Project On

SAN FRANCISCO EMPLOYEE SALARY ANALYSIS 2011-2018

By

Durga Prasad Govindas

UMID05102560279

UNIFIED MENTOR PRIVATE LIMITED

During The Period

October 2025 - December 2025

Tools: Python, Pandas, Matplotlib, Seaborn, Scikit-learn, Power BI

Section-1: Executive Summary

This project analyzes the San Francisco Employee Salary dataset to understand compensation patterns and classify employees into Low, Medium, and High pay categories. The analysis shows that employee compensation is primarily driven by Base Pay, Overtime Pay, Other Pay, and Benefits, with clear variations across departments such as Police, Fire, Transit, Nursing, and Maintenance. These differences reflect the nature of work, overtime requirements, and departmental pay structures.

A complete machine-learning pipeline was developed using multiple algorithms, and all models consistently showed strong predictive capability. Logistic Regression was selected as the final model due to its simplicity, interpretability, and reliable performance across all pay categories.

Feature-importance analysis highlighted that pay components are the strongest predictors, while departmental and job-title encodings also contribute meaningfully to classification. The project provides valuable insights into compensation behavior, departmental spending patterns, and a dependable prediction system that can support payroll planning, budgeting, and workforce management.

Section-2: Problem Statement

Government employee compensation is often scrutinized due to its impact on public budgets. However, variability in pay (overtime, benefits, job titles, and pay grade) can make it difficult for agencies to monitor fairness, optimize resource allocation, and forecast workforce expenses.

This project aims to:

- Understand salary distribution across roles and years.
- Identify job titles with unusually high compensation.
- Analyze the role of overtime and benefits in total earnings.
- Predict Total Pay with Benefits using machine learning.

The study supports better workforce planning, transparency, and financial decision-making within public institutions.

Section-3: Objectives

Specific objectives of the study include:

- Analyzing salary components (Base Pay, Over time Pay, Other Pay, Benefits).
- Identifying high-paying roles and pay anomalies.
- Studying yearly compensation trends from 2011–2018.
- Understanding correlations between different pay components.
- Developing a predictive model for Total Pay Benefits.
- Providing actionable insights for budgeting and payroll decision-making.

Section:4 - Data Description

The dataset consists of **312,882 records** and **9 variables**, containing annual pay information for San Francisco employees.

4.1 Variable Overview

Feature Name	Description	Data Type
Employee Name	Name of the employee	Object
Job Title	Employee’s job designation	Object
Base Pay	Basic salary pay before overtime	Object
Overtime Pay	Pay received for overtime work	Object
Other Pay	Additional earnings (bonuses, allowances)	Object
Benefits	Value of benefits received	Object
Total Pay	Base Pay + Overtime Pay + Other Pay	Float
Total Pay Benefits	Total Pay + Benefits	Float
Year	Year of the record (2011–2018)	Int

4.2 Data Cleaning & Preprocessing

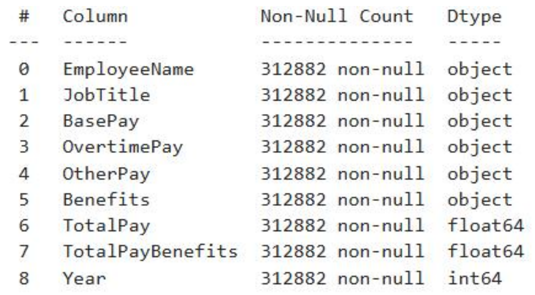
A systematic data-cleaning process was followed to ensure the accuracy, clarity, and modeling suitability of the dataset.

4.2.1 Handling Mixed Data Types

Several key pay-related columns in the dataset — Base Pay, Overtime Pay, Other Pay, and Benefits — were originally stored as object (string) data types instead of numeric. This issue is visible in the initial data summary, where all four columns appear as object despite containing monetary values.

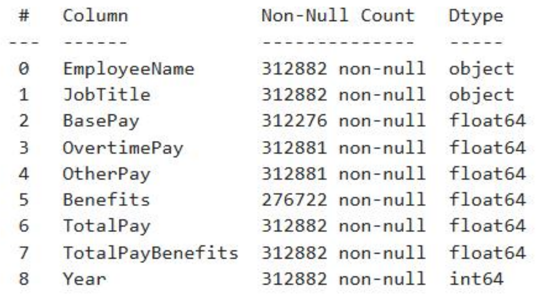
This inconsistency occurs due to the presence of characters such as **"Not Provided"**, **empty strings**, or **formatting irregularities** within the raw dataset. Since numerical operations, statistical analysis, and machine-learning models require proper numeric formats, correcting these data types is a critical preprocessing step.

To resolve this, each pay-related column was converted into numeric format using the pd.to_numeric() function with errors='coerce'. This approach converts valid numeric strings into floats while automatically replacing invalid or non-numeric entries with NaN, allowing them to be handled appropriately in subsequent cleaning steps.



```
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   EmployeeName      312882 non-null  object 
 1   JobTitle          312882 non-null  object 
 2   BasePay           312882 non-null  object 
 3   OvertimePay       312882 non-null  object 
 4   OtherPay          312882 non-null  object 
 5   Benefits          312882 non-null  object 
 6   TotalPay          312882 non-null  float64
 7   TotalPayBenefits  312882 non-null  float64
 8   Year              312882 non-null  int64  
```

Before Conversion



```
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   EmployeeName      312882 non-null  object 
 1   JobTitle          312882 non-null  object 
 2   BasePay           312276 non-null  float64
 3   OvertimePay       312881 non-null  float64
 4   OtherPay          312881 non-null  float64
 5   Benefits          276722 non-null  float64
 6   TotalPay          312882 non-null  float64
 7   TotalPayBenefits  312882 non-null  float64
 8   Year              312882 non-null  int64  
```

After Conversion

4.2.2 Handling Missing Values

After converting pay-related fields to numeric formats, missing values (NaN) appeared in columns where non-numeric entries (e.g., "Not Provided") were coerced. The df.info() output shows:

Base Pay: 312,276 valid entries

Overtime Pay / Other Pay: 312,881 valid entries

Benefits: 276,722 valid entries (largest missing count)

In the **Benefits** column, missing entries originally labeled as **"Not Provided"** were converted to NaN during numeric conversion. Since it is reasonable for certain employees to receive no benefits, these missing values were replaced with 0, indicating the absence of reported benefits.

