San Francisco Salaries

By: Keevey Song, Peng Zhao, Rasha Gargees, Andrew Pistole & James Bain

**A 2-5 paragraph description of what the analysis from the modified script shows.**

The original analysis was very thorough, so seeking modifications or tweaking of any sort given the original analysis was somewhat difficult. This meant that we had to bring in some variables from the outside in order for us to get anything interesting.

At end of the `Modified.Rmd` file, is an analysis that draws in gender as a new variable. In order to do this, we had to split the first name from the rest of the name in the *EmployeeName* column. We then used the `gender` package, which provides a gender prediction based on first name. However, the `gender()` function adds a lot of "fluff" (or what we are calling "fluff") to every single name such as *percent\_male* and *percent\_female* (Okay, these aren't as trivial as we are suggesting, but for this initial analysis we don't use these for reasons we will get to here shortly).

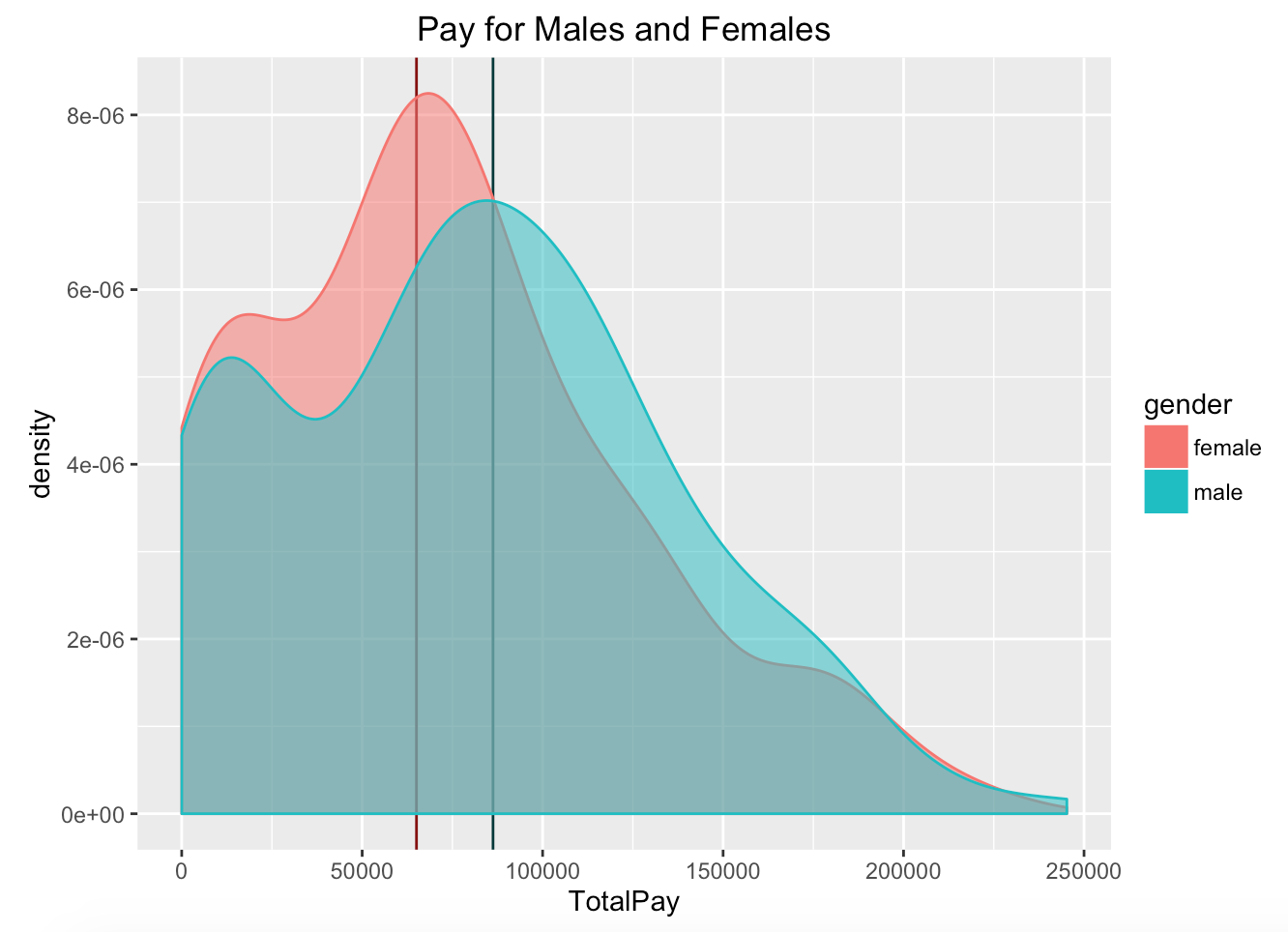
We were only interested in the projected gender of the each individual, so we created a function, `workerFunc()` to extract just the predicted gender of each name passed to it. I attempted to apply this to all 150,000 rows, and albeit small in comparison to some datasets, it took too long for an quick weekend analysis (> 10 hours). We decided to randomly sample the dataset instead, for which we extracted 1500 rows (a lot more time efficient!).We were interested in a simple pay comparison between genders in San Francisco, so I ran a simple summary on *TotalPay* for females and then again for males. Males, at least in the sample mean get paid $21,000 more.

