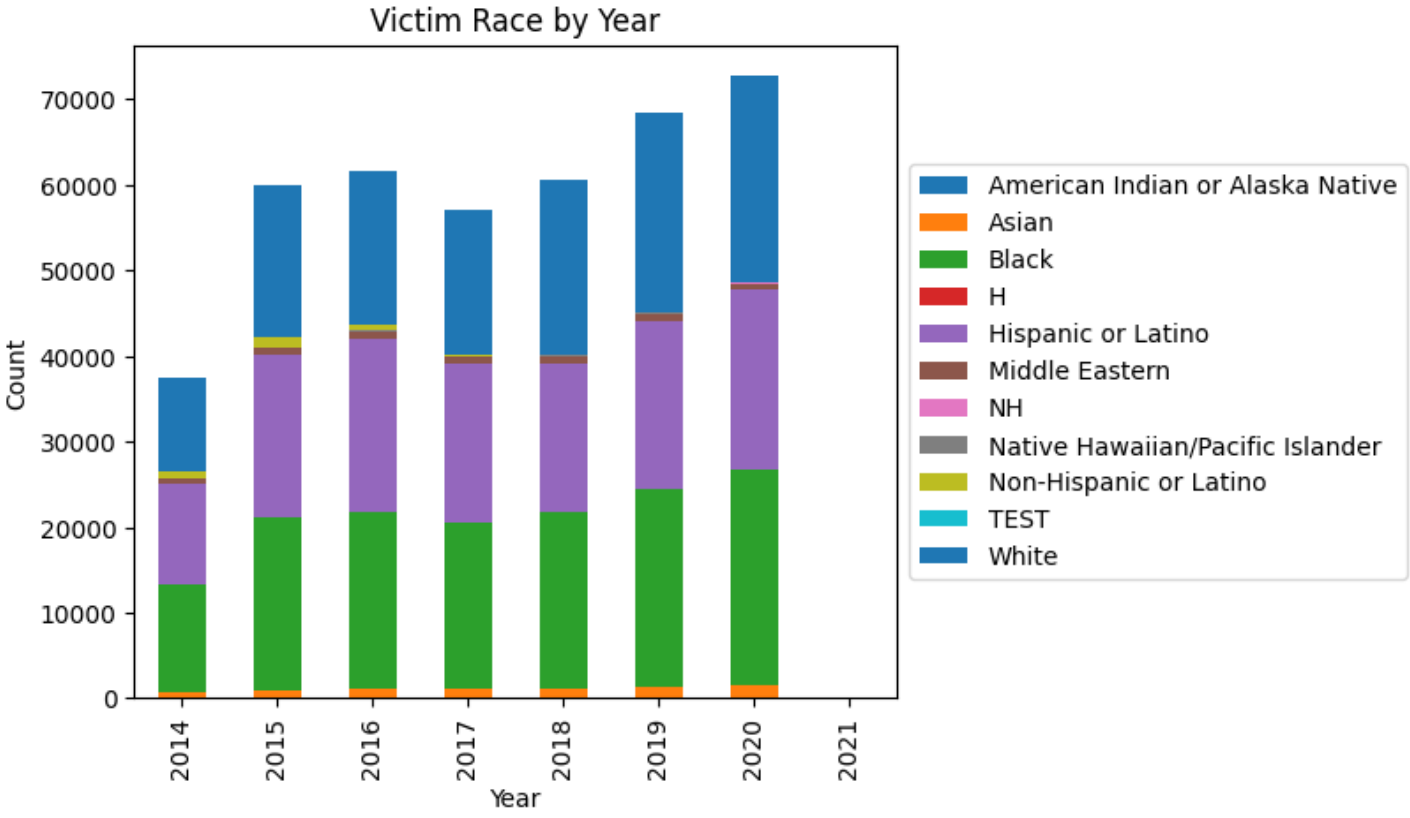
2.1 Dispatched Calls

In the context of crimes, dispatched calls are those that have been given to a police officer or another emergency responder. They are calls that law enforcement organizations have responded to. This graph takes a look at the dispatched calls made throughout each hour of the day in each location. Night corresponds to as 0–6, morning as 6–12, afternoon as 12–18, and evening as 18–0.



