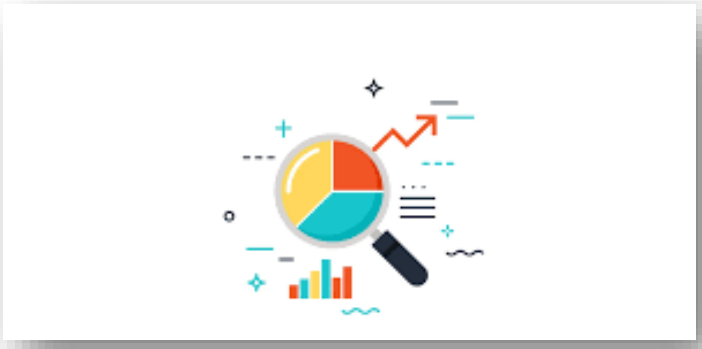
نادي شاي للذكاء الاصطناعي

# Task

Safa Abou Zaid



```
import pandas as pd
import numpy as np
df= pd.read_csv('Salaries.csv')
df.head()
```

أولاً: قمت باستيراد المكتبات اللازمة ، ثم قراءة البيانات من ملف csv حيث قمت بتضمينها في نفس المجلد لذلك كتبت الاسم وليس المسار ، بعدها طبعنا ال head بحيث نأخذ فكرة عن ال data .

```
print("the number of rows and columns in the dataset:")
df.shape
```

ثانياً : من خلال df.shape نحصل على عدد الاعمدة والاسطر في ال data set ،فكان عدد الأسطر 148654 وعدد الأعمدة هي 13

```
missing_values = df.isnull().sum()
print("Missing values in each column:")
print(missing_values)
```

ثالثا:

هنا قمنا بحساب القيم المفقودة من كل عمود من خلال التابع isnull() الذي يقوم بفحص ما اذا كانت القيم في كل خلية من البيانات مفقودة ام لا ، ثم يقوم التابع sum() بجمع عدد القيم المفقودة في كل عمود . وكانت النتيجة:

```
Missing values in each column:
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
```

نلاحظ أن العمود BasePay يحتوي على 609 قيمة مفقودة

والعمود OvertimePay و OtherPay كل منهم يحوي 4 قيم مفقودة

والعمود Benefits يحتوي 36163 قيمة مفقودة

أما العمودان Notes و Status فكل القيم مفقودة وبالتالي سنحتاج لحذفهم لاحقاً.

```
df.describe()
```

رابعاً: هذا الأمر يقوم بإعطاء ملخص احصائي عن البيانات يتضمن :

Count: العدد الإجمالي للقيم في كل عمود

Mean:المتوسط للقيم في كل عمود

std:الانحراف المعياري للقيم في كل عمود

Min:القيمة الأقل في كل عمود

25% ،50%،75% للقيم في كل عمود

Max:القيمة القصوى ي كل عمود

خامساً: نقوم بحذف عمود Notes,Statusلان جميع القيم فارغة لن تفيدنا هذه الاعمدة في تحليل البيانات .

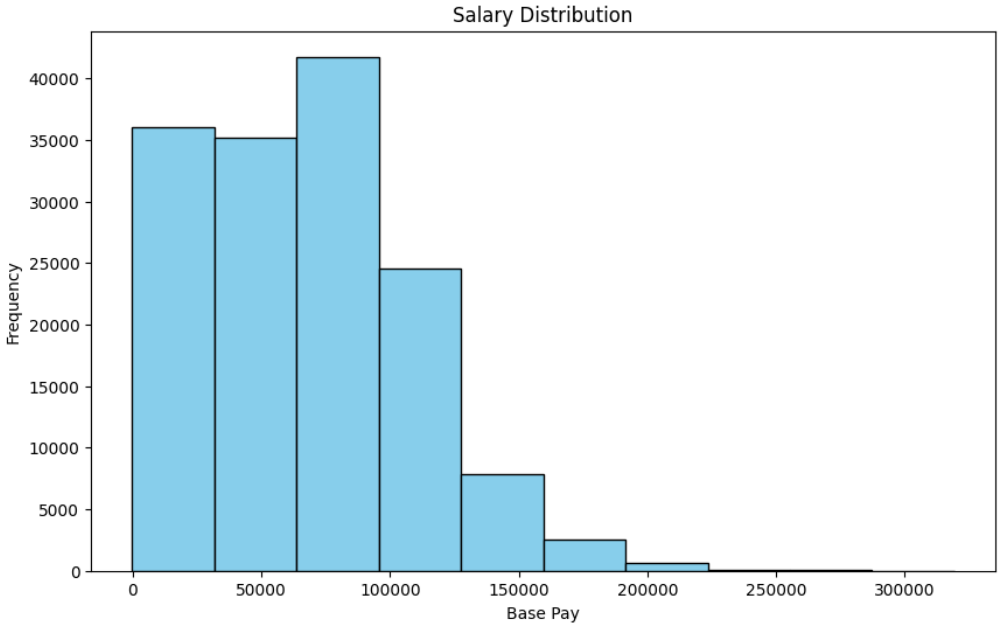
سادساً: نقوم بملأ القيم الفارغة بمتوسط قيم الحقل باستخدام fillnaهذا يساعد في الحفاظ على دقة البيانات وتجنب تأثير القيم الفارغة على التحليلات والاحصائيات التي نقوم بها .

