Crime Data Analysis and Visualization of Kansas City

Akhil Teja Kanugolu
Computer Science
UMKC
KCMO, United States
ak9bd@umsystem.edu

Sai Rohith Guntupally
Computer Science
UMKC
KCMO, United States
sgkxv@umsystem.edu

ABSTRACT

The primary focus to analyze crime Data in Kansas City because in current society there is a lot of incident's taking around, which makes the people fear of crime. There are numerous crime cases registering in Police Department per day in world, particularly focusing on Kansas City the rate of crime is bit high. To analyze the crime, data collected for years 2018 & 2019 from KCPD which was released for public interest to understand. Jupyter & Google colab are used for Machine learning using python and Microsoft Power Bi for Analysis. Attributes are selected based on the correlation between the features. Before selecting the features, they are preprocessed for removing the nulls to improve the efficiency of analysis. Later, we will visualize the Data using both technologies. After the preprocessing we will train the models using Randomforest classifier, KNN Classifier, Decision Tree Classifier and Logistic Regression try to predict. And in power Bi we will analyze the relation between the features. Finally host the visualization on S3 Bucket. By this model helps public to take the necessary precautions for their safety and for Police department helps to analysis the situation to take handy and helps to make the Kansas City as best law enforced city.

Crédible Source/ Data

Collecting data and source:

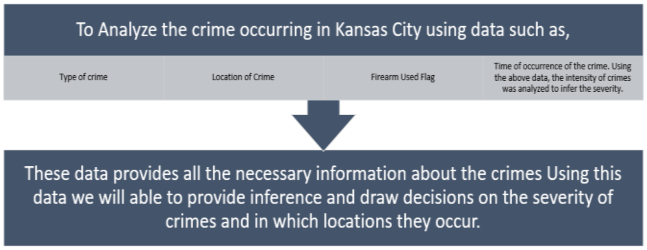
Data is collected from the following website – KCPD

(https://data.kcmo.org).

Record: 221k Rows of data with 23 features Currently we are analyzing data for crimes in Kansas City for year 2019 and 2018. This can be extended to the comparison for previous years (2014-2019).

1 Introduction

The criminal of any case can be analyzed based on the information from where the crime scene is happened which is then tested against previously done crime patterns which can be judged by the method, which was implied to test, and the required measures would be taken against it. This prediction can further be made useful in detecting these crimes in very advance and by deploying like more cops as well as patrols to these sensitive areas which can be identified by this system from Flow 1.



To Analyze the crime occurring in Kansas City using data such as,

Type of crime

Location of Crime

Firearm Used Flag

Time of occurrence of the crime. Using the above data, the intensity of crimes was analyzed to infer the severity.

These data provides all the necessary information about the crimes Using this data we will able to provide inference and draw decisions on the severity of crimes and in which locations they occur.

Flow 1: Analysis of Crime

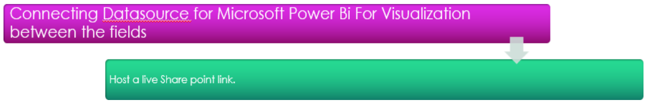
Crime Data Analysis and visualization for Kansas City crime data through Machine Learning techniques using the KCPD dataset. Will try to concade two years (2019 -20) data based on the features. The process as follows: We first collect the data and host the Data in S3 Bucket in AWS. Later, Using Jupyter Notebook/Google Colab we try to load the data and clean the data from Nulls/ Unused Features for prediction and finally, visualize the Data. Based on the Data processed we will understand the various factors like age, crime, sex which have an impact on the crime rate by visualization. Also, will perform various machine learning models like Random forest, Logistic Classifier, SVM, KNN on the given data for helping the prediction. Finally, Hosted the Visualized Data on the S3 Bucket. And using the same data set we will also Load in Microsoft Power Bi and Clean and process the data. Later, we'll un understand the relations between the various features by visualizing the data using MS Power Bi. (Plots Heat Map based on Location).

2 Methodology

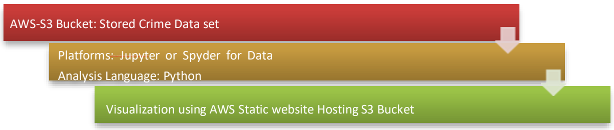
Used google colab or jupyter and power bi for visualizing the data. Trained and tested the models using machine learning. Can understand the implementation environments for Power bi and Machine learning from Flow 2 &3.

