PDA_3_1

SEABORN¶

In [185]:

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [55]:

```
mydata={'names':['RAM','SAAM','RAJ','VILLAN'],
       'AGE':[22,23,22,23],
       'salary':[20000,25000,21000,22000],
       'exc':[2,1,2,2],
      }
```

In [85]:

```
df=pd.DataFrame(mydata)
df.head()
```

Out[85]:

	names	AGE	salary	exc
0	RAM	22	20000	2
1	SAAM	23	25000	1
2	RAJ	22	21000	2
3	VILLAN	23	22000	2

HISTOGRAM¶

In [ ]:

In [61]:

```
plt.figure(figsize=(6,5))
sns.histplot(df["salary"],kde=True,bins=2)
plt.title("DISTRIBUTION OF SALARY")
plt.show()
```

```
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: 
FutureWarning: use_inf_as_na option is deprecated and will be removed 
in a future version. Convert inf values to NaN before operating 
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

No description has been provided for this image

1.postive skew,large salary value 2.no outler detected 3.Averge salary is about 21000 4.Majority salary are between 2000 and 22500

CORELATION MATRIX(HEAT MAP)¶

In [64]:

```
ndf=df.select_dtypes(include=["number"])
ndf.head()
```

Out[64]:

	AGE	salary	exc
0	22	20000	2
1	23	25000	1
2	22	21000	2
3	23	22000	2

In [66]:

```
plt.figure(figsize=(6,5))
sns.heatmap(ndf.corr(),cmap='plasma',annot=True)
plt.title("corelation between age,exp,sal")
plt.show()
```

No description has been provided for this image

In [ ]:

box plot¶

In [102]:

```
plt.figure(figsize=(6, 8))
sns.boxplot(x=df["AGE"])
plt.title('Age Distribution')
plt.show()
```

No description has been provided for this image

1.THE average age is 22.5 2.the abnormal value is around 23

In [105]:

```
temp=[21,47,39,22,31,33,29,26,27,25,49,46]
```

In [111]:

```
df=pd.DataFrame(temp)
df.head()
```

Out[111]:

	0
0	21
1	47
2	39
3	22
4	31

In [121]:

```
mydata1={'names':['RAM','SAAM','RAJ','VILLAN'],
       'AGE':[22,23,22,47],
       'salary':[20000,25000,21000,42000],
       'exp':[2,1,2,15],
        'g':['M','F','M','F'],
      }
df1=pd.DataFrame(mydata1)
```

In [123]:

```
plt.figure(figsize=(6,5))
sns.countplot(x=df1['exp'],palette='pastel',hue=df1['g'])
plt.title("count experience")
plt.show()
```

No description has been provided for this image

PAIR PLOT¶

In [127]:

```
sns.pairplot(df1)
```

```
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
```

```
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

Out[127]:

```
<seaborn.axisgrid.PairGrid at 0x265caaefb90>
```

No description has been provided for this image

IMPORTING LIBRARIES

In [ ]:

LOADING AND VERIFIYING DATA

In [234]:

```
sdf=pd.read_csv(r"C:\Users\DELL\Downloads\Salary_EDA.csv")
sdf
```

Out[234]:

	Age	Gender	Education Level	Job Title	Years of Experience	Salary
0	32.0	Male	Bachelor's	Software Engineer	5.0	90000.0
1	28.0	Female	Master's	Data Analyst	3.0	65000.0
2	45.0	Male	PhD	Senior Manager	15.0	150000.0
3	36.0	Female	Bachelor's	Sales Associate	7.0	60000.0
4	36.0	Female	Bachelor	Sales	7.0	60000.0

	Age	Gender	Education Level	Job Title	Years of Experience	Salary
			's	Associate		
...	...	...	...	...	...	...
370	35.0	Female	Bachelor's	Senior Marketing Analyst	8.0	85000.0
371	43.0	Male	Master's	Director of Operations	19.0	170000.0
372	29.0	Female	Bachelor's	Junior Project Manager	2.0	40000.0
373	34.0	Male	Bachelor's	Senior Operations Coordinator	7.0	90000.0
374	44.0	Female	PhD	Senior Business Analyst	15.0	150000.0

