TOTAL COMPENSATION ANALYSIS

Problem Statement

To analyse and identify Discrepancies in Compensation and Benefits for Employees working in various government organisations in the city of San Francisco, California.

The study aims to investigate the variations in salaries, benefits, and total compensation among different job positions, unions, and job families within the departments of these organisations.

The analysis will focus on understanding the factors contributing to these discrepancies, such as overtime, other benefits, and union affiliations, with the goal of ensuring fair and equitable compensation for all employees.

Business Context

- Equity and Fairness in Compensation: The primary goal is to ensure that compensation (including salaries and benefits) is fair and equitable across different job positions, departments, and union affiliations within the city's government organizations.
- Understanding Influencing Factors: To analyse how factors like overtime, other benefits, and union memberships contribute to variations in total compensation.

Introduction to the dataset

Organization Group Code	**4**	4
Job Family Code	2300	2700
Job Code	2320	2736
Year Type	Fiscal	Fiscal
Year	2022	2022
Organization Group	Community Health	Community Health
Department Code	DPH	DPH
Department	Public Health	Public Health
Union Code	791	250
Union	SEIU, Local 1021, RN	SEIU, Local 1021, Misc
Job Family	Nursing	Housekeeping & Laundry
Job	Registered Nurse	Porter
Employee Identifier	49309201	49282706
Salaries	95699.03	70468.4
Overtime	44216.85	37251.49
Other Salaries	5048.44	3367.87
Total Salary	144964.32	111087.76
Retirement	16165.6	14984.19
Health and Dental	0	0
Other Benefits	10280.79	8668.68
Total Benefits	26446.39	23652.87
Total Compensation	171410.71	134740.63

Variables: 22

Records: 4500

Below is a comprehensible detail about each dataset along with its datatypes.

1. **Organization Group Code**:
   Datatype - Categorical
   A code representing the organizational group to which the employee belongs.

2. **Job Family Code**:
   Datatype -Categorical
   A code specifying the job family to which the employee's job belongs.

3. **Job Code**:
   Datatype -Categorical
   A code indicating the specific job or role of the employee.

4. **Year Type**:
   Datatype -Categorical
   The type of year (e.g., "Fiscal" or "Calendar") to which the data pertains

5. **Year**:
   Datatype -Numerical
   The specific year for which the data is recorded.

6. **Organization Group**:
   Datatype -Categorical
   The name or label of the organizational group to which the employee belongs.

7. **Department Code**:
   Datatype -Categorical
   A code representing the department in which the employee works.

8. **Department**:
   Datatype - Categorical
   The name or label of the department in which the employee works.

9. **Union Code**:
   Datatype -Categorical
   A code identifying the employee's union affiliation.

10. **Union**:
    Datatype -Categorical
    The name or label of the union to which the employee belongs.

11. **Job Family**:
    Datatype -Categorical
    The name or label of the job family to which the employee's job belongs.

12. **Job**:
    Datatype -Categorical
    The name or label of the specific job or role of the employee.

13. **Employee Identifier**:
    Datatype -Numerical
    A unique identifier for each employee.

14. **Salaries**:
    Datatype -Numerical
    The amount of money paid as the base salary to the employee

15. **Overtime**:
    Datatype -Numerical
    The amount of money paid as overtime compensation to the employee

16. **Other Salaries**:
    Datatype -Numerical
    Additional salary payments or compensations received by the employee

17. **Total Salary**:
    Datatype -Numerical
    The total amount of money paid to the employee as salary

18. **Retirement**:
    Datatype -Numerical
    The amount allocated for the employee's retirement benefits

19. **Health and Dental**:
    Datatype -Numerical
    The amount allocated for the employee's health and dental benefits

20. **Other Benefits**:
    Datatype -Numerical
    Other benefits provided to the employee in addition to salary and retirement benefits.

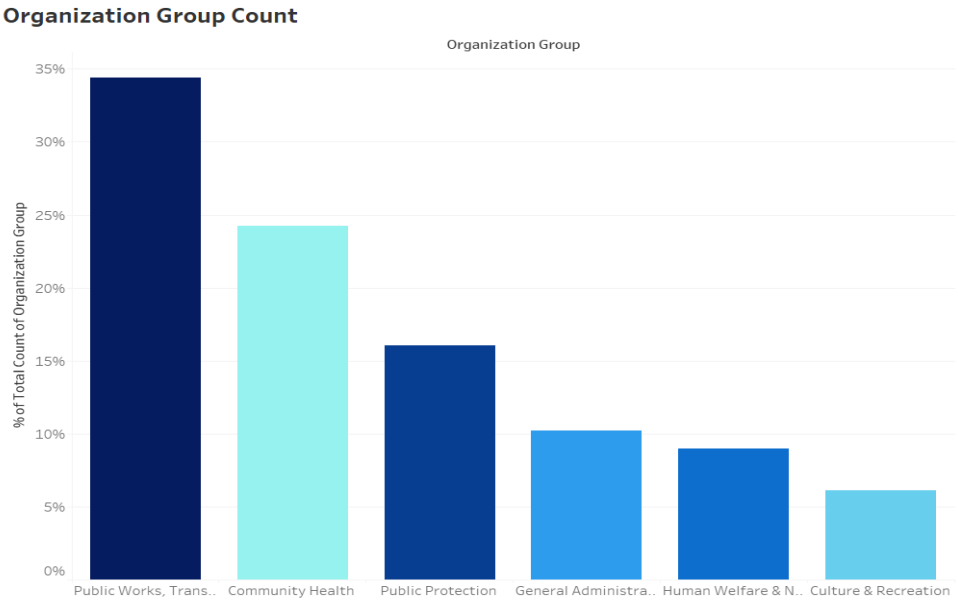
21. **Total Benefits**:
    Datatype -Numerical
    The total value of benefits provided to the employee, including retirement, health, dental, and other benefits.

22. **Total Compensation**:
    Datatype -Numerical
    The overall compensation package for the employee, including salary and all benefits.

Exploratory Data Analysis

Univariate Analysis

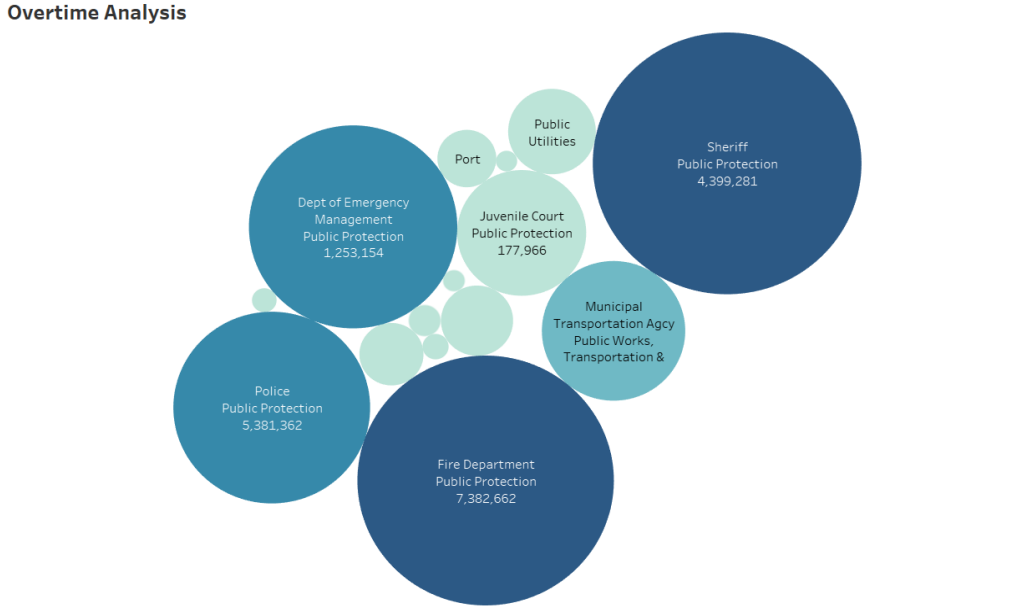
Number of Employees working in each Organization Groups.



Insights: Total Employees is Highest in Public Works, transportation & commerce.

Hypothesis & Insights

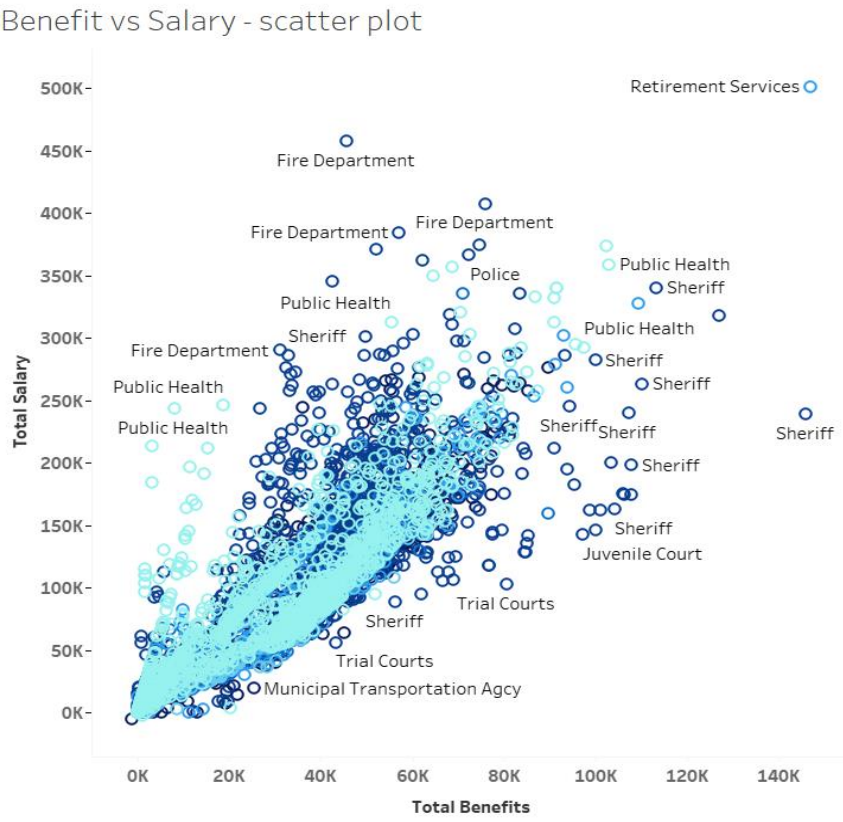
1. **Hypothesis**: Overworked Departments are mainly front-line workers (Multivariate Analysis)



