Boston PD Payroll Investigation Final Project -- Deliverable 2 & 3

Sandy Seedhom, Brandon Im, Nick Cheng-Yen Huang, Laura Reeve

Goals of the project:

Initially, we planned to look through data from various Massachusetts cities to look for overlap of officers that had been kicked off the force in one city and then moved to another city. However, the plans for the project changed midway through the semester, so we've been focusing on analyzing discrepancies in salary data based on race and gender. We analyzed state police data along with data from the following cities: Lowell, Quincy, Worcester, and New Bedford.

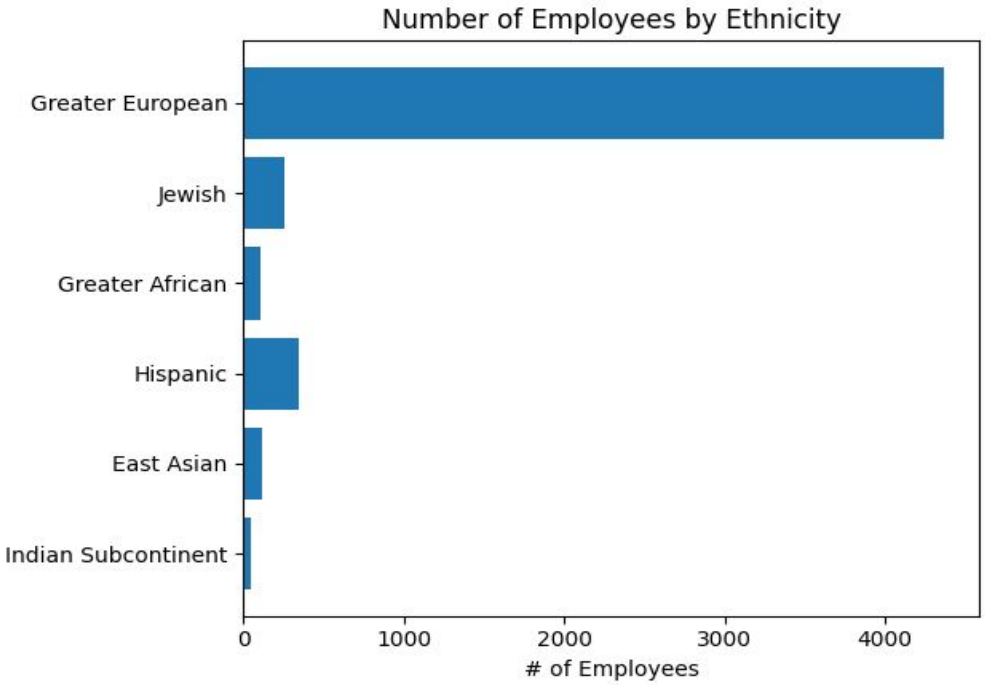
We did this to see if there were any major differences in pay based on officers' ethnicity or gender.

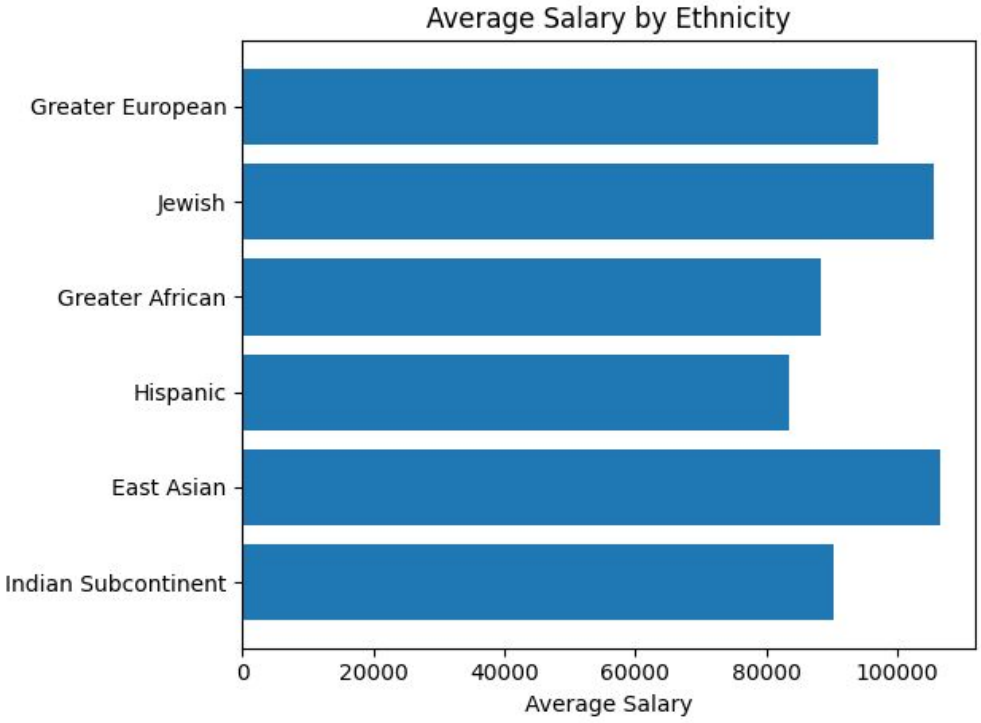
Methods Used:

To obtain the data, we used the scrapy Python library to crawl and scrape through govsalaries.com and output the relevant data into a CSV file. After processing the data to leave only the information that we needed, we used various APIs to try to determine race and gender for each officer.

