Final Report

1. Problem Statement

The objective of this analysis is to develop predictive models to estimate base salary for employees based on various features, such as 'Department Name’, 'Grade,' 'Overtime_Pay,' 'Longevity_Pay,', 'Gender,' 'Department,' and 'Division.' The goal is to identify the most accurate model that can be used for forecasting base salary.

2. Objective

Analyze the data to gain insights and compare the performance of different regression models in predicting base salary. The goal is to optimize these models to enhance predictive accuracy and implement data preprocessing techniques to improve model quality and interpretability.

3. Dataset

Annual salary information including gross pay and overtime pay for all active, permanent employees of Montgomery County, MD paid in calendar year 2023. This dataset is a prime candidate for conducting analyses on salary disparities, the relationship between department/division and salary, and the distribution of salaries across gender and grade levels.

The dataset includes the following features:

Numerical Features: 'Overtime_Pay,' 'Longevity_Pay,' 'Base_Salary.'

Categorical Features: 'Gender,' 'Department,' 'Division,' 'Grade.'

Target Variable: 'Base_Salary.'

Source: Employee Salaries Analysis (kaggle.com)

4. Methodology

Data Preprocessing:

Log Transformation: Applied to 'Overtime_Pay' and 'Longevity_Pay' to reduce skewness and stabilize variance

Label Encoding: Categorical columns such as 'Gender,' 'Department,' and 'Division' were encoded for model compatibility.

Ordinal Encoding: The 'Grade' column was encoded using a custom order based on the average base salary, ensuring that the ordinal nature of grades is preserved.

Modeling:

Developed and compared the following regression models: Linear Regression, Decision Tree, XGBoost, and Random Forest.

Conducted hyperparameter tuning using Grid Search CV for Decision Tree and Random Forest models to improve their performance.

Evaluated the models using performance metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared.

Evaluation Metrics:

MSE: Measures the average squared difference between predicted and actual base salaries.

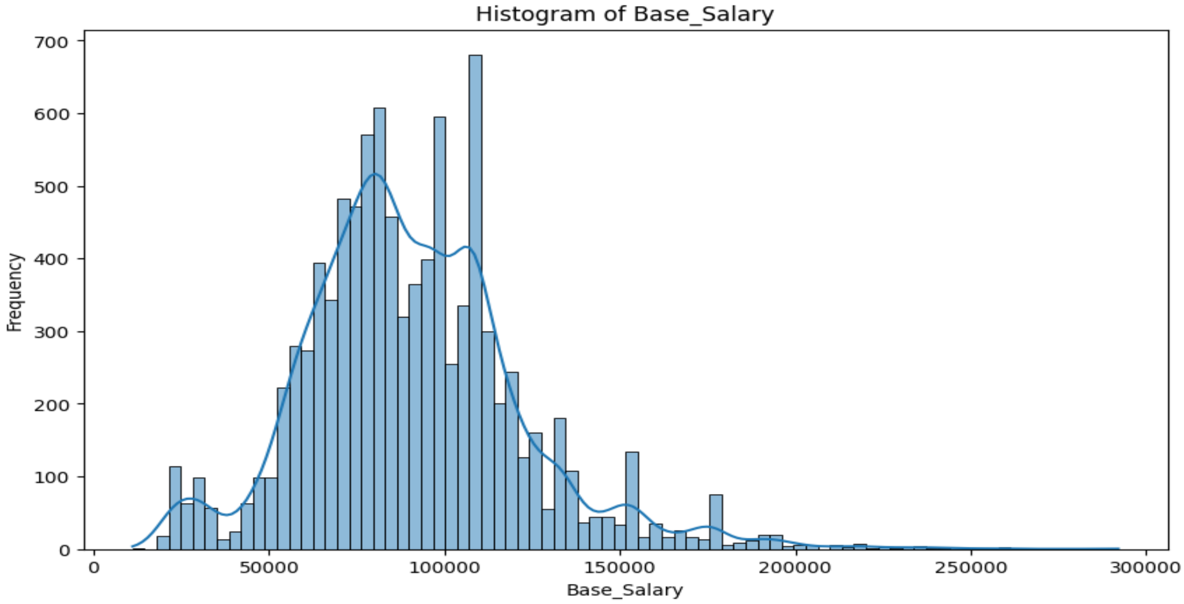
RMSE: The square root of MSE, indicating the average error magnitude in the predictions.