Insights:

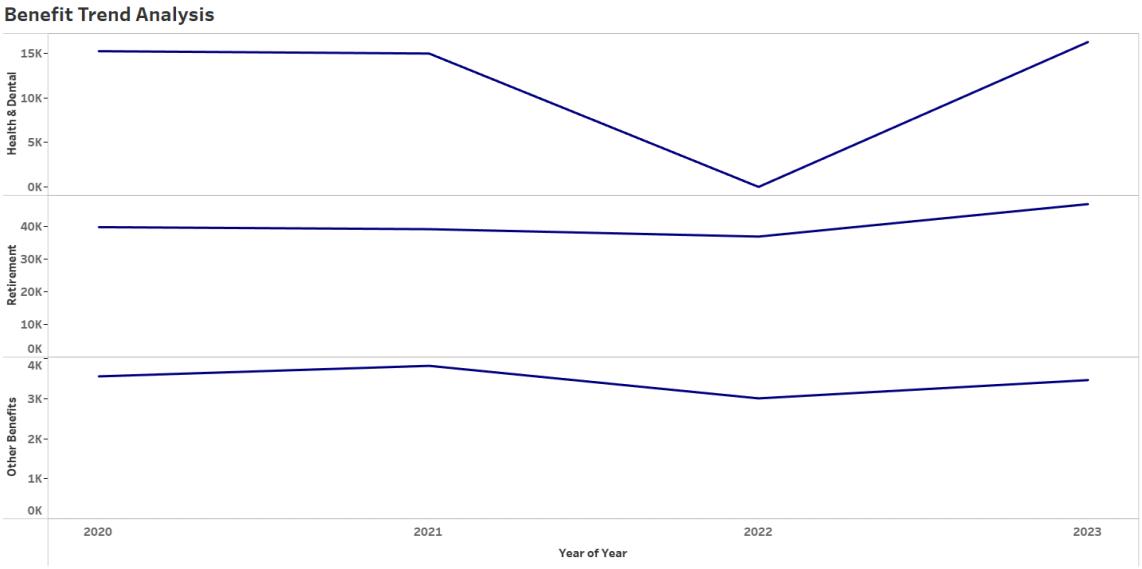
Top 4 departments providing highest average overtime are from public protection Organization groups and are namely **Sheriff, Fire department, Department of emergency management and police**

2. **Hypothesis**: Prestigious or long-term Jobs are given more Benefits (Multivariate Analysis)



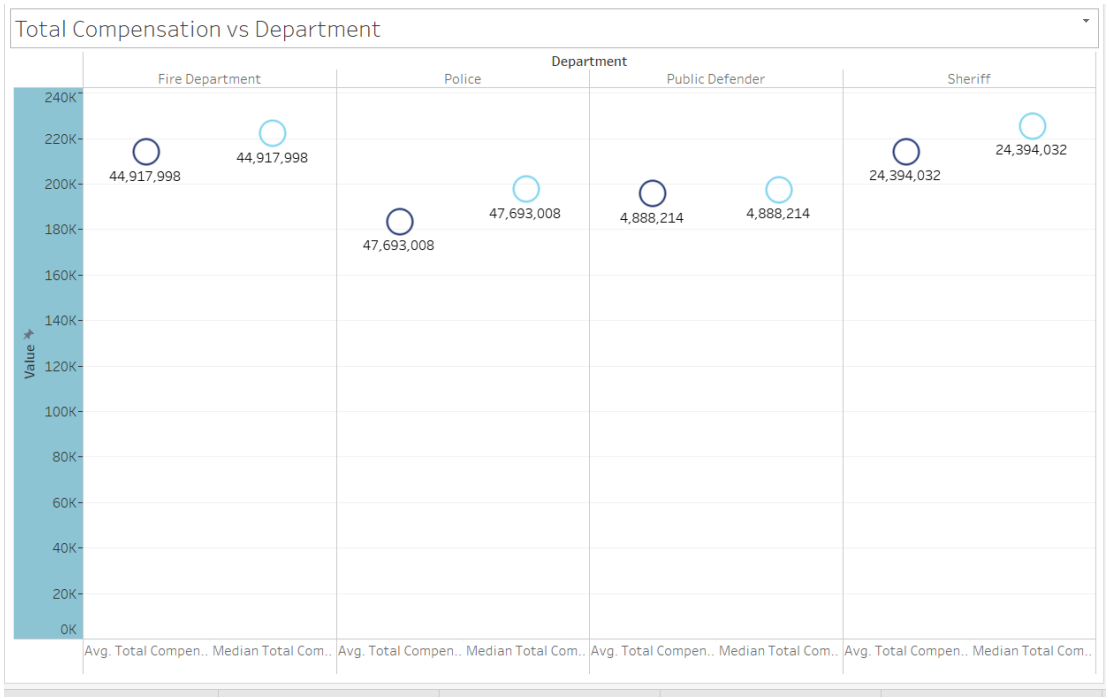
Insights: Sheriff is offered high Benefits and low salary

3. **Hypothesis**: There has been increasing trend in Employees covered in Benefits (Bivariate Analysis)



Insights: Sudden drop in Health & dental Benefits in 2022 might be because of Covid-19

4. **Hypothesis**: Average Total Compensation is Highest for Front Line Workers: (Bivariate Analysis)



Insights: Above represents the Top 4 Departments i.e.

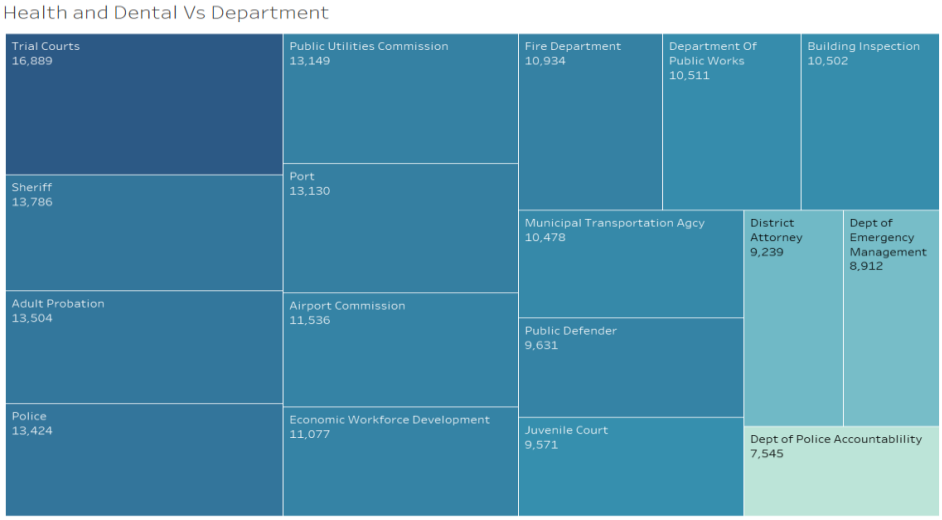
›Fire Department

›Police

›Public Defender

›Sheriff

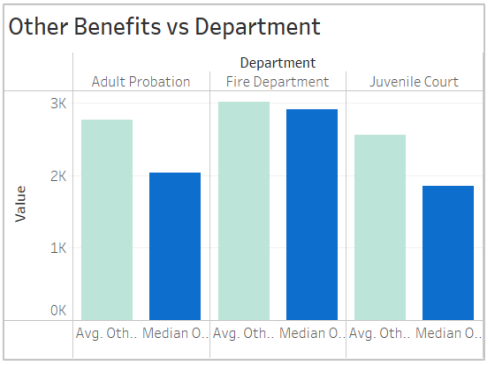
5. **Hypothesis**: Stressful Jobs have more Health Benefits (Bivariate Analysis)

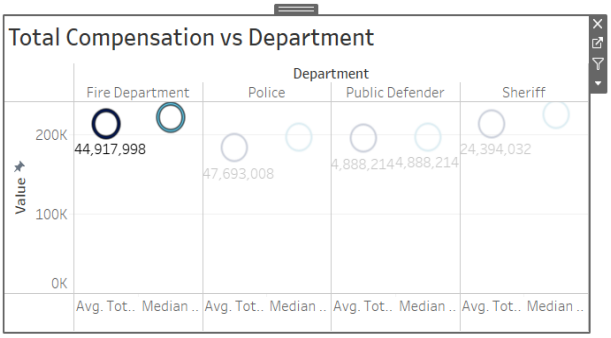


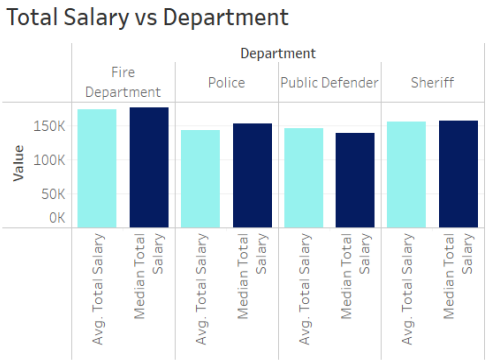
Insights:

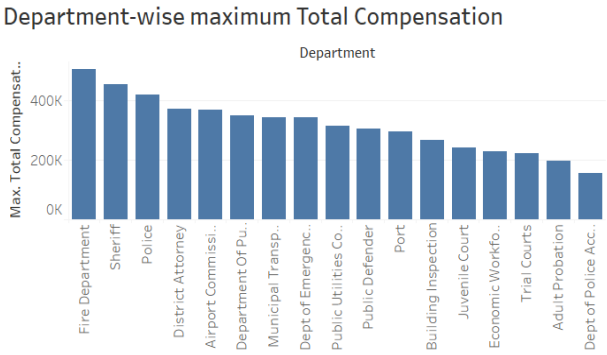
Trial court judges are exposed to more traumatic & stressful conditions hence they might be paid Higher Health & Dental Benefits.

