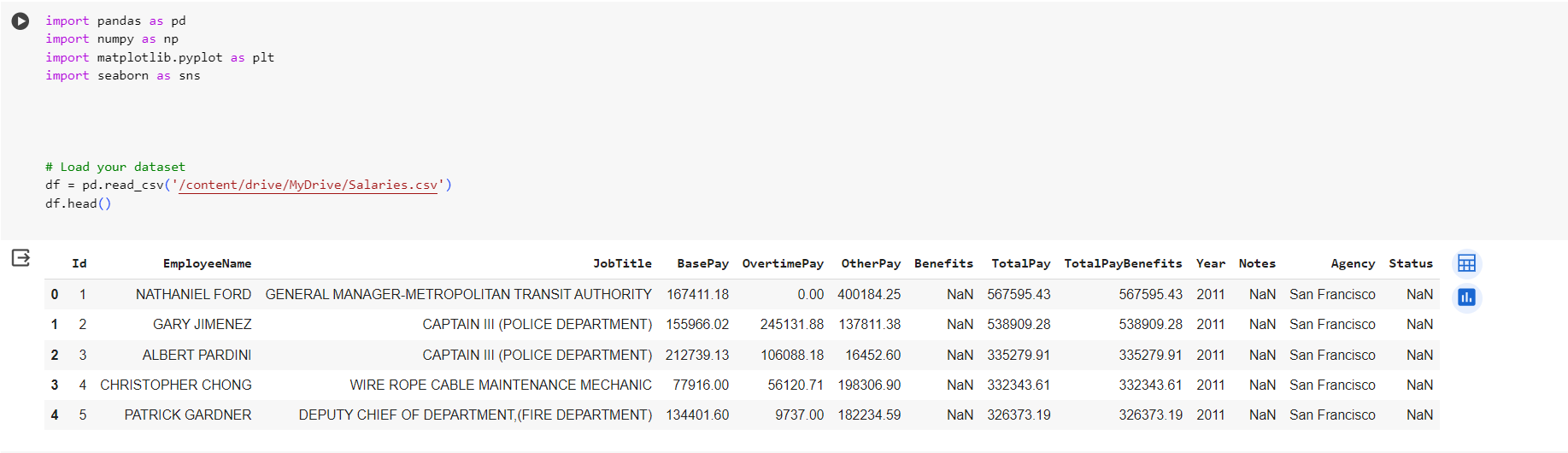


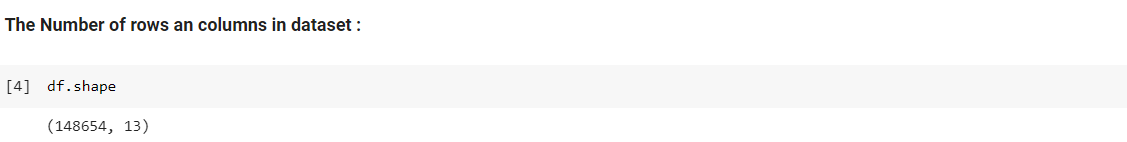
Rogina Alghizzawi

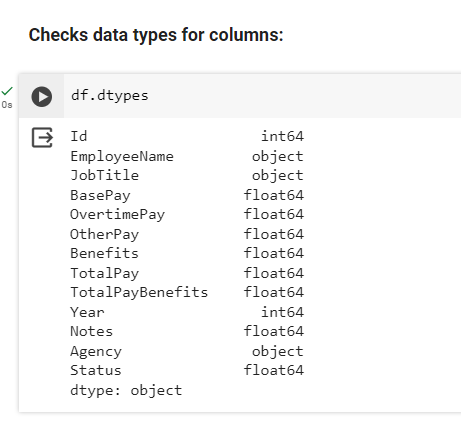
Data science and Artificial Intelligence

Explore and visualize

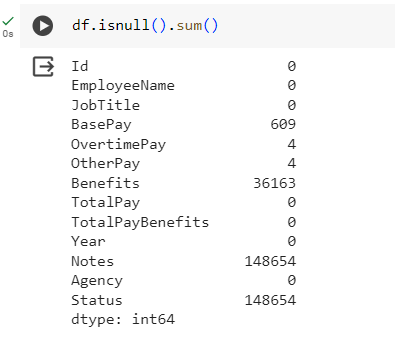
For Salaries dataset

1. We read data set and get first 5 rows using .head():

2.Show number of rows and columns in dataset, we have 148654 observation and 13 variable.

