NYU

The Anatomy of NYPD Crime Records:

Race, Gender, Politics and More

By: Jiaming Chen, Zizhen Chen, Jiayuan Huang, Yu Wu

Goals

- Intergroup Solidarity, Felt and Shared Foreignness between Asians and Hispanics in the U.S ✅
- They are under similar social pressure ？
  Similar crime behavior patterns in NYC (correlation based on crime amounts and types distribution) ❌

- **No hypothesis.**
- **All Races**
- **Various Analysis**
- **Direct Explaining**

PART 01

Data

Unbiased data = How?

- Not arrest data but Complaint Data
- Omit minor crimes such as marijuwana, loitering, minor valued theft

Cleaned dataset = what columns?

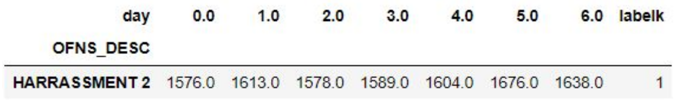
- Race, Gender, Age, Crime type, Victim, Location.....

Methodologies
&
Corresponding Results

PART 02

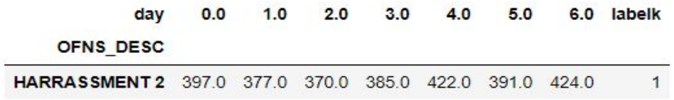
Clustering Analysis

Smallest Cluster for races



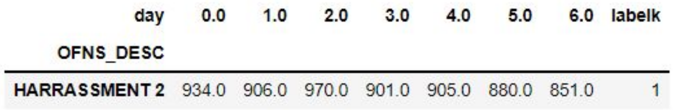
day	0.0	1.0	2.0	3.0	4.0	5.0	6.0	labelk
OFNS_DESC								
HARRASSMENT 2	1576.0	1613.0	1578.0	1589.0	1604.0	1676.0	1638.0	1

Hispanic



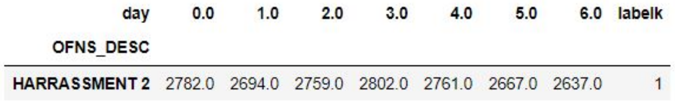
day	0.0	1.0	2.0	3.0	4.0	5.0	6.0	labelk
OFNS_DESC								
HARRASSMENT 2	397.0	377.0	370.0	385.0	422.0	391.0	424.0	1

Asian



day	0.0	1.0	2.0	3.0	4.0	5.0	6.0	labelk
OFNS_DESC								
HARRASSMENT 2	934.0	906.0	970.0	901.0	905.0	880.0	851.0	1

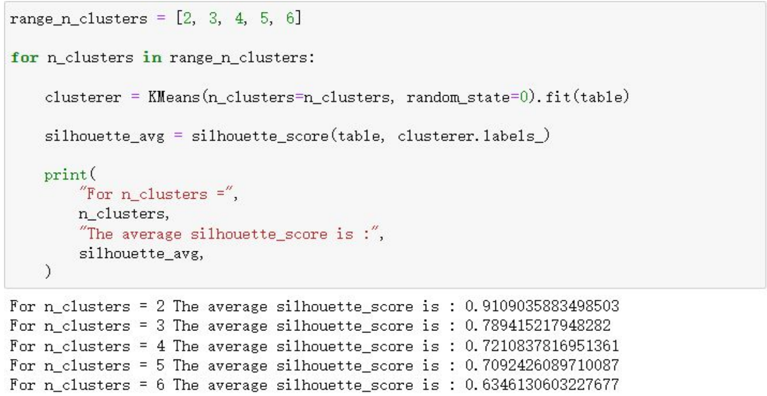
White



day	0.0	1.0	2.0	3.0	4.0	5.0	6.0	labelk
OFNS_DESC								
HARRASSMENT 2	2782.0	2694.0	2759.0	2802.0	2761.0	2667.0	2637.0	1

Black

Clustering using K-means with silhouette score



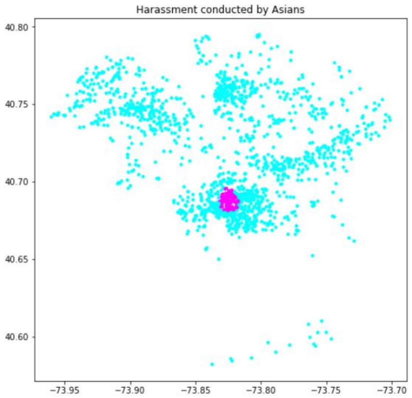
```
range_n_clusters = [2, 3, 4, 5, 6]

for n_clusters in range_n_clusters:

    clusterer = KMeans(n_clusters=n_clusters, random_state=0).fit(table)

    silhouette_avg = silhouette_score(table, clusterer.labels_)

    print(
        "For n_clusters =",
        n_clusters,
        "The average silhouette_score is :",
        silhouette_avg,
    )
```

```
For n_clusters = 2 The average silhouette_score is : 0.9109035883498503
For n_clusters = 3 The average silhouette_score is : 0.789415217948282
For n_clusters = 4 The average silhouette_score is : 0.7210837816951361
For n_clusters = 5 The average silhouette_score is : 0.7092426089710087
For n_clusters = 6 The average silhouette_score is : 0.6346130603227677
```

