The Wild Blazers Final Project Report

Team Members: Anish Bhardwaj

Project Introduction

In the past few years, we have seen racial issues evolve and adapt making issues that exist with law enforcement even more apparent with society. There has been an increasing lack of trust between civilians and police departments as a result of unfortunate events that have occurred in society. Simply taking a look at the different articles that surface on the daily gives substance to the idea that there is a lack of diversity in the police department and that there is inherent bias amongst the way officers treat individuals of different races.

To reinforce trust in law enforcement and to understand these biases better, we can leverage the power of data science. In this project, I take a look at Police and Civilian Demographics and try to understand whether there is evidence of bias amongst police officers in the Chicago Police Department. The objective of this project is to see whether it is possible to create a system based on both officer and victim information that allows us to recommend officers for sensitivity training and to provide a Proof of Concept for such a system. Having such systems in place would help improve trust in law enforcement and help them also better serve our communities. The project leverages different areas of data science ranging from sourcing data to visualizations and even Machine Learning models to craft such a system and understand along the way whether the claims and thoughts of the public hold merit.

To do so, I started off by initially exploring the Chicago Police Database which helped source data regarding factors like complaint rates, officer racial and gender distribution, district racial distribution and more. This information, once sourced, is further analyzed through interactive visualizations that helped see and understand the problem further. Finally, I use the collected information to create thresholds for multiple classification models and test different models to see which can truly give us a system to better understand and identify risk levels amongst officers. I also attempt to take a look at Unsupervised learning models over the sourced data to see whether there are certain features that help us group officers together. The results for this are seen below across all districts with relevant data available in Chicago.

Relational Analytics

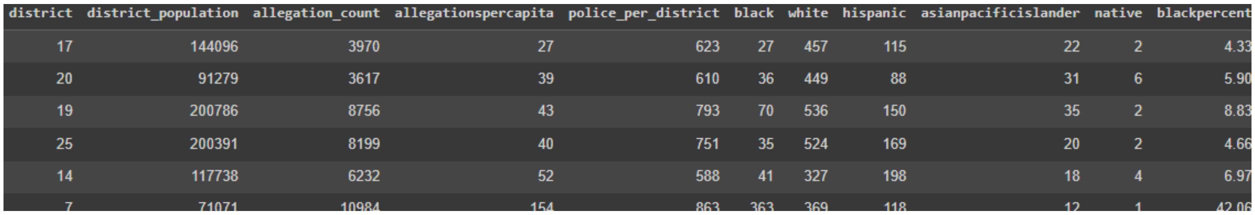
Questions Being Asked

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

Were we able to answer these questions?

Question 1: What is the Complaint Rate for police officers in the different districts of Chicago?

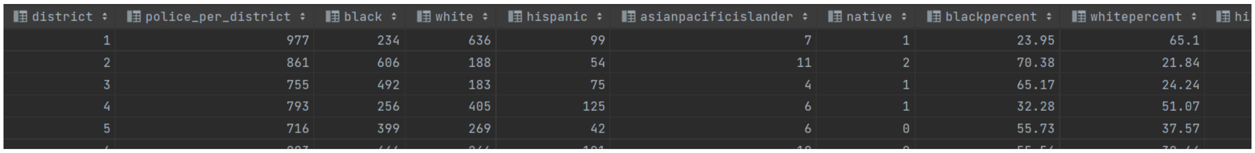
The purpose of this question is to see whether we can determine if there is truly a difference in the allegations per police officer in a district and whether this actually differs from district to district. Since our ultimate goal is to observe whether a bias is prevalent and provide suggestions for sensitivity training, this question is extremely important as we need to understand what the allegation rates are per district. To do this, I first queried the Police Officer numbers per district in the City of Chicago. I did so for each district and made sure that the district mapping was correct. I then queried the Population and Allegation data per district in the city of Chicago. The allegations and the police per district had now been queried but there was a problem with the identifier. This is because the district values are not mapped one to one and rather mapped in a bizarre method. I created another table thus of the district mappings to the district numbers which allowed me to create a dictionary which in turn allowed me to remap the values and allowed me to create the CSV file shown below. To fix the district values, I employed the help of python wherein I created a dictionary and then merged the data once I had mapped the values correctly. The final CSV can be seen here and looks like the image below (there are more features in the final combined CSV). I also normalized the allegations per capita in a district to see the distribution a little more clearly



