Data Science Project

The Wild Blazers
Team Members: Anish Bhardwaj

The Topic

- News reports indicate that police officers are biased towards certain racial groups
- Based on information in the dataset, and visualizations based on this data, this claim seems to hold substance
- Can we use the information in the CPD database to determine an officers risk potential and recommend them for sensitivity training?
- What other factors can we see that groups officers with allegations together?

Methodology and Data

- Took data from all 25 districts of Chicago
- Took Demographic Data of Civilians and Police Officers
  - Race
  - Gender
  - Allegation Rates
- Did some cool Data Science Stuff!

Tools Used

Police Diversity and Misconduct

Investigating with Data Science

Checkpoint 1 - Relational Analytics

Questions Being Asked

- What is the Complaint Rate for police officers in the different districts of Chicago?
- What is the racial distribution of officers in the districts of Chicago?
- What is the gender distribution of officers in the districts of Chicago?
- What is the racial distribution of citizens in the districts of Chicago?
- Combine the demographic distribution of a district correlate with the Officer Complaint Rate into one table.
- Combine the racial distribution of officers in a district alongside the racial distribution of the citizens of a district.

Question 1

- What is the Complaint Rate for police officers in the different districts of Chicago?
- Was able to query this information through the help of SQL and Python
- District Mapping needed to also be queried

```sql
SELECT data_policeunit.id - 1 AS district, count(*) AS police_per_district
FROM data_policeunit JOIN data_officer officer ON data_policeunit.id = officer.last_unit_id
JOIN data_officerallegation d on officer.id = d.officer_id
WHERE data_policeunit.description in ('District 001', 'District 002', 'District 003', 'District 004',
'District 005', 'District 006', 'District 007', 'District 008', 'District 009', 'District 010', 'District 011',
'District 012', 'District 013', 'District 014', 'District 015', 'District 016', 'District 017', 'District 018',
'District 019', 'District 020', 'District 021', 'District 022', 'District 023', 'District 024', 'District 025')
GROUP BY data_policeunit.id ORDER BY district ASC;
```

```sql
SELECT data_allegation_areas.area_id  as district, drp.area_population AS district_population, COUNT(*) AS allegation_count,
       round(COUNT(*)*1000/drp.area_population, 2) as allegationspercapita
FROM data_allegation_areas LEFT JOIN data_area area ON data_allegation_areas.area_id = area.id
JOIN (SELECT area_id, SUM(count) AS area_population FROM data_racepopulation group by area_id) drp
ON drp.area_id = area.id
WHERE area.area_type = 'police-districts' and drp.area_population is not null
GROUP BY data_allegation_areas.area_id, drp.area_population ORDER BY district ASC;
```

Question 2

- What is the racial distribution of officers in the districts of Chicago?
- BIG QUERY

```
SELECT data_policeunit.id - 1 AS district,
       count(*) AS police_per_district,
       count(*) filter (WHERE race = 'Black')  AS Black,
       count(*) filter (WHERE race = 'White')  AS White,
       count(*) filter (WHERE race = 'Hispanic')  AS Hispanic,
       count(*) filter (WHERE race = 'Asian/Pacific')  AS AsianPacificIslander,
       count(*) filter (WHERE race = 'Native American/Alaskan Native')  AS Native,
--         count(*) filter (WHERE race = 'Other/Unknown')  AS Other,
       round(count(*) filter (WHERE race = 'Black')*100.0/count(*), 2)  AS Blackpercent,
       round(count(*) filter (WHERE race = 'White')*100.0/count(*), 2)  AS Whitepercent,
       round(count(*) filter (WHERE race = 'Hispanic')*100.0/count(*), 2)  AS Hispanicpercent,
       round(count(*) filter (WHERE race = 'Asian/Pacific')*100.0/count(*), 2)  AS AsianPacificpercent,
       round(count(*) filter (WHERE race = 'Native American/Alaskan Native')*100.0/count(*), 2)  AS Nativepercent,
       100 - (round(count(*) filter (WHERE race = 'Black')*100.0/count(*), 2) +
              round(count(*) filter (WHERE race = 'White')*100.0/count(*), 2) +
              round(count(*) filter (WHERE race = 'Hispanic')*100.0/count(*), 2) +
              round(count(*) filter (WHERE race = 'Asian/Pacific')*100.0/count(*), 2) +
              round(count(*) filter (WHERE race = 'Native American/Alaskan Native')*100.0/count(*), 2)) AS Other
FROM data_policeunit JOIN data_officer officer ON data_policeunit.id = officer.last_unit_id
JOIN data_officerallegation d on officer.id = d.officer_id
WHERE data_policeunit.description in ('District 001', 'District 002', 'District 003', 'District 004',
'District 005', 'District 006', 'District 007', 'District 008', 'District 009', 'District 010', 'District 011',
'District 012', 'District 013', 'District 014', 'District 015', 'District 016', 'District 017', 'District 018',
'District 019', 'District 020', 'District 021', 'District 022', 'District 023', 'District 024', 'District 025')
GROUP BY data_policeunit.id ORDER BY district ASC;
```