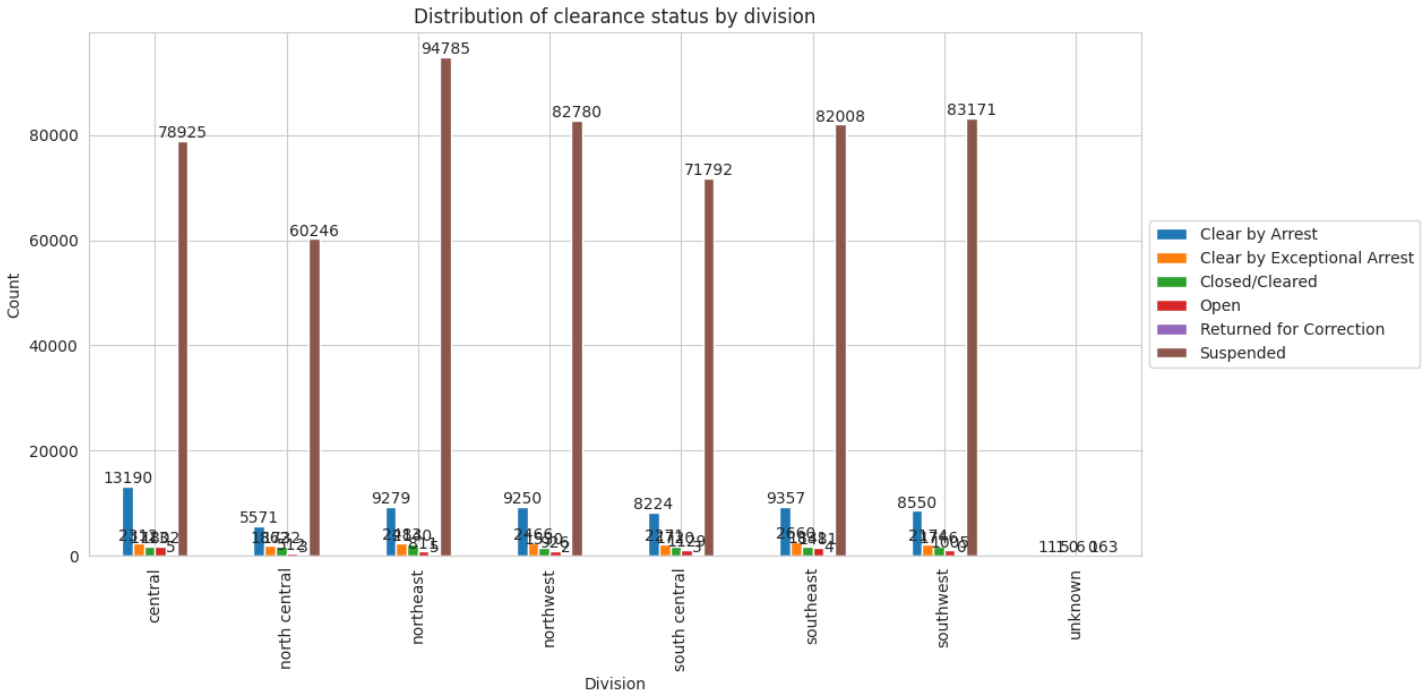
From this, I was able to see that the most freqent time of dispatched calls were in the afternoon.

