Boston PD Payroll Investigation Final Project -- Final Deliverable

CS 506 in collaboration with WGBH and BU Spark!

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

Goals of the project:

Initially, we planned to gather and analyze data from various Massachusetts cities to look for overlap of officers that had been decertified in one city and then moved to another. However, the plans for the project changed midway through the semester due to technical complications, so we've pivoted our goal towards analyzing discrepancies in salary data based on race and gender in the police force. We analyzed State police data along with data from the following cities: Lowell, Quincy, Worcester, and New Bedford.

Data Collection

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. We initially used beautiful soup, but found that we needed scrapy's crawling capabilities, and since data wasn't loaded dynamically on govsalaries.com, using it wouldn't be an issue. Thus, we built and ran a spider that would crawl through every page of employees for each year of the designated city. The fields of the spider were: Name, Year, Title, Employer, Monthly Wage, and Annual Wage. We programmed the spider to look into every employee's employment details page to retrieve their wage.

Data Preparation and Cleaning

After processing the data to leave only the information that we needed, we then cleaned the data by filtering out any employee without an occupation listed as "Lieutenant", "Captain", etc. We then filtered the data again, removing any employee whose occupation included "Fire" so that we could remove any state firefighters from our data. We then 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):

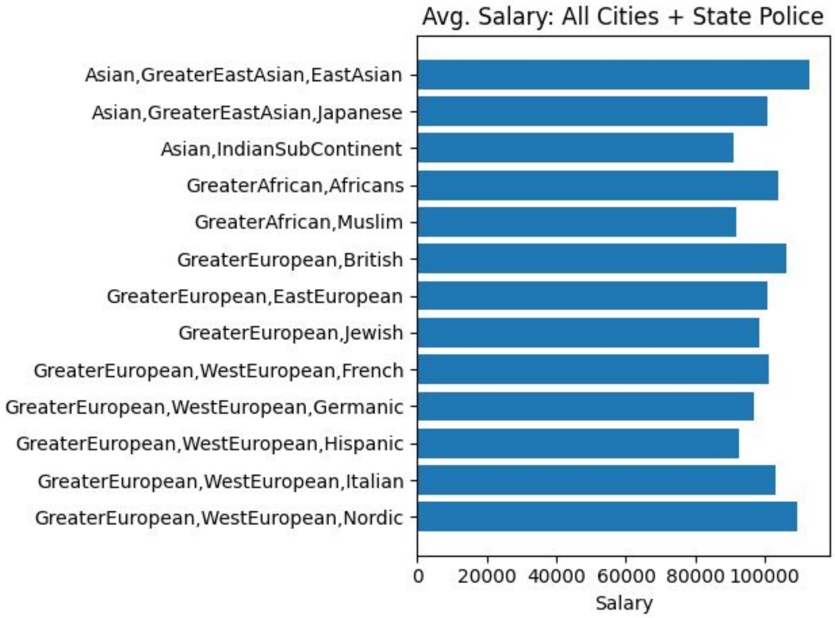
1. **Greater European:** British, EastEuropean, French, Germanic, Italian, Nordic (28756)