Before we were able to create the plot that shows year vs. total pay, we tried hard to build a plot that could show job field vs. total pay and separated by gender. The script in Kaggle did something similar but used sqlite file. So we tried to convert csv file to sqlite file but it failed. It turned out that sqlite can only save output as sql file, csv file, txt file, or db file but not sqlite file... We had no idea where sqlite file comes from. And then we tried to read db file or sql file in order to use sql query to classify the job categories but we failed again. Fortunately, this plot shows something interesting.

The original dataset only contains salaries in San Francisco. It would be interesting to see if there are any differences between major cities in California, such as San Francisco and Los Angeles. The original data source (<http://transparentcalifornia.com/agencies/salaries/>) contains salaries data of all cities and counties in California. The data of Los Angeles in 2011, 2012, 2013, 2014 were downloaded and processed/merged using the code that has been used to generate the SF data by Kaggle. Then this LA salaries file and the original SF salaries file were combined into a master file called “SF\_LA\_Salaries.csv” manually. Then categorize jobs in “SF\_LA\_Salaries.csv” with function “find\_job\_title” in the original script.

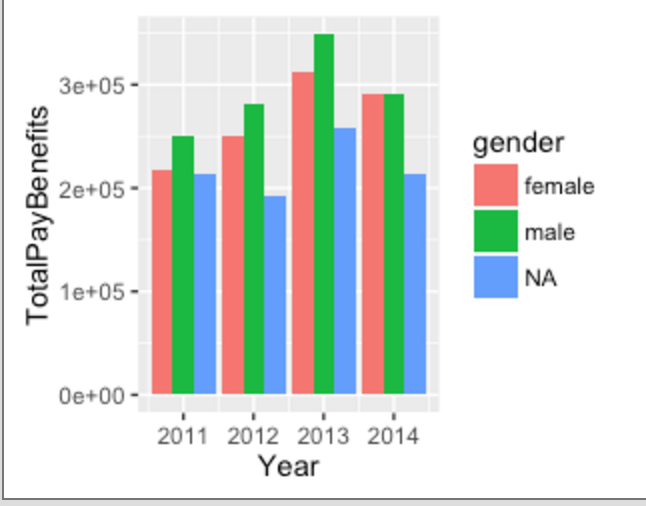
**A 2-5 paragraph description of what the analysis from the two additional visualizations would (does) show**

The following plot depicts the sampled distribution of total pay for males and females. The mean for both genders are also displayed as vertical lines on the graph.

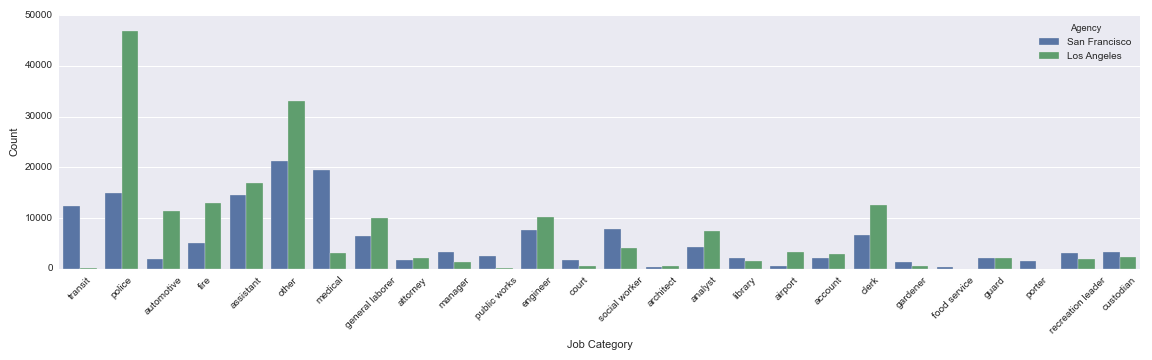


Notice where the frequencies of pay for males and females are concentrated. Males tend to be paid more.

This plot shows the trend of total pay of each gender from 2012 - 2014. Generally speaking, people would think that men would earn more money than women and the data before 2014 displays exactly this. However, in 2014, the pay of women is almost equal to pay of men.



Then compare the count of each category of jobs in SF and LA by the “countplot” function in the “Seaborn” library. The following visualization shows the comparison.



It's very interesting that the number of employees in police category in LA is almost 3 times that of SF. Does this mean LA is a more dangerous place to live compared to SF? According to California Department of Justice's record in 2013 (<https://en.wikipedia.org/wiki/California_locations_by_crime_rate>), the violent crime rates per 1,000 persons of LA and SF are 5.59 and 7.56 respectively, and the property crime rates per 1,000 persons of LA and SF are 15.90 and 13.96. They are pretty close. Then I realized that the population in LA is about 12 times larger than SF and things became reasonable. Many people work in transit field in SF and very few people in LA are in this field although LA has a much larger population. This is because SF has a much better public transportation system and people there almost do not need to drive. The public transportation system can create many positions in SF. Another interesting finding is there are much more people working in medical field in SF than LA given the fact that LA has a much larger population.