375 rows × 6 columns

In [203]:

```
sdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 375 entries, 0 to 374
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Age                  373 non-null    float64
 1   Gender               371 non-null    object 
 2   Education Level      372 non-null    object 
 3   Job Title            370 non-null    object 
 4   Years of Experience  373 non-null    float64
 5   Salary               372 non-null    float64
dtypes: float64(3), object(3)
memory usage: 17.7+ KB
```

HANDLING NULL VALUES

In [206]:

```
sdf.isnull().sum()
```

Out[206]:

```
Age                    2
Gender                 4
Education Level        3
Job Title              5
Years of Experience    2
Salary                 3
dtype: int64
```

In [208]:

```
sdf.dropna(inplace=True)
sdf.isnull().sum()
```

Out[208]:

```
Age                    0
Gender                 0
Education Level        0
Job Title              0
Years of Experience    0
Salary                 0
dtype: int64
```

conclusion: All null values are dropped,now features have non null

In [211]:

```
sdf.describe()
```

Out[211]:

	Age	Years of Experience	Salary
count	366.000000	366.000000	366.000000
mean	37.459016	10.045082	100492.759563
std	6.962303	6.517102	48013.732434
min	23.000000	0.000000	350.000000
25%	32.000000	4.000000	56250.000000
50%	36.000000	9.000000	95000.000000
75%	44.000000	15.000000	140000.000000

	Age	Years of Experience	Salary
max	53.000000	25.000000	250000.000000

In [213]:

```
sdf.describe(include='all')
```

Out[213]:

	Age	Gender	Education Level	Job Title	Years of Experience	Salary
count	366.000000	366	366	366	366.000000	366.000000
unique	NaN	2	3	169	NaN	NaN
top	NaN	Male	Bachelor's	Director of Marketing	NaN	NaN
freq	NaN	189	220	12	NaN	NaN
mean	37.459016	NaN	NaN	NaN	10.045082	100492.759563
std	6.962303	NaN	NaN	NaN	6.517102	48013.732434
min	23.000000	NaN	NaN	NaN	0.000000	350.000000
25%	32.000000	NaN	NaN	NaN	4.000000	56250.000000
50%	36.000000	NaN	NaN	NaN	9.000000	95000.000000
75%	44.000000	NaN	NaN	NaN	15.000000	140000.000000
max	53.000000	NaN	NaN	NaN	25.000000	250000.000000

conclusion¶

1.AGE minimum age is 23,maximum age is 53 majority of age falls between 32 and 34 few entites from 50s 2.GENDER .there are two unique value male and female .amoung 366,189 entries are male and 177 entries are female,so we can say male is dominating 3.EDUCATION LEVEL .most of the data concentrates on bachelor's(dominating) 4.JOB TITLE .amoung 366 ,12 times director of marketing is requested.others are repeated less than 12 timwes which means no job title is dominating in the dataset 5.YEARS OF EXPERIENCE .minimum experience is 0 ,maximum experience is 25,average experience is also a 25 .majority of people have experience between 4 and 15 6.SALARY .Minimum salary is 350,maximum experience is 25000,avareage salary isn 11 .majority salary is between 56000 and 1 .their might be outliers,min:350,avg:1 .

VISULIZATION¶

In [236]:

```
plt.figure(figsize=(6, 5))
sns.histplot(sdf["Age"], kde=True, bins=20)
plt.title("DISTRIBUTION OF AGE")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```

```
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

No description has been provided for this image

analyze salary usig bosplot

In [238]:

```
plt.figure(figsize=(6, 8))
sns.boxplot(x=sdf["Salary"])
plt.title('salary Distribution')
plt.show()
```

