**Analyzing Public Employee Salaries: Insights from the Salaries Dataset**

**Executive Summary**

This project seeks to analyze and derive reflections from the public employee salary dataset by base pay, overtime pay, benefits, and total compensation regarding job titles and years. The analysis is guided by data cleaning and transformation efforts through some key patterns and the development of predictive models to give insight into the drivers of total pay.

**Objectives**:

1. ***Data Cleaning***: Remove missing values and inconsistent value entries with their corresponding data types and prepare datasets for analysis.
2. ***Data Transformation***: Extract meaningful features and prepare the data for statistical and machine learning algorithms.
3. ***Data Visualization***: Using relevant Python libraries, visualize salary trends, distribution of pay across occupations, and the effect of benefits on total compensation package.
4. ***Data Modeling***: Build predictive models capable of detailing those key factors contributing to total pay and how these factors are likely to change over time. .

**Table of Contents**

[Introduction 4](#_Toc182751588)

[Dataset 4](#_Toc182751589)

[Data Transformation 6](#_Toc182751590)

[Dataset Analysis 7](#_Toc182751591)

[1. Analysis of Trends in Wages Over Time 7](#_Toc182751592)

[2. Pay distribution by job title 8](#_Toc182751593)

[3. Correlation matrix for pay-related components 10](#_Toc182751594)

[4. Pay distribution by job title 11](#_Toc182751595)

[5. Average total pay by job title over time 12](#_Toc182751596)

[6. Distribution of Employee Status Analysis 13](#_Toc182751597)

[Data Analysis and Results 15](#_Toc182751598)

[Discussion and Conclusion 16](#_Toc182751599)

[References 18](#_Toc182751600)

# ****Introduction****

Employee compensation forms a critical management function within an organization. Salaries thus influence staff satisfaction, retention, and productivity. Proper structuring should give more insight into salary disclosure and analysis. Trends from the pay structure can provide better channels in the allocation of resources while focusing on eliminating disparities. This study will specifically focus on the public employee salary dataset. It will contain such data as base pay, overtime pay, benefits data, and total compensation by job title and agency.

The dataset will serve as a great source for discerning pay trends to understand pay distribution across jobs and to understand how other elements of pay, like overtime and fringe benefits, fit into the full picture. Cleaning and transforming the data, along with the modeling process, will also yield certain trends and trends that might benefit stakeholders with decision-making.

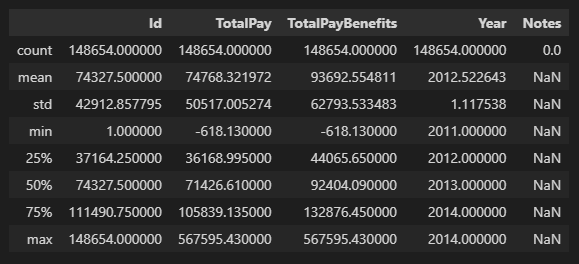
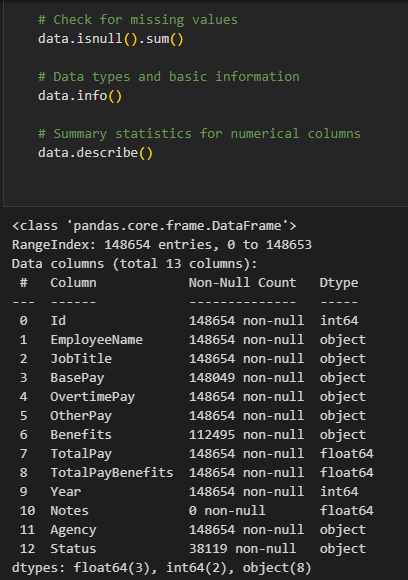
# Dataset

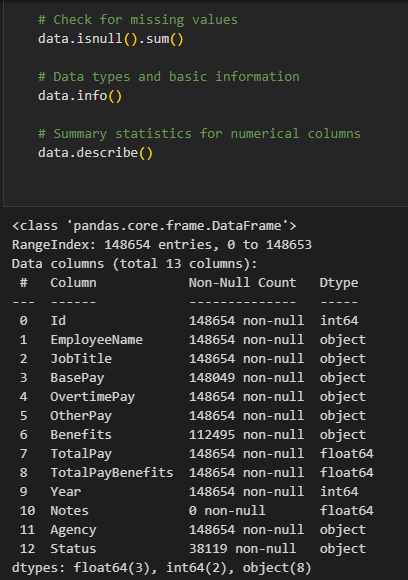
A data set contains extensive information about public employee salaries, detailing base pay, overtime pay, other pay types, and total compensation. It has 148,654 records with 13 attributes and provides elaborate compensation information across job titles, agencies, and years. Among the most critical columns are those including *Base Pay, Overtime Pay, Other Pay,* Benefits, and combining to provide *Total Pay and Total Pay Benefit*s. Other sections offer context for the roles and organizations tied to each entry, like *Job Title, Year, and Agency*.

