اول شي قرينا الداتا..

df.columns : هون استعرضنا أسماء الاعمدة بال داتا ....

df.shape تحدد ابعاد الداتا :

أسطر: 148654

اعمده: 13

زلنا عمودين غير مفيدين لانو كل قيمن nun وهنن: Notes....Status

عدد الأعمدة: 11

ضل عنا الاعمدة:

'Id', 'EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year', 'Agency'

df.head(): اعرضنا اول خمس اسطر بالداتا

df.describe(): اعرضنا بعض العمليات الإحصائية على القيم الرقمية بالداتا

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

Count: عدد القيم لكل عمود

Mean: المتوسط الحسابي للقيم لكل عمود

Std: الانحراف المعياري للقيم لكل عامود

Min: اصغر قيمة لكل عامود

25%: قيمة 25% لكل عمود

50%: قيمة 50% لكل عمود

75%: قيمة 75% لكل عمود

اكبر قيمة لكل عمود :Max

نوع القيم في كل عمود :df.dtypes

```
Id                   int64
EmployeeName        object
JobTitle            object
BasePay            float64
OvertimePay        float64
OtherPay           float64
Benefits           float64
TotalPay           float64
TotalPayBenefits   float64
Year                 int64
Agency              object
dtype: object
```

```
print(df.isnull().sum())
```

```
Id                     0
EmployeeName           0
JobTitle               0
BasePay              609
OvertimePay            4
OtherPay               4
Benefits           36163
TotalPay               0
TotalPayBenefits       0
Year                   0
Agency                 0
dtype: int64
```

عدد قيم nun في كل عمود

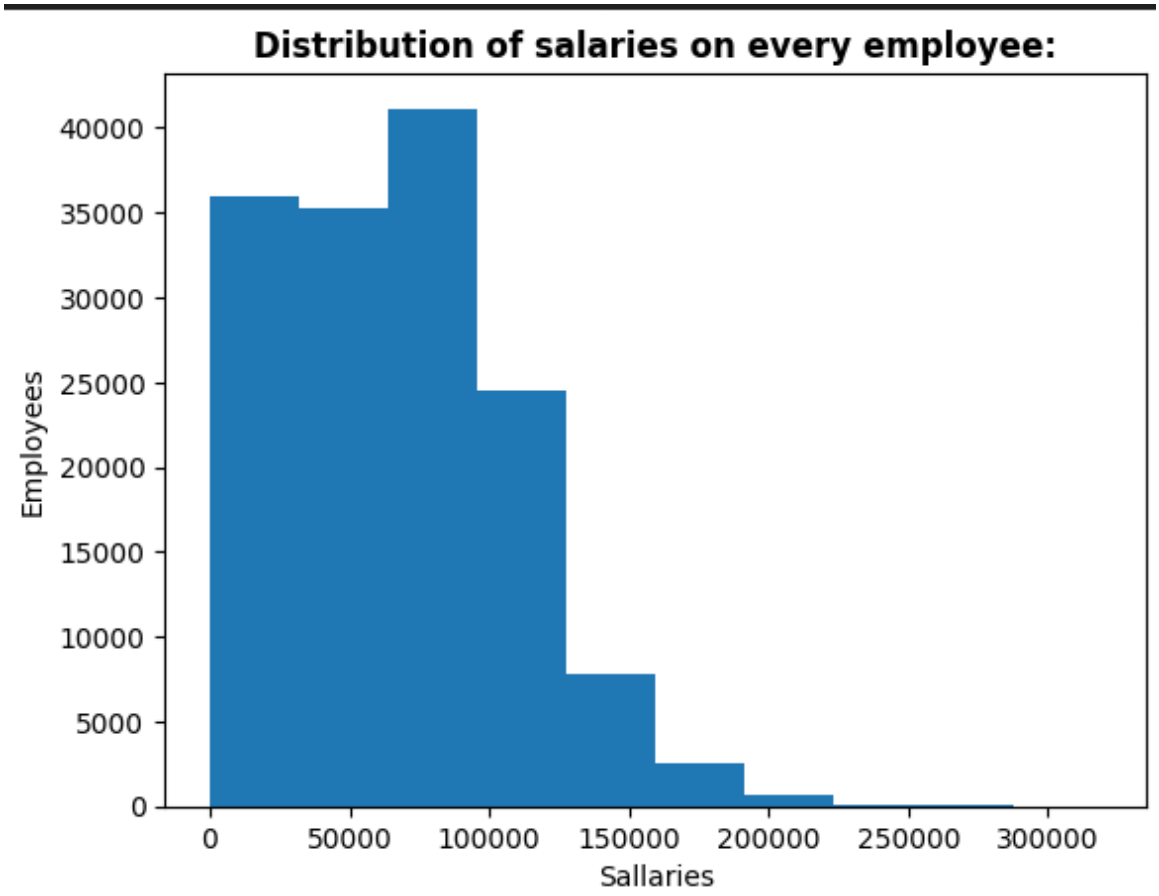
نستخدم تابع fillna ونستبدل قيم nun بال mean لقيم float واستبدلنا بالمتوسط حتى يكون انزياح المتوسط للداتا اقل ما يمكن.

