Note: every value less than 1000 handled as rubbish value

1-We have 13 feature, 148653 sample of data(rows) There is no missing data in any dependent column, but there is 609 values missing in basepay feature,4 missing values in overtimePay, 4 missing values in otherPay and 36163 missing values in benefits.

After deleting rubbish values: 145060 sample of data(rows) There is no missing data in any dependent column, but there is 215 values missing in basepay feature,0 missing values in overtimePay, 0 missing values in otherPay and 35363 missing values in benefits feature.

2-mean= 74768.32197169267

median= 71426.60999999999

mode= 0

min= -618.13

max= 567595.43

range= 568213.56

standard deviation= 50517.00527394987

after deleting rubbish values:

mean= 74954.38346771419

median= 71553.0

mode= 18594.0

min= 0.3

max= 567595.43

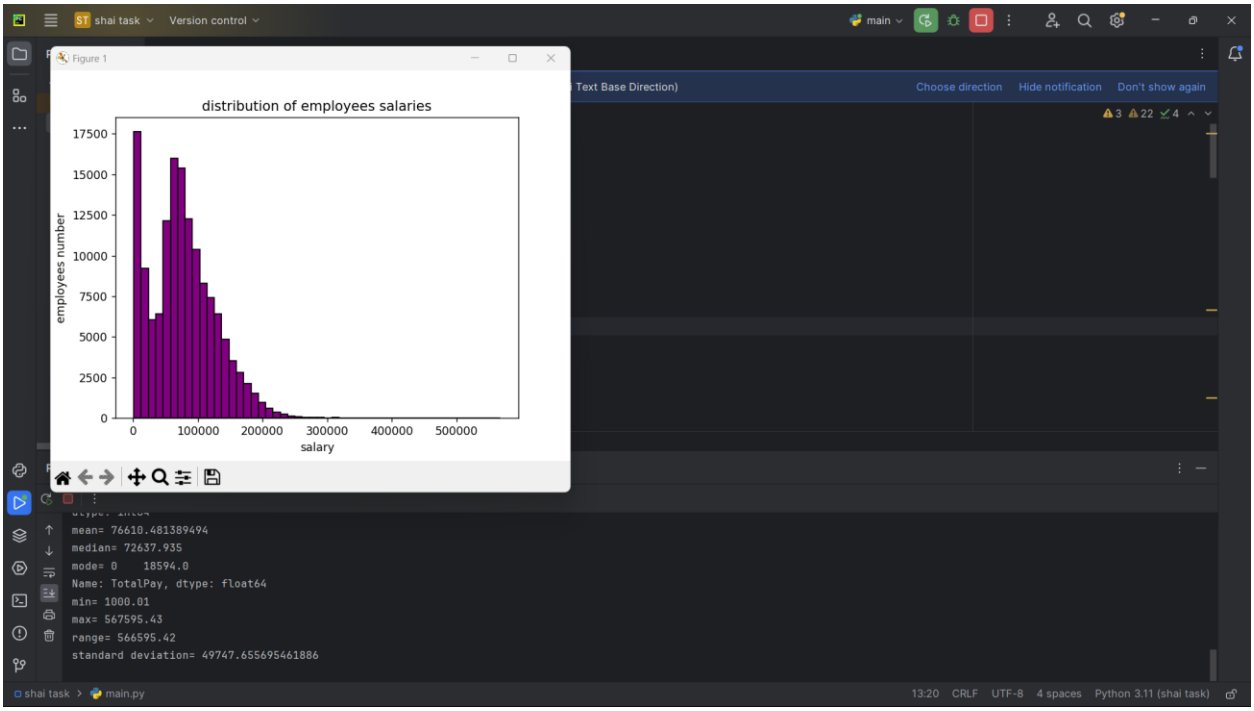
range= 567595.13

standard deviation= 50441.7662188465

3- I used mean to fill the missing data because using mode is not practical because its value=0, note: we can also use median

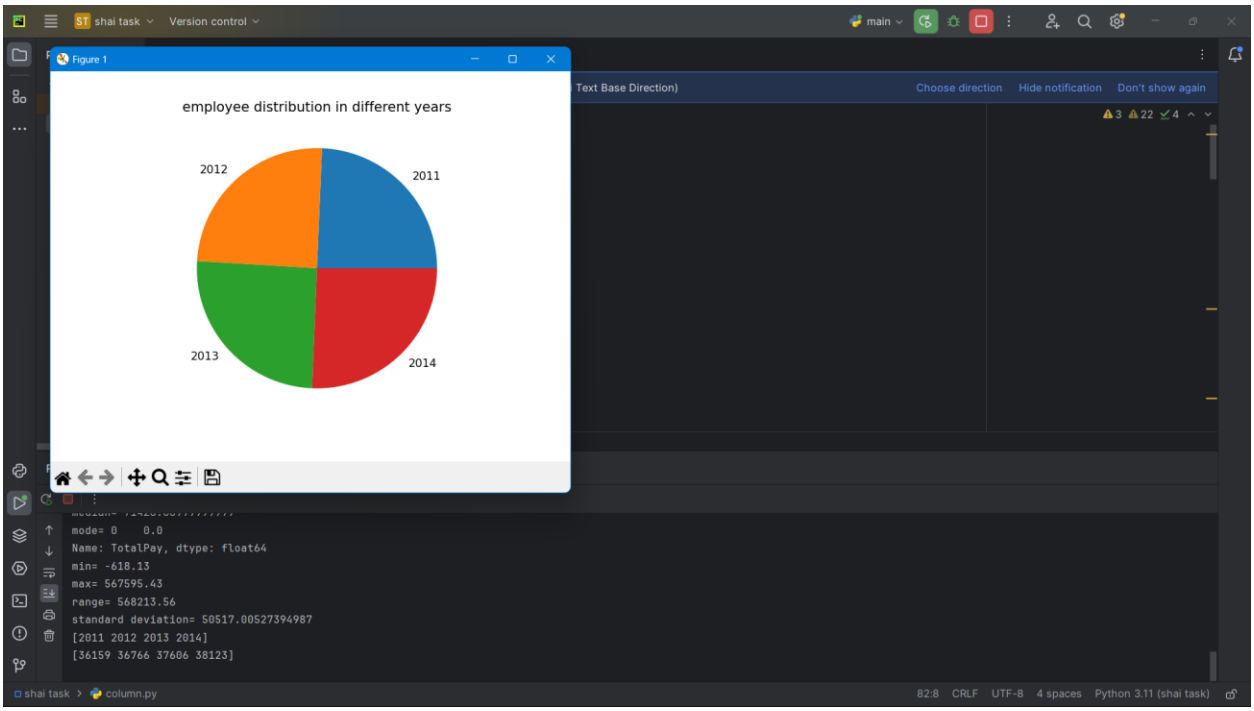
4-value 0 is the most frequent value and the period around 100000 containing the largest number of employees and after 220000 there is less than 1% of the total number of employees.

After deleting rubbish data:

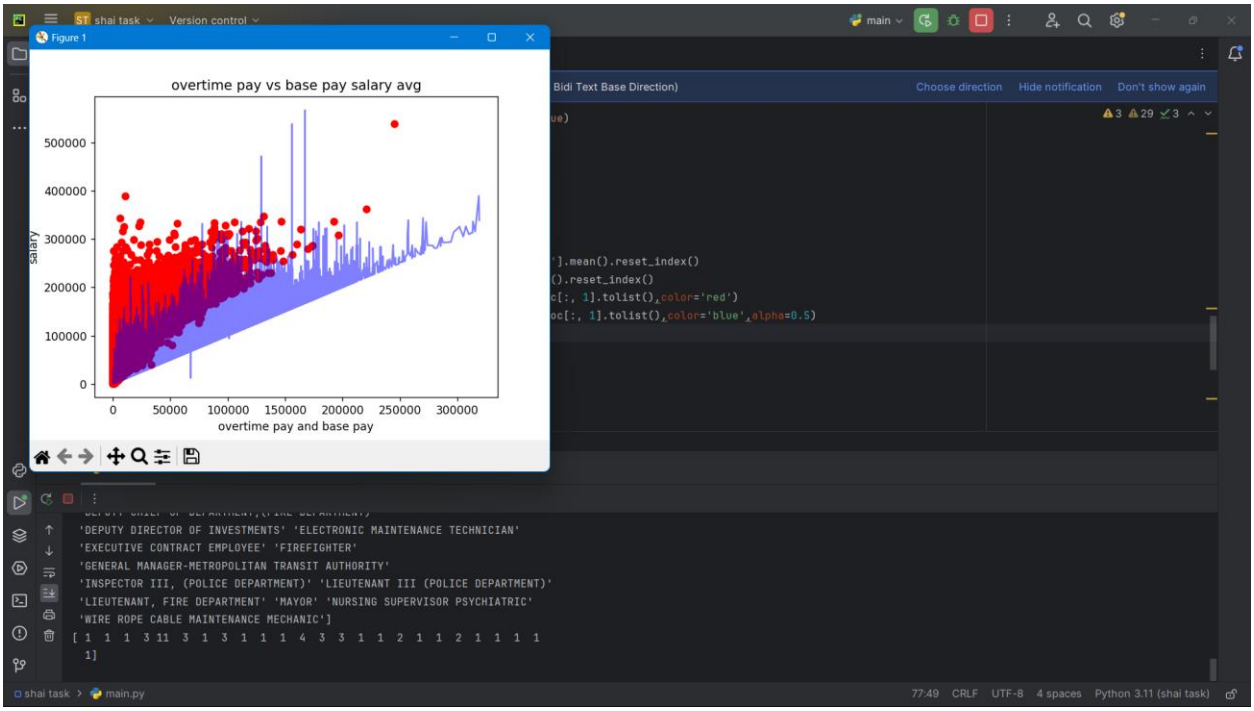


The rest of the charts havent any noticable change.

And the proportion of employees in different years is approximately one quarter per year from 2011 to 2014.



5)



6) total pay have high dependency with base pay but have much lower dependency with other pay.

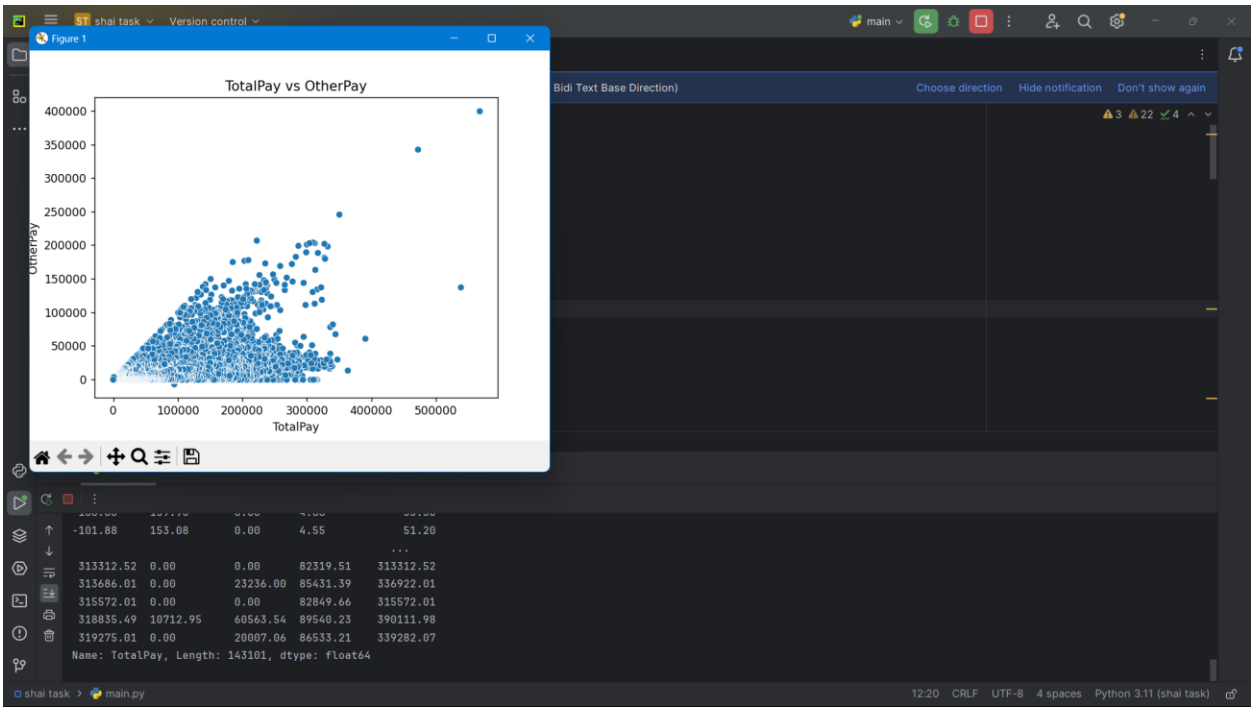


Figure 1
TotalPay vs OtherPay
OtherPay
TotalPay
0
50000
100000
150000
200000
250000
300000
350000
400000
500000
Bidi Text Base Direction)
Choose direction
Hide notification
Don't show again
-101.88 153.08 0.00 4.55 51.20
313312.52 0.00 0.00 82319.51 313312.52
313686.01 0.00 23236.00 85431.39 336922.01
315572.01 0.00 0.00 82849.66 315572.01
318835.49 10712.95 60563.54 89540.23 390111.98
319275.01 0.00 20007.06 86533.21 339282.07
Name: TotalPay, Length: 143101, dtype: float64
shai task
main.py
12:20
CRLF
UTF-8
4 spaces
Python 3.11 (shai task)


Figure 1
TotalPay vs BasePay
BasePay
TotalPay
0
50000
100000
150000
200000
250000
300000
400000
500000
x=4.47e+05 y=1.200e+05
Visual layout of bidirectional text can depend on the base direction (View | Bidi Text Base Direction)
Choose direction
Hide notification
Don't show again
import pandas as pd
import numpy as np
import seaborn as sn
from collections import Counter
-101.88 153.08 0.00 4.55 51.20
313312.52 0.00 0.00 82319.51 313312.52
313686.01 0.00 23236.00 85431.39 336922.01
315572.01 0.00 0.00 82849.66 315572.01
318835.49 10712.95 60563.54 89540.23 390111.98
319275.01 0.00 20007.06 86533.21 339282.07
Name: TotalPay, Length: 143101, dtype: float64
shai task
main.py
8:25
CRLF
UTF-8
4 spaces
Python 3.11 (shai task)