Implementing Several Techniques for Data Analytics

Name: Clinton Delvin

Student No: 10531003

Contents

[PART I – DATA WAREHOUSE / DATA MART 1](#_Toc39290279)

[1. Introduction 2](#_Toc39290280)

[1.1. Reasons for selecting the subject area AND DATA 2](#_Toc39290281)

[1.2. Vision and Goals 2](#_Toc39290282)

[1.3. Key StakeHolders 3](#_Toc39290283)

[2. Schema / Dimensional Model 3](#_Toc39290284)

[3. Implementation of Data Warehouse /Data House 3](#_Toc39290285)

[4. ETL to Popuplate the Data WareHouse / Data Mart 4](#_Toc39290286)

[5. Visualisations and Reports 7](#_Toc39290287)

[5.1. TABLEAU Visualizations 7](#_Toc39290288)

[5.2. Reports – SSRS 10](#_Toc39290289)

[6. Conclusions 13](#_Toc39290290)

[PART II – DATA MINING 14](#_Toc39290291)

[1. Choosing and Exploring Data 14](#_Toc39290292)

[1.1. Data Source 14](#_Toc39290293)

[1.2. Data Exploration 14](#_Toc39290294)

[2. Analysing, Cleaning and Preparing Data 14](#_Toc39290295)

[3. Model Building, Testing and Evaluation 17](#_Toc39290296)

[4. Conclusions 18](#_Toc39290297)

[References 18](#_Toc39290298)

[Appendix A – CREATE Table queries for Data Warehouse / Data Mart 19](#_Toc39290299)

[Appendix C – Python Code 20](#_Toc39290300)

# PART I – DATA WAREHOUSE / DATA MART

# 1. Introduction

Crime reports provide information about the various forms of crimes that have been committed-in all the relevant and essential descriptions. Specifically concentrating on the forms of crime, the number of victims, the nationality of the victim and the perpetrator and/or the nature of the offences. In view of the quantity of data in crime reports, data warehouse models can be used to access different sections of the reports and statistics can be used to summarize the information.

In the study of criminology, data storage and statistics may be very important. Using these instruments, our knowledge of the various types of crimes being committed, the nature of the crimes being committed, the race of people affected that can be compared with the race of people committing the crime, and the various features of the crimes will be nothing more than speculation.

Crime warehouse models augmented by statistics can theoretically be an significant predictor of a society's well-being. That crime rates could lead to misery in society, and abnormal victimization among groups may suggest that a society needs a governing body's attention. On the other hand, the reduction in the crime rate found in these studies and figures is an indication of better living conditions. Crime reports and statistics also have a major role to play in helping researchers, governments and other organizations to develop and test crime and crime theories and victims, in helping policymakers to devise policies to reduce crime, help victims of crime, and manage crime effectively. Without clear data and observations into crime, not only will initiatives designed to minimize crime and oppression fail but they would also represent unsuitable or inefficient assets and resources.

## 1.1. Reasons for selecting the subject area AND DATA

The primary aim of this analysis was to examine and explain the different forms of hate crimes committed over the years in various states of the United States. This analysis focused specifically on understanding and understanding how various factors interact. For example, if a correlation exists between the types of crimes against a specific race. So that proper rules are implemented or kept in place so that it would help to reduce the crime rate.

CRIME ANALYSIS IN THE STATES OF USA FROM THE YEAR 1997-2017.

## 1.2. Vision and Goals

1. Better Patrolling:

The rising risk for violence, with the aid of hate crime data, can be expected by criminal justice practitioners and policy makers. Suitable enforcement of law can be carried out in order to avoid the crimes predicted. The predictive data can be used to focus on certain areas and to optimize the resources available. The key aim is to predict and effectively eliminate these risks by forecasting this risk well in advance.

1. Improved Public Community Relations:

Crime data are an excellent way to improve relationships between various communities. It gives the police department and the people they represent an improved opportunity to look back at the ills of society. This helps us to improve transparency. The public will also be aware of how well politicians and law enforcement officials operate to protect the environment, create trust and ensure good working ties.

1. Monitoring Finances:

Budgets and financing are a key part of a local, state and national governing body and law enforcement. Suitable investment and distribution of monetary resources in the right locations and services will lead to improved protection in different communities. Without crime statics, the distribution of financial capital will be difficult to accurately view and forecast.

## 1.3. Key StakeHolders

The primary stakeholder is the Government in evaluating data regarding crimes, why the crime is predominantly committed, and in which region or state the crime rate is more relevant such that the government has a better understanding as to whether the crime rate can be reduced to a minimum.

# 2. Schema / Dimensional Model

* schema developed in visio or draw.io and discuss the reasons for the design.

# 3. Implementation of Data Warehouse /Data House

In order to examine violence in the United States, we have chosen the crime data package. We have segregated the data into following dimensions and fact as shown below:

|  |  |
| --- | --- |
| Calendar\_Dim | Shows date and year of the incident |
| Crime\_Scene\_Type\_Dim | Shows the crime area where crime had occurred |
| Crime\_Type\_Dim | Shows the type of crime |
| Race\_Dim | Shows the race of the offender and victim |
| State\_Dim | Shows the states in which crime happening |

One fact table used in star scheme to analyze.

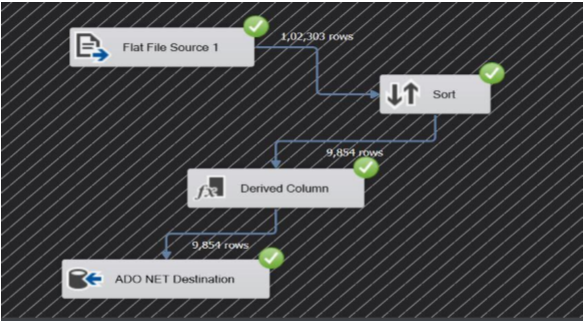
|  |  |
| --- | --- |
| Hate\_Crime\_Incident\_Fact | Holds information about all the dimensions |

# 4. ETL to Popuplate the Data WareHouse / Data Mart

Visual Studio SSDT has been used to load the details in SQL dimensional tables. Refer to Appendix A for further details.

CRIME ANALYSIS IN THE STATES OF USA FROM THE YEAR 1997-2017.

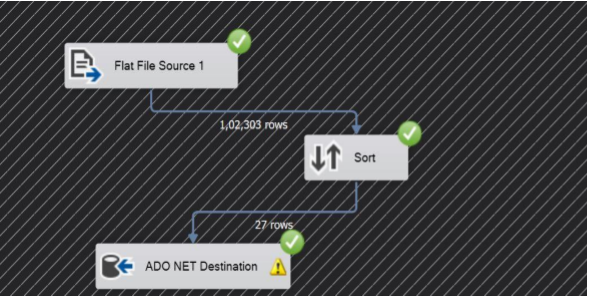
ETL PROCESS 1: CALENDAR\_DIM



Over here ther data is loaded into the table named as Calendar\_Dim from the crime dataset.

We used sort here to delete the few rows and the derivative column in order to replace the original column.

