shai for AI

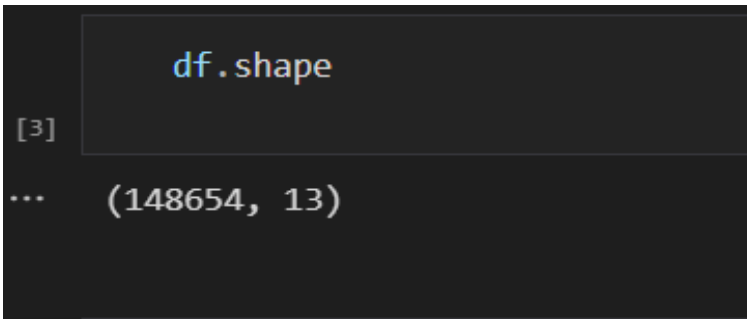
مجتمع شاي للذكاء الاصطناعي

إعداد : سيدرا الصباغ

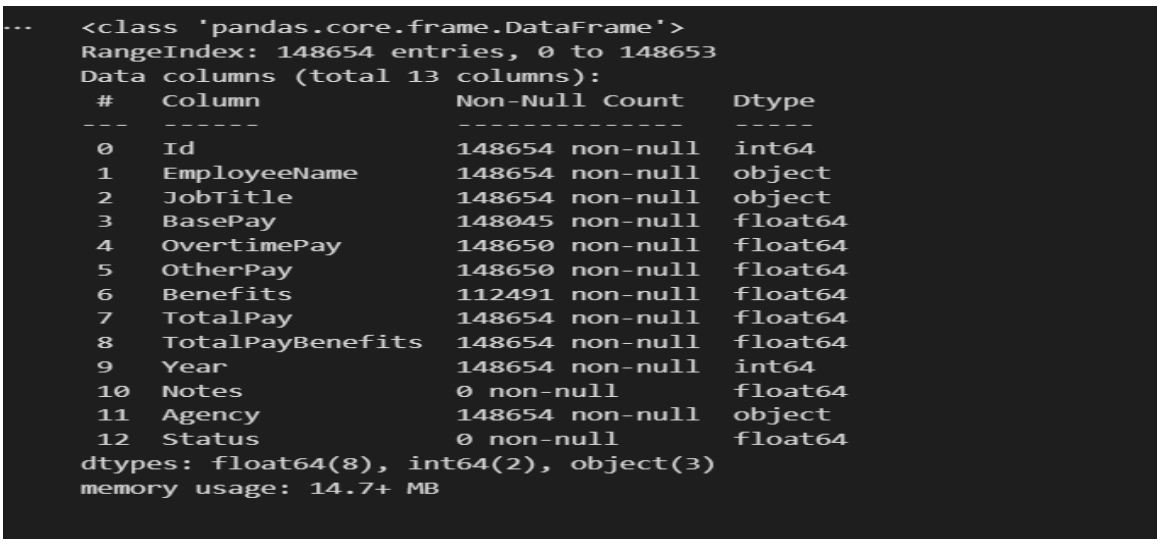
2024 م

1.حدد عدد الصفوف والاعمدة في مجموعة البيانات وحدد أنواع البيانات لكل عمود وتحقق من القيم المفقودة في كل عمود :

بعد قراءة مجموعة البيانات باستخدام مكتبة pandas وتحويلهم إلى بنية جدول dataframe نلاحظ النتائج التالية :
البيانات مؤلفة من 148654 صف و 13 عمود .



من أجل تحديد نوع البيانات نقوم باستخدام التابع info الذي يقوم بتحديد نوع البيانات لكل عمود مع حساب عدد الحقول التي ليست فارغة



نقوم باستخدام التابع isna لحساب عدد القيم غير المعتادة (NaN) ، و وظيفة isnan() تقوم بإرجاع قيمة بولية (True/False) لكل عنصر في المصفوفة تشير إلى ما إذا كان العنصر NaN أم لا، وبعد ذلك يتم استخدام .sum() لحساب عدد القيم التي تساوي True/false .
تظهر لدينا النتيجة التالية :

```
...   Id                      0
      EmployeeName            0
      JobTitle                0
      BasePay               609
      OvertimePay             4
      OtherPay                4
      Benefits            36163
      TotalPay                0
      TotalPayBenefits        0
      Year                    0
      Notes              148654
      Agency                  0
      Status             148654
      dtype: int64
```