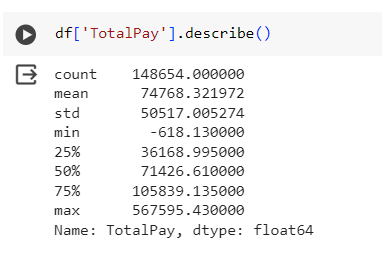
3.Check the data types of the variables:

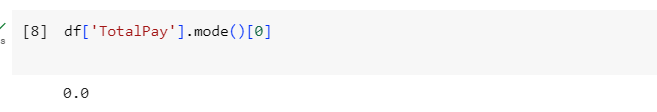
4. Check missing value : There is missing values in

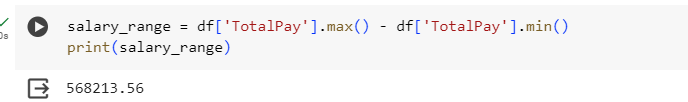
(BasePay, OtherPay, Benefits,Notes and Status):

5. There is two diffident way to check the descriptive statistics:

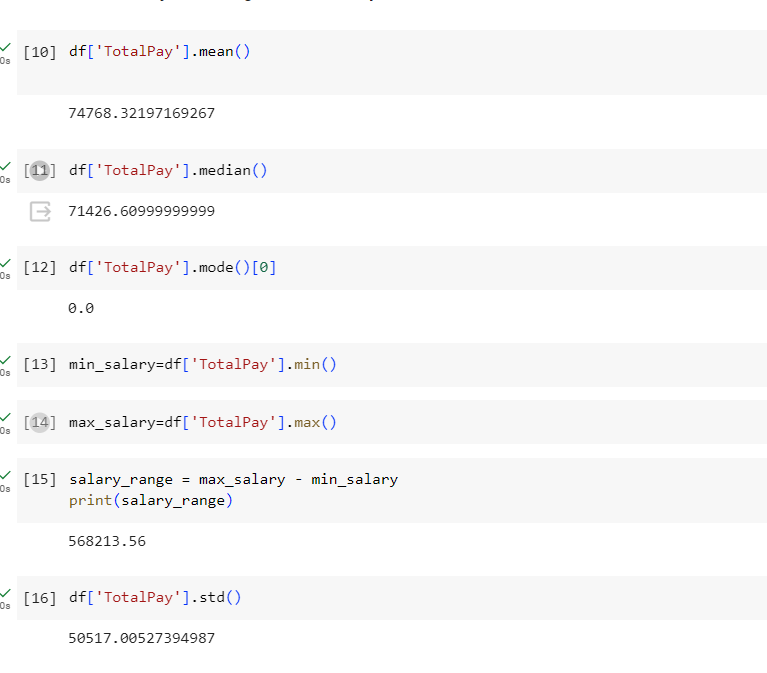
a. Use .describe():



* Count: The number of entries in the 'TotalPay' column. In this case, there are 148,654 entries.
* Mean: The average value of the 'TotalPay' column. The mean here is 74,768.32.
* Standard Deviation: 50,517.01. A higher standard deviation indicates greater variability in the data.
* Minimum: -618.13. negative values it is un real number for salary .
* 25th Percentile (Q1): 25% of salaries 36,168.99 and below.
* Median (50th Percentile): 71,426.61.
* **The mean grater than median so the data skewed to the right**
* 75th Percentile Q3: 75% of salaries 105,839.14 and below
* Maximum: The largest salary is 567,595.43.
* The mode is the value that appears most frequently in a dataset.it appears that the mode(s) for the 'TotalPay' column is 0
* Salary range showing the difference between the highest and lowest values.



