Analysis of Salary Variation in San Francisco

Department Of Information Systems, California State University Los Angeles
CIS 3200 Data Processing and Analytics - Professor Jongwook Woo, Phd
Group 3 | Authors: Julia Buano, Adony A. Oliva, Ismhael K. Asturias, Samuel Maingi, Nicholas Cun

Abstract

This group project examines compensation patterns among San Francisco city employees using the "SF Salaries" dataset, which includes salary records from 2011 to 2014. With a focus on overtime patterns and role-specific pay inequalities, our analysis focuses on how total compensation differs across departments and job titles. Our goal is to identify potential inefficiencies or injustices in the distribution of public funds by comparing pay systems over several years.

To achieve this, ElasticSearch was used for large-scale data storage and querying, as well as data cleansing, statistical analysis, and visualization. Our research identifies patterns in public-sector salaries, occupations that heavily rely on overtime, and department-level spending trends. Issues concerning personnel management, budgeting objectives, and fairness in public compensation systems can benefit from the knowledge gathered from this project.

1. Introduction

Understanding how public-sector salaries are allocated is crucial for efficiency and accountability, as they make up a sizable amount of local budgets. From 2011 to 2014, the "SF Salaries" dataset offers comprehensive payroll records for city employees, including basic pay, overtime, other pay, benefits, and total compensation for thousands of workers in various departments and job titles. The topic was chosen for the purpose to look into trends over time, overtime dependence, and possible wage inequities. We aim to determine whether some departments or jobs are overpaid in comparison to others by looking at these trends, which can help shape public policy and financial decisions. With precise aggregations and insights that a normal spreadsheet cannot achieve at this scale, ElasticSearch helps us to manage and analyze this dataset effectively.

2. Related Work

Several studies have investigated public-sector compensation patterns, pay disparities, and factors influencing total compensation. For example, Krueger (1988) used longitudinal data to compare private-sector workers who lost their jobs with those who took government jobs, finding that while federal employees enjoyed a 10–25% wage premium over their private-sector counterparts, state and local workers' wages were nearly similar or lower than those in the private sector. The study also highlighted parts like job security and benefits, by examining application "queues" for public jobs, suggesting that employee preference for stability and benefits might compensate for lower base wages. Another study, proven by Gittleman & Pierce (2012), extended this comparison to include employer-paid benefits and other compensation components, using nationally representative survey data. Their work found that, after accounting for worker skill and job characteristics, total compensation in the public sector is higher than in the private sector. The results highlight how crucial it is to take whole compensation into account when assessing public-sector pay because benefits like healthcare, retirement, and leave have a big impact on total compensation.

3. Our Work

By analyzing the SF Salaries dataset from 2011 to 2014 using ElasticSearch, our work expands on these insights by allowing for a detailed examination of compensation trends related to departments and jobs, overtime dependence, and yearly pay growth. In contrast to past studies that focus on survey data or aggregated trends, the method being used leverages actual payroll records, allowing accurate evaluation of total compensation components for thousands of employees. Our analysis is a more insightful contribution to the basic research of Krueger and Gittleman & Pierce because it offers extensive, data-driven information on pay discrepancies, operational challenges, and resource allocation within San Francisco's municipal workforce.

4. Background/Existing Work

It takes a combination of statistical approaches, structured data processing, and visualization tools to analyze municipal payroll data. Thousands of employee records were included in the SF Salaries dataset from 2011 to 2014. These records include the employee's name, department, job title, base pay, overtime pay, and total income, including benefits. Effective handling of this dataset requires both normalization to guarantee record comparability and data cleaning to eliminate duplicates, null entries, and outliers.

The foundation of our data processing process is ElasticSearch, an online analytics platform. Efficient aggregation, filtering, and search operations across several attributes, including department, job title, and year, are made possible by its indexing and query capabilities. This makes it possible to efficiently perform trend analysis, compensation distribution visualization, and descriptive statistical calculations that would be difficult to perform using standard spreadsheet applications because of the quantity of the datasets.