2. حساب الإحصائيات الأساسية للمتوسط والوسيط و الوضع والحد الأدنى و الحد الأقصى للراتب وتحديد نطاق الرواتب والعثور على الانحراف المعياري :

من اجل حساب الاحصائيات نقوم باستخدام دالة describe وتظهر النتائج التالية حيث أن

count: يُظهر عدد القيم غير المفقودة في كل عمود.

mean: يُظهر المتوسط الحسابي لكل عمود.

std: يُظهر الانحراف المعياري لكل عمود، وهو قياس لانتشار البيانات حول المتوسط كلما زادت قيمة الانحراف المعياري، زادت انتشار البيانات.

min: يُظهر أصغر قيمة في كل عمود.

25% , 50%, 75%: تُظهر الربع الأول (25th percentile)، الوسط (median)، والربع الثالث (75th percentile) لكل عمود. مثلاً، 25% من القيم في عمود "BasePay" أقل من 33588.2 و50% أقل من 65007.45.

max: يُظهر أكبر قيمة في كل عمود.

	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

3.تعامل مع البيانات المفقودة بالطريقة المناسبة مع شرح سبب استخدامها :

بعد عرض الاعمدة التي تحوي قيم فارقة يجب علينا تعبئة تلك القيم يوجد لدينا طرق عدة لحل تلك المشكلة منها simleimputer و استخدام تابع filnan الخ..
سوف اقوم باستخدام تقنية KNN Imputer لتعبئة القيم الفارغة في البيانات لان البيانات ذات طبيعة رقمية وتحتوي على علاقات معقدة بين السجلات. يعتمد هذا النوع من التعبئة على القرب الجغرافي للبيانات في الفضاء المتعدد الأبعاد.
في مكتبة scikit-learn. يمكن استخدامها لتعبئة القيم الفارغة في البيانات باستخدام أقرب الجيران (K Nearest Neighbors) لكل سجل.

وفيما يخص بارمتراتها قمت باختيار المسافة الاقليدية للحساب البعد بين الجيران و اختيار عدد الجيران يساوي 4 .

نلاحظ ايضاً يوجد اعمدة لا تحوي ولا قيمة مثل حقل notes و status و حقل benefits يحوي على قيم فارغة باعداد كبيرة لذلك قمت بحذفهم .

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

قبل التعبئة والحذف

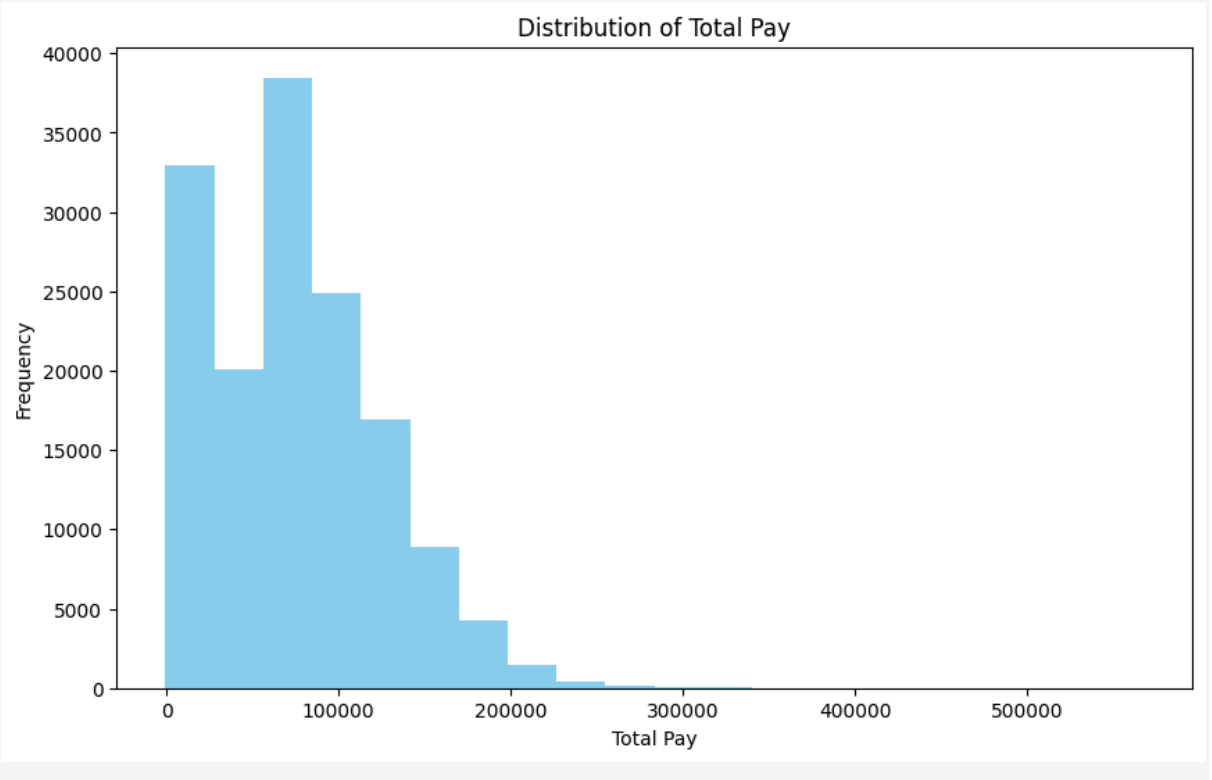
```
Id                0
EmployeeName      0
JobTitle          0
BasePay           0
OvertimePay       0
OtherPay          0
TotalPay          0
TotalPayBenefits  0
Year              0
Agency            0
dtype: int64
```