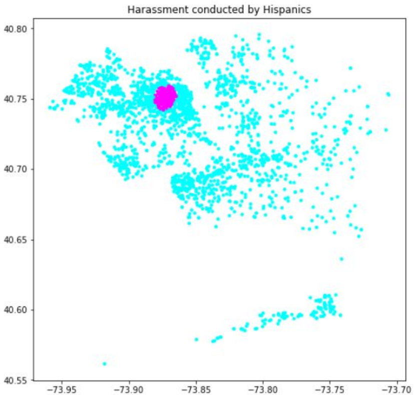
- As we can see, we will select 2 as the number of clusters
- The model cluster the harassment 2 into the smallest cluster for all the races

PART 02

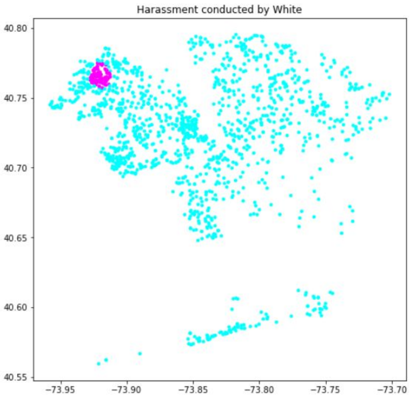
DBScan Analysis



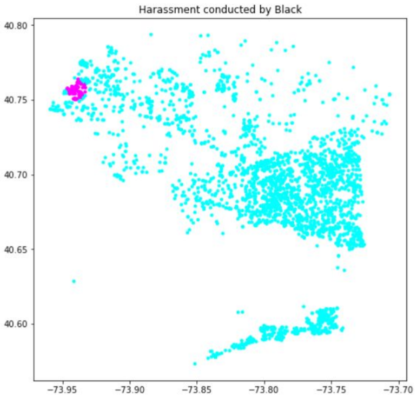
Asian



Hispanic



White

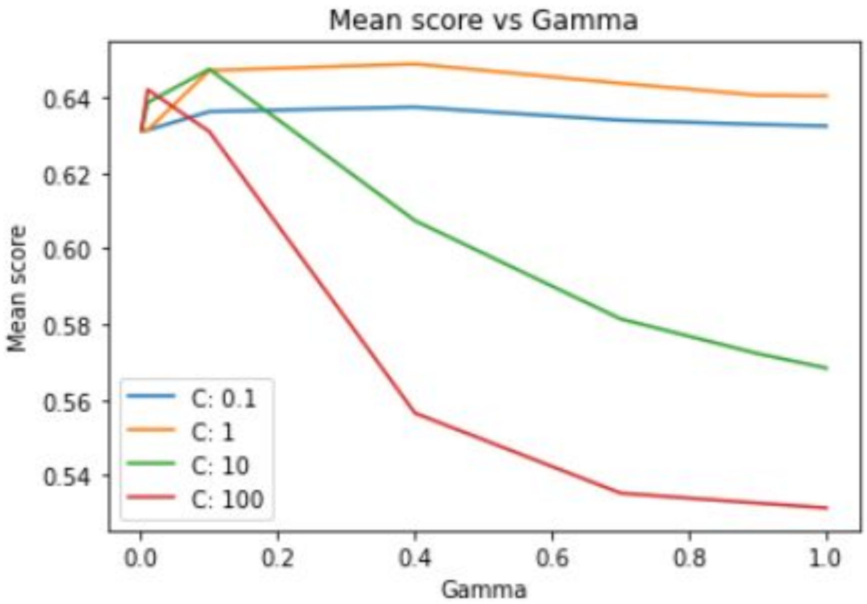


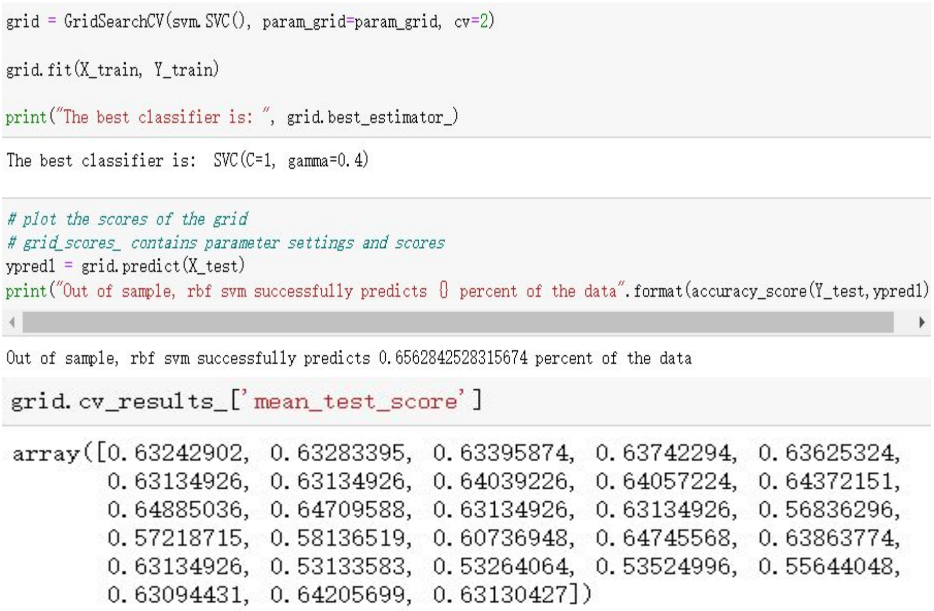
Black

- The lower part of Queens has the most serious asian harassment problem.
- Each race’s most serious harassment problem happened in different places in Queens.

PART 02