Our approach expands on findings from earlier research. The significance of looking at total compensation rather than just salaries is illustrated by Krueger (1988) and Gittleman & Pierce (2012). Our work establishes a framework for detailed analysis of municipal employee compensation, overtime use, and departmental inequities that previous studies were unable to do by combining the insights with scalable data-processing methods and city-level payroll data.

5. Working Process

ElasticSearch was used for the analysis of the SF Salaries data, and a specific procedure was followed. This approach produced visual reporting of workforce distribution, guaranteed data quality, and supported complex statistical analysis. In order to overcome the shortcomings of the text-based fields, the approach started with the necessary collection of the publicly available payroll records and involved crucial data cleansing through Dev Tools. To accurately carry out the regression modeling and produce data visualizations, specific procedures were followed in a specific order.

5.1 Data Type and Specifications

The dataset used in this analysis includes details of average salaries across multiple job titles and departments. The dataset includes the following key attributes:

- Job Information: Job Title/Department
- Employee Details: Employee ID, Employee Name
- Salary Details: Base Pay, Overtime Pay, Other Pay, Benefits
- The total size of the dataset is approximately 14 MB, comprising close to 150,000 rows.

5.2 Data Gathering

Finding and uploading the dataset required for the municipal salary research was the first step. The first step in the process was to get the salary dataset from the provided website. After downloading the dataset, it was uploaded to ElasticSearch within the machine learning/data visualizer. It was advised to enter this path into the search field in order to find it right away, rather than having to look for it by scrolling. Selecting the .csv file for the upload process was important.

5.3 Data Cleansing and Index Preparation

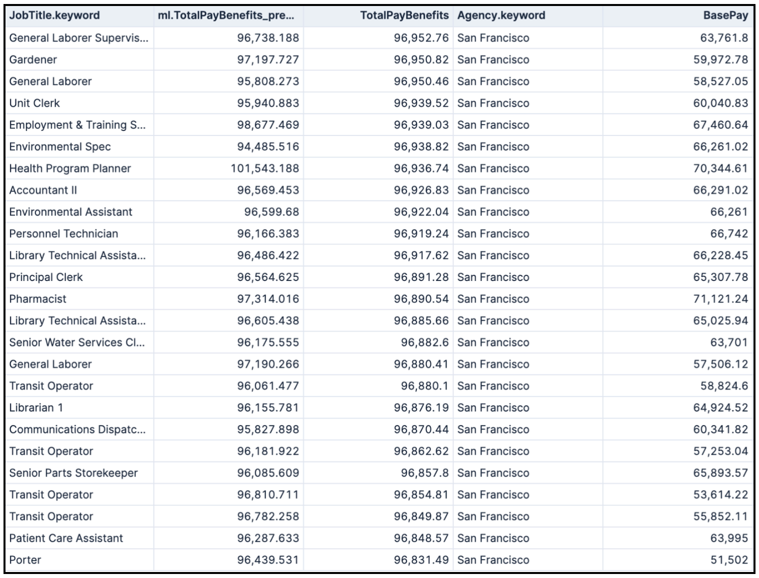
Using Dev Tools to execute two functions was a big part of data preparation, allowing for the incorporation of the job title into the ML model. This preparation was required since using the original JobTitle in its text form would generate an error in the machine learning model, requiring the production of a suitable keyword alternative. The mappings were defined using the PUT sf_salaries_fixed command, which indicated that the JobTitle field would have a sub-field called keyword. The data was then moved from the source index (sf_salaries) to the destination index (sf_salaries_fixed) by running the POST _reindex command. The field JobTitle.keyword was successfully established as a result, which meant it could be used in the next phase.

