University at Buffalo

Buffalo Crime and Response Time Predictions

Team 4

Team Members

Brandon Gerritz	50072844
Govarthanan Balasubramaniyan	50379755
Nishit Chaudhry	50354103
Rahul Pratap Singh	50370509
Riona Almeida	50412024
Utkarsh Dwivedi	50360426

MGS 616 - Predictive Analytics

Professor: Dr.Sanjukta Das Smith

Fall 2021

Index

1. Problem Statement/Business Problem

Crime is an ever-growing occurrence in the US with the FBI reporting a total of 2,109.9 property crimes per 100,000 people and 379.4 violent crimes per 100,000 people in 2019. For our project, we looked locally at the city of Buffalo to find patterns that could help law enforcement decrease the rate of violent and non-violent crime in the area. Upon our initial investigation, we discovered that response time by law enforcement officers was spread out over an extensive range. We decided to look into this further and come up with a predictive model that would predict the response time of law enforcement officials based on the input variables.

2. Data Collection and Preparation:

The data required for crime incident analysis for the city of Buffalo was procured from two sources.

1. Crime Incidents: This dataset contains crime incidents in the City of Buffalo and is provided by the Buffalo Police Department. This dataset is updated daily and offers a preliminary look at crime reports in the City of Buffalo, for the analysis data available as of October 12, 2021, was used and was filtered for a start date of January 1, 2018.
2. 2021 FFIEC Census Report: This data set was provided by the Federal Financial Institutions Examination Council, and it provides a detailed summary of Census Housing Information based on demographic, income, population, and housing.

The datasets were merged into a single dataset by using the field 'Census Block 2010' to establish a relationship between them and were further cleaned using Tableau Prep Builder 2021.3. In the cleaning step, null values were removed, columns with redundancies such as single repetitive values were removed, data types were reassigned appropriately and a

calculated field ‘Response_Time’ was created based on the difference between the values of ‘Incident_Datetime’ and ‘Created_at’ fields. Some discrepancies were noticed such as missing values for ‘Incident_Day_of_Week’ and they were rectified by calculating the appropriate day of the week.

The cleaned dataset contains 40 fields and 48724 records, below are the details in regards to the metadata for the dataset

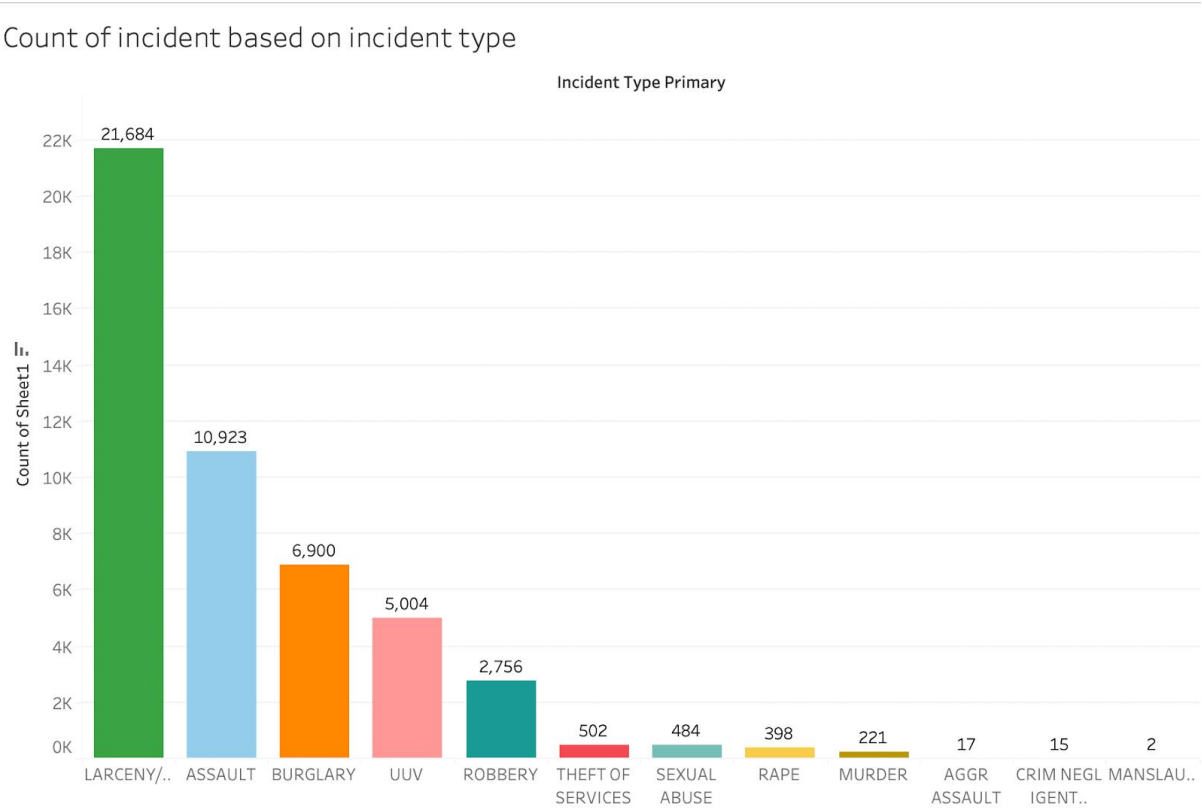
Column Name	Type and Description	Column Name	Type and Description
Incident_month	Text, describes the month of incident occurrence	**2021 Est. Tract Median Family Income**	Number, estimated median family income of the given tract for 2021
Incident_day_of_week	Text, describes the day of incident occurrence	**2015 Tract Median Household Income**	Number, median family income of the given tract for 2015
Longitude	Number, longitude of the location	**Tract Population**	Number, Population of the tract
Latitude	Number, latitude of the location	**Tract Minority %**	Number, minority percentage of the population
Response_time	Number, calculated field in seconds by difference between time stamp of created_at and incident_datetime	**Number of Families**	Number, count of families in that tract
Incident_datetime	Datetime, timestamp of when the incident occurred	**Non-Hisp White Population**	Number, Population of non-