Question 3

- What is the gender distribution of officers in the districts of Chicago?

```sql
SELECT data_policeunit.id - 1 AS district,
       count(*) filter (WHERE gender = 'F')  AS Femalepolice,
       count(*) filter (WHERE gender = 'M')  AS Malepolice,
       round(count(*) filter (WHERE gender = 'F')*100.0/count(*), 2)  AS Fpolicepercent,
       round(count(*) filter (WHERE gender = 'M')*100.0/count(*), 2)  AS Mpolicepercent
--        100 - (round(count(*) filter (WHERE gender = 'F')*100.0/count(*), 2) +
--               round(count(*) filter (WHERE gender = 'M')*100.0/count(*), 2)) AS Other
FROM data_policeunit JOIN data_officer officer ON data_policeunit.id = officer.last_unit_id
JOIN data_officerallegation d on officer.id = d.officer_id
WHERE data_policeunit.description in ('District 001', 'District 002', 'District 003', 'District 004',
'District 005', 'District 006', 'District 007', 'District 008', 'District 009', 'District 010', 'District 011',
'District 012', 'District 013', 'District 014', 'District 015', 'District 016', 'District 017', 'District 018',
'District 019', 'District 020', 'District 021', 'District 022', 'District 023', 'District 024', 'District 025')
GROUP BY data_policeunit.id ORDER BY district ASC;
```

Question 4

- What is the racial distribution of citizens in the districts of Chicago?

```sql
SELECT area_id AS district,
SUM(count) filter (WHERE race = 'Black') AS Blackpop,
SUM(count) filter (WHERE race = 'White')  AS Whitepop,
SUM(count) filter (WHERE race = 'Hispanic')  AS Hispanicpop,
SUM(count) filter (WHERE race = 'Asian/Pacific Islander')  AS AsianPacificIslanderpop,
SUM(count) filter (WHERE race = 'Native American/Alaskan Native')  AS Nativepop,
SUM(count) filter (WHERE race = 'Other/Unknown')  AS Otherpop,
SUM(count) filter (WHERE race = 'Black')*100.0 / (SUM(count))  AS Blackpoppercent,
SUM(count) filter (WHERE race = 'White')*100.0  / (SUM(count))  AS Whitepoppercent,
SUM(count) filter (WHERE race = 'Hispanic') *100.0 / (SUM(count)) AS Hispanicpoppercent,
SUM(count) filter (WHERE race = 'Asian/Pacific Islander')*100.0  / SUM(count) AS AsianPacificIslanderpoppercent,
SUM(count) filter (WHERE race = 'Native American/Alaskan Native')*100.0  / SUM(count)  AS NativeAmericanpoppercent,
SUM(count) filter (WHERE race = 'Other/Unknown')*100.0  / SUM(count)  AS Otherpoppercent
FROM data_racepopulation JOIN data_area area on data_racepopulation.area_id = area.id
WHERE area.area_type = 'police-districts'
GROUP BY area_id ORDER BY district ASC;
```

Question 5 and 6

- Combine the demographic distribution of a district with the Officer Complaint Rate into one table.
- Combine the racial distribution of officers in a district alongside the racial distribution of the citizens of a district?
- This question was important for the remaining parts of the project
- Used the data collected from the above statements (and modifications of said statements) to create a dataset that could be used further with the help of python
- Linked Here!