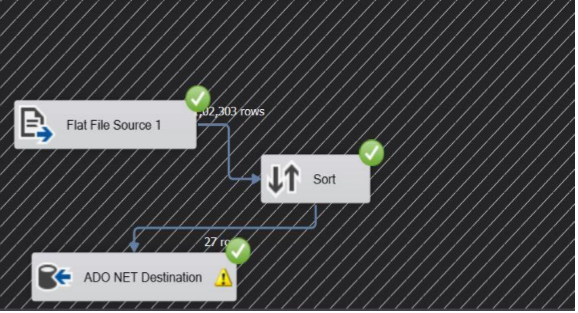
ETL PROCESS 2: CRIME\_SCENE\_TYPE\_DIM



Over here the data is loaded into the table named as Calendar\_Scene\_Type\_Dim from the crime dataset.

Over here sort function is used so that the duplicate values are removed and only the unique values are pushed into the table named as Crime\_Scene\_Type\_Dim.

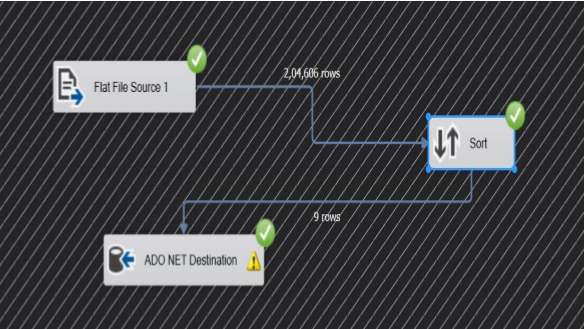
ETL PROCESS 3: CRIME\_TYPE\_DIM



Over here ther data is loaded into the table named as Calendar\_Type\_Dim from the crime dataset.

Over here sort function is used so that the duplicate values are removed and only the unique values are pushed into the table named as Crime\_Type\_Dim. In this table column named as Crime\_Type\_Id consists of unique values. So that this column can uniquely identify every crime type comitted in the state of USA.

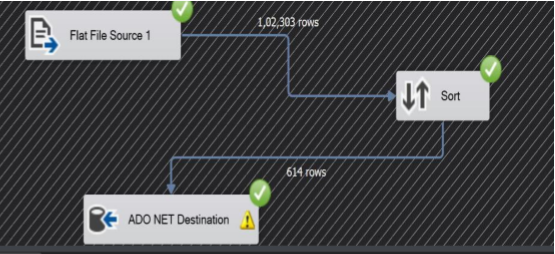
ETL PROCESS 4: RACE\_DIM



Over here the data is loaded into the table named as Race\_Dim from the crime dataset.

Over here sort function is used so that the duplicate values are removed and only the unique values are pushed into the table named as Race\_Dim. Race of the offender and victim are stored in this table.

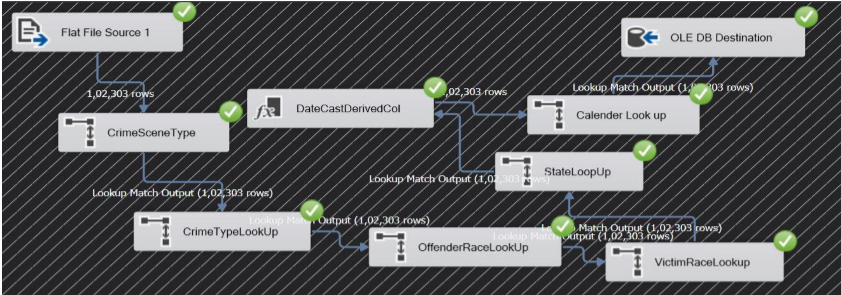
ETL PROCESS 5: STATE\_DIM



Over here the data is loaded into the table named as State\_Dim from the crime dataset.

Over here sort function is used so that the duplicate values are removed and only the unique values are pushed into the table named as State\_Dim. Over here State\_Id will be unique that will uniquely identify every state in USA.

ETL PROCESS FIVE: FACT\_DIM



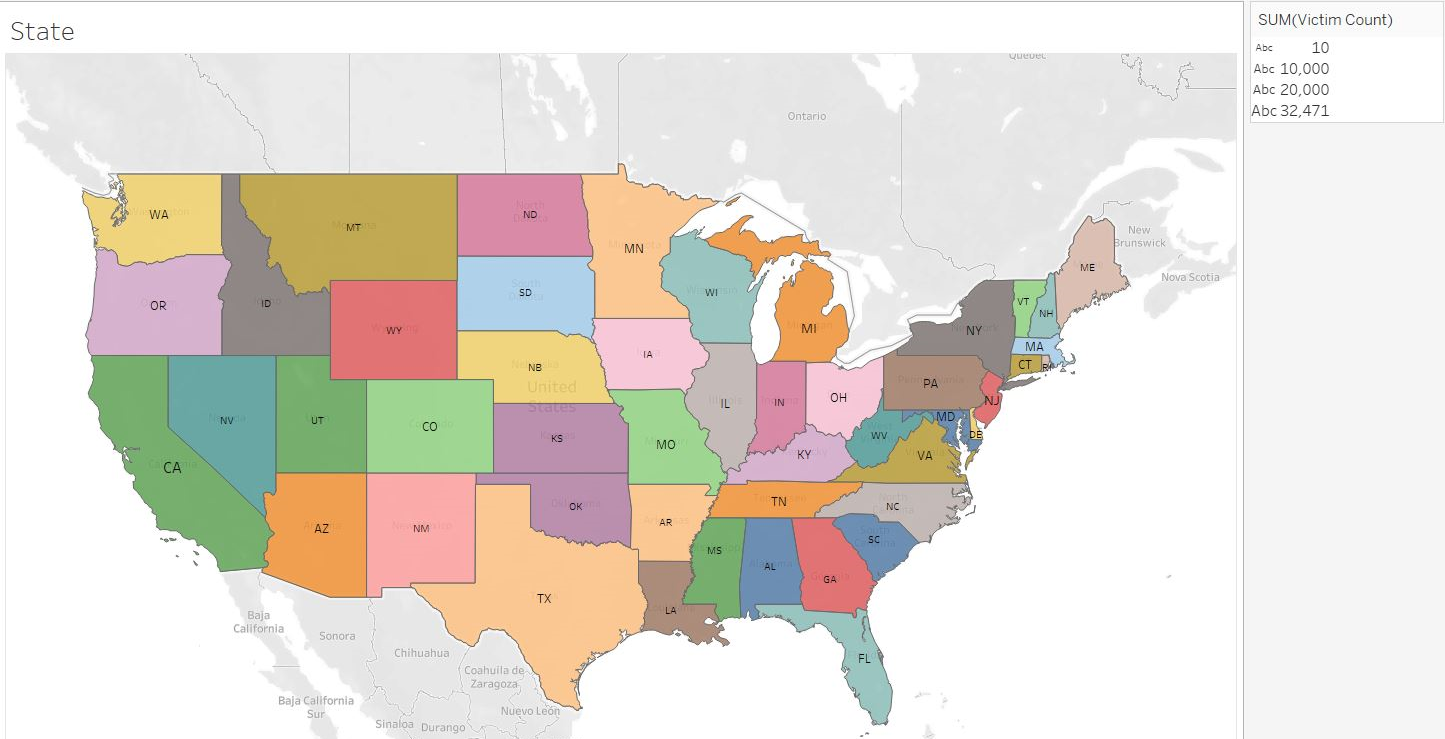
Over here the data is loaded into the table named as Fact\_Table from the crime dataset by connecting all the dimensions one to another using the lookup feature. Lookup is used over here to compare the results of the source and the destination. Lookup matches the data from the previous output that the other ETL processes gave and provides with a suitable result. This final ETL process will have the values of all the primary key combined and these primary keys of other tables will be stored as a foreign key in this table. Fact table will also have a primary key on its own so that it can uniquely identify every row in the table.