SVM





```
grid = GridSearchCV(svm.SVC(), param_grid=param_grid, cv=2)

grid.fit(X_train, Y_train)

print("The best classifier is: ", grid.best_estimator_)
```

```
The best classifier is:  SVC(C=1, gamma=0.4)
```

```
# plot the scores of the grid
# grid_scores_ contains parameter settings and scores
ypred1 = grid.predict(X_test)
print("Out of sample, rbf svm successfully predicts {} percent of the data".format(accuracy_score(Y_test,ypred1)
```

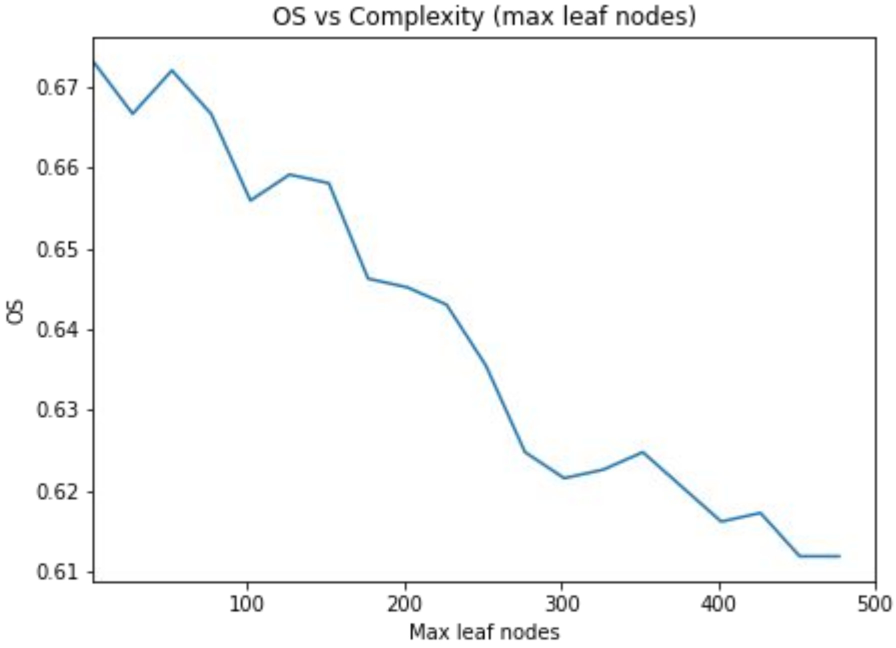
```
Out of sample, rbf svm successfully predicts 0.6562842528315674 percent of the data
```

