Explaining Factors Contributing to Total Compensation for San Francisco City Employees

The Outliers: Suhas Kurapati, Matias Pinto Chavez, Mason Wu, Kelly Xu

2024-12-12

Introduction

San Francisco is the United States city with the highest median household income. However, some have raised questions on whether San Francisco employees are overcompensated, with claims of excessive benefits and questionable overtime practices. A 2021 Forbes article highlighted this issue, noting that "one deputy sheriff in San Francisco earned $574,595 last year, which included $315,896 in overtime" (Andrzejewski 2021).

With the rising scrutiny over high compensation figures, our model aims to provide insight into the impacts of different factors on Total Compensation in San Francisco, to 1) support decisions about budget allocations, overtime policies, and resource management to ensure fair and sustainable compensation practices, and 2) to help job seekers make informed decisions about employment opportunities and better assess their potential earnings.

Research Question

What factors significantly influence total compensation for San Francisco public sector employees, and in what ways do they affect it?

Data

Our dataset contains salary and benefits data for City employees from fiscal year 2013 to the present, and is sourced from DataSF. (San Francisco Controller's Office 2024). The data is collected by the San Francisco Controller's Office, and is added on a bi-annual basis when available for each fiscal and calendar year. Each row in the dataset contains information about an employee and their work compensation, with columns such as total compensation, normal salary, overtime salary, hours worked, health and dental benefits, and more.

Key Variables

Our response variable is `Total.Compensation`.

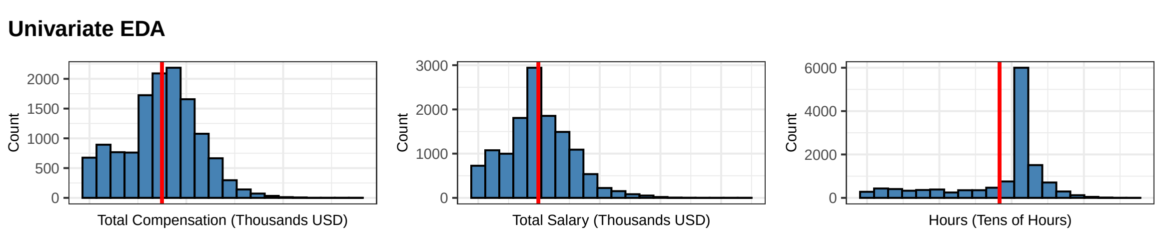
Variable_Name	Description	Datatype
Total.Compensation	The sum of all benefits & salaries paid to an employee, in thousands of USD.	Number
Salaries	Normal salaries paid to the employee (not including overtime pay, bonuses, etc.) in thousands of USD.	Number
Hours	The number of hours worked by the employee, in tens of hours.	Number
Employment.Type	Indicator for employment type: Permanent (ongoing) or NonPermanent (fixed-term/temporary).	Text
Overtime.Status	3 levels: NotEligible, Overtime (Eligible for overtime pay, and received >0 USD in overtime pay), and NoOvertime (eligible for overtime pay but did not receive overtime pay).	Text
Retirement	Indicator for whether an employee receives retirement benefits. 2 levels, Benefits and No Benefits	Text

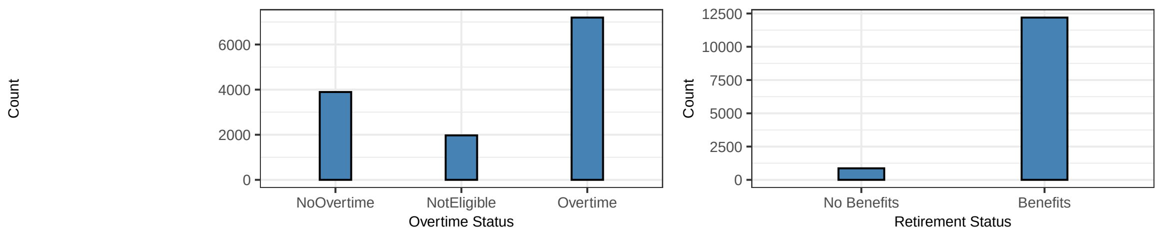
Data Considerations

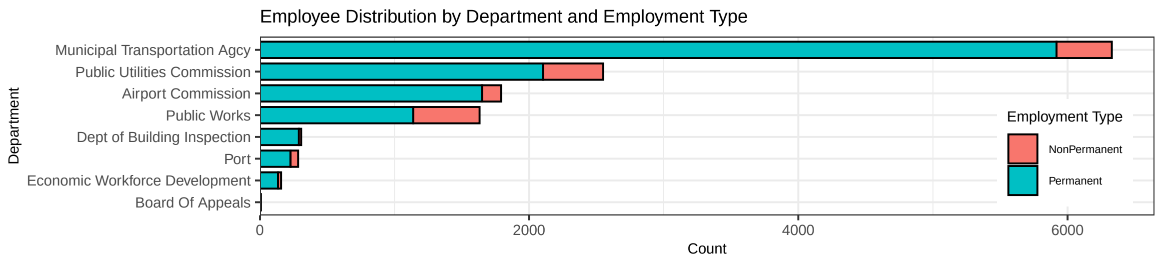
The dataset originally contained employee names, which we replaced with numerical IDs to ensure anonymity. We filtered out observations containing negative dollar amounts for `Total.Compensation` and/or `Salaries`, which represented retroactive adjustments to correct previous mistakes/overpayments. We then filtered out observations with zero-dollar amounts for `Total.Compensation` and/or `Salaries`, as these occur when employees receive one-time payouts or in similar special circumstances. Finally, we removed observations with repeat employee IDs to maintain independence among the observations. Situations like this represented individuals who changed roles or departments within the year. These transitions could indicate career mobility or temporary assignments, and removing these employees could potentially introduce a small bias.