# 5. Visualisations and Reports

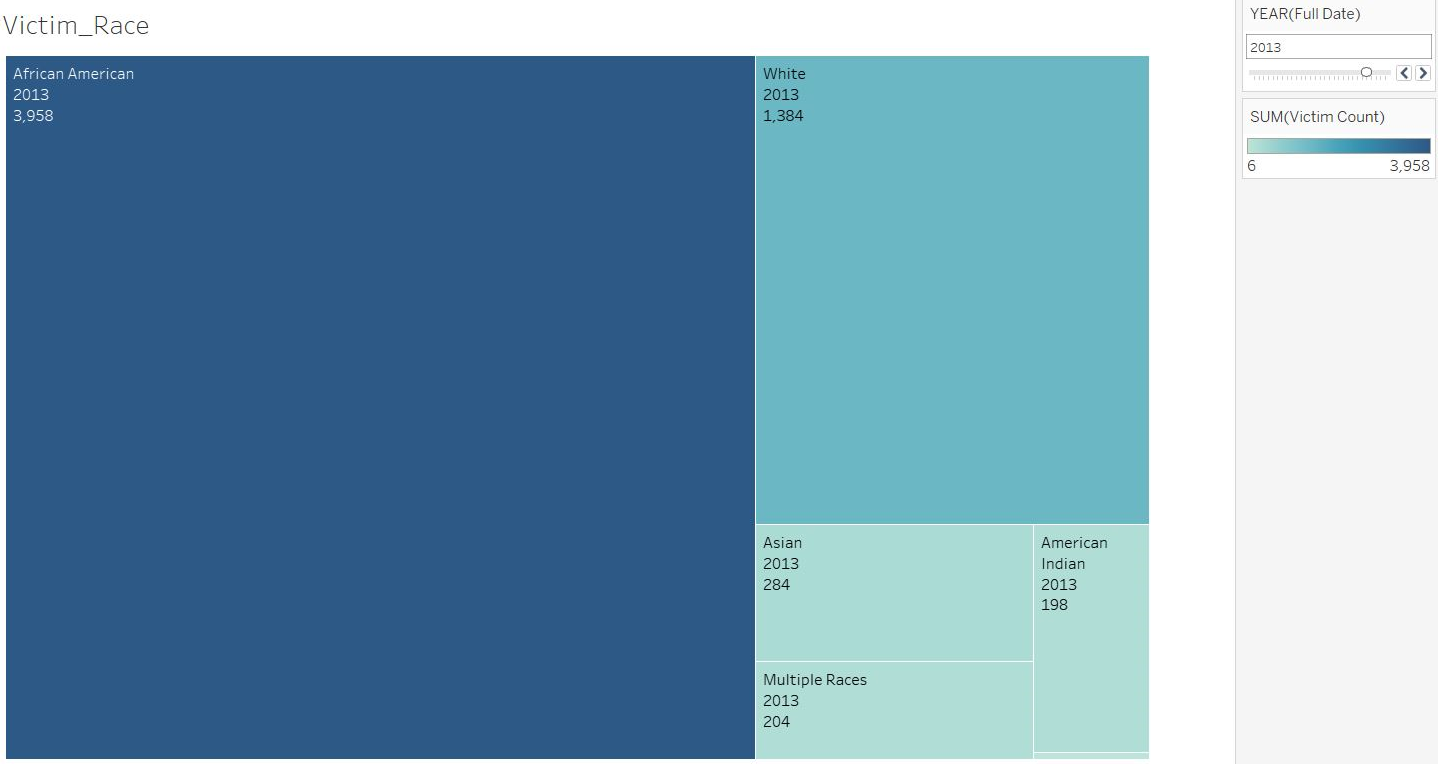
## 5.1. TABLEAU Visualizations

1) STATE:

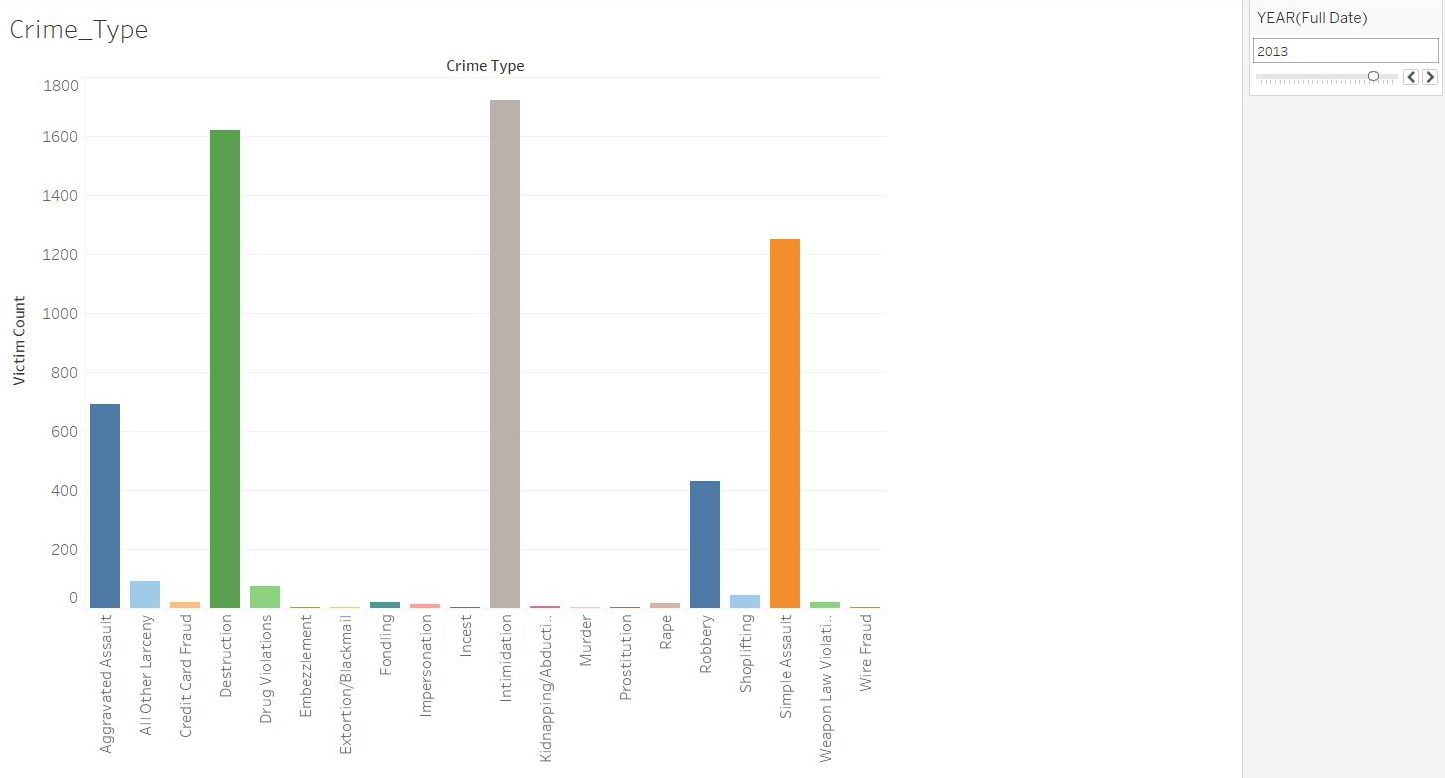
In this visualization, the total number of states that are in the US is displayed. It also indicates the states where the crime has occurred.