district	district_population	allegation_count	allegationspercapita	police_per_district	black	white	hispanic	asianpacificislander	native	blackpercent
17	144096	3970	27	623	27	457	115	22	2	4.33
20	91279	3617	39	610	36	449	88	31	6	5.90
19	200786	8756	43	793	70	536	150	35	2	8.83
25	200391	8199	40	751	35	524	169	20	2	4.66
14	117738	6232	52	588	41	327	198	18	4	6.97
7	71071	10984	154	863	363	369	118	12	1	42.06

Question 2: What is the racial distribution of officers in the districts of Chicago?

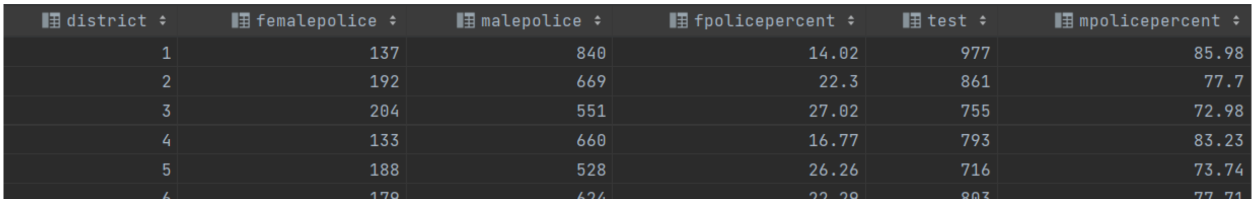
Since we are looking at the bias of the officers, I believe it is important to understand the racial composition of the districts of Chicago as well. This is because it is one thing to say that the police officers are biased or single out certain behavior based on demographic data but it is another to see this in action. Getting this data would allow me to compare and contrast with visualizations the officer racial distribution and the racial distribution of the communities these officers serve. To do this, I queried the total police officers per district and then the different racial groups (Black, White, Asian/Pacific Islander, Hispanic, and Native American/Alaskan Native) and then calculated the percentage composition of each race per district. I made sure to match officers specifically to their districts by using the description of the unit and tying that to the district id. This resulted in the creation of a table which had the numbers and the percentages per racial group. Through this, I was able to see that there was a large change between the officer distribution across districts and felt that this would give us some interesting correlations going forward



district	police_per_district	black	white	hispanic	asianpacificislander	native	blackpercent	whitepercent	hi
1	977	234	636	99	7	1	23.95	65.1	
2	861	606	188	54	11	2	70.38	21.84	
3	755	492	183	75	4	1	65.17	24.24	
4	793	256	405	125	6	1	32.28	51.07	
5	716	399	269	42	6	0	55.73	37.57	

Question 3: What is the gender distribution of officers in the districts of Chicago?

In the same way that race was analyzed, I was curious about how the gender distribution of police officers was across districts. The reason for analyzing this was that in many articles, we hear about white male police brutality and I wanted to see what the distribution of officers looked like per gender per district across the city of Chicago. To do this, I queried the data in a similar manner as to what I had done for the racial distribution except now I was looking at gender. Again, I made sure to divide the officers correctly by district and calculated the percentage of the officers in each gender bracket. This resulted in a table like the one that can be seen below. This data clearly showcased that while there are definitely some districts that are more male dominated than others, All districts had a larger percentage of males than females as police officers. had initially thought that gender could be an interesting attribute for prediction of the risk of a candidate but the extreme skew toward male police officers seems to invalidate that idea. I'll still explore this in the future to see whether further insights can be seen with this

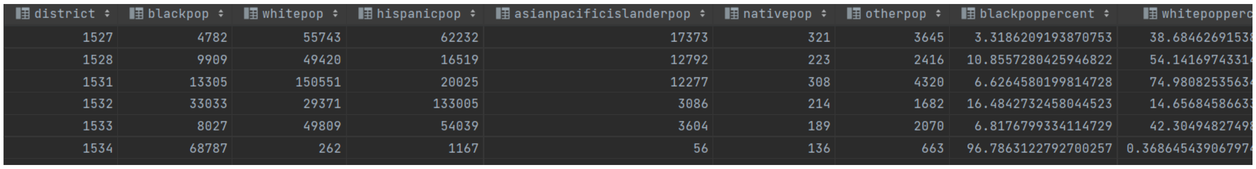


district	femalepolice	malepolice	fpolicepercent	test	mpolicepercent
1	137	840	14.02	977	85.98
2	192	669	22.3	861	77.7
3	204	551	27.02	755	72.98
4	133	660	16.77	793	83.23
5	188	528	26.26	716	73.74
6	179	624	22.29	803	77.71

Question 4: What is the racial distribution of citizens in the districts of Chicago?

