FDSL / Assignment2_PartB.ipynb

Fizashaikh63 Add files via upload ec73b55 · 2 minutes ago

1112 lines (1112 loc) · 123 KB

Preview Code Blame Raw

In [52]:
```
import pandas as pd
data=pd.read_csv("employee_dataset.csv",header=0)
data
```

Out[52]:

	EmpID	Name	Age	Gender	Department	Salary	JoiningDate	Per
0	1	Employee_1	50	Female	Sales	90000.0	2015-01-01	
1	2	Employee_2	36	Male	Finance	62500.0	2015-01-02	
2	3	Employee_3	29	Male	Finance	39500.0	2015-01-03	
3	4	Employee_4	42	Male	Sales	35000.0	2015-01-04	
4	5	Employee_5	40	Male	Finance	41500.0	2015-01-05	
...	...	...	...	...	...	...	...	...
995	996	Employee_996	34	Female	HR	31000.0	2017-09-22	
996	997	Employee_997	51	Female	IT	56500.0	2017-09-23	
997	998	Employee_998	44	Male	Finance	98000.0	2017-09-24	
998	999	Employee_999	40	Female	Sales	64500.0	2017-09-25	
999	1000	Employee_1000	53	Female	Sales	86000.0	2017-09-26	

1000 rows × 9 columns

In [53]:
```
print(data.isnull().sum())
#print(data.info('Salary'))
```

```
EmpID                0
Name                 0
Age                  0
Gender               0
Department           0
Salary               5
JoiningDate          0
PerformanceScore   177
WorkHours           37
dtype: int64
```

In [49]:
```
data['Salary'] = pd.to_numeric(data['Salary'])
```

In [50]:
```
print(data.info('Salary'))
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   EmpID             1000 non-null   int64
 1   Name              1000 non-null   object
 2   Age               1000 non-null   int64
 3   Gender            1000 non-null   object
 4   Department        1000 non-null   object
```

```
 5   Salary           1000 non-null   float64
 6   JoiningDate      1000 non-null   object
 7   PerformanceScore 1000 non-null   float64
 8   WorkHours        1000 non-null   float64
dtypes: float64(3), int64(2), object(4)
memory usage: 70.4+ KB
None
```

In [44]:
```
#print(data["Salary"].fillna(0))
data["Salary"] = data["Salary"].fillna(data["Salary"].mean())
data["Salary"]
```

Out[44]:
```
0      90000.0
1      62500.0
2      39500.0
3      35000.0
4      41500.0
        ...
995    31000.0
996    56500.0
997    98000.0
998    64500.0
999    86000.0
Name: Salary, Length: 1000, dtype: float64
```

In [46]:
```
data = data.dropna(subset=["Salary"])
data
```

Out[46]:

	EmpID	Name	Age	Gender	Department	Salary	JoiningDate	Per
0	1	Employee_1	50	Female	Sales	90000.0	2015-01-01	
1	2	Employee_2	36	Male	Finance	62500.0	2015-01-02	
2	3	Employee_3	29	Male	Finance	39500.0	2015-01-03	
3	4	Employee_4	42	Male	Sales	35000.0	2015-01-04	
4	5	Employee_5	40	Male	Finance	41500.0	2015-01-05	
...	...	...	...	...	...	...	...	...
995	996	Employee_996	34	Female	HR	31000.0	2017-09-22	
996	997	Employee_997	51	Female	IT	56500.0	2017-09-23	
997	998	Employee_998	44	Male	Finance	98000.0	2017-09-24	
998	999	Employee_999	40	Female	Sales	64500.0	2017-09-25	
999	1000	Employee_1000	53	Female	Sales	86000.0	2017-09-26	

1000 rows × 9 columns

In [22]:
```
print(data.fillna(0, inplace=True))
```

```
None
```

In [25]:

In [25]:
```
data.duplicated().sum()
```

Out[25]: np.int64(0)

In [27]:
```
print(data.drop_duplicates(inplace=True))
```

```
None
```

In [54]:
```
data1 = pd.DataFrame(data)
print("Before:\n", data1)
```

```
Before:
      EmpID           Name  Age  Gender Department   Salary JoiningDate  \
0         1     Employee_1   50  Female      Sales  90000.0  2015-01-01
1         2     Employee_2   36    Male    Finance  62500.0  2015-01-02
2         3     Employee_3   29    Male    Finance  39500.0  2015-01-03
3         4     Employee_4   42    Male      Sales  35000.0  2015-01-04
4         5     Employee_5   40    Male    Finance  41500.0  2015-01-05
..      ...            ...  ...     ...        ...      ...         ...
995     996   Employee_996   34  Female         HR  31000.0  2017-09-22
996     997   Employee_997   51  Female         IT  56500.0  2017-09-23
997     998   Employee_998   44    Male    Finance  98000.0  2017-09-24
998     999   Employee_999   40  Female      Sales  64500.0  2017-09-25
999    1000  Employee_1000   53  Female      Sales  86000.0  2017-09-26

     PerformanceScore  WorkHours
0                 3.0       43.0
1                 2.0       54.0
2                 1.0       54.0
3                 4.0       37.0
4                 4.0       37.0
..                ...        ...
995               2.0       36.0
996               1.0       44.0
997               4.0       51.0
998               1.0       53.0
999               5.0       40.0

[1000 rows x 9 columns]
```