MAE: The average absolute difference between predicted and actual base salaries.

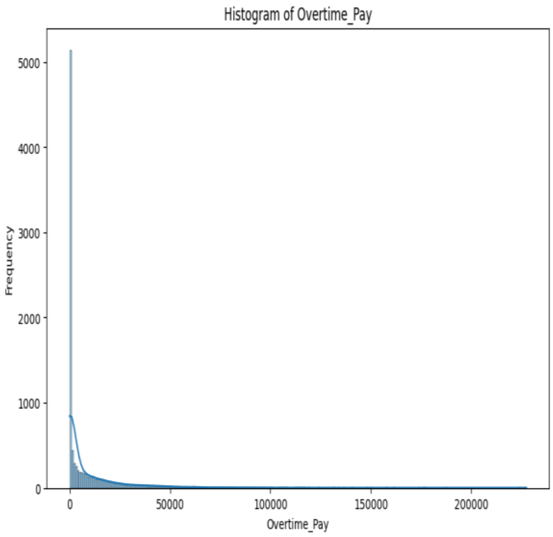
R-squared: Represents the proportion of variance in the base salary explained by the independent variables.

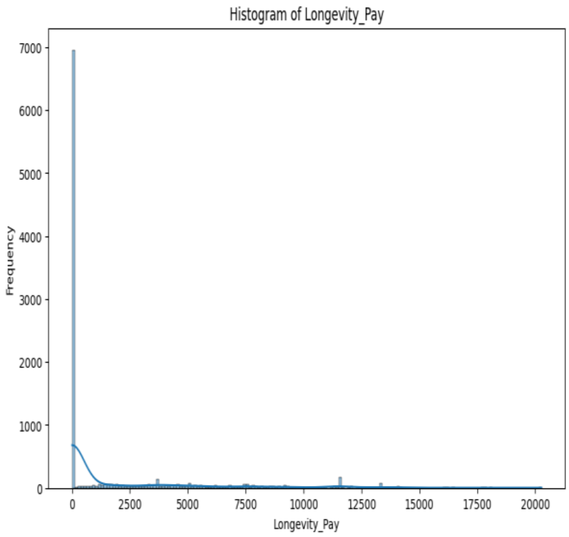
Adjusted R-squared: a statistical measure that indicates how well a regression model explains the variability of the dependent variable, accounting for the number of independent variables in the model.