			Hispanic white population
incident_type_primary	Text, sub-classification of the incident	**Tract Minority Population**	Number, population of the minority for the given tract
Zip	Number, zip code where the incident occurred	**American Indian Population**	Number, Population of American Indians in that tract
created_at	Datetime, timestamp of when the police responded	**Asian/ Hawaiian/ Pacific Islander Population**	Number, Population of Asian, Hawaiian and Pacific islanders
hour_of_day	Number, what hour of the day the incident occurred	**Black Population**	Number, Population of the people who are black
parent_incident_type	Text, parent category of the incident	**Hispanic Population**	Number, population of people who are Hispanic
Neighbourhood	Text, neighbourhood the incident occurred in	**Other Population/ Two or More Races**	Number, population of people belonging to other minority groups
police district	Text, police district the incident took place in	**Total Housing Units**	Number, total housing units in that tract

council district	Text, council district the incident took place in	**1- to 4- Family Units**	Number, number of houses in the tract which have at most 4 rooms
census tract 2010	Census tract, Redundant	**Median House Age (Years)**	Number, median age of the households in that tract
census block group 2010	Redundant	**Owner Occupied Units**	Number, number of housing units occupied by owners
census block 2010	Redundant	**Vacant Units**	Number, number of vacant house units in that tract
Tract Income Level	Text, income level of the census tract	**Owner Occupied 1- to 4- Family Units**	Number, number of houses in the tract which have at most 4 rooms and occupied by the owner
% Below Poverty Line	Number, percentage of people living below poverty line for that tract	**Renter Occupied Units**	Number of houses on rent

3. Exploratory Data Analysis:

Exploratory Data Analysis is a key … On performing optimum EDA, it becomes easier to choose necessary explanatory variables. A lot of interesting insights were observed based on descriptive analytics. The census data also contributed to some notable findings.

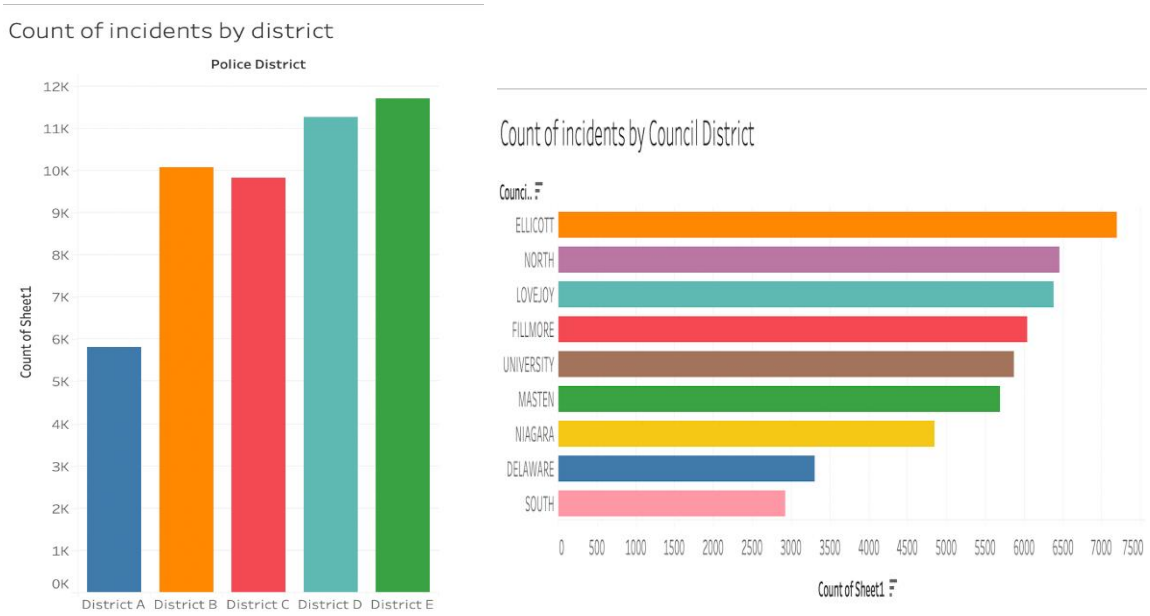


Visualization of Count of incident-by-incident type primary

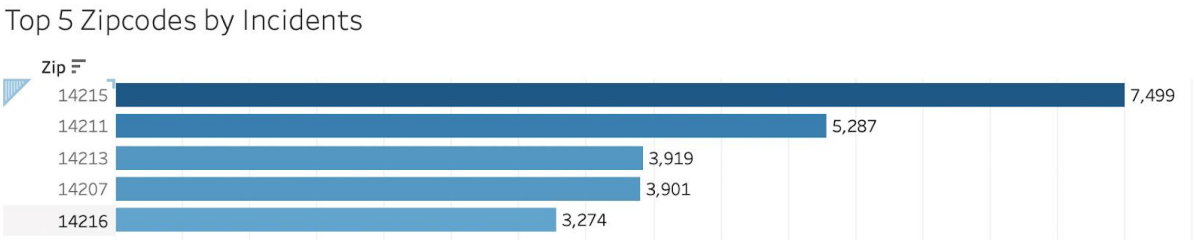
It is observed that Larceny/ Theft, Assault and Burglary are some of the top types of incidents that frequent Buffalo localities.

3.1 Observations based on Location

The dataset included two types of districts - Police and Council. As far as Police Districts are concerned, **District B, D and E** account for the most incident calls**.** By Council district, **Ellicott, North, Lovejoy and Filmore** seem to report a lot of incidents. Furthermore, we can also see the most badly affected zip codes.



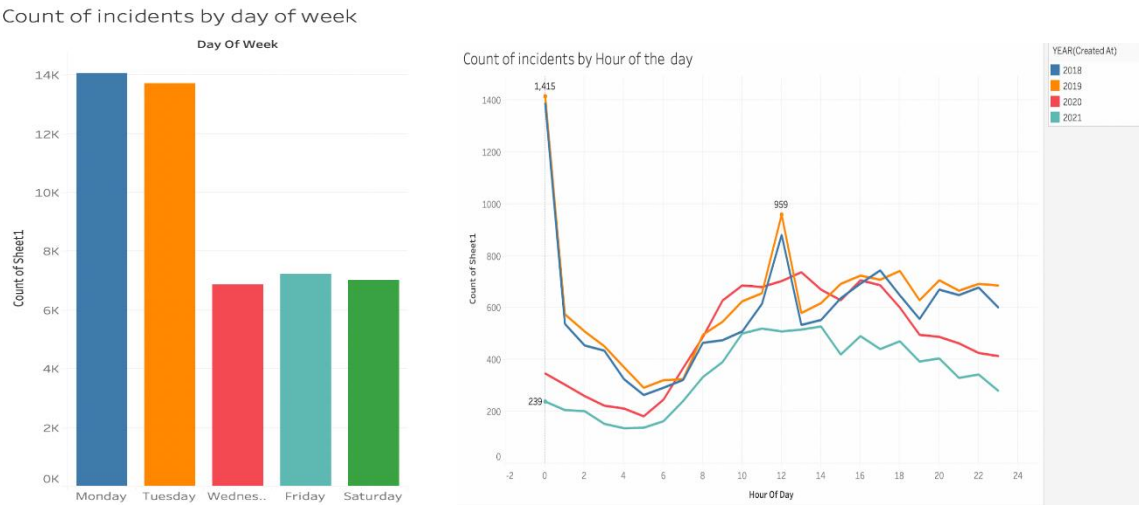
Visualization of Count of incident-by-Council district and police districts