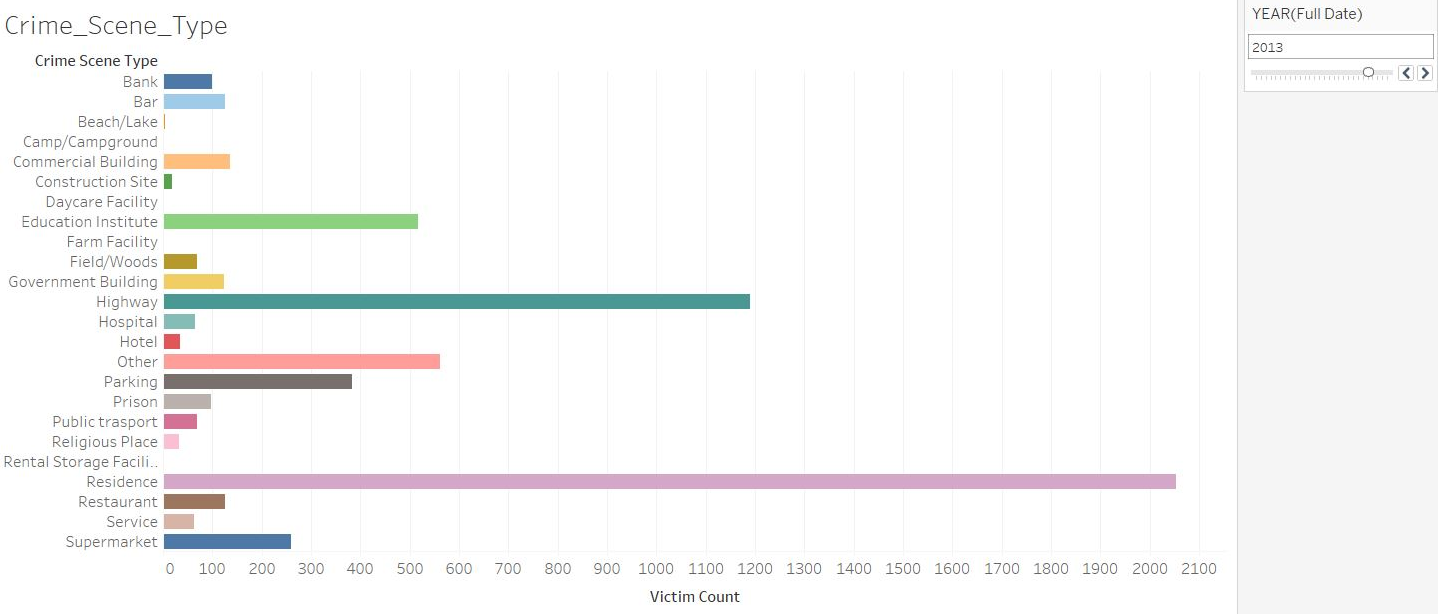
2) Here we see the violence that takes place in 2013 in various races. In this table, we used to illustrate the crimes that arise in various races more clearly. The graph clearly indicates that American African people are generally the victim. A slider functionality is also added to check the victims as per the year.



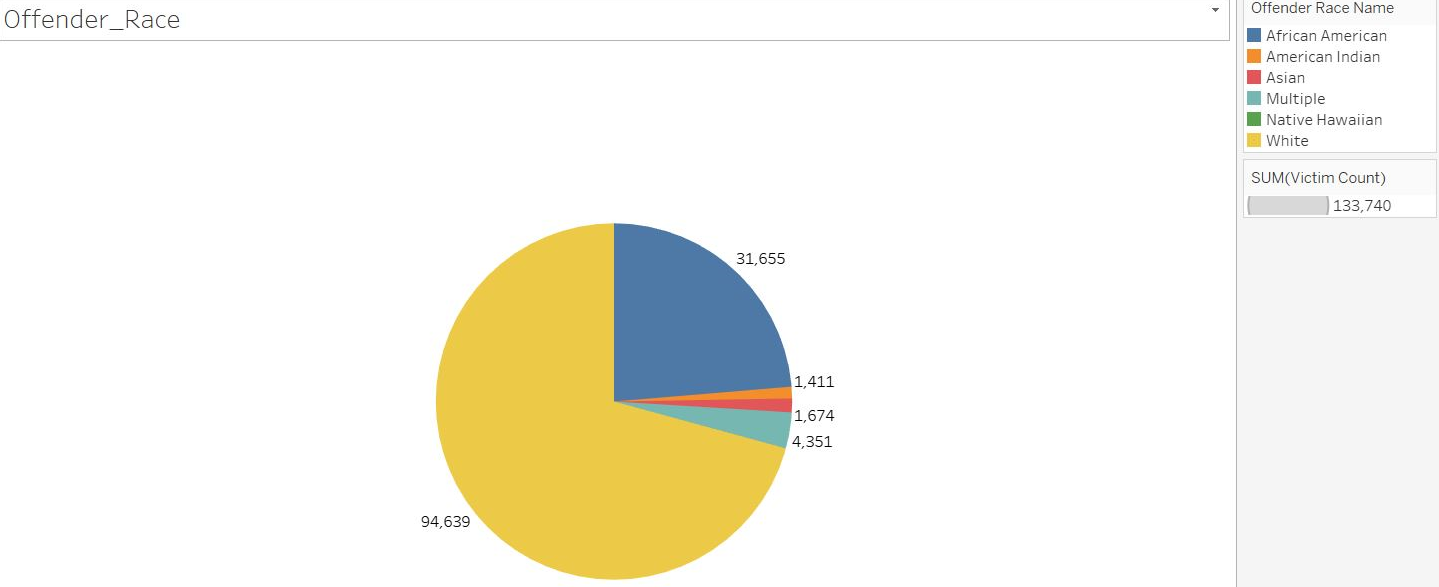
3) Below you can see the various forms of crimes in 2013 that are happening more frequently. We used bar chart here to illustrate various forms of crimes. We can see that intimidation is the major crime which occurred more frequently than other crimes. A slider functionality is also added here to check the crime type as per year so that proper analysis could be done.



4) Here, we examine the occurrence of crimes in various places or climates in the States of the United States. This is a frequency diagram to evaluate the crime level in different places or in various environments. From this, we can see that in the Residential area followed by the highway more crimes have occurred.