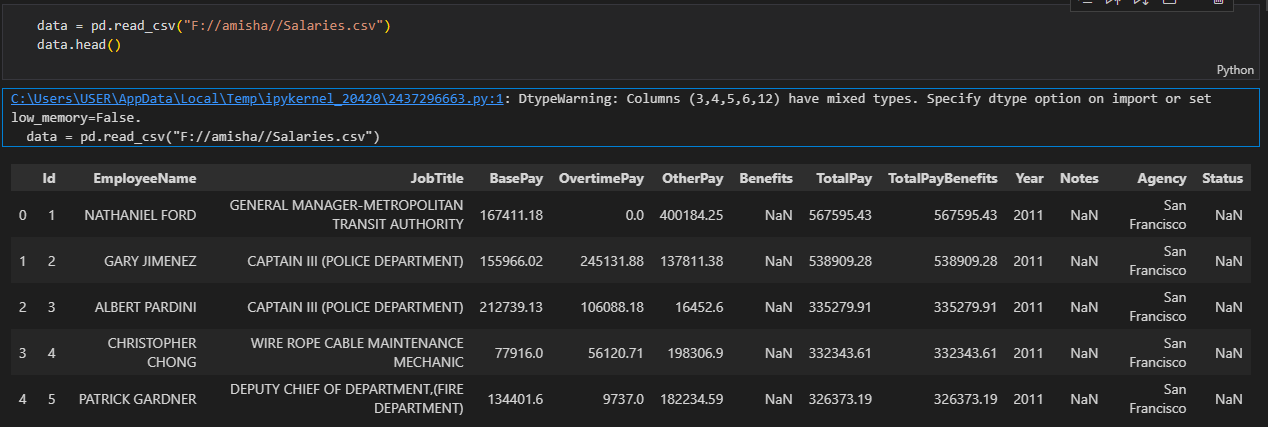
Although the dataset is rich, it also presents some limitations. It has missing entries, for example, in the core columns of BasePay (605 missing entries) and Benefits (36,159 missing entries). More importantly, one field, Notes, does not have any meaningful data and would thus be excluded from any analysis. Data types of important salary-related columns - that is, BasePay and OvertimePay - are stored as objects and must hence be converted to numeric if meaningful analysis is to be derived.

Therefore, the ideal source for analyzing trends of salary and identifying high and low earners would be this dataset. It can also be used further to understand overtime and benefits on total compensation. This project aimed to address missing values and ensure data consistency to draw actionable insights regarding optimal salary and transparency.

The following screenshots offer a clear view of the dataset







# Data Transformation

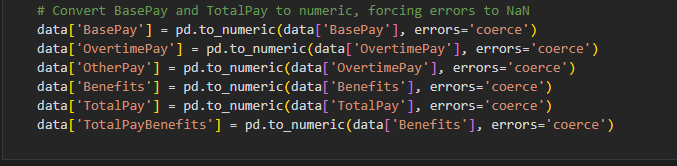
Data transformation is an important step in preparing raw data for analysis. In this dataset, this type of transformation mainly refers to converting all salary columns into a uniform format, removing invalid values, and getting the data ready for visualization and modeling.

**Important transformations done are**:

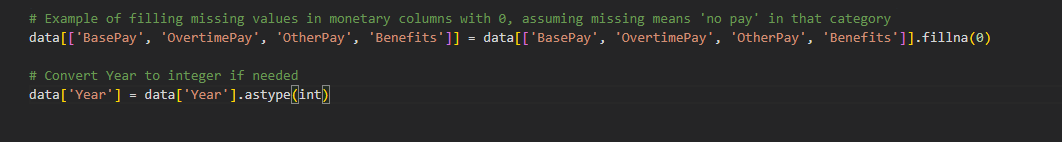
1. ***Conversion of Object Data to Numerical***: Columns that include BasePay, OvertimePay, OtherPay, Benefits, TotalPay, and TotalPayBenefits are object types, as they are completely numeric values. Using the function pd.to\_numeric() those columns were converted to numerical types with invalid or unprocessable input entries coerced into NaN. This helps ensure proper calculations and usability in the important statistical and machine learning tools.
2. ***Missing Values***:

* After the conversion, missing or invalid numerical inputs were found to be NaN.
* For columns of BasePay, OtherPay, OvertimePay, and Benefits, missing values were replaced by 0. To be expected, missing entries are likely unpaid or unreported values.