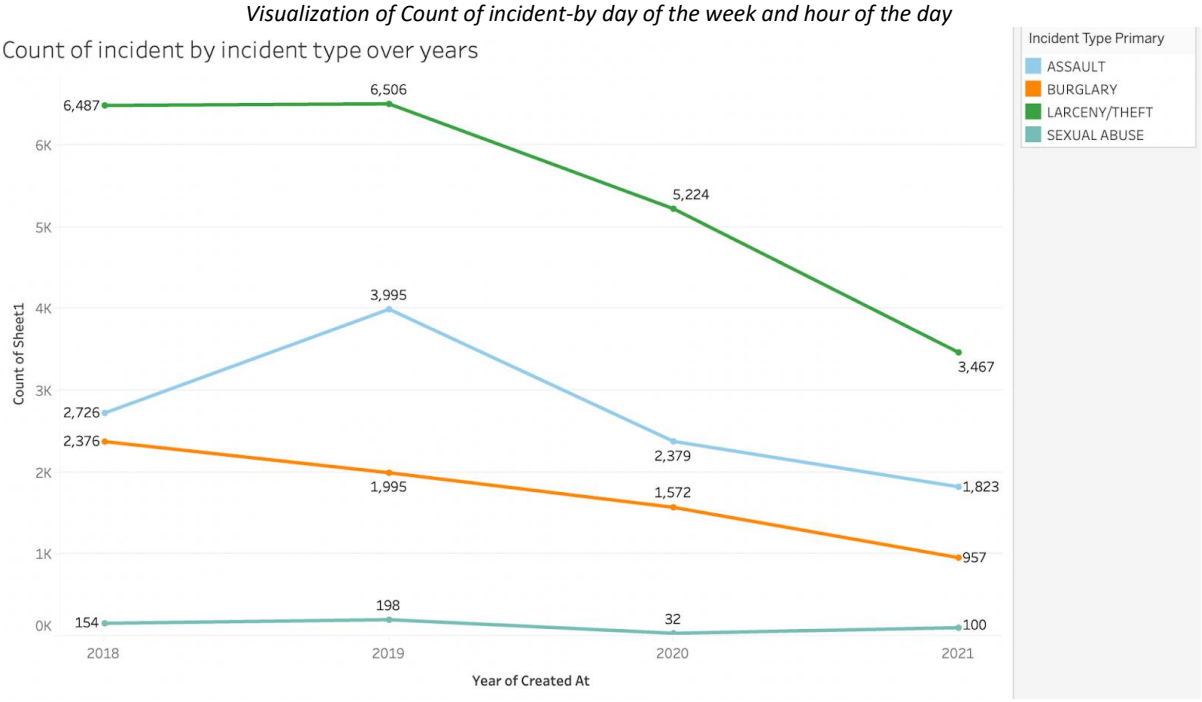
Visualization of Count of incident-by-Zip codes

3.2 Observations related to Incident time:

Mondays have the greatest number of incidents in the week. Incidents seem to also be noted and recorded most at 12 AM. Overall there is a decline in the number of cases in Buffalo over the years (2018-2021). But as far as types of incidents goes, Sexual Abuse is seeing a steady increase in the number of cases in 2021.



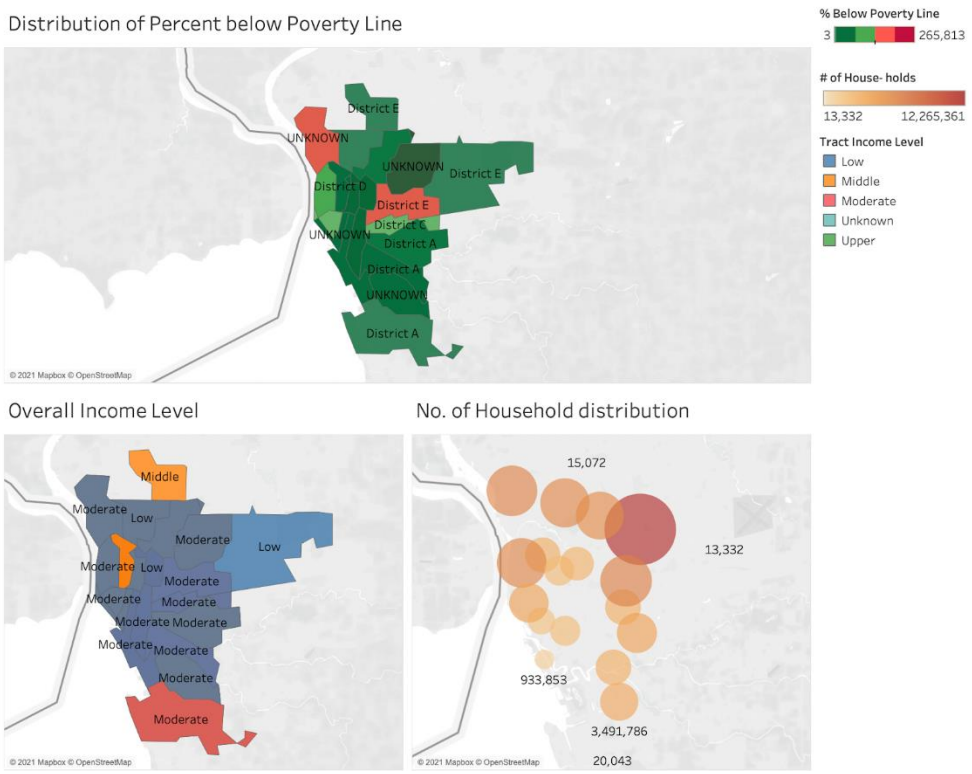
Visualization of Count of incident-by day of the week and hour of the day