Insights/ Learnings

- Data Querying is not easy!
- Takes a lot of effort but can be done in an optimal manner
- The data I got allowed to preemptively see that the Gender Bias of officers was greatly skewed towards males
- The data also seemed to show that districts police racial demographics were not always representative of the population
- The first layer of the onion has been peeled

Checkpoint 2 - Visualizations

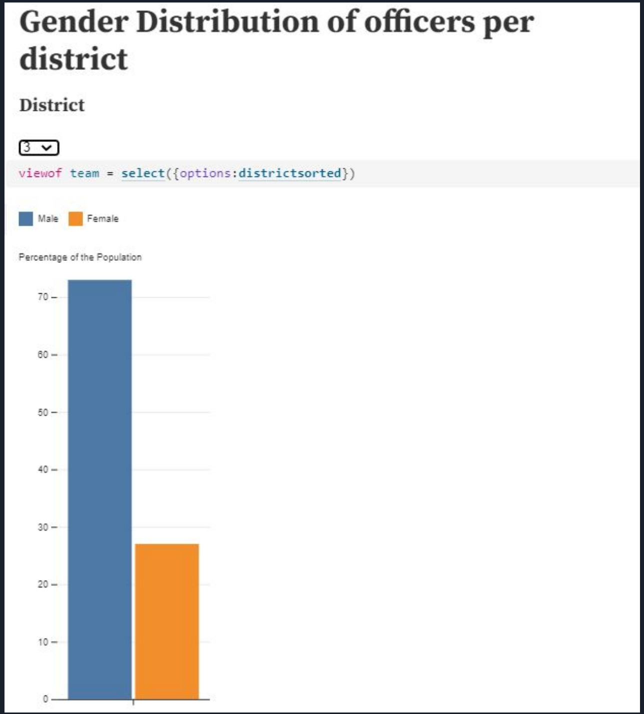
The Visualizations being made

- Gender Distribution of Police Officers per District
- Population Race Distribution vs Police Race Distribution per district
- Population to Police Officer Ratio and Allegation Rate per police officer per district (Combined Interactive Visualisation)

Visualization 1

- Gender Distribution of Police Officers per District
- Why?
  - To visualize the gender distribution more clearly
  - Helped me understand d3 better
- Can choose district to see the distribution
- Linked Here!