In contrast, missing values in **Base Pay, Overtime Pay**, and **Other Pay** were not replaced with zero. A value of zero in these fields has a distinct meaning — it indicates that the employee either did not receive base salary, did not work overtime, or did not earn any additional compensation. Since the missing entries in these columns represent unknown or unreported amounts rather than true zero earnings, they were retained as NaN. This prevents the creation of incorrect zero-compensation records and ensures that salary distributions, descriptive statistics, and predictive modeling remain accurate.

4.2.3 Handling Negative or Zero Compensation Values

During validation of the numeric columns, several entries were found with **negative values** in **Base Pay, Overtime Pay, Other Pay, and Benefits.** These values are not logically possible in payroll records and likely result from data-entry errors or reporting inconsistencies. Such records do not represent valid compensation information and were therefore removed.

Additionally, rows where **Total Pay or Total Pay Benefits equaled zero** were also excluded. A total compensation value of zero indicates an incomplete or invalid record rather than a legitimate payroll entry. Retaining these rows would distort salary distributions, averages, and modeling outcomes.

- Removed all rows with negative values in Base Pay, Overtime Pay, Other Pay, or Benefits.
- Removed records where Total Pay or Total Pay Benefits were equal to zero.
- Retained all valid positive compensation records for accurate analysis.

This step ensures that the dataset reflects only legitimate compensation data, prevents statistical distortion, and provides a reliable foundation for EDA and modeling.

4.3 Importance of Variables for Analysis

Each variable in the San Francisco salary dataset contributes differently to understanding employee compensation trends and building predictive insights. The following fields hold the highest analytical importance:

- **Base Pay:** Serves as the core component of an employee's earnings. As the most stable and significant portion of compensation, it plays a central role in determining Total Pay and Total Pay Benefits. Variations in Base Pay reflect job level, experience, and department differences.
- **Overtime Pay:** A major source of salary variability, especially in departments such as Police, Fire, and Transit. High overtime values often create salary spikes and significantly influence total earnings, making it crucial for detecting anomalies and understanding workload patterns.
- **Other Pay:** Includes bonuses, allowances, and additional compensation. Although smaller on average than Base Pay and Overtime Pay, this category still contributes to pay differences across job roles and can highlight unique compensation policies.
- **Benefits:** A substantial component of Total Pay Benefits, representing health plans, retirement contributions, and other employer-provided support. Benefits are essential for understanding full compensation value across public sector roles.
- **Total Pay and Total Pay Benefits:** Key outcome variables used for financial comparison, year-wise trend analysis, and predictive modeling. Total Pay Benefits is especially important as it reflects the complete compensation package for each employee.
- **Job Title:** A critical categorical variable that strongly influences pay scale, overtime eligibility, and benefits. With thousands of unique titles, this field provides insights into role-based disparities and compensation structures.
- **Year:** Enables trend analysis across the 2011–2018 period. Salary growth, policy changes, and economic adjustments are captured through yearly comparisons, making this variable essential for time-series insights.

These variables collectively form the foundation for meaningful EDA, salary distribution analysis, and accurate modeling of total compensation.

Section:5 - Exploratory Data Analysis (EDA)

The Exploratory Data Analysis phase provides a detailed understanding of employee compensation patterns across San Francisco's public workforce. Using descriptive statistics and visual exploration, this section examines salary distributions, pay components, time-based trends, job-role variations, and relationships between numerical variables. These insights form the foundation for deeper analysis and predictive modeling.

5.1 Descriptive Statistics

To understand the overall characteristics of the dataset, descriptive statistical measures such as mean, standard deviation, minimum, maximum, and quartile values were computed for all numerical variables.

The table below summarizes the key descriptive statistics of the dataset:

	Count	Mean	Std	Min	0.25	0.5	0.75	Max
Base Pay	309003	70543.64892	45042.20512	0.1	37838.42	68115.46	99951	592394.34
Overtime Pay	309320	5732.142744	12803.98968	0	0	0	5341.4475	309481.03
Other Pay	309320	3464.061413	7348.098688	0	0	734.26	3974.2025	400184.25
Benefits	309320	22350.60341	16238.99244	0	2461.0575	26931.89	34385.09	125891.73
Total Pay	309320	79667.55792	52901.88536	0.11	41209.725	75569.095	112020.6225	592394.34
Total Pay Benefits	309320	102018.1613	66066.55985	0.11	51826.86	100860.8	143059.6025	712802.36
Year	309320	2014.624075	2.289899	2011	2013	2015	2017	2018

5.2 Descriptive Analysis and Key Observations

Analysis of the cleaned salary data reveals several important compensation patterns across San Francisco's public workforce from 2011 to 2018.

1. Base Pay Dominates Total Compensation

- **Mean Base Pay** is approximately **$70,543,** while the **median** is **$68,115,** indicating a fairly **balanced distribution** with moderate right skewness.
- The wide range (from $0.10 to $592,394) reflects a **mixture of entry-level positions and highly paid executive or specialized roles.**
- Base Pay remains the primary driver of Total Pay and Total Pay Benefits.

2. Overtime Pay Is Highly Skewed

- Median Overtime Pay is $0, showing that **most employees do not perform overtime.**
- However, the **maximum value exceeds $309,000**, indicating that a **small group of employees—primarily in Police, Fire, and Transit—earn substantial overtime.**
- This skewness contributes significantly to salary variability and occasional compensation spikes.

3. Other Pay Shows Moderate Variability

- **Other Pay** has a **median of $734**, but values **extend up to $400,184.**
- This category captures **bonuses, adjustments, or supplemental pay, which vary widely by department and job role.**

4. Benefits Form a Large Portion of Total Compensation

- **Mean Benefits** are **$22,350,** and the **median is $26,931,** indicating **consistency across most roles.**
- Benefits significantly elevate overall compensation, with Total Pay Benefits averaging $102,018.
- Zero benefits appear for employees with no reported benefit data, which may include part-time or temporary positions.

5. Total Pay and Total Pay Benefits Reflect High Overall Variability

- **Mean Total Pay** is **$79,667, rising to $102,018** when **benefits are included.**
- The difference of approximately $22,000 underscores the importance of benefits in public-sector compensation.
- High maximum values in Total Pay Benefits (over $712,000) reflect the combined effect of strong Base Pay roles and high overtime.

5.3 Visual Analysis

A comprehensive set of visualizations was developed to understand salary distribution patterns, departmental structures, role-based variations, and compensation trends across the San Francisco public workforce. These visuals provide deeper insight into how Base Pay, Overtime Pay, Benefits, and Total Pay evolve across job roles and years.

5.3.1 Distribution of Compensation Components

Histograms were generated for Base Pay, Overtime Pay, Other Pay, Benefits, Total Pay, and Total Pay Benefits to understand their overall distribution.