However, we determined that filtering these extreme observations out likely would not affect the power of our model. Within the context of our dataset, all of the data filtering only removed 494 observations from an initial 13,554 observations–an amount unlikely to affect the model's ability to generalize well.

Exploratory Data Analysis



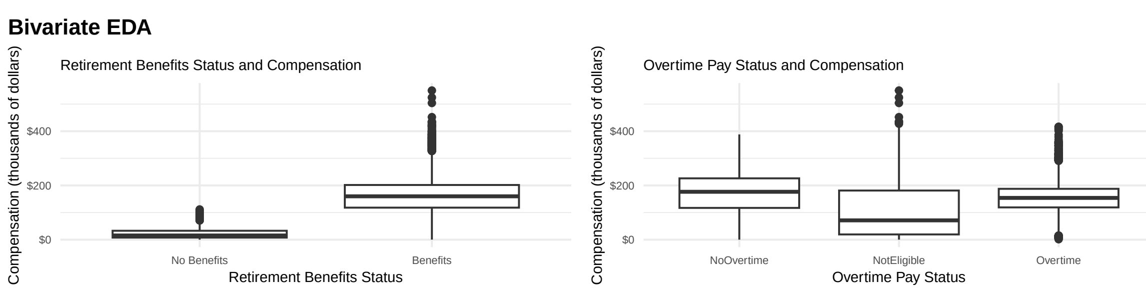


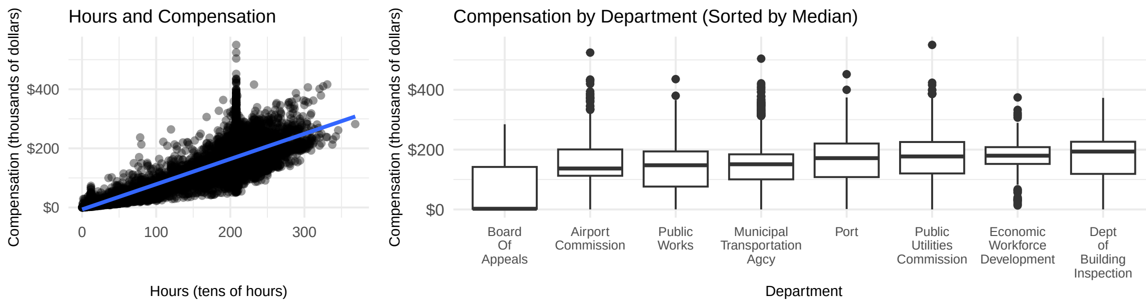


The total compensation and wage charts show relatively symmetric and right-skewed distributions, indicating that most employees follow a normal distribution, with some employees having higher than average wages and salaries. The work hours chart is mostly flat except for two values with large counts that reflect considerable variation. Most employees work similar hours, and a few employees work more.

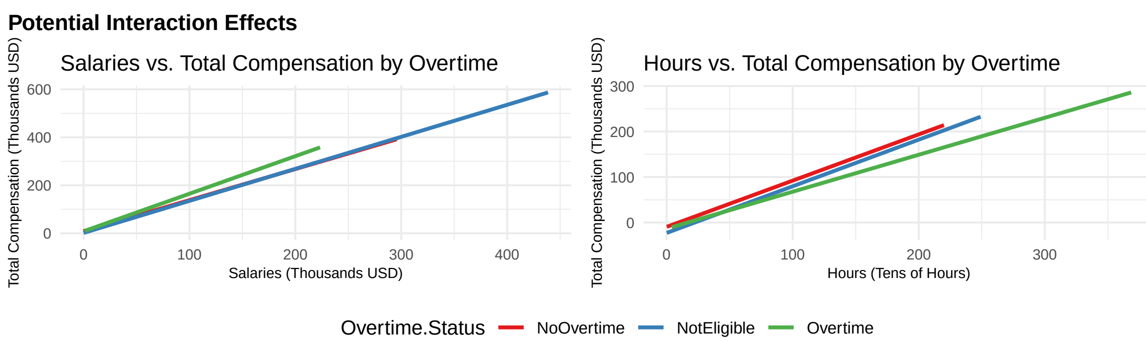
The peak in total hours worked corresponds to employees working around 40 hours per week in one year (2080 hours). This is demonstrated by the red dashed line in the histogram at 2080, where 31.1% of employees are in a range close to this value (2000-2160 hours), suggesting compliance with full-time schedules.

The overtime and retirement benefits charts show that most employees work overtime and receive retirement benefits. The employee department and employment type chart reveals that most employees work in the Municipal Transportation Agency and Public Utilities Commission departments, with the majority in permanent positions and a small fraction in non-permanent (temporary/fixed-term) positions.





The retirement plot shows that employees with retirement benefits generally have higher compensation. In the overtime graph, employees with and without hourly pay have similar compensation, but those who are not eligible for overtime pay have much lower compensation and a higher RIQ. The work hours graph indicates a strong positive linear relationship between hours and compensation, suggesting that compensation increases with work hours. In the department graph, median compensations are close to $200,000 in most departments except the Board of Appeals. However, the compensation differential (RIQ) varies across departments.