القيام ببعض العمليات الاحصائية على رواتب الموظفين:

القيم الأكثر تكرارا:Mode

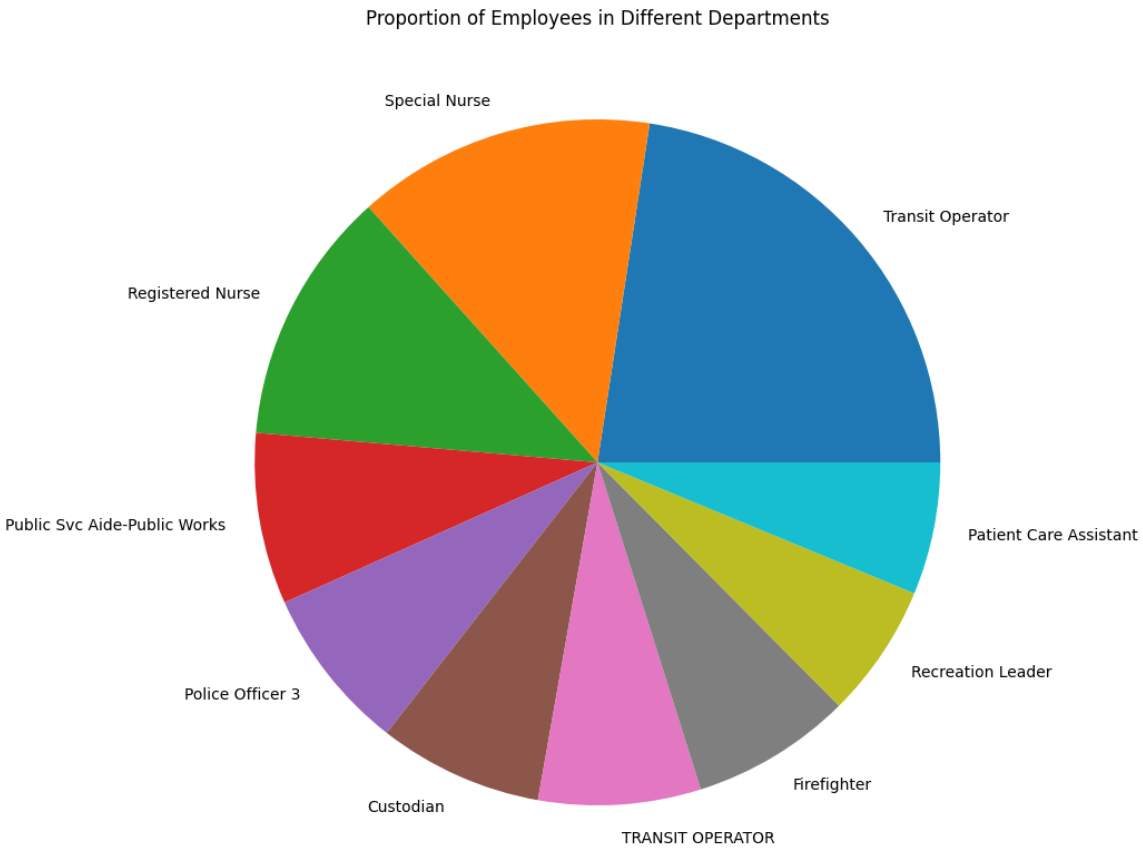
```
Mean:  66325.44884048769
Median:  65092.19
Minimum:  -166.01
Maximum:  319275.01
Range:  319441.02
Standard Deviation:  42676.946744797686
Mode:  0    0.0
Name: BasePay, dtype: float64
```