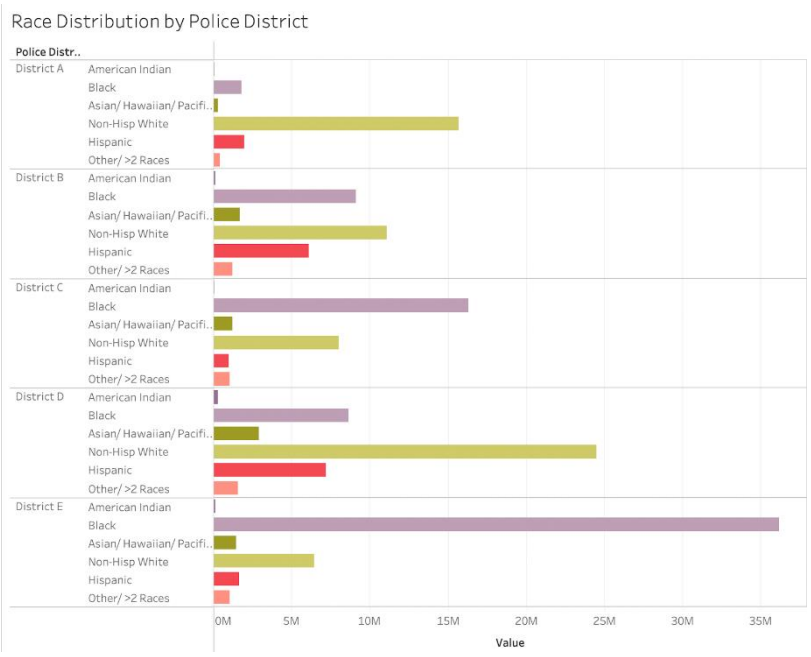
Trend of top incidents based on the year

3.3 Observations based on Census Data -

The census data also accounted for some interesting observations. The tract-level of income seems to be lowest towards District E, which also has an unusually large number of households per area. We were also able to establish a distinguishment of the racial distribution in each police district.



Buffalo Census Data Visualization



Race Distribution by Police district

4. Model Building

Predicting Response Time

We imported the file that we have prepared from the tableau prep builder with 48724 rows and 40 columns.

Feature Selection:

We have made the role-based assignment for each variable. We have rejected Created at, Incident date-time fields as we have extracted the month, hour of the day columns out of it. We have also rejected the 2021 Estimated tract median family income as it's an estimated value. The Calculated variable Response time is made as the target variable.

Variables - FIMPORT

(none) not Equal to ...

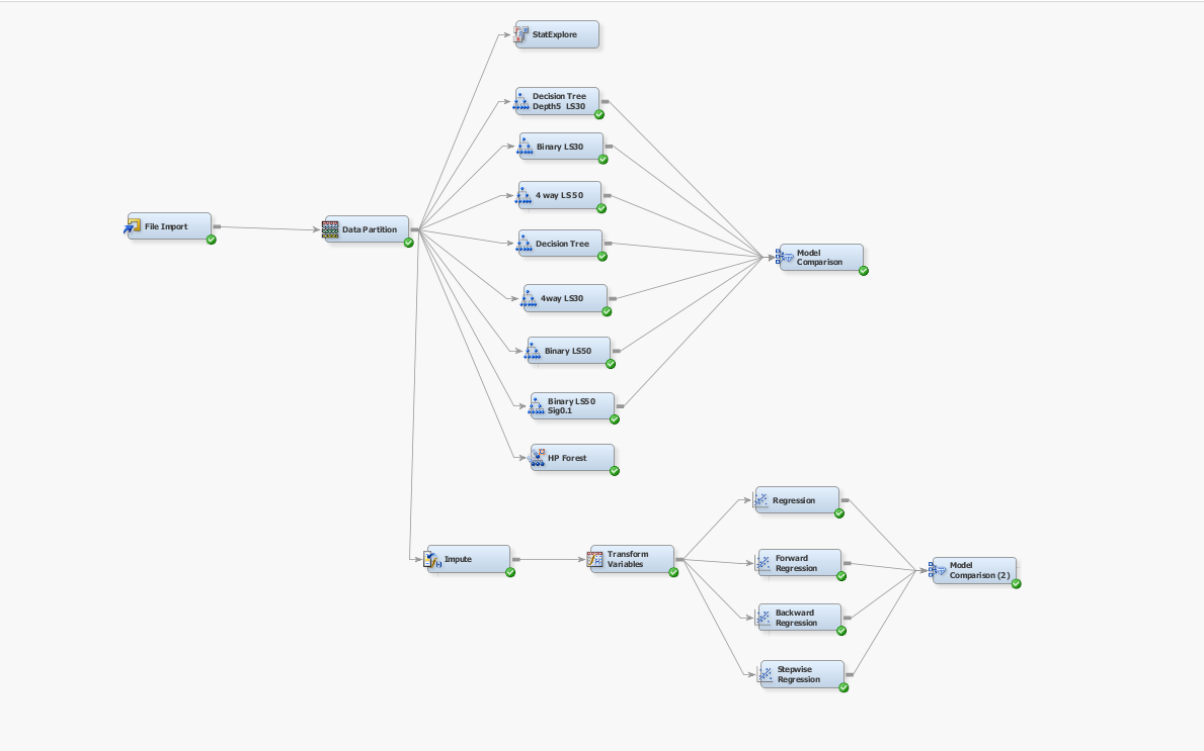
Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
__of_House__holds	Input	Interval	No		No	.	.
__Below_Poverty_Line	Input	Interval	No		No	.	.
_1__to_4__Family_Units	Input	Interval	No		No	.	.
_2015_Tract_Median_Household_Inc	Rejected	Nominal	No		No	.	.
_2021_Est__Tract_Median_Family_I	Rejected	Nominal	No		No	.	.
American_Indian_Pop__ulation	Input	Interval	No		No	.	.
Asian__Hawaiian__Pacific_Islande	Input	Interval	No		No	.	.
Black_Pop__ulation	Input	Interval	No		No	.	.
Hispanic_Population	Input	Interval	No		No	.	.
Incident_day_of_week	Input	Nominal	No		No	.	.
Latitude	Input	Interval	No		No	.	.
Longitude	Input	Interval	No		No	.	.
Median_House_Age__Years_	Input	Interval	No		No	.	.
Non_Hisp_White_Population	Input	Interval	No		No	.	.
Number_of_Families	Input	Interval	No		No	.	.
Other_Population__Two_or_More_Ra	Input	Interval	No		No	.	.
Owner_Occupied_1__to_4__Family_U	Input	Interval	No		No	.	.
Owner_Occupied_Units	Input	Interval	No		No	.	.
Renter_Occupied_Units	Input	Interval	No		No	.	.
Response_time	Target	Interval	No		No	.	.
Total_Housing_Units	Input	Interval	No		No	.	.
Tract_Income_Level	Input	Nominal	No		No	.	.
Tract_Minority__	Input	Interval	No		No	.	.
Tract_Minority_Population	Input	Interval	No		No	.	.
Tract_Population	Input	Interval	No		No	.	.
Vacant_Units	Input	Interval	No		No	.	.
census_block_2010	Input	Nominal	No		No	.	.
census_block_group_2010	Input	Interval	No		No	.	.
census_tract_2010	Input	Nominal	No		No	.	.
council_district	Input	Nominal	No		No	.	.
created_at	Rejected	Interval	No		No	.	.
hour_of_day	Input	Nominal	No		No	.	.
incident_datetime	Rejected	Interval	No		No	.	.
incident_type_primary	Input	Nominal	No		No	.	.
neighborhood	Input	Nominal	No		No	.	.
parent_incident_type	Input	Nominal	No		No	.	.
police_district	Input	Nominal	No		No	.	.
zip	Input	Nominal	No		No	.	.