2.1 Power Bi

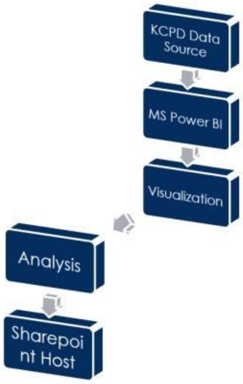
Data set from KCPD was taken and imported into Power Bi environment which loads as table data. Once the data loaded then it will be tuned according to the requirement. After that created different visual for the analyzed with different features of crime data. Finally, after the analysis of data on power bi hosted the report on to service using SharePoint. Flow 4 helps to understand power bi chart.


Connecting Datasource for Microsoft Power Bi For Visualization between the fields
Host a live Share point link.

Flow 2: Power Bi Implementation Environment


AWS-S3 Bucket: Stored Crime Data set
Platforms: Jupyter or Spyder for Data
Analysis Language: Python
Visualization using AWS Static website Hosting S3 Bucket

Flow 3: Machine learning Implementation Environment



Flow 4: Analysis Flow Using Power BI

2.1.1 Analysis using Power BI

By observing the Figure 1, white and black are the top2 races which are actively involved the crime compared to the rest of the races.

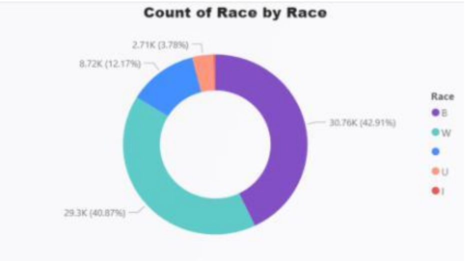


Figure 1: Count of crime by Race

Figure 2 gives the crime count based on the zone in the Kansas City. According to the figure EPD, CPD & MPD zone have more threat of crime

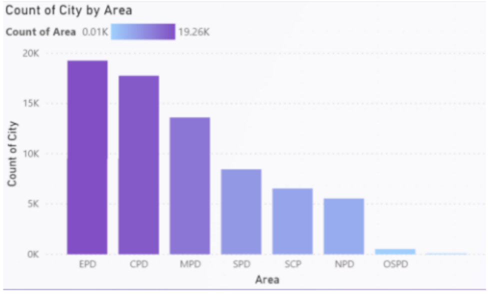


Figure 2: Crime by Area

From Figure 3, One can draw the knowledge that Male Gender are highly involved in the crime.

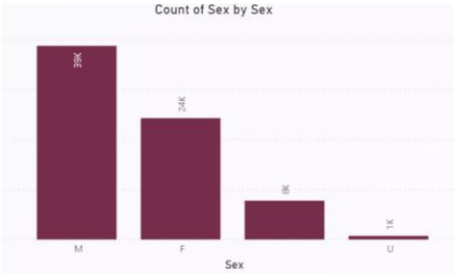


Figure 3: Density of crime by Sex

From the heat map of crimes, can conclude highest rate of crime in Kansas City are more in central part between the years 2018 and 2019. One need to be careful when they want to relocate to central part of Kansas City based on highest rate of crime incidents.

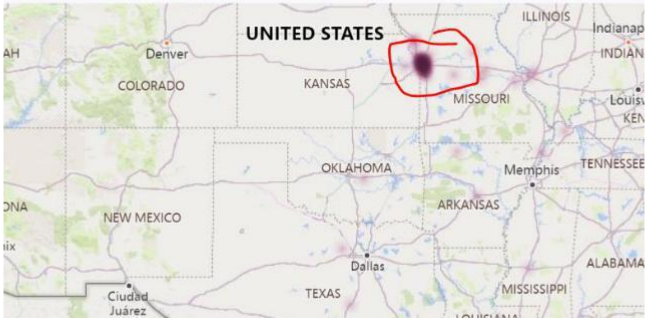


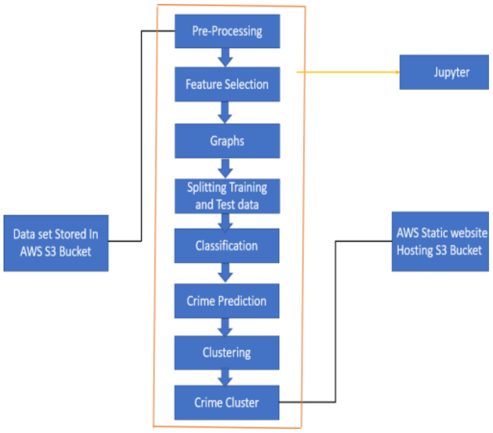
Figure 4: Heat Map at Kansas City

2.2 Machine Learning

Similarly, with power bi initially the data was loaded as data frame. After loaded 2018 and 2019 data are merged up together for improving the efficiency of the Data. Merging the together led to mismatch of columns which got dropped from the dataset.