2.2 Victim Race by Year



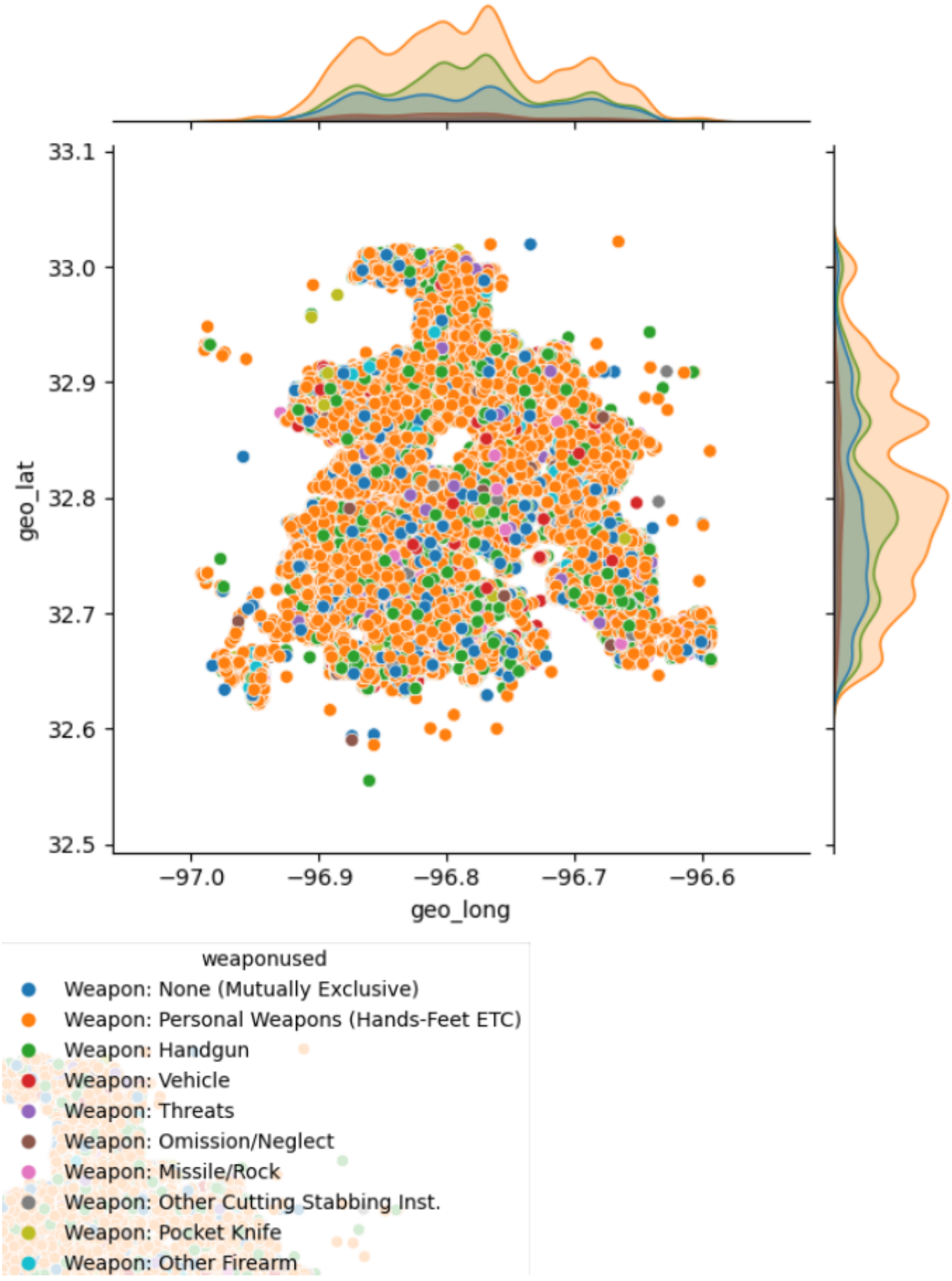
According to this graph, crime rates increased each year with American Indians/Alaska Natives, Hispanics/Latinos, and Blacks being the main victims.

2.3 Clearance Status



This graph revealed that most offenses had been suspended or temporarily stopped due to an absence of evidence or proof. The second most frequent offense status denotes that the arrest was successful and the defendant was charged with the crime, by beling cleared by arrest. The graph demonstrates that fewer individuals are being arrested than would be expected based on the ratio of arrests to offenses that have been put on hold.