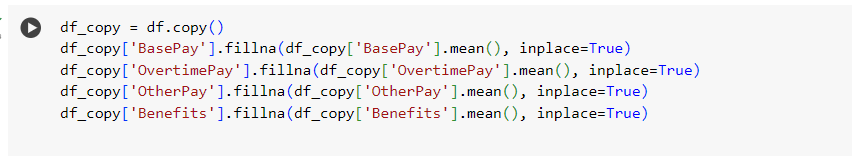
b. I use other way that can give us a descriptive statistics:



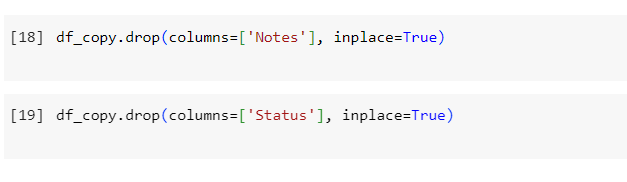
6. Handling Missing values:

1.I used .mean() function: The mean is not the robust way because it sensitive to outliers. So it is not the perfect way to handling missing value.

* I made a copy because I do not want use this use case but I just want to show how to handling missing values.

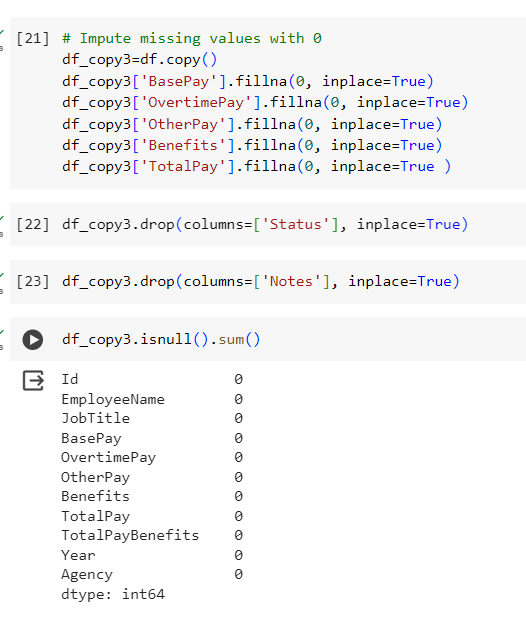


* There is no way to fill the missing value in Notes and Status columns so I dropped it:



2. fill missing value with zeros:

* I made a copy because I do not want use this use case but I just want to show how to handling missing values.