بعد التعبئة والحذف

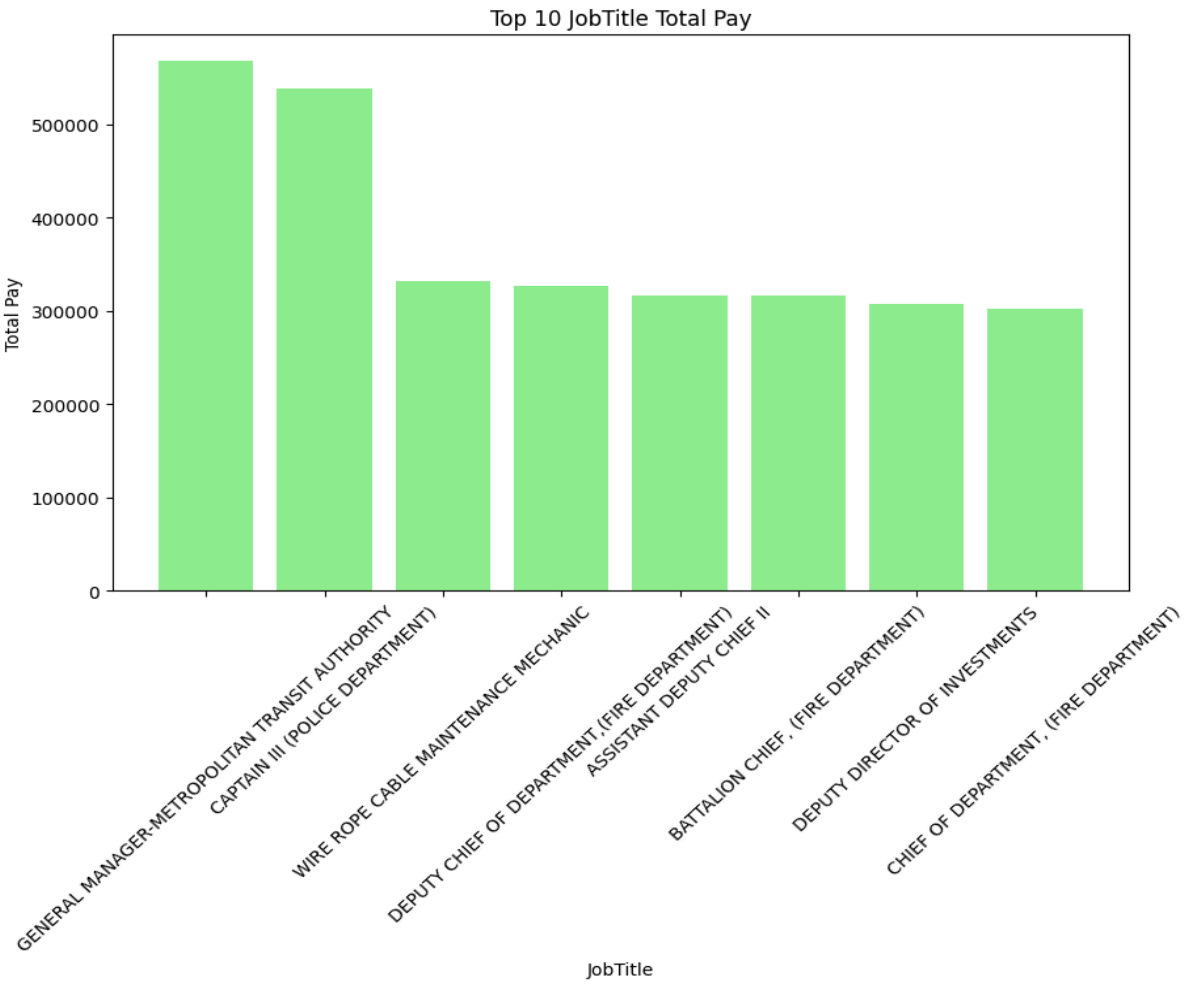
4.قم بإنشاء رسوم بيانية أو مخططات شريطية لتصوير توزع الرواتب واستخدم المخططات الدائرية لتمثيل نسبة الموظفين في الاقسام المختلفة :