عرض توزع الرواتب على عدد الموظفين



أكبر عدد للموظفين تقريبا 40000 يقبضون الرواتب تقريبا 100000

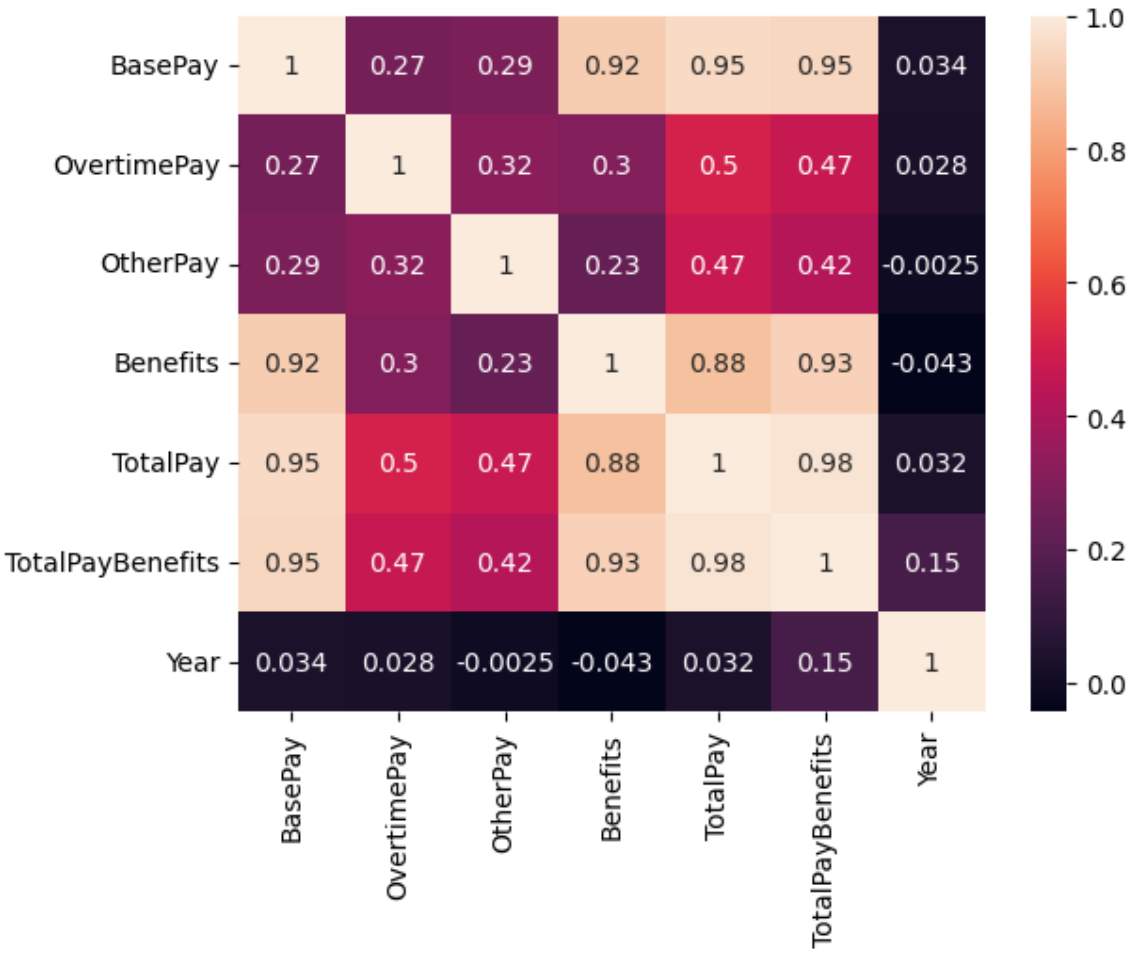
اقل عدد من الموظفين تقريبا اقل من 5000 يقبضون بين 150000 و200000



أكبر عدد للموظفين في اول عشر اقسام:

أكبر عدد للموظفين هو في قسم Transit Operator

أصغر عدد للموظفين في اكبر عشر اقسام هو قسم Patient Care Assistant



هذا الجدول يعبر عن مدى الارتباط الاحصائي بين كل عمود وعمود اخر للأعمدة الرقمية:

كل ما كان الرقم أقرب من 1 و1- العمود يكون أكثر ارتباطا احصائيا من بقية العواميد

-1 يعني ارتباط عكسي

كل ما كان الرقم أقرب من الصفر يكون العمود ابعد من ان يكون مرتبطا احصائيا من بقية العواميد

مثلا:

OtherPay & Year: -0.0025

هون ارتباط عكسي بس القيمة أقرب للصفر وهاد بيعني انو اكتر عمودين استقلالا احصائيا عن غيرة من الاعمدة

TotalPay & TotalPayBenefits: 0.98

هون ارتباط إيجابي لانوا العامودين ابعد ما يمكن عن الصفر وأقرب للواحد