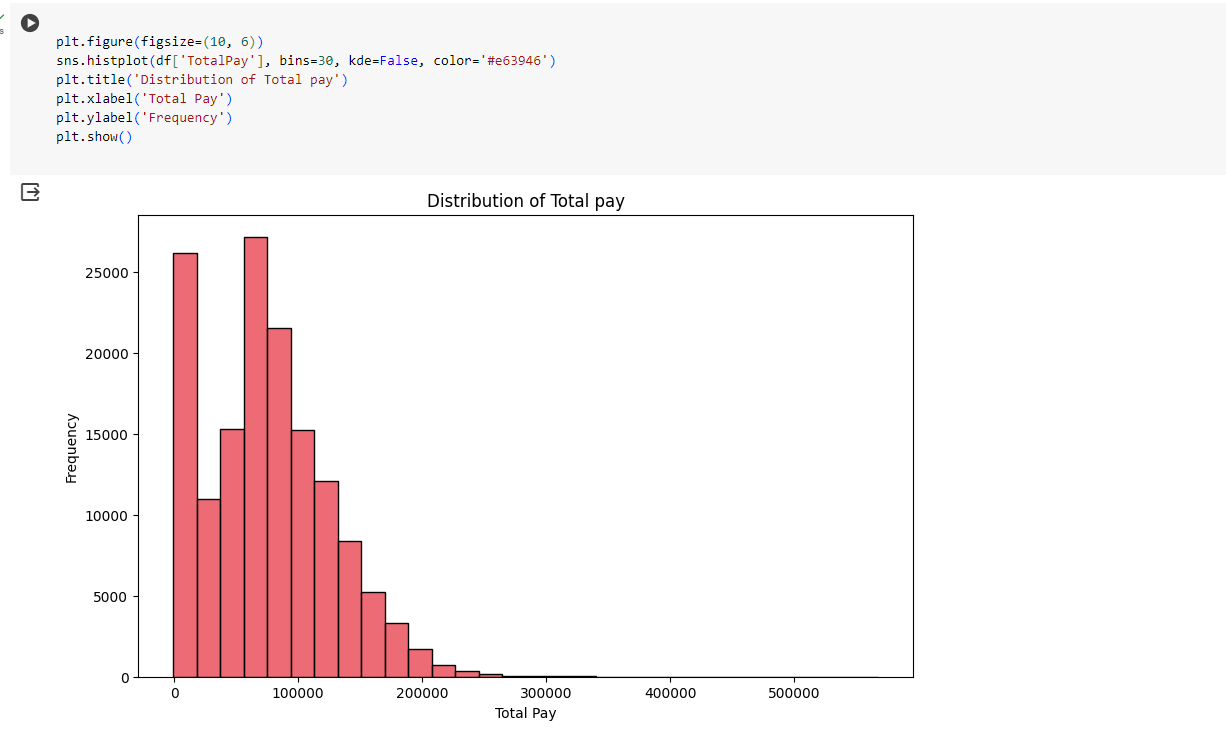
3.I used .median() function:

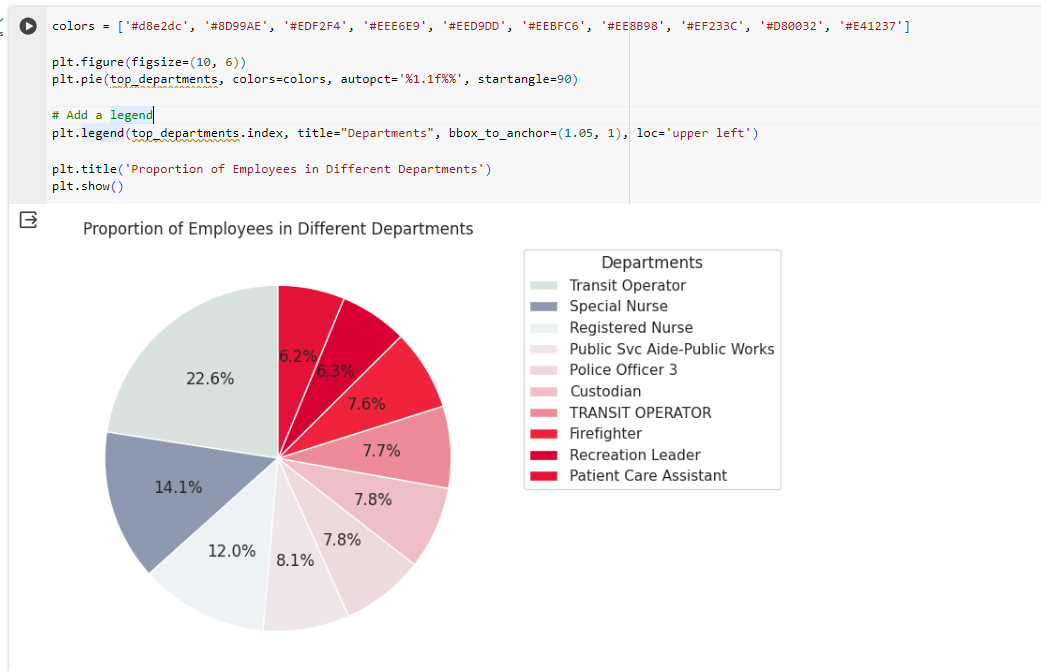
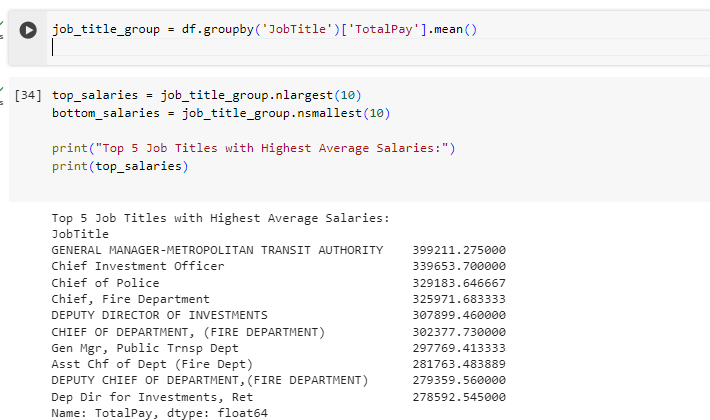
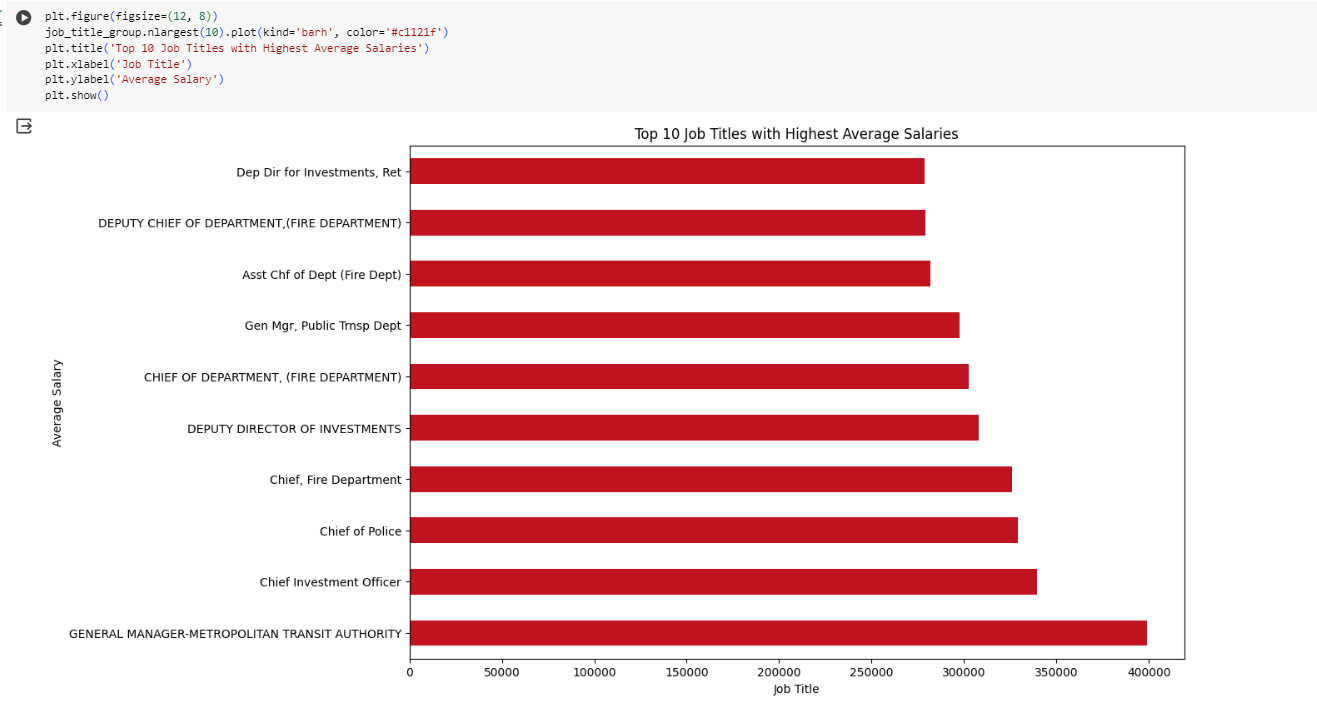
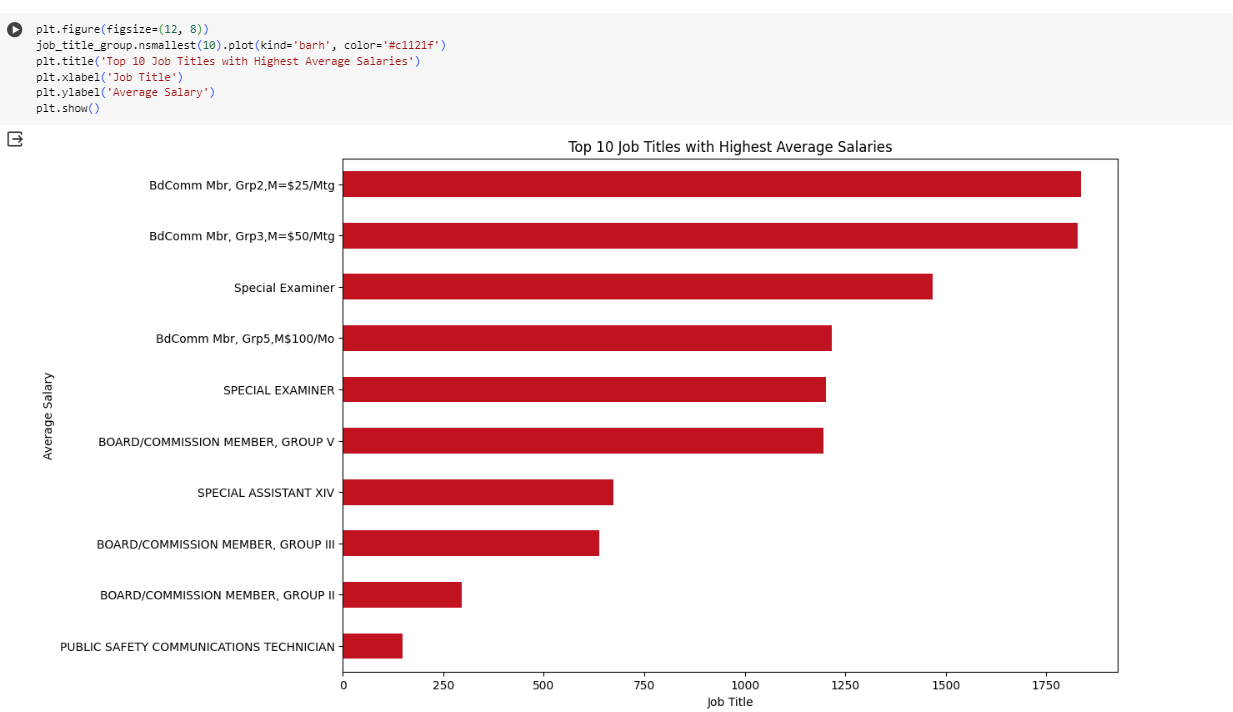
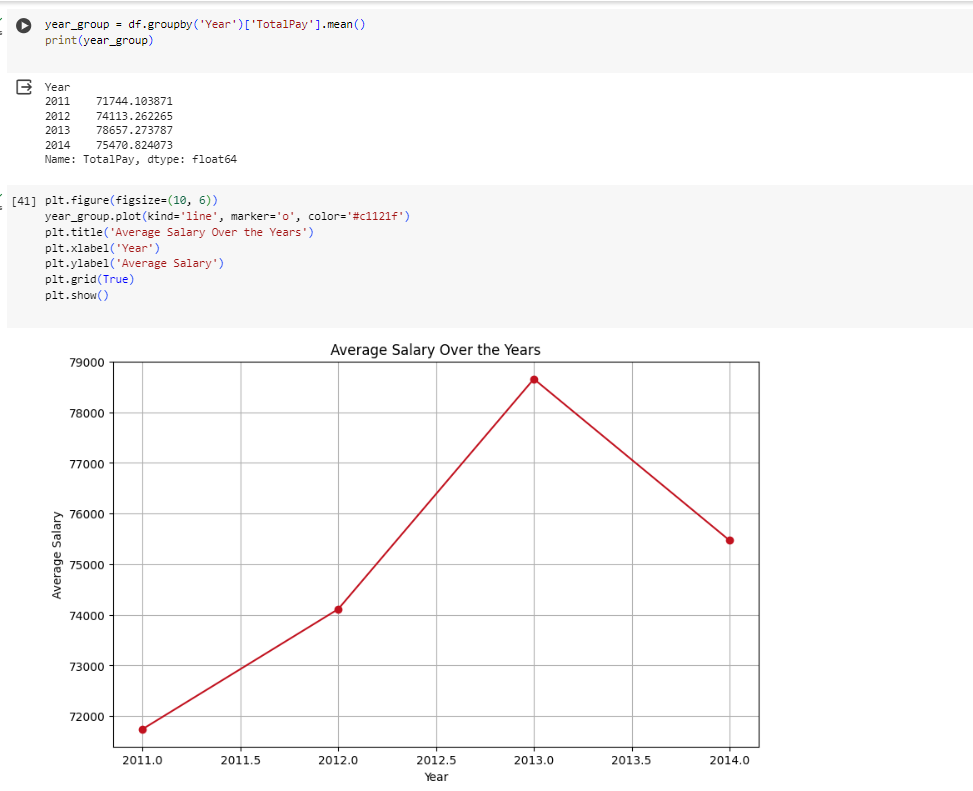
* it is more robust ,less affected by extreme
* it is the best way from my opinion to fill missing value
* fill missing value from other columns like totalpay column
* Totalpay(Salary):it's the sum of three columns (BasePay,OvertimePay,OtherPay)