1. ***Standardization for Analysis***: The dataset is now ready for statistical computations, correlation analysis, and model development because all the columns that describe the pay are now in numeric format. This removes inconsistencies, thus more reliable and meaningful insights will be developed. These transformations provide a clean foundation, making the dataset ready for visualization, exploratory data analysis, and predictive modeling.



On the issue of missing values, all monetary columns such as `BasePay`, `OvertimePay`, `OtherPay`, and `Benefits` are filled with 0 since presumably missing entries denote no payment in those categories. Another column, `Year` is converted to an integer for uniformity and preparation for analysis and visualization of time-based trends.

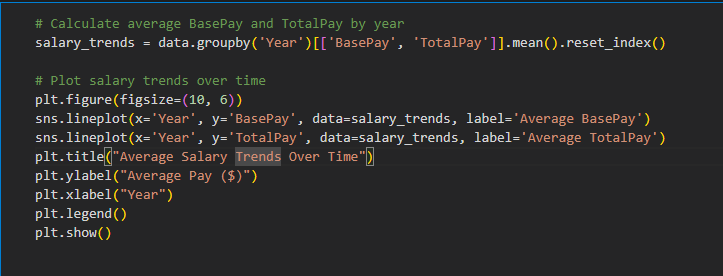


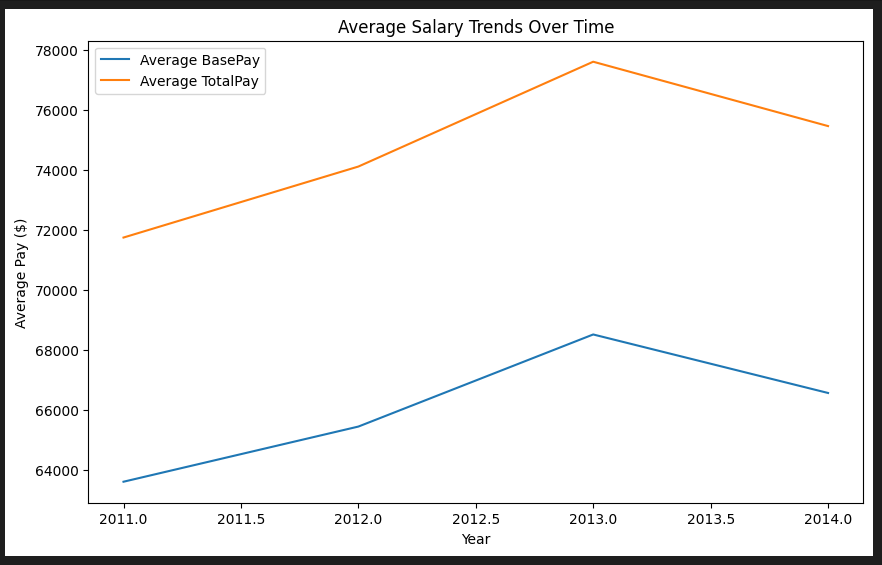
# Dataset Analysis

Dataset analysis involves exploring the cleaned and transformed data to uncover patterns, trends, and relationships between variables. Below are key steps to guide the analysis:

## 1. Analysis of Trends in Wages Over Time

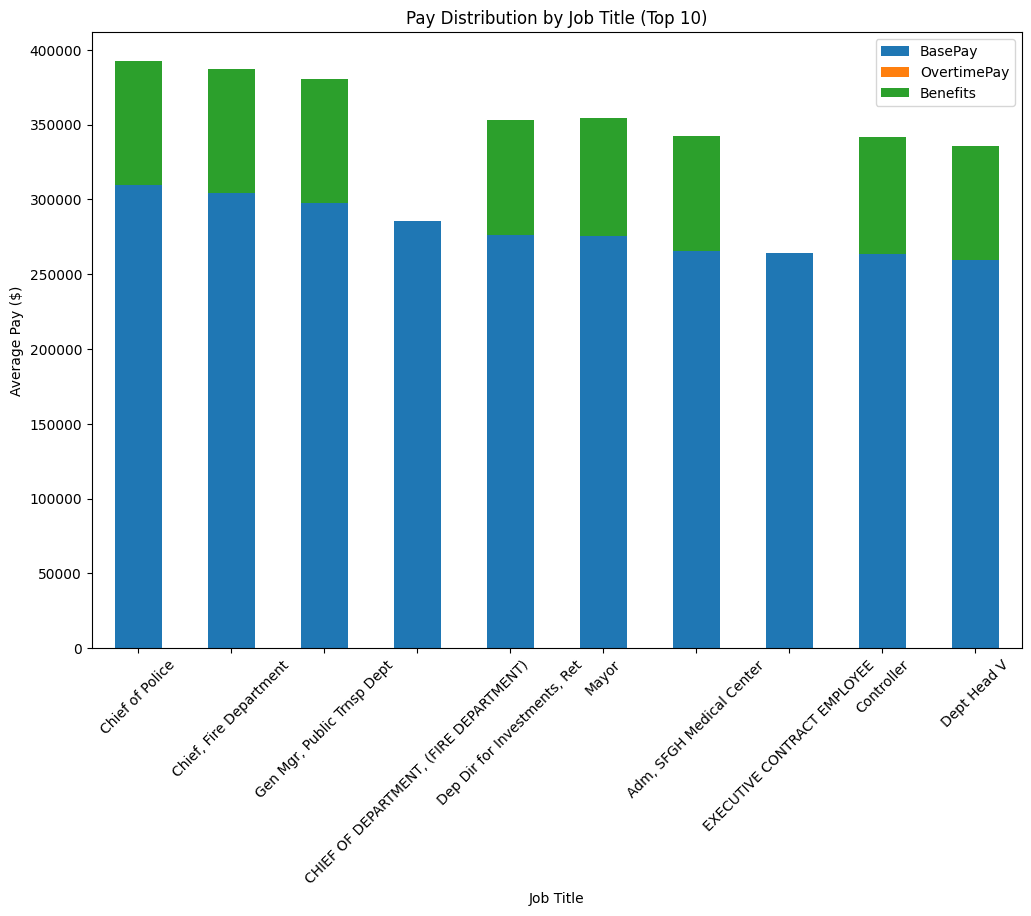
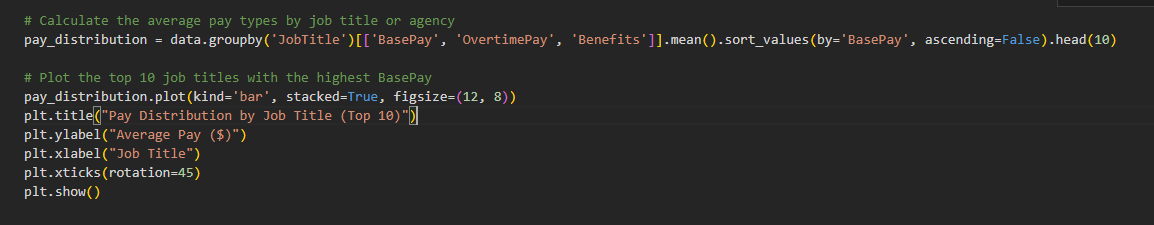
This analysis plotted the time series of average BasePay and TotalPay across years; it shows how salaries have been changing through the years. With data grouped into Years, calculating the average of each group presented a nice plot of the trend of salary changes using line plots. The BasePay trend alone presents the stability or growth in base salary over the years; the TotalPay trend will present overtime and extra compensation that makes a great difference if the two differ significantly. Now, fluctuation or growth can give market feedback regarding certain economic factors or policy changes, or sometimes budget adjustments in answering the compensation trends in public sector employment.