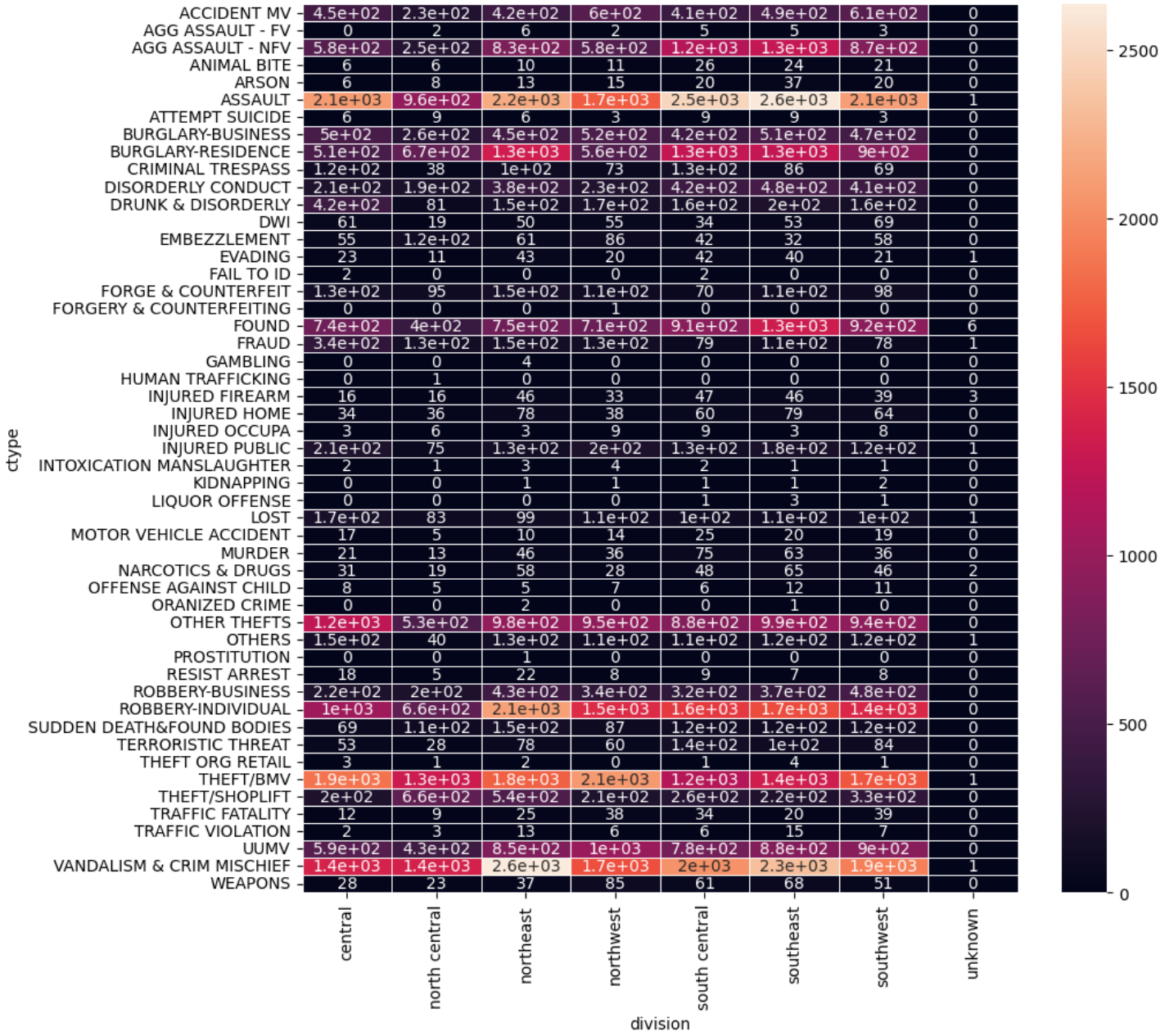
2.4 Top 10 weapons used throughout Dallas



It can bee seen that the top ten weapons used in Dallas are 'None (Mutually Exclusive)','Personal Weapons (Hands-Feet ETC)', Handgun','Vehicle','Threats', 'Omission/Neglect','Missile/Rock','Other Cutting Stabbing Inst.','Pocket Knife', and 'Other Firearm'.