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

[21]  ✓ 0.0s

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

سابعاً: رسم مخطط histogramتوزيع الرواتب باستخدام histogram



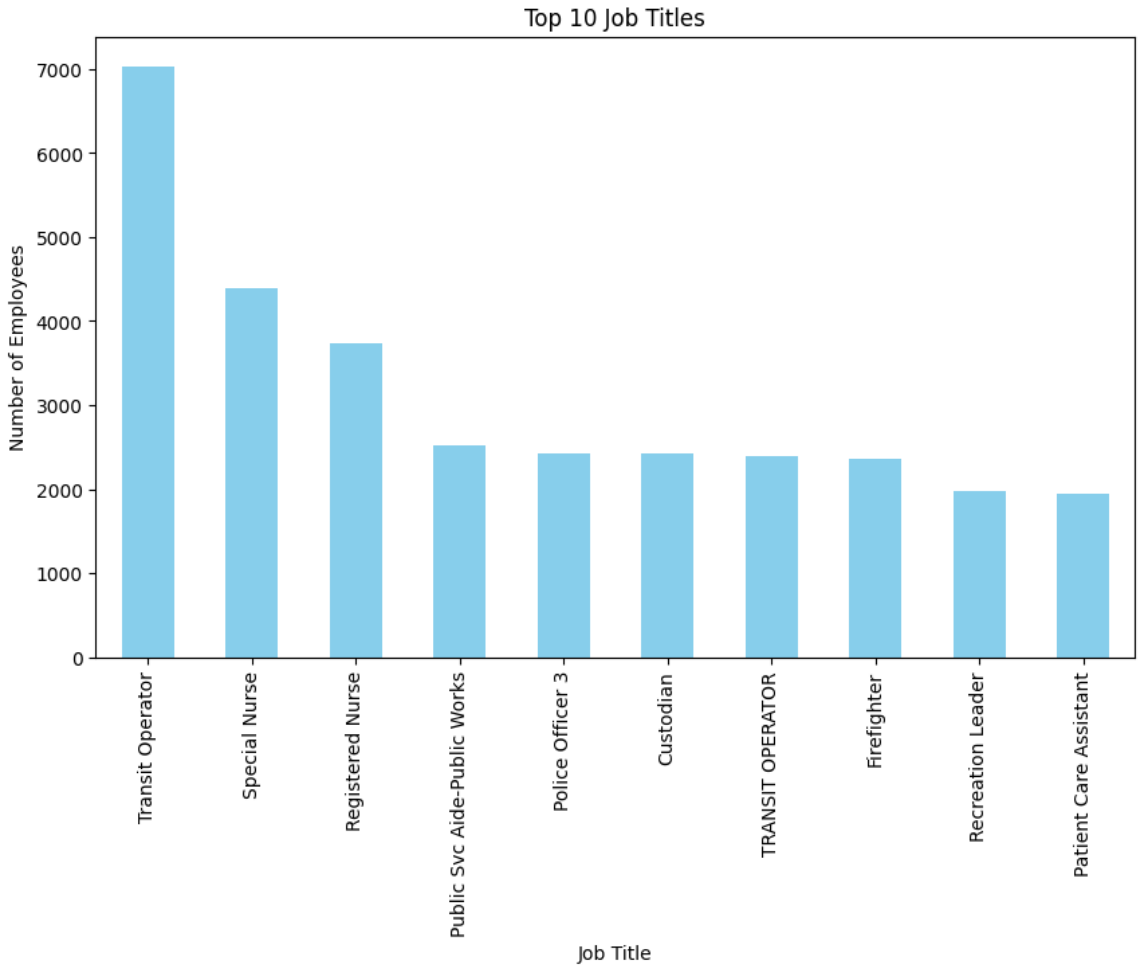
-نلاحظ أن 40000 موظفين رواتبهم هي 100000