In [56]:
```
data2 = pd.DataFrame(data)
print(data2.groupby('Gender')['Salary'].mean())
```

```
Gender
Female    61910.10101
Male      63161.00000
Name: Salary, dtype: float64
```

In [57]:
```
data.groupby('Department')['Salary'].mean()
```

Out[57]:
```
Department
Finance      62483.173077
HR           64689.473684
IT           59035.502959
Marketing    63711.165049
Sales        62328.828829
Name: Salary, dtype: float64
```

In [58]:
```
data.groupby('Department')['Salary'].agg(['mean','max','min','count'])
```

Out[58]:

Department	mean	max	min	count
Finance	62483.173077	99500.0	25000.0	208
HR	64689.473684	100000.0	25000.0	190
IT	59035.502959	97500.0	25000.0	169
Marketing	63711.165049	100000.0	25000.0	206
Sales	62328.828829	100000.0	25500.0	222

In [60]:
```
data.groupby('Department')['PerformanceScore'].agg(['mean','max','min','cou
```

Out[60]:

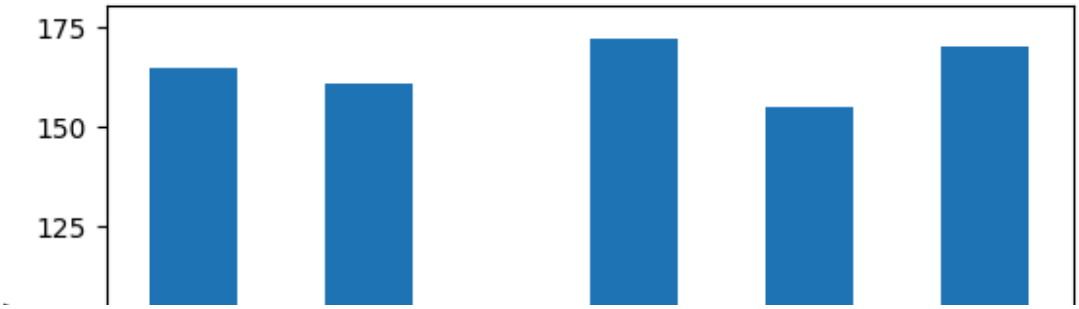
Department	mean	max	min	count
Finance	2.981928	5.0	1.0	166
HR	2.957055	5.0	1.0	163
IT	3.082192	5.0	1.0	146
Marketing	2.964706	5.0	1.0	170
Sales	3.044944	5.0	1.0	178

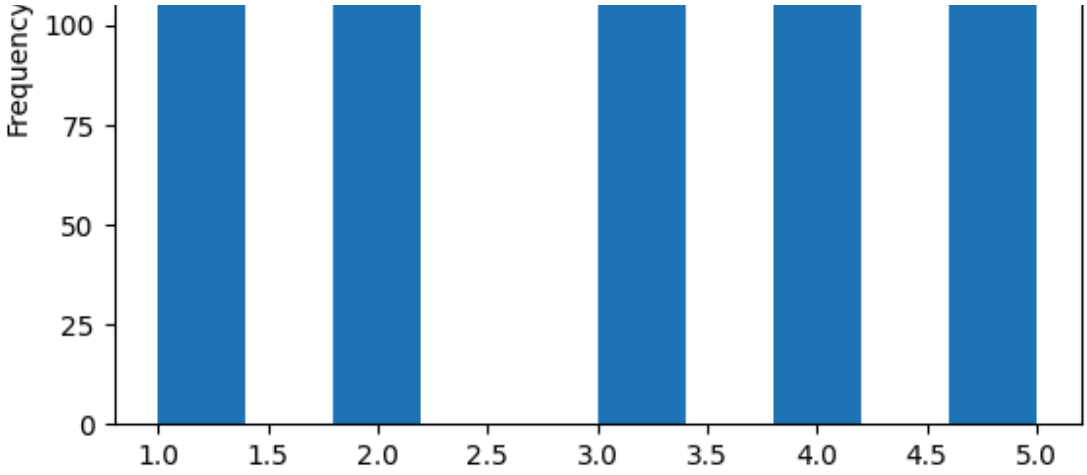
In [61]:
```
#group by multiple columns
data.groupby(['Department','Gender'])['Salary'].mean()
```