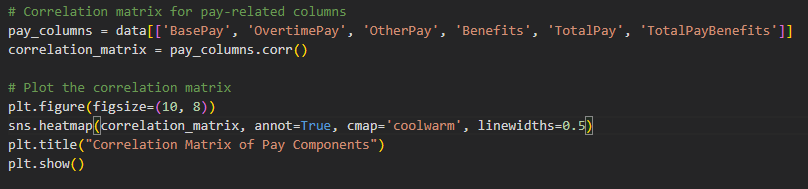
## 2. Pay distribution by job title

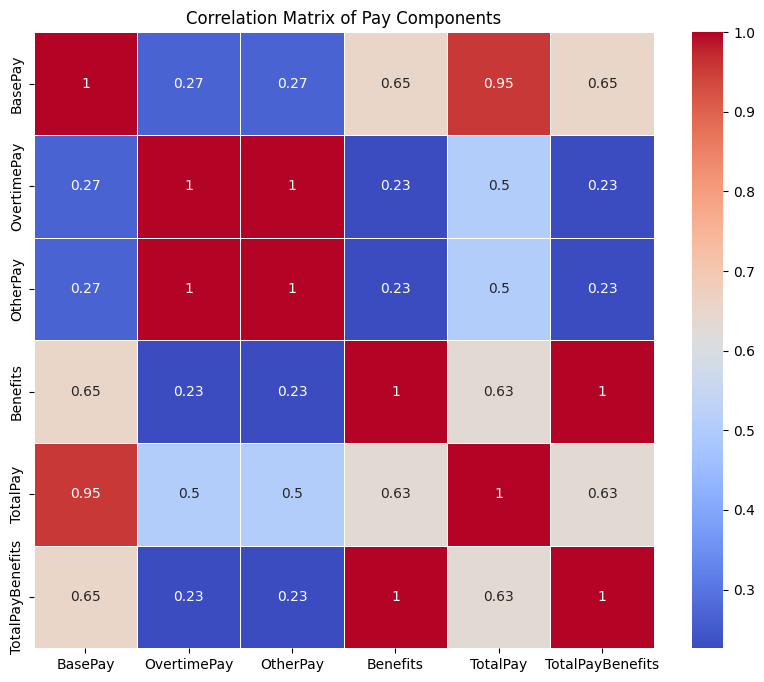
This paper explores the mean pay distribution across job titles based on `BasePay`, `OvertimePay`, and `Benefits`. Data was aggregated by `JobTitle` and ranked in the top 10 roles by `BasePay`. A stacked bar chart of all pay types implies the relative percentage that is due to overtime and benefits. The chart provides information on which roles carry the highest salaries and how their compensations are structured. This analysis helps to identify high-paying roles and understand the relative importance of base pay, overtime, and benefits in the making of total compensation for different job titles.



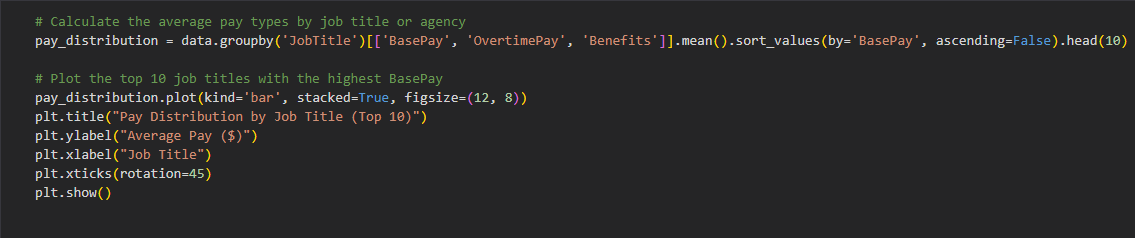
## 3. Correlation matrix for pay-related components