وان 35000 موظف تقريبا رواتبهم بين 0 و 50000

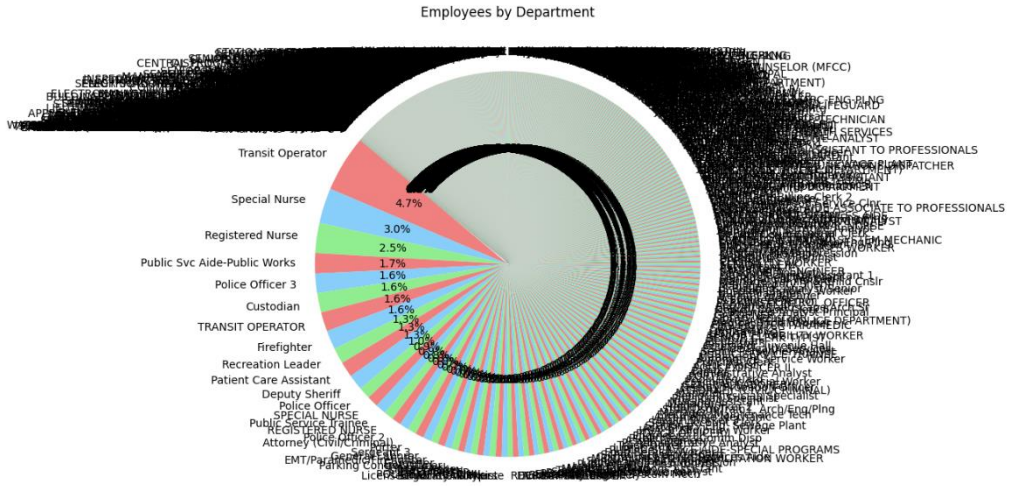
_25000 موظف تقريباً تتراوح رواتبهم بين 90000و160000

ثامناً:رسم مخططbar chart لتمثيل نسبة الموظفين حسب الفئة الوظيفية

نلاحظ أن أكثر وظيفة شيوعا هيtransit operator بعدد موظفين 7000




تاسعاً: رسم مخطط البيتشارت لتمثيل الموظفين في الأقسام المختلفة




عاشراً: تجميع البيانات حسب عمود واحد أو أكثر واحتساب اللإحصائيات القصوى لكل مجموعة وقارن المتوسطات الراتبية عبر المجموعات المختلفة.

النتائج التي تم طباعتها تمثل مجموعة البيانات المجمعة والتي تم تجميعها حسب العناوين الوظيفية (JobTitle) والسنة (Year).

تم حساب الإحصائيات المختلفة مثل المتوسط (mean) لكل من الرواتب الأساسية (BasePay)، رواتب العمل الإضافي (OvertimePay)، ورواتب أخرى (OtherPay)، والفوائد (Benefits)، والرواتب الإجمالية (TotalPay)، والرواتب الإجمالية بالمزايا (TotalPayBenefits).

على سبيل المثال، لوظيفة "ACCOUNT CLERK" في العام 2011، تم حساب المتوسط للرواتب الأساسية وهو 43300.806506، والمتوسط لرواتب العمل الإضافي وهو 373.200843، كذلك الأمر لوظيفة "ACCOUNT CLERK" في العام 2011، تمثل متوسط رواتب العمل الإضافي 373.200843. وبالمثل، يتم عرض متوسط رواتب العمل الإضافي لبقية الوظائف والسنوات.

هذه البيانات تساعد في فهم متوسط الرواتب والمزايا لكل وظيفة عبر السنوات المختلفة.

أخيراً: ارتباط بين عمود totalpay ,otherpay