CS506 Coursework Report

--Massachusetts Police Department Payroll Investigation

Anzhe Meng, U50590533
Ruizhi Jiang, U17637349
Jiahao Song, U57363411

Our team was assigned to collaborate with BU Spark! And Mr. Paul Singer, trying to do some exploratory data analysis (*abbr*: EDA) inside the police officers' payroll around the state of Massachusetts. We were in charge of looking into five cities, including Boston, Brockton, Cambridge, Lynn and Springfield. Now over the past couple of months we've collected some interesting patterns in this field. Upon our latest observation and discovery, a policeman is more likely to be better paid than a policewoman according to our observation; and the salary is highly related to the job title, which is somehow affected by a police department employee's race and ethnicity.

In the rest of our report, we are going to present our methodology to acquire our datasets, pre-process the data and implement the EDA.

Data Acquisition

At first, Mr.Singer and BU Spark! provided us the payroll data in different cities. But they were in different formats, which took us several weeks to figure out how to extract these data files to comma-separated values (CSV) format. Some third-party Python packages were imported in order to transform PDF into CSV, such as tabular-py, pyPdf. However, it turned out that around 100 rows were missing in one transformed data table. Even if we could fill out these rows, we were still facing the fatal problem: the data table in different

cities include different columns, thus impossible to sum up those columns and merge them together.

What if we acquisite the payroll data from the government webpage by ourselves? Enlightened by our Project Manager Gowtham Asokan, we managed to scrape the website https://govsalaries.com/, which records all the governmental or public employees' salary over the last few years.

Since we were only given five cities to handle with, take City of Boston as an example to elaborate our web-scraping experience.

In general we implemented the scraping bots with the help of scrapy, a powerful python library that provides users with an established web scraping infrastructure. Even though scrapy is powerful, we still encountered trouble when we visited this website too frequently. So it is obvious that the website is equipped with some sort of anti-scraper mechanism and we were categorized as unlawful visitors.

This problem troubled us quite a while until we found Scrapinghub. It helped us to draw out all the desired data in just a couple of minutes, without spending one cent. So until now, our data basically looked as shown in fig. 1.

Annual_Wage	Employer	Job_Title	Monthly_Wage	Name	Year	_type
159,764	City Of Boston	Fire Captain Admn-Advance Tech	13,314	John Forristall J	2015	GovsalariesItem
146,160	City Of Boston	Dep Supn	12,180	Nora Baston L	2015	GovsalariesItem
140,052	City Of Boston	Fire Captain	11,671	Jamie Walsh J	2015	GovsalariesItem
135,818	City Of Boston	Operational Leader	11,318	Ann Callahan B	2015	GovsalariesItem
130,680	City Of Boston	Director	10,890	Grace Diggs V	2015	GovsalariesItem
128,336	City Of Boston	Principal Elementary	10,695	Soo Cynthia Ann Hoo	2015	GovsalariesItem
125,979	City Of Boston	Police Lieutenant	10,498	William Slavin J	2015	GovsalariesItem
125,384	City Of Boston	Small Learning Comm Leader	10,449	Charles Eudene Cauley	2015	GovsalariesItem
124,427	City Of Boston	Asst Headmaster	10,369	Zayda Cruz-gonzalez	2015	GovsalariesItem
122,994	City Of Boston	Headmaster	10,250	Troy Henninger	2015	GovsalariesItem
122,231	City Of Boston	Police Lieutenant	10,186	Kenneth Macmaster A	2015	GovsalariesItem
121,017	City Of Boston	Prin Dp Sys Anal-Dp	10,085	Jonathan Handy D	2015	GovsalariesItem
120,380	City Of Boston	Director Of Instruction	10,032	Jessica Madden-fuoco R	2015	GovsalariesItem
119,883	City Of Boston	Teacher	9,990	Lindsay Chaves R	2015	GovsalariesItem
117,584	City Of Boston	Police Sergeant/Hdq Dispatcher	9,799	Joseph Maguire M	2015	GovsalariesItem
115,466	City Of Boston	Police Lieutenant	9,622	Charles Kelly G	2015	GovsalariesItem
114,451	City Of Boston	Director	9,538	Shakera Walker A	2015	GovsalariesItem
114,451	City Of Boston	Director	9,538	Jonathan Sproul Galli	2015	GovsalariesItem
114,451	City Of Boston	Manager	9,538	Peter Crossan A	2015	GovsalariesItem
114,487	City Of Boston	Police Lieutenant	9,541	Richard Driscoll J	2015	GovsalariesItem
114,530	City Of Boston	Registrar	9,544	Kenny Chin	2015	GovsalariesItem
114,572	City Of Boston	Police Sergeant	9,548	Gary Barker	2015	GovsalariesItem

Fig.1 Snapshot of Raw Data

Data Preprocessing

Because there was a change of our target, we could describe our data preprocessing in two phrases.

In the beginning, we planned to find out the decertified cops in the payroll. We assumed a policeman/woman was decertified if the same person showed up in different regions. In this case, we needed to join the records together if the names in the records are the same. Thus we just simply utilized *join* in the python library pandas, concatenating records as we desired. Then our data was like fig. 2.