Now we've answered the questions regarding the racial and gender distribution of the officers in the city of Chicago but one very important thing that we need to look at is the distribution of the citizens in the different districts. The allegation rate is very different across the different districts but to truly see whether a bias exists, we need to be able to study the racial demographics of the population that resides in a district.

To do so, I queried the Population statistics across each district and then calculated the percentage of the total population per race as well. This resulted in a table as shown below:



district	blackpop	whitepop	hispanicpop	asianpacificislanderpop	nativepop	otherpop	blackpoppercent	whitepopperc
1527	4782	55743	62232	17373	321	3645	3.3186209193870753	38.68462691538
1528	9909	49420	16519	12792	223	2416	10.8557280425946822	54.14169743314
1531	13305	150551	20025	12277	308	4320	6.6264580199814728	74.98082535634
1532	33033	29371	133005	3086	214	1682	16.4842732458044523	14.65684586633
1533	8027	49809	54039	3604	189	2070	6.8176799334114729	42.30494827498
1534	68787	262	1167	56	136	663	96.7863122792700257	0.368645439067974

Where we can see the districts and the population numbers. I fixed the district values later on using the same dictionary mentioned before.

Task 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.

Both of these tasks are extremely important for the visualizations that were planned as well as the machine learning task. While we were able to query all the data in the previous steps well, the issue with combining all the data into one table was that the districts didn't match since the numbering was skewed (As can be seen with the last SQL Query). Thus, I queried all the data, downloaded the CSVs and created a dictionary in python to map the district values (1527 onwards) to the correct district values. This allowed me to merge the different tables into one combined table which allowed me to then move forward with the visualizations.

The CSVs can be found here:
https://drive.google.com/drive/folders/1H2zUJB3dZlQk_iQCvfE9rqaorApjSH8g?usp=share_link

The code used to create the combined CSV can be found here:
https://colab.research.google.com/drive/18NLG5uzqU9E2IOALMT805pcCbMoLVZz7#scrollTo=GNn-yx3SxowX

The original questions these were based on ended up being more clearly answered in the Visualization section.

Final Thoughts on Relational Analytics

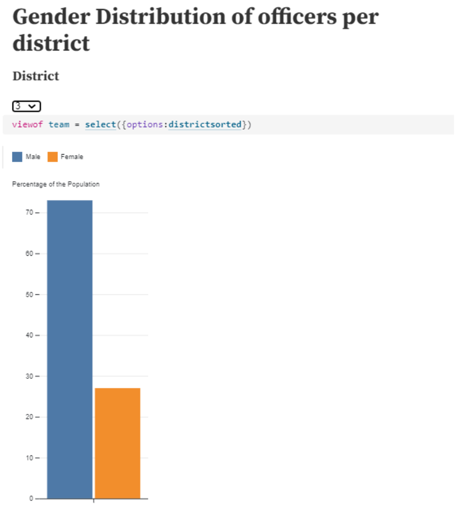
- The data I got allowed to preemptively see that the Gender Bias of officers was greatly skewed towards males
- The data also seemed to show that districts police racial demographics were not always representative of the population
- The combined data allowed us to create effective visualizations in the next part of the project

Interactive Visualizations

Visualizations Being Done

1. Gender Distribution of Police Officers per District
2. Population Race Distribution vs Police Race Distribution per district
3. 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



