Data analysis of salary data set

In this report, I started to understand the dataset and what it is. The dataset includes details such as employee names, job titles, base pay, overtime pay, benefits, total pay, and other relevant information.

My analysis aims to know ideas related to the distribution of salaries, average salaries over the years, and possible relationships between salaries, departmental ratios, and other variables.

The dataset contains 148,654 rows and 13 columns. I have explained the type of each column, as most columns have numeric data types (int64 or float64).

```
df.dtypes
```

```
Id                 int64
EmployeeName      object
JobTitle          object
BasePay          float64
OvertimePay      float64
OtherPay         float64
Benefits         float64
TotalPay         float64
TotalPayBenefits float64
Year               int64
Notes            float64
Agency            object
Status           float64
dtype: object
```

Of course, I found missing values in many columns, in "BasePay", "OvertimePay", "OtherPay" and "Benefits". In addition, the "Notes" and "Status" columns do not contain any value and are completely empty columns.

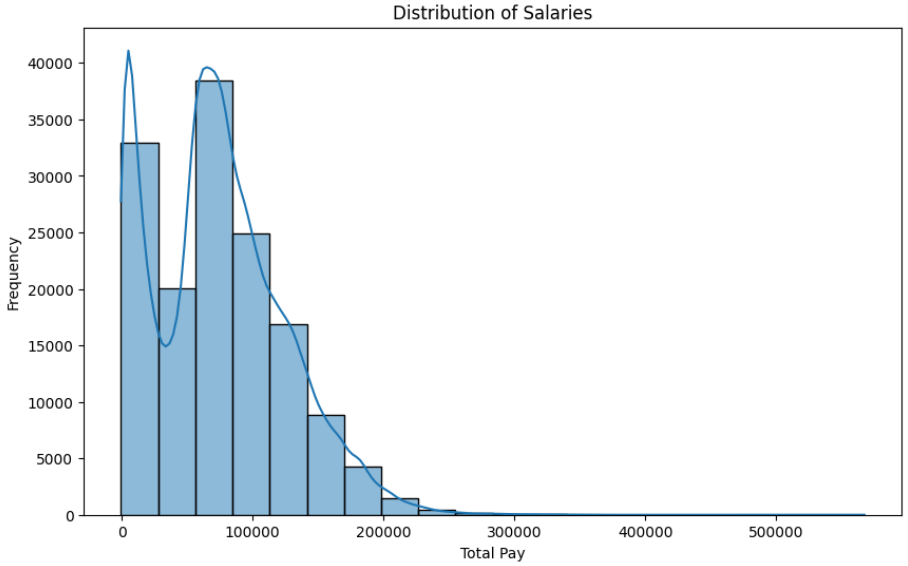
```
df.isna().sum()
```

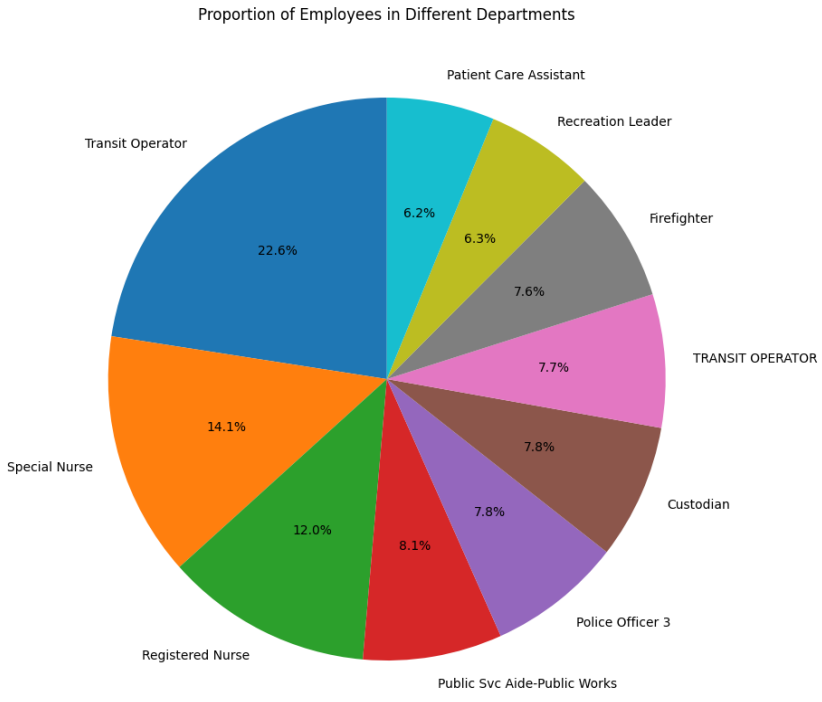
```
Id                     0
EmployeeName           0
JobTitle               0
BasePay              609
OvertimePay            4
OtherPay               4
Benefits           36163
TotalPay               0
TotalPayBenefits       0
Year                   0
Notes             148654
Agency                 0
Status            148654
dtype: int64
```

Then I calculated the mean, median, mode, minimum, and maximum salary to gain insights into salary distributions by (df.descrube()).

I dropped unnecessary columns that would not affect the quality of the dataset (“Status” and “Notes”) and fill the missing values by mean “Benefits” and “BasePay”.
In addition, some rows with missing data were removed from the dataset.

Histogram used to visualize the distribution of salaries, providing insight into the spread and shape of salary data. Pie charts were also used to represent the percentage of employees in different departments, highlighting the distribution of job titles within the organization and used matplotlib library.





Correlation analysis helps to understand the relationship between salary and other numerical columns such as BasePay, OtherPay, TotalPay, and TotalPayBenefits .
A correlation matrix was created and the correlation coefficient ranges from -1 to 1.
1 indicates a perfect positive linear relationship , -1 indicates a perfect negative linear relationship and 0 indicates no linear relationship between the variables.

In conclusion, this exploratory data analysis provided valuable insights into the salary data set, including salary distributions, departmental ratios, average salaries over the years, and potential relationships between salary and other variables. The results of this analysis will inform the decision-making process within the organization and avoid errors. Which may occur or reduce it.