5. Analysis and Insights:

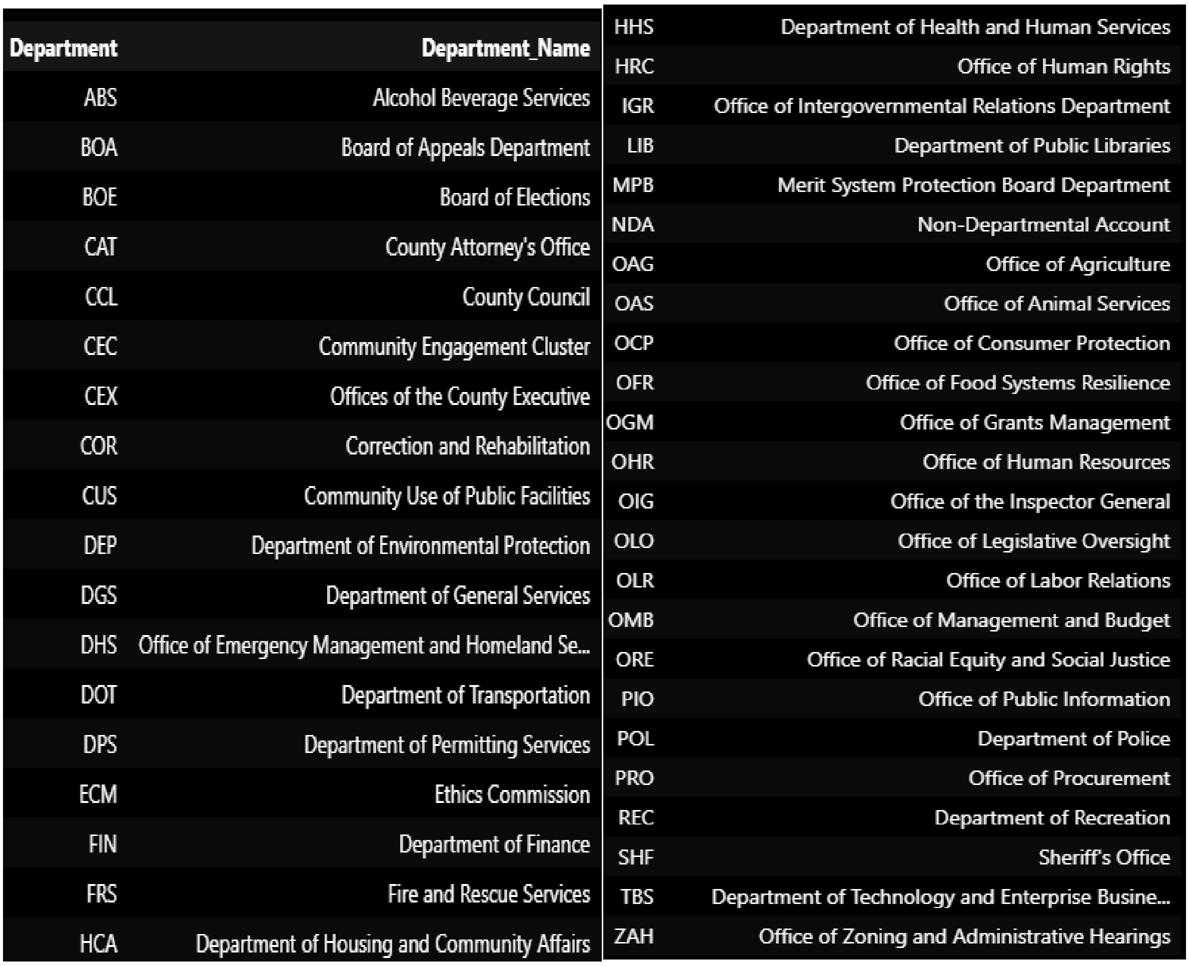


From this, we can see that the distribution follows a normal distribution, with a higher number of employees receiving a medium base salary.



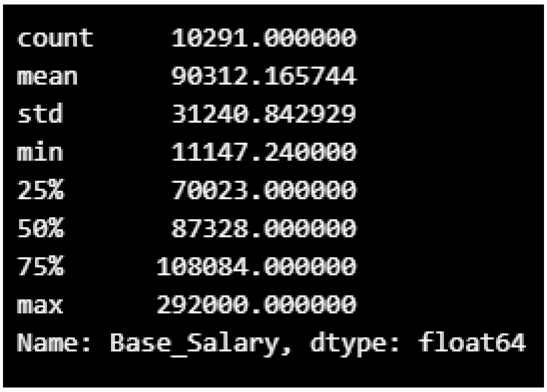


These are the distributions of overtime pay and longevity pay. As seen from the charts, both distributions are positively skewed. This skewness can be addressed using log transformation. Additionally, the charts indicate that the number of employees receiving overtime and longevity pay is relatively low. It can be inferred that, particularly for longevity pay, the number of employees staying in the same job for an extended period is limited.