Handled the features/ Attributes based on the data type of the features. For numeric feature, the null values are replaced by the mean of the features. And String data type handled by mode operation. Once preprocessing was done to handle the description, we used stemmer operation from nltk library. Next standard scaling applied on the features to maintain the features range at same point using Euclidean distance. Later based on the correlation between the features we selected the features which have major impact on the crime.

After feature selection we visualized different graphs using pandas and seaborn libraries. After the visualization on Google colab, split the data into training and test data to create different machine learning classification models and clustering. Finally hosted the model and visualization on S3 bucket. Follow Flow 5 for machine learning process.


Pre-Processing
Feature Selection
Jupyter
Graphs
Splitting Training and Test data
Data set Stored In AWS S3 Bucket
Classification
AWS Static website Hosting S3 Bucket
Crime Prediction
Clustering
Crime Cluster

Flow 5: Analysis Flow Using Power BI

2.2.1 Machine Learning Visualization

Figure5 gives the comparison of crime rate between 2019 and 2018. From the figure can observe crime rate of 2018 was lot more than 2019.

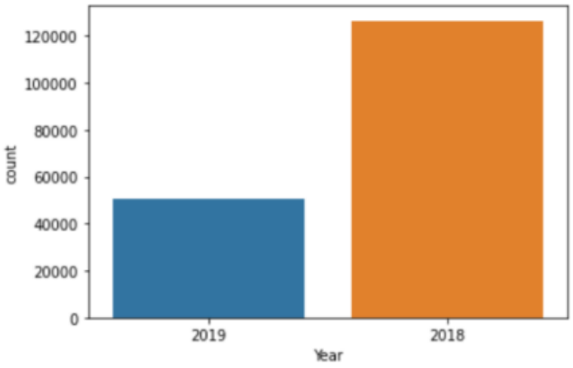


Figure 5: 2018 Vs 2019 Crime Data

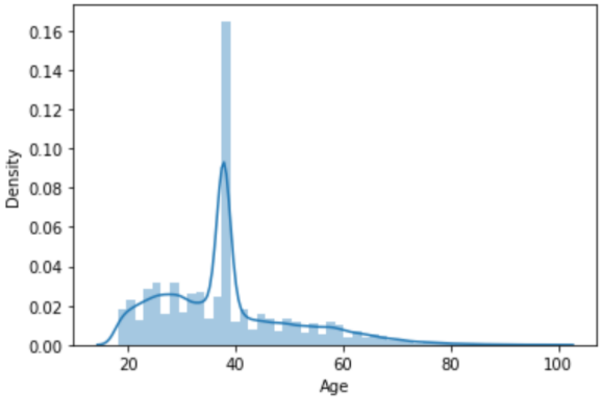


Figure 6: Age Frequency Plot

Figure 6 shows people with age range 20-80 have mostly Influenced of crime. The peak is noted at age 38.

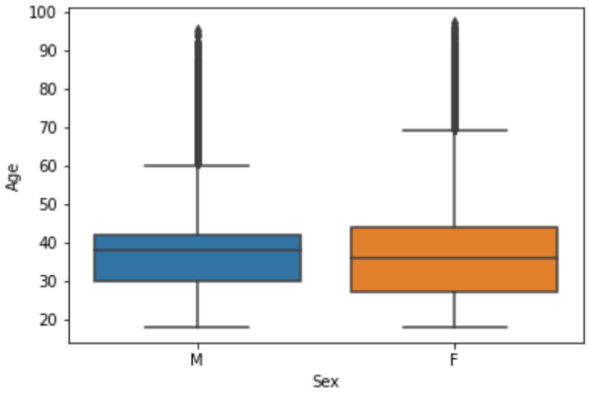


Figure 7: Observation between Age and Sex

Figure 7 can observe that Male at the age range 30-43 committed more crime and median age is recorded at 38. Female with age range 28-45 committed more range and median age is recorded at 36. Also observed that Female age range is greater than male age range.

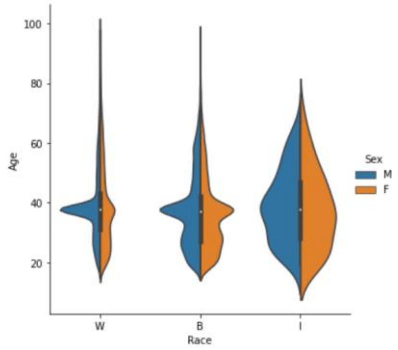


Figure 8: Catplot Observation between Age and Race

Observation from Figure 8 shows that Whites, blacks and American Indian or Alaska native have equal percent in Male and Female category. All have the highest peak point at age 38.