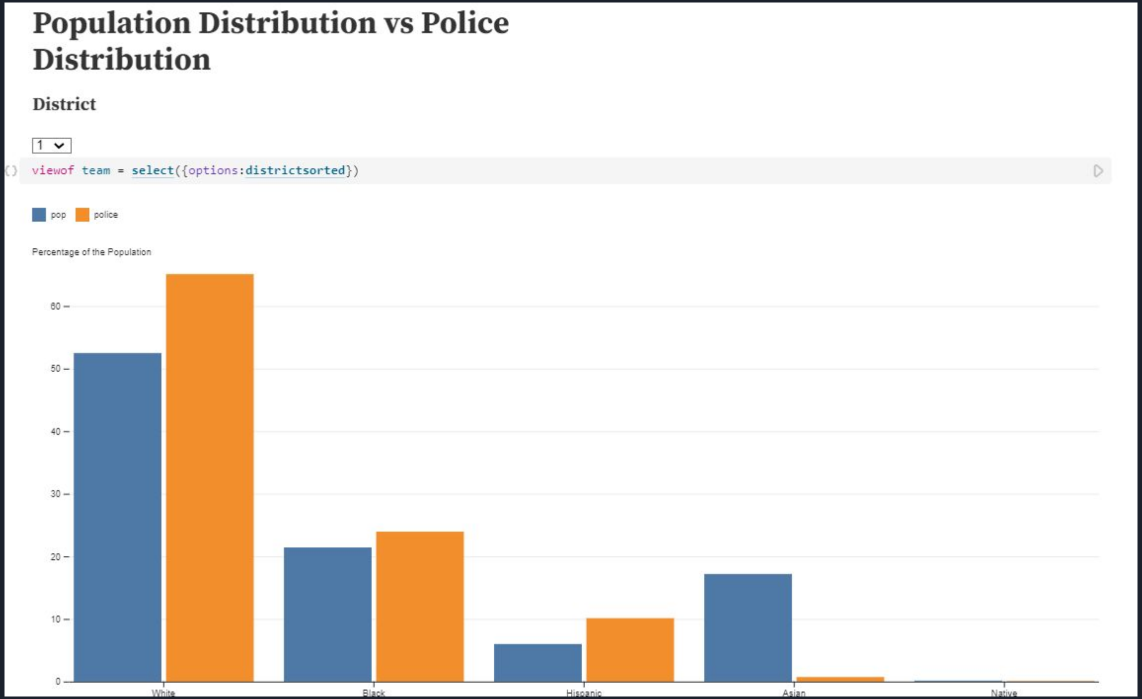
Visualization 1



Visualization 2

- Population Race Distribution vs Police Race Distribution per district
- Why?
  - To see whether the population race distribution matches the police race distribution
  - Use insights from this and the next visualization to confirm a potential scope of bias
- Example: District 11 and District 20
- Can choose district to see the distribution
- Linked Here!

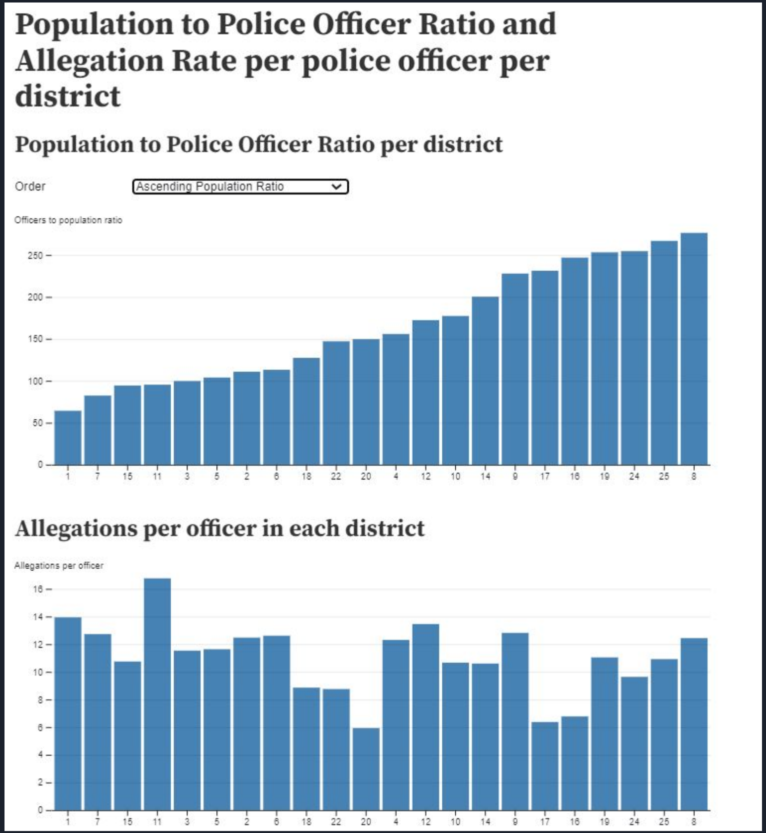
Visualization 2



Visualization 2

- Population to Police Officer Ratio and Allegation Rate per police officer per district
- Why?
  - See the allegation rate per district
  - Confirm whether or not allegations increase if the Police Officer Ratio increases
  - Combine with insights from previous visualization
- Can sort in different ways (Police Officer Ratio, District wise, Allegations per officer per district)
- Linked Here!