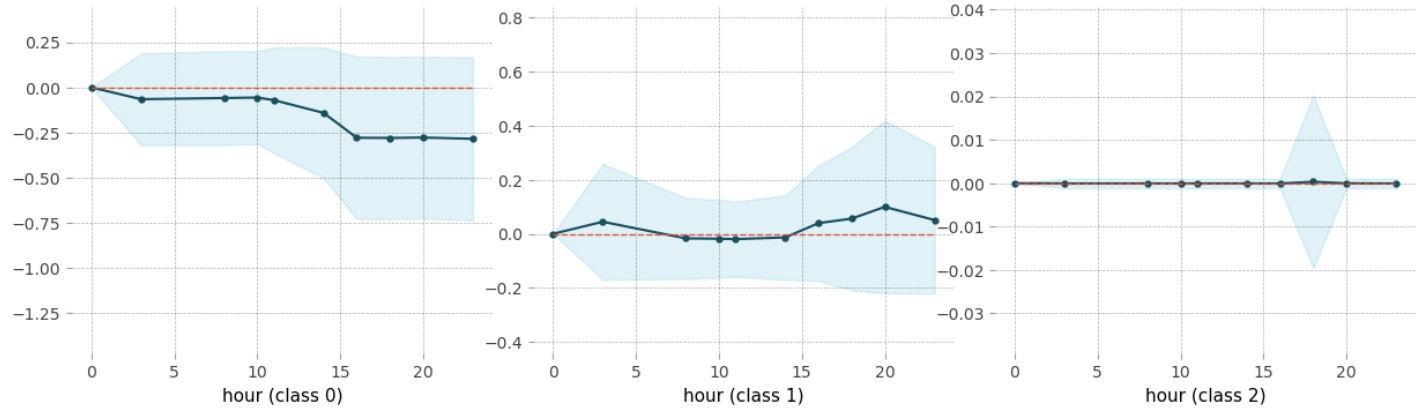
3. Multivariate analysis

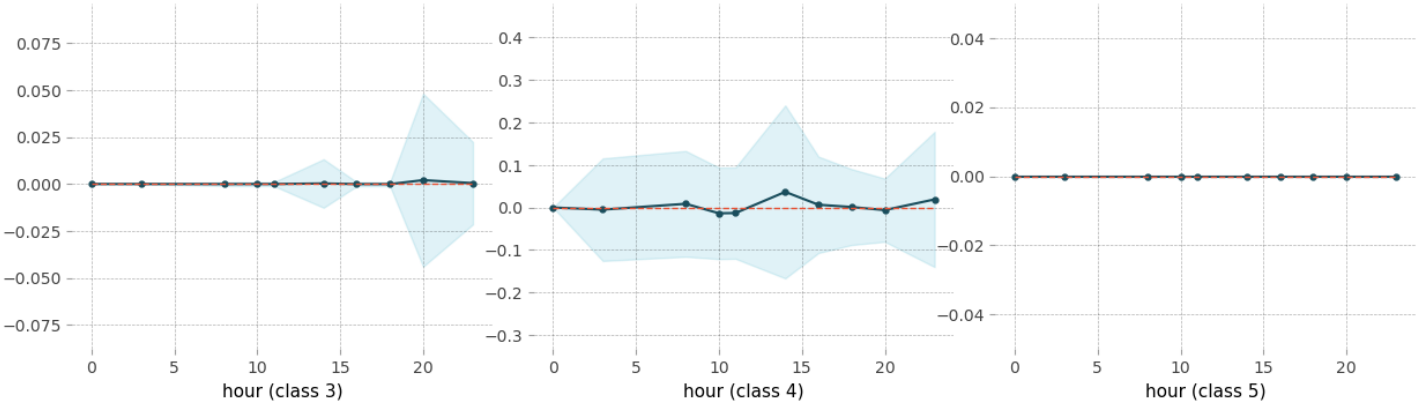
3.1 Crimes by Division

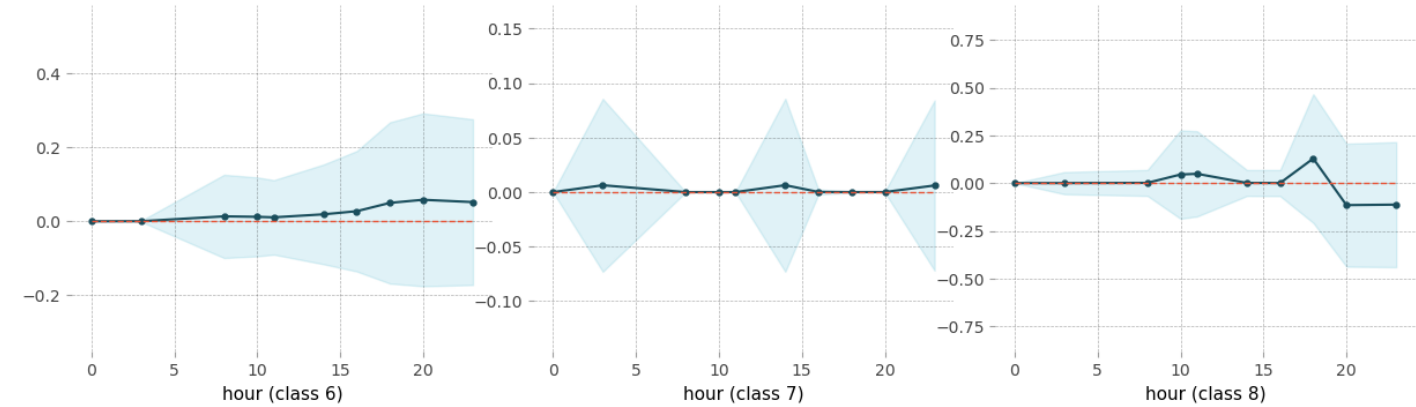


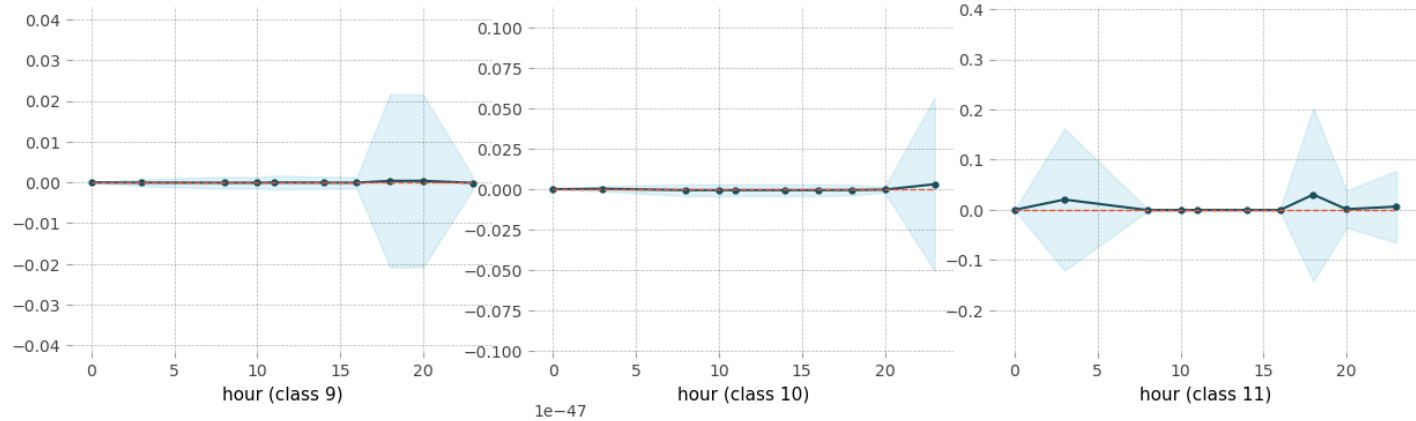
Using this heatmap, it was possible to examine how different sorts of crimes are distributed throughout the different geographical