“Let’s talk about Equal Pay:” An Analysis of Salaries in San Francisco

Travis Obregon

Institute of Technology Tralee

Dromthacker

Tralee, Ireland

Travis.Obregon@students.ittralee.ie

Abstract

Equal pay has been an issue within society for many years, and women are not able to reap the same benefits as men. Research has shown that females are earning wages that are mere fractions of what men are. Over the years, endless amounts of data have been collected, and it is important to analyze it so that new insights can be made. This project aims to determine whether men and women are treated equally in the workforce based on their wages.

To fully evaluate whether discrimination exists in the workforce, the research conducted was complimented by an implementation of an Apache Hadoop MapReduce job on a collection of salaries in San Francisco between 2011 and 2014. If the difference in average pay between a male and female was more than $5,000 for a job, it is assumed that they are being treated fairly.

The conclusion of the project found there to be no proof of sexual discrimination in the workforce for a given job. It is suggested that future iterations of this project make use of more salary samples that are spread across multiple cities so that the status of equal pay can be analyzed at a global level.

# Introduction

Sexual discrimination has been going on for many years with men always obtaining higher salaries than women for the same work. In the United States of America (USA), the year 1963 saw the formulation of the Equal Pay Act (EPA). The EPA aims to forbid sex-based wage discrimination among men and women who work in the same organization, and operate in jobs that demand essentially equivalent expertise, effort and responsibility under identical working environments (U.S. Equal Employment Opportunity Commission, n.d.). Even though the EPA has been around for over 50 years, women in America are still struggling to make as much money as men. Currently, on average, a woman is making 82 cents for every dollar a man receives (Farber, 2017).

In today’s world, there is an ever-increasing amount of data that needs to be stored. The ability to analyze data so that new insights can be made is crucial for many companies and organizations. This leads to improved decision making, and strategic business actions (SAS, 2017). The research conducted in this paper will be to determine whether there is still discrimination regarding equal pay for men and women. The data set that will be used is salaries for jobs in San Francisco between 2011 and 2014 (Kaggle Inc, 2017).

The ambition of this thesis is to evaluate the salaries amongst men and women to see if sexual discrimination is still a major problem in the 21st century, based on the data collected in San Francisco. To come up with this result, the average salary for a male and female pertaining to a job will be calculated. For a given job, if the male’s average salary is greater than the female’s average salary by $5,000 or more, it will be deemed that sexual discrimination exists for that profession.

# Approach

The path chosen to perform this research is to use Apache Hadoop, commonly referred to as just Hadoop. Hadoop “is a framework that allows for the distributed processing of large data sets across clusters of computers” (Apache Hadoop, 2017). Hadoop provides a distributed file system, known as HDFS, that grants high-throughput access to application data, as well as MapReduce capabilities. MapReduce is a programming model that allows large data sets to be processed and generated, and the idea of this was first introduced by Google (Dean & Ghemawat, 2004). The programming language R will also be used for preprocessing and analysis of the output of the Hadoop MapReduce job.

## Preprocessing

The salaries data set contained 148,654 records in a CSV file, which made it easy to import into RStudio to start preprocessing the data. The following columns are included in the data set:

* Id
* EmployeeName
* JobTitle
* BasePay
* OvertimePay
* OtherPay
* Benefits
* TotalPay
* TotalPayBenefits
* Year
* Notes
* Agency
* Status

The table below shows the columns that were removed from the data set immediately and the reasons for doing so.

Table 1 Columns that were immediately removed

|  |  |
| --- | --- |
| Column | Reason |
| Id | It has no importance in calculating the average salaries. |
| BasePay, OvertimePay, OtherPay, Benefits, TotalPay | The only column needed is TotalPayBenefits, which the sum of all these columns produces. |
| Notes | The value was missing for each row. |
| Agency | The value was “San Francisco” for each row. This is meaningless because it is known that all the salaries in the data set come from San Francisco. |
| Status | This field represented whether the job was full time or part time. However, it was only available for records that had a Year of 2014. |

The columns in the data set were also modified to comply with Google’s coding standards for R (Google, n.d.). In addition to the removal of columns, rows were also removed if the total pay for the employee was not greater than $10,000. The reasoning behind doing this is that some records were shown to have negative values for total pay. After this was done, the next step was to make sure that the values of the job title are uppercase because these values will represent the keys in the MapReduce job. As of right now, there is no knowledge of the gender of each employee. To accomplish this, an R package called “gender” was used. A temporary column to house the employee’s first name was created because the function to calculate the gender requires a single name. The gender package can calculate gender based on years ranging from 1932 to 2012. The year of each record in the data set was used in the gender function except for the years 2013 and 2014, which had to resort to 2012 due to the packages limitations. The code snippet below illustrates how each employee’s gender was determined.

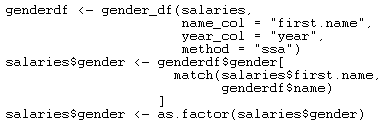


Figure 1 Determine an employee's gender

The gender of some employee’s names was unable to be resolved. Therefore, these records were removed since knowing the gender is required to answer the motive of this project. The preprocessing was now complete and some general clean up was done to remove the employee name, year, and first name columns because their use is no longer needed. A summary of the data to this point shows that there are 115,597 records left with 48,671 females, and 66,926 males. The boxplot that follows depicts the salaries based on gender.

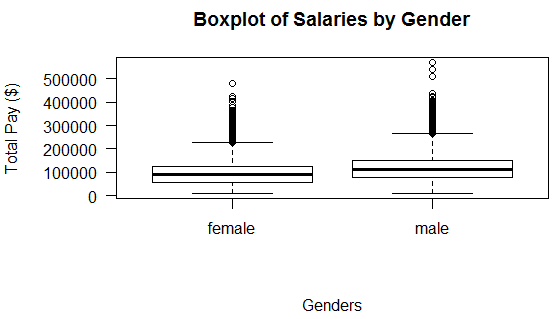


Figure 2 Boxplot of salaries by gender

The mean salary for all male’s is $115,029.70, in comparison to only $96,735.46 for females. A histogram of the male salaries is represented next.

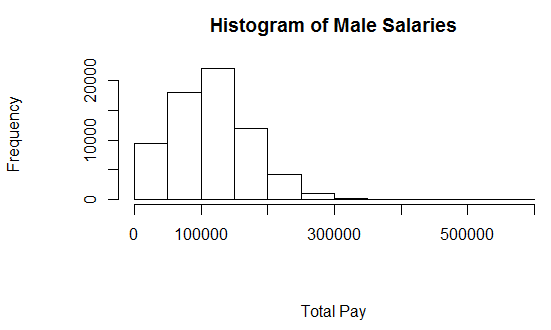


Figure Histogram of male salaries between 2011 and 2014

Female salaries are expressed with the following histogram.

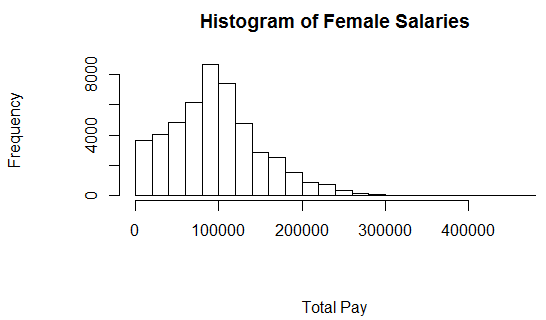


Figure Histogram of female salaries between 2011 and 2014

Finally, after preprocessing was done, the data was written to a text file, which will be passed on to Hadoop.

## Hadoop MapReduce

A MapReduce job typically splits the input data into individual chunks, which are processed by the tasks of the mapper in a fully parallel manner. The MapReduce framework’s next step is to sort the output of these tasks and pass the data along to the reduce tasks. In most situations, the input and the output of the job are stored in a file-system, and this would be HDFS. The framework is responsible for scheduling tasks, monitoring them and re-executing the failed tasks (Apache Hadoop, 2013).

To gain a better understanding of Hadoop and MapReduce, a free video series published by Udacity was watched (Udacity, Inc., n.d.). The instructors of the course work at Cloudera, which is an industry leader in the world of data science. They offer a virtual machine that is running Hadoop and related technologies for big data. The virtual machine is running in pseudo-distributed mode, which means that is a full cluster, but it is just running on one machine. Additionally, the virtual machine supports Hadoop streaming. This allows any programming language to be used for the mapper and reducer scripts, and not just Java.

Python was used for the scripts of the mapper and reducer. The mapper first read the text file from standard input and emitted records that had the job title as a key, and the total pay and gender as part of the value. These records were then sent to the shuffle and sort phase in the MapReduce job where the records were placed in alphabetical order by their key. Next, the reducer was responsible for turning the results into a CSV file with the following headings:

* Job Title – Represents the key for each record
* Male Average Pay – The average salary for a male for a given job
* Female Average Pay – The average salary for a female for a given job
* Discrimination – Set to True if the male average pay for the current job is greater than the female average pay by $5,000 or more, and false otherwise.

Rows were only added to the CSV file if there were both salaries for a male and female for a given job position. This was required so that a calculation can be made to see if sexual discrimination existed for a job. This could not be done if a job had only salaries for one gender. To run the Hadoop MapReduce job, the following command needed to be run inside of the terminal.

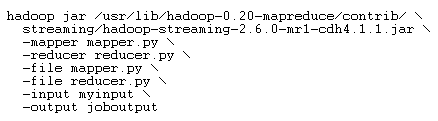


Figure Command used to run a Hadoop MapReduce job

Hadoop first needs a location of a jar file so it can being execution of the job. Next, the files for the mapper and reducer scripts must be supplied. These scripts must also be passed along to the “-file” option so the MapReduce job knows which files it needs to get access to. Then, the input directory needs to be specified. This is where the data is stored in HDFS. In this case, this is where the salaries text file was located. Finally, the output option dictates what directory the MapReduce job should place the results when it is finished. The above command can be cumbersome to write. Fortunately, the virtual machine that was used came with an alias that allows a much shorter command to be written.



Figure Alias command used to run a Hadoop MapReduce job

Now, only the mapper script, reducer script, input directory, and output directory need to be specified as arguments. After a MapReduce job is run, the output directory is populated with 3 items. They are “\_SUCCESS”, which is an empty file that simply states that the job was successful, a directory called “\_logs”, which contains some information about what happened while the job was running, and a file called “part-00000” that contains the results from the reducer script.

# Discussion

In April of 2013, according to the National Partnership for Women & Families, women who have a full-time job earn an average of $52,301 per year, while men who have a full-time job earn an average of $62,269 per year. This amounts to a yearly gap of $9,968, which means women are paid 84 cents for every dollar given to men. For the entire USA, this number is decreased to just 77 cents. Digging further, African American women are compensated 64 cents, and Latinas are compensated 55 cents (National Partnership for Women & Families, 2013). If the salary gap were removed in San Francisco, women would have more funds to have one of the following:

* 72 more weeks of food
* 4 more months of mortgage and utilities payments
* 7 more months of rent
* 2,462 additional gallons of gas

Looking at the data from the National Partnership for Women & Families, there is no way of determining what jobs the women and men work. For instance, if the sampled jobs for women were cashiers, while the sampled jobs for men were store managers, then it would be more likely to expect there to be a difference between the average salaries. In the sample data set that this paper utilizes, the average salary for a male in 2013 in San Francisco is $124,834.80, compared to $105,124 for females. This leads to the result of pay gap of $19,710.81, which is more than the double the amount calculated by the National Partnership for Women & Families. It is hard to conclude on whether, the National Partnership for Women & Families had a larger data set than the one used for this research. In 2013, this research had 29,172 jobs. Of these jobs, 16,843 were male, and 12,329 were female. In the year 2013 alone, it is evident that men have a significant increase in wages over females. There are many possible reasons as to why the gender gap is so difficult to close. According to Olivia Mitchell, three of the most compelling contributors are the penalty women face for becoming mothers, the bias women face by their employers, and the lack of negotiating skills that women have (Farber, 2017). If an individual does not negotiate their first salary, they are subject to losing at least $500,000 by the age of 60 (Babcock & Sara, n.d.). Women may even be denied jobs because of their gender (Berger, 1971). In California, there was a law that prohibits women from lifting 10 pounds or more on stairways which increase more than 5 feet from their bases. This law was put in place to “protect” women. However, it only hindered them from obtaining better opportunities to meet the needs of their families. On average, women are considerably less experienced than men, which leads them into lower-paying jobs and industries (Blau & Kahn, 2000). Another factor as to why the gender pay gap still exists, is that employers are unwilling to disclose the gender pay gap reviews that are formally conducted on their company. Approximately 3.7% of organizations will report this data internally, and only 1.3% will do so externally (Adams, et al., 2010). Although there is plenty of reports and journals stating the existence of sexual discrimination in the workforce, the salaries that were investigated in San Francisco showed there to only be a couple thousand dollars’ difference between the average salary of a male and the average salary of a female. It is not known what years of experience each employee had in the data set. This would play a pivotal role in the amount of money earned by everyone regardless of gender. For example, there may have been only two employees in the salaries data who are Accountants, one male and one female. If the male earned $50,000 and the female earned $30,000, but the male had 10 years more experience, then the expectation is that the male should earn significantly more. However, because experience was not known, a true understanding of sexual discrimination in the workforce could not be made. Despite experience, it could also be the case that the two Accountants work at separate firms, and one firm is more successful than the other. Therefore, the successful firm can offer a higher salary to its employees.

# Findings

The CSV that was created as a result of the MapReduce job was loaded into RStudio for further analytics. After some initial sanitizing of the data, such as trimming white space from columns and setting the discrimination column to be a factor, the summary function was used to get an overview of the data.

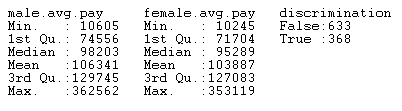


Figure 7 Results of the summary function on the output from the MapReduce job

At a first glance, the figure above illustrates that 633 out of 1001 jobs in the data set show no sexual discrimination in the salaries made by employees. The values of the minimum, 1st quartile, median, mean, and 3rd quartile for males and females are all within $5,000 of each other. However, the maximum male average pay and the maximum female average pay has a difference of almost $10,000. Hence, there appears to be a problem with giving woman high salaries. The correlation between a male’s average pay and a female’s average pay is 0.8824544. This signifies a positive linear relationship that is very strong (Rumsey, n.d.). The scatterplot below represents the correlation between the aforementioned variables.

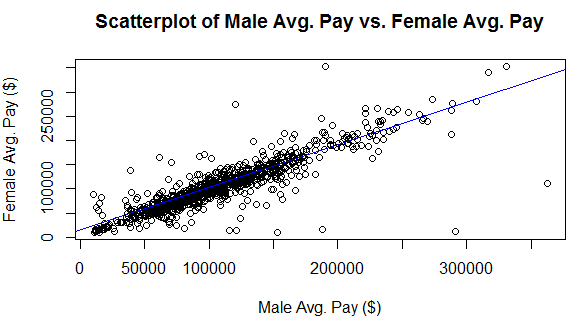


Figure 8 Scatterplot of male avg. pay vs. female avg. pay

Out of the 1001 jobs, 414 of them pay females more than males, 586 of them pay males more than females, and only 1 of them pay males and females equally. A final piece of analysis was done to find the difference in the average total pay between males and females for each job. For example, for an Accountant, if the male average pay was $40,000 and the female average pay was $30,000, then the total pay difference would be $10,000. To achieve this, a new column was added to the data set to hold the difference for each job. The sum of this column equated to $2,456,511.89. First thoughts indicated that this was a magnificent display of sexual discrimination. However, the mean was calculated, and it turns out that on average, each job is paying males $2454.06 more than females. This value falls within the $5,000 threshold that was defined earlier.

