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: Status.... Notes

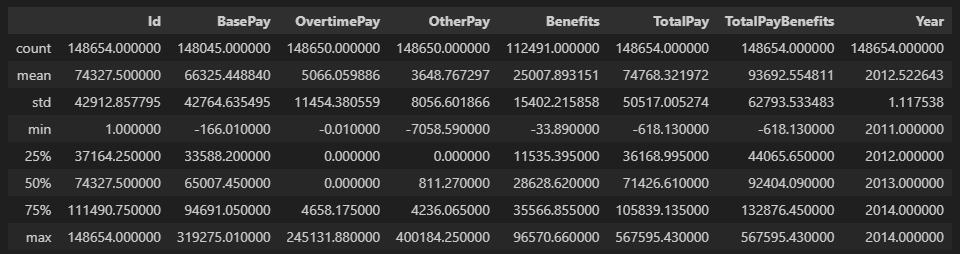
عدد الأعمدة: 11

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 :



Count: عدد القيم لكل عمود

Mean: المتوسط الحسابي للقيم لكل عمود

Std: الانحراف المعياري للقيم لكل عامود

Min: اصغر قيمة لكل عامود

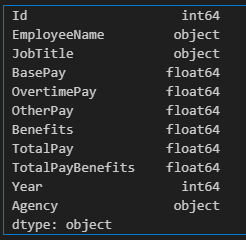
25%: قيمة 25% لكل عمود

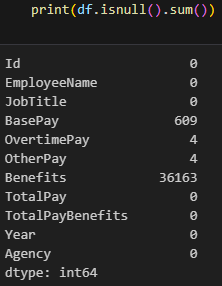
50%: قيمة 50% لكل عمود

75%: قيمة 75% لكل عمود

Max: اكبر قيمة لكل عمود

df.dtypes: نوع القيم في كل عمود



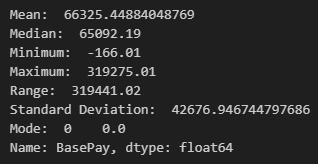


عدد قيم nun في كل عمود

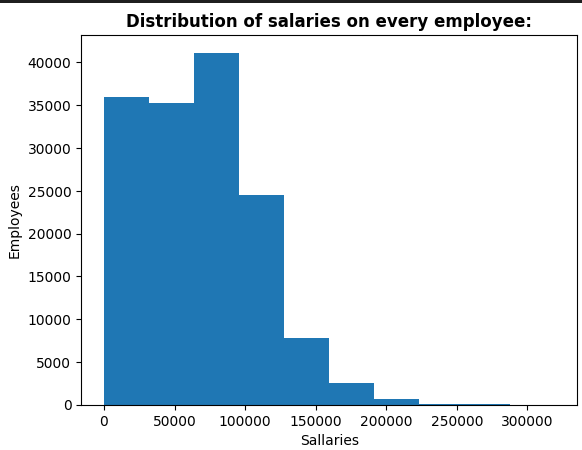
نستخدم تابع fillna ونستبدل قيم nun بال mean لقيم float واستبدلنا بالمتوسط حتى يكون انزياح المتوسط للداتا اقل ما يمكن.

القيام ببعض العمليات الاحصائية على رواتب الموظفين:

Mode القيم الأكثر تكرارا:

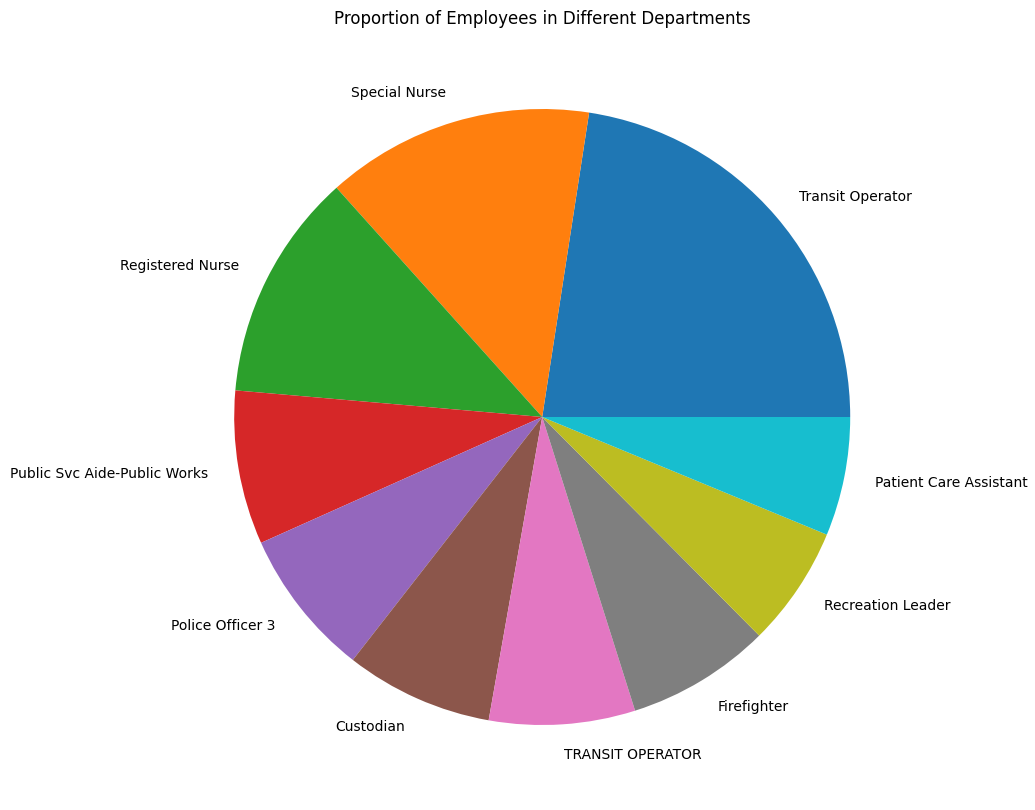


عرض توزع الرواتب على عدد الموظفين



أكبر عدد للموظفين تقريبا 40000 يقبضون الرواتب تقريبا 100000

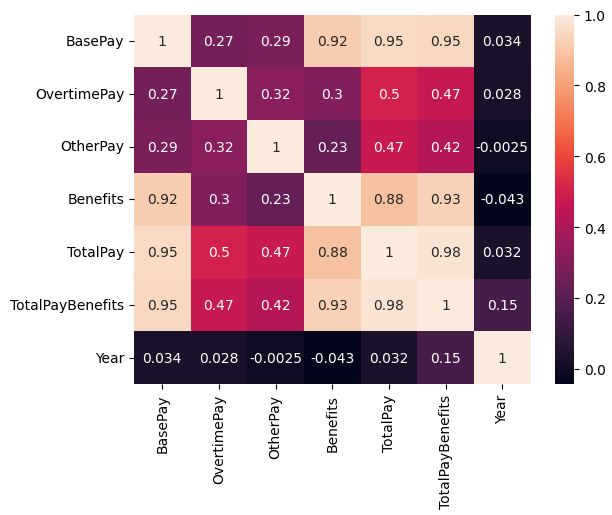
اقل عدد من الموظفين تقريبا اقل من 5000 يقبضون بين 150000 و200000



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