Out[61]:
```
Department  Gender
Finance     Female    62026.881720
            Male      62852.173913
HR          Female    62595.744681
            Male      66739.583333
IT          Female    58164.772727
            Male      59981.481481
Marketing   Female    67171.296296
            Male      59897.959184
Sales       Female    59107.142857
            Male      65609.090909
Name: Salary, dtype: float64
```

In [63]:
```
data['PerformanceScore'].plot(kind='hist')     # histogram
```

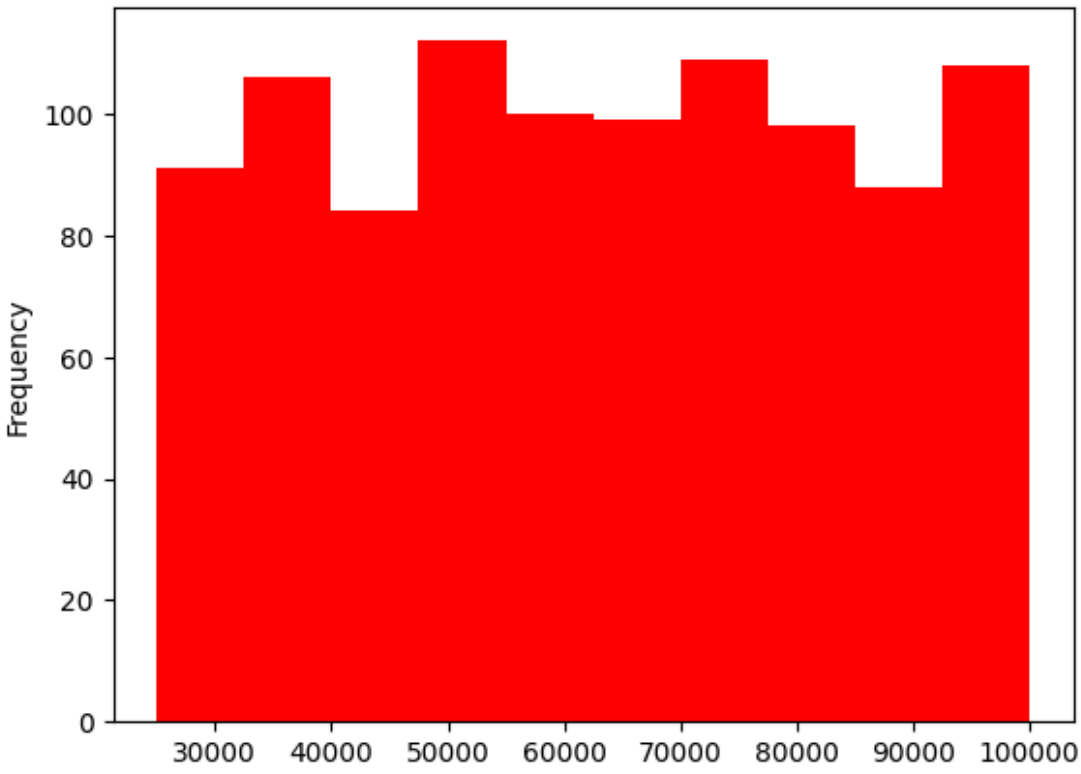
Out[63]:
```
<Axes: ylabel='Frequency'>
```





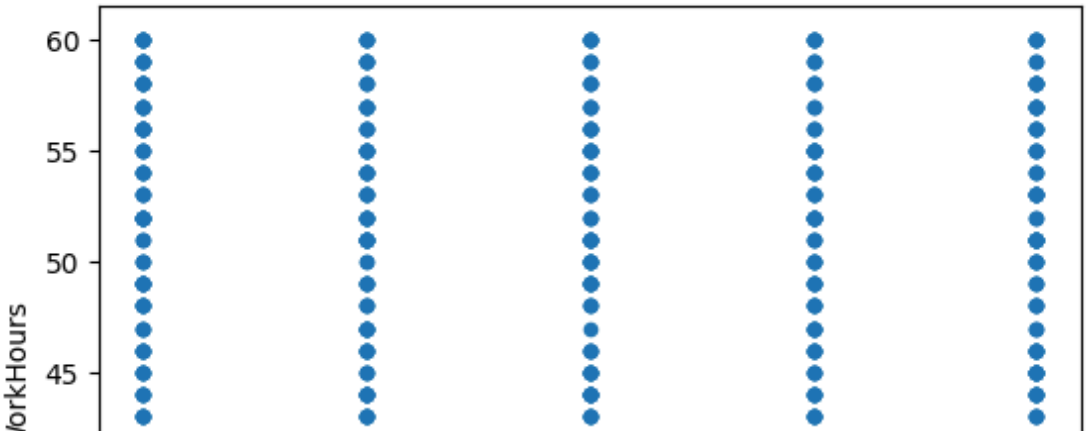
In [70]:
```
data['Salary'].plot(kind='hist',color='red')
```