إنشاء رسوم بيانية

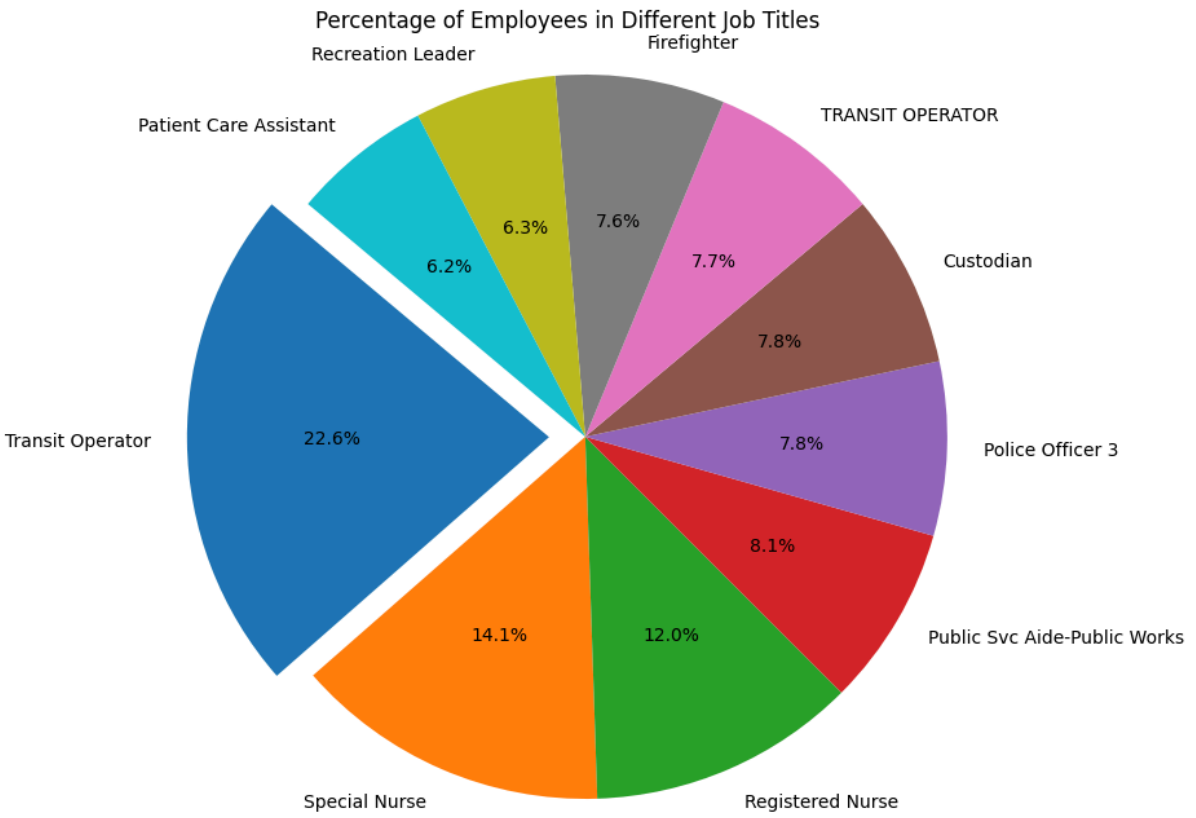
بعد رسم total pay نلاحظ أكثر قيمة متكررة قريبة من 60000