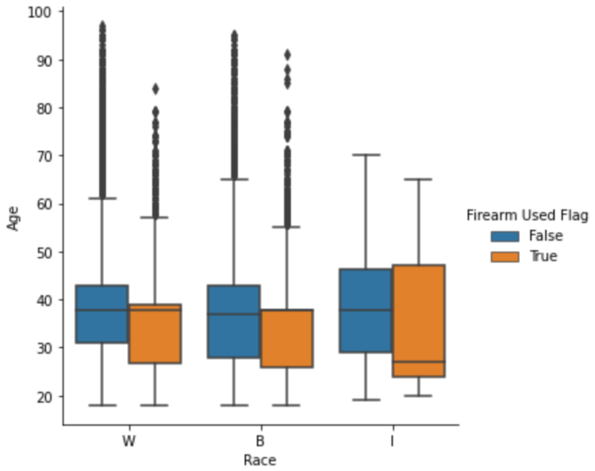


Figure 9: Catplot for the Firearm used by Race

In Figure9, Whites category Male and Female has equal percentage of using firearms, range 32-43 in male and 26-38 in female category but according to graph the median of age used firearm is 38 in male and 38 in female. In Blacks category Male used more firearm than female, range 28-45 in male and 25-38 in female category but according to graph the median of age used firearm is 37 in male and 39 in female. In American Indian or Alaska naive category Female used more firearm than male, range 29-47 in male and 26-48 in female category but according to graph the median of age is 38 in male and 28 in female.

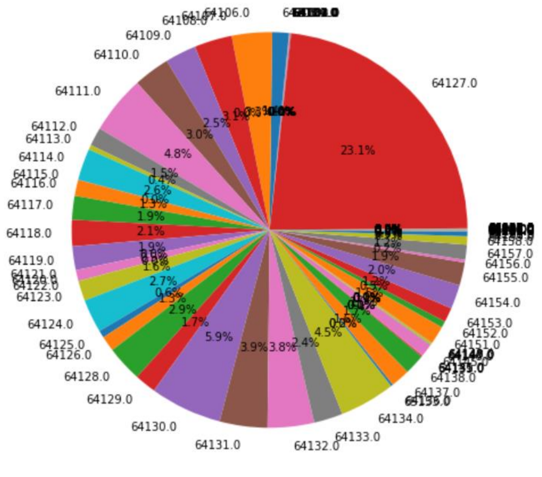


Figure 10: Crime rate according to Zip code

According to the graph the highest recorded crime is in area with zip code '64127' and the percentage is 23.1%.

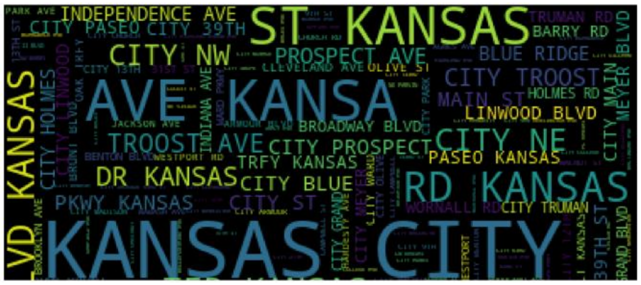


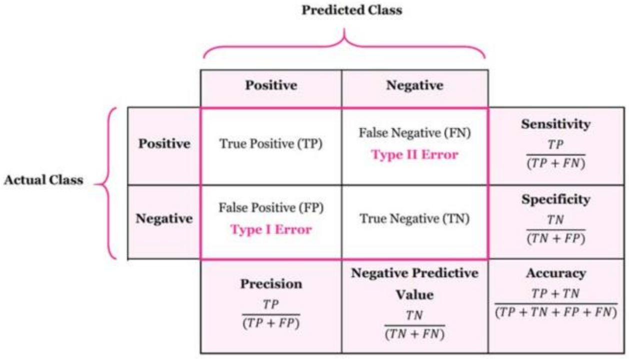
Figure 11: Locations of the Crime

The areas which are considered as red zone i.e., crime reported areas are Independence ave, Prospect ave, Paseo Kansas, Westport, Troost, Main ST, Amour, Meyer Blvd.

2.2.2 Machine Learning Models

Before going into models first the concepts of conclusion matrix, F1 score, RMSE required to select the best performing model among the classifier models.

Confusion Matrix is a tool to decide the performance of classifier. It consists of statistics approximately real and anticipated classifications. If the model want best performance count of Type I error should be less. Figure 12, Sensitivity, precision and Accuracy of the models are more important.



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 12: Confusion Matrix

F1 Score helps to seek balance among the precision and recall of the classifier model.