Data Partition:

We have made the split as 70% of data for training and 30 % of the data for Validation purposes.

4.1. Decision Trees

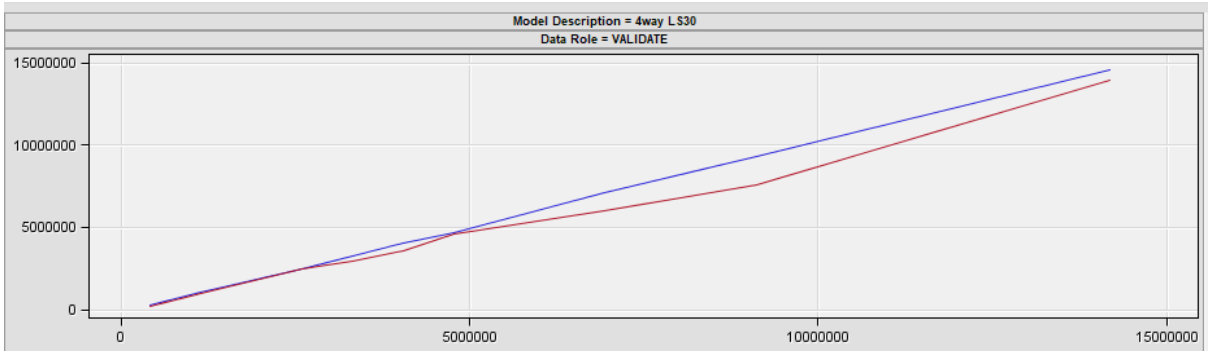


According to the interval target variable, we have made the decision tree properties. We have created multiple decision trees with different pre-pruning techniques like maximum branch, maximum depth, leaf size. We have set all the decision trees to follow Use in search condition for missing values and ProbF is set as splitting rule for Interval target criterion.

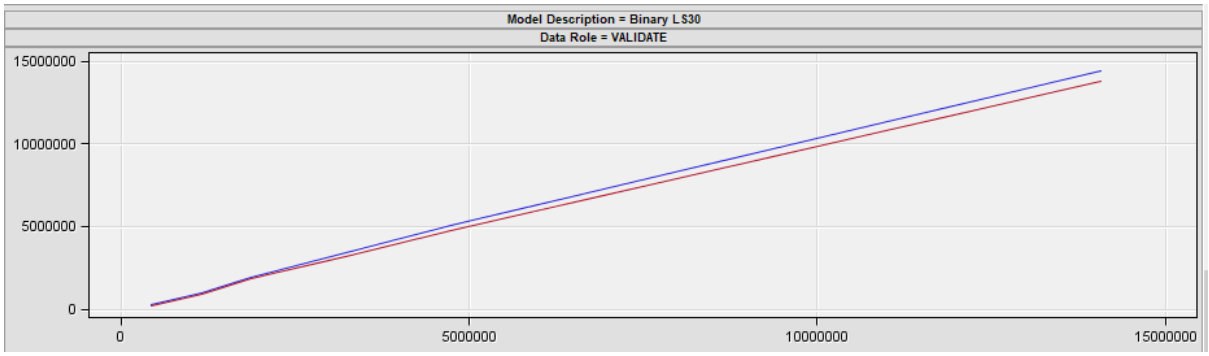
Models	**Maximum Branch**	**Maximum Depth**	**Leaf Size**	**Significance level**
Decision Tree	2	6	5	0.2
Binary LS_30	2	6	30	0.2
Binary LS_30_5	2	5	30	0.2
Binary LS_50	2	6	50	0.2
Binary LS_50 0.1	2	6	50	0.1

4-way LS_30	4	6	30	0.2
4-way LS_50	4	6	50	0.2

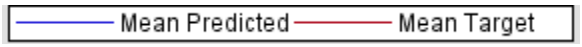
4-way LS 30 is the best model suggested by SAS based on Average Squared Error score. After checking Score Distribution based on Response Time (Target Variable), we concluded that Binary LS30 fits the validation dataset better than 4-way LS30 with average squared error score difference of 0.005.

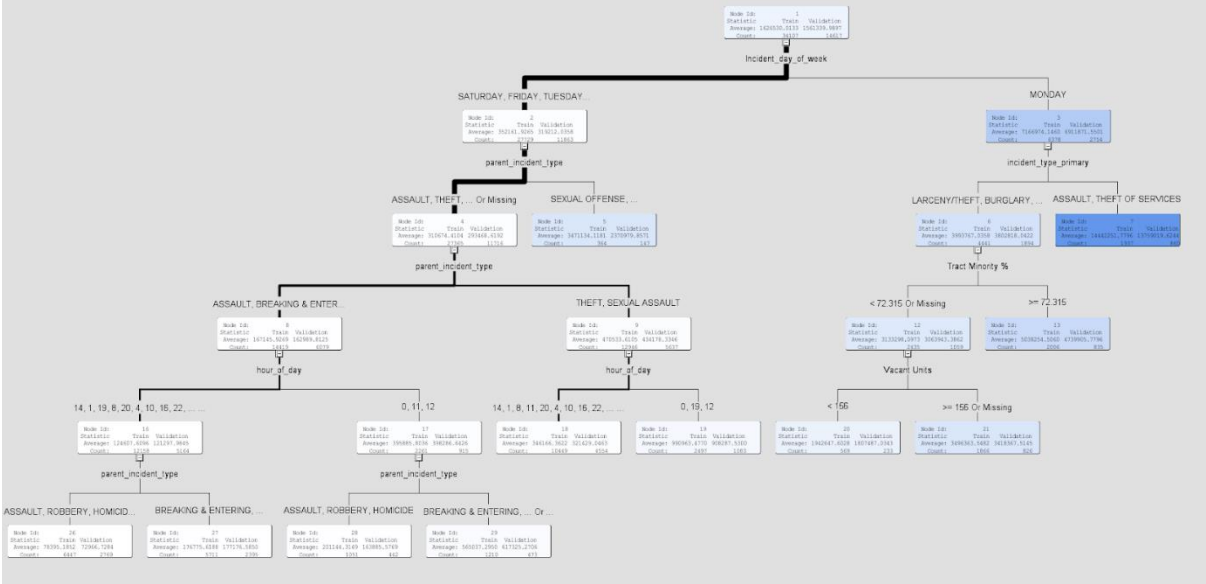


Validation score distribution 4Way LS30



Validation score distribution Binary LS30





Binary LS30 Decision tree

Important Inferences from Binary LS30 Decision Tree:

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Incident_day_of_week		1	1.0000	1.0000	1.0000
incident_type_primary		1	0.7819	0.7754	0.9917
parent_incident_type		4	0.1339	0.0866	0.6468
Tract_Minority__	Tract Minority %	1	0.1287	0.1158	0.8995
Vacant_Units	Vacant Units	1	0.0661	0.0694	1.0499
hour_of_day		3	0.0649	0.0616	0.9489

Important takeaways from decision tree analysis:

1. Incident day of the week is the most important variable with Monday (close to ~18-19% of incidents) showing different characteristics than the rest of the days and hence the tree splits the way it does.
2. Under parent incident type, Sexual Offenses is the highlighted category that exhibits different behavior than the rest of the incident types for the rest of the days. 'Assault and Theft of Services' is an important category in incident type primary having a large count of around 2797 making it the largest and common crime type on Mondays**.**

4.2. Random Forest

To further assess our features and understandings from decision tree models, we used the default HP Random Forest model to confirm the variable importance for the target variable. The following features were important in this case:

Variable Name	Number of Splitting Rules	Train: Mean Square Error
Incident day of week	463	4.61E12
parent incident type	440	9.305E11
hour of day	413	1.203E11
incident type primary	367	9.634E11
census block 2010	142	2.039E10
neighborhood	141	5.707E10
census tract 2010	118	2.024E10
council district	114	1.048E10

Important feature from HP Random Forest

We can infer that Incident-day of week, parent incident type, hour of day, incident type primarily provides information gain for the target variable.

4.3. Regression Models

Using SAS

As our project has the Continuous target variable, we made the model selection to be a linear regression. We have imputed the dataset as it has got some of the missing values on Census block group 2010 and Census Block 2010, we have used Tree imputation method as it is the best on the imputation among others as it imputes values based on the values from the dataset. The Imputed Dataset is given as input to different regression nodes that we have built which differs slightly based on the model selection and selection criterion

1. **Regression - Criteria set as 'default:**

   As it is a default regression node there is not much to do regarding the properties of regression. As a default regression model, it considers all the variables for its prediction

2. **Backward Regression**

   We have set the model Validation error. The backward regression ran through 20 steps and has rejected features that don't help with prediction. We finally ended up with 8 important features selection as Backwards and the selection criteria to be

Step 20: Effect Tract_Income_Level removed.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	352	2.931553E17	8.3282756E14	19.91	<.0001
Error	33754	1.4120038E18	4.1832191E13		
Corrected Total	34106	1.7051591E18			

The DMREG Procedure

Model Fit Statistics

R-Square	0.1719	Adj R-Sq	0.1633
AIC	1070106.5496	BIC	1070115.6693
SBC	1073084.9016	C(p)	365.4699

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
IMP_census_block_2010	265	1.3601E16	1.23	0.0071
Incident_day_of_week	6	2.23794E17	891.63	<.0001
Tract_Minority__	1	1.40958E14	3.37	0.0664
council_district	9	8.23319E14	2.19	0.0200
hour_of_day	23	2.85436E15	2.97	<.0001
incident_type_primary	11	2.68093E16	58.26	<.0001
neighborhood	35	2.30583E15	1.57	0.0166
parent_incident_type	2	7.52221E14	8.99	0.0001

Backward Regression Output

3. Forward Regression

We have set the model selection as Forward and the selection criteria to be Validation error. The regression selected the model, based on the error rate for the validation data, it ran through 4 steps and gave us the important features with respect to specific categorical value as they have been converted to dummy variables.

The selected model, based on the error rate for the validation data, is the model trained in Step 4. It consists of the following effects:

Intercept Incident_day_of_week Tract_Minority__ hour_of_day incident_type_primary

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	41	2.7601642E17	6.7321077E15	160.47	<.0001
Error	34065	1.4291427E18	4.1953403E13		
Corrected Total	34106	1.7051591E18			

Model Fit Statistics

R-Square	0.1619	Adj R-Sq	0.1609
AIC	1069896.0476	BIC	1069897.8772
SBC	1070250.4124	C(p)	153.3269

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Incident_day_of_week	6	2.27347E17	903.17	<.0001
Tract_Minority__	1	6.08642E14	14.51	0.0001
hour_of_day	23	3.00413E15	3.11	<.0001
incident_type_primary	11	2.98646E16	64.71	<.0001

Forward Regression Output

4. Stepwise Regression:

We have set the model selection as Stepwise and the selection criteria to be Validation error. The regression selected the model, based on the error rate for the validation data, it ran through 4 steps and gave us the important features It consists of the following effects: Intercept, Incident_day_of_week, Tract_Minority, hour_of_day, incident_type_primary

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	41	2.7601642E17	6.7321077E15	160.47	<.0001
Error	34065	1.4291427E18	4.1953403E13		
Corrected Total	34106	1.7051591E18			

Model Fit Statistics

R-Square	0.1619	Adj R-Sq	0.1609
AIC	1069896.0476	BIC	1069897.8772
SBC	1070250.4124	C(p)	153.3269

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Incident_day_of_week	6	2.27347E17	903.17	<.0001
Tract_Minority__	1	6.08642E14	14.51	0.0001
hour_of_day	23	3.00413E15	3.11	<.0001
incident_type_primary	11	2.98646E16	64.71	<.0001

Stepwise Regression Output

Using Python

Jupyter Notebook link:

https://github.com/nishitchaudhry/Buffalo_Crime_Response_Time/blob/main/Buffalo_Crime_Response_Time.ipynb

After the data pre-processing using Train-test split (70:30), categorical dummy variable creation and scaling numerical data using Minmax, we created models using Ordinary Least Squares algorithm from stats model library with Linear Regression from Sci-kit Learn in Python. The following different strategies for model building were used:

1. Using all features
2. Top 15 features using Recursive Feature Selection

3. Backward Regression on top 15 features using P-value (variable significance) and Variance Inflation Factor(multicollinearity)

The importance of features according to the final model after removing multicollinearity and including all the features with high statistical significance are:

	Features	VIF
0	parent_incident_type_Other Sexual Offense	1.00
1	incident_type_primary_THEFT OF SERVICES	0.26
2	neighborhood_Upper West Side	0.18
3	incident_type_primary_RAPE	0.07
4	Incident_day_of_week_Monday	0.04
5	incident_type_primary_ASSAULT	0.03
6	Incident_month_January	0.00