طريقة أخرى للرسم باستخدام bar char لتوضيح توزع الرواتب وبما أن الرواتب عددها كبير جداً قمت باختيار أكثر نوع عمل متكرر

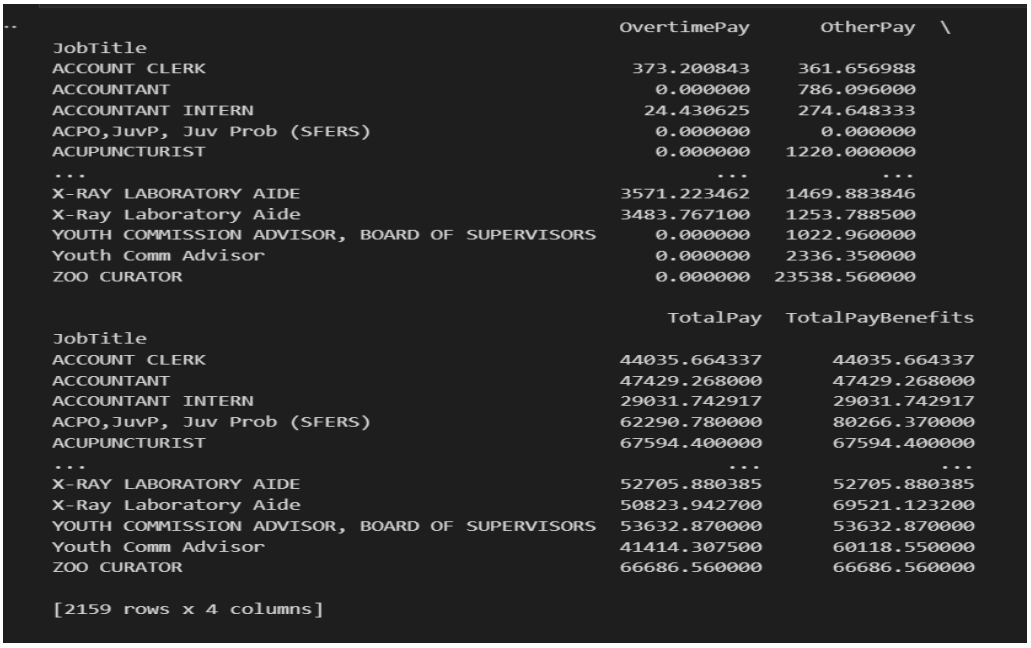


من خلال استعراض المخطط الدائري ،يتبين أن الوظيفة transit operator هي الأكثر شيوعاً بنسبة تقدر حوالي 22.6% من إجمالي الوظائف العشرة الأعلى تكراراً .



5.قم بتجميع البيانات حسب عمود واحد أو أكثر وحساب الاحصائيات لكل مجموعة ومقارنة متوسط الرواتب :

تجميع البيانات في عمود واحد يسهل عملية تحليل البيانات وفهمها بشكل أفضل ، لذلك قمت بتجميع حقول التالية overtimepay - otherpay -totalpay -totalpaybenefits
ثم قمت بعرض متوسط الحسابي لكل مجموعة من اجل المقارنة

