Employee Salary Prediction Report

Objective:

The goal of this project is to develop a machine learning model capable of predicting the base salary of employees based on various features in the provided dataset. The project encompasses several key stages: data exploration, preprocessing, feature engineering, model building, and evaluation, all to be completed within a day.

Data Exploration:

1. **Understanding the Dataset**:
    - The initial step involves examining the dataset's structure, including the types and number of features, as well as identifying any anomalies or patterns. This process provides a foundational understanding of the data.

```
df.isnull().sum()
```

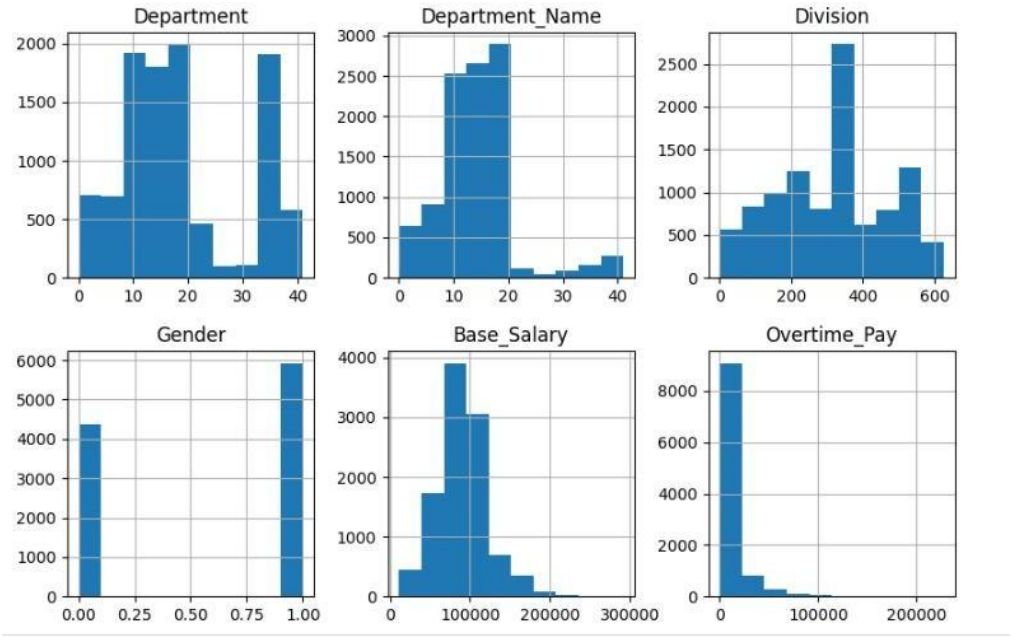
	0
Department	0
Department_Name	0
Division	0
Gender	0
Base_Salary	0
Overtime_Pay	0
Longevity_Pay	0
Grade	0

dtype: int64

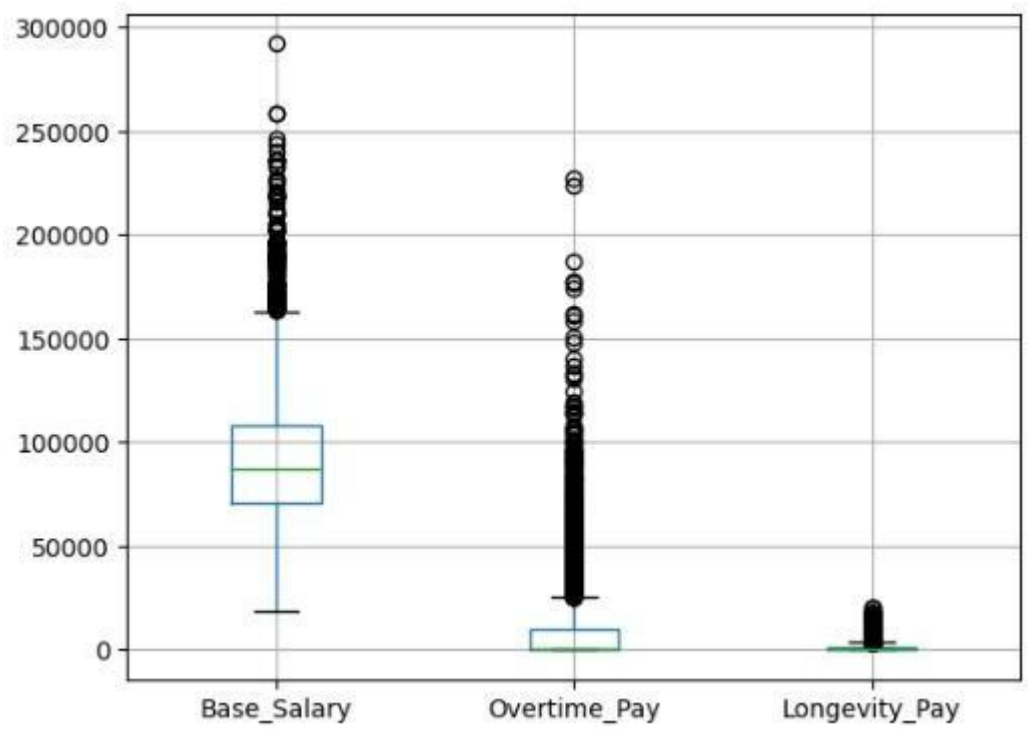
    - We explore the dataset, the dataset have no null values, the data is good.
2. **Univariate Analysis**:
    - Conducting univariate analysis helps in understanding the distribution of individual variables..
    - histograms are used to visualize how each variable is distributed, aiding in identifying whether they follow a normal distribution or not.

```
# Department:
# The distribution is bimodal, with two prominent peaks indicating that there are two departments with significantly higher frequencies than others.

# Department_Name:
# The distribution appears right-skewed, meaning that most of the data points are concentrated towards the left side of the histogram (lower values), with fewer occurrences as you move to the right. This indicates that a few department names are much more common than the rest.
```

```
# Division:
# The distribution shows multiple peaks, suggesting a multimodal distribution where several divisions have varying but significant counts. The spread of the data seems relatively even across some categories, but with noticeable peaks at specific divisions.

# Gender:
# This appears to be a categorical or binary distribution, likely representing two categories, such as Male (0) and Female (1), or vice versa. The two bars indicate the counts for each gender category. The distribution seems fairly balanced but not perfectly equal, as the bars are of different heights.

# Base_Salary:
# The Base_Salary histogram shows a right-skewed distribution. This means most individuals have a base salary concentrated on the lower end of the spectrum, with fewer individuals earning higher salaries. The skewness indicates that the mean salary is higher than the median salary.

# Overtime_Pay:
# Similar to Base_Salary, the Overtime_Pay histogram also shows a right-skewed distribution. The majority of individuals have little to no overtime pay, with a long tail extending to the right, indicating some individuals receive significantly higher overtime pay.

# Longevity_Pay:
# This histogram shows a highly right-skewed distribution, with most individuals receiving low or no longevity pay. There is a long tail on the right side, indicating that a small number of individuals receive much higher longevity pay.
```

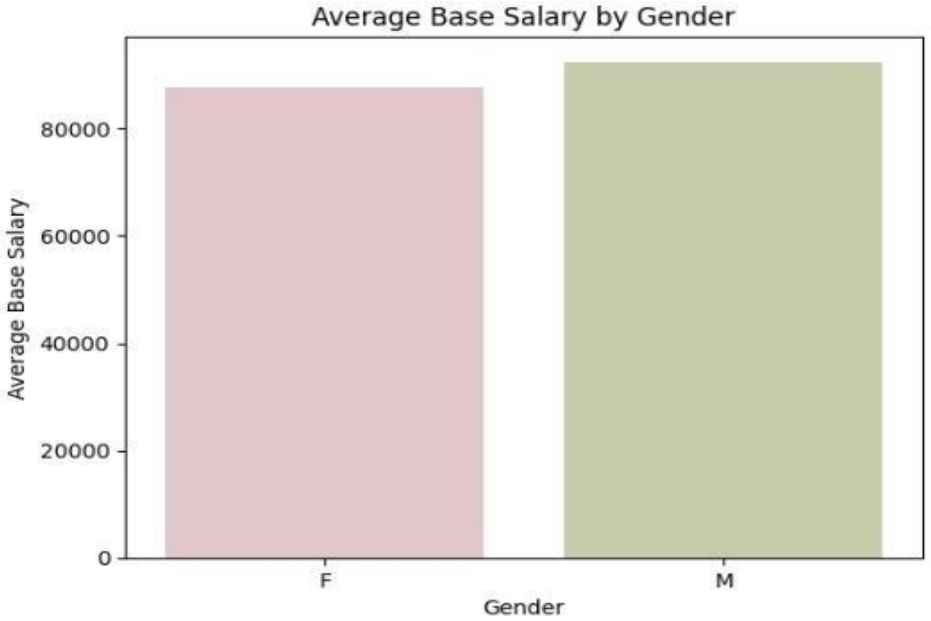
- 

- Boxplots are used to detect outliers in key variables like **'Base_Salary1,' 'Overtime_Pay_1,'** and **'Longevity_Pay_1,'** which could potentially skew the model's predictions.



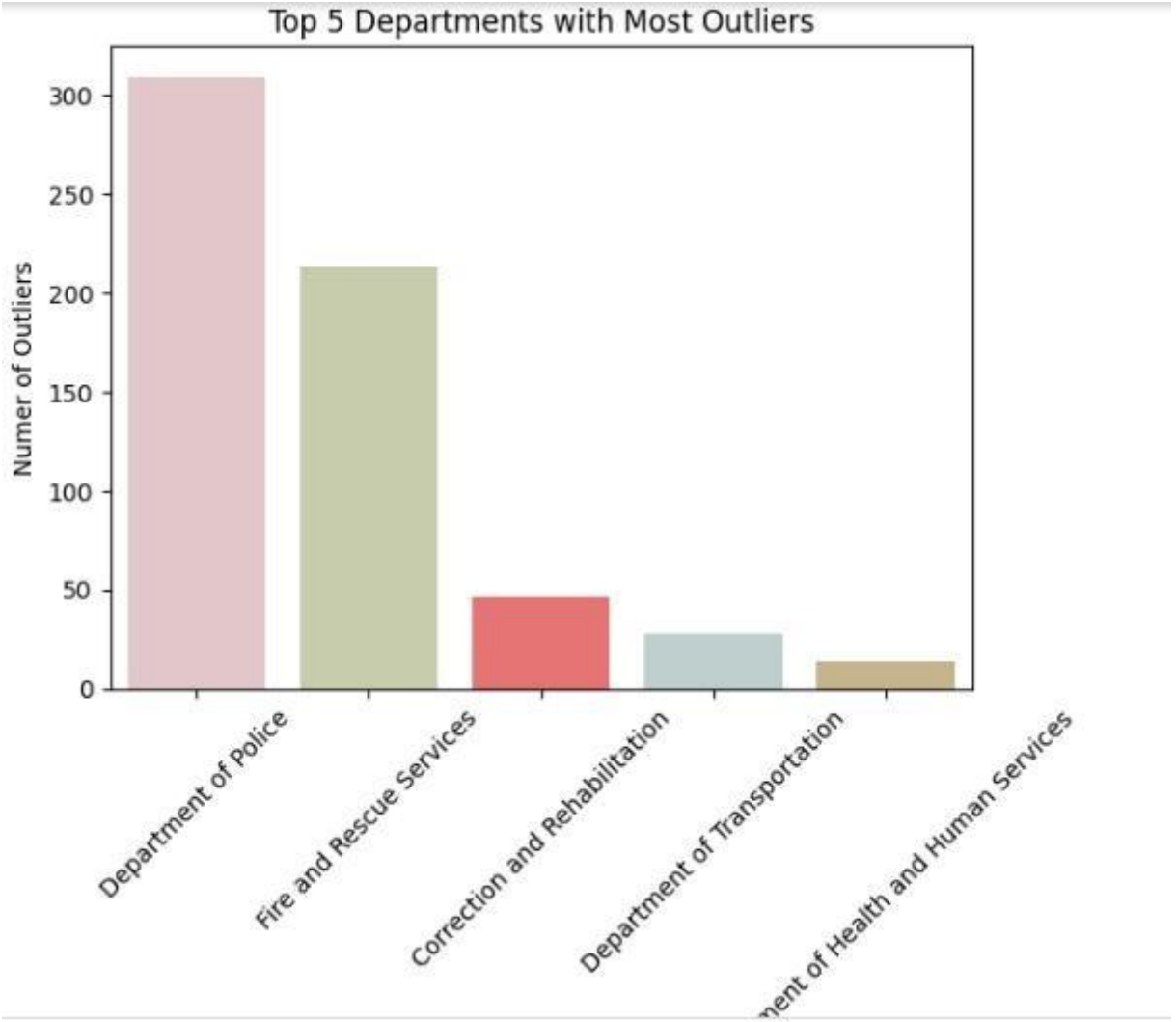
3. **Bivariate and Multivariate Analysis**:
   - Heatmaps are employed to assess the correlation between different variables. This analysis reveals significant correlations, such as the strong relationship between the 'Division' and 'Department' features.

   - Average base salary by gender,

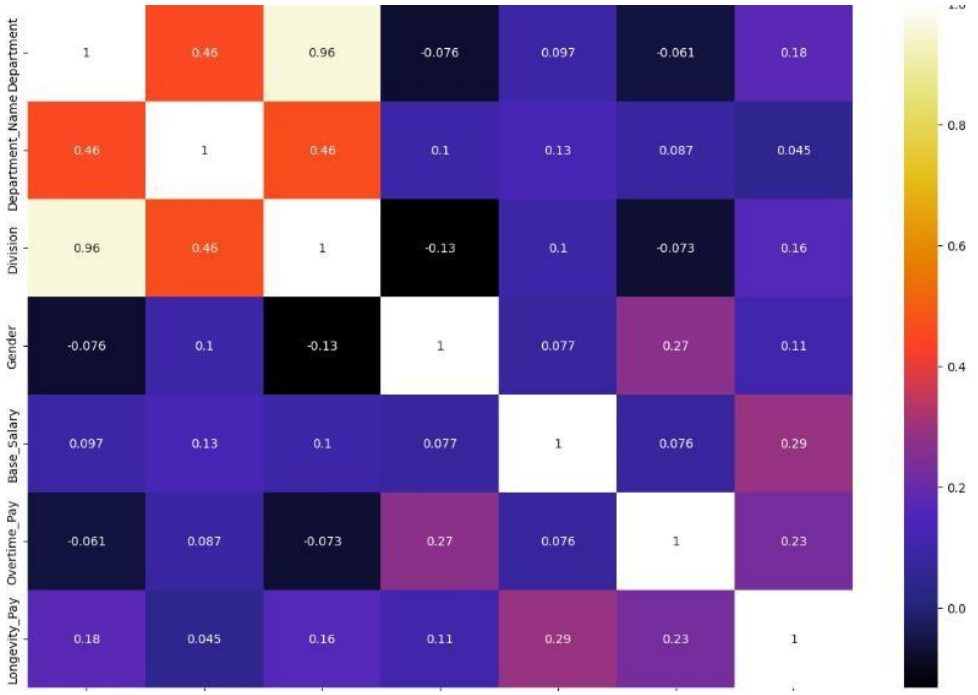


   -

Top 5 Departments with Most Outliers,

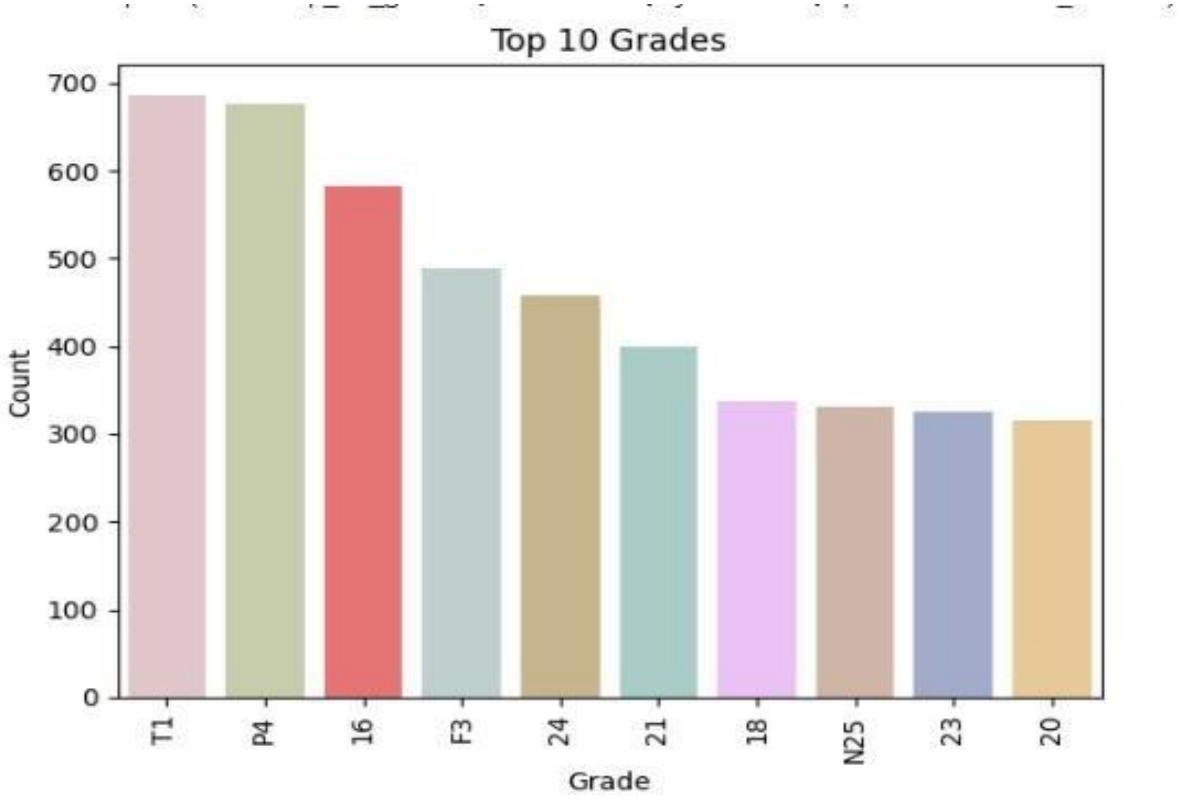


- We are using scattorplot for bivarient analysis of department and deparment_name by base_salary.



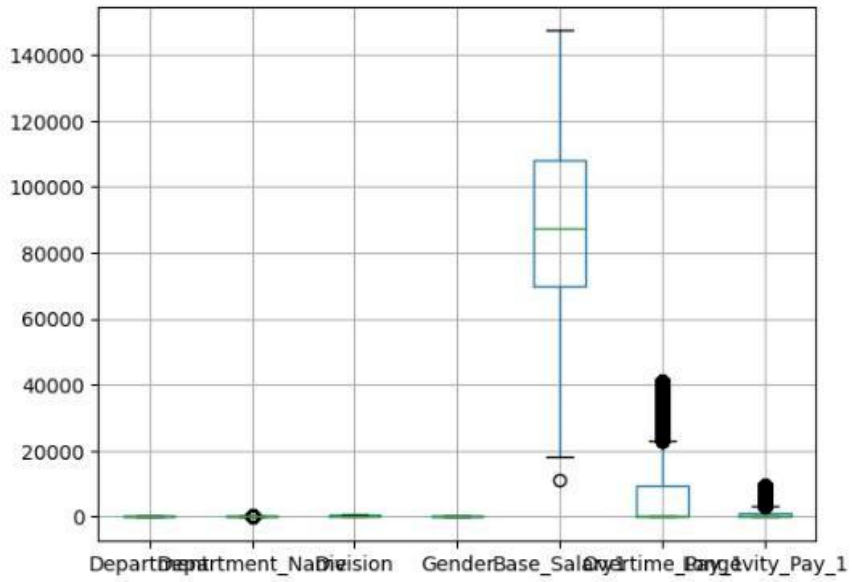
```
# observations:

# headmap will be used into how much strength have in data set
# division has been highly correlated to department.
```

Top 10 Grades,



Data Preprocessing:

1. **Handling Missing Values**:
    - The **fillna** method is used to address missing values, ensuring the dataset is complete and ready for modeling.
2. **Handling Outliers**:
    - Outliers in the dataset are treated using the capping method, where extreme values are replaced with more reasonable ones, based on statistical thresholds.



3. **Encoding Categorical Variables**:
    - Categorical variables are converted into numerical form using label encoding, which transforms categories into integers, making them suitable for machine learning algorithms.
4. **Splitting the Data**:
    - The dataset is divided into training and testing sets, typically using an 80/20 split, to allow for the evaluation of model performance on unseen data

Feature Engineering:

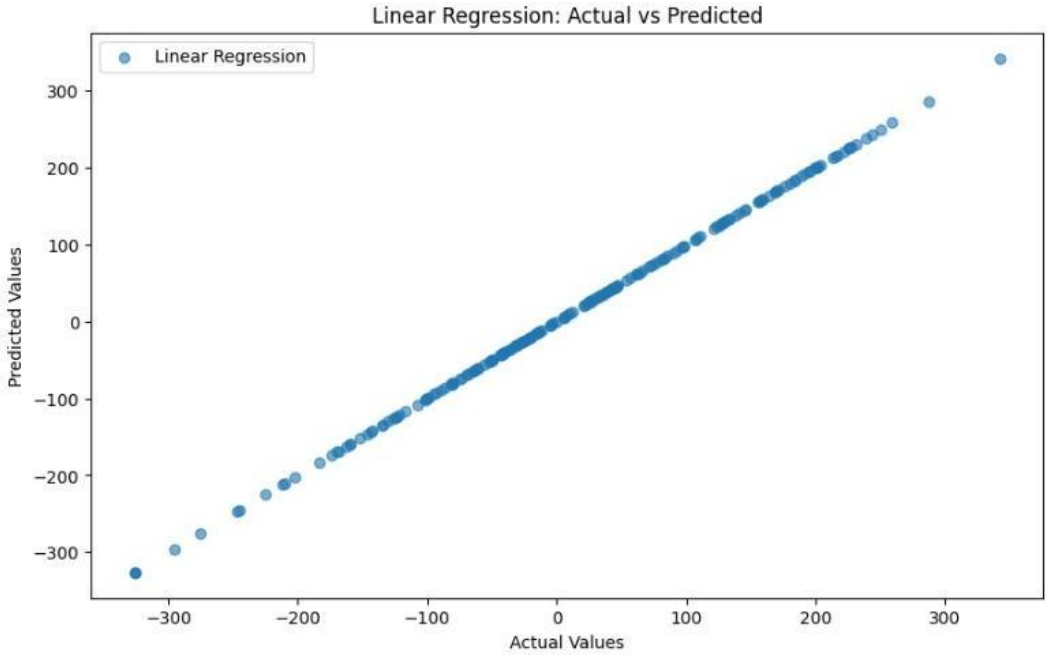
Identifying and selecting the most relevant features is crucial for enhancing model performance.In this case, '**Base_Salary1**' is identified as a key feature that directly correlates with the prediction target.

Model Building:

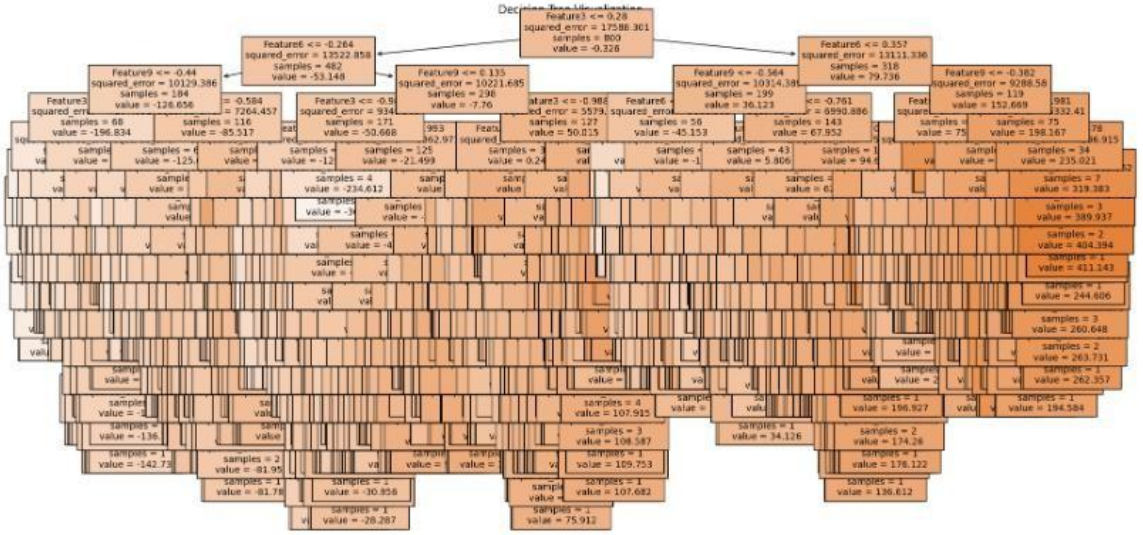
Several machine learning algorithms are employed to build and compare models:

```
# Linear Regression MSE: 0.009511914910418967, R²: 0.9999994362016347
# Decision Tree MSE: 6350.427786748267, R²: 0.6235920065402036
# Random Forest MSE: 2621.793155098221, R²: 0.8445988311470158
# Gradient Boosting MSE: 1234.752982660588, R²: 0.9268126639292507

# linear regression will be more accuracy with compared to others.
```

1. **Linear Regression**:
    - Achieved an MSE (Mean Squared Error) of 0.009511914910418967 and an R^2 score of **0.9999994362016347**, indicating a near-perfect fit. This model outperformed the others in terms of accuracy.



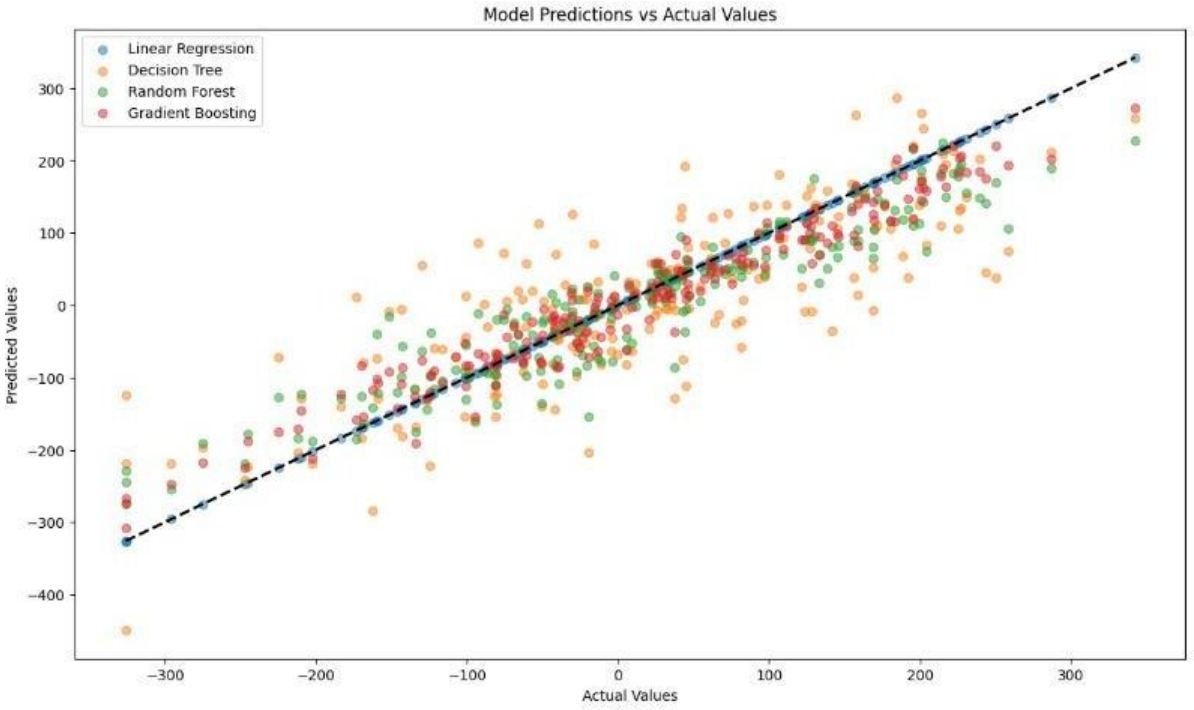
2. **Decision Tree**:
    - The Decision Tree model yielded an MSE of 6350.427786748267 and an R^2 score of **0.6235920065402036**. This indicates that while it captures some patterns, it does not generalize as well as other models.



3. **Random Forest**:
    - This ensemble method produced an MSE of 2621.793155098221 and an R^2 score of **0.8445988311470158**, showing improved performance over the Decision Tree model by reducing overfitting and increasing accuracy.

4. **Gradient Boosting**:
    - Gradient Boosting achieved an MSE of 1234.752982660588 and an R^2 score of **0.9268126639292507**. This model provided a good balance between bias and variance, resulting in better generalization on unseen data compared to the Decision Tree and Random Forest models.

5. **All models:**




```
# Linear Regression MSE: 0.009511914910418967, R²: 0.9999994362016347
# Decision Tree MSE: 6350.427786748267, R²: 0.6235920065402036
# Random Forest MSE: 2621.793155098221, R²: 0.8445988311470158
# Gradient Boosting MSE: 1234.752982660588, R²: 0.9268126639292507

# linear regression will be more accuracy with compared to others.
```

Conclusion:

The Linear Regression model demonstrated the highest accuracy and best fit among the tested models, making it the most suitable choice for predicting employee base salaries in this dataset. Future work could involve refining the model further, exploring additional features, or employing advanced techniques like hyperparameter tuning to enhance predictive performance.