We will explore a possible interaction effect between Hours Worked and Overtime Status, as the slope of Hours Worked vs. Total Compensation appears to vary by Overtime Status. We will also explore a possible interaction effect between Overtime status and Normal Salaries Earned, as Normal Salary appears to have a stronger positive impact on Total Compensation for employees who work overtime than for those who do not.

Methodology

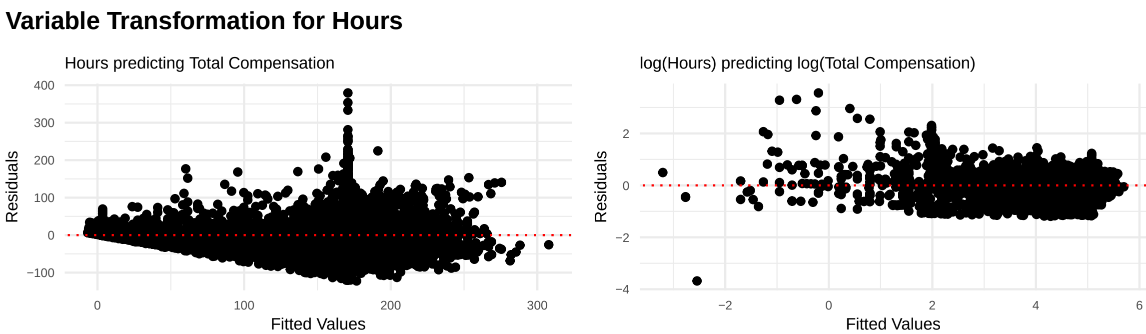
Choosing Modelling Strategy and Initial Variables

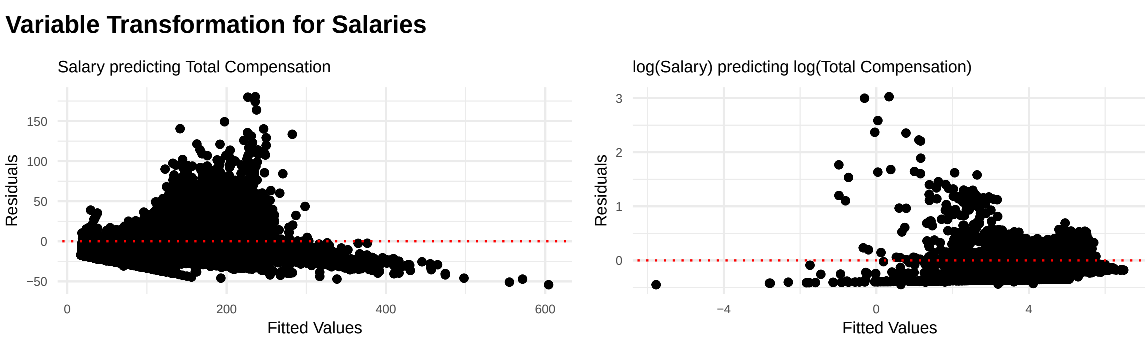
Since our response variable, `Total.Compensation` is numerical, we will fit multiple linear regression models. From the EDA, `Salaries` is by far the strongest predictor of `Total.Compensation`, which is naturally expected since normal salaries make up the most of total compensation, with extra benefits and bonuses contributing less.

To best analyze the effects of all the variables, we fit one MLR model with `Salaries` as a predictor, and one MLR model without `Salaries` as a predictor. Since our project motivations and research question center largely around overtime and hours worked, we also choose `Hours` and `Overtime.Status` as predictors in the model.

Variable Transformations

To assess the constant variance condition between the quantitative variables `Salaries` and `Hours`, and `Total.Compensation` , we fit a linear regression model using just `Salaries` to predict `Total.Compensation`, and another linear regression model using just `Hours` to predict `Total.Compensation`. From the plot of residuals vs. fitted values for each model, we decided to construct our model using `log(Total.Compensation)`, `log(Hours)`, and `log(Salaries)`, to better satisfy the constant variance condition.





Interaction Effects

We conducted nested F-Tests to explore potential interaction effects `Overtime.Status*Log.Hours` and `Overtime.Status*Log.Salaries`. The p-values for both F tests are well below our alpha level of 0.05, so we can conclude that both interaction terms are statistically significant, and should be added to the model [See Nested F-Test For Interaction Terms for the outputs of the nested F tests].

Final Model Selection: Nested F-Test and AIC

To determine whether any of `Department`, `Retirement`, or `Employment.Type` should be included in the model, we first perform a nested F-test. Since the p-value is well below our alpha level of 0.05, we conclude that at least one of the coefficients associated with one of the variables, is not 0.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
13051	311.186	NA	NA	NA	NA
13042	258.358	9	52.828	296.308	0

We proceed the variable selection process by calculating the AIC for a model with predictors `Log.Salaries`, `Log.Hours`, `Overtime.Status`, `Log.Salaries*Overtime.Status`, and `Log.Salaries*Hours`, with every different combination of `Department`, `Retirement`, and/or `Employment.Type`. We use AIC instead of BIC because we do not wish to impose a harsher penalty for models with more terms, based on our motivation to explain Total Compensation for employees in SF. The model with the lowest AIC (-14133.21) includes all of `Department`, `Retirement`, and `Employment.Type` , in addition to the variables we have already selected. [See AIC Analysis for all AIC calculations]

