**The report for ShAi task**

* First, I downloaded the data and displayed it.
* Make an initial look at the data to identify the values ​​in the data in general.
* Identify the columns we have and know the number of rows and columns in the data.
* **Know the types of data in each column.**

Id int64

Employee Name object

Job Title object

Base Pay float64

Overtime Pay float64

Other Pay float64

Benefits float64

Total Pay float64

TotalPayBenefits float64

Year int64

Notes float64

Agency object

Status float64

* **Finding the missing values in the data and knowing their number in each column.**  
  Id 0  
  EmployeeName 0  
  JobTitle 0  
  BasePay 609  
  overtimePay 4  
  OtherPay 4  
  Benefits 36163

TotalPay 368

TotalPayBenefits 0

Year 0

Notes 148654

Agency 0

Status 148654

* **Find the arithmetic mean, median, smallest value, and largest value in the Salary field.**

Mean salary: 74768.32197169267

Mean salary: 74768.32197169267

Median salary: 71426.60999999999

Mode salary: 0.0

Minimum salary: -618.13

Maximum salary: 567595.43

Salary range: 568213.56

Salary standard deviation: 50517.00527394987

* **Data Cleaning and Preprocessing:**

1-Handled missing values and zero values in 'TotalPay' column.

2-Imputed missing values in numerical columns using mean imputation.

3-Dropped unnecessary columns and irrelevant rows.

* **A model was built using linear regression:**  
  1-Built a linear regression model to predict 'TotalPay' based on relevant features.

2-Addressed the issue of zero values in 'TotalPay' by predicting them using the model.

3-utilized sklearn's Linear Regression for modeling.

**Why Linear Regression?**

Linear regression is used in this case because it assumes a linear relationship between the input features ('BasePay', 'OvertimePay', 'OtherPay', 'Benefits', 'Year') and the target variable ('TotalPay'). It's a straightforward model that works well when there is a linear correlation between the features and the target. However, it's essential to note that the choice of the model depends on the characteristics of the data, and other models like decision trees or ensemble methods could also be considered based on the specific context and performance requirements.

* **Basic Data Visualization:**

1-Created histograms and bar charts to visualize the distribution of salaries.

2-Employed pie charts to represent the proportion of employees in different departments.

* **Simple Correlation Analysis:**

1-Explored the correlation between 'TotalPay' and 'BasePay'.

2-Plotted a scatter plot to visualize the relationship.