Summary:

In this task:

I chose to focus on the data cleaning and exploratory analysis of a dataset containing salary information, named 'Salaries.csv'.

Here are the steps I did:

1. Importing Libraries: The code begins by importing essential libraries, including Pandas for data manipulation, NumPy for numerical operations, SciPy for statistical functions, and Matplotlib for data visualization. These libraries are chosen for their efficiency and versatility in handling and analyzing structured data.

2. Reading the Dataset: The Pandas library is employed to read the CSV file into a DataFrame (df). Pandas is a preferred choice for handling tabular data due to its powerful and intuitive data structures.

3. Basic Data Exploration: The code performs initial exploration to understand the dataset's structure. Functions such as shape, dtypes, and isnull().sum() are used to retrieve information about the dataset's dimensions, data types of columns, and the count of missing values in each column, respectively. This information is vital for assessing data quality.

4. Handling Missing Values: To ensure data integrity, missing values in the 'TotalPay' column are replaced with the median salary using the fillna method. This choice is based on the robustness of the median in handling skewed data compared to the mean.
and I delete the empty columns as the dataset is large deleting empty columns or rows would not affect .

5. Descriptive Statistics: Descriptive statistics, such as mean, median, minimum, maximum, and histograms, are calculated and plotted using Matplotlib. These statistics provide a snapshot of the salary distribution, aiding in identifying central tendencies and potential outliers.

6. Pie Chart and Scatter Plot: A pie chart depicting the distribution of employees by year and a scatter plot visualizing the relationship between 'OvertimePay' and 'TotalPay' are created. These visualizations offer insights into yearly trends and the correlation between overtime pay and total pay.

7. Grouping and Correlation Analysis: Grouping the data by 'Agency' and calculating mean salaries provides insights into the average salary per agency. The correlation between 'TotalPay' and 'OvertimePay' is computed to understand the relationship between these variables.

8. Exporting Cleaned Data: The cleaned DataFrame is exported to a new CSV file named 'cleaned_salaries.csv'. This step ensures that the cleaned data can be used for further analysis or shared with others.