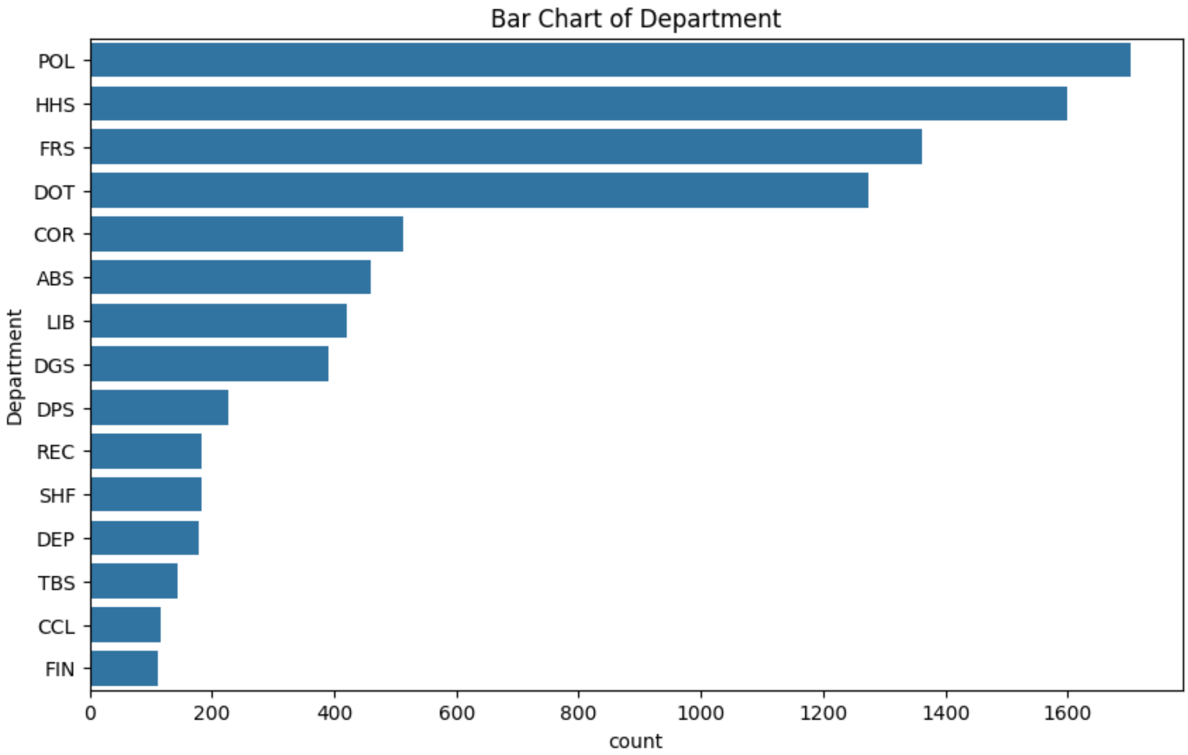
Department	Department_Name
ABS	Alcohol Beverage Services
BOA	Board of Appeals Department
BOE	Board of Elections
CAT	County Attorney's Office
CCL	County Council
CEC	Community Engagement Cluster
CEX	Offices of the County Executive
COR	Correction and Rehabilitation
CUS	Community Use of Public Facilities
DEP	Department of Environmental Protection
DGS	Department of General Services
DHS	Office of Emergency Management and Homeland Se...
DOT	Department of Transportation
DPS	Department of Permitting Services
ECM	Ethics Commission
FIN	Department of Finance
FRS	Fire and Rescue Services
HCA	Department of Housing and Community Affairs
HHS	Department of Health and Human Services
HRC	Office of Human Rights
IGR	Office of Intergovernmental Relations Department
LIB	Department of Public Libraries
MPB	Merit System Protection Board Department
NDA	Non-Departmental Account
OAG	Office of Agriculture
OAS	Office of Animal Services
OCP	Office of Consumer Protection
OFR	Office of Food Systems Resilience
OGM	Office of Grants Management
OHR	Office of Human Resources
OIG	Office of the Inspector General
OLO	Office of Legislative Oversight
OLR	Office of Labor Relations
OMB	Office of Management and Budget
ORE	Office of Racial Equity and Social Justice
PIO	Office of Public Information
POL	Department of Police
PRO	Office of Procurement
REC	Department of Recreation
SHF	Sheriff's Office
TBS	Department of Technology and Enterprise Busine...
ZAH	Office of Zoning and Administrative Hearings

These are the department and their respective department names. Since both "department" and "department name" refer to the same thing, I removed the "department" column during preprocessing.