7.Data Visualization:

* Histogram plot to visualize the distribution:The skewed to the right mean>meadian
* Histogram plot shows there is an outlier.



* Pie chart shows Proportion of Employees in Different Departments:
* Transact operator is the largest department
* Create new dayaframe by grouping the original dataframe by the 'JobTitle' column and then calculating the mean of the 'TotalPay' column for each group.
* top\_salaries = job\_title\_group.nlargest(10): selects the top 10 job titles with the highest average total pay.
* bottom\_salaries = job\_title\_group.nsmallest(10): selects the bottom 10 job titles with the lowest average total pay
* The top job title with the highest average total pay is GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY:
* The bottom job title with the lowest average total pay is PUBLIC SAFETY COMMUNICATIONS TECHNICIAN
* df.groupby('Year'): This part groups the original DataFrame df based on in the 'Year' column
* fter grouping by 'Year', the code then calculates the mean (average) of the 'TotalPay' column for each group
* in 2011, the average total pay was $71,744.10, and in 2014, it was $75,470.82.
* In 2013 the average total pay was the highest
* Positive Correlation: There appears to be a positive correlation between BasePay and TotalPay. This means that, generally, as BasePay increases, TotalPay also increases.
* Strength of the Relationship: The points tend to follow a linear pattern, but there is some spread. This suggests a moderate to strong linear relationship.
* Outliers: There are some points that are far removed from the main cluster, especially at the higher end of the scale for TotalPay. These outliers may represent cases where TotalPay is much higher than what would be typically expected based on BasePay.