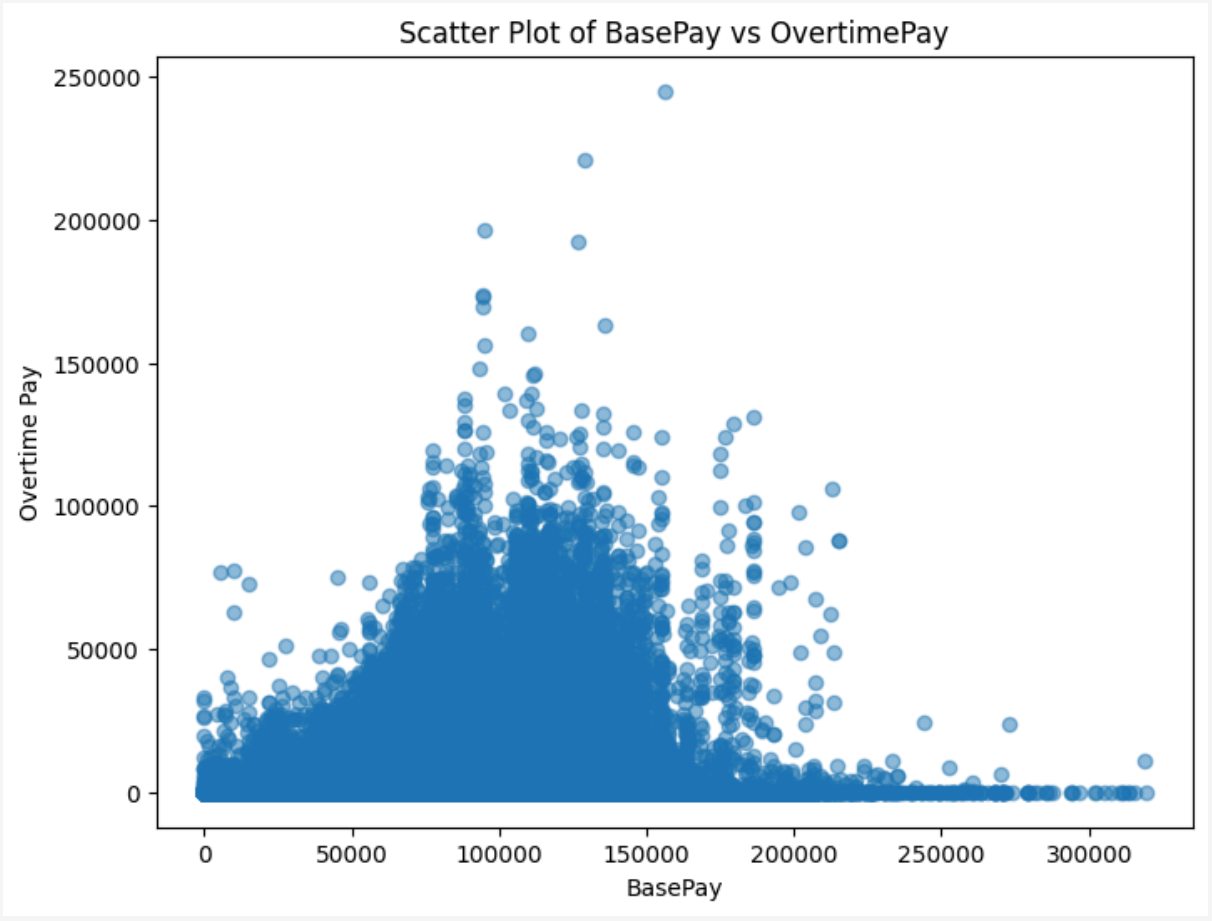


JobTitle	OvertimePay	OtherPay	TotalPay	TotalPayBenefits
ACCOUNT CLERK	373.200843	361.656988	44035.664337	44035.664337
ACCOUNTANT	0.000000	786.096000	47429.268000	47429.268000
ACCOUNTANT INTERN	24.430625	274.648333	29031.742917	29031.742917
ACPO,JuvP, Juv Prob (SFERS)	0.000000	0.000000	62290.780000	80266.370000
ACUPUNCTURIST	0.000000	1220.000000	67594.400000	67594.400000
...	...	...	...	...
X-RAY LABORATORY AIDE	3571.223462	1469.883846	52705.880385	52705.880385
X-Ray Laboratory Aide	3483.767100	1253.788500	50823.942700	69521.123200
YOUTH COMMISSION ADVISOR, BOARD OF SUPERVISORS	0.000000	1022.960000	53632.870000	53632.870000
Youth Comm Advisor	0.000000	2336.350000	41414.307500	60118.550000
ZOO CURATOR	0.000000	23538.560000	66686.560000	66686.560000

[2159 rows x 4 columns]

6.حدد أي ارتباط بين الرواتب وعمود رقمي أخر :

بناءً على الحقول المتاحة في مجموعة البيانات ، يمكننا تحديد الارتباط بين basepay و overtimepay وسبب اختيار هو أنه يمكن أن يكون لديه تأثير كبير على الراتب الإجمالي للموظفين .


Scatter Plot of BasePay vs OvertimePay
Overtime Pay
BasePay
250000
200000
150000
100000
50000
0
0
50000
100000
150000
200000
250000
300000