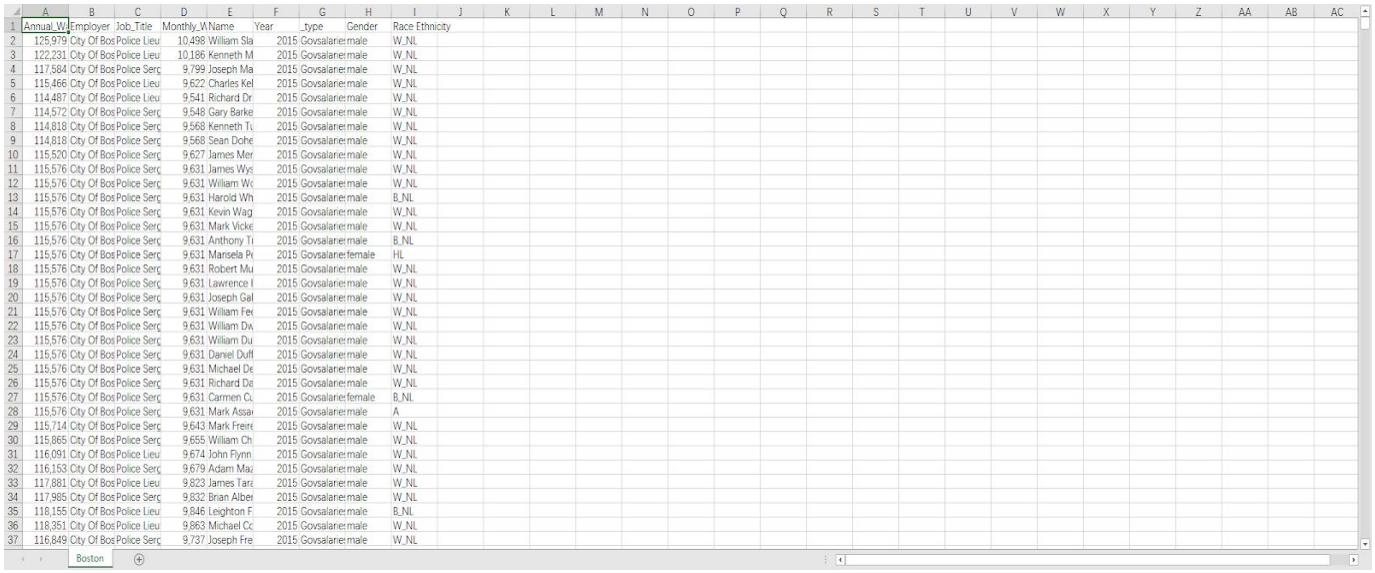
Employer	Job_Title	Monthly_Wage	Name	Year	State:	Agency:	Year decertified:	Unnamed: 4
City Of Boston	Wkg Frpr Linepr & Cablesplicer	10,182	Paul Kelly	2015	Pennsylvania	Not identified	2004.0	https://www.u Pen...
City Of Boston	Wkg Frpr Linepr & Cablesplicer	11,269	Paul Kelly	2016	Pennsylvania	Not identified	2004.0	https://www.u Pen...
City Of Boston	Manager	10,834	Robert Smith	2015	Georgia	Not identified	2006.0	https://www.u Geo...
City Of Boston	Manager	10,942	Robert Smith	2016	Georgia	Not identified	2006.0	https://www.u Geo...
City Of Boston	Police Offc Acad Instr	8,267	William Smith	2015	Kansas	Shawnee Police	NaN	https://www.u Kan...
City Of Boston	Police Offc Acad Instr	11,129	William Smith	2016	Kansas	Shawnee Police	NaN	https://www.u Kan...
City Of Boston	Fire Fighter	8,472	Michelle Johnson	2015	Texas	Dallas County Sheriff	NaN	https://www.u Tex...
City Of Boston	Fire Fighter	9,651	Michelle Johnson	2016	Texas	Dallas County Sheriff	NaN	https://www.u Tex...

Fig.2 Merged Dataframe

But we discarded these ones later because one same name couldn't help us recognize whether they referred to the same person.

In Phrase 2, we broke down our target, only discovering the patterns within the payroll of police. As we expected, we needed to distinguish a person's gender and ethnicity/race only based on his/her name. We introduced the help of namsor. The dataset after this process is shown like Fig.3.

Powered by Namsor, we have two extra columns in our dataset. One column is the Gender. Two values of this column are 'male' and 'female'. This result is very accurate with accuracy rate 99%. The other column is the Race Ethnicity. There are four values of this column. They are 'W_NL'(white, non latino), HL (hispano latino), A (asian, non latino), B_NL (black, non latino). But the result of the race is not very accurate according to the doc of Namsor.



Annual_W	Employer	Job_Title	Monthly_W	Name	Year	_type	Gender	Race Ethnicity
125,979	City Of Bos	Police Lieu	10,498	William Sla	2015	Govsalarie	male	W_NL
122,231	City Of Bos	Police Lieu	10,186	Kenneth M	2015	Govsalarie	male	W_NL
117,584	City Of Bos	Police Serg	9,799	Joseph Ma	2015	Govsalarie	male	W_NL
115,466	City Of Bos	Police Lieu	9,622	Charles Kel	2015	Govsalarie	male	W_NL
114,487	City Of Bos	Police Lieu	9,541	Richard Dr	2015	Govsalarie	male	W_NL
114,572	City Of Bos	Police Serg	9,548	Gary Barke	2015	Govsalarie	male	W_NL
114,818	City Of Bos	Police Serg	9,568	Kenneth Tı	2015	Govsalarie	male	W_NL
114,818	City Of Bos	Police Serg	9,568	Sean Dohe	2015	Govsalarie	male	W_NL
115,520	City Of Bos	Police Serg	9,627	James Mer	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	James Wys	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	William Wc	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Harold Wh	2015	Govsalarie	male	B_NL
115,576	City Of Bos	Police Serg	9,631	Kevin Wag	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Mark Vicke	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Anthony Ti	2015	Govsalarie	male	B_NL
115,576	City Of Bos	Police Serg	9,631	Marisela Pe	2015	Govsalarie	female	HL
115,576	City Of Bos	Police Serg	9,631	Robert Mu	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Lawrence I	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Joseph Gal	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	William Fe	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	William Da	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	William Du	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Daniel Duff	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Michael De	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Richard Da	2015	Govsalarie	male	W_NL
115,576	City Of Bos	Police Serg	9,631	Carmen Cu	2015	Govsalarie	female	B_NL
115,576	City Of Bos	Police Serg	9,631	Mark Assa	2015	Govsalarie	male	A
115,714	City Of Bos	Police Serg	9,643	Mark Freire	2015	Govsalarie	male	W_NL
115,865	City Of Bos	Police Serg	9,655	William Ch	2015	Govsalarie	male	W_NL
116,091	City Of Bos	Police Lieu	9,674	John Flynn	2015	Govsalarie	male	W_NL
116,153	City Of Bos	Police Serg	9,679	Adam Maz	2015	Govsalarie	male	W_NL
117,881	City Of Bos	Police Lieu	9,823	James Tara	2015	Govsalarie	male	W_NL
117,985	City Of Bos	Police Serg	9,832	Brian Alber	2015	Govsalarie	male	W_NL
118,155	City Of Bos	Police Lieu	9,846	Leighton F	2015	Govsalarie	male	B_NL
118,351	City Of Bos	Police Lieu	9,863	Michael Cc	2015	Govsalarie	male	W_NL
116,849	City Of Bos	Police Serg	9,737	Joseph Fre	2015	Govsalarie	male	W_NL

Boston

Fig.3 Example of Data After NamSor

Data Analysis

To begin with, we analyzed the payrolls mainly in two aspects: gender, and ethnicity. And now we are going to illustrate our thoughts in Cambridge and this will also be our sample to analyze other cities.