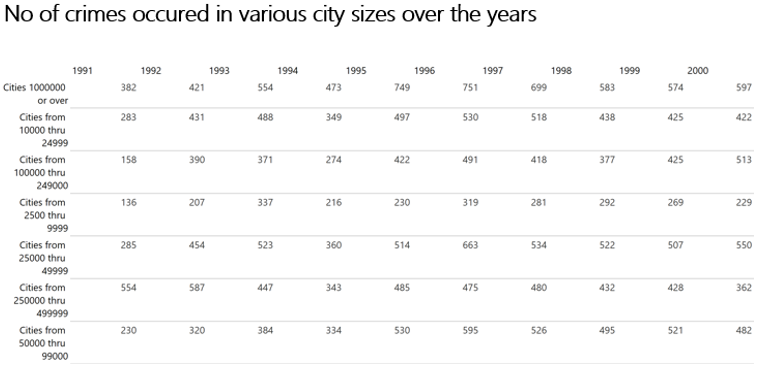
5) Here, we have visualized on the offender race and we conclude from the report that the offender race is generally White followed by African American.



See Appendix A for code

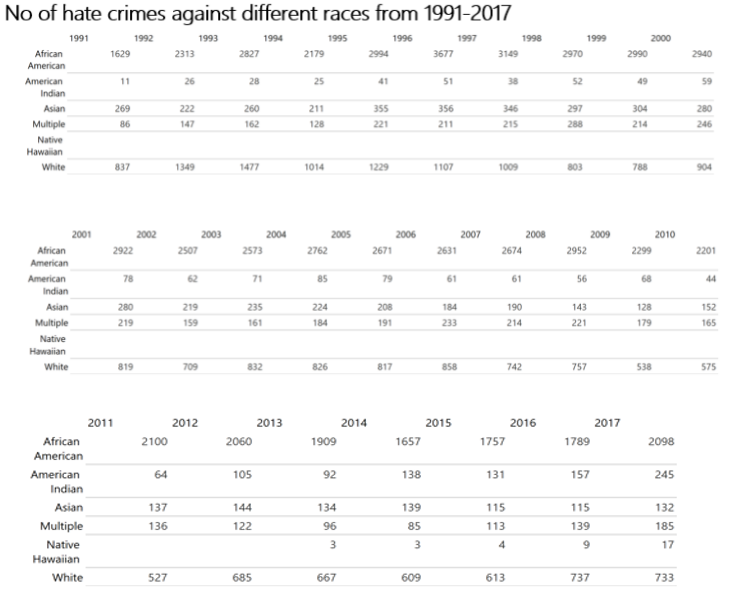
## 5.2. Reports – SSRS

1)



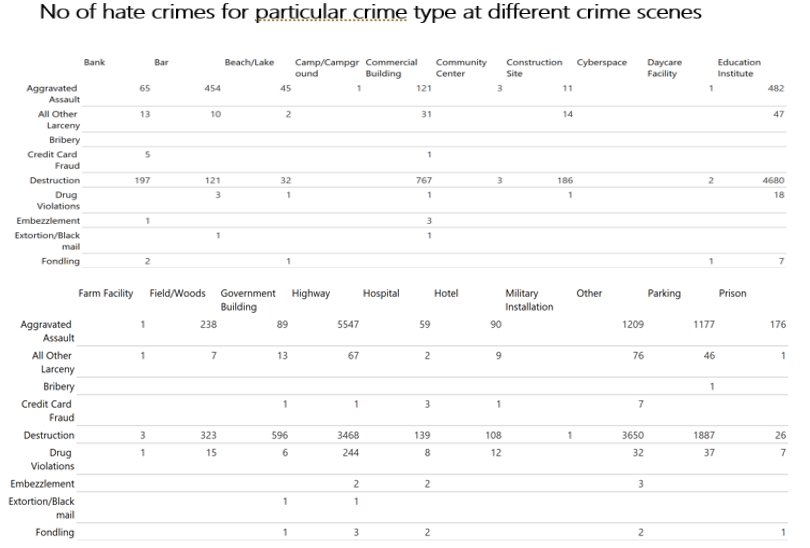
The first report is a matrix report which shows the number of crimes that took place over the years 1991 to 2017 in cities with different sizes with population. One insight I discovered was that in large cities with populations larger than 1 million, the number of crimes remained low over the years, but in smaller populated towns, the number of hate crimes was rising. This may reflect the notion of transparency in larger cities than in smaller cities.

2)



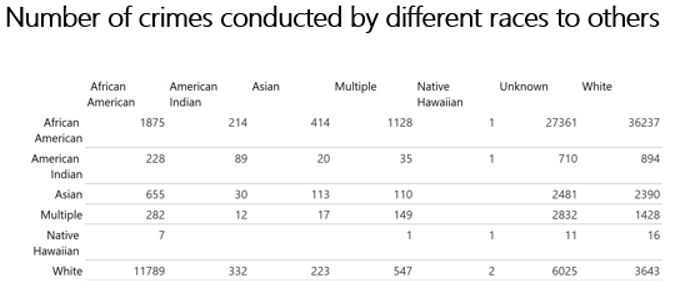
The second study reveals the number of victims of hate crimes in the years from various races. This graph reveals that African Americans are the most impacted over the last few years.

3)



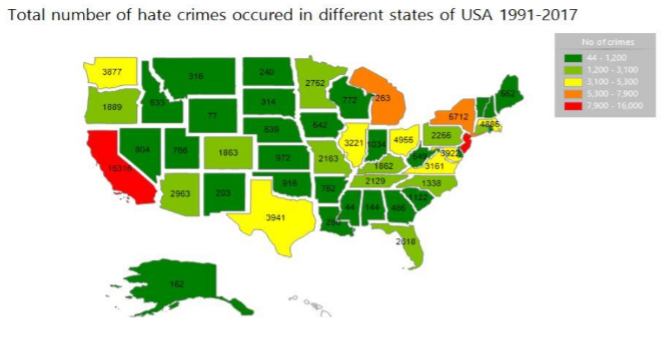
Third report attempts to find a correlation by means of a matrix report between the crime type and the crime scene type. Here, most people are targeted at schools and commercial buildings.

4)



The fourth study is another matrix which attempts to establish a connection between the perpetrator race and the victim race. The inhabitants of Africa and the Americas are the two most criminally opposed groups.

5)



The fifth study is an illustrative map of the United States showing the different nations and the number of crimes that have taken place. The states are coded in a red-yellow-gray color scheme with the most crime in red and the lowest in gray. In order for us to see in depth the consistency of the different states of hate crimes, this study is extremely useful.

# 6. Conclusions

From 1991-2017, we analyzed the FBI statistical data on all cases of hate crime in the United States. We also studied the characteristics of the incidents and searched for similarity of some aspect in the dimension’s crime form, crime scene, race and date. We find that the African American and White minority have committed the largest number of hate crimes over the years in the USA. Central states have the lowest number of racial crimes reported over the years. The most notorious sites of crime are the sites of education and economic activity in which crimes of aggravated assault and destruction are the most committed. Over the years in the US, hate crime cases have risen and the warehouse is the ultimate USA environment for violence.