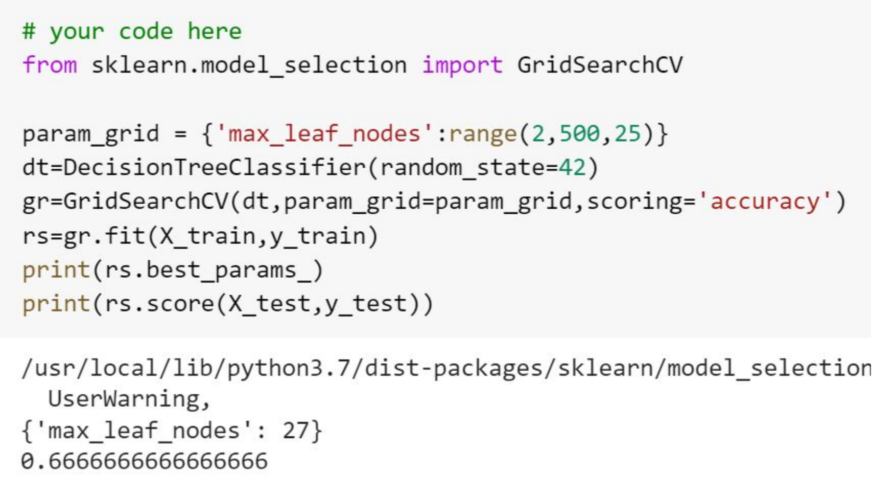
```
grid.cv_results_['mean_test_score']
```

```
array([0.63242902, 0.63283395, 0.63395874, 0.63742294, 0.63625324,
       0.63134926, 0.63134926, 0.64039226, 0.64057224, 0.64372151,
       0.64885036, 0.64709588, 0.63134926, 0.63134926, 0.56836296,
       0.57218715, 0.58136519, 0.60736948, 0.64745568, 0.63863774,
       0.63134926, 0.53133583, 0.53264064, 0.53524996, 0.55644048,
       0.63094431, 0.64205699, 0.63130427])
```

PART 02

Decision Tree



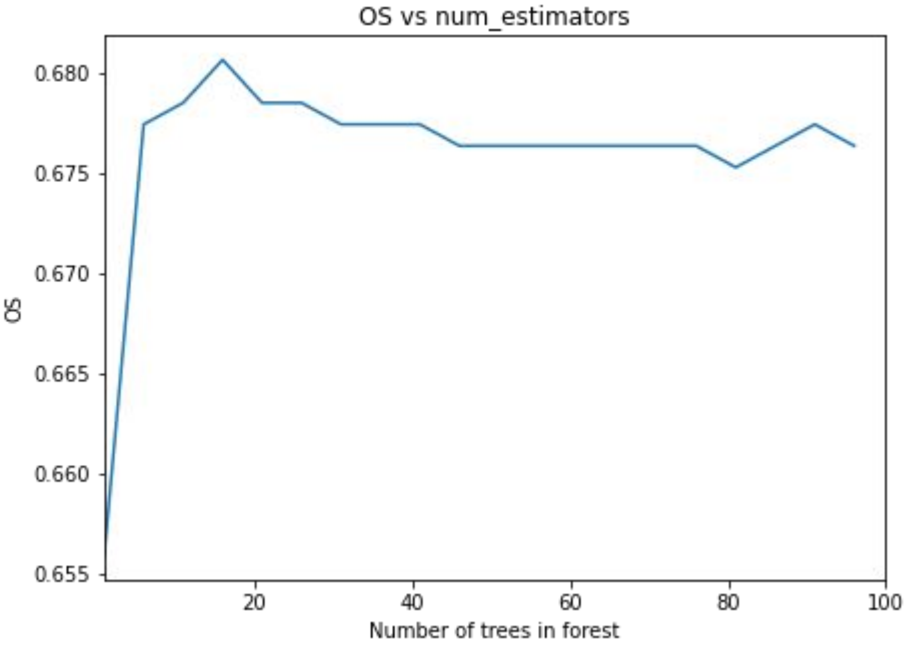


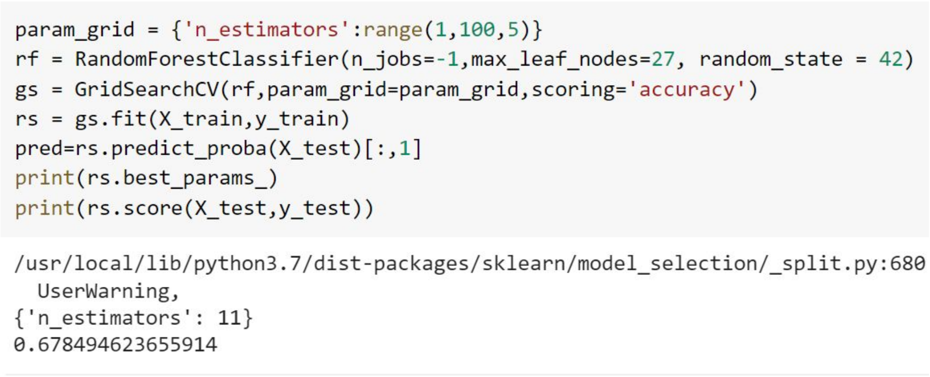
```
# your code here
from sklearn.model_selection import GridSearchCV

param_grid = {'max_leaf_nodes':range(2,500,25)}
dt=DecisionTreeClassifier(random_state=42)
gr=GridSearchCV(dt,param_grid=param_grid,scoring='accuracy')
rs=gr.fit(X_train,y_train)
print(rs.best_params_)
print(rs.score(X_test,y_test))
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection
  UserWarning,
{'max_leaf_nodes': 27}
0.6666666666666666
```