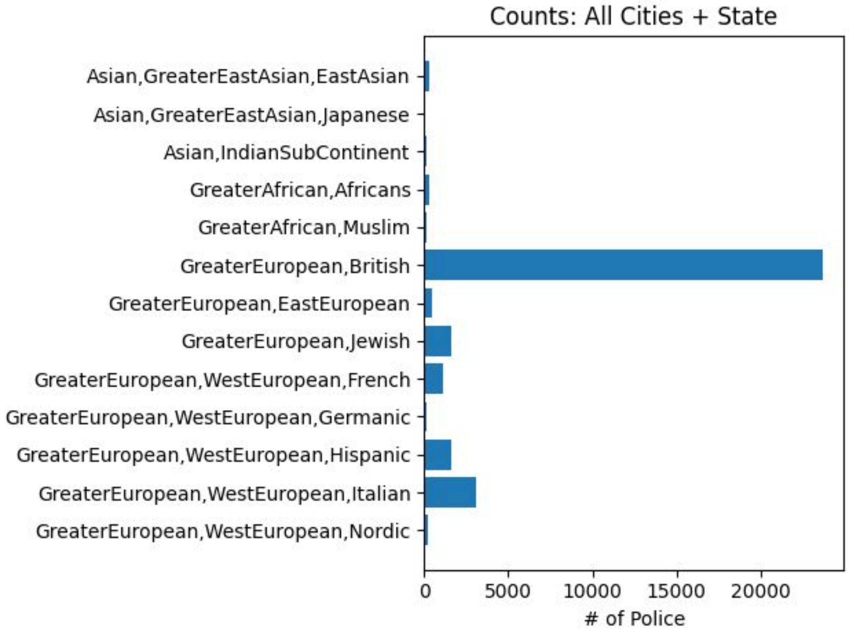
2. **Jewish:** GreaterEuropean,Jewish (1600)
3. **Greater African:** GreaterAfrican,Africans, GreaterAfrican,Muslim (367)
4. **Hispanic:** GreaterEuropean,WestEuropean,Hispanic (1621)
5. **East Asian:** Asian,GreaterEastAsian,EastAsian, Asian,GreaterEastAsian,Japanese (331)
6. **Indian Subcontinent:** Asian,IndianSubContinent (141)

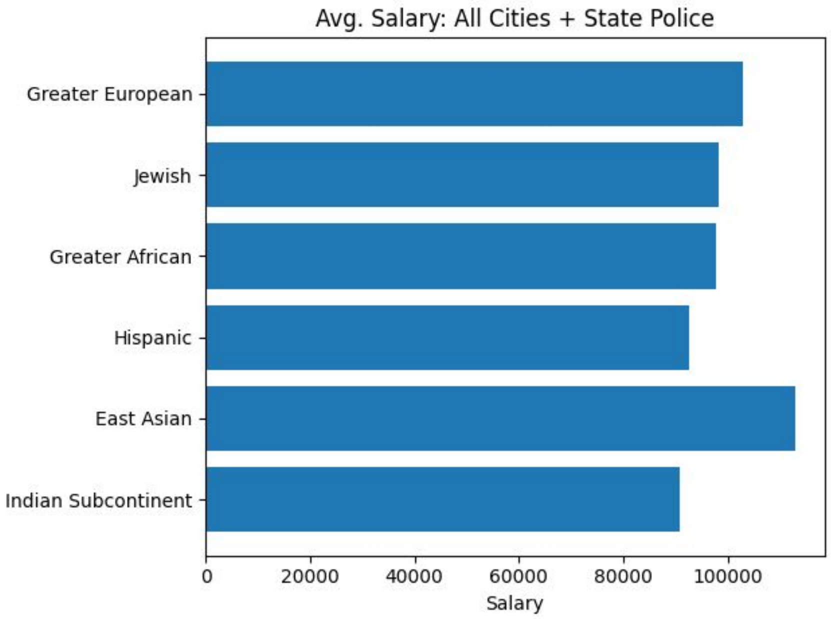
NameSor: We used the Namsor API to try to understand the breakdown of genders and the corresponding salaries in our data. To do this, we parsed the first and last names from the data we scraped and found the most likely gender for each person using the API. We graphed the data using matplotlib for each city. The data below is represented as (average salary, number of employees).

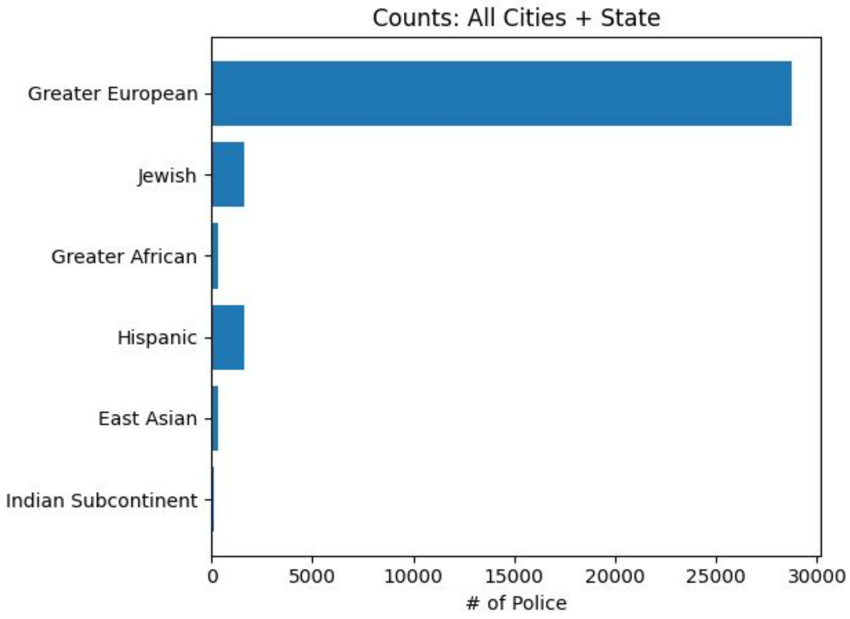
Analysis:

Ethnicity vs. Salary results: Ethnicolr:

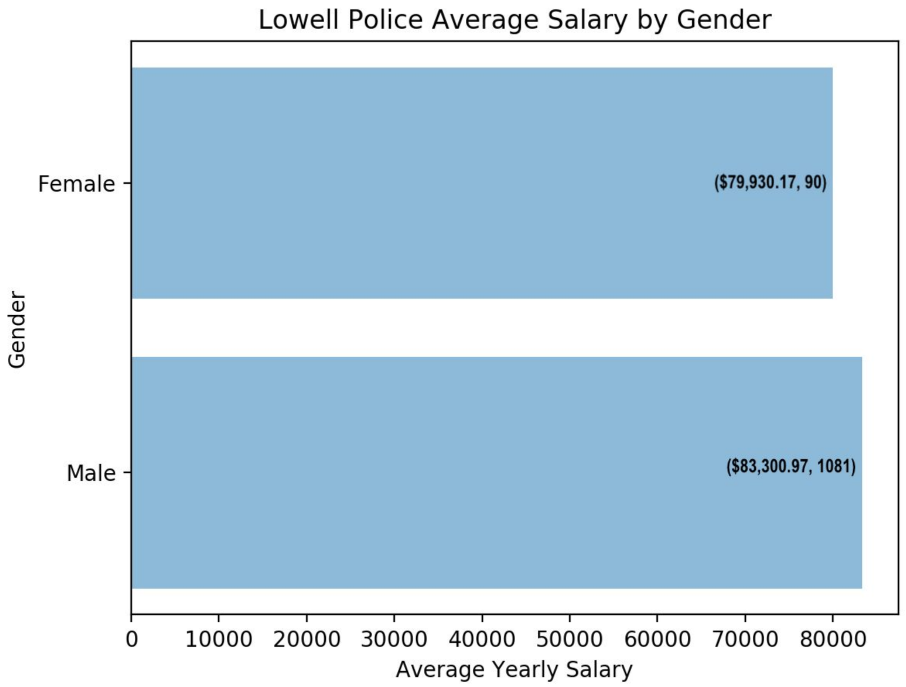


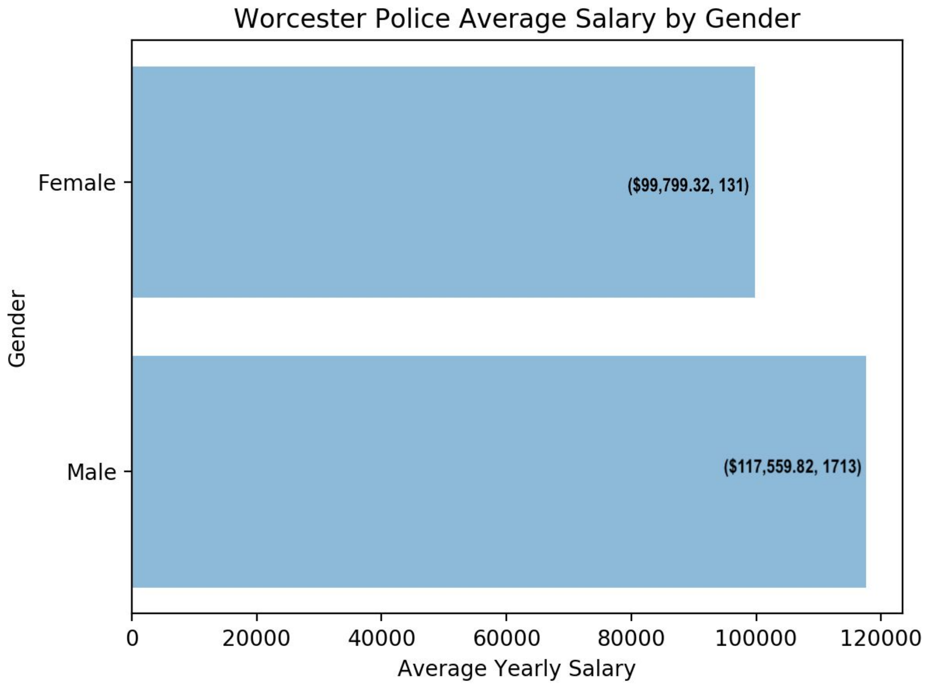


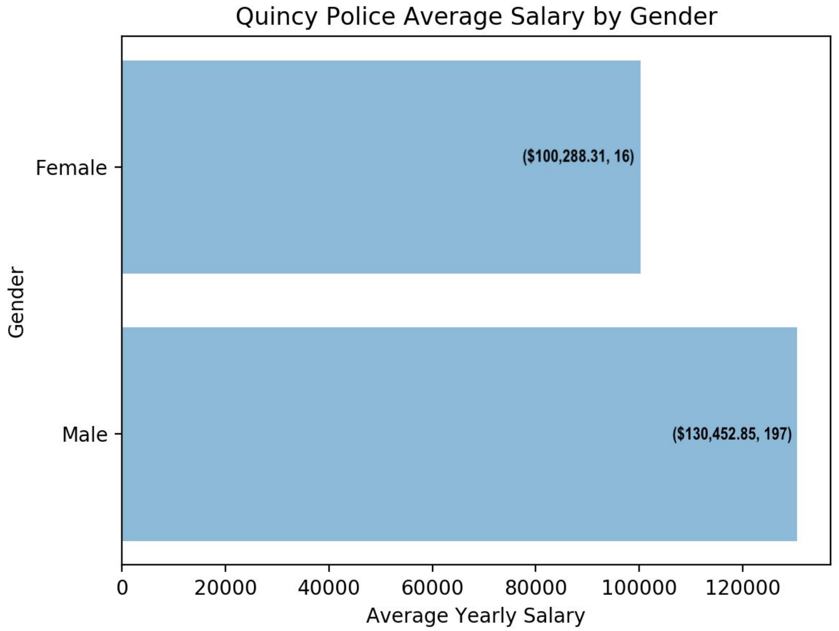

Avg. Salary: All Cities + State Police
Greater European
Jewish
Greater African
Hispanic
East Asian
Indian Subcontinent
0
20000
40000
60000
80000
100000
Salary