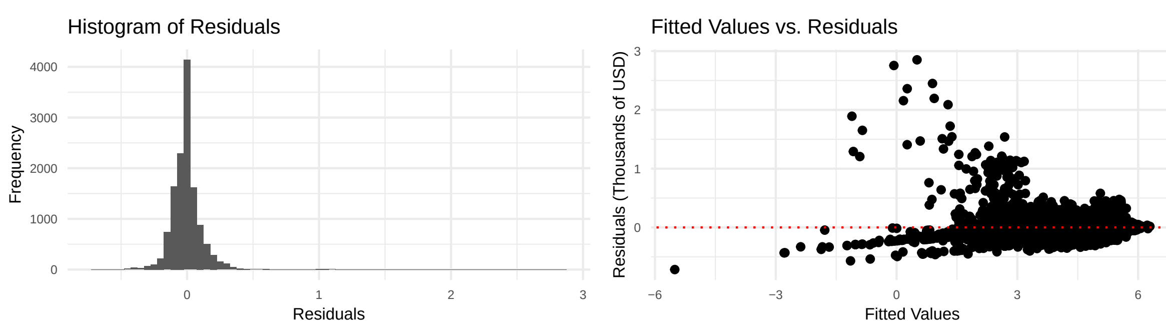
Final Model Selection: Model Diagnostics and Model Conditions

Final Model 1: With log(Salaries) as a predictor

Fitting a model that included interaction terms increased multicollinearity without improving the adjusted R squared value [see Testing Model 1 + interaction terms]. To precisely interpret the coefficients, it was decided not to include the interaction terms to address the multicollinearity issue. For Final Model 1, we will fit a main effects model predicting `Log.Total.Compensation` using `Log.Salaries`, `Log.Hours`, `Overtime.Status`, `Retirement` , `Employment.Type`, and `Department`.

Model Conditions

The constant variance condition could be improved, but the normality condition is satisfied.



Model Diagnostics

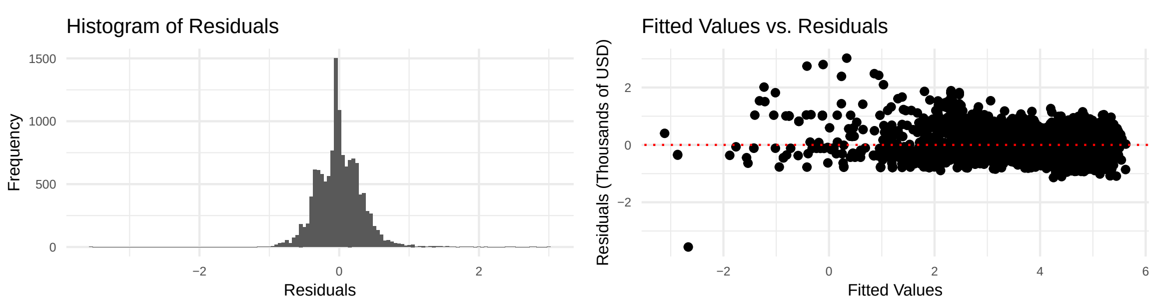
The model with Log(Salaries) shows some multicollinearity between Log(Salaries) and Log(Hours), with VIF values of 9.263 and 9.335, respectively—high but below the threshold of 10. This likely stems from fixed hourly wages and linear dependence in compensation packages driven by salary. Cook's distance values are all well below 0.5, indicating no influential outliers or high-leverage points [see Model Diagnostics: Final Model 1].

Final Model 2: Without log(Salaries) as a predictor

Similarly to Final Model 1, fitting a model including interaction terms introduced high multicollinearity, while not improving the adjusted R Squared value [see Testing Model 2 + interaction terms], so it was decided not to include interaction terms. For Final Model 2, we will fit a main effects model predicting `Log.Total.Compensation` using `Log.Hours`, `Overtime.Status`, `Retirement` , `Employment.Type`, and `Department`.

Model Conditions

The normality and constant variance conditions are satisfied for this model.



Model Diagnostics

All the VIFs of the model are well below 10, indicating no issue with multicollinearity. The Cook's Distance values for all observations are well below 0.5, indicating no observations with large leverage. [see Model Diagnostics: Final Model 2].

Results

Final Model with Log(Salaries) as a predictor:

term	estimate	std.error	statistic	p.value
(Intercept)	0.225	0.011	19.772	0.000
Log.Salaries	0.910	0.004	219.212	0.000

term	estimate	std.error	statistic	p.value
Log.Hours	0.029	0.005	6.048	0.000
Overtime.StatusNotEligible	-0.008	0.006	-1.371	0.170
Overtime.StatusOvertime	0.126	0.004	35.122	0.000
RetirementBenefits	0.343	0.008	43.084	0.000
Employment.TypePermanent	0.026	0.007	3.762	0.000
DepartmentBoard Of Appeals	-0.073	0.056	-1.297	0.195
DepartmentDept of Building Inspection	0.013	0.009	1.452	0.147
DepartmentEconomic Workforce Development	0.026	0.013	2.014	0.044
DepartmentMunicipal Transportation Agcy	0.045	0.004	11.252	0.000
DepartmentPort	-0.002	0.010	-0.228	0.820
DepartmentPublic Utilities Commission	0.014	0.005	3.091	0.002
DepartmentPublic Works	0.018	0.005	3.417	0.001

The Adjusted R^2 of 0.976 indicates the model explains 97.6% of the variability in log(Total Compensation). However, high VIF values for predictors like Log(Salaries) and Log(Hours) suggest multicollinearity, limiting the model's statistical reliability. We aim to address this issue in the second model.