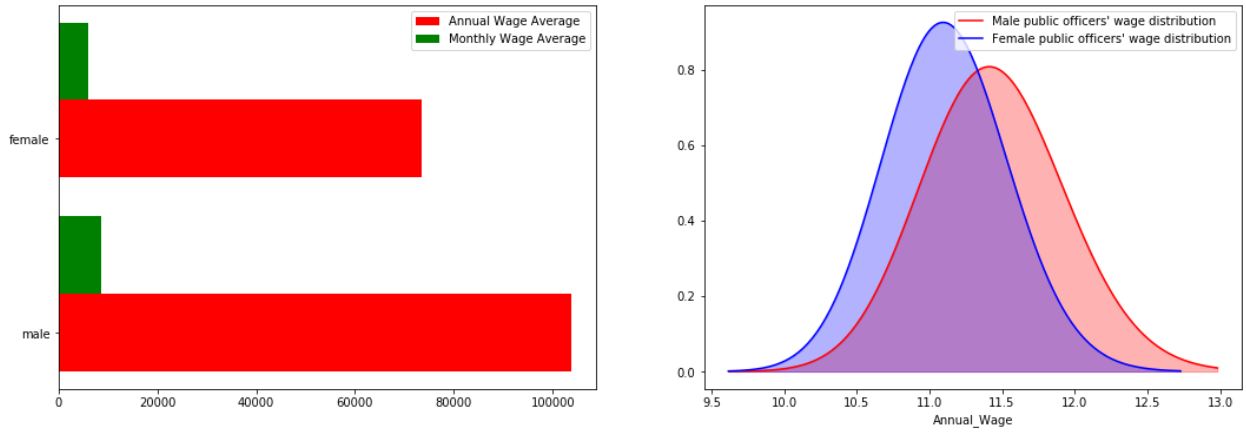


Fig.4 (a) Comparison between Average Salaries of Males and Females;
(b) Distribution of Annual Salaries of Different Genders

As fig.4(a) shows, no matter monthly or annually, a policeman is approximately 30% better paid than his female colleagues if he works

for City of Cambridge. And the ranges of both genders' wages are quite large, which could be shown by fig.4(b).

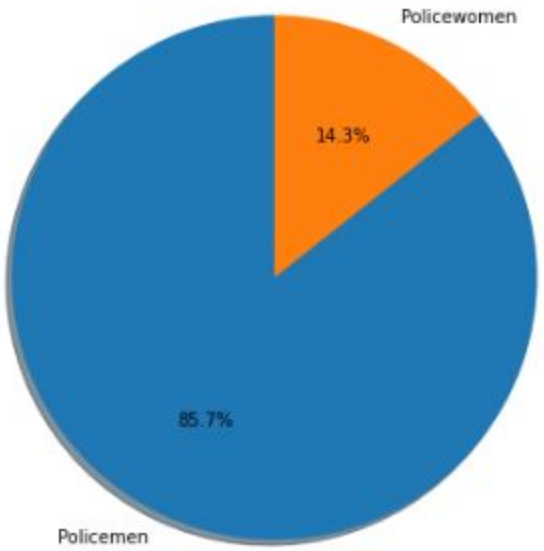


Fig. 5(a) The Distribution of Gender of Police Officers

Fig.5(a) shows that 85% police officers are men, 15% are women.

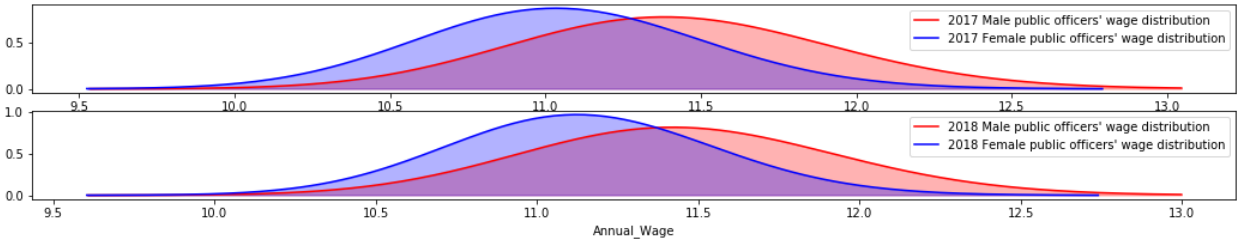


Fig. 5(b) Annual wages over Year 2017-2018

Meanwhile, we could safely conclude that there was no huge variableness in the salary over the period 2017-2018, due to the fact that both graphs look almost the same in fig. 5(b).

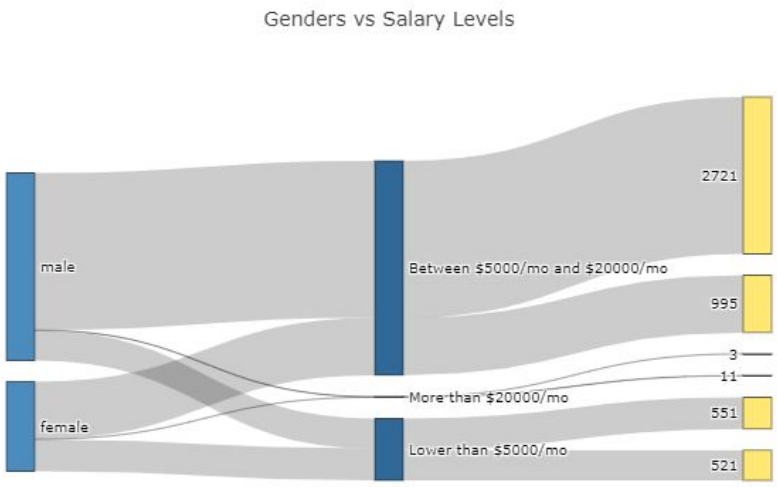


Fig. 6 Genders vs. Salary Ranges

Furthermore, it is found that the reason why the average of female's salary is lower than that of male's is directly because of the big gap of the middle class, while the number of upper class and lower class of employees are the same.

In Fig.7, we have the distribution of wage in different cities. In Boston, the distribution of man and woman is very similar. In Brockton, policemen have the higher average salaries and the lower variance. The same in Cambridge. The variance of distribution of the wage of policemen in Springfield is the smallest among the four cities.