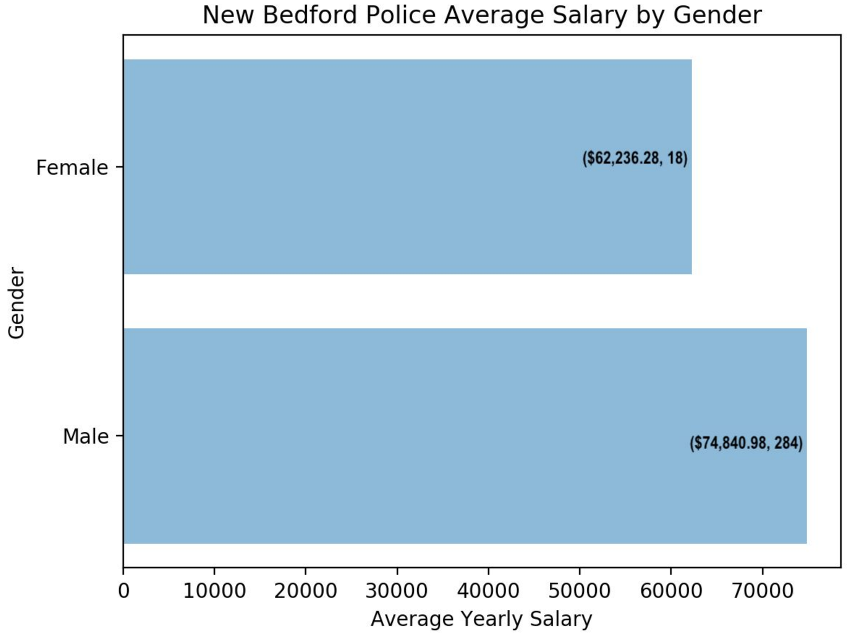

Counts: All Cities + State
Greater European
Jewish
Greater African
Hispanic
East Asian
Indian Subcontinent
0
5000
10000
15000
20000
25000
30000
# of Police

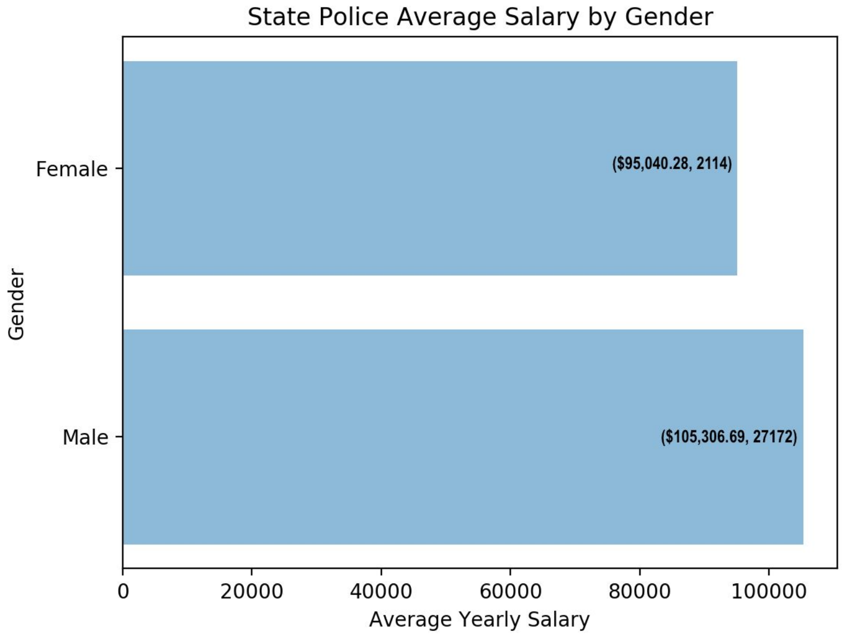
Gender vs. Salary Results: Namsor:

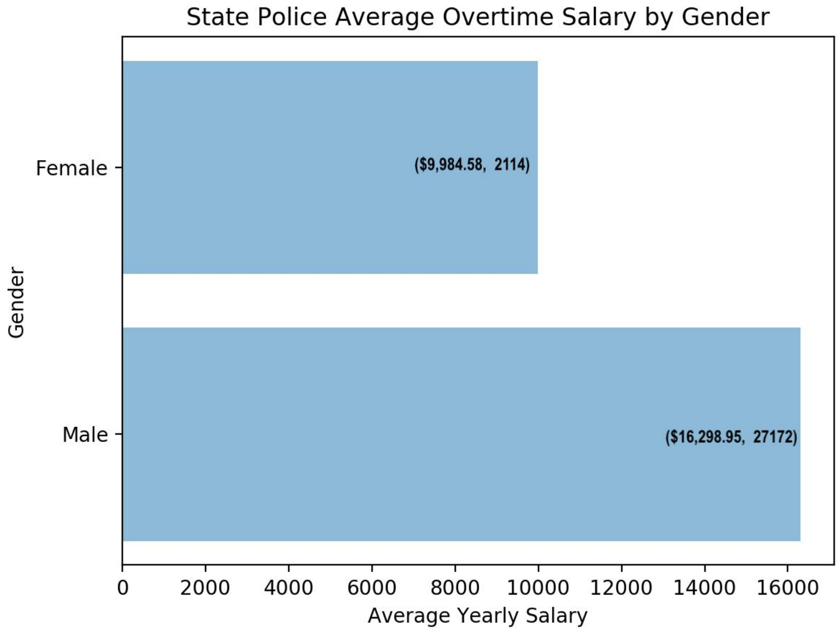


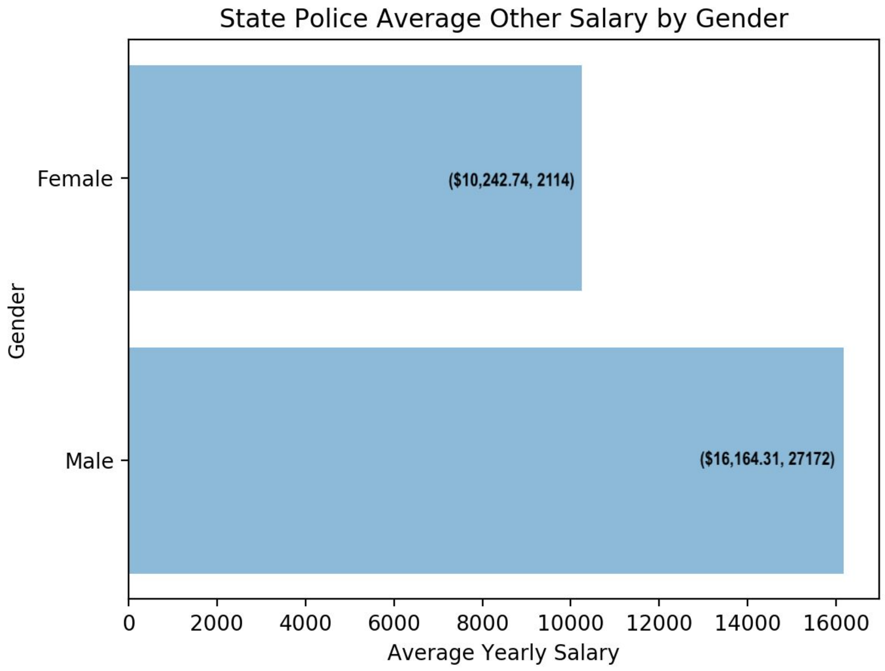


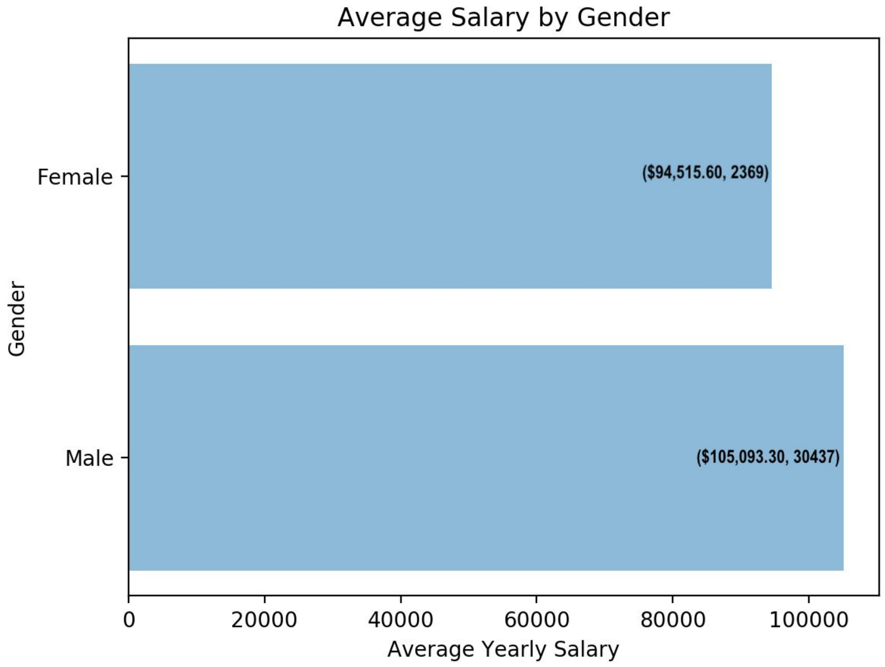

Quincy Police Average Salary by Gender
Gender
Female
Male
($100,288.31, 16)
($130,452.85, 197)
0
20000
40000
60000
80000
100000
120000
Average Yearly Salary