areas. Theft and BMV are widespread across the country, although they are particularly prevalent in the northwest. The next most frequent crimes in Dallas were vandalism and criminal mischief and residential burglary, respectively.

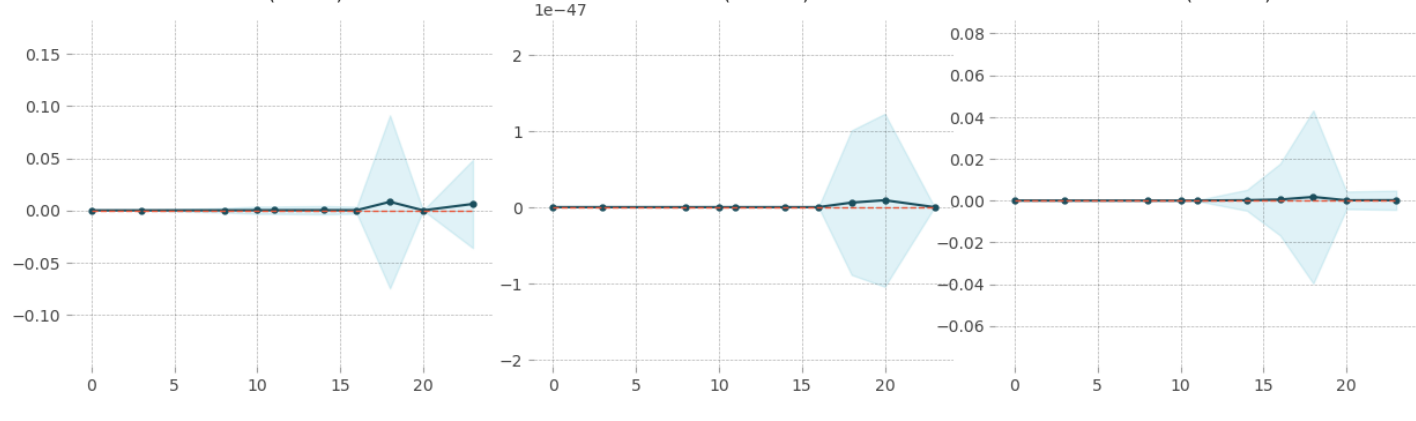
3.2 Prediction


hour (class 0)
hour (class 1)
hour (class 2)