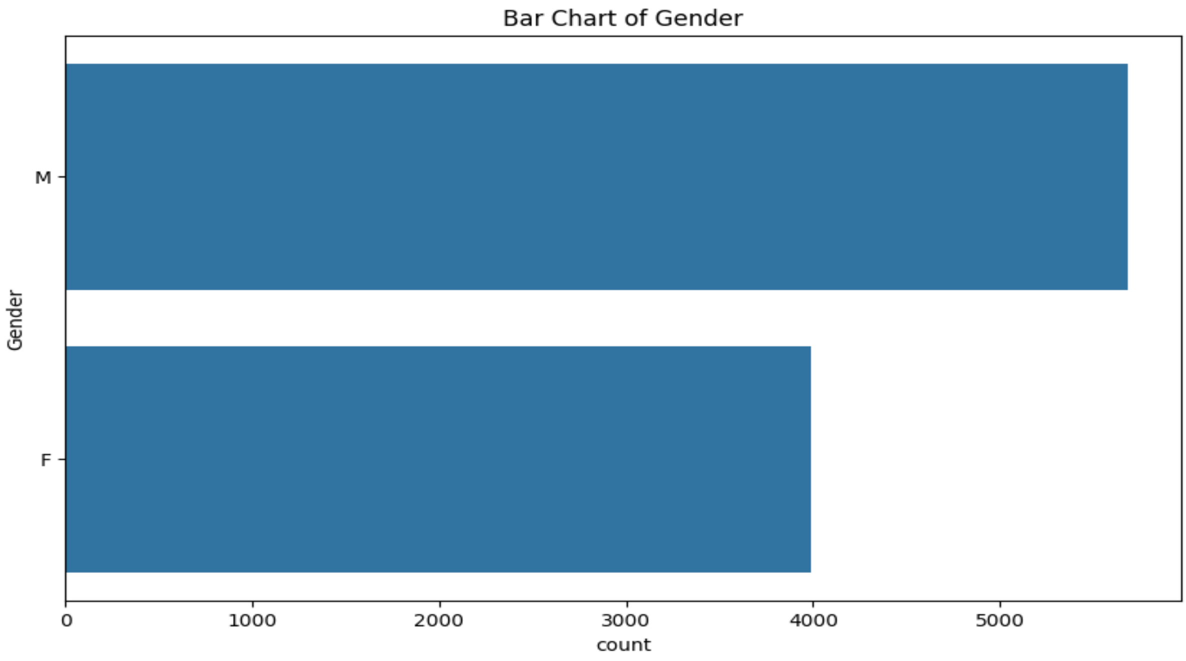
```
count      10291.000000
mean       90312.165744
std        31240.842929
min        11147.240000
25%        70023.000000
50%        87328.000000
75%       108084.000000
max       292000.000000
Name: Base_Salary, dtype: float64
```

From the above, we can see that the average base salary is $90,312, with a salary range between $11,147 and $292,000 across various departments. This indicates that there is a wide variation in base salaries, which could be attributed to differences in roles, responsibilities, and seniority levels within the departments.

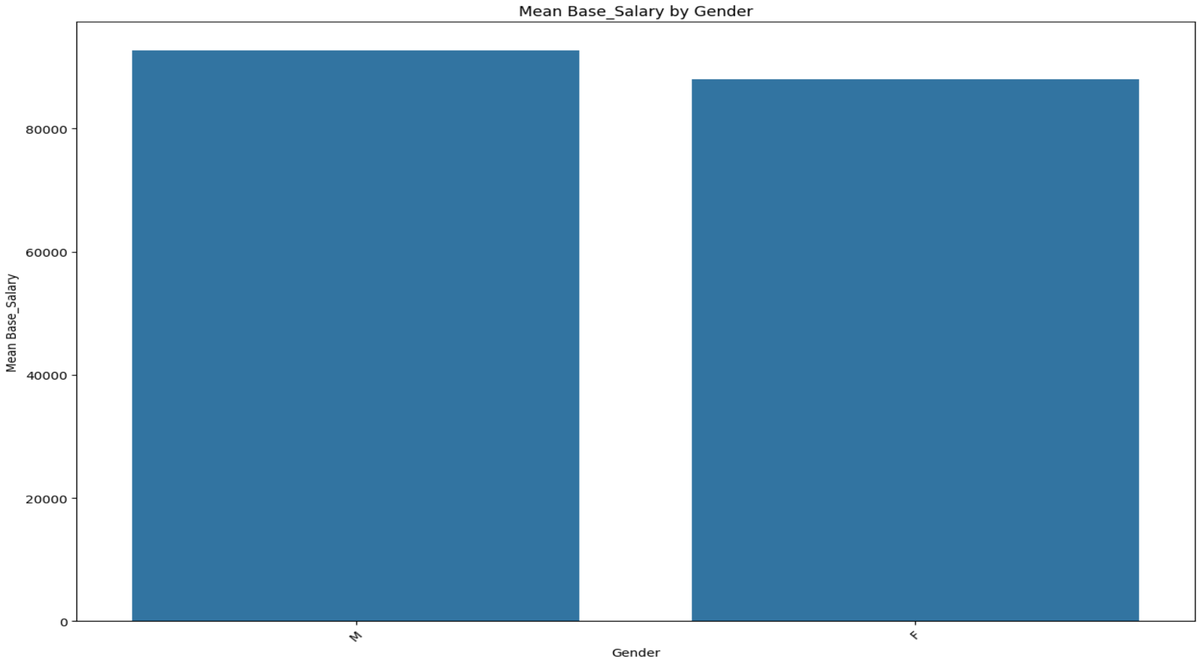


[POL refers to the Department of Police, HHS to the Department of Health and Human Services, FRS to Fire and Rescue Services, and DOT to the Department of Transportation.]