Gender Distribution of officers per district

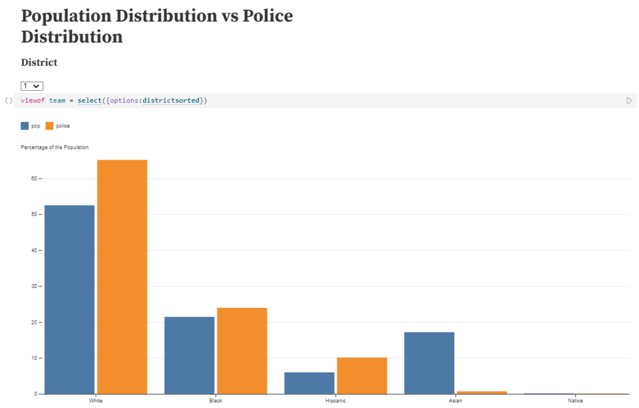
District

3

```
viewof team = select({options:districtsorted})
```

The gender distribution graph for each district shows us that there is a clear disparity in the gender distribution of officers in a district. If we comb through each district's values, we can see that most times the male officers outnumber female officers in a nearly 4 to 1 disparity. This opens the floor to further questions in researching this data that could be conducted which could relate to whether based on population demographics of police officers in a district, do districts with better gender and race distribution amongst the officers lead to a lower allegation rate per police officer? This first visualization was very good for me to also get familiar with the d3.js environment and as a result of it, I was able to easily navigate the code and mechanisms for my other graphs. The graph allows us to see the different districts by selecting a district and then going over each individual district's breakup of gender across the police force.

Visualization 2: Population Race Distribution vs Police Race Distribution per district



Population Distribution vs Police Distribution

District

1

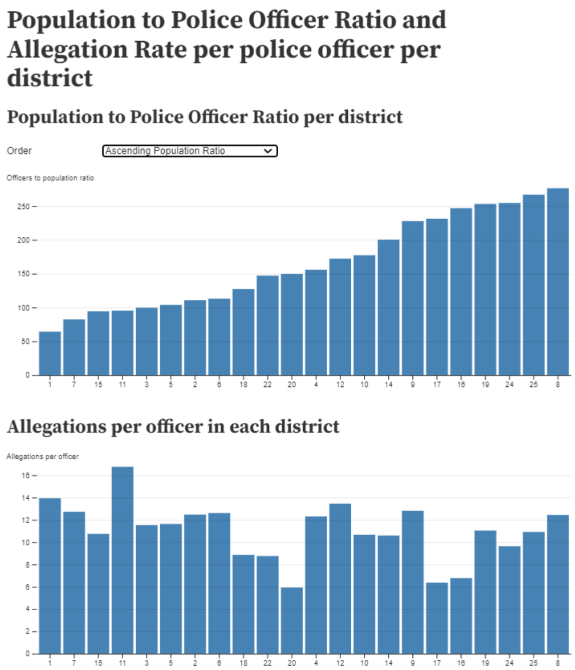
```
viewof team = select({options:districtsorted})
```

The racial distribution of police officers vs the districts they serve was a very interesting graph. In this, we can see the clear distribution of officers and district demographics as well as where disparities exist. If we dive deeper into district 11, (which has the highest number of allegations per officer as seen in the 3rd visualization), we see that there is an overwhelmingly large number of white officers in this district. The district is also majority black in terms of population and the ratios of the police officers to the district does not match. This showcases a potential of racial bias in the district. The converse can be seen in district 20 which has the lowest allegations per police officer. The distribution of police officers in this district is closer to the distribution of the population. The district also has a white majority of civilians. It is data like this that shows that while principles are in place to reduce bias, there is still a potential bias evident in the districts across Chicago. As seen in the sample graph above, we can select the district to see the distributions in each individual district and thus visualize potential disparity in the police and civilian distributions.

Visualization 3: Population to Police Officer Ratio and Allegation Rate per police officer per district (Combined Interactive Visualisation)

This visualization is a very important one that was used to work against the hypothesis that the allegation rate is greater when there are more police officers in the district. In this interactive visualization, users can take from multiple different sorting options (Ascending and Descending Police per capita, Ascending and Descending Allegations per police officer in a district, and District wise) to see whether this hypothesis is true. Based on the visualization, it is evident that there is 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. This visualization, in conjunction with the 2nd visualization, also allows us to see how officer and district racial demographics seem to potentially play a role in the allegations made.