6. **Hypothesis**: High risk jobs are provided with Higher Salary & Benefits (Bivariate Analysis)









Insights: Fire Department a High risk Department have High average Other Benefits, Total Compensation, Total Salary & maximum Total Compensation

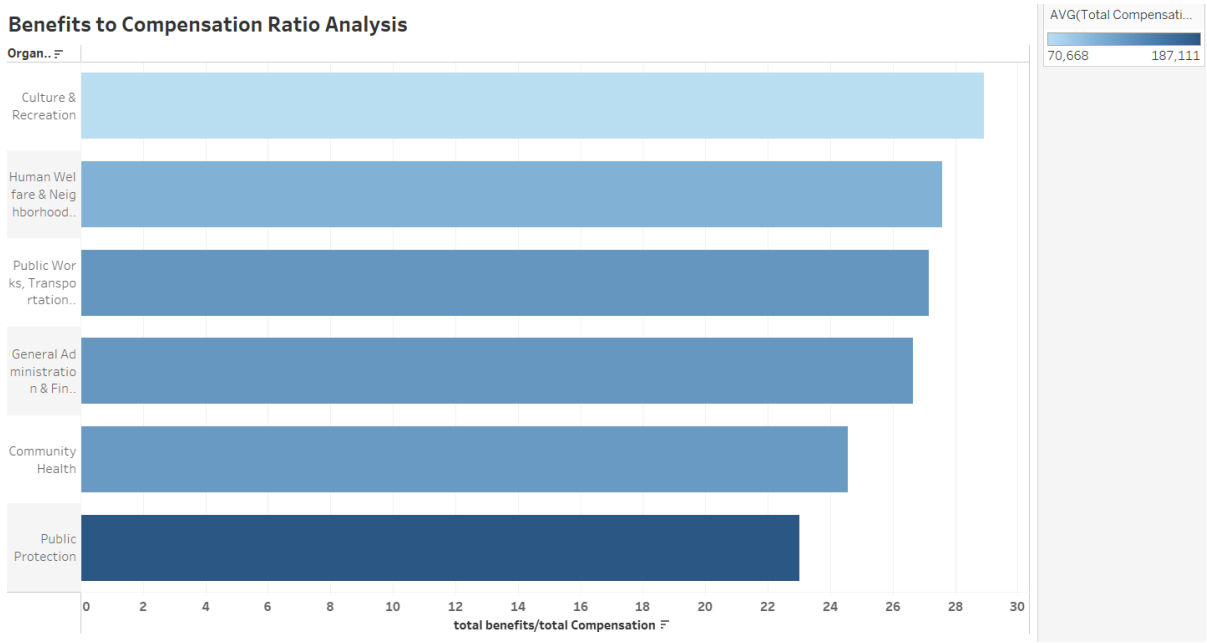
7. **Hypothesis**: Departments prioritize direct monetary compensation (salary) over non-monetary benefits (Benefits)

Composition Of Benefits to total

Departme..	Avg. Total Com..	total benefit..	total salary/tot..	Avg. Total Bene..
Trial Courts	115,676	33	67	38,093
Juvenile Court	140,738	31	69	43,378
Port	138,243	30	70	40,978
Adult Probation	135,802	30	70	40,149
Public Utilities Commission	149,376	28	72	41,762

Insights: Trial courts, is having the highest percentage, average total compensation offered is the lowest among top 5.

Similar trend can be observed in Organization Groups that inspite of benefits provided were being highest for **Culture and recreation,** the average total compensation offered is the lowest among 5.



Recommendations and Interventions:

1. Overworked Departments are Mainly Front-Line Workers:

- Intervention: Implement workload management strategies, such as optimizing shifts, hiring additional staff, or redistributing tasks to ensure a balanced workload.

- Recommendation Conduct regular assessments of workload distribution and adjust staffing levels accordingly.

2. Prestigious or Long-Term Jobs are Given More Benefits

- Intervention Review and adjust the benefits packages for jobs perceived as prestigious or long-term to ensure fairness and equity across all positions.

- Recommendation Conduct a comprehensive review of benefits across job categories and make adjustments to align with job responsibilities and market standards.

3. Increasing Trend in Employees Covered in Benefits

- Intervention Investigate the reasons behind the sudden drop in health and dental benefits in 2022, considering external factors such as the impact of COVID-19.

- Recommendation Develop contingency plans to address potential disruptions in benefits due to external factors, ensuring employee well-being is a priority.

4. Average Total Compensation is Highest for Front Line Workers

- Intervention Review and potentially adjust the compensation structure to ensure that front-line workers receive fair and competitive total compensation.

- Recommendation Conduct regular market salary surveys to benchmark compensation against industry standards and make adjustments as needed.

5. Stressful Jobs Have More Health Benefits

- Intervention Implement targeted wellness programs and support mechanisms for employees in stressful roles, focusing on mental health and stress reduction.

- Recommendation Provide stress management resources, counselling services, and promote a supportive work environment to address the unique challenges of high-stress positions.

6. High-Risk Jobs are Provided with Higher Salary & Benefits

- Intervention Conduct a thorough analysis of the compensation structure for high-risk jobs to ensure it aligns with the level of risk and responsibility involved.

- Recommendation Collaborate with relevant departments to regularly review and update compensation for high-risk roles based on industry benchmarks and risk assessments.

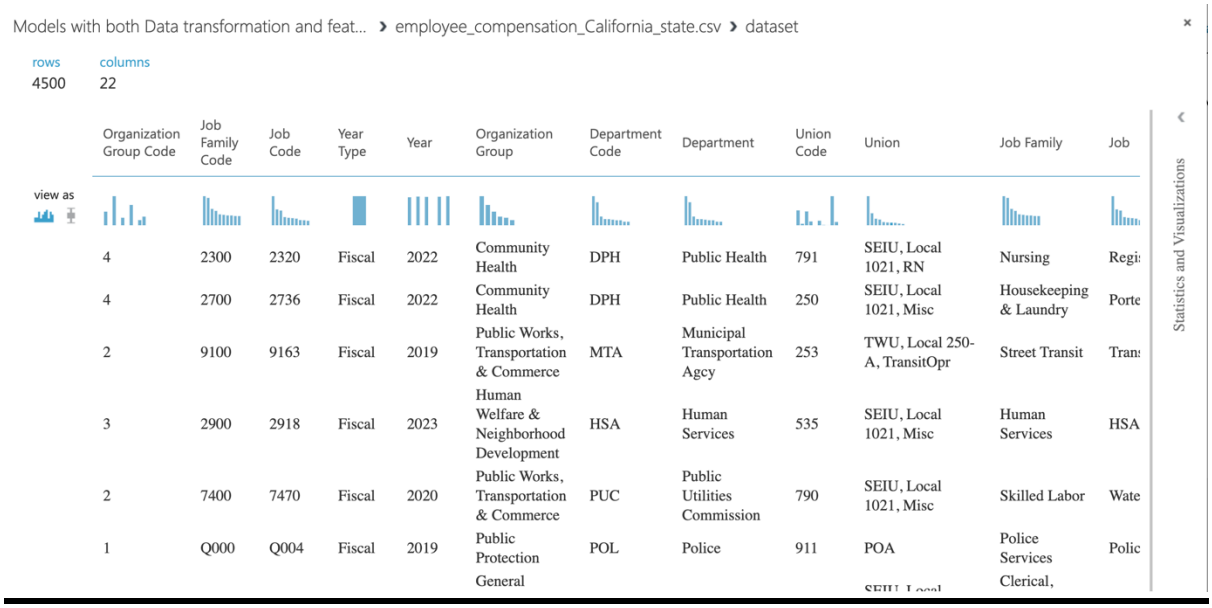
7. Departments Prioritize Direct Monetary Compensation Over Non-Monetary Benefits

- Intervention Educate departments on the importance of a balanced compensation and benefits approach to attract and retain top talent.

- Recommendation Conduct workshops and training sessions for departmental heads to emphasize the value of non-monetary benefits in employee satisfaction and retention.