Interpreting Results

Salaries: A 10% raise in salary (in thousands USD) corresponds to a 9.1% expected increase in compensation, holding all else constant. The nearly identical factor indicates that salary is the most correlated factor in determining compensation.

Hours: When the working hours (in 10s of hours) of an employee double, it is expected that their compensation will increase by 2.0%, holding all else constant.

Overtime: Employees ineligible for overtime are expected to earn 0.79% less than those eligible but not working overtime, while those working overtime are expected to earn 13.4% more than those eligible but not working overtime, holding all else constant.

Retirement Benefits: Receiving retirement benefits is expected to increase an employee's compensation by 40.9%, holding all else constant, indicating that it may be a major component of the employee compensation package.

Employment Type: Permanent employees are expected to earn 2.6% more than temporary or fixed-term employees, holding all else constant.

Department: Compensation varies by department, with employees from the most compensated department, the Municipal Transportation Agency, expected to earn 9.1% more than those in the least compensanted department, the Board of Appeals.

Final Model 2: Model without Log(Salaries) as a predictor:

term	estimate	std.error	statistic	p.value
(Intercept)	-0.938	0.022	-42.993	0.000
Log.Hours	1.006	0.004	241.611	0.000
Overtime.StatusNotEligible	0.004	0.012	0.369	0.712
Overtime.StatusOvertime	-0.186	0.007	-26.106	0.000
RetirementBenefits	0.369	0.017	21.433	0.000
Employment.TypePermanent	0.401	0.015	27.150	0.000
DepartmentBoard Of Appeals	0.416	0.122	3.409	0.001
DepartmentDept of Building Inspection	0.073	0.020	3.672	0.000
DepartmentEconomic Workforce Development	0.110	0.028	3.972	0.000
DepartmentMunicipal Transportation Agcy	-0.025	0.009	-2.908	0.004
DepartmentPort	0.187	0.021	9.081	0.000
DepartmentPublic Utilities Commission	0.136	0.010	13.573	0.000
DepartmentPublic Works	0.030	0.011	2.693	0.007

The Adjusted R^2 of 0.889 is lower than the previous model but still explains most of the variability in log(Total Compensation). Residual plots show better distribution, and VIF values (all <4) confirm multicollinearity is resolved (Model Diagnostics: Final Model 2). This indicates salary adds limited explanatory power relative to its contribution to VIF.

Interpreting Results

Hours: When the working hours (in 10s of hours) of an employee double, it is expected that their compensation (thousands of USD) will double, holding all else constant. This approximately identical factor indicates that working hours is the most correlated factor in determining compensation.

Overtime: Employees ineligible for overtime are expected to earn 0.4% more than those eligible but not working overtime, while those working overtime are expected to earn 17% less than those eligible but not working overtime, holding all else constant.

Retirement Benefits: Employees receiving retirement benefits are expected to earn 44.6% more than those who do not, holding all else constant.

Employment Type: Permanent employees are expected to earn 49.3% more in total compensation than temporary, holding all else constant.

Department: The Municipal Transportation Agency has the lowest expected compensation, while the Board of Appeals has the highest, holding all else constant.

Discussion + Conclusion

To answer our research question, the most significant factor that is expected to increase an employee's total compensation is retirement benefits. From both of our models, receiving retirement benefits is expected to significantly increase total compensation by about 40%. Permanent employees are expected to make more than temporary employees, and working more hours is expected to result in making more money. Model 1 predicts a smaller impact of these factors on total compensation, while Model 2 predicts a much greater impact.

From our models, working overtime is not expected to lead to extreme increases in total compensation, which suggests that employees making hundreds of thousands of dollars in overtime are likely not represenative of the total population. However, more analysis is necessary before we can draw conclusions about the specific impacts of working overtime on total compensation. From Model 1, employees who work overtime are expected to earn about 14% more than employees who do not, while from Model 2, employees who work overtime are expected to earn about 17% less than employees who do not.

More analysis is also necessary before we can draw conclusions about the impact of department on total compensation. From Model 1, it is expected that the Municipal Transportation Agency is the most highly compensated, and the Board of Appeals is the least compensated. This is in direct contrast to the results from Model 2, likely due to the high multicollinearity in Model 1, which causes high variability in model coefficients.

