|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Salaries DataSet | | | | |
|  |  |  |  |
| Show Data | |  | |

Steps :

1. **Data Exploration**
2. Descriptive statistics
3. Data cleaning
4. Basic Data Visualization
5. Grouped Analysis
6. Simple correlation Analysis

**Data Exploration**

1- It show that the columns : (Notes, Status) are all null so

there is no inforamtion we will get from this columns.

2- there is some columns contain null but it's not many.

3- DataType of numerical columns is float so that is right,

String columns as Object and it is right so our dataType of our data is good.

**Descriptive Statistics:**

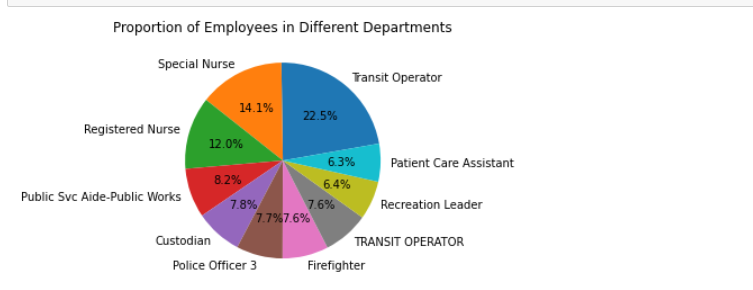
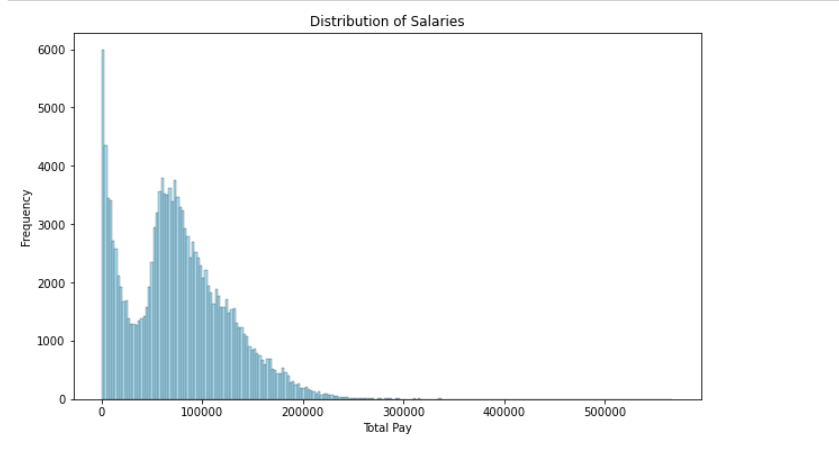
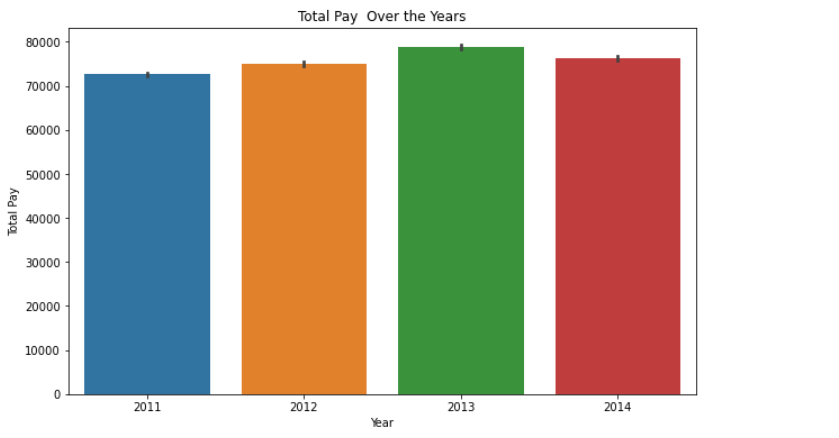
* Defined a function, **statistics**, to calculate and print mean, median, mode, minimum, maximum, range, and standard deviation for a specified column.
* Utilized the function to analyze the "BasePay" column.
* Used the **describe** method to obtain statistical summaries for numerical columns.
* Examined both numeric and non-numeric columns with the **describe** method.

**Data Cleaning:**

* Removed columns "Notes" and "Status" due to all-null values.
* Excluded the "Agency" column as it was identified as a leak column.
* Filled null values in the "Benefits" column with 0.
* Filtered out rows with negative values in various salary-related columns.
* Dropped remaining rows with null values.

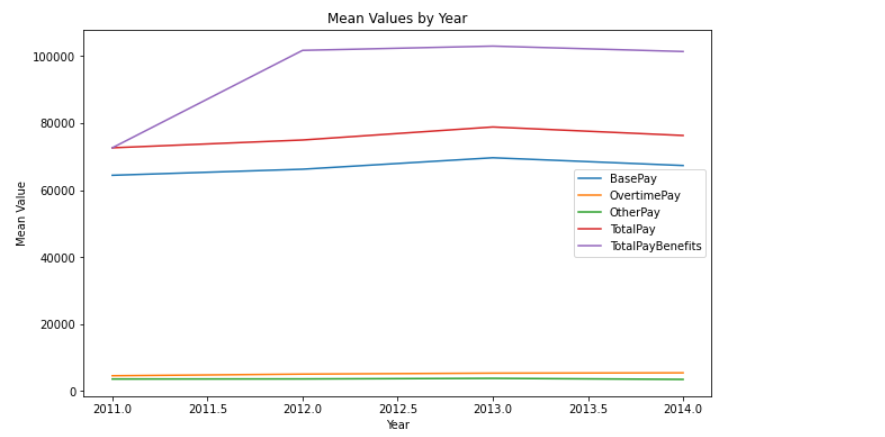
**Basic Data Visualization:**

* Plotted a histogram to visualize the distribution of salaries ("TotalPay").
* Created a pie chart to represent the proportion of employees in different job titles.
* Displayed a bar chart to illustrate the total pay over the years.
* Utilized a pair plot to show the distribution of each column and correlations between them.



**Grouped Analysis:**

* Grouped the data by year and computed the average salaries for different salary components.
* Plotted mean values for each salary component over the years.



**Simple Correlation Analysis:**

* Plotted a heatmap to visualize the correlation matrix between numerical columns.
* Observed correlations, particularly noting the correlation between "TotalPay" and "BasePay," "OvertimePay," and "OtherPay."
* Demonstrated the correlation between "TotalPay" and "BasePay" using a scatter plot.

