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 policeman 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. (**NOTE**: Temporary conclusions, might be changed after deeper research)

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

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, let us 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 a 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. (**NOTE:** need a figure to elaborate, need more details to describe)

Exploratory Data Analysis

To begin with, we analyzed the payrolls mainly in two aspects: gender, and ethnicity. Until we finish the report for our deliverable 2, we have

only looked into the difference of genders in terms of salary in City of Cambridge. We will finish our rest of study in the following weeks. And now we are going to illustrate our thoughts in Cambridge and this will also be our sample to analyze other cities.

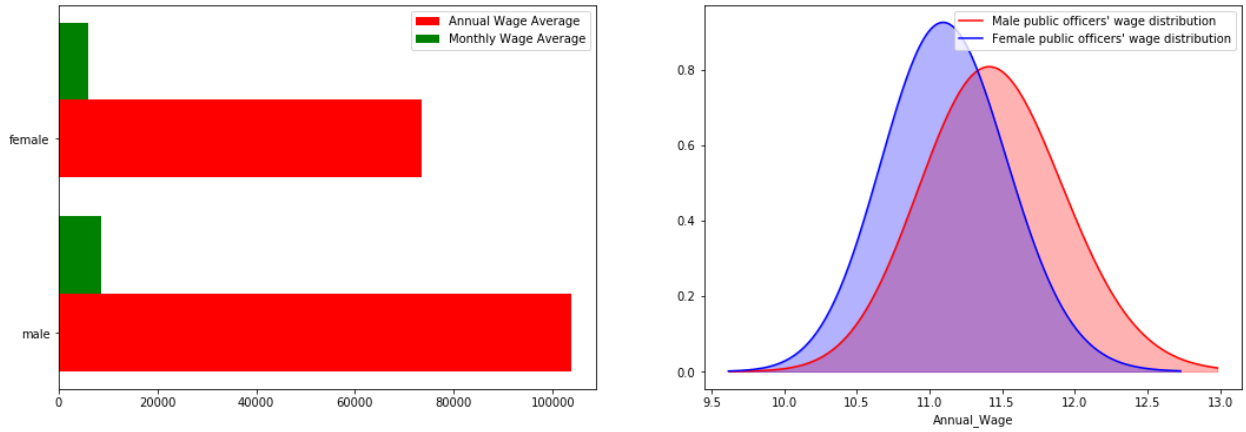

Fig.4 (a) Comparison between Average Salaries of Males and Females;
(b) Distribution of Annual Salaries of Different Genders

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).

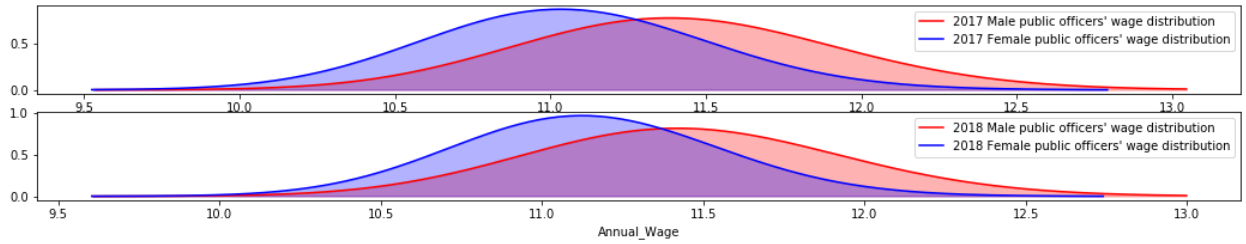

Fig. 5 Annual wages over Year 2017-2018

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.

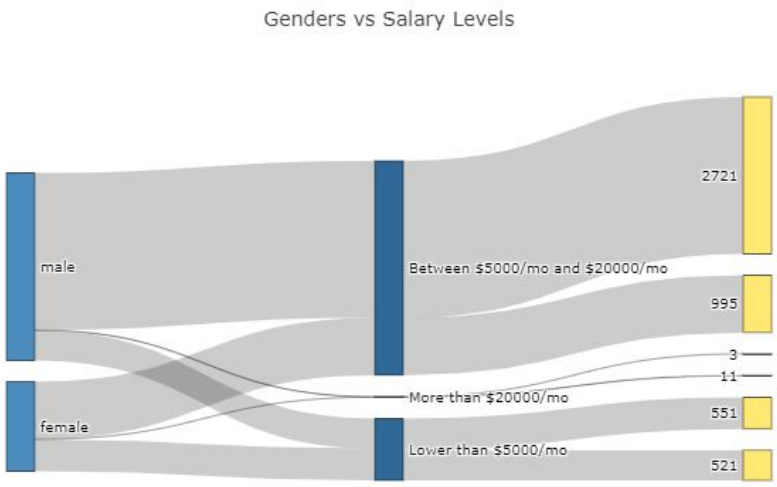

Fig. 6 Genders vs. Salary Ranges

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.

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.

Miscellaneous

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.