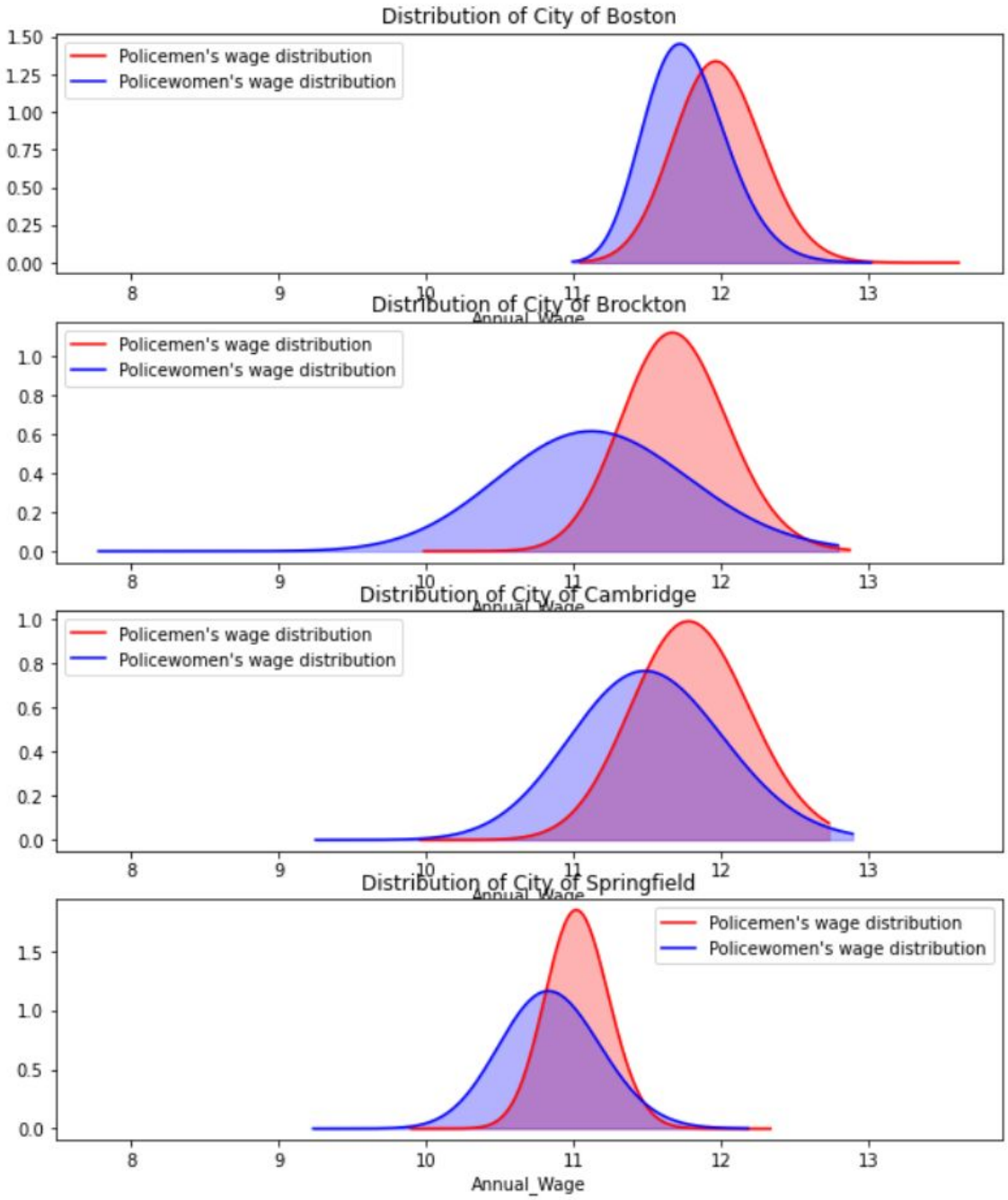


Fig. 7 Distribution of Wage in Different City

Besides, we also do the race breakdown of our dataset. We first want to check the composition of race in our data. Powered by Namsor, we get the composition. Fig.8 shows that white people are more than a

half. The number of black people and hispano people are similar. The asian are only 3.7%.

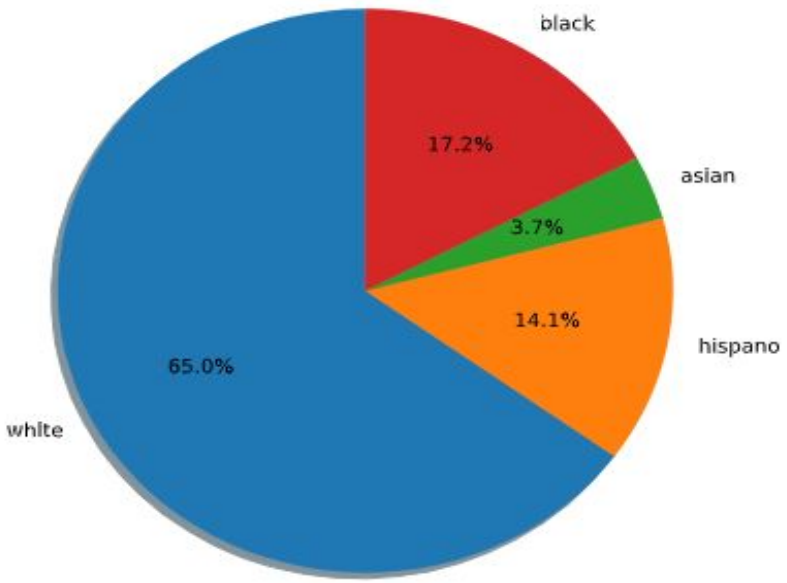


Fig. 8 Composition of Race of Our Data

Fig. 9 is the average salary of different races. Hispano have the lowest average annual and month salary. The average salary of the rest races are very similar.

Fig. 10 shows that the wage distributions of different races are similar in the shape of curves, which means the average of different races are similar, and the variances are similar, too.