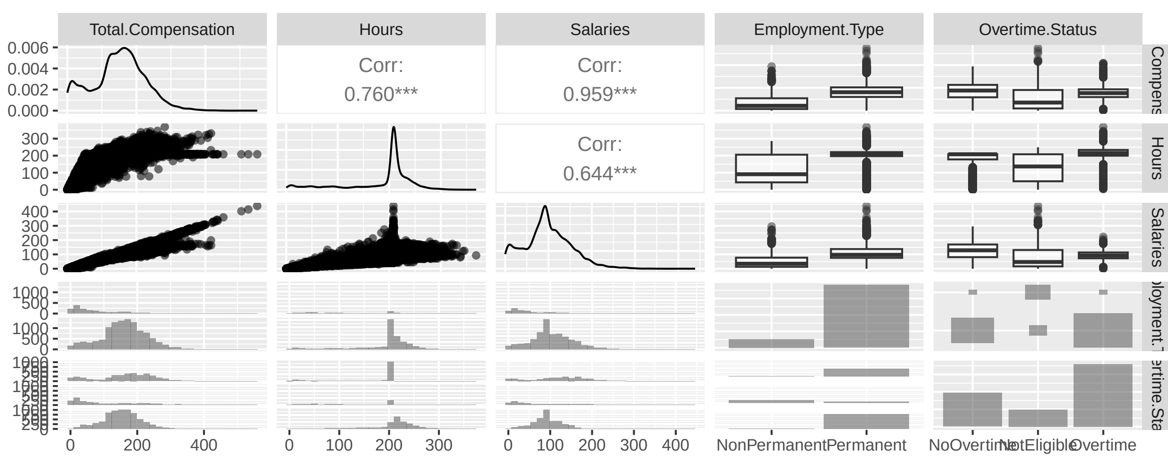
Generally speaking, most quantitative predictors were positively correlated with total salary, and most categorical predictors contextually related to total compensation as expected (ex: retirement benefits were indicative of higher average total compensations). Even though log(Salaries) is the most significant predictor for total compensation, the inclusion of base salaries introduced multicollinearity, likely because it is a large component of compensation. This redundancy inflates variance in coefficient estimates, as seen in the residual plots.

The dataset reflects real-world constraints in data sampling. While the it provided some predictors, it lacked more critical information (ex: education, demographics, etc) that likely influence compensation. Because of these omissions in the data, we had to use available data and rigorously compare models to identify the most reliable. Our report shows while perfect data is rarely available, rigorous statistical methods can still yield valuable insights. Future research could explore further data sampling to incorporate additional predictors to better capture nuances in employment. Even within its limitations, this analysis provides a solid understanding the some drivers (hours, employment type, overtime status, retirement, and department) of public employee compensation in San Francisco.

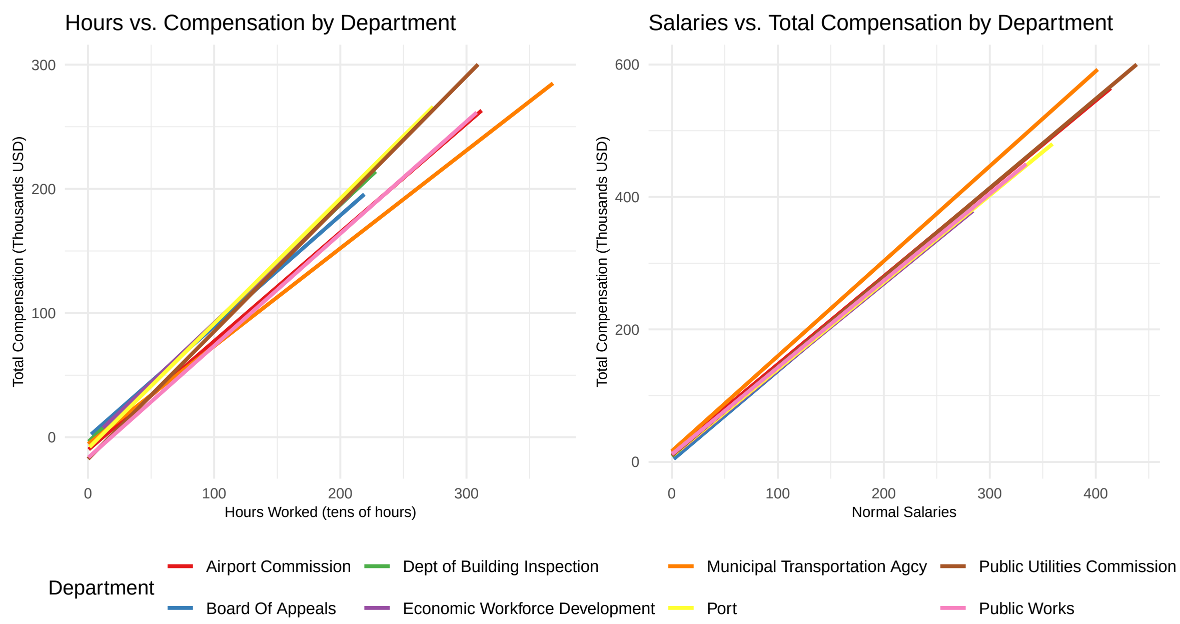
Appendix

Additional EDA

We observe that compensation levels vary across different employment types, suggesting an uneven distribution of compensation. The median compensation is higher for permanent compared to non permanent employment types.