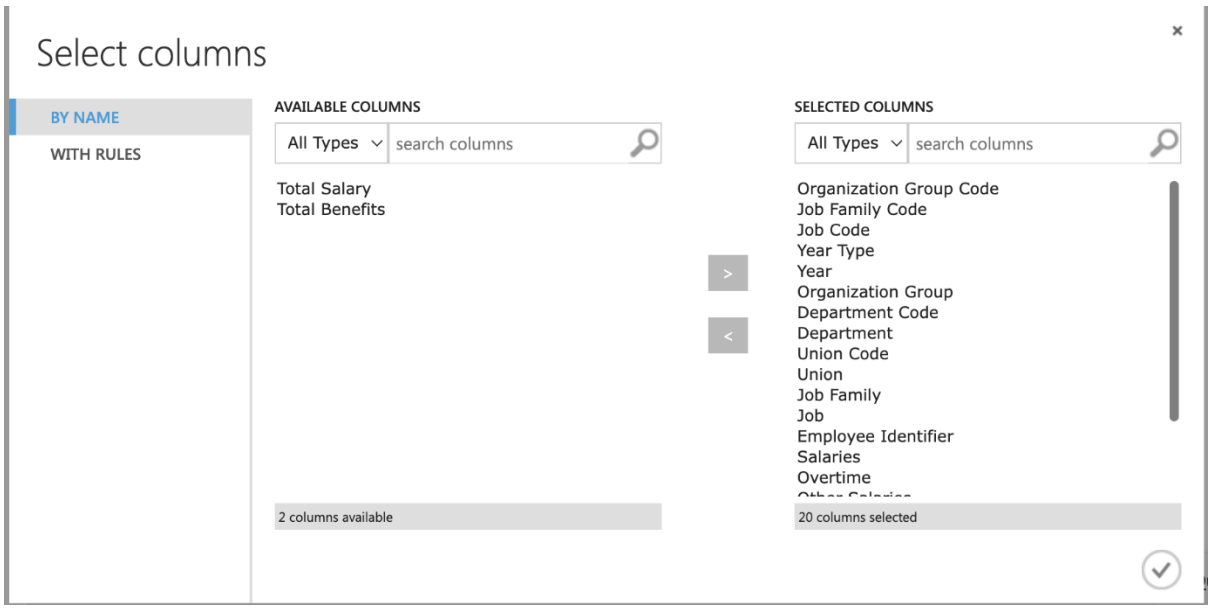
Azure – Model Building for Predictive Analysis

- Import the dataset
- Summarize the dataset to visualise the distribution of the dataset
- Splitting the dataset into 60% training and 30% testing dataset.

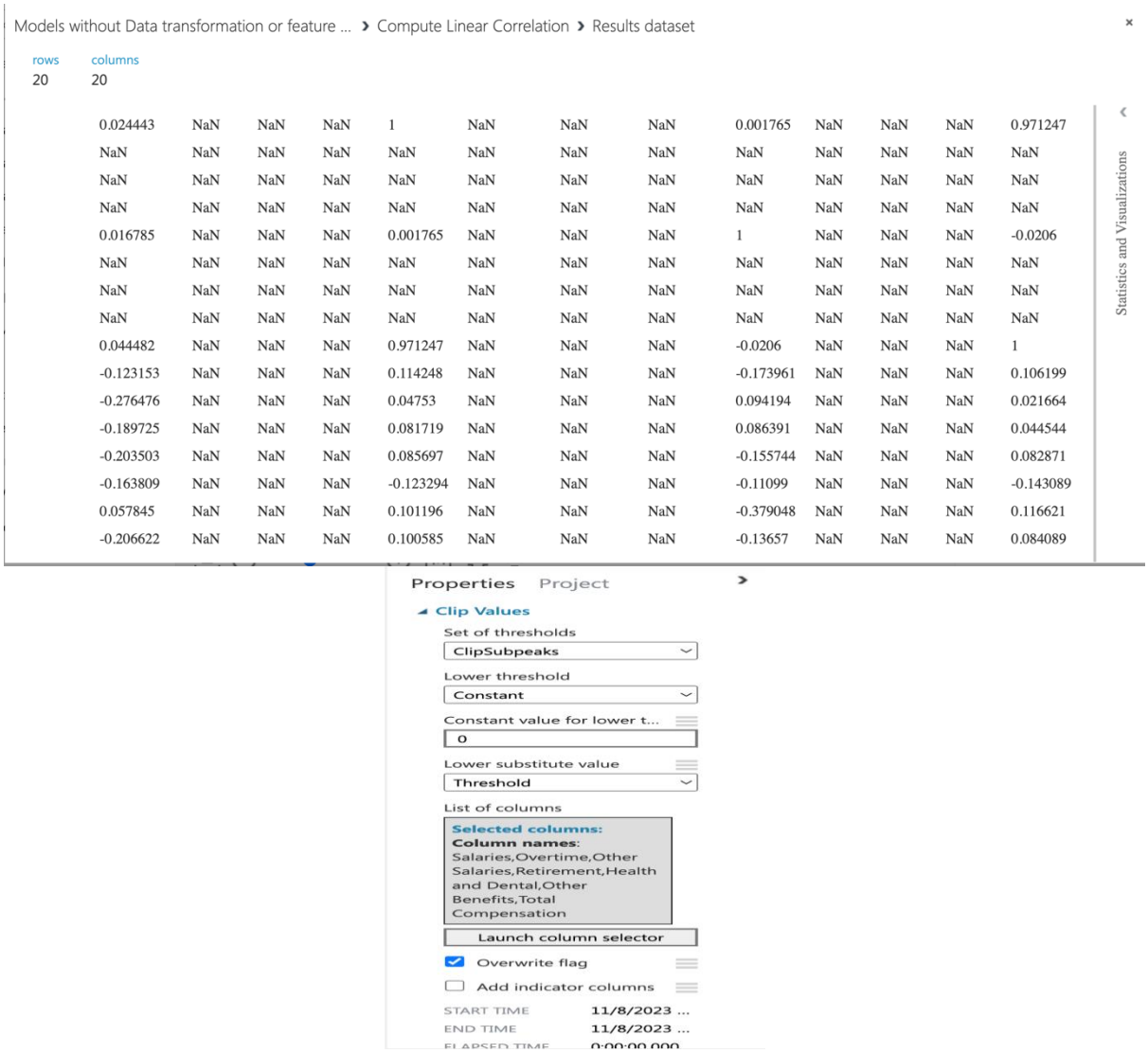


Organization Group Code	Job Family Code	Job Code	Year Type	Year	Organization Group	Department Code	Department	Union Code	Union	Job Family	Job
4	2300	2320	Fiscal	2022	Community Health	DPH	Public Health	791	SEIU, Local 1021, RN	Nursing	Regi
4	2700	2736	Fiscal	2022	Community Health	DPH	Public Health	250	SEIU, Local 1021, Misc	Housekeeping & Laundry	Porte
2	9100	9163	Fiscal	2019	Public Works, Transportation & Commerce	MTA	Municipal Transportation Agcy	253	TWU, Local 250-A, TransitOpr	Street Transit	Trans
3	2900	2918	Fiscal	2023	Human Welfare & Neighborhood Development	HSA	Human Services	535	SEIU, Local 1021, Misc	Human Services	HSA
2	7400	7470	Fiscal	2020	Public Works, Transportation & Commerce	PUC	Public Utilities Commission	790	SEIU, Local 1021, Misc	Skilled Labor	Wate
1	Q000	Q004	Fiscal	2019	Public Protection	POL	Police	911	POA	Police Services	Polic
					General					Clerical,	