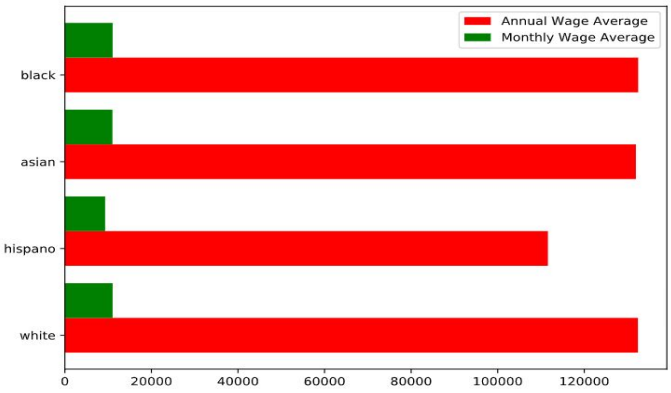


Fig.9 Comparison between Average Salaries of Different Races

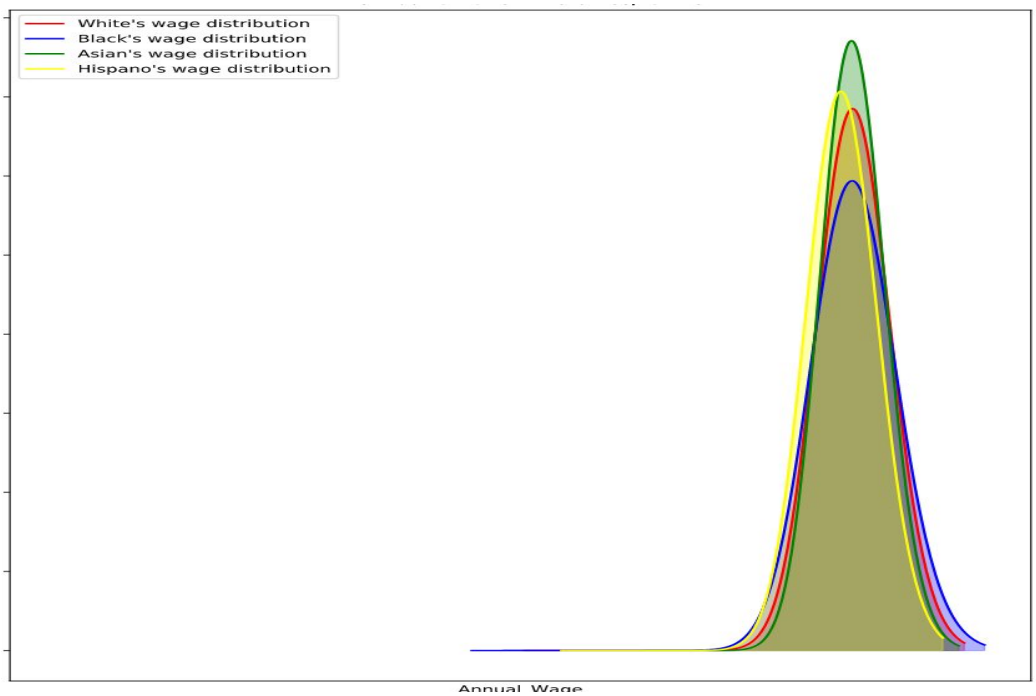


Fig.10 Comparison between Average Salaries of Different Races

Fig.11(a) is the distribution of wages of different races in the city of boston. So in Boston, variance of asian is the biggest, which means some asian get high salaries but some asian get the low salary at the same time.

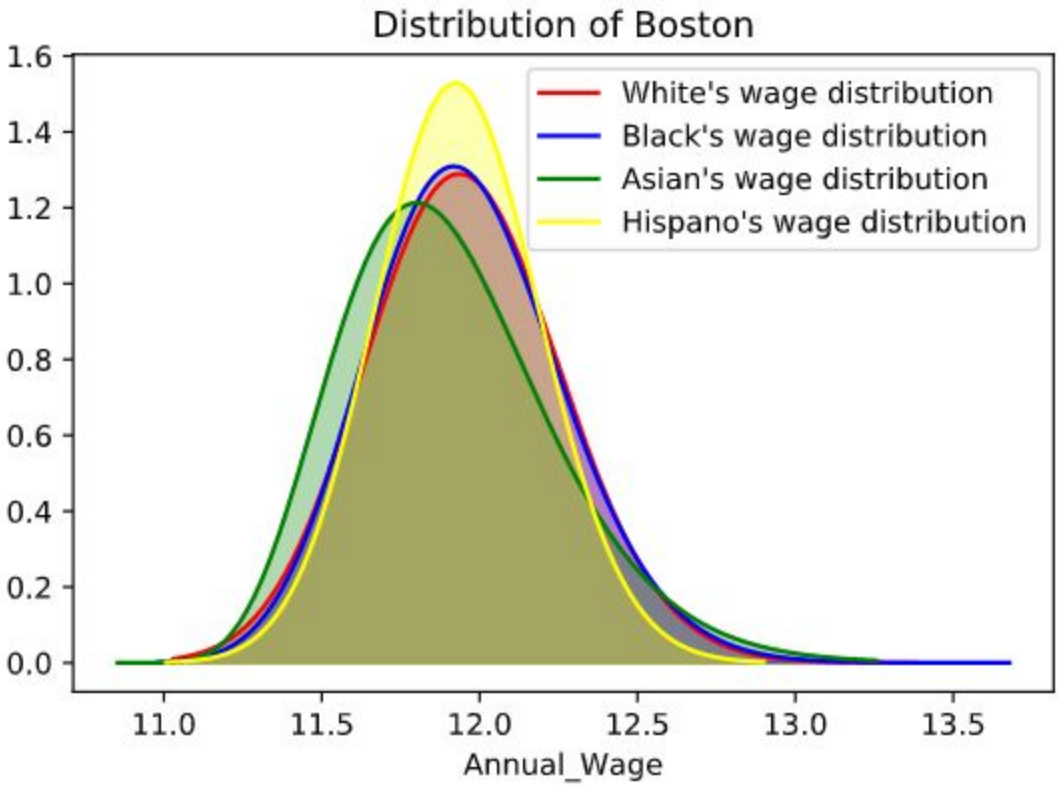


Fig.11(a) Distribution of Wage of Different Races in Boston

In Brockton, according to Fig.11(b), the average annual salary of black is the lowest. The asian have the smallest variance.