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

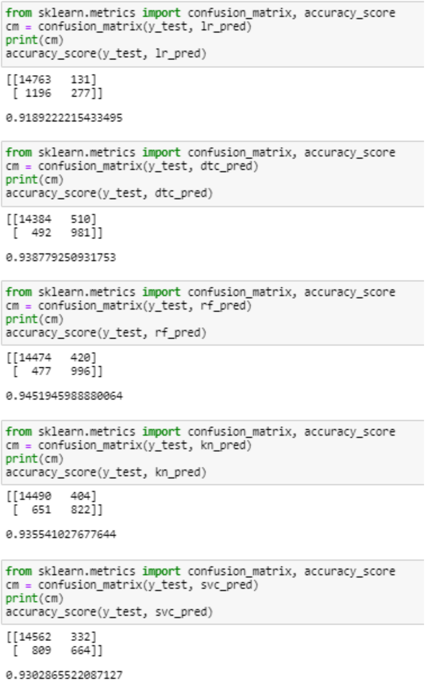
Figure 13: F1 Score

RMSE estimates the value difference between the prediction of the classifier model and the original train (actual) value.

It ranges from -1 to 1 and preferably should be near to 0 will be considered as better performing model. However, low value will be sufficient.

3 Model Construction

After preprocessing the data and feature selection based on the correlation. Consider Inputs: IBRS, Involvement, Race, Gender, Age, Description, Dvflag, Month and later the data got splited into test and train data, then the train data feed the different Classifier models Logistic Regression, Decision Tree, Random forest, K Neighbors and Support vector machine (In order) and tried to predict the Firearm used or not at crime scene. Parallelly also computed the classification report, confusion matrix and RMSE to get the best working model. Based on those validations we also try to predict and compare the models for test data.

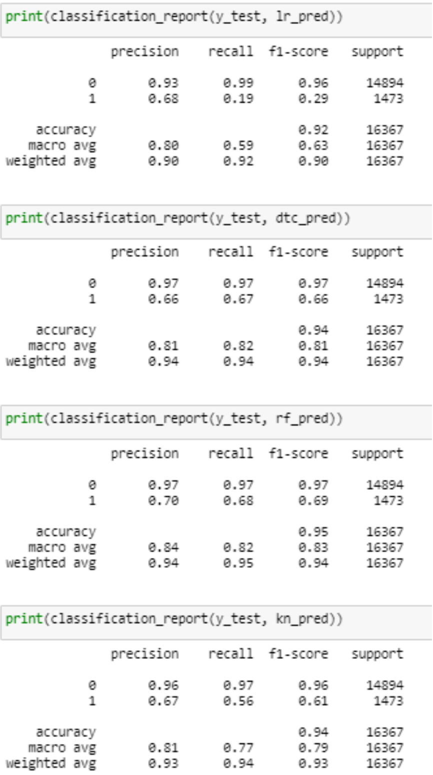


```
from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, lr_pred)
print(cm)
accuracy_score(y_test, lr_pred)

[[14763   131]
 [ 1196   277]]

0.9189222215433495

from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, dtc_pred)
print(cm)
accuracy_score(y_test, dtc_pred)

[[14384   510]
 [  492   981]]

0.938779250931753

from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, rf_pred)
print(cm)
accuracy_score(y_test, rf_pred)

[[14474   420]
 [  477   996]]

0.9451945988880064

from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, kn_pred)
print(cm)
accuracy_score(y_test, kn_pred)

[[14490   404]
 [  651   822]]

0.935541027677644

from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, svc_pred)
print(cm)
accuracy_score(y_test, svc_pred)

[[14562   332]
 [  809   664]]

0.9302865522087127
```

Result 1: Confusion Matrixes for 5 models

From the Result 1, By comparing the Type I and Type II errors among the 5 models, can select Random Forest and Decision Tree because those are having less Type I and Type II errors which helps to increase the sensitivity and the precision of the model. To be more specific Random forest model is combination of several Decision trees. So can say Random forest is effective than Decision tree based on precision and sensitivity.