More Interaction Effects



Nested F-Test For Interaction Terms

Nested F Test for `Overtime.Status*Log.Hours`

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
13055	351.222	NA	NA	NA	NA
13053	323.989	2	27.233	548.582	0

Nested F Test for `Overtime.Status*Log.Salaries`

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
13055	351.222	NA	NA	NA	NA
13053	321.653	2	29.569	599.973	0

AIC Analysis

AIC for model with none of the new variables:

```
# A tibble: 1 x 1
      AIC
    <dbl>
1 -11721.
```

AIC for model with Department added:

```
# A tibble: 1 x 2
      AIC     BIC
    <dbl>   <dbl>
1 -11869. -11741.
```

AIC for model with Retirement added:

```
# A tibble: 1 x 2
      AIC     BIC
    <dbl>   <dbl>
1 -13315. -13232.
```

AIC for model with Employment Type added:

```
# A tibble: 1 x 2
      AIC     BIC
    <dbl>   <dbl>
1 -12171. -12089.
```

AIC for model with Retirement, Employment Type, and Department added:

```
# A tibble: 1 x 2
      AIC     BIC
    <dbl>   <dbl>
1 -13443. -13301.
```

AIC for Retirement and Department added:

```
# A tibble: 1 x 2
      AIC     BIC
    <dbl>   <dbl>
1 -13435. -13300.
```

AIC for Retirement and Employment Type added:

```
# A tibble: 1 x 2
      AIC     BIC
    <dbl>   <dbl>
1 -13328. -13238.
```

AIC for Department and Employment Type added:

```
# A tibble: 1 x 2
      AIC     BIC
    <dbl>   <dbl>
1 -12286. -12152.
```

Testing Model 1 + interaction terms

Model Output when including

`Log.Salaries*Overtime.Status` **and** `Log.Hours*Overtime.Status`.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.938	0.022	-42.993	0.000
Log.Hours	1.006	0.004	241.611	0.000
Overtime.StatusNotEligible	0.004	0.012	0.369	0.712
Overtime.StatusOvertime	-0.186	0.007	-26.106	0.000
RetirementBenefits	0.369	0.017	21.433	0.000
Employment.TypePermanent	0.401	0.015	27.150	0.000
DepartmentBoard Of Appeals	0.416	0.122	3.409	0.001
DepartmentDept of Building Inspection	0.073	0.020	3.672	0.000
DepartmentEconomic Workforce Development	0.110	0.028	3.972	0.000
DepartmentMunicipal Transportation Agcy	-0.025	0.009	-2.908	0.004
DepartmentPort	0.187	0.021	9.081	0.000
DepartmentPublic Utilities Commission	0.136	0.010	13.573	0.000
DepartmentPublic Works	0.030	0.011	2.693	0.007

VIF

Notice very high VIF for Hours, Overtime, Salaries, and all interaction terms.

	x
Log.Hours	30.302691
Overtime.StatusNotEligible	41.950284
Overtime.StatusOvertime	107.670362
Log.Salaries	29.781856
RetirementBenefits	2.966575
Employment.TypePermanent	3.470890
DepartmentBoard Of Appeals	1.019167
DepartmentDept of Building Inspection	1.171710
DepartmentEconomic Workforce Development	1.150679
DepartmentMunicipal Transportation Agcy	2.433739
DepartmentPort	1.154154
DepartmentPublic Utilities Commission	2.035919
DepartmentPublic Works	1.759346
Overtime.StatusNotEligible:Log.Salaries	126.925208
Overtime.StatusOvertime:Log.Salaries	301.881266

	x
Log.Hours:Overtime.StatusNotEligible	239.613681
Log.Hours:Overtime.StatusOvertime	589.909355

Adjusted R Squared

```
[1] 0.9774542
```

Testing Model 2 + interaction terms

Model Output when including `Log.Hours*Overtime.Status`

term	estimate	std.error	statistic	p.value
(Intercept)	-0.938	0.022	-42.993	0.000
Log.Hours	1.006	0.004	241.611	0.000
Overtime.StatusNotEligible	0.004	0.012	0.369	0.712
Overtime.StatusOvertime	-0.186	0.007	-26.106	0.000
RetirementBenefits	0.369	0.017	21.433	0.000
Employment.TypePermanent	0.401	0.015	27.150	0.000
DepartmentBoard Of Appeals	0.416	0.122	3.409	0.001
DepartmentDept of Building Inspection	0.073	0.020	3.672	0.000
DepartmentEconomic Workforce Development	0.110	0.028	3.972	0.000
DepartmentMunicipal Transportation Agcy	-0.025	0.009	-2.908	0.004
DepartmentPort	0.187	0.021	9.081	0.000
DepartmentPublic Utilities Commission	0.136	0.010	13.573	0.000
DepartmentPublic Works	0.030	0.011	2.693	0.007

VIF

Notice very high VIF for the Overtime variable, and the interaction terms.

	x
Log.Hours	2.424180
Overtime.StatusNotEligible	33.749502
Overtime.StatusOvertime	91.637640
RetirementBenefits	2.962155
Employment.TypePermanent	3.035438
DepartmentBoard Of Appeals	1.016523
DepartmentDept of Building Inspection	1.163147
DepartmentEconomic Workforce Development	1.149149
DepartmentMunicipal Transportation Agcy	2.409020
DepartmentPort	1.144680
DepartmentPublic Utilities Commission	1.998608
DepartmentPublic Works	1.752881
Log.Hours:Overtime.StatusNotEligible	29.875730
Log.Hours:Overtime.StatusOvertime	95.143474