5.4 Creating and Running Regression Models

After successfully preparing the data and creating the JobTitle.keyword field, a statistical framework was set up to determine the factors that influence salary. This was achieved by navigating to Machine Learning / Data Frame Analytics Jobs and building a data frame analytics regression model. TotalPayBenefits was chosen as the model's key dependent variable, which fully matched the project's objective of examining entire compensation. Many fields, such as ID, EmployeeName, and Benefits.keyword, were specifically unchecked from the "is included" section in order to streamline the ML model and make it safer because it has fewer unneeded variables to work with. As the model relies exclusively on the prepared field, JobTitle.keyword, the original JobTitle had to be left unchecked because failing to do so would result in an error. Finally, the percentage of data utilized to train the predictive model before analysis was determined by setting the model's training percentage to 90.

5.5 Model Results

The job was started when all variables and parameters were configured, which led to waiting patiently for every phase to be completed. The model processed 90% of the training data and determined the statistical weights of each input variable during this execution period. The quantitative output was made available through the Data Frame Analytics/Results Explorer, which allowed users to view the ML model's findings after all phases had been successfully completed and the model had stopped operating. As shown in Figure 1, Results Explorer displays the impact of variables, such as JobTitle.keyword, on the expected outcome, TotalPayBenefits, and provides analytical measures related to it.



JobTitle.keyword	ml.TotalPayBenefits_pre...	TotalPayBenefits	Agency.keyword	BasePay
General Laborer Supervis...	96,738.188	96,952.76	San Francisco	63,761.8
Gardener	97,197.727	96,950.82	San Francisco	59,972.78
General Laborer	95,808.273	96,950.46	San Francisco	58,527.05
Unit Clerk	95,940.883	96,939.52	San Francisco	60,040.83
Employment & Training S...	98,677.469	96,939.03	San Francisco	67,460.64
Environmental Spec	94,485.516	96,938.82	San Francisco	66,261.02
Health Program Planner	101,543.188	96,936.74	San Francisco	70,344.61
Accountant II	96,569.453	96,926.83	San Francisco	66,291.02
Environmental Assistant	96,599.68	96,922.04	San Francisco	66,261
Personnel Technician	96,166.383	96,919.24	San Francisco	66,742
Library Technical Assista...	96,486.422	96,917.62	San Francisco	66,228.45
Principal Clerk	96,564.625	96,891.28	San Francisco	65,307.78
Pharmacist	97,314.016	96,890.54	San Francisco	71,121.24
Library Technical Assista...	96,605.438	96,885.66	San Francisco	65,025.94
Senior Water Services Cl...	96,175.555	96,882.6	San Francisco	63,701
General Laborer	97,190.266	96,880.41	San Francisco	57,506.12
Transit Operator	96,061.477	96,880.1	San Francisco	58,824.6
Librarian 1	96,155.781	96,876.19	San Francisco	64,924.52
Communications Dispatc...	95,827.898	96,870.44	San Francisco	60,341.82
Transit Operator	96,181.922	96,862.62	San Francisco	57,253.04
Senior Parts Storekeeper	96,085.609	96,857.8	San Francisco	65,893.57
Transit Operator	96,810.711	96,854.81	San Francisco	53,614.22
Transit Operator	96,782.258	96,849.87	San Francisco	55,852.11
Patient Care Assistant	96,287.633	96,848.57	San Francisco	63,995
Porter	96,439.531	96,831.49	San Francisco	51,502

Figure 1. Effect on Total Pay Benefits

In order to meet the objective of providing data-driven information on pay discrepancies and the distribution of resources, the data obtained from this research determined the extent to which certain job titles and departments drive the overall salary.

5.6 Visualization For Job Distribution

The final task involved constructing a comprehensive Treemap visualization to illustrate the distribution and frequency of the provided job titles across the municipal workforce. The process began by selecting "Home" from the hamburger menu of ElasticSearch, then navigating to Analytics > Dashboard. We then select Create Dashboard first, followed by Create Visualization. The visualization used the prepared field by clicking the "+" icon that appears while hovering over "JobTitle.keyword". The axis configuration "Top 5 values of jobTitle.keyword" was selected. To create a full and detailed visual map of the workforce, the "Number of values" was increased to 1000, much higher than the default selection. At the same time, the title was changed to "Top 1000 Job Titles in San Francisco". The visualization type was validated by selecting "Treemap" from the available options, which is best suited for displaying hierarchical data or distributions of frequencies. Eventually, the treemap is done, which leads to pressing "Save and return" and then "Save" in the top right corner, naming the new dashboard “Top 1000 Jobs in SF: Treemap”. Based on what can be seen in Figure 2, this approach resulted in a complete and scalable visual representation of worker distribution, which can be used to examine departmental concentrations and the corresponding popularity of municipal jobs.