In the sample Visualization Below, I have set it to an ascending population ratio which means that as we go from left to right in this graph, the number of officers relative to the population decreases. As can be seen, there is no linear correlation between these 2 graphs and adjusting it in other ways shows the same thing. This showcases that the police officers serving a district are not getting more allegations as a result of an increased population and points to the presence of an underlying bias.


Population to Police Officer Ratio and Allegation Rate per police officer per district

Population to Police Officer Ratio per district

Order: Ascending Population Ratio

Allegations per officer in each district

What did these Visualizations help us understand?

- Helped Visualize data sourced in tables from the first Checkpoint (Visualizations 1, 2, and 3)
- Helped Identify and Visualize disparities related to Gender Distribution amongst officers in a district (Visualization 1)
- Helped Identify and Visualize the racial distribution of Population vs Police in a district (Visualization 2)
- Helped disprove the idea that Allegations only increase when Police per capita increases (Visualization 3)
- Helped Understand and see the correlation between Population vs Police Distribution and the Allegations in a district (Visualization 2 and 3)

Final Thoughts

These visualizations were extremely fun to make and helpful in showcasing that there is a potential for bias amongst police officers and the districts they serve. By going through these visualizations, we can see factors such as the gender gap and the Population vs Police Racial distribution more clearly. By combining insights across different visualizations, we can further see potential for inherent bias showing that we are moving closer to our overarching theme of determining risk amongst officers and showcasing a clear path toward creating Machine Learning Models that can identify risk and factors that lead to bias.

Machine Learning

Goals

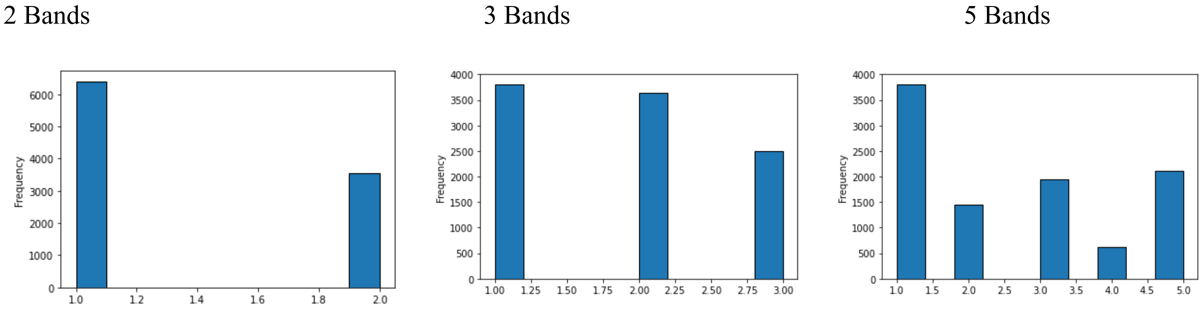
1. *Supervised Machine Learning* - Determine an officer's potential for risk spread across different bands (upto 5 bands) based on the information in data_officer combined with district information for each district. The potential for such a system is to recommend officers for sensitivity training based on their and their district's information in case they seem to exhibit high risk.

2. *Unsupervised Machine Learning* - Run K Means Clustering on officer data to see what factors seem to be important and influence officers being grouped together. This will help see if there are some standout characteristics that seem to divide officers in different groups.

Data Cleaning

To Clean this data and to make sure that I had all the bands setup for the classification process, I used the Data_Cleaning.ipynb jupyter notebook. In this notebook which is linked here. I combined multiple different CSVs that had been sourced and created throughout the project. I calculated the mean and standard deviation of allegations per district so as to get the metrics with which I could create the bands. Graphs showing the distribution of these risk bands can be seen below

Graphs Showing Bands



2 Bands

3 Bands

5 Bands

2 Bands - *(1 - Low Risk, 2 - High Risk)*

3 Bands - *(1 - Low Risk, 2 - Medium Risk, 3 - High Risk)*

5 Bands - *(1 - Very Low Risk, 2 - Low Risk, 3 - Medium Risk, 4 - High Risk, 5 - Very High Risk)*

Did we accomplish these Goals? What did we learn?

Supervised Machine Learning

Now that I had cleaned the data and created these different bands, to accomplish this goal, I would have to train a few models with the data (district, gender, race, allegation_count, sustained_count, unsustained_count, birth_year, black, white, hispanic, blackpop, whitepop, hispanicpop) corresponding to the officers with the band values as the target labels. The code for this can be seen by opening Models_for_classification.ipynb or by clicking here.