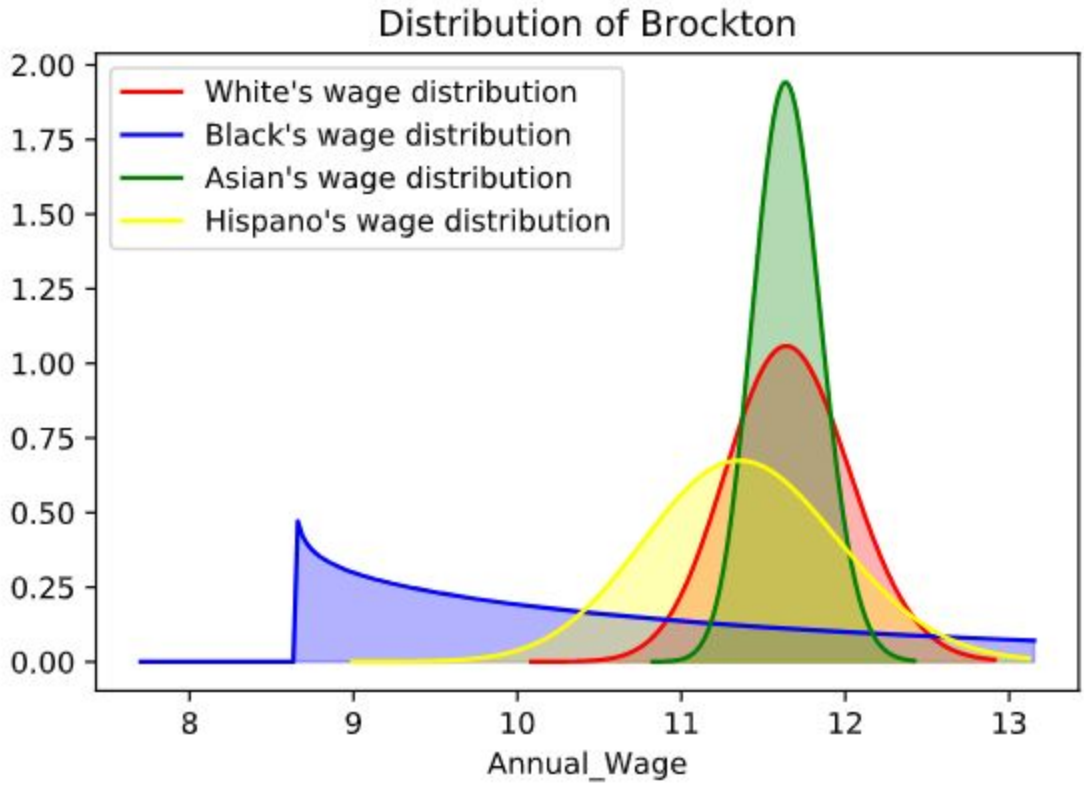


Fig.11(b) Distribution of Wage of Different Races in Brockton

Fig.11(c) shows that in Cambridge, asian have the highest average salary.

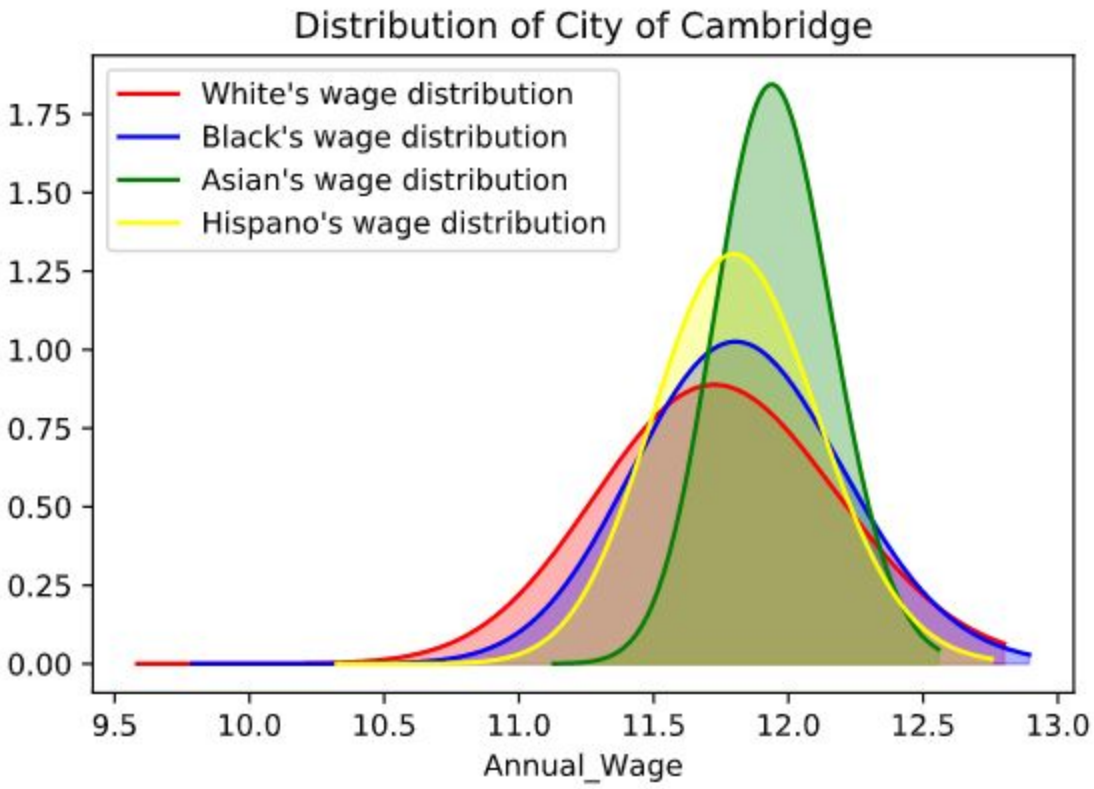


Fig.11(c) Distribution of Wage of Different Races in Cambridge

In the city of Springfield, the hispano have the lowest average salary. The average salary of black, asian and white are very similar.

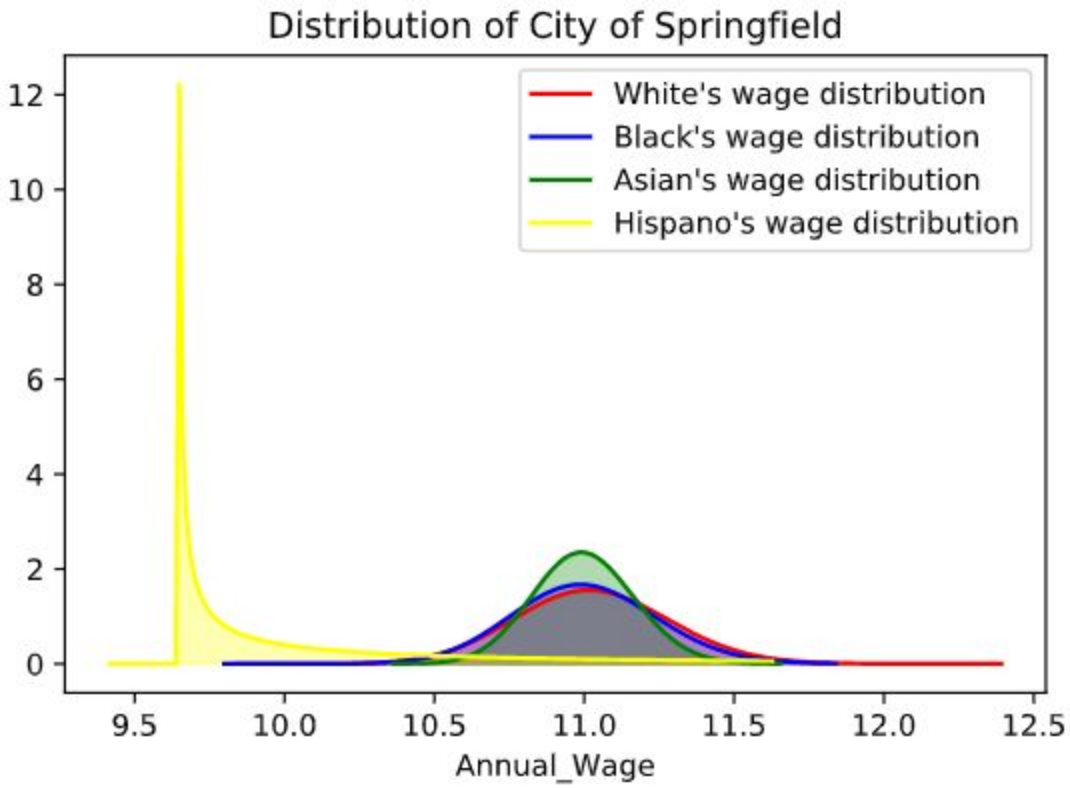


Fig.11(d) Distribution of Wage of Different Races in Springfield

Finally, we do the graph of race vs income. As shown in Fig.12, few white and black people get very high salaries.