Summary of Whole Dataset

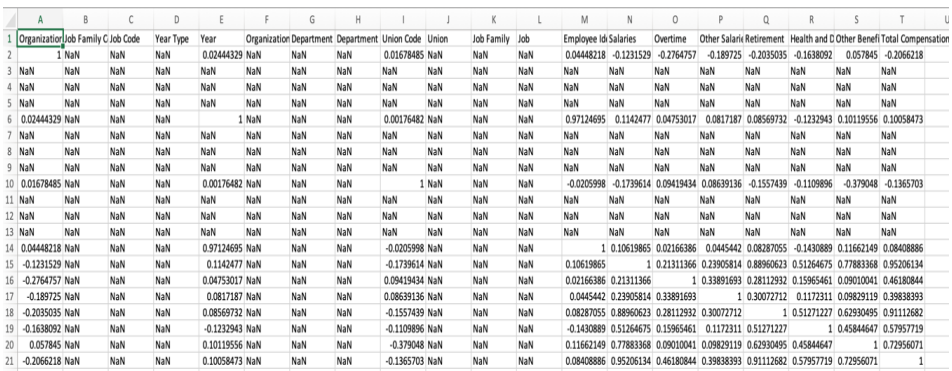


Selecting Columns for Dataset



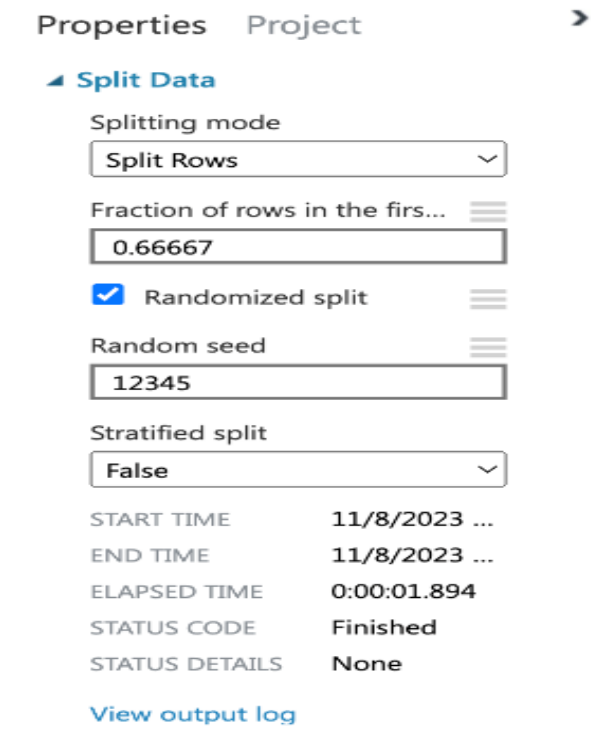
Clip Values

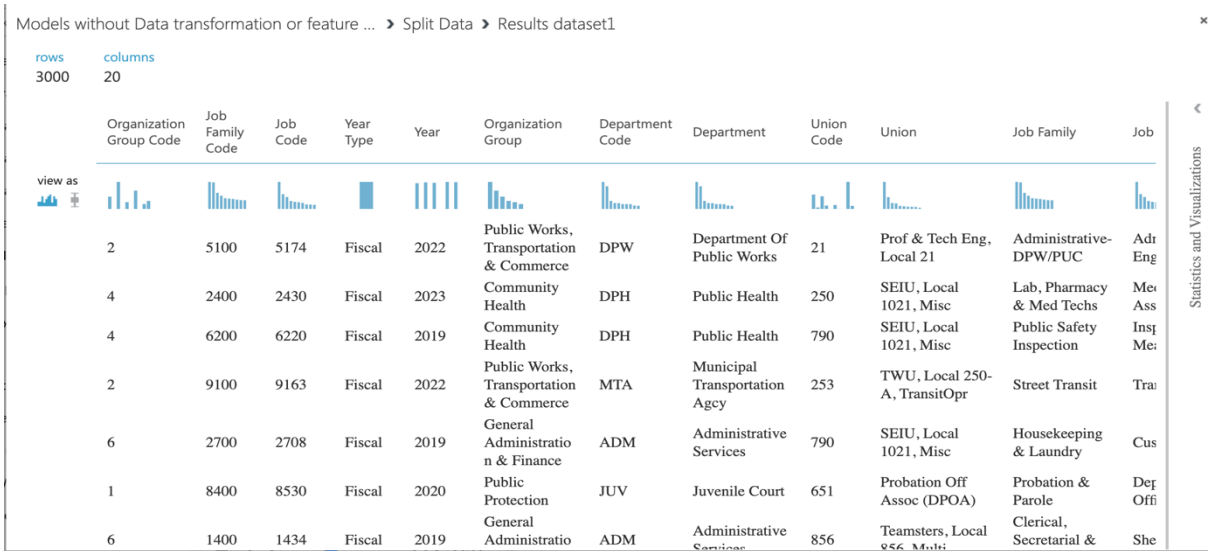
Correlation Matrix



Correlation Matrix

Splitting the Dataset





Summary of Training Dataset

Models without Data transformation or feature ... › Split Data › Results dataset2

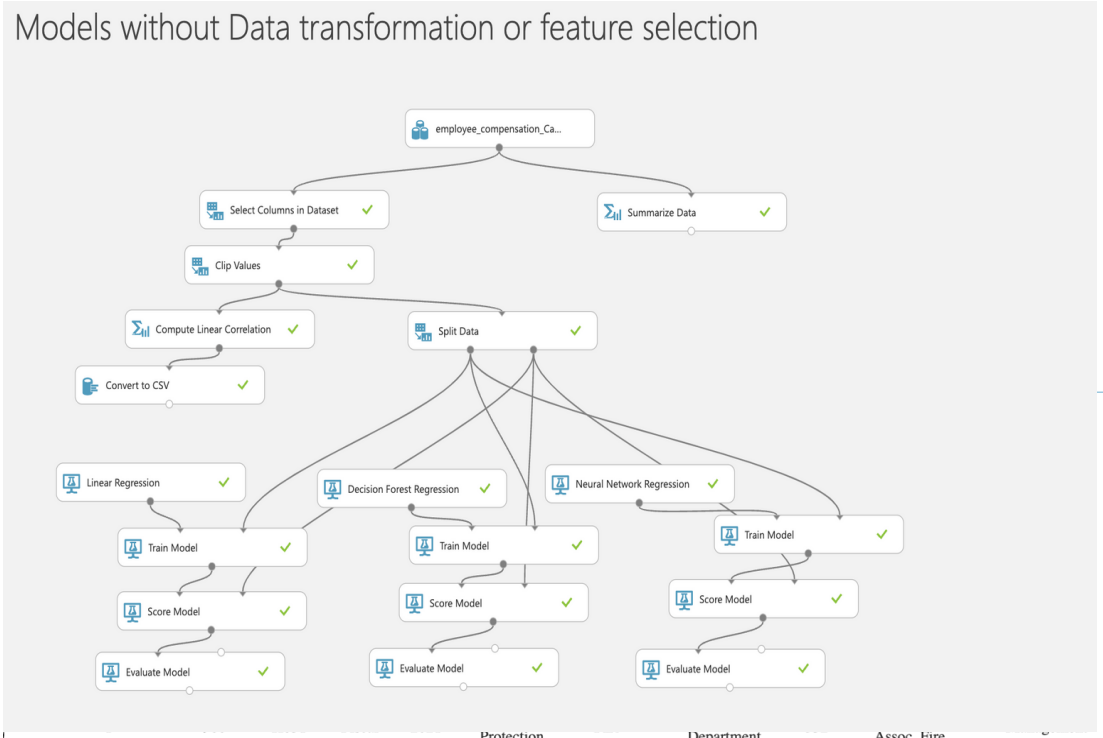
rows 1500 columns 20

	Organization Group Code	Job Family Code	Job Code	Year Type	Year	Organization Group	Department Code	Department	Union Code	Union	Job Family	Job
view as												
	1	8300	8300	Fiscal	2020	Public Protection	SHF	Sheriff	790	SEIU, Local 1021, Misc	Correction & Detention	Sheri
	2	2700	2718	Fiscal	2020	Public Works, Transportation & Commerce	AIR	Airport Commission	790	SEIU, Local 1021, Misc	Housekeeping & Laundry	Custo Supe
	1	8400	8530	Fiscal	2023	Public Protection	ADP	Adult Probation	651	Probation Off Assoc (DPOA)	Probation & Parole	Depu Proba Offic
	1	8100	8177	Fiscal	2022	Public Protection	DAT	District Attorney	311	Municipal Attorneys Assoc	Legal & Court	Attor (Civi
	4	2300	2320	Fiscal	2023	Community Health	DPH	Public Health	791	SEIU, Local 1021, RN	Nursing	Regis Nurs
	1	900	H051	Fiscal	2021	Public Protection	FIR	Fire Department	352	Municipal Exec Assoc, Fire	Management	Assis Depu
	6	7300	7344	Fiscal	2021	General Administration & Finance	ADM	Administrative Services	236	Carpenters, Local 22	Journeyman Trade	Carp

Statistics and Visualizations

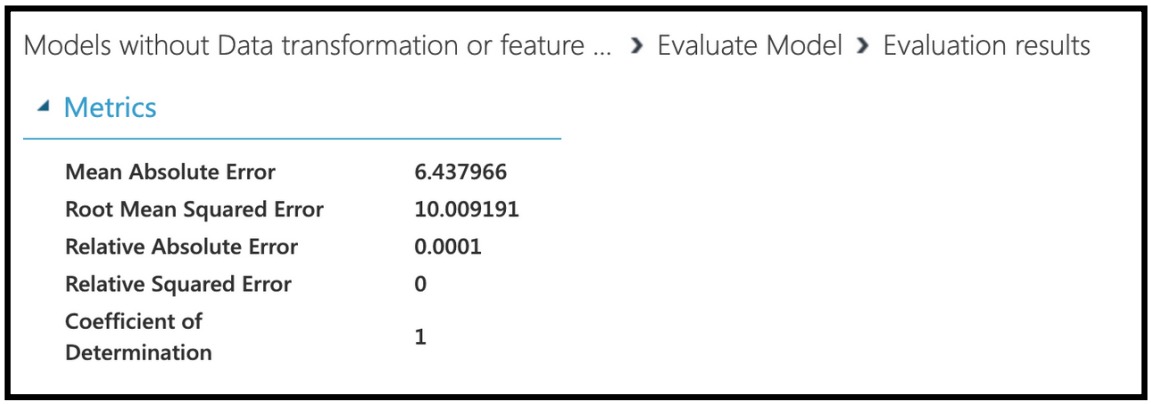
Summary of Testing Dataset

a) Without Data transformation or feature selection – Model 1



Without Data Transformation or Feature selection

Model Evaluation



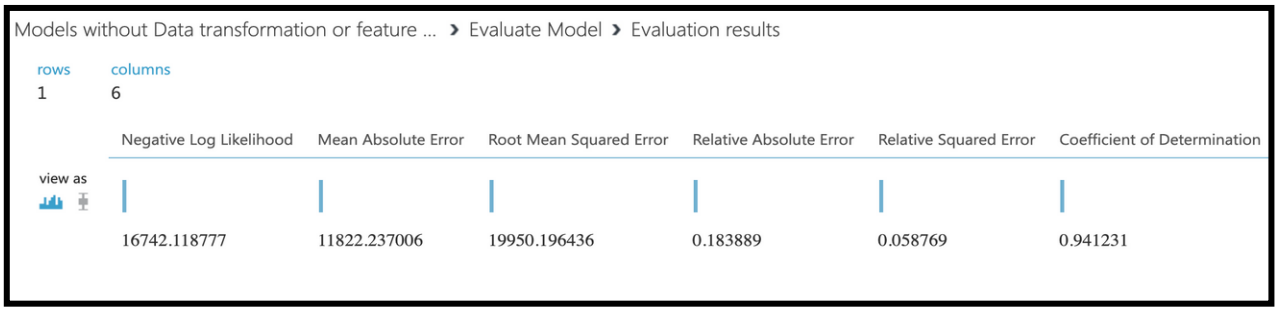
Linear Regression

Model Evaluation Summary:

- MAE (6.44) - On average, the model's predictions differ by approximately 6.44 units from the actual values.
- RMSE (10.01) - The square root of the average squared differences between predictions and actual values is 10.01, indicating the average magnitude of errors. A lower RMSE signifies better model accuracy.
- Relative Absolute Error (0.0001) - The small value suggests the model's predictions closely align with actual values on a relative scale.
- R^2 (1.0) : With an R^2 of 1.0, the model achieves a perfect fit, explaining 100% of the variance in the dependent variable. This signals excellent overall model performance.