New Bedford Police Average Salary by Gender
Gender
Female
Male
($62,236.28, 18)
($74,840.98, 284)
0
10000
20000
30000
40000
50000
60000
70000
Average Yearly Salary

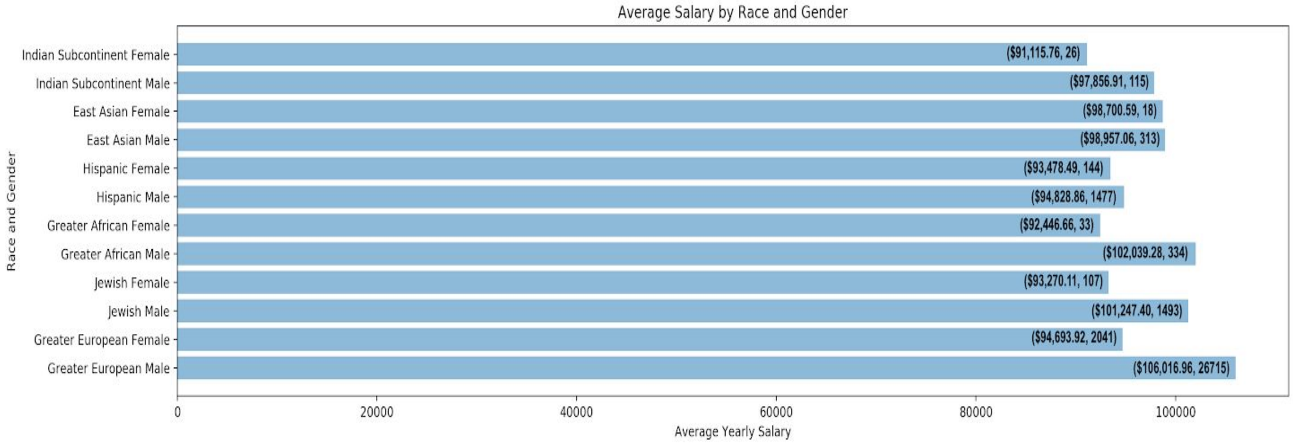

State Police Average Salary by Gender
Gender
Female
Male
($95,040.28, 2114)
($105,306.69, 27172)
0
20000
40000
60000
80000
100000
Average Yearly Salary


State Police Average Overtime Salary by Gender
Gender
Female
Male
($9,984.58, 2114)
($16,298.95, 27172)
0
2000
4000
6000
8000
10000
12000
14000
16000
Average Yearly Salary


State Police Average Other Salary by Gender


Average Salary by Gender

Overall Results:



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.
For the gender data, we found that a majority of police in the cities we analyzed are males. On average, male members of the police force have a higher annual salary than female police officers do, which is shown consistently amongst all the cities we analyzed. The biggest discrepancy between gender and pay was within Quincy (~30,000), and the smallest was within Lowell (~$4,000). Also, we found that male State police officers make about 150% more average overtime income yearly than their female colleagues.
Combining gender and race, we found that Greater European males make the highest average salary, while Indian Subcontinent females make the smallest average salary.

Project Complexity:

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 were initially included in our data. This required us to further clean the data to ensure that only police members were a part of our analysis. This was still difficult, though, because we are not familiar with certain police position titles listed as employees' occupations.

Additionally, certain cities formatted their data differently which caused complications when gathering and cleaning the data. Cities like New Bedford also didn't list employees' occupations for certain years, and the difference in the number of listed employees for each year made it so that we had to exclude the data gathered for the years where the occupations were missing.

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 Namesor API can only be called per 5000 names, and so we had to create a new account every time we processed 5000 names, which happened many times since we're dealing with such large datasets.

Conclusions:

Based on the collected data, we've concluded that there is a significant gap in average salary between genders. Police officers of the male gender tend to receive higher pay than their female coworkers. Likewise, there is a similar discrepancy among ethnic groups, where officers of Greater European ancestry received a higher average salary than those of any other race.

While we tried our best to avoid biases, possible biases in our data may stem from not taking into account differences in part-time and full-time workers due to it not showing in the officers employment details, and not taking into account the difference in titles, since they were so different amongst the cities.