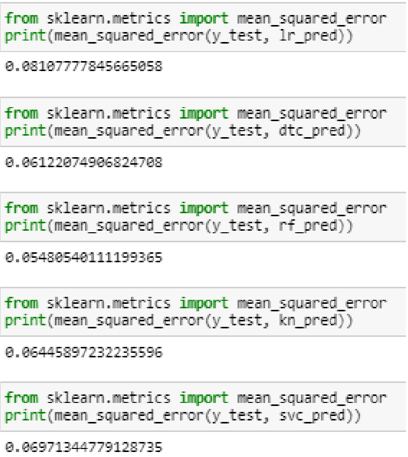
Next biasing Result2, Considering the classification report Random forest is having 95% accuracy of predicting the values. And splitting down based on using firearm accuracy is 69% and without firearm used 97%. From here we can see under sampling takes place on predicting with firearm. Although under sampling takes place the accuracy rate is more for Random forest compared to other classification models.



```
print(classification_report(y_test, lr_pred))

              precision    recall  f1-score   support

           0       0.93      0.99      0.96     14894
           1       0.68      0.19      0.29      1473

    accuracy                           0.92     16367
   macro avg       0.80      0.59      0.63     16367
weighted avg       0.90      0.92      0.90     16367

print(classification_report(y_test, dtc_pred))

              precision    recall  f1-score   support

           0       0.97      0.97      0.97     14894
           1       0.66      0.67      0.66      1473

    accuracy                           0.94     16367
   macro avg       0.81      0.82      0.81     16367
weighted avg       0.94      0.94      0.94     16367

print(classification_report(y_test, rf_pred))

              precision    recall  f1-score   support

           0       0.97      0.97      0.97     14894
           1       0.70      0.68      0.69      1473

    accuracy                           0.95     16367
   macro avg       0.84      0.82      0.83     16367
weighted avg       0.94      0.95      0.94     16367

print(classification_report(y_test, kn_pred))

              precision    recall  f1-score   support

           0       0.96      0.97      0.96     14894
           1       0.67      0.56      0.61      1473

    accuracy                           0.94     16367
   macro avg       0.81      0.77      0.79     16367
weighted avg       0.93      0.94      0.93     16367
```

Result 2: Classification Report for 5 models

Finally based on Result3, RMSE we have Random forest with 0.055 which is ~ 0 (near to zero), which would be preferable when RMSE value near to zero. So, by comparing Confusion matrix, Classification Report and RMSE we found the better precision, sensitivity, accuracy, F1 score and RMSE for Random forest model for predicting about the firearm. Summarizing Random forest was best suited for current scenario. Can use this model by cops at crime incident helps to taking relative measures, if cops know the criminal have firearm at the scene.



```
from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test, lr_pred))

0.08107777845665058

from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test, dtc_pred))

0.06122074906824708

from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test, rf_pred))

0.05480540111199365

from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test, kn_pred))

0.06445897232235596

from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test, svc_pred))

0.06971344779128735
```

Result 3: RMSE for 5 models

3.1 Clustering

K Means clustering works with process of vector quantization. Clustering is done for numeric features as it need to calculate the centroid for different clusters based on Euclidean distance. Number of clusters are calculated based on elbow method.

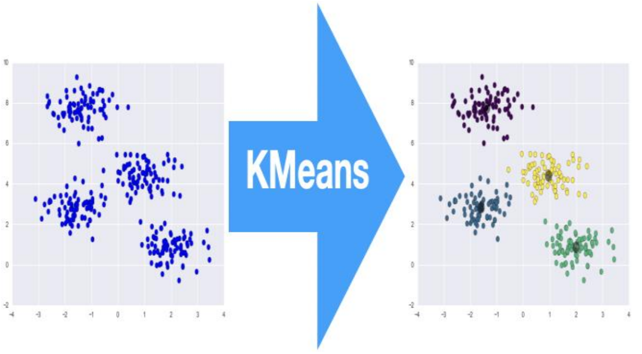
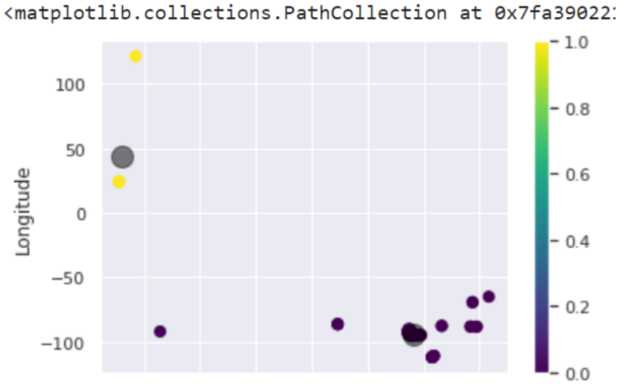


Figure 14: K Means clustering

As clustering need to do for numeric data, as we do not have more numeric features on crime data. So, tried to cluster by latitude and longitude



```
<matplotlib.collections.PathCollection at 0x7fa39022
```

Result 4: Clustering Based on Longitude and Latitude

By the observing Result 4, cluster 1 have highest crime rate at longitude 0 to – 100 and latitude 90-100. So, necessary precautions need to be held in the area to handle the crime rate situation

Even though clustered for longitude and latitude desired results are not achieved. So, moved on to K prototype. In K prototype categorical data can be processed. Here it clusters numeric and categorical data into clusters by using mode operation instead of using Euclidean distance.

