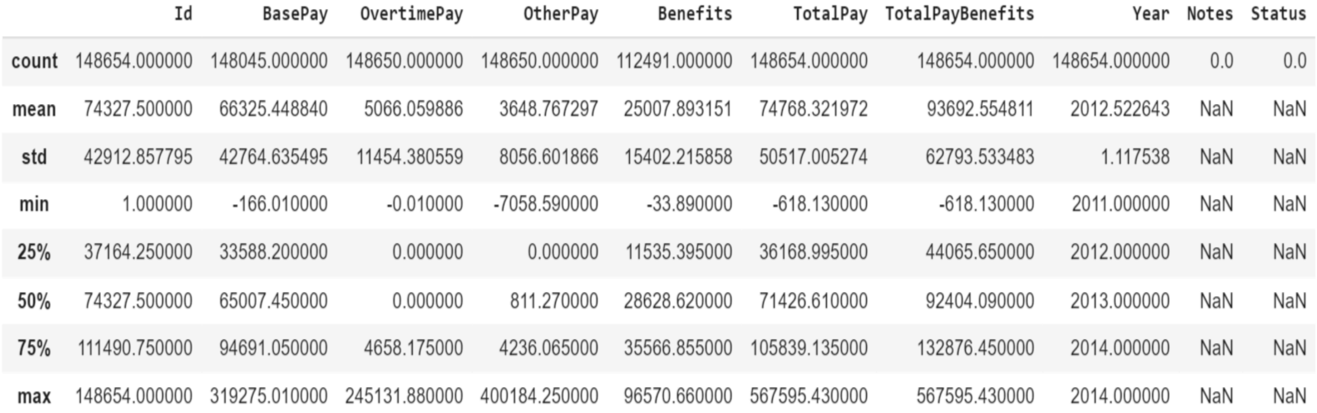
The provided employee compensation data from 2011 contains records for 148654 employees across various agencies. The data includes fields such as employee name, job title, base pay, overtime pay, benefits, total compensation, agency, and employment status.

I analyzed the data to gain insights into average salaries, highest and lowest paid roles and variability in total compensation amounts, and here's the key findings in each task:

- **Basic data exploration insights :**

1. We have 148654 raw in our dataset and 13 column.
2. The numerical data types that represents the salaries are of type float64.
3. This dataset has 609 missing value in **'BasePay'**, 4 in **'overtimePay'** , 4 in **'otherPay'**, 36163 in **'Benifits'** and two columns that have all missing values : **'Notes'** and **'status'.**

- **Descriptive statistics insights :**



	Id	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Status
count	148654.000000	148045.000000	148650.000000	148650.000000	112491.000000	148654.000000	148654.000000	148654.000000	0.0	0.0
mean	74327.500000	66325.448840	5066.059886	3648.767297	25007.893151	74768.321972	93692.554811	2012.522643	NaN	NaN
std	42912.857795	42764.635495	11454.380559	8056.601866	15402.215858	50517.005274	62793.533483	1.117538	NaN	NaN
min	1.000000	-166.010000	-0.010000	-7058.590000	-33.890000	-618.130000	-618.130000	2011.000000	NaN	NaN
25%	37164.250000	33588.200000	0.000000	0.000000	11535.395000	36168.995000	44065.650000	2012.000000	NaN	NaN
50%	74327.500000	65007.450000	0.000000	811.270000	28628.620000	71426.610000	92404.090000	2013.000000	NaN	NaN
75%	111490.750000	94691.050000	4658.175000	4236.065000	35566.855000	105839.135000	132876.450000	2014.000000	NaN	NaN
max	148654.000000	319275.010000	245131.880000	400184.250000	96570.660000	567595.430000	567595.430000	2014.000000	NaN	NaN

From the previous description of the data frame we can see that:

1. The min salary of employees was a negative number , which is something weird, i thought about removing those raws that has totalpay of negative number, but it turns out that this employee has othersPay value of negative number and all other numerical values are 0, so it may be a bonus or a commission that the company have to pay to that employee.
2. The max salary of employees is 567595.430000 , and it turns out that the job title of that employee is **general manager-metropolitan transit authority** and it was in 2011.

3. The mean value of salary is 74768.321972.
4. The range of salary after subtracting the min value from the max value is 568213.56.

- **Data cleaning insights :**

1. There's no duplicated raws in the dataset.
2. When i was trying to check the outliers for the totalPay , i found some salaries that goes beyond the normal range, and it can be normal as we can have different unique job titles in the dataset, so removing those raws that goes beyond the normal range wasn't a good idea in order not to lose valuable information.
3. There are 4 raws found that have **'Not provided'** information, so i dropped them.
4. The two columns **'notes'** and **'status'** are dropped, as they have no valuable information to our dataset.
5. It was found that the **'Basepay'** and **'Benifits'** columns have null values , so i decided to fill these cells with **0**, as it won't affect the submission of the **totalPay**.

- **Basic data visualization insights:**

1. I used histogram to represent the distribution of salaries among all employees, and it turns out that around more than 60,000 employee have salary less than 100,000.
2. Around 30,000 have range of salaries from +100,000 to 150,000.
3. Almost 10,000 employee have range of salaries from 150,000 to 200,000.
4. A few percentage of employees take more that 200,000.
5. The next task was representing the distribution of the employees in different departments, for this task i used pie chart, the dataset hasn't included the **deparetemt** column, so i extracted it from the **jotTitle** column if it's mentioned , and for

all job titles that didn't include the department, i gave those cells **'no department sepicified'** string.

6. There were some departments that has the same meaning but written in different formats, so i brought together those ones, ex: **fire dept** and **fire department.**
7. Finally, from the pie chart i've concluded that the vast majority of employees belong to **civil& criminal** departement.
8. The next departement that represented a big number of employees belonging to it was the **police deparetement**.
9. The next departments were almost equally partitioned.

- **Basic data visualization insights:**

1. The first group made was the average salaries by department, and turns out that **fire department** has the biggest average salary among all other departments.
2. The second group made was the average salaries by department, and from the analysis we cas see that the job title that has the biggest average salary was **general manager-metropolitan transit authority.**

- **Simple correlation analysis insights:**

1. From the correlation matrix, we can see that the **totalPay** is very correlated to **totalPayBenifits** and **BasePay**.
2. From the scatter Plots , we can see that **totalPay** has actually linear relationship with **totalPayBenifits** and **BasePay** , and slight linearity with other kind of pays.