Conclusion:

The model demonstrates high accuracy, minimal errors, and a perfect fit according to R^2. However, thorough validation within the specific problem context is crucial, considering potential overfitting and other factors influencing generalizability.

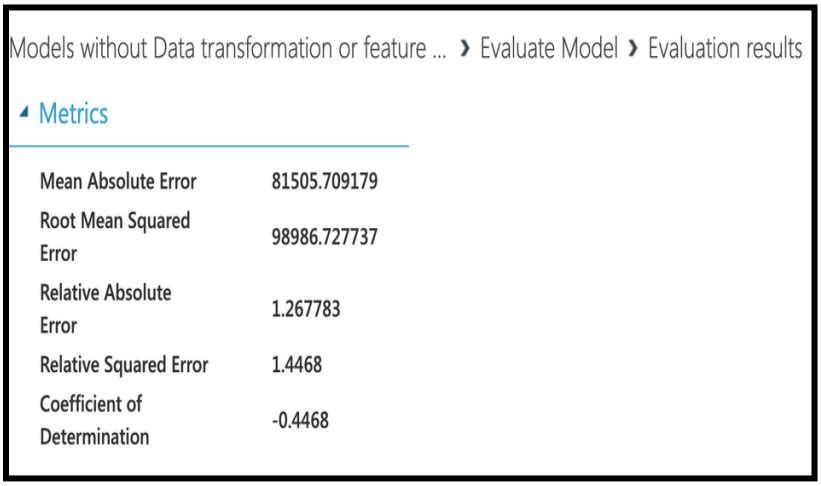


Decision Tree Regression

Model Evaluation Summary:

- Negative Log Likelihood: A measure of the model's likelihood to predict the observed values. A lower value is desirable, indicating higher likelihood.

- Mean Absolute Error (MAE): The average absolute difference between predicted and actual values is 16742.12. Lower values signify better model accuracy.
- Root Mean Squared Error (RMSE): The square root of the average squared differences between predictions and actual values is 11822.24. Lower RMSE indicates better accuracy with 11822.24 being the average magnitude of errors.
- Relative Absolute Error : Interpretation: Relative to the scale of the predicted and actual values, the error is 0.183889. A smaller value suggests closer alignment between predictions and actual values.
- Relative Squared Error : Interpretation: The error, relative to the squared scale of predictions and actual values, is 0.058769. A lower value indicates better model performance.
- Coefficient of Determination (R^2) : Interpretation: R^2 is 0.941231, indicating that the model explains 94.12% of the variance in the dependent variable. A high R^2 signifies a strong fit of the model to the data.



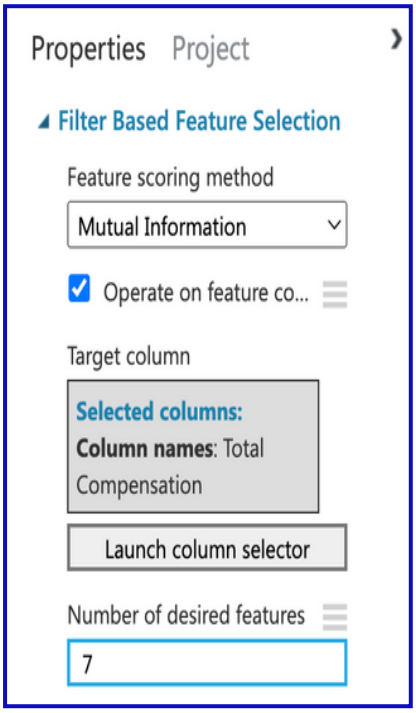
Neural Network

- Mean Absolute Error (MAE): 81505.71 On average, the model's predictions deviate by approximately 81505.71 units from the actual values.
- Root Mean Squared Error (RMSE): 98986.73 : The square root of the average squared differences between predictions and actual values is 98986.73. A lower RMSE indicates better model accuracy.
- Relative Absolute Error: 1.267783 : The error, relative to the scale of the predicted and actual values, is 1.267783. A smaller value suggests closer alignment between predictions and actual values.
- Relative Squared Error: 1.4468 : The error, relative to the squared scale of predictions and actual values, is 1.4468. A lower value indicates better model performance

- Coefficient of Determination (R^2): -0.4468 : The negative R^2 indicates that the model does not fit the data well, explaining less variance than a horizontal line. It implies a poor model fit.

Conclusion:The Neural Network Regression model, without data transformation or feature selection, demonstrates considerable errors and a negative R^2, suggesting a suboptimal fit to the data.

b. Model without Data Transformation But Feature Selection



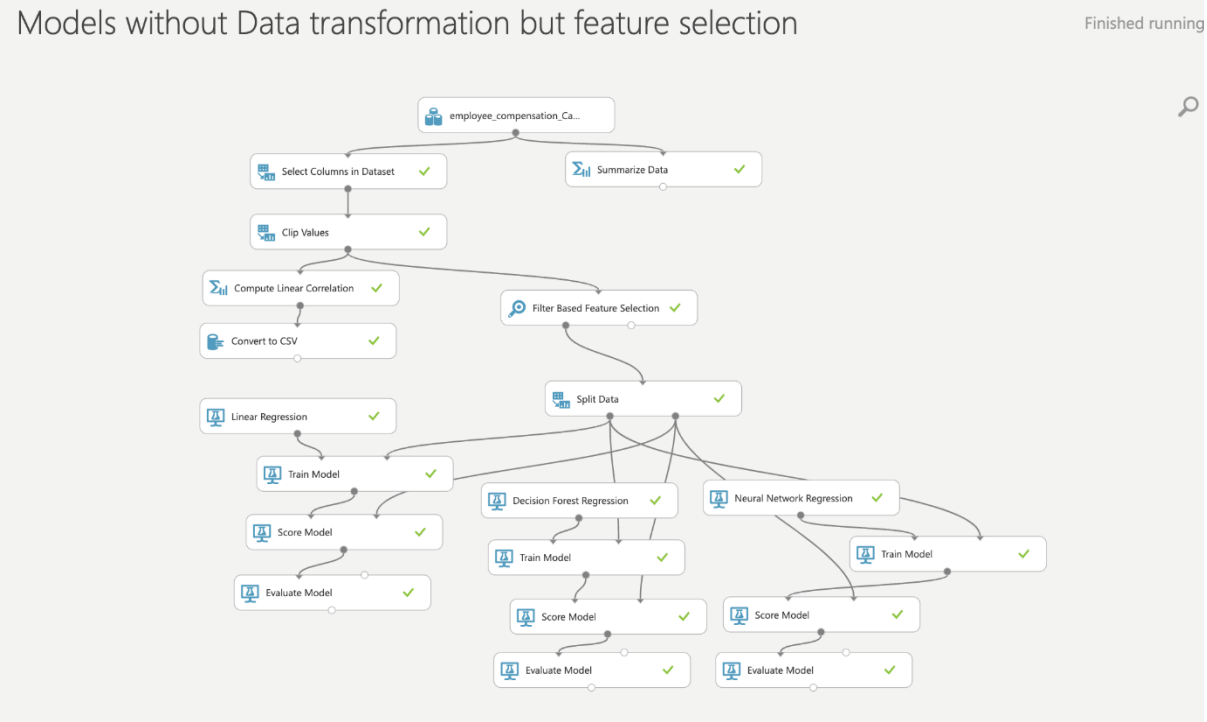
Total Compensation	Other Benefits	Salaries	Retirement	Health and Dental	Job Family Code	Job Family	Union
171410.71	10280.79	95699.03	16165.6	0	2300	Nursing	SEIU, Local 1021, RN
134740.63	8668.68	70468.4	14984.19	0	2700	Housekeeping & Laundry	SEIU, Local 1021, Misc
193310.97	11276.5	80054.13	17714.81	16495.5	9100	Street Transit	TWU, Local 250-A, TransitOpr
118524.24	6287.23	78514.79	14686.38	16719.41	2900	Human Services	SEIU, Local 1021, Misc
121341.19	6391.86	78437.83	16350.73	19012.79	7400	Skilled Labor	SEIU, Local 1021, Misc
228419.64	3083.35	129778.72	23729.69	15525.74	Q000	Police Services	POA
13485.81	971.32	9996.72	0	0	1400	Clerical, Secretarial & Steno	SEIU, Local 1021, Misc
211795.78	10334.79	140426.69	21598.5	16471.7	2300	Nursing	SEIU, Local 1021, RN
136719.44	7243.48	85432.01	17811.46	15088.45	1000	Information Systems	Prof & Tech Eng, Local 21
198220.44	11488.01	116464	26210.1	17191.44	7300	Journeyman Trade	Auto Machinist, Local 1414
78541.45	5561.74	69776.17	0	0	2300	Nursing	SEIU, Local 1021, RN

Features Selected are specified below:

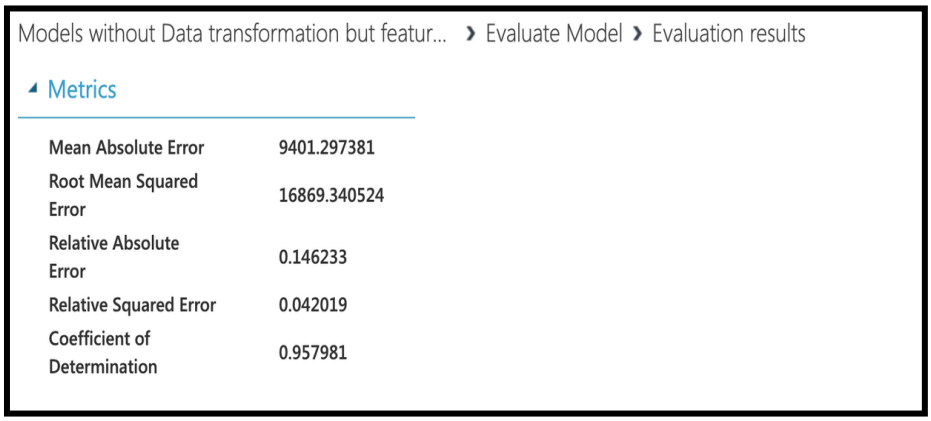
- Other Benefits
- Salaries

- Retirement
- Health and Dental
- Job Family Code
- Job Family

These top 7 features are determined by mutual information, measures the level of dependency between variables, highlighting those with the most informative power in predicting the target variable.