Visualization 2



Insights/ Learnings

- D3 is powerful
- The male officers outnumber female officers in a nearly 4 to 1 disparity
- No clear correlation between the ratio of police officers to the citizens in the district and the Allegation rate per police officer in the district
    - GOOD…Means that the onion peeling is working and we are moving towards a goal!
- There seems to be a potential for racial bias evident by the visualizations

Checkpoint 4 - Machine Learning

Questions Being Asked

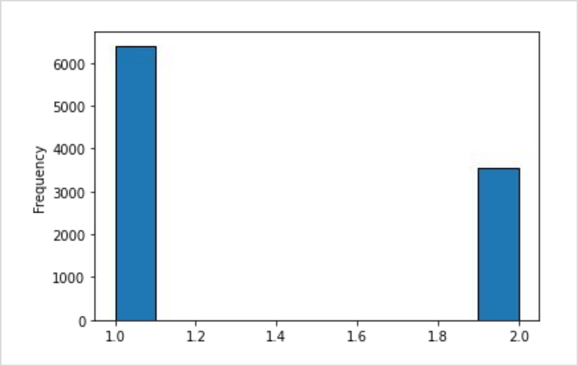
- Can we model the risk potential of an officer based on their demographic data? (Supervised Learning)
    - Can we do this effectively enough to support having such a model determine whether an officer needs sensitivity training?
- Are there specific features that can be used to determine clusters of officers with allegations?

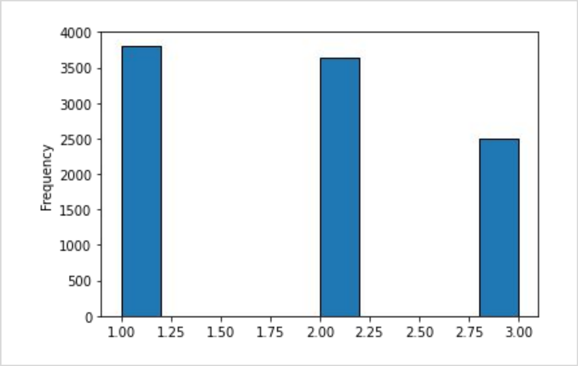
Supervised Learning

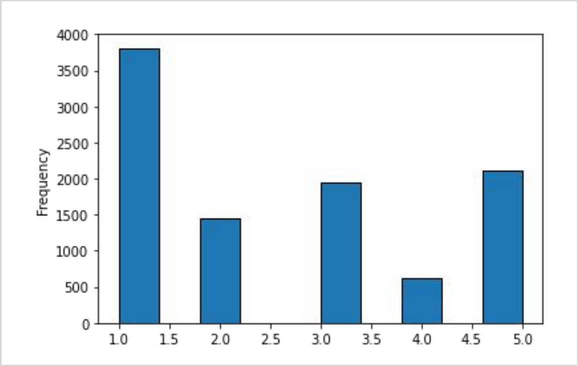
- Can we model the risk potential of an officer based on their demographic data?
- Use officer data, allegation statistics, and district data to create bands
  - Binary Classification Band
    - 0 = Low Risk, 1 = High Risk
  - 3-Class Band
    - 0 = Low Risk, 1 = Medium Risk, 2 = High Risk
  - 5-Class Band
    - 0 = Very Low Risk, 1 = Low Risk, 2 = Medium Risk, 3 = High Risk, 4 = Very High Risk

How Bands were Determined

- Combination of Allegation data per district with Mean and Standard Deviation Values
- Done to ensure a decent number of individuals in each band







Models Used

- K-Nearest Neighbours
- Decision Tree Classifier
- Logistic Regression
- All were tested using the K-Fold Approach (K == 5)

Best Performing Models

- Binary Classification
  - Best Model - 95.4% Accuracy with Logistic Regression
- 3-Class Classification
  - Best Model - 90.5% Accuracy with Logistic Regression
- 5-Class Classification
  - Best Model - 81.4% Accuracy with Decision Tree