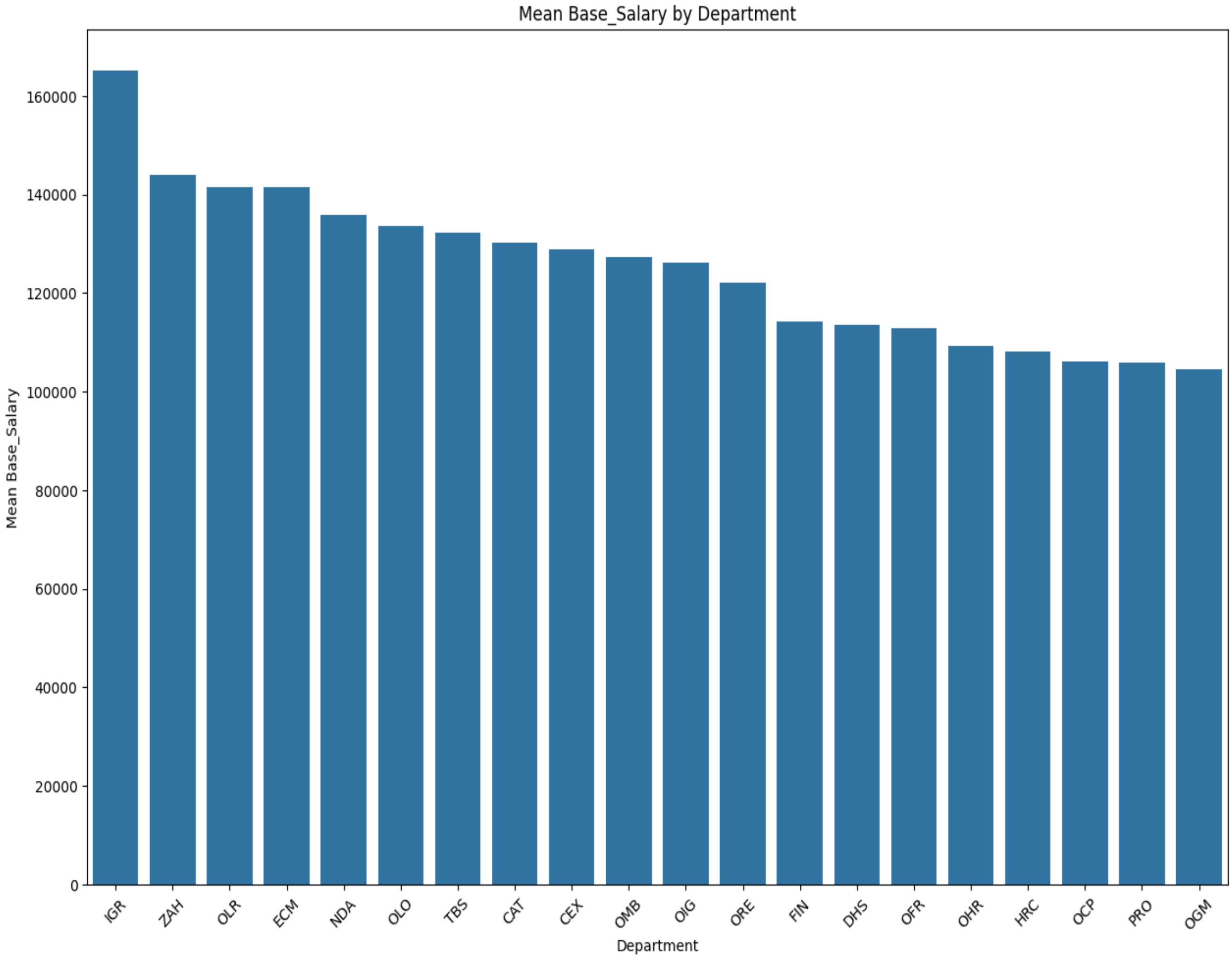
We can see from the charts that the number of employees in the Police Department (POL) is the highest. This suggests that Montgomery County places a high priority on public safety and security. These departments play critical roles in Montgomery County. The Department of Health and Human Services (HHS) focuses on public health and social services, ensuring the well-being of the community. Fire and Rescue Services (FRS) provide emergency response and fire safety, crucial for protecting lives and property. The Department of Transportation (DOT) manages transportation infrastructure and services, facilitating mobility and connectivity within the county. Their prominence reflects the county's commitment to public health, safety, and efficient transportation.