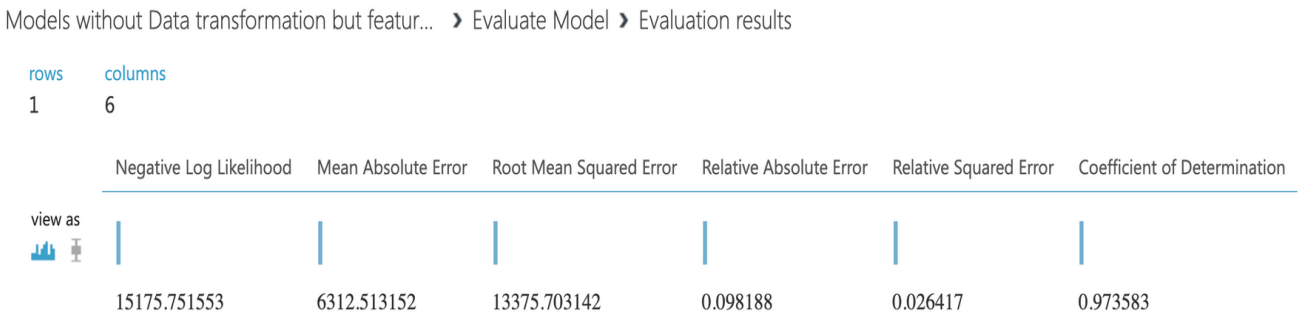
Evaluation Summary:



Linear Regression

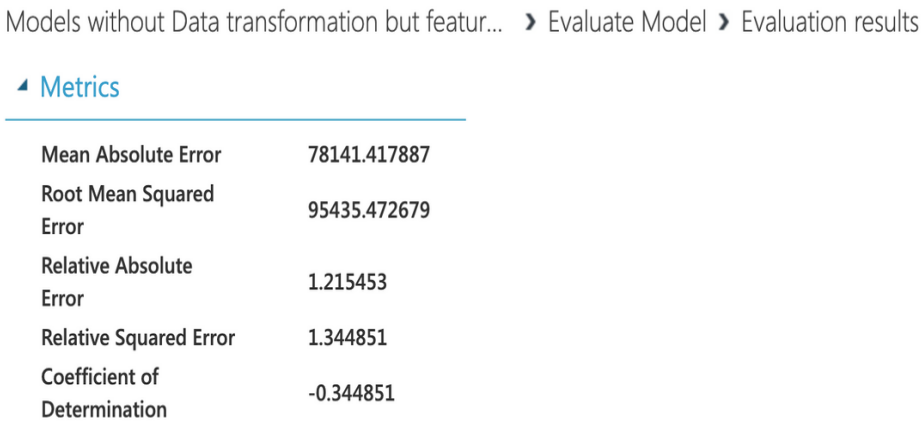
- Mean Absolute Error (MAE): 9401.30 : On average, the model's predictions deviate by approximately 9401.30 units from the actual values. A lower MAE signifies improved accuracy

- Root Mean Squared Error (RMSE): 16869.34 : The square root of the average squared differences between predictions and actual values is 16869.34. A lower RMSE indicates better accuracy.
- Relative Absolute Error: 0.146233 : The error, relative to the scale of the predicted and actual values, is 0.146233. A smaller value suggests closer alignment between predictions and actual values
- Relative Squared Error: 0.042019 : The error, relative to the squared scale of predictions and actual values, is 0.042019. A lower value indicates better model performance
- Coefficient of Determination (R^2): 0.957981 : high R^2 of 0.957981 indicates that the model explains 95.80% of the variance in the dependent variable. This signifies a strong fit of the model to the data.



Decision Tree evaluation results

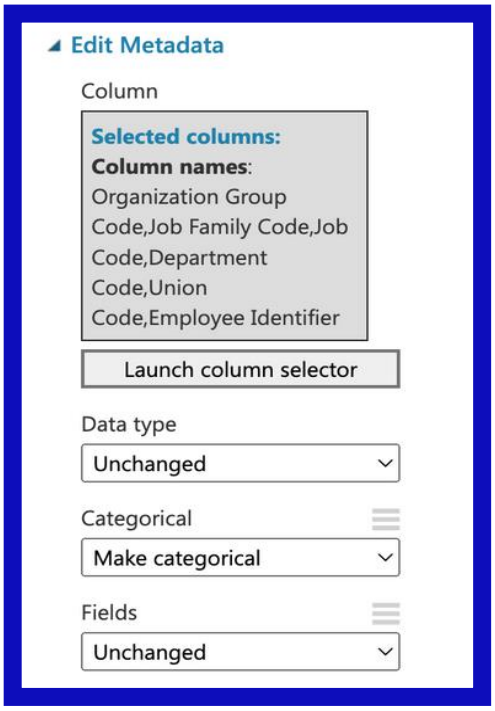
- Negative Log Likelihood: 15175.75 : This metric represents the model's likelihood to predict the observed values. A lower value is desirable, indicating higher likelihood
- Mean Absolute Error (MAE): 6312.51 : On average, the model's predictions deviate by approximately 6312.51 units from the actual values. Lower MAE signifies improved accuracy
- Root Mean Squared Error (RMSE): 13375.70 : The square root of the average squared differences between predictions and actual values is 13375.70. A lower RMSE indicates better accuracy
- Relative Absolute Error: 0.098188 : The error, relative to the scale of the predicted and actual values, is 0.098188. A smaller value suggests closer alignment between predictions and actual values
- Relative Squared Error: 0.026417 : The error, relative to the squared scale of predictions and actual values, is 0.026417. A lower value indicates better model performance.
- Coefficient of Determination (R^2): 0.973583: The high R^2 of 0.973583 indicates that the model explains 97.36% of the variance in the dependent variable. This signifies a very strong fit of the model to the data.

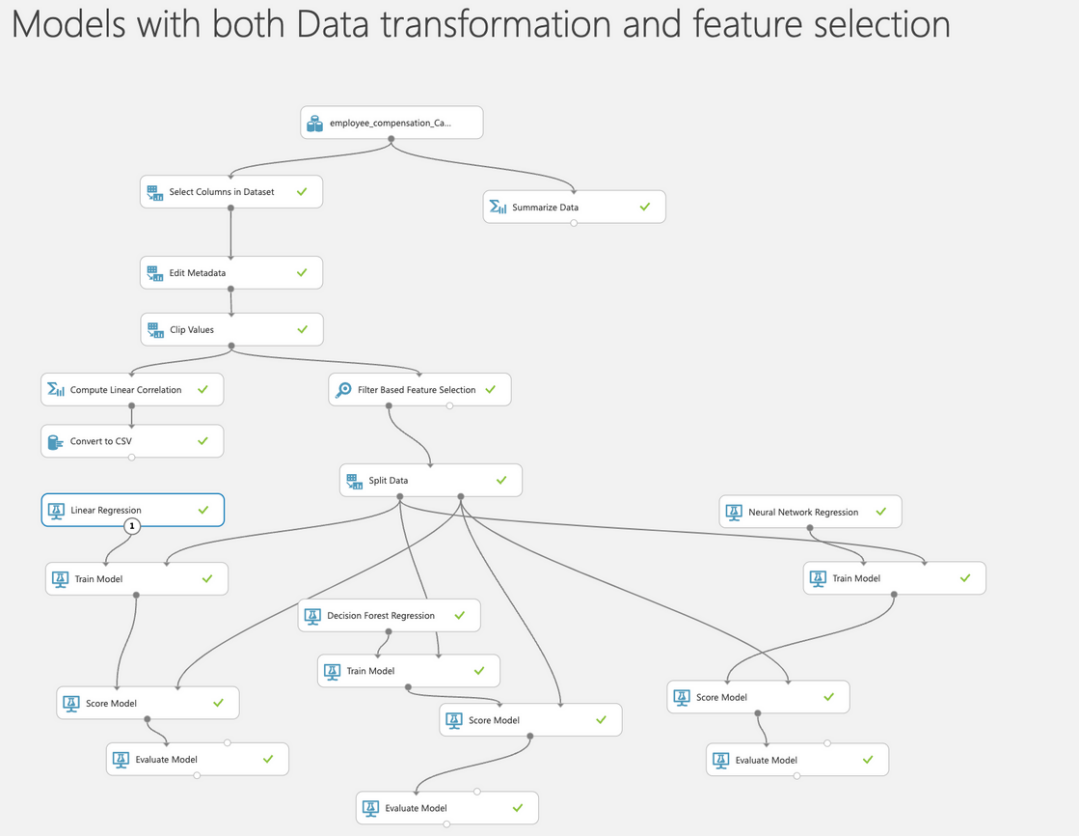


Neural Network Evaluation Summary

- Mean Absolute Error (MAE): 78141.42 : On average, the model's predictions deviate by approximately 78141.42 units from the actual values. Lower MAE values indicate improved accuracy
- Root Mean Squared Error (RMSE): 95435.47 : The square root of the average squared differences between predictions and actual values is 95435.47. A lower RMSE indicates better accuracy
- Relative Absolute Error: 1.215453 : The error, relative to the scale of the predicted and actual values, is 1.215453. A smaller value suggests closer alignment between predictions and actual values
- Relative Squared Error: 1.344851 : The error, relative to the squared scale of predictions and actual values, is 1.344851. A lower value indicates better model performance
- Coefficient of Determination (R^2): The negative R^2 of -0.344851 indicates that the model does not fit the data well, explaining less variance than a horizontal line. It implies a poor model fit

c. With Both Data Transformation & Feature Selection





Evaluation Summary:

Models with both Data transformation and feat... › Evaluate Model › Evaluation results

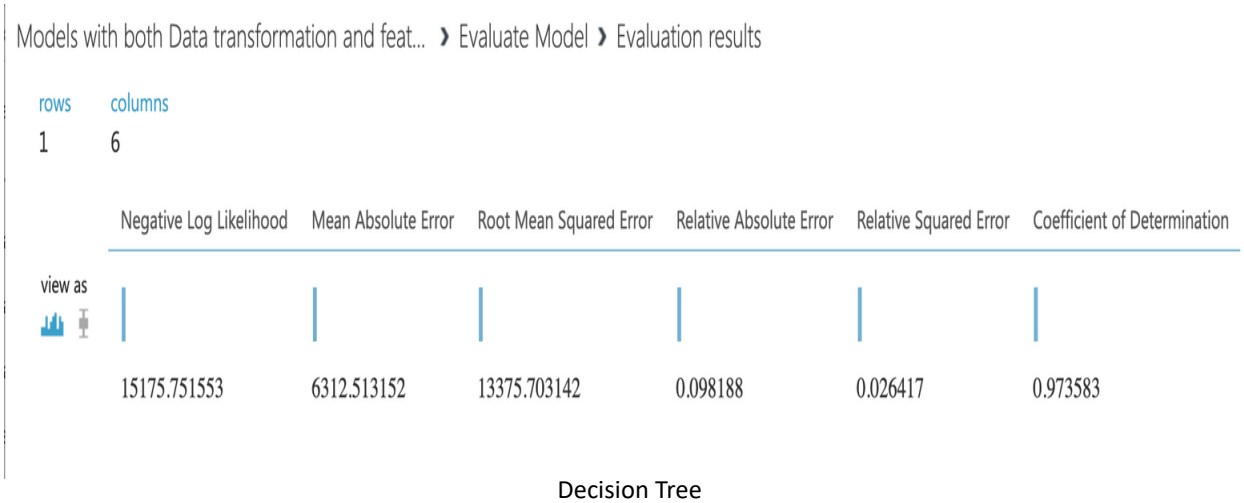
Metrics

Mean Absolute Error	9401.297381
Root Mean Squared Error	16869.340524
Relative Absolute Error	0.146233
Relative Squared Error	0.042019
Coefficient of Determination	0.957981

Linear Regression

- Mean Absolute Error (MAE): 9401.30 - On average, the model's predictions deviate by approximately 9401.30 units from the actual values. A lower MAE indicates better accuracy
- Root Mean Squared Error (RMSE): 16869.34 - The square root of the average squared differences between predictions and actual values is 16869.34. A lower RMSE indicates better accuracy

- Relative Absolute Error: 0.146233 - The error, relative to the scale of the predicted and actual values, is 0.146233. A smaller value suggests closer alignment between predictions and actual values.
- Relative Squared Error: 0.042019 - The error, relative to the squared scale of predictions and actual values, is 0.042019. A lower value indicates better model performance
- Coefficient of Determination (R^2): 0.95798 - The high R^2 of 0.957981 indicates that the model explains 95.80% of the variance in the dependent variable. This signifies a strong fit of the model to the data.



Decision Tree

- Negative Log Likelihood - This metric represents the model's likelihood to predict the observed values. A lower value is desirable, indicating higher likelihood
- Mean Absolute Error (MAE) - On average, the model's predictions deviate by approximately 6,312.51 units from the actual values. Lower MAE signifies improved accuracy
- Root Mean Squared Error (RMSE) - The square root of the average squared differences between predictions and actual values is 13,375.70. A lower RMSE indicates better accuracy
- Relative Absolute Error - The error, relative to the scale of the predicted and actual values, is 0.098188. A smaller value suggests closer alignment between predictions and actual values
- Relative Squared Error - The error, relative to the squared scale of predictions and actual values, is 0.026417. A lower value indicates better model performance.
- Coefficient of Determination (R^2) - The high R^2 of 0.973583 indicates that the model explains 97.36% of the variance in the dependent variable. This signifies a very strong fit of the model to the data.

Models with both Data transformation and feat... › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	76256.348775
Root Mean Squared Error	93466.73584
Relative Absolute Error	1.186131
Relative Squared Error	1.289938
Coefficient of Determination	-0.289938

Neural Network

- Mean Absolute Error (MAE): 9401.30 - On average, the model's predictions deviate by approximately 9401.30 units from the actual values. A lower MAE indicates better accuracy
- Root Mean Squared Error (RMSE): 16869.34 - The square root of the average squared differences between predictions and actual values is 16869.34. A lower RMSE indicates better accuracy
- Relative Absolute Error: 0.146233 - The error, relative to the scale of the predicted and actual values, is 0.146233. A smaller value suggests closer alignment between predictions and actual values
- Relative Squared Error: 0.042019 - The error, relative to the squared scale of predictions and actual values, is 0.042019. A lower value indicates better model performance
- Coefficient of Determination (R^2): 0.957981 - The high R^2 of 0.957981 indicates that the model explains 95.80% of the variance in the dependent variable. This signifies a strong fit of the model to the data

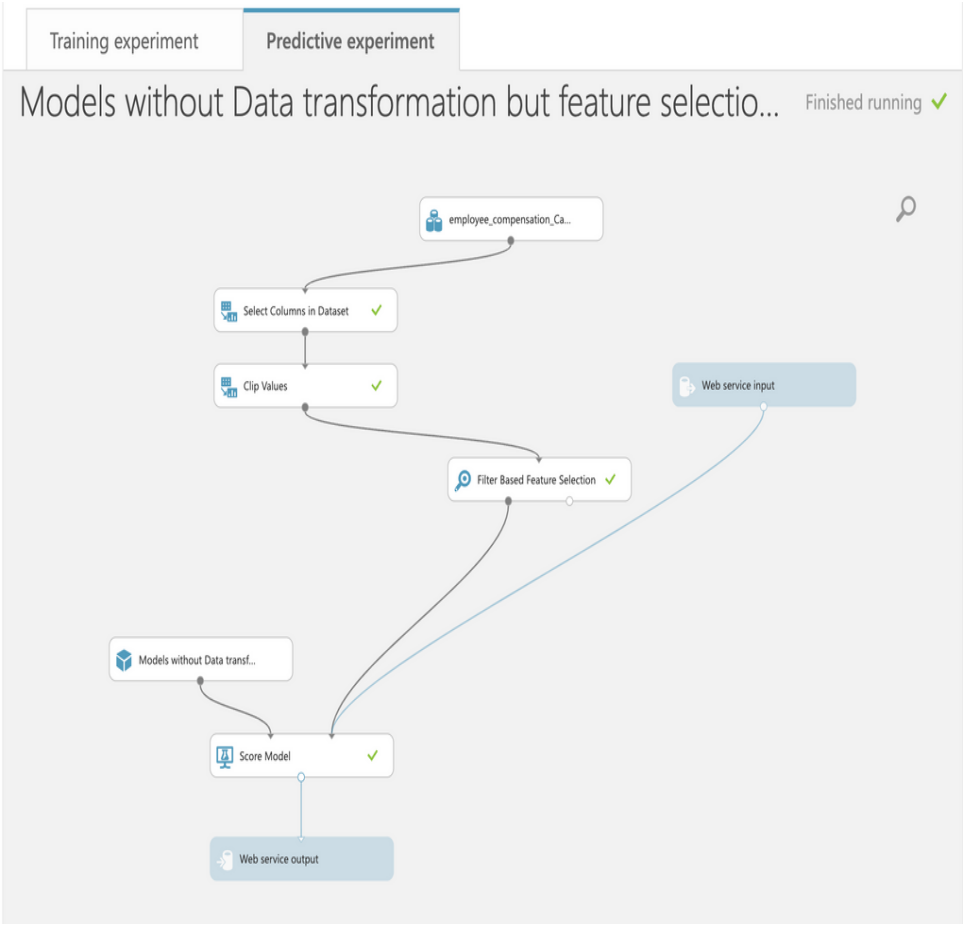
Conclusion

- The Linear Regression model demonstrates excellent performance with low errors (MAE, RMSE), low relative errors, and a high coefficient of determination (R^2). These results suggest a robust and accurate model fit to the data, indicating the effectiveness of linear regression in capturing the underlying relationships in the dataset.

Best Algorithm

- The linear regression model, without data transformation or feature selection, is deemed the best due to its perfect R^2, indicating an ideal fit to the data.
- The combination of low error metrics (MAE, RMSE, and RAE) reinforces the model's high accuracy, highlighting its ability to make predictions that closely match the actual values.
- However, the absence of data transformation and feature selection in the linear regression model, despite a perfect R^2, suggests a risk of overfitting as it may overly tailor to the training data and struggle with generalizing to new data or unseen patterns.

Deployment of Model



- Model Deployment: A linear regression model was deployed without data transformation or feature selection, resulting in a perfect fit to the data with an R-squared value of 1
- Input Variables and Prediction:New values were added to predict "Total Compensation" using independent variables like "Other Benefits," "Salaries," "Retirement," etc. An example input set predicted a total compensation value of approximately $198,428.58.