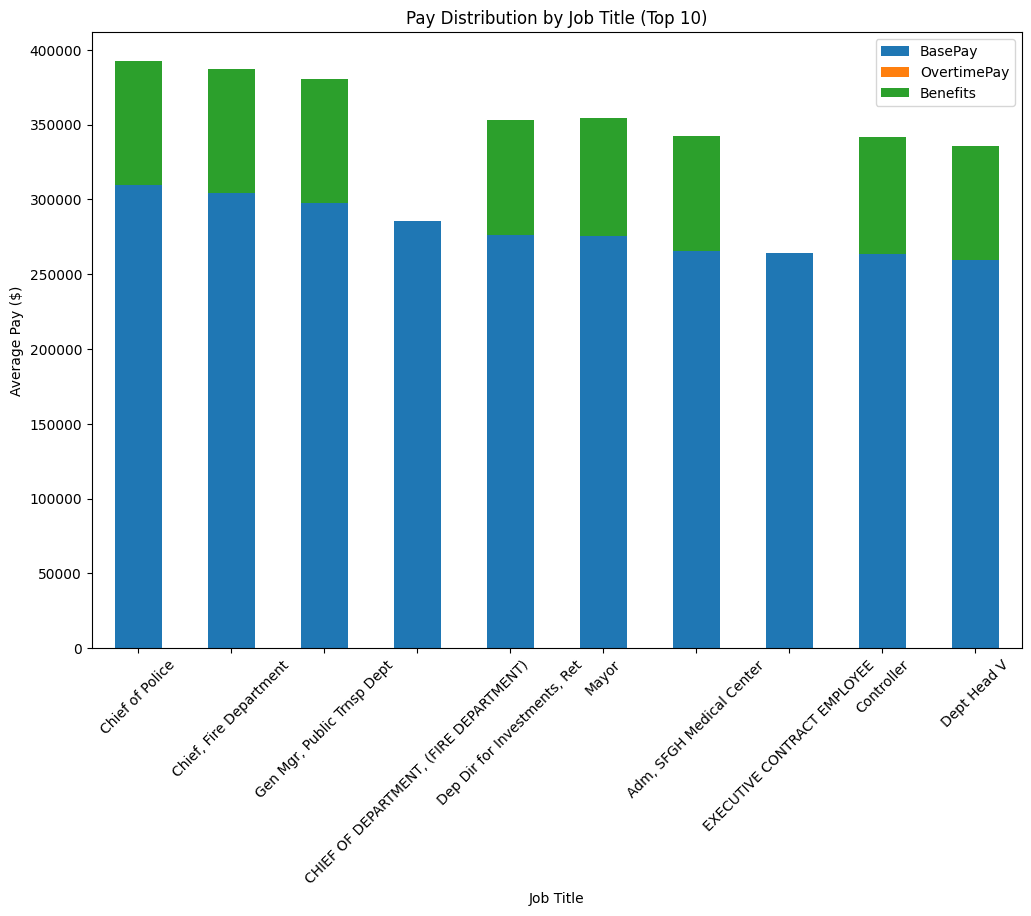
A correlation matrix was generated to find relations among pay-related components, namely `BasePay`, `OvertimePay`, `OtherPay`, `Benefits`, `TotalPay`, and `TotalPayBenefits`. The heatmap visualization depicts how these variables interact within a range of -1 to 1, correlating. High positive correlation values-for example, the one between `TotalPay` and `BasePay`-indicate that the base pay is primarily driving total compensation. Weaker correlations simply highlight regions where other factors, such as benefits or overtime, mainly drive independently. This analysis elicits interdependencies between pay components, which will identify the main drivers for total employee compensation.





## 4. Pay distribution by job title

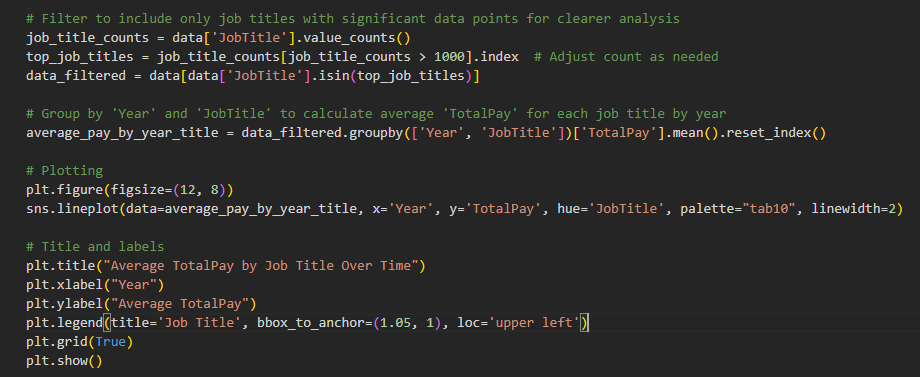
This section discusses how to analyze the average pay distribution for the top 10 job titles with the highest average Base Pay. It does so by grouping the dataset on JobTitle and computing the mean values of BasePay, OvertimePay, and Benefits to gain insight into how compensation varies by role.