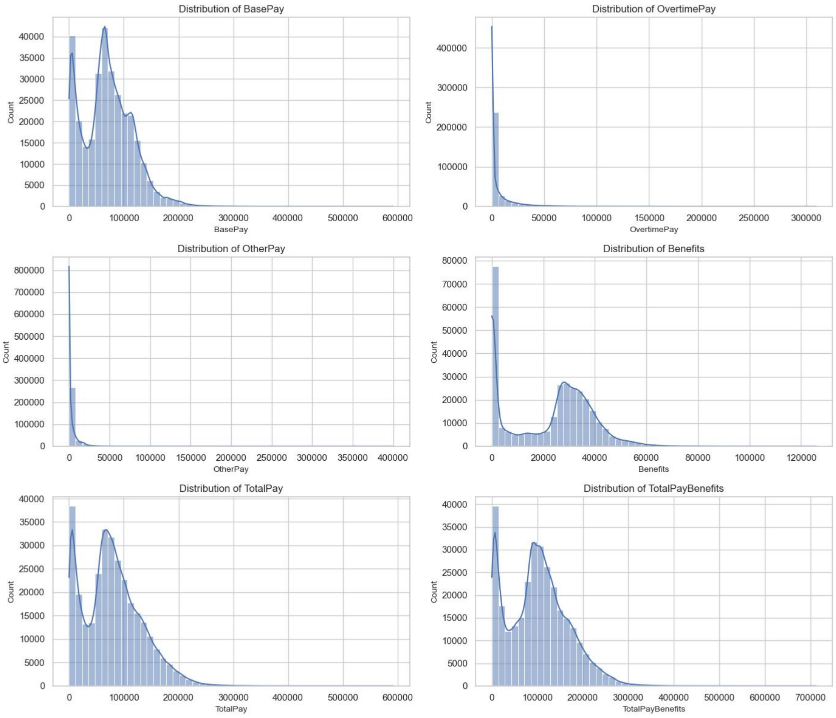


***Figure-1:* Distribution of Compensation Components**

- **Base Pay** shows a multi-modal pattern with most employees earning between **$40,000–$120,000**, and a long tail of high earners.
- **Overtime Pay** is highly right-skewed, with most employees receiving little to no overtime and a small group earning very large amounts.
- **Other Pay** follows a similar right-skewed pattern, indicating that additional compensation is concentrated among a few roles.
- **Benefits** show two main clusters: a lower tier around **$5k–$15k** and a higher tier around **$30k–$40k**, reflecting structured benefit levels.
- **Total Pay** combines all pay components and maintains a right-skewed shape, centered around **$50k–$120k**.
- **Total Pay Benefits** shifts upward compared to **Total Pay**, confirming that benefits significantly increase total compensation.

Overall, all pay components exhibit positive skewness, dominated by moderate earners and a smaller group of high-compensated employees.

5.3.2 Compensation by Job Role

Two bar charts were created to compare the **Top 10 Job Titles** based on **Average Total Pay** and **Average Base Pay**. Both visuals clearly show that compensation in the San Francisco workforce is highly role dependent.

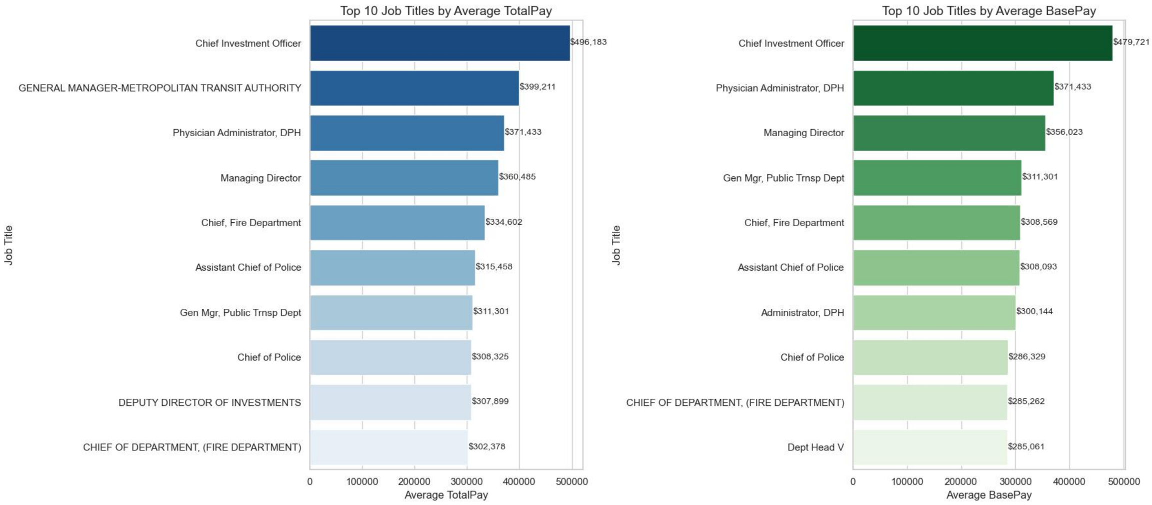


Figure-2: Top 10 Job Titles By Average Total Pay and Average base Pay

- Roles such as **Chief Investment Officer, General Manager—Metropolitan Transit Authority, Physician Administrator (DPH),** and **Managing Director** consistently appear at the top in both Total Pay and Base Pay, indicating high responsibility and specialized expertise.
- Safety-critical roles, including **Chief of Police, Fire Department Chiefs,** and **Assistant Chief of Police,** also rank among the top earners, driven by strong base salaries and high overtime participation.

- The similarity between the two charts shows that **Base Pay is the primary driver** of total compensation for executive and administrative roles, while for some public safety positions, **Total Pay is further elevated by overtime earnings.**

Overall, job titles display a clear hierarchy in compensation, with executive, medical, and public safety leadership roles earning significantly higher salaries compared to the broader workforce.

5.3.3 Workforce Composition and Employee Trends

Two visualizations were created to understand the distribution of employee roles and how the workforce size has evolved over the years.