No description has been provided for this image

In [240]:

```
plt.figure(figsize=(6,5))
sns.heatmap(sdf.corr(),cmap='plasma',annot=True)
plt.title("corelation between age,exp,sal")
plt.show()
```

```
----------------------------------------------------------------------
-----
ValueError                                Traceback (most recent call
last)
Cell In[240], line 2
      1 plt.figure(figsize=(6,5))
----> 2 sns.heatmap(sdf.corr(),cmap='plasma',annot=True)
      3 plt.title("corelation between age,exp,sal")
      4 plt.show()

File ~\anaconda3\Lib\site-packages\pandas\core\frame.py:10704, in
DataFrame.corr(self, method, min_periods, numeric_only)
  10702 cols = data.columns
  10703 idx = cols.copy()
> 10704 mat = data.to_numpy(dtype=float, na_value=np.nan, copy=False)
  10706 if method == "pearson":
  10707     correl = libalgos.nancorr(mat, minp=min_periods)

File ~\anaconda3\Lib\site-packages\pandas\core\frame.py:1889, in
DataFrame.to_numpy(self, dtype, copy, na_value)
   1887 if dtype is not None:
   1888     dtype = np.dtype(dtype)
-> 1889 result = self._mgr.as_array(dtype=dtype, copy=copy,
na_value=na_value)
   1890 if result.dtype is not dtype:
   1891     result = np.array(result, dtype=dtype, copy=False)

File ~\anaconda3\Lib\site-packages\pandas\core\internals\
managers.py:1656, in BlockManager.as_array(self, dtype, copy,
na_value)
   1654         arr.flags.writeable = False
   1655 else:
-> 1656     arr = self._interleave(dtype=dtype, na_value=na_value)
   1657     # The underlying data was copied within _interleave, so no
need
   1658     # to further copy if copy=True or setting na_value
   1660 if na_value is lib.no_default:

File ~\anaconda3\Lib\site-packages\pandas\core\internals\
managers.py:1715, in BlockManager._interleave(self, dtype, na_value)
   1713     else:
   1714         arr = blk.get_values(dtype)
-> 1715     result[rl.indexer] = arr
   1716     itemmask[rl.indexer] = 1
   1718 if not itemmask.all():

ValueError: could not convert string to float: 'Male'
```

```
<Figure size 600x500 with 0 Axes>
```

In [251]:

```
plt.figure(figsize=(6,5))
sns.countplot(x=sdf['Gender'],palette='pastel',hue=sdf['Gender'])
plt.title("GENDER COUNT")
plt.show()
```

No description has been provided for this image

In [257]:

```
plt.figure(figsize=(6,5))
sns.countplot(x=sdf['Education
Level'],palette='pastel',hue=sdf['Education Level'])
plt.title("Education Level COUNT")
plt.show()
```

No description has been provided for this image

In [271]:

```
sns.pairplot(sdf,hue="Education Level")
```

```
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

Out[271]:

```
<seaborn.axisgrid.PairGrid at 0x265ceb63810>
```

No description has been provided for this image

OBSERVATION¶

1.PEEK SALARY IS GIVEN TO BACHELOR STUDENTS 2.empolys of bachelors having more experience 3.salary is alos effected by years of experience

group education level and find average salary in every categories

In [281]:

```
g=sdf.groupby("Education Level")['Salary'].mean()
g
```

Out[281]:

```
Education Level
Bachelor's     74465.848214
Master's      129583.333333
PhD           157843.137255
Name: Salary, dtype: float64
```

filter the data set in which gender is female and education level is master send find the avg salary on that set

In [290]:

```
g=sdf[(sdf["Years of Experience"]>20)]
g.head()
```

Out[290]:

	Age	Gender	Education Level	Job Title	Years of Experience	Salary
19	51.0	Male	Bachelor's	Sales Director	22.0	180000.0
30	50.0	Male	Bachelor's	CEO	25.0	250000.0
39	49.0	Male	Bachelor's	Sales Executive	21.0	160000.0
50	51.0	Female	Bachelor's	Customer Service Manager	22.0	130000.0
60	51.0	Female	Master's	Director of Operations	23.0	170000.0

In [ ]: