Shai assignment

Task 1:

There was 148654 record and 13 attribute, the ‘Notes’ and ‘Status’ attribute all there values were missing the ‘BasePay’ had 609 missing values the attributes [‘OvertimePay, OtherPay’] and had missing values in the same 4 attributes and ‘Benefits’ had 36163.here is the data type of each attribute .

```
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
```

Task 2:

Here is some basic statistics for the salaries

```
range     568213.56
mode      0.0
count     148654.000000
mean       74768.321972
std        50517.005274
min         -618.130000
25%        36168.995000
median     71426.610000
75%       105839.135000
max       567595.430000
```

Task 3:

I decided to get rid of the attributes that all there values were missing

Also the 4 records that had [‘OvertimePay, OtherPay’] missing values

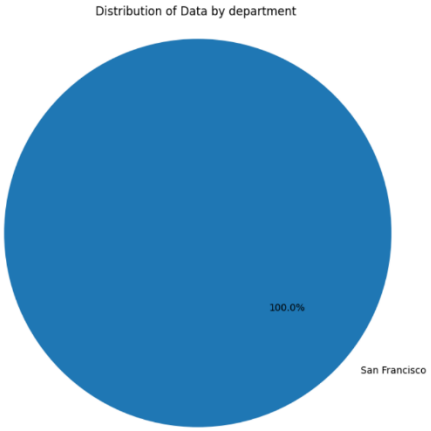
Then filled the null values in ‘Benefits’ by (TotalPayBenefits – TotalPay)

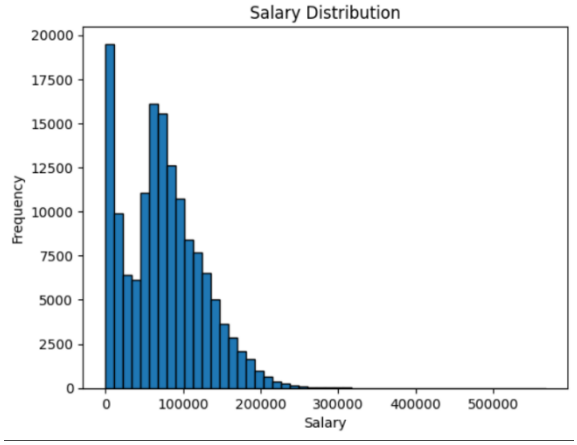
And the missing values in BasePay by (TotalPay-(OtherPay+OvertimePay)) here is the missing values in the data and shape now :

```
Id                 0
EmployeeName       0
JobTitle           0
BasePay            0
OvertimePay        0
OtherPay           0
Benefits           0
TotalPay           0
TotalPayBenefits   0
Year               0
Agency             0
dtype: int64
(148650, 11)
```

Task 4:

The salary distribution in right skewed there for I concluded the most the employees are in the low end of the salaries .and according to the pie chart all the employees from the data are in the San Francisco department .



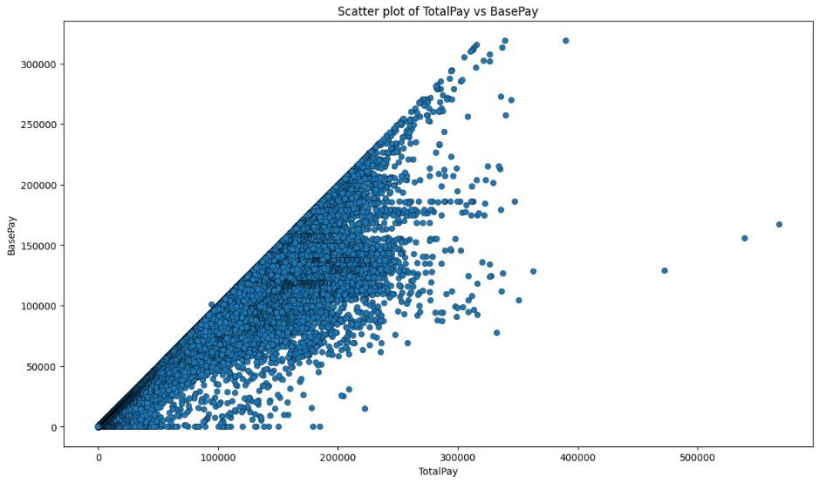


Task 5:

After grouping the data by the year I found that the year with the highest average salaries was 2013.

Task 6:

the attribute with highest correlation to salary was ‘BasePay’.



BasePay	0.954690
OvertimePay	0.504859
OtherPay	0.470496
Benefits	0.632187
TotalPay	1.000000
TotalPayBenefits	0.977312
Year	0.032145