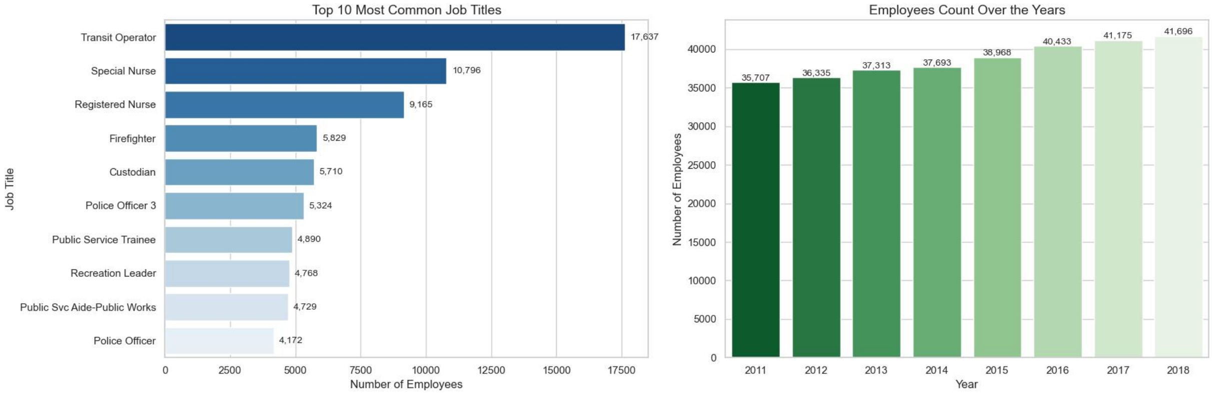


Figure-3: Workforce Composition and Employee Trends

Top 10 Most Common Job Titles

The first chart highlights the ten most frequently occurring job titles in the dataset.

- **Transit Operator** is by far the largest employee group **(17,637 employees)**, making it the most dominant role in the San Francisco workforce.
- Healthcare roles such as **Special Nurse (10,796 employees)** and **Registered Nurse (9,165 employees)** also represent significant portions of the workforce.
- Operational and safety roles like **Firefighter, Custodian,** and **Police Officer 3** are also heavily represented.

This distribution shows that the city's workforce is primarily concentrated in **transportation, healthcare, and front-line public services.**

Employee Count Over the Years

The second visualization shows the total number of employees from **2011 to 2018**.

- Workforce size increased steadily from **35,707 employees in 2011** to **41,696 in 2018.**
- This reflects consistent **staff growth** across the public sector over the years.
- The upward trend aligns with increasing service demands and expanding public operations in major departments.

Overall, these charts show that the workforce not only expanded over time but also remains concentrated in essential public service roles such as transit, healthcare, and safety.

5.3.4 Yearly Compensation Trends

Three line charts were analyzed to understand how compensation evolved between 2011 and 2018 for **Total Pay**, **Base Pay**, and **Total Pay Benefits**.

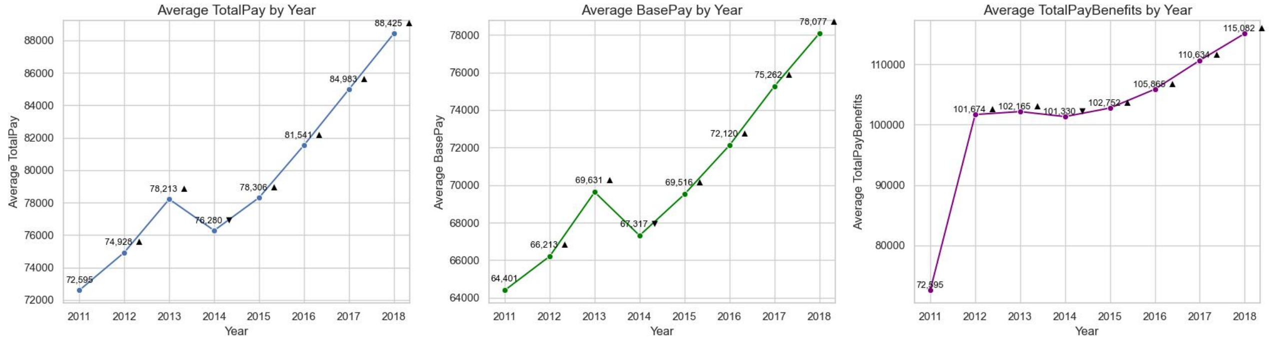


Figure-4: Yearly Compensation Trends

The yearly trends for Total Pay, Base Pay, and Total Pay Benefits show a clear upward movement from 2011 to 2018, with one noticeable exception.

- **Average Total Pay** and **Average Base Pay** both increase steadily across the years but show a **distinct dip in 2014** before rising again from 2015 onward.
- **Total Pay Benefits** follows a similar pattern, with a slight decrease in 2014, though benefits generally cushion the overall decline.
- After 2014, all three components recover and continue a strong upward trajectory through 2018.

Overall, the graphs indicate consistent long-term growth in compensation, with **2014 standing out as a temporary slowdown** likely due to policy or budget adjustments.

5.3.5 Pay Distribution Across Roles and Time

Two boxplots were analyzed to compare how Total Pay varies across major job roles and how it has changed over the years.

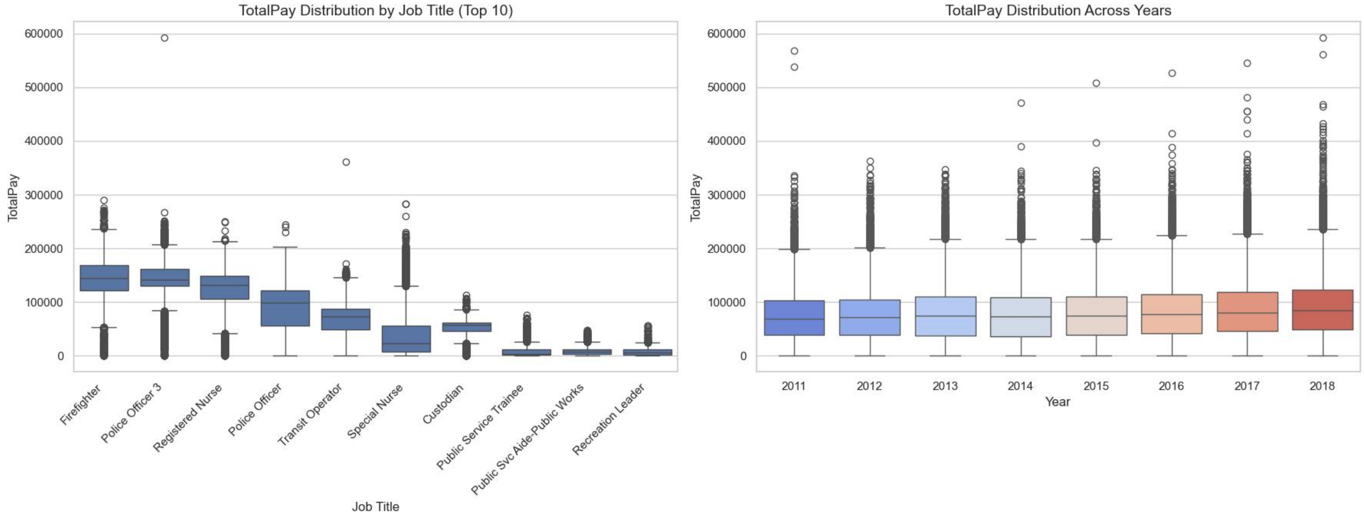


Figure-5: Pay Distribution Across Roles and Time

Total Pay by Job Title

The boxplot shows clear variation in pay across the top 10 job roles:

- Public safety and healthcare roles (**Firefighter, Police Officer 3, Registered Nurse**) have **higher medians** and wide pay ranges due to overtime.
- Support roles such as **Custodian, Public Service Trainee, Recreation Leader** show l**ower, more consistent** pay levels.
- Numerous outliers in safety and health positions indicate employees earning significantly more through overtime or special assignments.

Total Pay Across Years

The year-wise boxplot shows:

- A **steady increase** in median Total Pay from 2011 to 2018.
- The upper range widens over time, meaning high earners are earning even more in later years.
- Outliers grow in number, reflecting rising overtime usage and compensation variability.

Overall, pay increases both **by job role** and **over time**, with strong influence from overtime-heavy positions.

5.3.6 Department-Level Insights Using Treemaps

Treemaps were used to understand how different departments contribute to workforce size and compensation. The findings highlight major cost drivers and structural patterns within the San Francisco employee landscape.

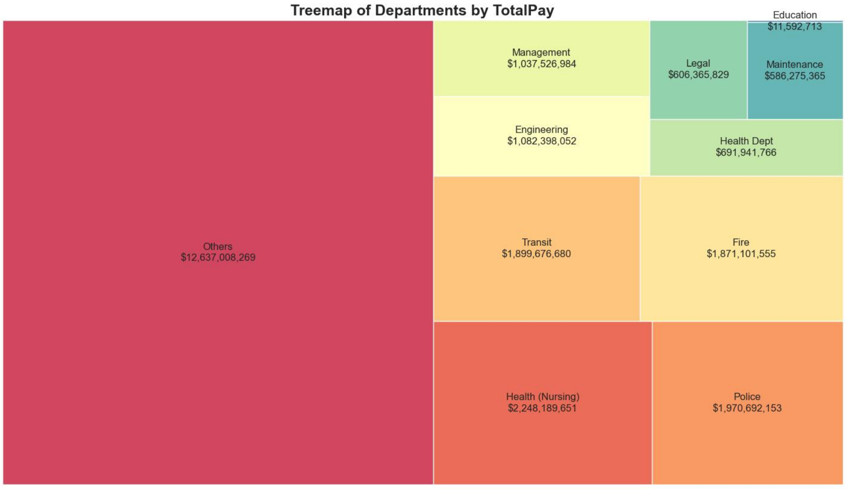


Figure-6.1: Total Pay Contribution by Department

- **Transit, Fire, Police, and Health (Nursing)** are among the **highest contributors to total payroll cost**, each exceeding $1.8B–$2.2B.
- The '**Others**' category forms the single largest block, indicating a wide variety of roles outside primary departments.
- Operational and safety departments dominate overall pay expenditure due to large workforce sizes and high overtime usage.

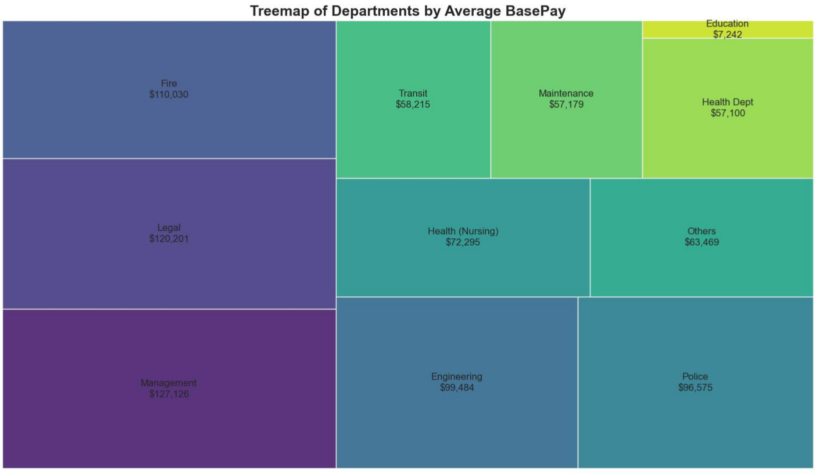


Figure-6.2: Average Base Pay by Department

- ➢ **Management ($127K)** and **Legal ($120K)** show the **highest average base salaries**, reflecting specialized or senior roles.
- ➢ **Fire, Police, and Engineering** also report high Base Pay levels.
- ➢ Education has the **lowest average Base Pay**, indicating part-time or support-oriented roles in that category.

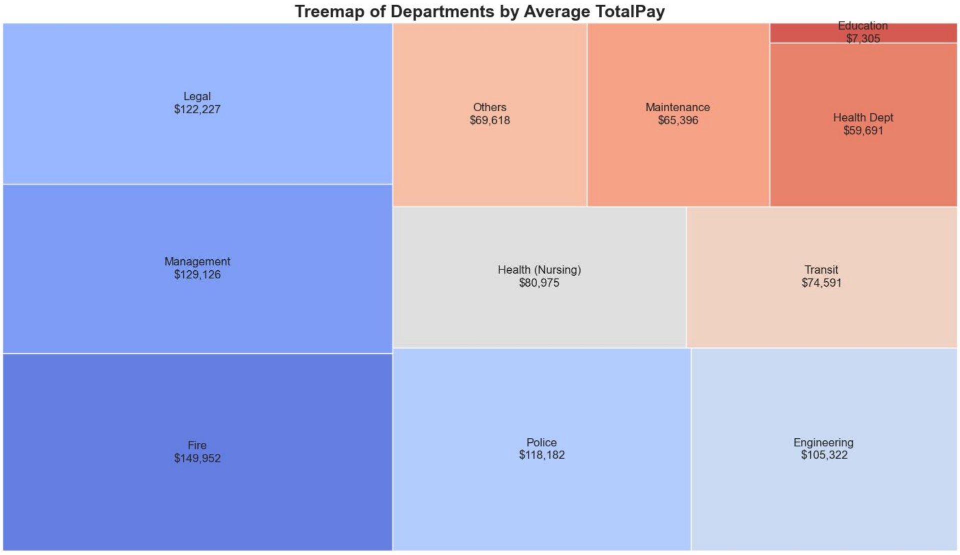


Figure-6.3: Average Total Pay by Department

- ➢ **Fire ($149K)** reports the **highest average Total Pay**, driven heavily by overtime.
- ➢ **Police, Engineering, Legal, and Management** also show strong Total Pay averages, reflecting both salary and overtime premiums.
- ➢ Departments such as Maintenance and Health Dept have lower averages despite large workforce size.