Ethnicolr: We used the Ethnicolr API to try to understand the breakdown of ethnicities and the corresponding salaries in our data. To do this, we parsed the first and last names from the data we scraped and found the most likely ethnicity for each person. Then we combined the datasets from the towns and graphed the data using matplotlib. Since Ethnicolr has many different options for ethnicities, we broke these down into the following 5 categories (in parentheses is the number of people in each group):

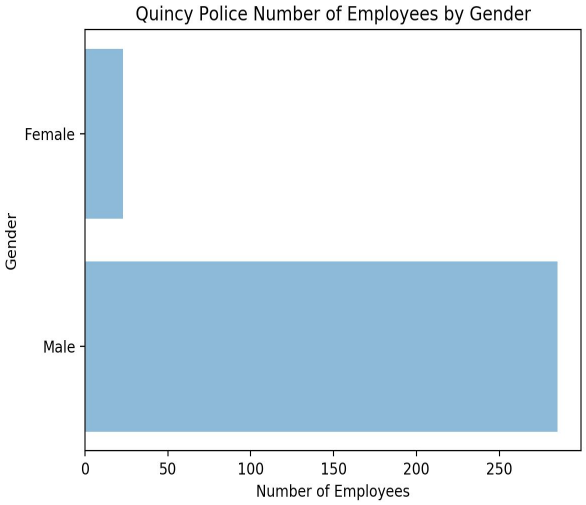
1. **Greater European:** British, EastEuropean, French, Germanic, Italian, Nordic (4367)
2. **Jewish:** GreaterEuropean,Jewish (259)
3. **Greater African:** GreaterAfrican,Africans, GreaterAfrican,Muslim (112)
4. **Hispanic:** GreaterEuropean,WestEuropean,Hispanic (344)
5. **East Asian:** Asian,GreaterEastAsian,EastAsian, Asian,GreaterEastAsian,Japanese (121)
6. **Indian Subcontinent:** Asian,IndianSubContinent (45)

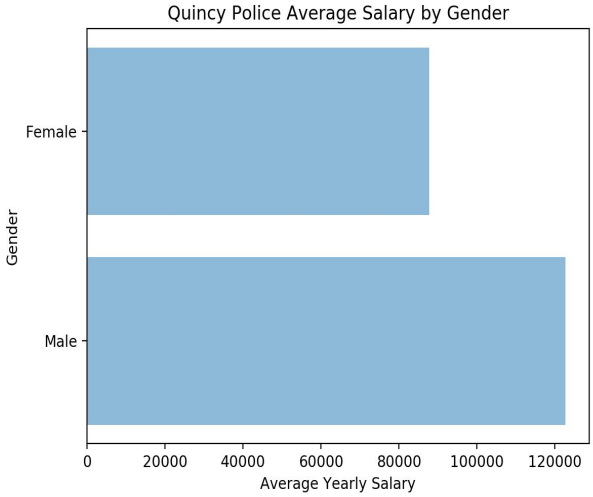


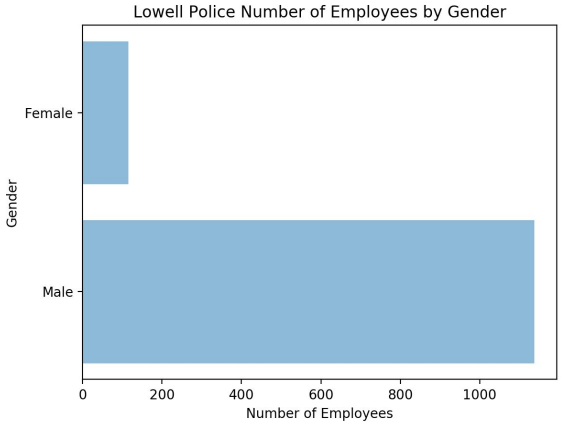


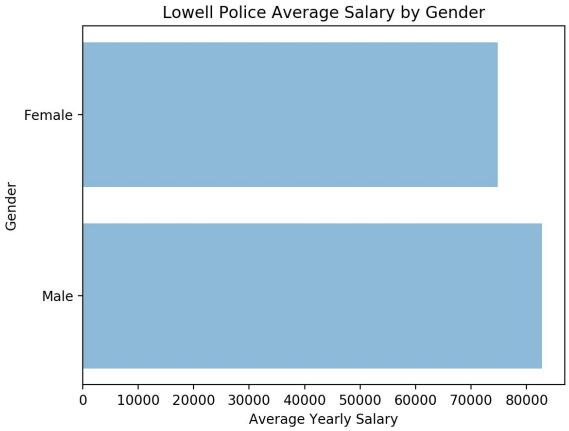
As we can see, people in the East Asian and Jewish groups have the highest average salaries, while people in the Greater African and Hispanic groups have the lowest average salaries. This may be due to differences in