PART 02

Random Forest



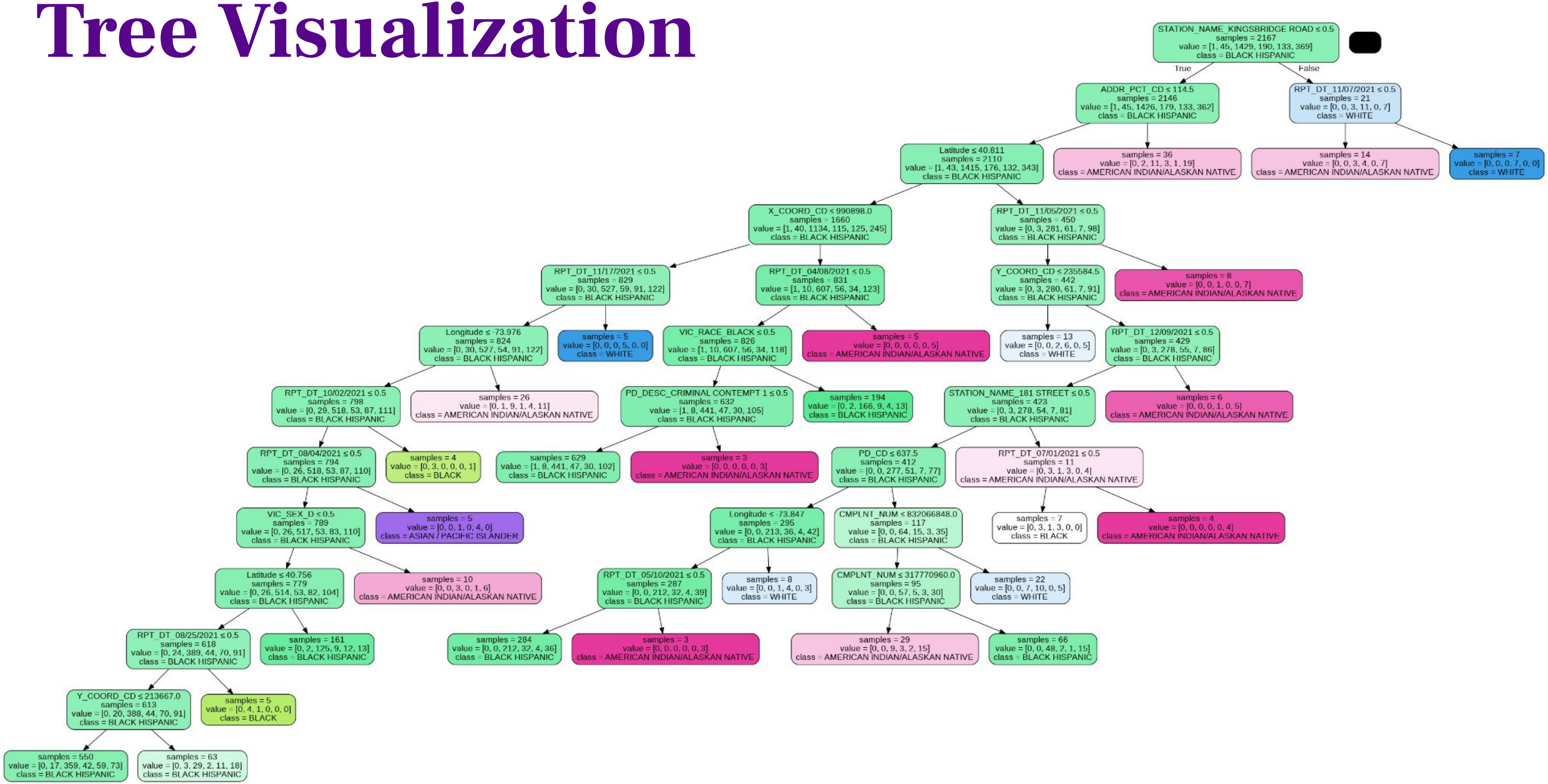


```
param_grid = {'n_estimators':range(1,100,5)}
rf = RandomForestClassifier(n_jobs=-1,max_leaf_nodes=27, random_state = 42)
gs = GridSearchCV(rf,param_grid=param_grid,scoring='accuracy')
rs = gs.fit(X_train,y_train)
pred=rs.predict_proba(X_test)[:,1]
print(rs.best_params_)
print(rs.score(X_test,y_test))
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/model_selection/_split.py:680
  UserWarning,
{'n_estimators': 11}
0.678494623655914
```