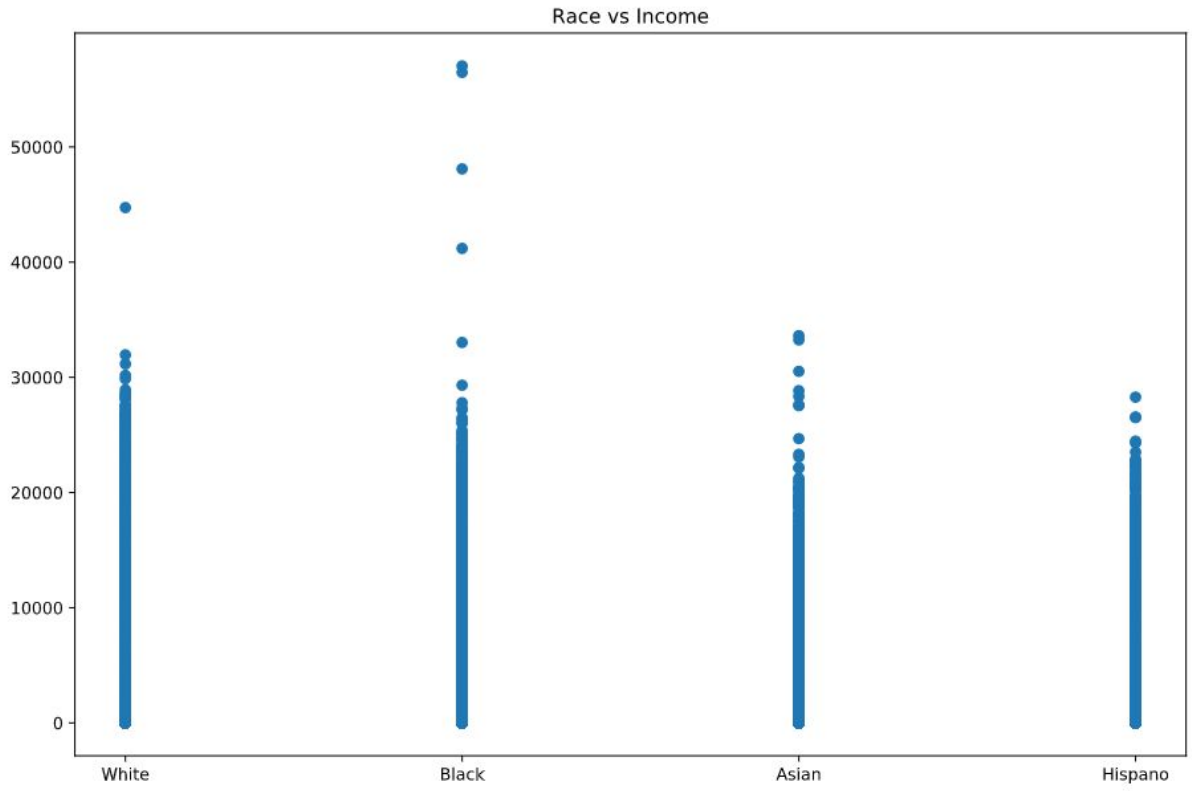


Fig.12 Race vs Income

In Fig.13, we do the rank vs income. We can see that some police officers get really high salaries among different ranks. And some lieutenants and some sergeants get really low salaries.

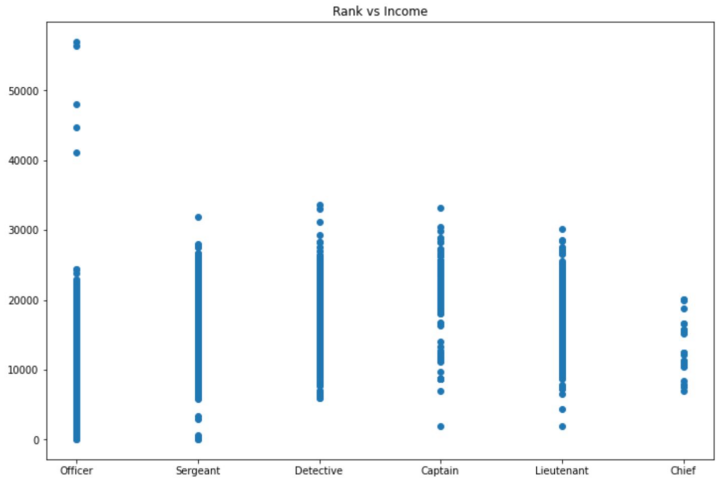


Fig.13 Rank vs Income

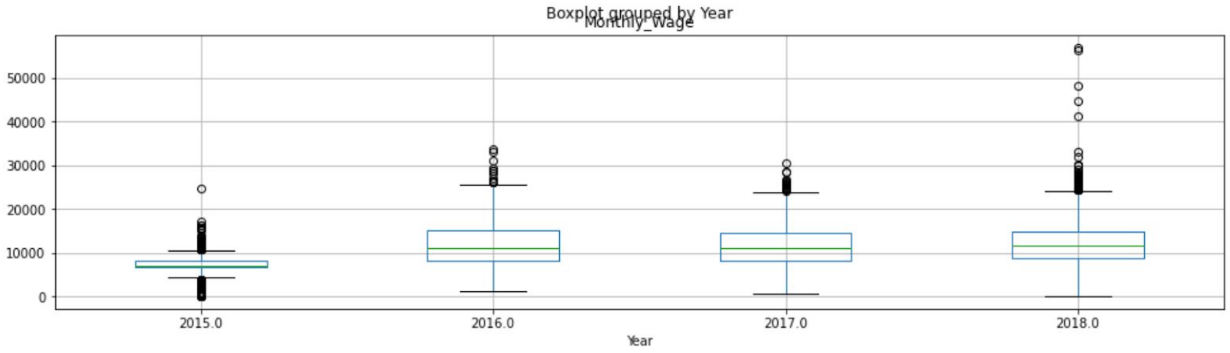


Fig.14 Monthly Wage in Different Years

As shown in Fig.14, in 2018, a few people get the highest average monthly salary among those four years.