hour (class 3)
hour (class 4)
hour (class 5)


hour (class 6)
hour (class 7)
hour (class 8)


hour (class 9)
hour (class 10)
hour (class 11)


1e−47

legend: {'': 0, 'ALL OTHER OFFENSES': 1, 'ANIMAL OFFENSES': 2, 'ARSON': 3, 'ASSAULT OFFENSES': 4, 'BRIBERY': 5, 'BURGLARY/ BREAKING & ENTERING': 6, 'COUNTERFEITING / FORGERY': 7, 'DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY': 8, 'DISORDERLY CONDUCT': 9, 'DRIVING UNDER THE INFLUENCE': 10, 'DRUG/ NARCOTIC VIOLATIONS': 11, 'EMBEZZELMENT': 12, 'EXTORTION/ BLACKMAIL': 13, 'FAMILY OFFENSES, NONVIOLENT': 14, 'FRAUD OFFENSES': 15, 'GAMBLING OFFENSES': 16, 'HOMICIDE OFFENSES': 17, 'HUMAN TRAFFICKING': 18, 'KIDNAPPING/ ABDUCTION': 19, 'LARCENY/ THEFT OFFENSES': 20, 'LIQUOR LAW VIOLATIONS': 21, 'MISCELLANEOUS': 22, 'MOTOR VEHICLE THEFT': 23, 'PEEPING TOM': 24, 'PORNOGRAPHY/ OBSCENE MATERIAL': 25, 'PUBLIC INTOXICATION': 26, 'ROBBERY': 27, 'SEX OFFENSES': 28, 'STOLEN PROPERTY OFFENSES': 29, 'TRAFFIC VIOLATION - HAZARDOUS': 30, 'TRAFFIC VIOLATION - NON HAZARDOUS': 31, 'TRESPASS OF REAL PROPERTY': 32, 'WEAPON LAW VIOLATIONS': 33}

Using the LightGBM, a machine learning gradient boosting classifier that uses tree-based learning techniques, it was possible to predict the kinds of crimes that would occur at certain times of the day.

The following variables were used: beat (the smallest police geographic region), division, year, and date the incident. In order to analyze the dataset, the dataset was first split into X_train, X_test, Y_train, and Y_test. Nibrscrimecategory was set as the target variable.

These were the user defined hyperparameters: {'boosting':'gbdt', 'objective':'multiclass', 'num_class':39, 'max_delta_step':0.9, 'min_data_in_leaf': 21, 'learning_rate': 0.4, 'max_bin': 465, 'num_leaves': 41}.

The aforementioned parameters were used to build a LightGBM classifier model, which was then trained using the X_train and y_train datasets. The model generated y_pred predictions based on the X_test data. Since the accuracy rating was 66.6725%, the parameters probably need to be adjusted.

The graphs above are PDPs(Partial Dependence Plots), which visualizes the relationship between a target variable and a feature of interest. This allows us to visualize the effect of a feature on a predicted outcome of a model, while other features are held constant. In the case above, the 'hour' variable was used to visualize the relationship between the hour of a day and a probability of each category of nibrscrimecategory. The green line shows the average relationship between the hour feature and the model's predicted outcome, and the blue shading around the green line shows the level of confidence. A wider blue shading indicated more uncertainty, because it shows wider variance.

4. Suggestion

For the PDP above, I used th hour variable because from all the feature variables, hour had one of the highest scores on the weights. This feature importance was calculated with the eli5 library. This means that the hour feature is one of the most important variables for predicting crime categories. For example,driving under the influence, liquor law violations,and robbery all tend to happen during the

evenings and night, so the police department could take this into account to place their staff members at the adequate time and adequate period. Also, the citizens of Dallas should be aware of the types of crime that is more probable at each hour.