

missing data by suitable method know the two columns i will handle the missing data in it are BasePay and Benefits and i will leave OvertimePay ,OtherPay because the NAN value is only 4 and that willn't effect the data that much while Notes ,Status can be leaved or dropped as these two columns have no values at all and it doesn't affect the analysis of the salaries the basepay i will replace it by the mean or the median and the benefits i will replace it by 0

after knowing that there was missing data in basepay, benefits, overtime, otherpay, notes, status in benefits I filled the NAN columns by zero and in basepay I filled it by the median after plotting it and realized that the frequency at zero is very high and it is skewed and dropped the status and notes columns as they don't affect the data. Meanwhile the overtime and otherpay there was only 4 NAN so I left them

when plotting the totalpay we can see that most of the salaries are between 5000 and 100000 and most of the employees are transit operator

and the GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY has the highest mean between the other jobs

and the correlation between the totalpay and basepay is positive linear relationship