PART 02

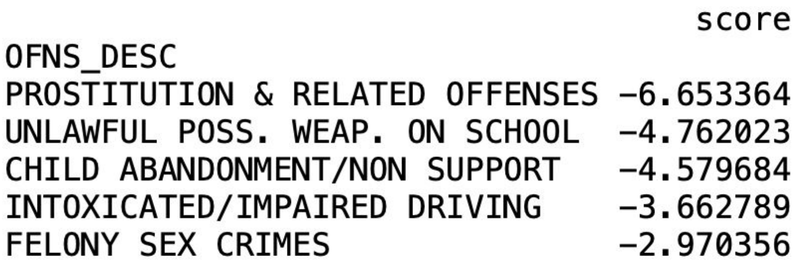
Tree Visualization



STATION_NAME_KINGSBRIDGE ROAD ≤ 0.5 samples = 2167 value = [1, 45, 1429, 190, 133, 369] class = BLACK HISPANIC
[redacted]
True
False
ADDR_PCT_CD ≤ 114.5 samples = 2146 value = [1, 45, 1426, 179, 133, 362] class = BLACK HISPANIC
RPT_DT_11/07/2021 ≤ 0.5 samples = 21 value = [0, 0, 3, 11, 0, 7] class = WHITE
Latitude ≤ 40.811 samples = 2110 value = [1, 43, 1415, 176, 132, 343] class = BLACK HISPANIC
samples = 36 value = [0, 2, 11, 3, 1, 19] class = AMERICAN INDIAN/ALASKAN NATIVE
samples = 14 value = [0, 0, 3, 4, 0, 7] class = AMERICAN INDIAN/ALASKAN NATIVE
samples = 7 value = [0, 0, 0, 7, 0, 0] class = WHITE
X_COORD_CD ≤ 990898.0 samples = 1660 value = [1, 40, 1134, 115, 125, 245] class = BLACK HISPANIC
RPT_DT_11/05/2021 ≤ 0.5 samples = 450 value = [0, 3, 281, 61, 7, 98] class = BLACK HISPANIC
RPT_DT_11/17/2021 ≤ 0.5 samples = 829 value = [0, 30, 527, 59, 91, 122] class = BLACK HISPANIC
RPT_DT_04/08/2021 ≤ 0.5 samples = 831 value = [1, 10, 607, 56, 34, 123] class = BLACK HISPANIC
Y_COORD_CD ≤ 235584.5 samples = 442 value = [0, 3, 280, 61, 7, 91] class = BLACK HISPANIC
samples = 8 value = [0, 0, 1, 0, 0, 7] class = AMERICAN INDIAN/ALASKAN NATIVE
Longitude ≤ -73.976 samples = 824 value = [0, 30, 527, 54, 91, 122] class = BLACK HISPANIC
samples = 5 value = [0, 0, 0, 5, 0, 0] class = WHITE
VIC_RACE_BLACK ≤ 0.5 samples = 826 value = [1, 10, 607, 56, 34, 118] class = BLACK HISPANIC
samples = 5 value = [0, 0, 0, 0, 0, 5] class = AMERICAN INDIAN/ALASKAN NATIVE
samples = 13 value = [0, 0, 2, 6, 0, 5] class = WHITE
RPT_DT_12/09/2021 ≤ 0.5 samples = 429 value = [0, 3, 278, 55, 7, 86] class = BLACK HISPANIC
RPT_DT_10/02/2021 ≤ 0.5 samples = 798 value = [0, 29, 518, 53, 87, 111] class = BLACK HISPANIC
samples = 26 value = [0, 1, 9, 1, 4, 11] class = AMERICAN INDIAN/ALASKAN NATIVE
PD_DESC_CRIMINAL CONTEMPT 1 ≤ 0.5 samples = 632 value = [1, 8, 441, 47, 30, 105] class = BLACK HISPANIC
samples = 194 value = [0, 2, 166, 9, 4, 13] class = BLACK HISPANIC
STATION_NAME_181 STREET ≤ 0.5 samples = 423 value = [0, 3, 278, 54, 7, 81] class = BLACK HISPANIC
samples = 6 value = [0, 0, 0, 1, 0, 5] class = AMERICAN INDIAN/ALASKAN NATIVE
RPT_DT_08/04/2021 ≤ 0.5 samples = 794 value = [0, 26, 518, 53, 87, 110] class = BLACK HISPANIC
samples = 4 value = [0, 3, 0, 0, 0, 1] class = BLACK
samples = 629 value = [1, 8, 441, 47, 30, 102] class = BLACK HISPANIC
samples = 3 value = [0, 0, 0, 0, 0, 3] class = AMERICAN INDIAN/ALASKAN NATIVE
PD_CD ≤ 637.5 samples = 412 value = [0, 0, 277, 51, 7, 77] class = BLACK HISPANIC
RPT_DT_07/01/2021 ≤ 0.5 samples = 11 value = [0, 3, 1, 3, 0, 4] class = AMERICAN INDIAN/ALASKAN NATIVE
VIC_SEX_D ≤ 0.5 samples = 789 value = [0, 26, 517, 53, 83, 110] class = BLACK HISPANIC
samples = 5 value = [0, 0, 1, 0, 4, 0] class = ASIAN / PACIFIC ISLANDER
Longitude ≤ -73.847 samples = 295 value = [0, 0, 213, 36, 4, 42] class = BLACK HISPANIC
CMPLNT_NUM ≤ 832066848.0 samples = 117 value = [0, 0, 64, 15, 3, 35] class = BLACK HISPANIC
samples = 7 value = [0, 3, 1, 3, 0, 0] class = BLACK
samples = 4 value = [0, 0, 0, 0, 0, 4] class = AMERICAN INDIAN/ALASKAN NATIVE
Latitude ≤ 40.756 samples = 779 value = [0, 26, 514, 53, 82, 104] class = BLACK HISPANIC
samples = 10 value = [0, 0, 3, 0, 1, 6] class = AMERICAN INDIAN/ALASKAN NATIVE
RPT_DT_05/10/2021 ≤ 0.5 samples = 287 value = [0, 0, 212, 32, 4, 39] class = BLACK HISPANIC
samples = 8 value = [0, 0, 1, 4, 0, 3] class = WHITE
CMPLNT_NUM ≤ 317770960.0 samples = 95 value = [0, 0, 57, 5, 3, 30] class = BLACK HISPANIC
samples = 22 value = [0, 0, 7, 10, 0, 5] class = WHITE
RPT_DT_08/25/2021 ≤ 0.5 samples = 618 value = [0, 24, 389, 44, 70, 91] class = BLACK HISPANIC
samples = 161 value = [0, 2, 125, 9, 12, 13] class = BLACK HISPANIC
samples = 284 value = [0, 0, 212, 32, 4, 36] class = BLACK HISPANIC
samples = 3 value = [0, 0, 0, 0, 0, 3] class = AMERICAN INDIAN/ALASKAN NATIVE
samples = 29 value = [0, 0, 9, 3, 2, 15] class = AMERICAN INDIAN/ALASKAN NATIVE
samples = 66 value = [0, 0, 48, 2, 1, 15] class = BLACK HISPANIC
Y_COORD_CD ≤ 213667.0 samples = 613 value = [0, 20, 388, 44, 70, 91] class = BLACK HISPANIC
samples = 5 value = [0, 4, 1, 0, 0, 0] class = BLACK
samples = 550 value = [0, 17, 359, 42, 59, 73] class = BLACK HISPANIC
samples = 63 value = [0, 3, 29, 2, 11, 18] class = BLACK HISPANIC