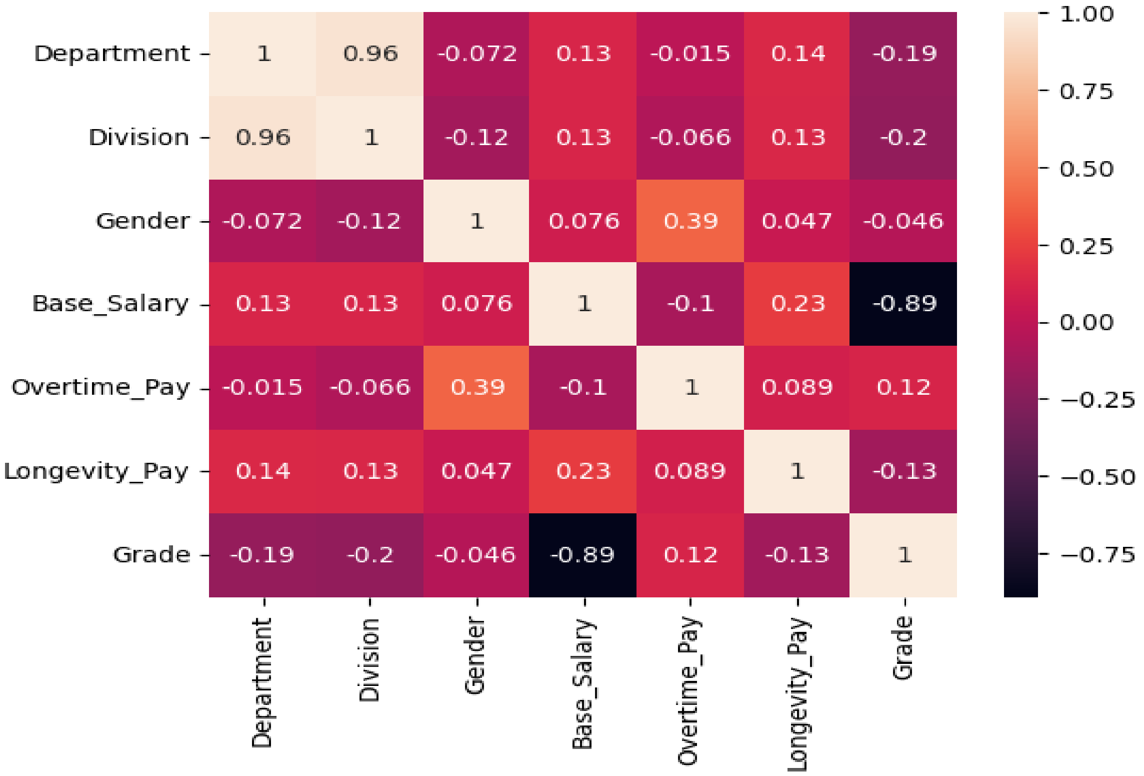
From the above chart, we can see that the number of male employees is slightly greater than the number of female employees. This means there is a small gender imbalance in the workforce, with a marginally higher number of male employees compared to female employees.



From the above chart, we can assume that there is not much difference in the base salary between male and female employees, as both groups are receiving similar salaries. This suggests a level of gender equality in terms of base salary within the organization, indicating fair compensation practices across genders.



The Office of Intergovernmental Relations Department has the highest average base salary, followed by the Office of Zoning and Administrative Hearings. The Office of Labor Relations and the Ethics Commission receive similar base salaries. This suggests that the Office of Intergovernmental Relations may have roles with higher responsibilities or specialized skills that warrant higher compensation. The relatively similar base salaries for the other offices indicate a more uniform salary structure among those departments, possibly reflecting similar job roles or levels of responsibility.