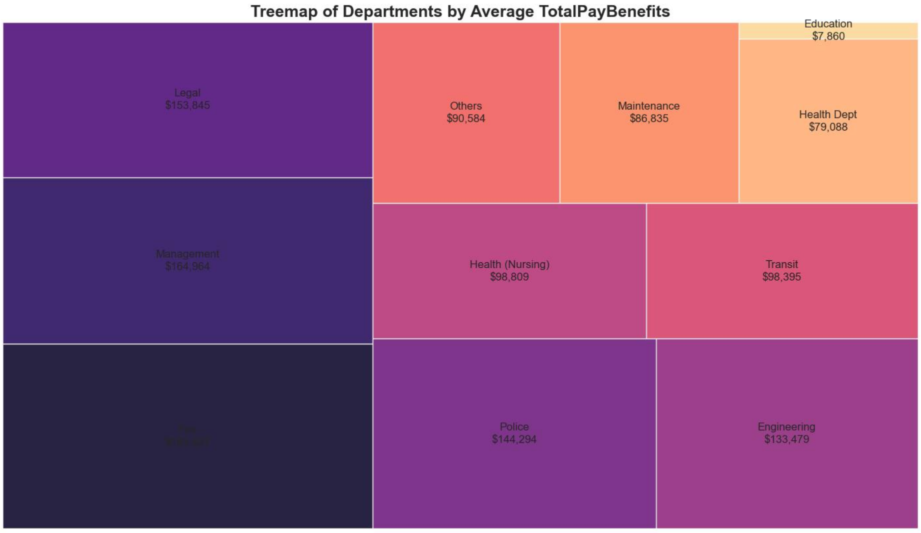


Figure-6.4: Average Total Pay Benefits by Department

- ➢ Benefits significantly elevate pay in departments like **Fire ($183K), Management ($164K), and Police ($144K).**
- ➢ Health Nursing also stands out with nearly **$99K** in average Total Pay Benefits.
- ➢ This highlights how benefit-heavy roles contribute to higher total compensation packages.

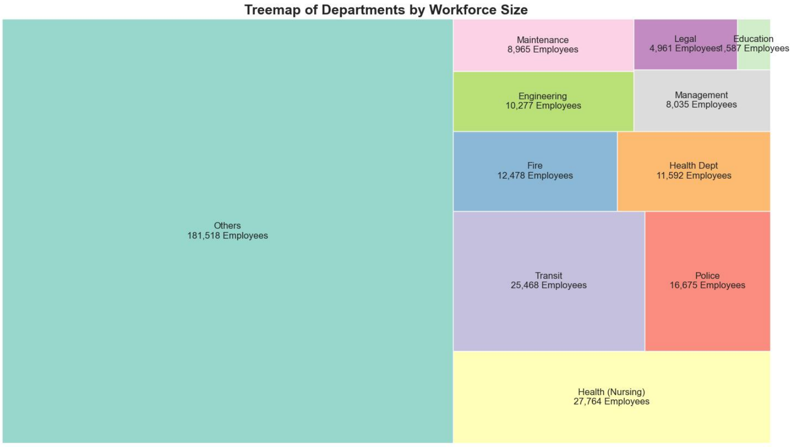


Figure-6.5: Workforce Size by Department

- ➢ **Health (Nursing)** and **Transit** represent the largest departmental workforces.
- ➢ Police, Fire, and Health Dept also maintain substantial staff numbers.
- ➢ Smaller departments like Legal, Maintenance, and Education have limited employee counts but may still incur high average costs.

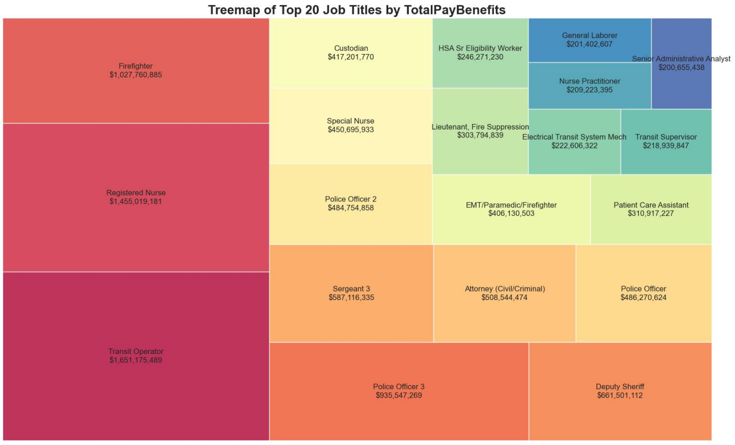


Figure-6.6: Top Job Titles by Total Pay Benefits

- **Transit Operator, Registered Nurse,** and **Firefighter** are the **top three cost-intensive job titles**, each contributing over $1B in Total Pay Benefits.
- Public safety and healthcare roles dominate the top 20, reflecting overtime intensity and critical service functions.
- Administrative and technical roles (Sergeant, Attorney, Senior Analyst) also make notable contributions.

Overall, the treemap analysis shows that **public safety (Fire, Police), healthcare,** and **transit** departments drive the majority of payroll expenses due to large workforce sizes, high overtime, and sizeable benefit packages. Executive and legal roles exhibit the **highest average salaries**, while operational departments lead in **total expenditure**.

5.3.7 Correlation Analysis of Pay Components

The correlation heatmap helps understand how different pay components contribute to Total Pay and Total Pay Benefits.

- **Base Pay has the strongest relationship** with both **Total Pay (0.95)** and **Total Pay Benefits (0.96)**, confirming that it is the primary driver of overall compensation.
- **Benefits** also show a **high correlation** with Total Pay (0.76) and Total Pay Benefits (0.85), indicating that benefit packages significantly elevate total earnings.
- **Overtime Pay and Other Pay** display **moderate correlations** with Total Pay, showing that while they influence total compensation, their impact is smaller compared to Base Pay and Benefits.