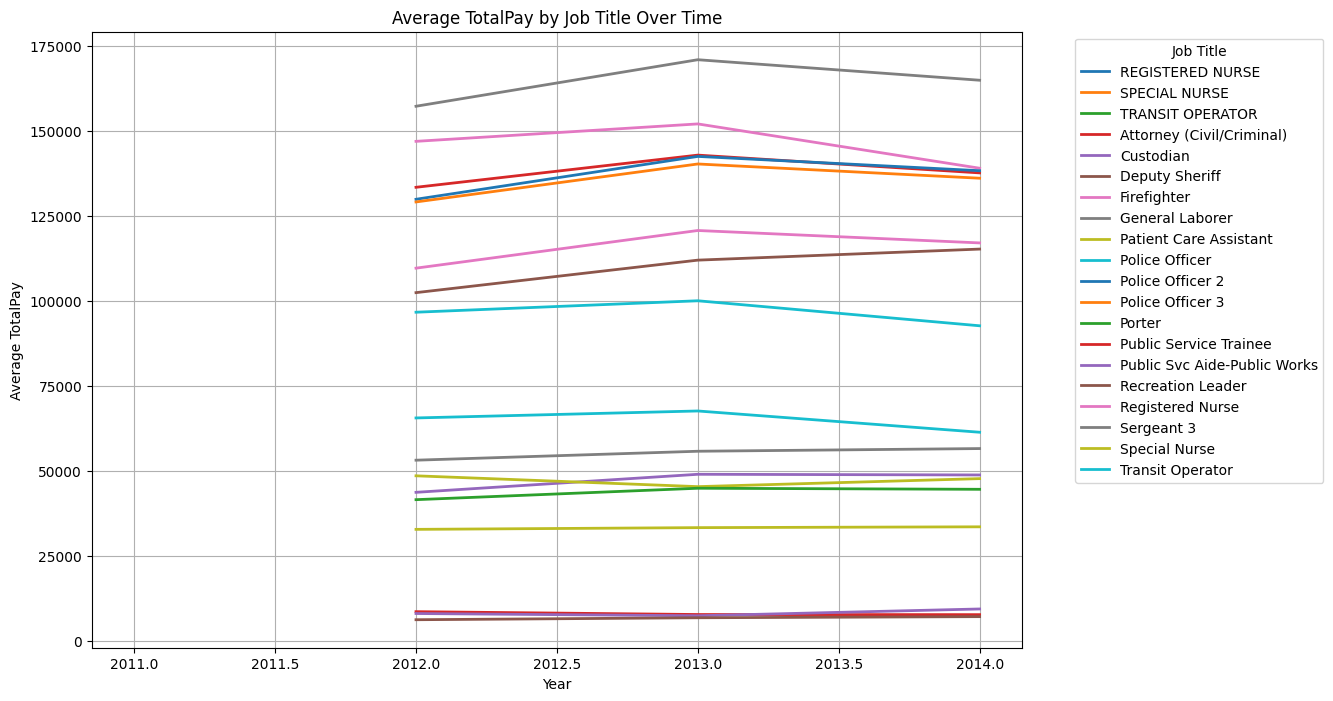


## 5. Average total pay by job title over time

Analytical view helps one in determining the trends in compensations for high titles jobs across time. This analysis filters out job roles having more than 1,000 entries and thus reduces noise entries that are few. It helps in making finer insights about total average pay, how it might shift over years, grouped by the combination of Year and Job Title.

The line plot shows consistent growth, stagnation, or fluctuations related to the economy, policy changes, or other organizational structures. Noted trends, such as a sudden spike in TotalPay for specific positions, could suggest higher demand, increased overtime utilization, and higher reliance on their pay for benefits. This data is great support for discussions on workforce planning, budgeting, and fair pay policies.