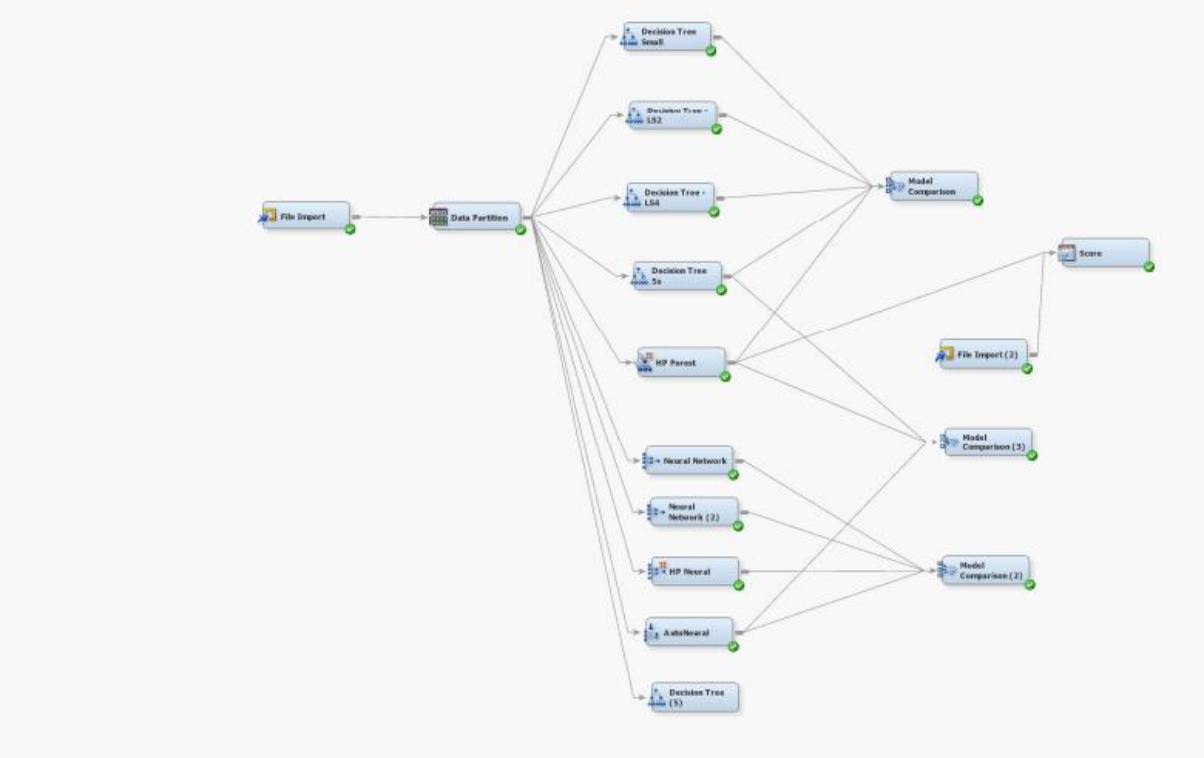
Important features from Python OLS

We can infer from this model results that it is similar to our earlier decision tree and regression models. Hence, confirming the important features for our target variable.

4.4. Classification model

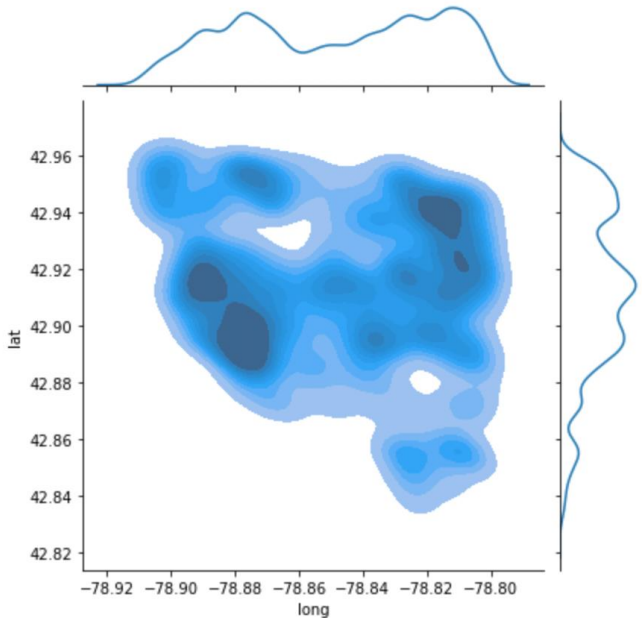
As an additional part of our project, we looked into classifying and predicting the type of incident that would occur based on the input variables such as time and neighbourhood. However, due to the fact that violent crimes like murder and sexual assault accounted for less than 5% this created an issue for the prediction model as it was misclassifying some events as Larceny/Theft and Assault which together make up over 60% of the data. (Figure 1.1) To create a predictive classification model we used multiple decision trees of varying leaf sizes and branches to get a variation of different results from the trees. Upon model comparison of the built trees and a random forest model we determined that the random forest model. We also ran the data through varying neural network nodes with different settings selected. These neural networks were then compared

to find the best fitting with our AutoNeural node performing the best. Upon comparison of the best models from the neural networks and decisions tree, again the random forest model was selected as the best fitting. This was then scored against a modified scoring dataset to get our predicted result which is where problems arose due to the issue mentioned above. The predictions returned by the model misclassified incidents that we know are homicide and murder as Larceny/Theft and Assault. However, the model does accurately perform a descriptive analysis telling us the likelihood of which type of incident can occur based on the input variable of time, day and location.



5.5. Clustering model - Neighbourhood Planning for Emergencies

Clustering is an unsupervised machine learning method typically used to group objects of similar attributes together. In our dataset of police incidents, we have location-wise longitude and latitude values for every emergency call recorded. Based on this information, we built Kernel Density Estimate (KDE) plots for overall incidents

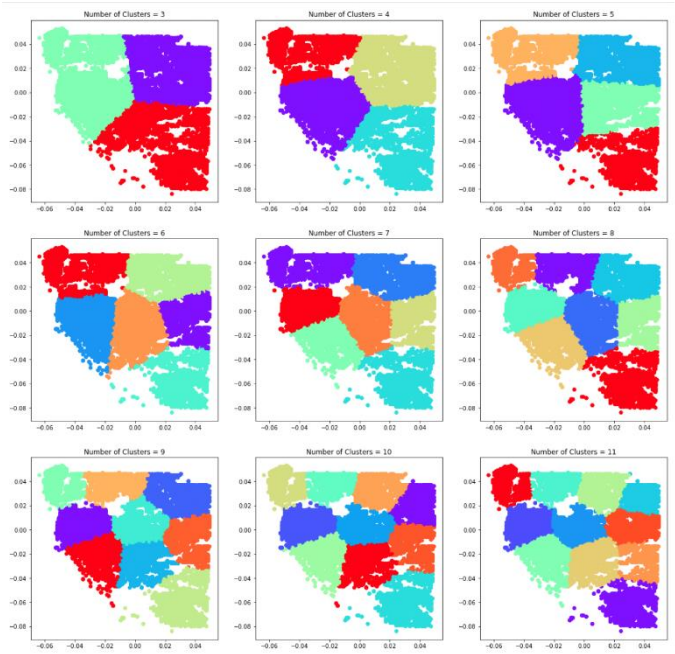


Density Distribution for all incidents

Similarly, the incident distribution for some major primary incident types are as follows