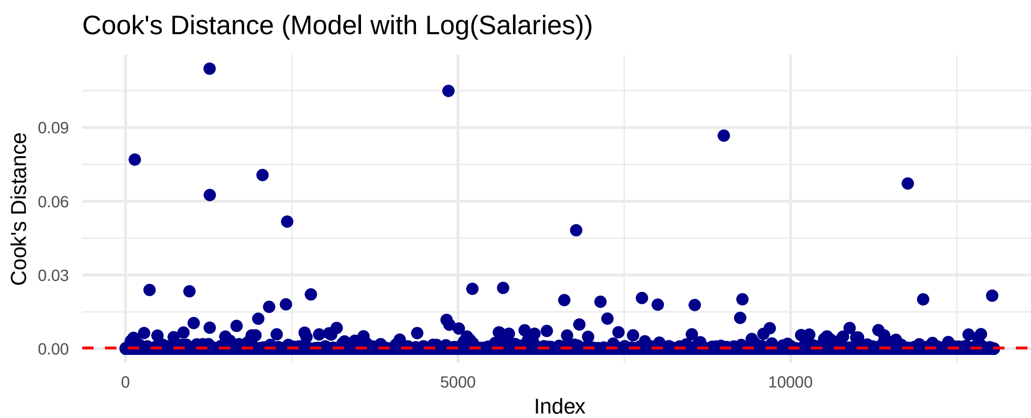
Adjusted R Squared

```
[1] 0.8890332
```

Model Diagnostics: Final Model 1

Variable	VIF
Log.Salaries	9.263
Log.Hours	9.335
Overtime.StatusNotEligible	2.402
Overtime.StatusOvertime	1.892
RetirementBenefits	2.327
Employment.TypePermanent	3.146
DepartmentBoard Of Appeals	1.011
DepartmentDept of Building Inspection	1.164
DepartmentEconomic Workforce Development	1.146
DepartmentMunicipal Transportation Agcy	2.420
DepartmentPort	1.152
DepartmentPublic Utilities Commission	2.025
DepartmentPublic Works	1.751

Cook's Distance



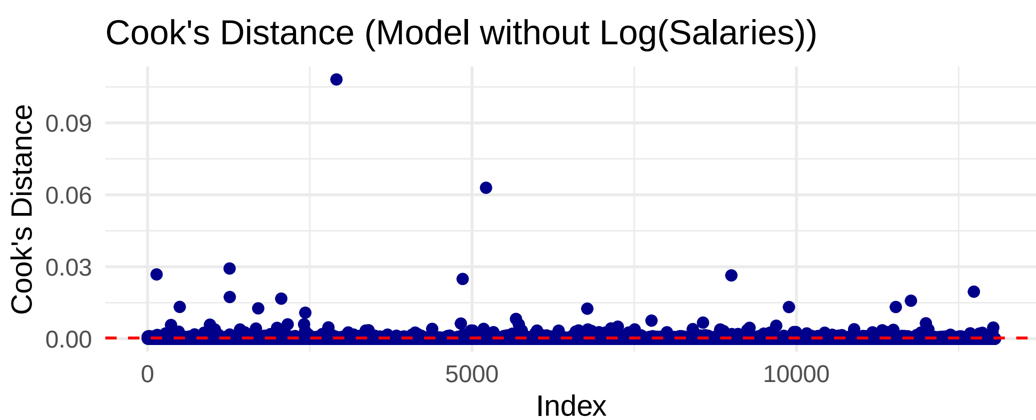
The Cook's distance plot reveals no particular influential points, with the highest values reaching approximately 0.12. The majority of observations have Cook's distance values well below 0.05, clustering near zero. A few notable spikes appear at different indices, particularly around indices 0-2500, indicating some potentially influential observations in the early part of the dataset. However, none of these points exceed the traditional threshold of 1.0 for high influence, suggesting that while there are some moderately influential points, none are severely impacting the model estimates. Noting that Cook's distance measure both point leverage and how well the observation fits the general trend of the data, we can say the observations adhere to the model's trend well.

Model Diagnostics: Final Model 2

VIF

Variable	VIF
Log.Hours	1.467
Overtime.StatusNotEligible	2.402
Overtime.StatusOvertime	1.594
RetirementBenefits	2.326
Employment.TypePermanent	2.960
DepartmentBoard Of Appeals	1.009
DepartmentDept of Building Inspection	1.163
DepartmentEconomic Workforce Development	1.145
DepartmentMunicipal Transportation Agcy	2.405
DepartmentPort	1.142
DepartmentPublic Utilities Commission	1.997
DepartmentPublic Works	1.751

Cook's Distance



Multicollinearity

Variable	VIF
Employment.TypePermanent	3.137
Hours	2.842
Salaries	2.578
DepartmentMunicipal Transportation Agcy	2.417
Overtime.StatusNotEligible	2.414
RetirementBenefits	2.272
DepartmentPublic Utilities Commission	2.013
Overtime.StatusOvertime	1.997
DepartmentPublic Works	1.751
DepartmentDept of Building Inspection	1.164
DepartmentPort	1.147
DepartmentEconomic Workforce Development	1.145
DepartmentBoard Of Appeals	1.007

Based on a model with no interaction terms, there does not seem to be evidence of multicollinearity among the variables. The Variance Inflation Factor (VIF) values of all variables are below the threshold of 10.

References

Andrzejewski, Adam. 2021. *Forbes*. https://www.forbes.com/sites/adamandrzejewski/2020/09/01/why-san-francisco-is-in-trouble--19000-highly-compensated-city-employees-earned-150000-in-pay--perks/.

San Francisco Controller's Office. 2024. "Employee Compensation Dataset." https://data.sfgov.org/City-Management-and-Ethics/Employee-Compensation/88g8-5mnd/about_data.