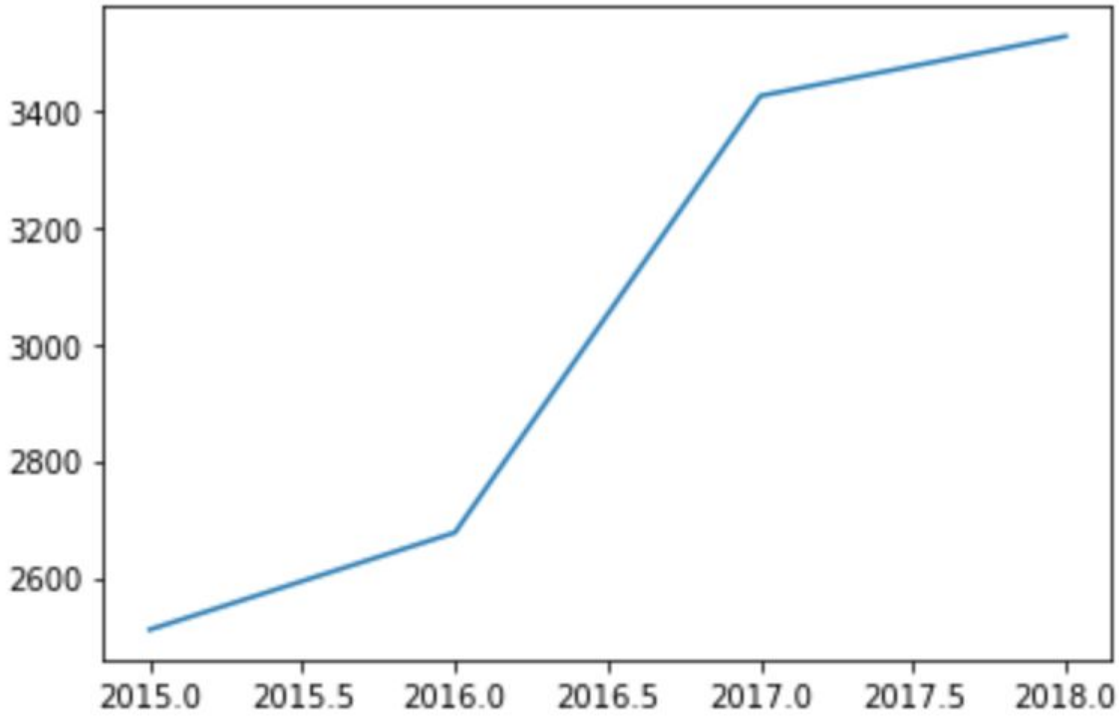


Fig.15 Monthly Wage Plot by Year

Fig.15 shows that the average monthly salary increases from 2015 to 2018.

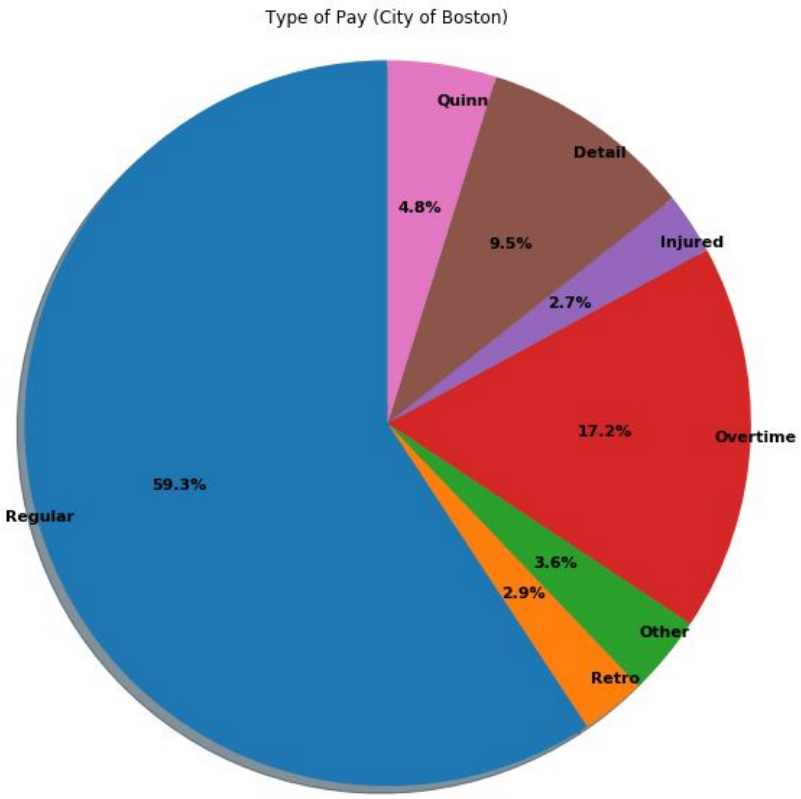


Fig.16 Type of Pay Breakdown

Fig.16 tells us over half is the regular pay. And one over fifth is the overtime pay. Only 3% is the injured pay. So there is not so much injured pay.

Conclusion

- Policemen are more likely to be better paid than their female colleagues
- City of Boston is best employer among the five in terms of salary
- No big difference in the salary between ethnicities/races
- No big gaps observed between different titles in the police system

- There was an increase in the salary over the years, though the scale of police force is getting bigger

Challenges & Space to improve

In a nutshell, we have been faced with three major challenges in total:

1. As we mentioned above, we failed to confirm two same names both belong to one same person. Honestly we were aware of this even when we got started since the available data is really limited. But given the accountability of the police system of the United States and the confidentiality of some data, this result was satisfactory to both us and our client. So we simply skipped this part and changed our target.
2. We were blocked by the targeted website when collecting data. Fortunately we found the useful scrapinghub, which saved us plenty of time.
3. When we are using namsor to identify people's gender and ethnicity, we are faced with a quota limit. In other words, this data service is not free of charge. In order to save our money, we are trying to log in as many accounts as possible so that we could gain more quotas. We are still dealing with this challenge. Hopefully the way we guess will work and we can move forward to the next step.

As for any improvement to our research, a major one we can come up with is that we need more data. For example, now we can only access the relevant data in 2015-2018. As far as we are concerned, such size of the dataset is far from enough if we research the change of wage over periods.

Gratitude

We are grateful to our client and PM for their endeavor and trust. We really appreciate their help when our process was staggering. We thank them for providing us with a really real-world working atmosphere and environment by holding discussions on a regular basis, even during the coronavirus pandemic. So all of us think this is a really meaningful coursework.