What can we infer from this

- We can use the allegation rate of officers combined with the standard deviation and mean to create bands for risk amongst officers
- We can train models that allow us to determine an officers level of risk quite well even with 5 bands (playing around with model parameters further and maybe using neural networks could help the accuracy even higher)
    - The models are very accurate for 2 and 3 bands of risk
- In Production, such a model could be deployed and individuals in the highest band of risk could be given sensitivity training following which their allegation rate is closely monitored by a team

Unsupervised Learning

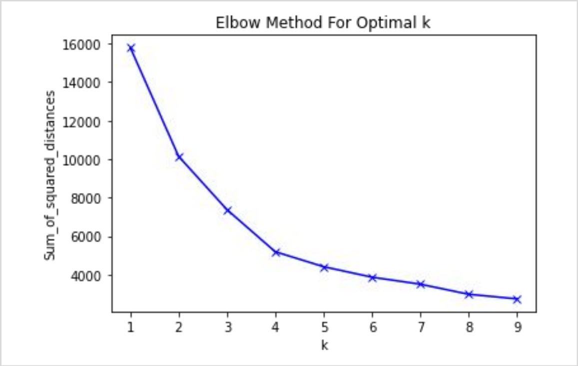
- 2 Variations of the Dataset
  - district, gender, race, major_award_count, allegation_count, sustained_count, unsustained_count, birth_year, black, white, hispanic, blackpop, whitepop, hispanicpop
    - Black, white, hispanic are one_hot_encoded values of which population is the majority for the police demographics in each district
    - Blackpop, whitepop, hispanicpop are one_hot_encoded values of which population is the majority for the civilian demographics in each district
    - The other values (Asian/Pacific Islander, Native American) are not here in these as they are never the majority population for either the police or population race demographics
  - district, gender, race, major_award_count, allegation_count, sustained_count, black, white, hispanic, blackpop, whitepop, hispanicpop
    - Removed birth_year and unsustained_count as they skewed the data heavily

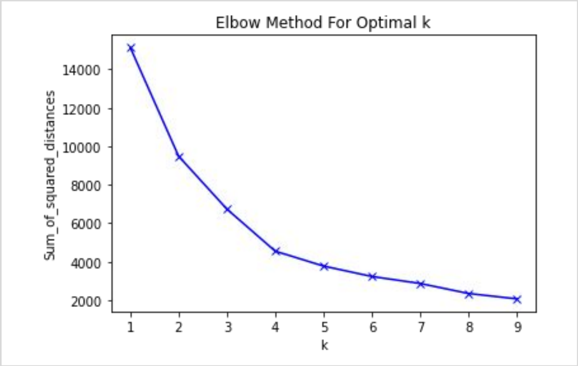
Methodology

- Performed K-Means Clustering
- Performed Elbow Method analysis
- Determined most important features using KMeans-Feature-Importance wrapper Class
- Divided into optimal clusters
- Graphed the 3 most important features as per each individual cluster
- Determined whether similarities or differences were there in the clusters

Elbow Method

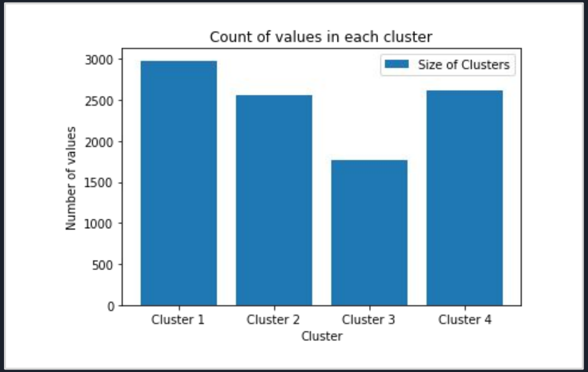
- From Left to Right, Each of the different Elbow Method graphs for each different data set (1, 2)
- Both had an optimal cluster number of 4