# PART II – DATA MINING

# 1. Choosing and Exploring Data

## 1.1. Data Source

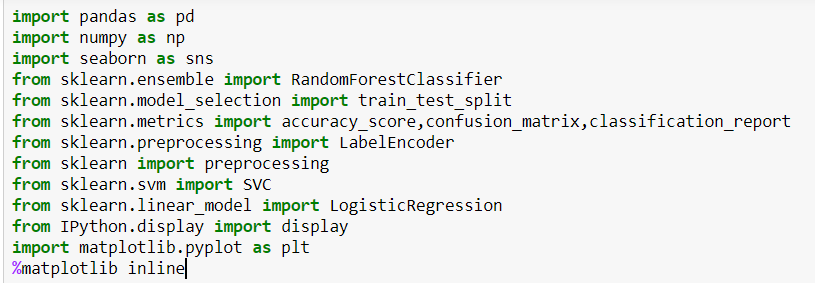
As of 26 June 2019, the Navigable Waterways dataset is part of the U.S. Department of Transportation (USDOT)/Office of Transportation Statistics (BTS) National Transportation Atlas Database (NTAD). The National Waterway Network is a comprehensive network database of national navigable waterways. The data set covers the 48 contiguous states plus the District of Columbia, Hawaii, Alaska, Puerto Rico and the water links. The nominal size of the dataset varies with the source material. The majority of the information is 1:100,000 with larger scales used in harbor / bay / port areas and smaller scales used in open waters. The data can be used for empirical waterway efficiency studies, for the collection and visualization of product flow statistics.

## 1.2. Data Exploration

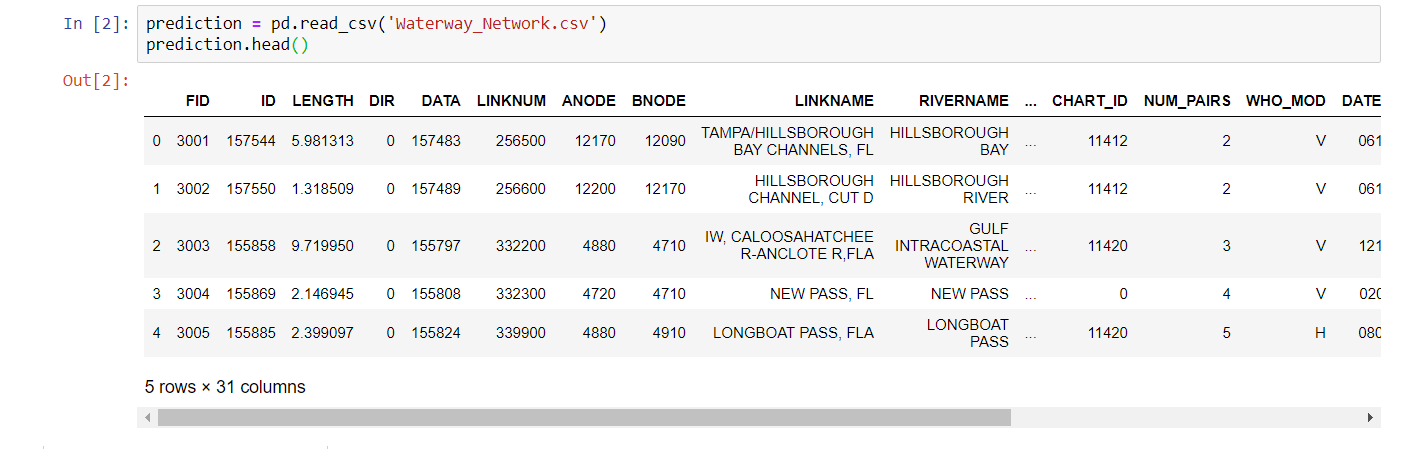
Over here the various data is explored and a prediction is done to check the geographical location where the waterways are flowing from in the different part of united states. Null values are eliminated from this data set and then different algorithms are applied to this dataset to check the accuracy of this algorithm so that effective prediction of these data sets could be done.

# 2. Analysing, Cleaning and Preparing Data

Over here to analyze the data all the required packages are imported into python code.

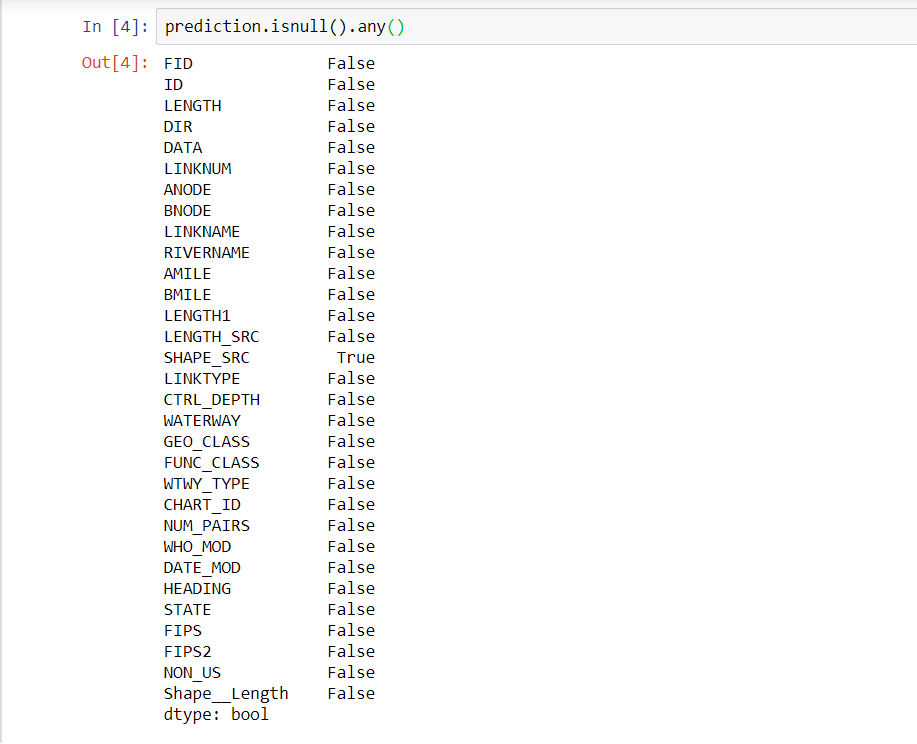


Once all the required packages are loaded into the python code then the entire CSV file is read and is inserted into a variable named as prediction.



Once the data is loaded into a variable named as predictions.