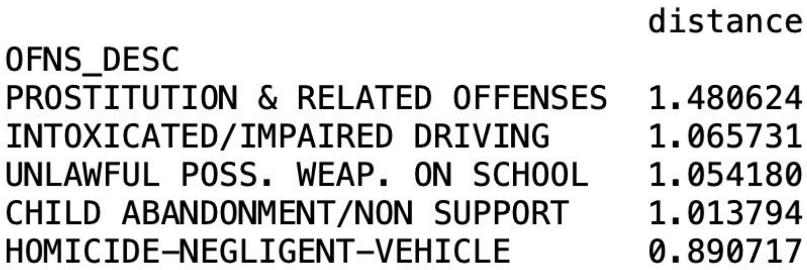
Anomaly Detection

Result of Gaussian Mixture:



```
                                  score
OFNS_DESC
PROSTITUTION & RELATED OFFENSES -6.653364
UNLAWFUL POSS. WEAP. ON SCHOOL  -4.762023
CHILD ABANDONMENT/NON SUPPORT   -4.579684
INTOXICATED/IMPAIRED DRIVING    -3.662789
FELONY SEX CRIMES               -2.970356
```

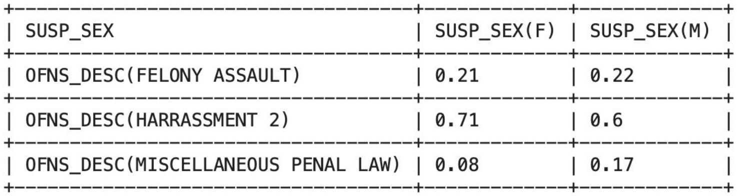
Result of K-Means:



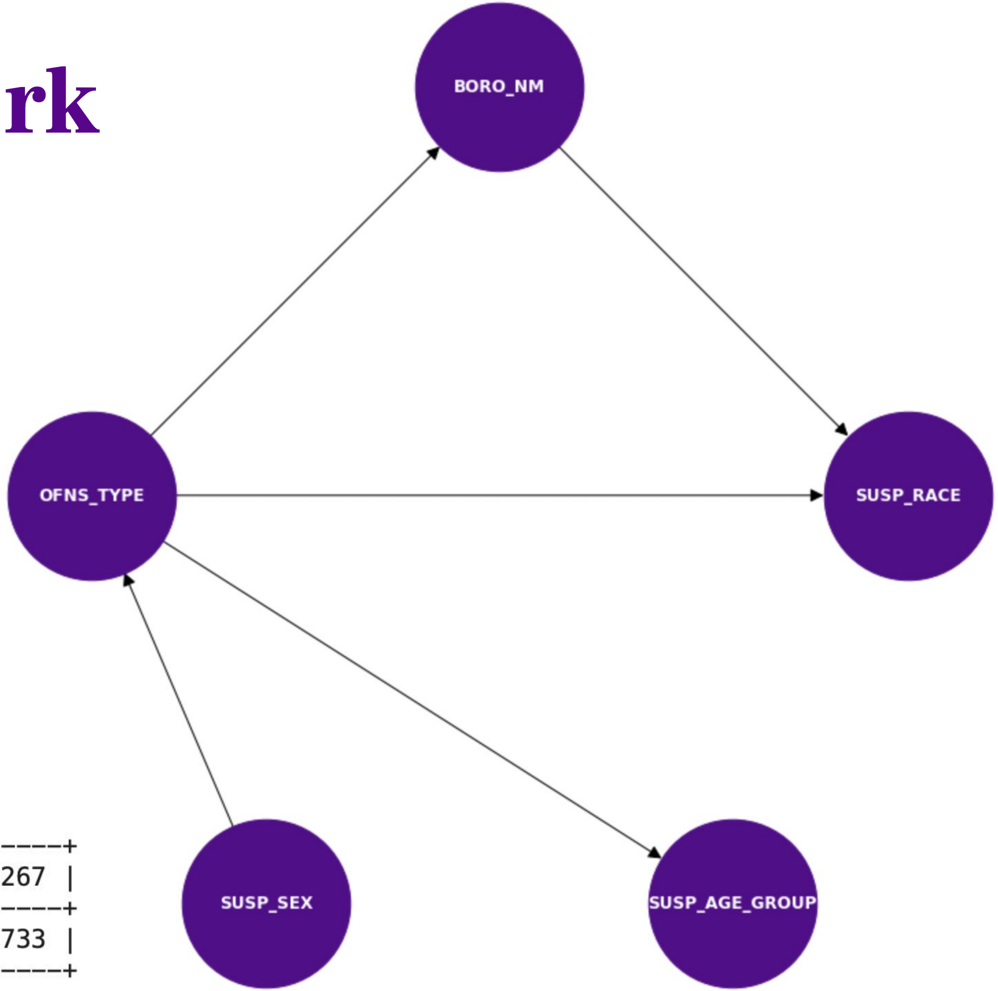
```
                                 distance
OFNS_DESC
PROSTITUTION & RELATED OFFENSES  1.480624
INTOXICATED/IMPAIRED DRIVING     1.065731
UNLAWFUL POSS. WEAP. ON SCHOOL   1.054180
CHILD ABANDONMENT/NON SUPPORT    1.013794
HOMICIDE-NEGLIGENT-VEHICLE       0.890717
```

PART 02

Bayesian Network



```
+-----------------------------------+-------------+-------------+
| SUSP_SEX                          | SUSP_SEX(F) | SUSP_SEX(M) |
+-----------------------------------+-------------+-------------+
| OFNS_DESC(FELONY ASSAULT)         | 0.21        | 0.22        |
+-----------------------------------+-------------+-------------+
| OFNS_DESC(HARRASSMENT 2)          | 0.71        | 0.6         |
+-----------------------------------+-------------+-------------+
| OFNS_DESC(MISCELLANEOUS PENAL LAW) | 0.08        | 0.17        |
+-----------------------------------+-------------+-------------+
```



```
CPD of SUSP_SEX:
+-------------+---------+
| SUSP_SEX(F) | 0.27267 |
+-------------+---------+
| SUSP_SEX(M) | 0.72733 |
+-------------+---------+
```

Conclusions

- Victims: Asians, American Indians are similar, Hispanics and Blacks are similar when they are victims

- SVM and Random Forest: decent accuracy could be improved by tuning more precisely

- Servere harassment is most popular for all races. Crime locations are discrete for all races

- Offense types affect community's formation