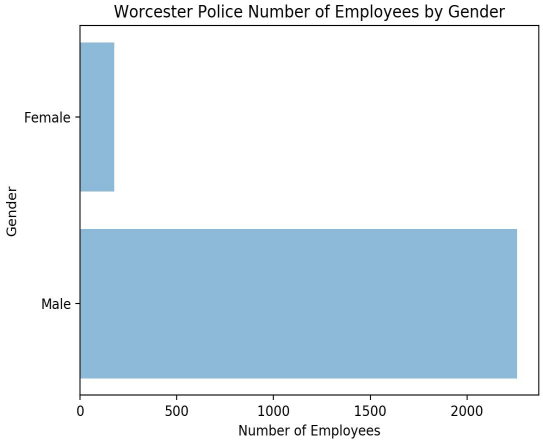
NameSor: We used the Namsor API to try to understand the breakdown of genders and the corresponding salaries in our data. To do this, we parsed the first and last names from the data we scraped and found the most likely gender for each person using the API. We graphed the data using matplotlib for each city (so far only Lowell, Quincy, and Worcester are complete). The data is graphed below:

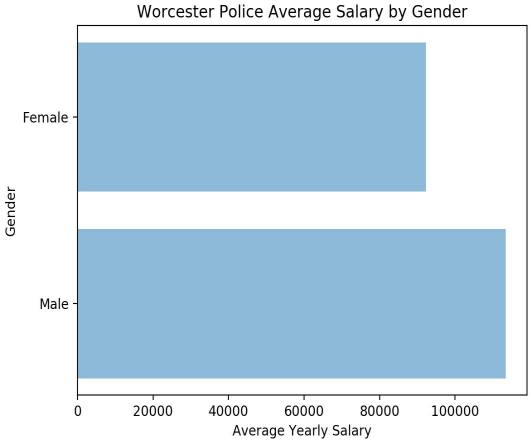












Analysis:

After collecting race and gender data, and later using python libraries such as matplotlib to visualize the data, we found that, in general, employees of Greater European descent had a greater representation in the police force than any other race. Likewise, those of East Asian, Jewish, and Greater European ancestry tended to command more pay. In contrast, those of Hispanic and Greater African descent tended to make the least. In our final deliverable, we'll also be including the data from State workers.
For the gender data, we found that a majority of police in the cities we analyzed are males. On average, male police also make more salary annually than female police do. The data is summarized below:

City/Gender	**Male**	**Female**
Quincy	$122,705.66	$87,860.74
Lowell	$82,772.43	$74,843.83
Worcester	$11,3388.68	$92,264.63

Challenges:

We had a few challenges while working with the various APIs and trying to get accurate predictions from the data. Firstly, we had to sort through all of the employees' occupations for police related words like 'Captain', 'Officer', 'Chief', etc, which could also coincide with terms used within the fire department or other law enforcement occupations. So, employees that may not be part of the police department may have been included in our data.

Regarding our challenges with Ethnicolr, it is a great API, but it's very limited since it is trying to determine ethnicity based solely on a person's name. For some names, such as many east Asian names, there's a fairly clear origin and you can assume that the person is from that area, but that's not the case for all names. For example, a lot of African-Americans have names with European origins, given the history of the US. Thus, they wouldn't fall into the "Greater African" category but would be classified as "Greater European", even if they have no European blood. Thus, we need to look at the Ethnicolr data with this knowledge and not take the resulting analyses at face value.

Another challenge we faced is that the Mamesor API can only be called per 5000 names, and so we had to create a new account everytime we processed 5000 names, which happened many times since we're dealing with such large datasets. Additionally, certain cities' employees occupations were unlisted in our data, which will have to be fixed in the future.

There are a few additions that we'll be making in our final deliverable. Firstly, we'll be running Ethnicolr on the State Police dataset and Namsor on New Bedford and State dataset, which will give us more to analyze. On top of that, we'll be looking at ways that we can separate this data by job level, because it would make sense for a captain to make more than a first year officer, and this may account for some of the pay differences we see.

Conclusions:

So far, it seems as if there is a large gap between average salary and gender, as well as average salary and ethnicity. We hope that once we complete the analysis of all of our data, it will give us a clearer picture of the real payroll discrepancies.