From Result 5, Cluster 0 by K prototype was impacted by

age-25,

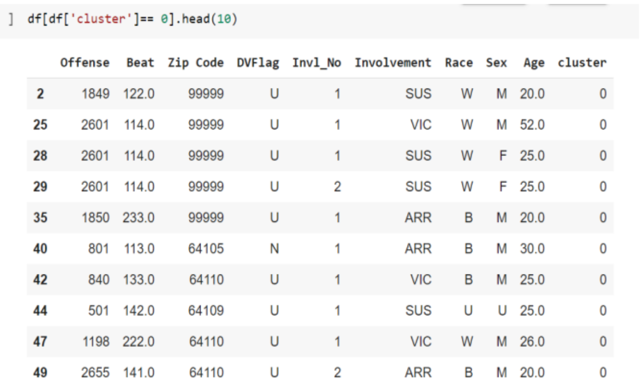
zipcode-64110

beat 120-140 and involvement 1,2 and

Race – white and

offence – 1800-2600

This cluster having a greater number of crimes means these features are having more impacting on crime which should be looked after.



```
] df[df['cluster']== 0].head(10)
```

	Offense	Beat	Zip Code	DVFlag	Invl_No	Involvement	Race	Sex	Age	cluster
2	1849	122.0	99999	U	1	SUS	W	M	20.0	0
25	2601	114.0	99999	U	1	VIC	W	M	52.0	0
28	2601	114.0	99999	U	1	SUS	W	F	25.0	0
29	2601	114.0	99999	U	2	SUS	W	F	25.0	0
35	1850	233.0	99999	U	1	ARR	B	M	20.0	0
40	801	113.0	64105	N	1	ARR	B	M	30.0	0
42	840	133.0	64110	U	1	VIC	B	M	25.0	0
44	501	142.0	64109	U	1	SUS	U	U	25.0	0
47	1198	222.0	64110	U	1	VIC	W	M	26.0	0
49	2655	141.0	64110	U	2	ARR	B	M	20.0	0

Result 5: Cluster 0 for K prototype

From Result 6, Cluster 1 by K prototype was impacted by

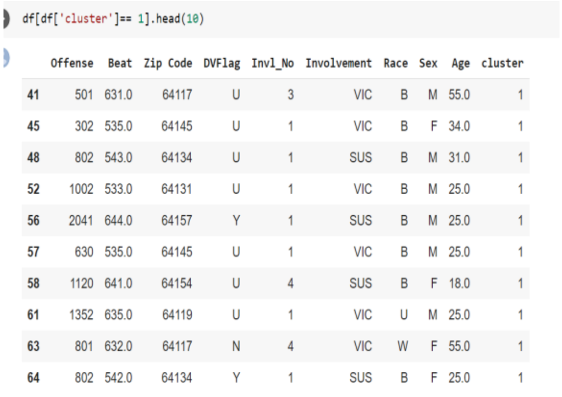
age-30-50,

zipcode-64117 to 64145

beat 540-630 and involvement :1,3,4 and Race: Black.

Offense- 300-1300

Which are having less crime compared to cluster 0.



```
df[df['cluster']== 1].head(10)
```

	Offense	Beat	Zip Code	DVFlag	Invl_No	Involvement	Race	Sex	Age	cluster
41	501	631.0	64117	U	3	VIC	B	M	55.0	1
45	302	535.0	64145	U	1	VIC	B	F	34.0	1
48	802	543.0	64134	U	1	SUS	B	M	31.0	1
52	1002	533.0	64131	U	1	VIC	B	M	25.0	1
56	2041	644.0	64157	Y	1	SUS	B	M	25.0	1
57	630	535.0	64145	U	1	VIC	B	M	25.0	1
58	1120	641.0	64154	U	4	SUS	B	F	18.0	1
61	1352	635.0	64119	U	1	VIC	U	M	25.0	1
63	801	632.0	64117	N	4	VIC	W	F	55.0	1
64	802	542.0	64134	Y	1	SUS	B	F	25.0	1