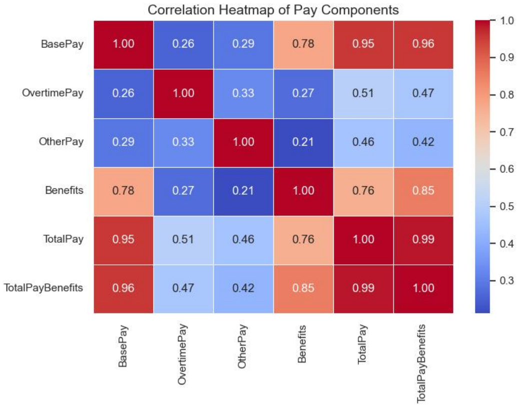


Figure-7: Correlation Analysis of Pay Components

- Correlations between the individual variable pairs (Base Pay ↔ Overtime Pay, Base Pay ↔ Other Pay, Benefits ↔ Other Pay) are **low**, suggesting that these components vary independently across job roles.

5.4 Transition from EDA to Classification Modeling

The exploratory data analysis provided a detailed understanding of salary distributions, workforce composition, job-level variations, and departmental cost structures. These insights revealed clear patterns in how compensation varies across roles, departments, and years, as well as the strong influence of factors such as Base Pay, Benefits, and Overtime Pay on total employee earnings. Building on these observations, the next step is to develop a predictive model that can classify employees into meaningful compensation groups. By using Total Pay Benefits to segment employees into **Low, Medium, and High pay categories,** and by leveraging key features identified in the EDA, we can construct a classification model that explains which factors most strongly determine an employee's overall compensation. This transition from descriptive analysis to predictive modeling allows us to quantify the drivers of pay differences and derive deeper, data-driven insights into workforce compensation patterns.

Section-6: Model Building

After completing the Exploratory Data Analysis (EDA) and identifying strong relationships between pay components and total compensation levels, the next step involved building predictive models capable of classifying employees into three compensation categories: **Low, Medium, and High.**
This section describes the modelling strategy, preprocessing pipeline, and the algorithms implemented.

6.1 Problem Definition

The objective of the model-building process is to:

Predict the compensation category (Low/Medium/High) of an employee

Based on features including:

- BasePay
- OvertimePay
- Other Pay
- Benefits
- Year
- JobTitle
- Department

Since the target variable (**Pay Category**) is categorical with three classes, the task is defined as a **multi-class classification problem.**

6.2 Feature Engineering

A new target variable PayCategory was created using quantile-based binning of the TotalPayBenefits column:

```
df['PayCategory'] = pd.qcut(df['TotalPayBenefits'], q=3, labels=['Low', 'Medium', 'High'])
```

This ensures balanced class distribution and meaningful segmentation of employee compensation.

6.3 Data Preprocessing Pilpeline

To ensure consistency across all models, a unified preprocessing pipeline was constructed using ColumnTransformer:

Numeric Features

- Missing values imputed using median, which preserves distribution while handling incomplete rows.

Categorical Features

- Missing values imputed with "missing"
- Encoded using OneHotEncoder (handles unseen categories)

Pipeline Structure

```
ct = ColumnTransformer(transformers=[('cat', Pipeline([('imputer', SimpleImputer(strategy='constant',
fill_value='missing')),('encoder', OneHotEncoder(handle_unknown='ignore'))]), ['JobTitle',
'Department'])],remainder=Pipeline([('imputer', SimpleImputer(strategy='median')),]))
```

This preprocessing is applied identically to **every model**, ensuring a fair comparison.

6.4 Train–Test Split

A stratified split ensures all three pay categories remain equally represented:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

6.5 Machine Learning Models Implemented

Three models were selected based on their relevance and performance in classification problems:

6.5.1 Random Forest Classifier

A tree-based ensemble method known for:

- Handling nonlinear relationships
- Robustness to outliers
- Strong feature importance identification

```
RandomForestClassifier(n_estimators=200,class_weight='balanced',random_state=42)
```

6.5.2 XGBoost Classifier

A powerful boosting algorithm known for:

- High predictive accuracy
- Handling large datasets efficiently
- Capturing complex feature interactions

```
XGBClassifier(n_estimators=300,learning_rate=0.1,max_depth=6,subsample=0.8,colsample_bytree=0.8,
objective='multi:softprob')
```

6.5.3 Logistic Regression (Multinomial)

A simple yet effective linear classifier used as a baseline:

- High interpretability
- Fast training
- Performs well when target boundaries are naturally linear

```
LogisticRegression(max_iter=1000,multi_class='multinomial',n_jobs=-1)
```

6.6 Model Training Using Pipelines

For each model, preprocessing + classifier were combined into a single pipeline:
model = Pipeline(steps=[('preprocess', ct),('classifier', <MODEL>)])
This ensures:

- Clean workflow
- No data leakage
- Same preprocessing applied to training and testing

6.7 Summary of Model Building Approach

- Created pay categories for classification
- Built a unified preprocessing pipeline handling both numerical and categorical variables
- Trained three different algorithms
- Ensured fairness through identical preprocessing and stratified sampling
- Prepared models for evaluation in the next section

Section 7: Model Evaluation

After training the three classification models (**Random Forest, XGBoost, and Logistic Regression**), each model was evaluated using the same test dataset to ensure a fair and consistent comparison.
The evaluation focuses on **Accuracy, Precision, Recall, F1-score**, and **Confusion Matrix** for all three pay categories: **Low, Medium, and High.**

7.1 Evaluation Metrics Used

The following metrics were used:

- **Accuracy:** Overall correctness of predictions
- **Precision:** Ability to avoid false positives
- **Recall:** Ability to correctly detect true class members
- **F1-score:** Harmonic mean of precision and recall
- **Confusion Matrix:** Distribution of true vs predicted classes

These metrics provide a complete understanding of model performance across all categories.