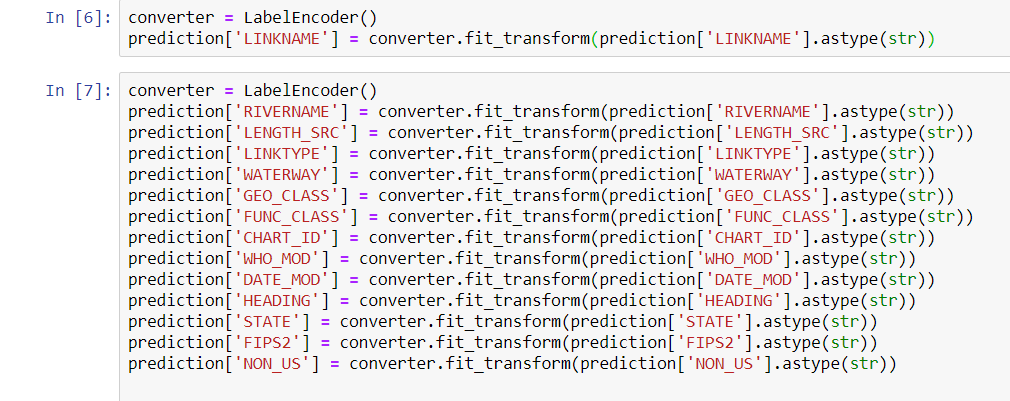
Then data inside the prediction is checked and verified if the data inside the prediction variable contains any NULL value or not by running a python code that is shown below :



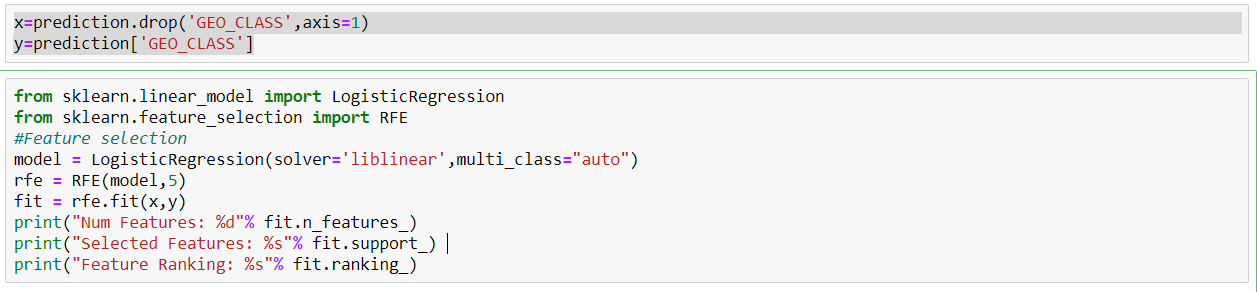
As the column named as SHAPE\_SRC contains NULL value, the column is dropped by running the below code :



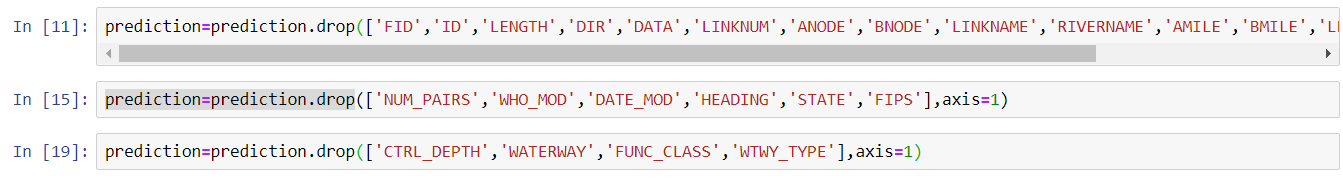
Once the NULL values are eliminated then the data type of these columns are checked and then conversion of these data type to integer so that analysis could be done on the data provided. Below is the snipped of the code where a “convert” function is used to convert the data from one format to another:



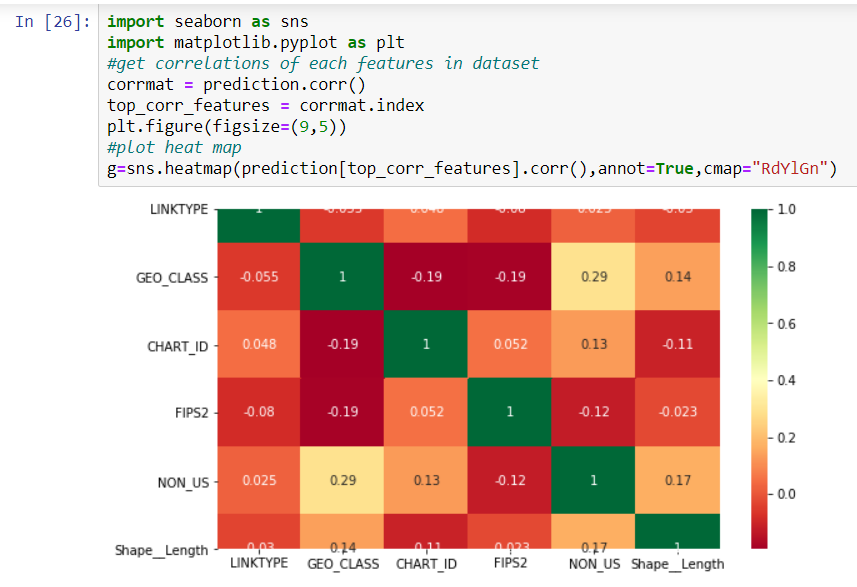
Then once the data is converted to a analytical format. Then the data that has to be predicted will be dropped from the X axis and then the same data will be inserted in the Y- axis with the following command:



Once the code is run. All the False column has to be dropped from the X axis so that prediction can be carried out.

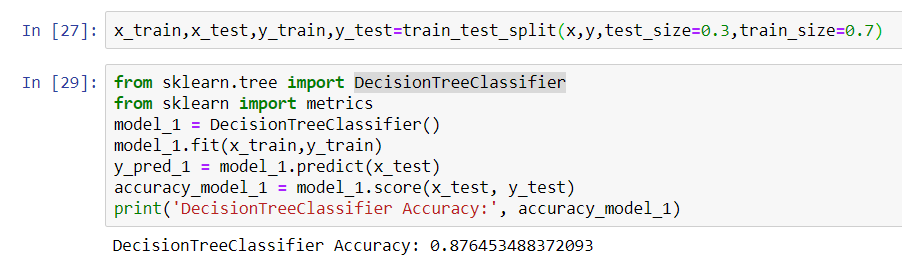


A correlation heat map is then generated.

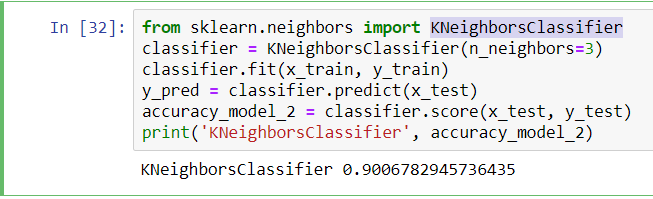


# 3. **Model Building, Testing and Evaluation**

The X and Y axis that is taken into consideration for prediction here. Algorithms like DecisionTreeClassifier and KNeighborsClassifier are used on the dataset. The score for this data is calculated.



Accuracy of 87% is achieved if Decision Tree Classifier is implemented here.



Accuracy of 90% is achieved if KNeighborsClassifier is implemented here.

# 4. Conclusions

After running various algorithms on the data set. It was found out that Decision Tree Classifier and KNeighborsClassifier is best suited for the dataset. Due to which these two algorithms are used for prediction in this dataset.

# References

Anon., 2018. *Crime Data Explorer.* [Online]   
Available at: https://crime-data-explorer.fr.cloud.gov/  
[Accessed 29 April 2020].

Anon., 2018. *Crime Data of USA.* [Online]   
Available at: https://crime-data-explorer.fr.cloud.gov/downloads-and-docs  
[Accessed 28 April 2020].