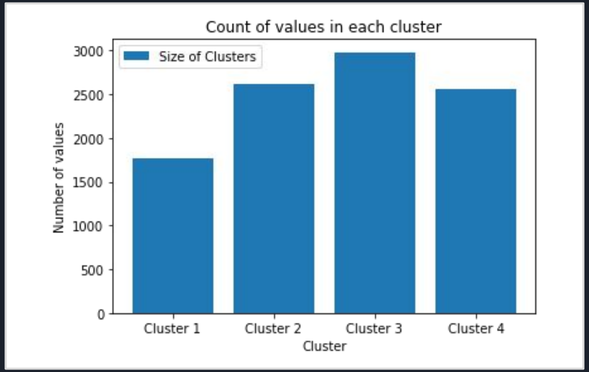




Cluster Sizes

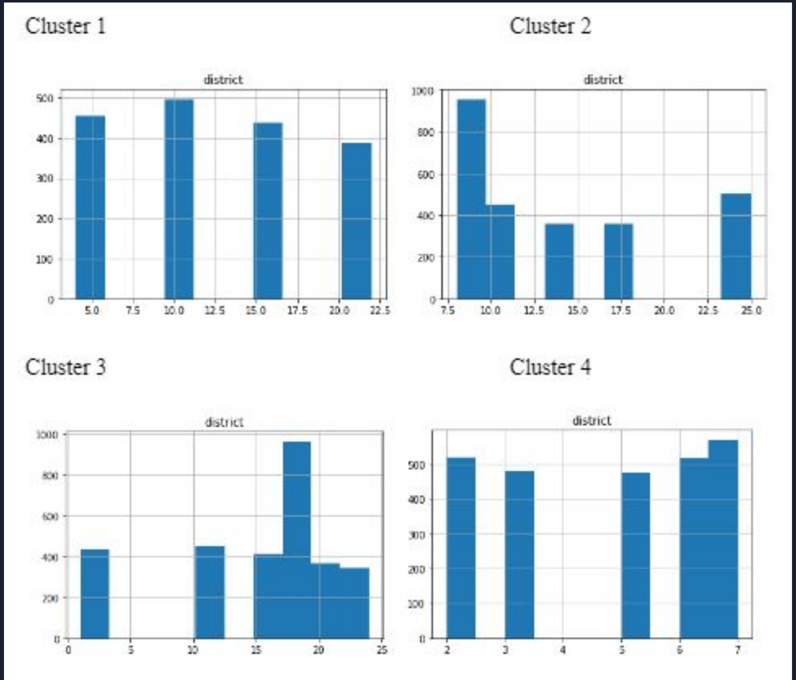
- From Left to Right, Each of the different Elbow Method graphs for each different data set (1,2,3)
- All had decently distributed Clusters





Sample Graphs across Clusters

- Calculated Feature Importances using KMeans-Feature-Importance wrapper library on sklearn
- Plotted 3 most important features for each dataset created



Insights

- Allegation Count, District, and Race of the officer were deemed important in the second dataset
- Birth_year and Unsustained_count were also important from the first dataset
- Showcases potential for some mapping across these various data variables
- I was able to find 2 interesting links
    - The race distributions in the second dataset had some major differences
    - The districts were all different and unique to each cluster in the second dataset clusters

```
[124] data0['district'].value_counts()

11.0    496
4.0     454
15.0    438
22.0    387
Name: district, dtype: int64
```

```
data1['district'].value_counts()

25.0    501
8.0     497
9.0     458
10.0    452
14.0    358
17.0    357
Name: district, dtype: int64
```

```
[126] data2['district'].value_counts()

19.0    485
18.0    480
12.0    453
1.0     435
16.0    414
20.0    369
24.0    346
Name: district, dtype: int64
```

```
[127] data3['district'].value_counts()

7.0    570
2.0    519
6.0    516
3.0    482
5.0    476
Name: district, dtype: int64
```

Future Work

Future Work

- Playing even further with the ML models based on risk and incorporating data that work in specific with violent interactions to better sensitize such officers
- Continuing to work on clustering to see whether some connection can be made across the aforementioned promising data variables
- Working with more variables and hyperparameters across the current created models mentioned here to improve them even further

Thank You!