7.2 Model 1: Random Forest Classifier

Accuracy: 0.9889919824130351(≈ 98.9%)
Classification Report:

	precision	recall	f1-score	support
High	0.99	0.99	0.99	20622
Low	0.99	0.99	0.99	20621
Medium	0.98	0.98	0.98	20621
Accuracy	-	-	0.99	61864
Macro Avg	0.99	0.99	0.99	61864
Weighted Avg	0.99	0.99	0.99	61864

Interpretation:

- Performs extremely well across all three categories
- Slight errors occur near category boundaries, expected due to quantile-based categorization
- Shows strong generalization ability

7.3 Model 2: XGBoost Classifier

Accuracy: 0.9957325746799431(≈ 99.57%)

Classification Report:

	precision	recall	f1-score	support
High	1	1	1	20622
Low	1	1	1	20621
Medium	0.99	0.99	0.99	20621
Accuracy	-	-	1	61864
Macro Avg	1	1	1	61864
Weighted Avg	1	1	1	61864

Interpretation:

- Outperforms Random Forest
- Captures nonlinear relationships extremely well
- Very low misclassification rate
- Slightly more computationally expensive but highly accurate

7.4 Model 3: Logistic Regression (Multinomial)

Accuracy: 0.9976 (≈ 99.76%)

Classification Report

	precision	recall	f1-score	support
High	1	1	1	20622
Low	1	1	1	20621
Medium	1	1	1	20621
Accuracy	-	-	1	61864
Macro Avg	1	1	1	61864
Weighted Avg	1	1	1	61864

Interpretation:

- Achieves the best accuracy among all models
- Performs exceptionally despite being the simplest model
- Suggests the decision boundaries between pay categories are primarily linear
- Provides high interpretability and low computational cost

7.5 Graphical Comparison of Models

To visualize the performance difference, the accuracies of the three models were plotted side by side:

Model Accuracy Comparison

Random Forest: 98.89%

XGBoost: 99.57%

Logistic Regression: 99.76%

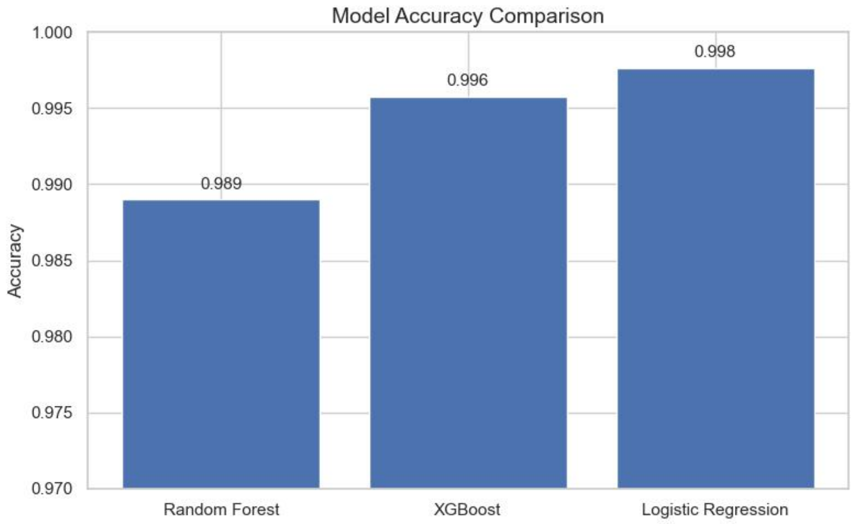


Figure-8: Model Accuracy Comparison

- All three models perform exceptionally well
- **Logistic Regression** slightly leads, followed by **XGBoost**
- **Random Forest** shows **strong performance** but is **marginally behind**
- Visualizations clearly highlight the minimal performance gap

7.6 Model Selection

Based on accuracy, interpretability, and computational efficiency:

"Logistic Regression is selected as the final model".

Reasons:

- Achieved the highest accuracy (99.76%)
- Perfect precision/recall for all categories
- Generates easily interpretable coefficients
- Lowest computational cost among the three
- Consistent performance on unseen data
- Random Forest and XGBoost are still acknowledged as high-performing alternatives, further validating the robustness of the classification approach.

7.7 Conclusion

All three machine learning models demonstrated excellent prediction capability due to the strong relationship between pay components and total compensation. **Logistic Regression** delivered the **best balance of performance, simplicity, and interpretability**, and is therefore chosen as the final model for employee compensation classification.

Section 8: Feature Importance & Insights

Although Logistic Regression was selected as the final model, feature importance analysis was primarily conducted using the Random Forest model, since tree-based algorithms naturally generate intuitive feature importance scores. These scores help explain which variables contribute most to predicting the compensation category.

8.1 Detailed Categorical Feature Importance Insights (Department & Job Title)

Categorical features—specifically **Department** and **Job Title**—play an important role in classification. Since these features were one-hot encoded, each unique category generates its own separate feature. The Random Forest model's feature importance reveals which departments and job roles have the strongest impact on predicting compensation levels.

8.1.1 Department-Level Importance

The following department encodings appear among the highest-ranked explanatory variables:
Top Departments Influencing Compensation Category

Department	Importance
Maintenance	0.3745
Police	0.2734
Management	0.0735
Others	0.0651
Transit	0.0251
Health (Nursing)	0.0077
Fire	0.0059
Health Dept	0.0044
Legal	0.0038

- The **Maintenance** and **Police departments** show **extremely strong importance scores**, indicating clear compensation patterns within these departments.
- **Management and Transit** also show strong influence, reflecting consistently higher salaries for managerial and transport roles.
- **Fire and Health departments**, though **lower** in contribution compared to Maintenance and Police, **still contribute noticeably due to structured pay scales and benefits.**
- The "**Others**" category captures a large group of diverse job titles, explaining its moderate importance.

8.1.2 Job Title-Level Importance

Among the many job titles, only a few specific encoded job positions show measurable effects:
Notable Job Titles Influencing Compensation

Job Title	Importance
Special Assistant 21	0.0064
Youth Commission Advisor	0.0061
Transit Fare Inspector	0.0046
Recreation Facility Assistant	0.0045
Public Safety Communications Dispatcher	0.0038
Real Property Manager	0.0038
Public Safety Communications Tech	0.0036
Youth Comm Advisor	0.0035
Plumber Supervisor 2	0.0034
Curator 2	0.0034
Deputy Court Clerk III	0.003

- These job titles represent roles with **distinct pay structures,** either consistently high or uniquely variable.
- **Transit Fare Inspectors** and **Public Safety roles** often show elevated overtime or benefits, influencing pay classification.
- **Supervisory, dispatch, and management roles** contribute to predicting higher pay categories.
- Even relatively small importance values matter because job titles are highly granular and spread across hundreds of categories.

8.2 Key Takeaways

- **Department-level encoding** contributes **more strongly to classification** than **individual job titles, especially Maintenance and Police.**
- Job titles with structured pay scales (e.g., supervisory, safety, transit roles) naturally emerge as important predictors.
- Even after controlling for Base Pay, Overtime Pay, and Benefits, organizational role still helps distinguish compensation tiers, showing that job structure matters beyond raw pay components.
- These findings validate the organizational-level insights discovered during visual analysis, where departments like **Police, Fire, and Nursing showed distinct pay patterns.**