From the above, we can see that the division and department columns are highly correlated with each other. Therefore, we can keep either one of these columns for training the model because they provide similar information. With respect to base salary, which is our target variable, the grade is highly negatively correlated with the target. This means that as the grade increases, the base salary tends to decrease

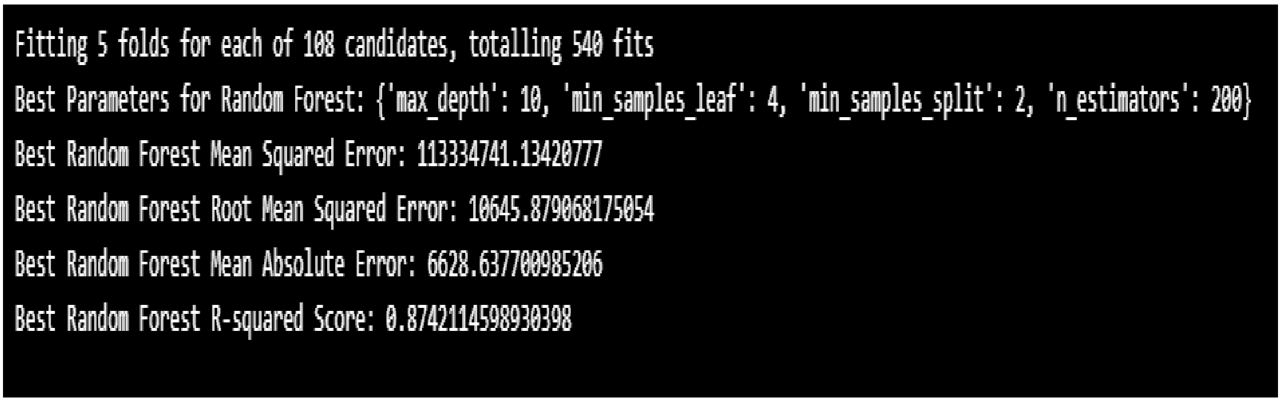
6. Inference

Linear Regression provided a baseline with moderate performance in predicting base salary.

Decision Tree models, particularly after hyperparameter tuning, demonstrated improved accuracy, highlighting the importance of model tuning.

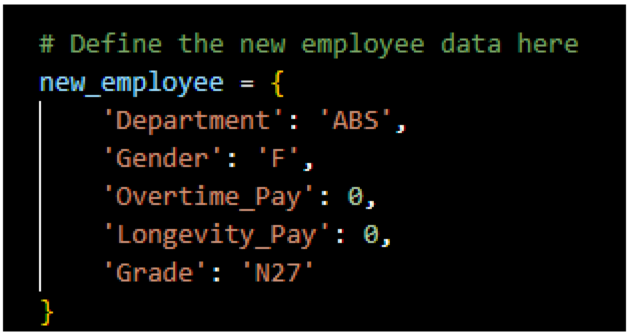
XGBoost emerged as a strong candidate, offering a good balance of accuracy and interpretability.

Random Forest (with Grid Search CV) performed the best, achieving the lowest error metrics and highest R-squared value, indicating a strong fit to the data. Below is the mse, mae, rmse, r2 score got from the random forest model:

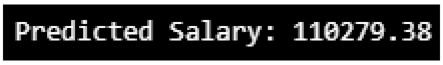


```
Fitting 5 folds for each of 108 candidates, totalling 540 fits
Best Parameters for Random Forest: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200}
Best Random Forest Mean Squared Error: 113334741.13420777
Best Random Forest Root Mean Squared Error: 10645.879068175054
Best Random Forest Mean Absolute Error: 6628.637700985206
Best Random Forest R-squared Score: 0.8742114598930398
```

7. Prediction



```
# Define the new employee data here
new_employee = {
    'Department': 'ABS',
    'Gender': 'F',
    'Overtime_Pay': 0,
    'Longevity_Pay': 0,
    'Grade': 'N27'
}
```



```
Predicted Salary: 110279.38
```

8. Conclusion

The analysis concluded that the Random Forest model with Grid Search CV is the most accurate model for predicting base salary, followed closely by XGBoost. The thorough preprocessing steps, including log transformations and ordinal encoding, contributed significantly to the accuracy of the predictions.

These models are recommended for deployment in predicting base salary, subject to further validation and testing to ensure their reliability in practical applications. The findings underscore the necessity of proper data preprocessing and model tuning in developing accurate predictive models. Further refinement, including additional feature engineering and data collection, may enhance the models' predictive power and applicability.