Anon., 2018. *Navigable Waterway Nodes.* [Online]   
Available at: https://data-usdot.opendata.arcgis.com/datasets/navigable-waterway-lines  
[Accessed 20 April 2020].

# Appendix A – CREATE Table queries for Data Warehouse / Data Mart

APPENDIX A.1: QUERY FOR TABLE CREATION

CREATE TABLE Calendar\_Dim ( CalendarKey INT NOT NULL IDENTITY, FullDate DATE,

HATE CRIME ANALYSIS IN THE STATES OF USA FROM THE YEAR 1997-2017 24

Year\_ Date, PRIMARY KEY (CalendarKey)); GO

CREATE TABLE Crime\_Scene\_Type\_Dim (Crime\_Scene\_ID INT NOT NULL IDENTITY, CrimeType varchar (500), PRIMARY KEY (Crime\_Scene\_ID)); GO

CREATE TABLE Crime\_Type\_Dim (Crime\_Type\_ID INT NOT NULL IDENTITY, Crime\_Type varchar (500), PRIMARY KEY (Crime\_Type\_ID)); GO

CREATE TABLE State\_Dim (State\_Id varchar (100), State\_Abbr varchar (100), State\_Name varchar (200), Population\_Group\_Description varchar (200), PRIMARY KEY (State\_Id)); GO

CREATE TABLE Race\_Dim (Race\_Id INT NOT NULL IDENTITY, Race\_Name varchar (500), PRIMARY KEY (Race\_Id)); CREATE TABLE Hate\_Crime\_Incident\_Fact ( CalendarKey INT, Crime\_Scene\_ID INT, Crime\_Type\_ID INT, Victim\_Race\_Id INT, Offender\_Race\_Id INT, State\_Id VARCHAR (100), Victim\_Count INT,

HATE CRIME ANALYSIS IN THE STATES OF USA FROM THE YEAR 1997-2017 25

PRIMARY KEY (CalendarKey, Crime\_Scene\_ID, Crime\_Type\_ID, Victim\_Race\_Id, Offender\_Race\_Id, State\_Id), FOREIGN KEY (Calendarkey) REFERENCES Calendar\_Dim (CalendarKey), FOREIGN KEY (Crime\_Scene\_ID) REFERENCES Crime\_Scene\_Type\_Dim (Crime\_Scene\_ID), FOREIGN KEY (Crime\_Type\_ID) REFERENCES Crime\_Type\_Dim (Crime\_Type\_ID), FOREIGN KEY (Victim\_Race\_Id) REFERENCES Race\_Dim (Race\_Id), FOREIGN KEY (Offender\_Race\_Id) REFERENCES Race\_Dim (Race\_Id), FOREIGN KEY (State\_Id) REFERENCES State\_Dim (State\_Id), );

**OR**

# Appendix C – Python Code

import pandas as pd

import numpy as np

import seaborn as sns

from sklearn.ensemble import RandomForestClassifier

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score,confusion\_matrix,classification\_report

from sklearn.preprocessing import LabelEncoder

from sklearn import preprocessing

from sklearn.svm import SVC

from sklearn.linear\_model import LogisticRegression

from IPython.display import display

import matplotlib.pyplot as plt

%matplotlib inline

prediction = pd.read\_csv('Waterway\_Network.csv')

prediction.head()

prediction.isnull().any()

prediction=prediction.drop('SHAPE\_SRC',axis=1)

converter = LabelEncoder()

prediction['LINKNAME'] = converter.fit\_transform(prediction['LINKNAME'].astype(str))

converter = LabelEncoder()

prediction['RIVERNAME'] = converter.fit\_transform(prediction['RIVERNAME'].astype(str))

prediction['LENGTH\_SRC'] = converter.fit\_transform(prediction['LENGTH\_SRC'].astype(str))

prediction['LINKTYPE'] = converter.fit\_transform(prediction['LINKTYPE'].astype(str))

prediction['WATERWAY'] = converter.fit\_transform(prediction['WATERWAY'].astype(str))

prediction['GEO\_CLASS'] = converter.fit\_transform(prediction['GEO\_CLASS'].astype(str))

prediction['FUNC\_CLASS'] = converter.fit\_transform(prediction['FUNC\_CLASS'].astype(str))

prediction['CHART\_ID'] = converter.fit\_transform(prediction['CHART\_ID'].astype(str))

prediction['WHO\_MOD'] = converter.fit\_transform(prediction['WHO\_MOD'].astype(str))

prediction['DATE\_MOD'] = converter.fit\_transform(prediction['DATE\_MOD'].astype(str))

prediction['HEADING'] = converter.fit\_transform(prediction['HEADING'].astype(str))

prediction['STATE'] = converter.fit\_transform(prediction['STATE'].astype(str))

prediction['FIPS2'] = converter.fit\_transform(prediction['FIPS2'].astype(str))

prediction['NON\_US'] = converter.fit\_transform(prediction['NON\_US'].astype(str))

prediction.info()

x=prediction.drop('GEO\_CLASS',axis=1)

y=prediction['GEO\_CLASS']

from sklearn.linear\_model import LogisticRegression

from sklearn.feature\_selection import RFE

#Feature selection

model = LogisticRegression(solver='liblinear',multi\_class="auto")

rfe = RFE(model,5)

fit = rfe.fit(x,y)

print("Num Features: %d"% fit.n\_features\_)

print("Selected Features: %s"% fit.support\_)

print("Feature Ranking: %s"% fit.ranking\_)

prediction=prediction.drop(['FID','ID','LENGTH','DIR','DATA','LINKNUM','ANODE','BNODE','LINKNAME','RIVERNAME','AMILE','BMILE','LENGTH1','LENGTH\_SRC'],axis=1)

prediction=prediction.drop(['NUM\_PAIRS','WHO\_MOD','DATE\_MOD','HEADING','STATE','FIPS'],axis=1)

prediction=prediction.drop(['CTRL\_DEPTH','WATERWAY','FUNC\_CLASS','WTWY\_TYPE'],axis=1)

import seaborn as sns

import matplotlib.pyplot as plt

#get correlations of each features in dataset

corrmat = prediction.corr()

top\_corr\_features = corrmat.index

plt.figure(figsize=(9,5))

#plot heat map

g=sns.heatmap(prediction[top\_corr\_features].corr(),annot=True,cmap="RdYlGn")

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.3,train\_size=0.7)

from sklearn.tree import DecisionTreeClassifier

from sklearn import metrics

model\_1 = DecisionTreeClassifier()

model\_1.fit(x\_train,y\_train)

y\_pred\_1 = model\_1.predict(x\_test)

accuracy\_model\_1 = model\_1.score(x\_test, y\_test)

print('DecisionTreeClassifier Accuracy:', accuracy\_model\_1)

from sklearn.neighbors import KNeighborsClassifier

classifier = KNeighborsClassifier(n\_neighbors=3)

classifier.fit(x\_train, y\_train)

y\_pred = classifier.predict(x\_test)

accuracy\_model\_2 = classifier.score(x\_test, y\_test)

print('KNeighborsClassifier', accuracy\_model\_2)