I start in this notebook by first uploading the cleaneddata_2.csv file as a dataframe. I then dropped the unnecessary columns from the cleaned data (std, mean, etc.) leaving only the columns mentioned above. I then iterated through the different target columns (y, y2, and y3) for models on different bands. Considering this was a classification task, I chose the DecisionTree, KNeighbours, and Logistic Regression Classifiers for training the models.

To test each of these models on the different datasets, I also used a library called cross-validate. This library allowed me to perform K-fold validation which essentially splits the data into K parts (in this case 5) and iterates over taking 4 parts for training and 1 part for testing with each part getting a chance to be a validation part. This is a great way to test a models potential as it ensures that the model isn't just performing well for a specific set of data.

The results of training these models across the different number of bands can be seen in the table below. The best performing model has its accuracy in bold for each set of bands.

	Decision Tree Classifier	K Neighbours Classifier	Logistic Regression Classifier
2 Bands	94.93%	94.62%	**95.44%**
3 Bands	89.60%	88.50%	**90.50%**
5 Bands	**81.42%**	79.22%	77.08%

As can be seen in the table above, we can get quite good accuracy across the board even with 5 bands. The best performing models for both 2 and 3 bands was the logistic regression classifier. This was expected because Logistic Regression performs very well for a small number of classes. When the number of bands was increased to 5, the Logistic Regression Classifier did not perform as well but the Decision tree classifier performed very well.

Final Thoughts and what we learnt with Supervised Machine Learning

This was a very interesting task and showed that given officer data, we can create bands that determine, to a reasonably high degree, even with 5 bands the risk level of an officer. The higher the risk of an officer is the greater the chance is for the officer to have allegations against them and thus officers with a high band can be recommended sensitivity training that would help them become better officers with a less chance of being biased. I was able to accomplish my goal of creating effective models that could model an officers potential for risk

Unsupervised Machine Learning

For this Goal, I wanted to take the same officer data I already had and then, using different factors (different data columns) across 2 clustering scenarios, see what features stood out in terms of importance for them. I changed the dataframe from the first to the second scenario based on the reasoning given below.

Clustering 1: Linked Here

Labels: district, gender, race, major_award_count, allegation_count, sustained_count, unsustained_count, birth_year, black, white, hispanic, blackpop, whitepop, hispanicpop

Reasoning: Simply using all the data that wasn't mean/stdev or class data for Clustering

Clustering 2: Linked Here

Labels: district, gender, race, major_award_count, allegation_count, sustained_count, black, white, hispanic, blackpop, whitepop, hispanicpop

Reasoning: Removing Birth_year and unsustained_count as they skewed the results in a manner that felt like there were no clear insights

For Each of these, I first imported the dataset that we had created when cleaning the data. I then dropped all the columns except those specified here. Then, I followed this by performing one round of clustering. I went from 1 to 10 clusters each time following which I graphed the sum of squared distances. I did this to use the elbow method through which I was able to find the optimal number of clusters each time. I then used the Kmeans feature importance library (Linked Here) which is a wrapper class for the sklearn KMeans system to determine the most important features for each clustering set with the optimal number of clusters.

Within this I used the WCSS method which determines which feature had the maximum impact on the Within-Cluster Sum of Squares which is what KMeans attempts to minimize. I then graphed the 3 most important features across each cluster that had been created and drew inferences from that.

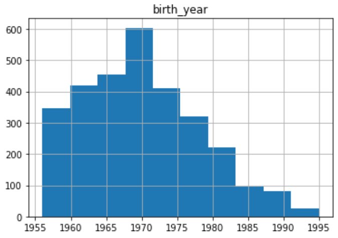
Clustering 1

Optimal Clusters = 4 (Based on the Elbow Method)

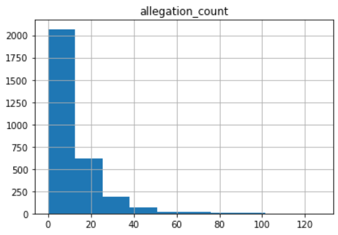
Most Important Features from Highest to Lowest Importance: Birth Year, Allegation_count, Unsustained_count

Sample Graphs from these areas

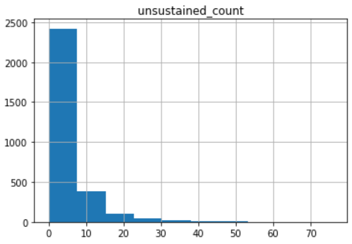
Birth Year



Allegation Count



Unsustained Count



Inferences Drawn from this Data

Based on this clustering mechanism, I was unable to truly get much information. As a result, I decided to perform clustering again but this time omit the birth_year and the unsustained count class which seemed to influence the clusters heavily

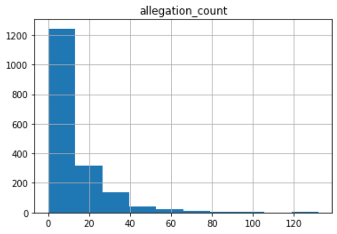
Clustering 2

Optimal Clusters = 4 (Based on the Elbow Method)

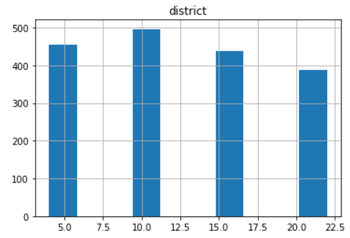
Most Important Features from Highest to Lowest Importance: Allegation count, District, Race of Officer

Sample Graphs from these areas

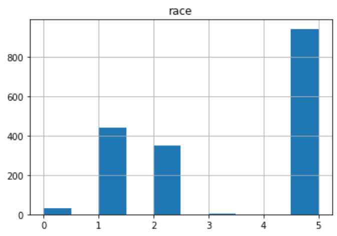
Allegation Count



District



Officer Race



Inferences from this Clustering

Well on looking at the clusters across this dataset further, unfortunately, while it did bin multiple specific districts together and had no overlap, there seems to be no relation to the allegation count or other important features. However, all hope is not lost when this is seen, it’s just an indication that we still may not have explored the perfect level of labels and clusters to find this out. I still believe that there is scope for showcasing something with this type of clustering as can be seen with the differences across the race based clusters in the second clustering experiment. It does seem to showcase that there is some indication of a reasoning to which allegations, race, and districts are important to clustering. The findings just seem to be inconclusive for me. (An unfortunate consequence that sometimes happens with Unsupervised Machine Learning). While I was able to accomplish my goal of performing unsupervised machine learning on Officer Data and also determine important features that lead to clustering, I could not go beyond that. This does however provide an interesting angle for future work.

Final Thoughts with Machine Learning

- I created some pretty nifty models that can classify officers based on risk perception quite well across even 5 bands of risk. The models showcase potential for creating a Machine Learning based risk perception system which can flag individuals for sensitivity training to have officers that are less prone to allegations against them and thus resulting in officers who better serve our communities
- I tried to perform clustering on two sample datasets modified against the data_officer combined with district data CSV. The findings were slightly inconclusive but do seem to show that allegation rate, district, race of the officer, birth year, and unsustained count may have potential in showcasing some connection across officers

Overall Project Discussion and Scope for Future Work

Through this project, I learnt a lot about different Data Science methods and how to employ them to effect a holistic outcome. Starting off, I queried data directly from the Chicago Police database and began to see important features and different things with respect to what factors showcased differences between civilians and police demographics. Visualizing this data helped these factors become more evident and helped see these disparities further as well as showcase a very important factor which was that officer allegation rates are not just correlated with the number of officers serving a district. This showed that there was potential for bias amongst officers and by using machine learning, I was able to create a Proof of Concept for a system that would classify officers by risk across different bands of risk. Even the clustering methods, while presenting somewhat inconclusive results, showed that there were some features that had a larger impact on grouping officers together including the districts they serve and an officer’s race. The project unearthed a lot of information and the scope adapted along the way but I feel that I was able to accomplish the task I set off to accomplish and that my findings showcase potential for working with the Chicago police so that officer’s can reduce inherent bias and better serve the community

Scope for 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
- Investigating the impact that sensitivity training has on officers and their allegations and complaints moving forward
- Investigating the direct relation between an officer’s race and the allegations that they have received to create even more refined systems for risk detection