Result 6: Cluster 1 for K prototype

From Result 7, Cluster 2 by K prototype was impacted by

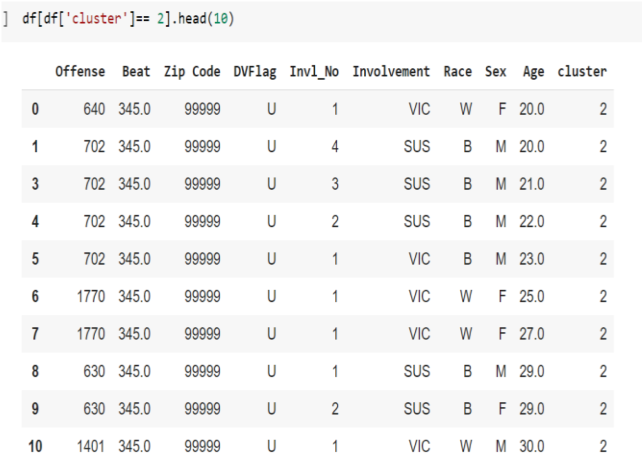
age-20-30,

zipcode-99999 beat 345 and involvement 1,2,3,4 and Race Black.

Offense- 640-1700

Dvflag-U

Which are having moderate prone crime compared to cluster 0,1.



```
] df[df['cluster']== 2].head(10)
```

	Offense	Beat	Zip Code	DVFlag	Invl_No	Involvement	Race	Sex	Age	cluster
0	640	345.0	99999	U	1	VIC	W	F	20.0	2
1	702	345.0	99999	U	4	SUS	B	M	20.0	2
3	702	345.0	99999	U	3	SUS	B	M	21.0	2
4	702	345.0	99999	U	2	SUS	B	M	22.0	2
5	702	345.0	99999	U	1	VIC	B	M	23.0	2
6	1770	345.0	99999	U	1	VIC	W	F	25.0	2
7	1770	345.0	99999	U	1	VIC	W	F	27.0	2
8	630	345.0	99999	U	1	SUS	B	M	29.0	2
9	630	345.0	99999	U	2	SUS	B	F	29.0	2
10	1401	345.0	99999	U	1	VIC	W	M	30.0	2

Result 7: Cluster 2 for K prototype

Conclusion

Collected the Data set from KCMO and hosted the Data in S3 Bucket. Later Using Google Colab, we tried to visualize the Data To visualize the data first we preprocessed the data by removing nulls and using method mode median to replace the nulls. Based on the Observations can determine that white Race of Male gender people with age around 40 yrs. in 5th & 7th Month in Area CPD&EPD mostly in zip code 64127 with 23.1% are prone for crime. And most of the crimes are not armed. Later for description we used the tokenization and tokenizer Standard Scaling and Feature Reduction used to normalize and select the features Trained the models Based on Area, DV Flag, Month, Sex, Race, Area, IBRS Code we predicted whether Firearm Used or not with Random forest, Decision Tree, Logistic Classifier, SVM, KNN. Finalized the Random forest model using confusion matrix, classification report and RMSE. Using The prediction can be further made useful for detecting the crimes in advance or by deploying more cops and patrols to the sensitive areas which are identified by the system. Finally Hosted the Visualized Data and Machine learning Model on the S3 Instance which helps the tenants if they want to relocate by checking the crime prone area. Clustering helps find out the crime hotspots throughout the longitudinal and latitudinal positions over a map. This plot will assist the police branch in figuring out which region calls for greater attention and consequently larger security forces might be deployed at that crime hotspot where firearms are used by criminals.

Hosted Model: https://atozcue.s3.amazonaws.com/index.html

In future, Deep learning methods like CNN can be done for future analysis.

Author Contribution

Akhil Teja Kanugolu: Data Preprocessing for 2019 Data set, MS Power Bi, S3 Bucket Utilization, Analysis using jupyter on Visualization, Machine learning models – Decision Tree, Random Forest, SVM. Hosted on S3 bucket, K prototype clustering.

Sai Rohith Guntupally: S3 Hosting of Data set, Kmeans Prototype, Machine learning logistic classifier, KNN.

REFERENCES

[1] Christian Tabedzki, Amruthesh Thirumalaiswamy and Paul van Vliet. 2018. Yo Home to Bel-Air: Predicting Crime on The Streets of Philadelphia

[2] Varvara Ingilevich and Sergey Ivanov. 2018. Crime rate prediction in the urban environment using social factors, Procedia Computer Science, 136, 472–478

[3] Jay Patel, Haresh Wala, Deepak Shahu, Hazel Lopes, "Intellectual and Enhance Digital Solution For Police Station", Smart City and Emerging Technology (ICSCET) 2018 International Conference on, pp. 1-4, 2018.

[4] Swati Nair, Saloni Soniminde, Sruthi Sureshbabu, Apurva Tamhankar, Sagar Kulkarni, "Assist Crime Prevention Using Machine Learning", SSRN Electronic Journal, 2019.

[5] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano et.al, “Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data” (2014), Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, pp. 427-434.