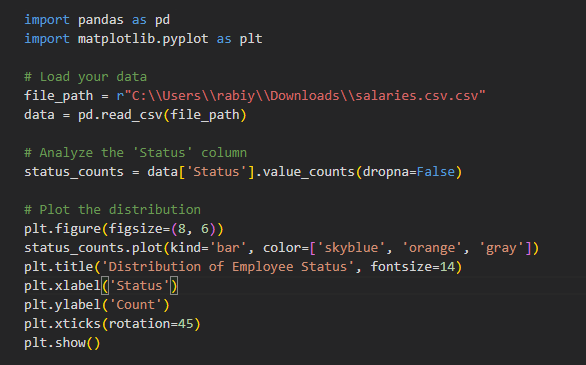


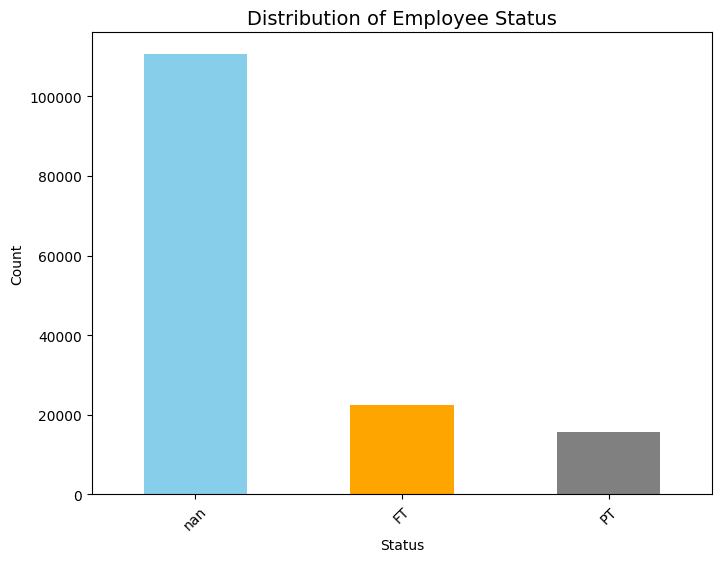


## 6. Distribution of Employee Status Analysis

The analysis sees the 'employee status' distribution, plotting a bar chart with the count for each category, along with missing values if any, indicating the proportion of employees with various statuses, such as active, retired, or terminated. Now, with the inclusion of missing data, it provides transparency and completeness to the analysis.

The results may reflect workforce trends, such as an increased active workforce that characterizes ongoing operations or a population of retiree workforce significant to the organization due to historical growth. These insights guide workforce management and planning and can initiate efforts in adjusting organizational policies.





# Data Analysis and Results

The dataset was analyzed in terms of trends and insights relevant to public employee compensation. Specifically, this included consideration of Base Pay, Overtime Pay, Benefits, and Total Compensation across the job titles, years, and statuses. Data cleaning and transformation were integral for accuracy as missing values were accounted for and columns about salaries were turned into numeric types.

Its main conclusion is that within the ***Pay Distribution by Job Title***, the skills are special or the responsibilities critical to garner the high Base Pay. The distribution of base pay, overtime pay, and benefits across total earnings was also shown well with the help of a stacked bar chart.

***Average TotalPay Over Time*** was a forecasting of trends in the top job titles with over 1,000 data points. The line graph showed year-over-year fluctuations of change; some jobs had increased over time, and others had very little to no movement in pay. This could be because of economic change, budget modification, or some other organizational priority.

Inspecting the ***Distribution of Employee Status***, a bar graph showed active, retired, and terminated employees, highlighting workforce dynamics. This distribution gives an understanding of organizational growth, retirement patterns, and turnover rates.

The results allow stakeholders to pinpoint high-value jobs, understand the structure of pay, and assess employee distribution to make better decisions. Such knowledge supports workforce planning, pay optimization, and even fair pay practices, which should be aligned with organizational objectives and economic circumstances.

# Discussion and Conclusion

An analysis of public employee salary data, mainly public employee compensation trends and workforce dynamic analysis, was presented. A closer look at Base Pay, Overtime Pay, and Benefits indicated that high-paying jobs are tied up to specific skills and organizational priorities. The visual pay distribution also showed how extra elements, such as overtime and benefits, significantly contributed to aggregate earnings and were a suitable starting point in considering compensation equity and resource allocation. Trend analysis of the timeline disclosed dynamic changes in Total Pay across various job titles, reflecting economic changes, policy implementation, or changing needs in organizational dynamics. The constant watch to periodically adjust salary structures should be based on maintaining competitiveness and equity. Similarly, the composition of distribution of statuses of employees informed workforce structure, retirement trends, and turnover rates, which form very important undertakings in workforce planning and sustainability.

While the findings are actionable, there are limitations. Missing or incomplete data and absence of some information relating to Benefits the robustness of conclusions. Analysis has only been done within the confines of historical data and would not be able to reflect the contribution of external factors such as economic downturns or legislative changes without supplementary context.

In summary, these analyses provide a good insight into public employee wages to assist decision-making regarding equitable pay structures, workforce enhancement, and future planning. However, for further strengthening of insights, include going external with the data and pursuing expanded analysis that would include predictive modeling, allowing organizations to foresee trends and be proactive about workforce challenges.

# References

1. Dataset- <https://www.kaggle.com/datasets/kaggle/sf-salaries?resource=download>
2. Scipy Documentation. (2023). *Statistical Correlation Analysis*. Retrieved from https://docs.scipy.org/doc/scipy/reference/
3. Storytelling with Data. (2021). *Best Practices in Data Visualization*. Retrieved from [https://www.storytellingwithdata.com](https://www.storytellingwithdata.com/)
4. Tidy Tuesday Data Science Community. (2022). *Public Sector Pay Analysis and Visualization Techniques*. Retrieved from [https://www.tidytuesday.com](https://www.tidytuesday.com/)
5. Python documentation for each library:
   * 1. Pandas: <https://pandas.pydata.org/docs/>
     2. Matplotlib: <https://matplotlib.org/stable/contents.html>
     3. Scikit-learn: <https://scikit-learn.org/stable/>