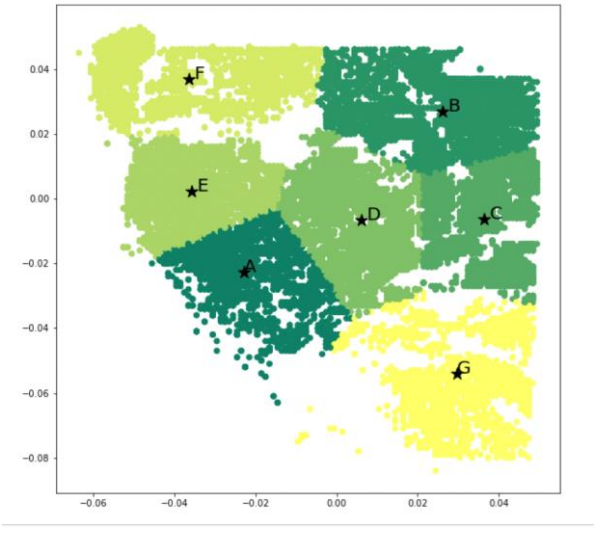
From the above KDE plots, we observe certain areas that are highly prone to incidents. We would expect the police enforcement to be especially vigilant and sufficient in such areas. Currently, Buffalo city planning is based on the five Police Districts (here A, B, C, D, E). For this we can refer to our exploratory data analysis where we have the current geo-planning of Buffalo law enforcement. With the dataset available with us which has records dated from 2018, we wish to establish whether Buffalo city planning based on the five Police Districts (here A, B, C, D, E) is adequate for serving the Buffalo city population. We do this by performing a cluster analysis based on the number of incidents. Here, we have used the **k-means clustering model** to achieve our purpose.

On applying k-means clustering, we tried to figure out the ideal number of clusters required for better geo-spatial city planning. We ended up with the following model output based on number of incidents -



K-means clustering based on map coordinates and incident density

Since in the original police districts E and B (refer Fig 5.6.1 and Fig 5.6.2) , we know that a more granular split of districts is definitely required here. This is deemed possible when the number of clusters are greater than or equal to 7. Here, the lighter colour shades indicate a low population density and incidents churn overall. We take the final number of police districts based on population density as 7.



Proposed City Planning for Emergency Incident Handling

6. Model Evaluation

The following are the regression model evaluation scores for the continuous target variable (Response Time):

Models	Mean Squared Error	Absolute Mean Error	R-Squared	Adjusted R-Squared
Binary LS_30	3.572E13	6.09E7	-	-
Binary LS_50	3.572E13	6.09E7	-	-
4-way LS_30	3.567E13	6.11E7	-	-
HP Forest	4.654E13	-	-	-
Forward Regression	4.195E13	-	0.1619	0.1609
Backward Regression	4.183E13	-	0.1719	0.1633
Stepwise Regression	4.195E13	-	0.1619	0.1609
Python OLS	3.9E13	2.7E6	0.167	0.167

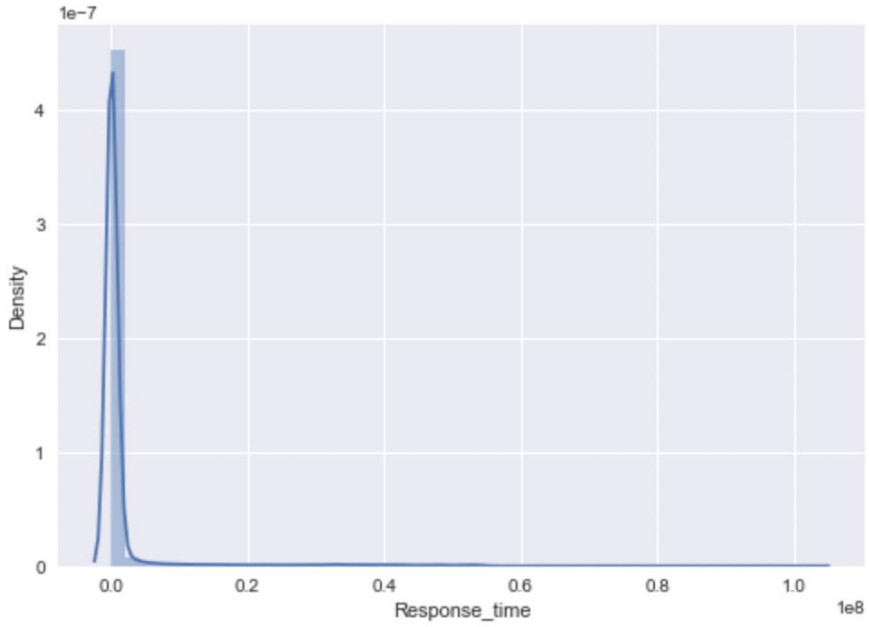
6..Challenges/bottlenecks

1. Response Time Limitation:

- From our assumptions and understandings, we have derived Response time from the difference of 'incident datetime' and 'Created_at' column
- The target variable (response time) is highly-right skewed, making our model prediction as seen below from the figure with a range of 60-38,000,000 seconds.
- Even after omitting the outliers and considering 95% target response time with a range of 60-6,000,000 seconds (python model), R-square remains low
- 0-5,000 seconds consists of 8535 rows and 20k-35k seconds consist of 7962 rows. The segmented data points in our target variable like above might have hampered our model predictions

2. Limitation of our dataset (Endogeneity):

There are some underlying hidden features which might bring more information and help our models to fit the data points well for higher R-squared score.



7. Citations:

- https://data.buffalony.gov/Public-Safety/Crime-Incidents/d6g9-xbgu
- https://www.ffiec.gov/census/report.aspx?year=2021&state=36&msa=15380&county=&tract=&report=housing&page=3

8. Appendix:

- Python Regression Jupyter Notebook
  https://github.com/nishitchaudhry/Buffalo_Crime_Response_Time/blob/main/Buffalo_Crime_Response_Time.ipynb
- Clustering Jupyter Notebook
  https://github.com/r10na/Emergency-Geo-Planning-for-Buffalo/blob/main/Cluster%20Analysis%20-%20Buffalo%20Crime%20Incidents%20Emergency%20Planning.ipynb