# Conclusion

Being a woman today can be very challenging, especially when it is hard to obtain the same affordances that men are given. Research shows that women are being paid less than men for equal work. As aforementioned, women are being paid 82 cents for every dollar that men make. This is a sizable improvement from what it was in 1963, where women made a mere 54 cents to every dollar men did. Even though the pay gap is decreasing, it will still take 70 years to fully abolish the gender wage gap according to a recent study if the current upward trend remains (Anon., 2016). Nonetheless, this research paper could conclude that there are some places in the world that are closer to equal pay than others. This thesis defined that there could be no accusation of sexual discrimination in the workplace if the average salary between a man and woman for the same job was less than $5,000. Since the findings informed that the average difference is $2,454.06, it is safe to say that there is no evidence of gender bias in the workforce in San Francisco.

This data analysis only explored salaries in San Francisco. In the future, data be collected to evaluate more cities, states, and countries so that equal pay can be analyzed at a more higher level.

# References

Adams, L., Gore, K. & Shury, J., 2010. *Gender pay gap reporting survey 2009,* Manchester: Equality and Human Rights Commission Research Report Series .

Anon., 2016. *Large gender gaps remain across global labour market, UN agency reports.* [Online]   
Available at: http://www.un.org/apps/news/story.asp?NewsID=53383#.WQe50GhtnIV  
[Accessed 1 May 2017].

Apache Hadoop, 2013. *MapReduce Tutorial.* [Online]   
Available at: https://hadoop.apache.org/docs/r1.2.1/mapred\_tutorial.html#Inputs+and+Outputs  
[Accessed 2 May 2017].

Apache Hadoop, 2017. *What Is Apache Hadoop?.* [Online]   
Available at: http://hadoop.apache.org  
[Accessed 1 May 2017].

Babcock, L. & Sara, L., n.d. *Negotiation and the Gender Divide.* [Online]   
Available at: http://www.womendontask.com/stats.html  
[Accessed 2 May 2017].

Berger, C. G., 1971. *Equal Pay, Equal Employment Opportunity and Equal Enforcement of the Law for Women,* s.l.: Valparaiso University.

Blau, F. D. & Kahn, L. M., 2000. *GENDER DIFFERENCES IN PAY,* Cambridge: NATIONAL BUREAU OF ECONOMIC RESEARCH.

Dean, J. & Ghemawat, S., 2004. *MapReduce: Simplified Data Processing on Large Clusters,* San Francisco: Google.

Farber, M., 2017. *3 Reasons Why the Gender Pay Gap Still Exists.* [Online]   
Available at: http://fortune.com/2017/04/03/equal-pay-day-2017-wage-gap/  
[Accessed 1 May 2017].

Google, n.d.. *Google's R Style Guide.* [Online]   
Available at: https://google.github.io/styleguide/Rguide.xml  
[Accessed 1 May 2017].

Kaggle Inc, 2017. *SF Salaries.* [Online]   
Available at: https://www.kaggle.com/kaggle/sf-salaries  
[Accessed 1 May 2017].

National Partnership for Women & Families, 2013. *San Francisco Women and the Wage Gap.* [Online]   
Available at: http://go.nationalpartnership.org/site/DocServer/Wage\_Gap\_SanFrancisco.pdf  
[Accessed 2 May 2017].

Rumsey, D. J., n.d.. *HOW TO INTERPRET A CORRELATION COEFFICIENT R.* [Online]   
Available at: http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/  
[Accessed 1 May 2017].

SAS, 2017. *What Is Big Data?.* [Online]   
Available at: https://www.sas.com/en\_ie/insights/big-data/what-is-big-data.html  
[Accessed 1 May 2017].

U.S. Equal Employment Opportunity Commission, n.d.. *The Equal Pay Act of 1963.* [Online]   
Available at: https://www.eeoc.gov/laws/statutes/epa.cfm  
[Accessed 1 May 2017].

Udacity, Inc., n.d. *Intro to Hadoop and MapReduce.* [Online]   
Available at: https://www.udacity.com/course/intro-to-hadoop-and-mapreduce--ud617  
[Accessed 1 May 2017].