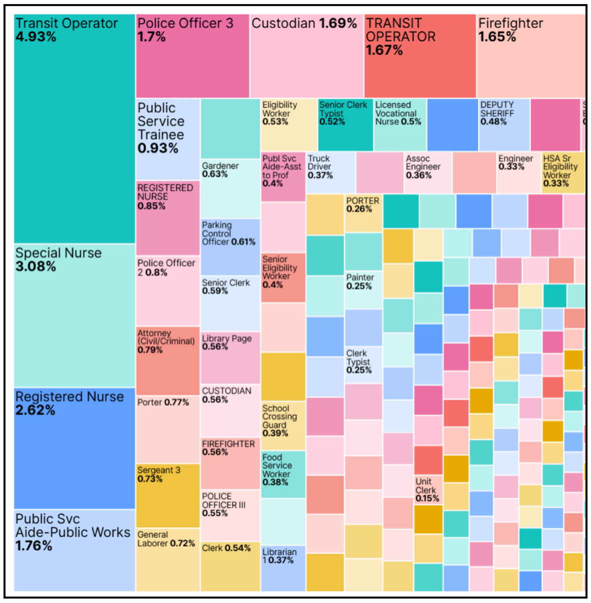


Figure 2. Percentage of Common Jobs Accounted for in Wages

6. Conclusion

The analysis of San Francisco city employee compensation from 2011 to 2014 involved the use of ElasticSearch to manage, query, cleanse, and analyze the dataset, which contained thousands of rows with job, employee, and salary details. The data preparation phase included steps, like creating a JobTitle.keyword field and reindexing to make sure that there is compatibility with machine learning models. We then built regression models to examine the factors influencing total compensation (TotalPayBenefits), selecting variables, and excluding unnecessary fields to optimize model performance. The regression results revealed fluctuations in salary across the years and highlighted inequalities based on position and gender, providing quantitative insights into pay patterns across departments and roles. Using the Data Frame Analytics Results Explorer, we measured the impact of each variable, demonstrating how specific job titles contribute to overall compensation trends and supporting the aim of uncovering pay disparities.

In addition to predictive modeling, we created a treemap visualization to represent the distribution of job titles across the municipal workforce, showing that Transit Operator was the most common role, representing 4.93% of employees. This combination of visualizations and regression analysis provided both qualitative and quantitative understanding of public-sector compensation. Through this work, we learned not only technical skills but also gained insights into the real-world implications of workforce composition and pay inequality. For example, disparities in pay based on position and gender highlight the need for more equitable compensation policies, and our findings suggest potential fixes, such as adjusting pay scales, managing overtime distribution, and implementing transparent salary review processes. Throughout it all, our analysis presents the value of combining data visualization with predictive modeling to inform evidence-based decisions in public administration and provides a lens for addressing equality and efficiency in municipal salary systems across departments.

References

Elastic
https://www.elastic.co/guide/en/kibana/current/reporting-getting-started.html

Gittleman, M., & Pierce, B. (2012). Compensation for state and local government workers. Journal of Economic Perspectives, 26(1), 217–242. https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.26.1.217

Kaggle Data set
https://www.kaggle.com/datasets/kaggle/sf-salaries

Krueger, A.B. (1988). Are Public Sector Workers Paid More Than Their Alternative Wage? Evidence From Longitudinal Data and Job Queues. *NBER.* https://www.nber.org/system/files/working_papers/w2500/w2500.pdf.