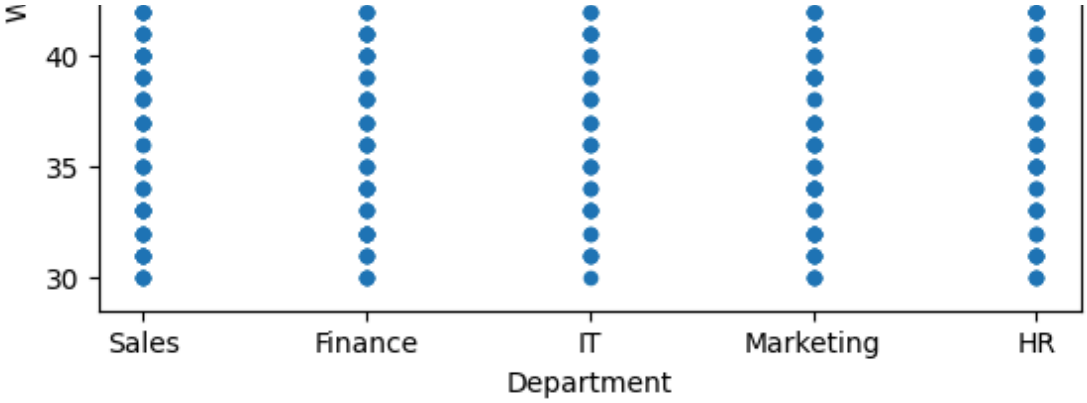
Out[70]: <Axes: ylabel='Frequency'>



In [65]:
```
data.plot(x='Department', y='WorkHours', kind='scatter')  # scatter plot
```

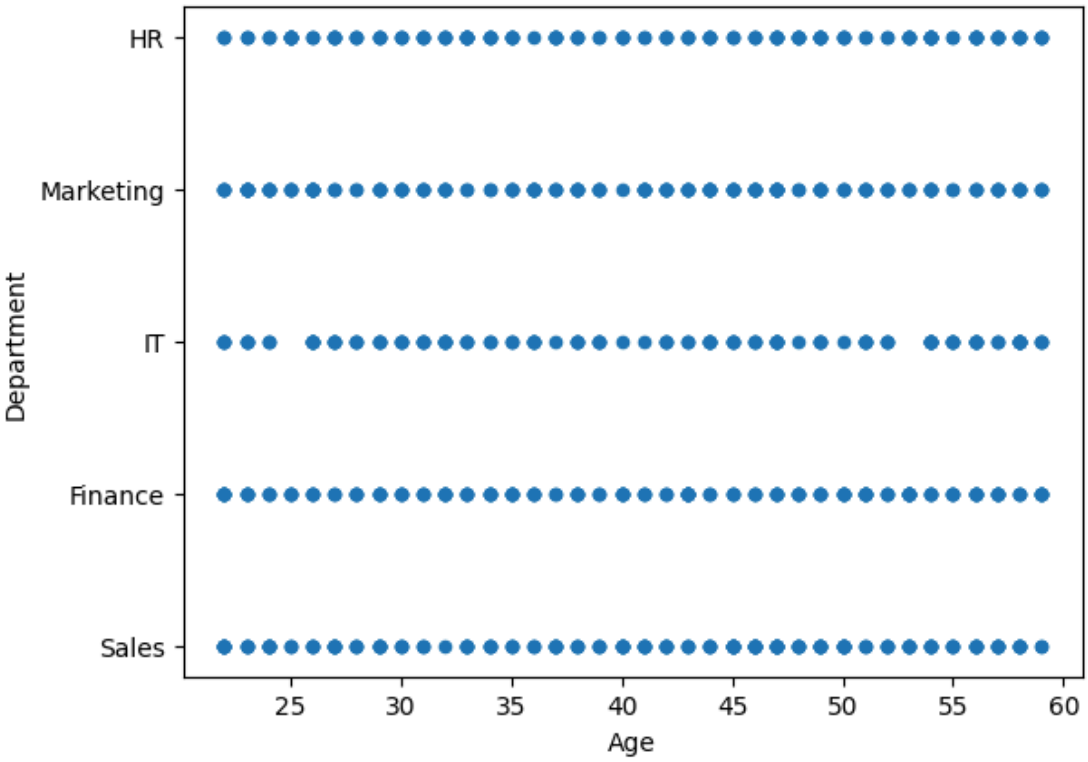
Out[65]: <Axes: xlabel='Department', ylabel='WorkHours'>





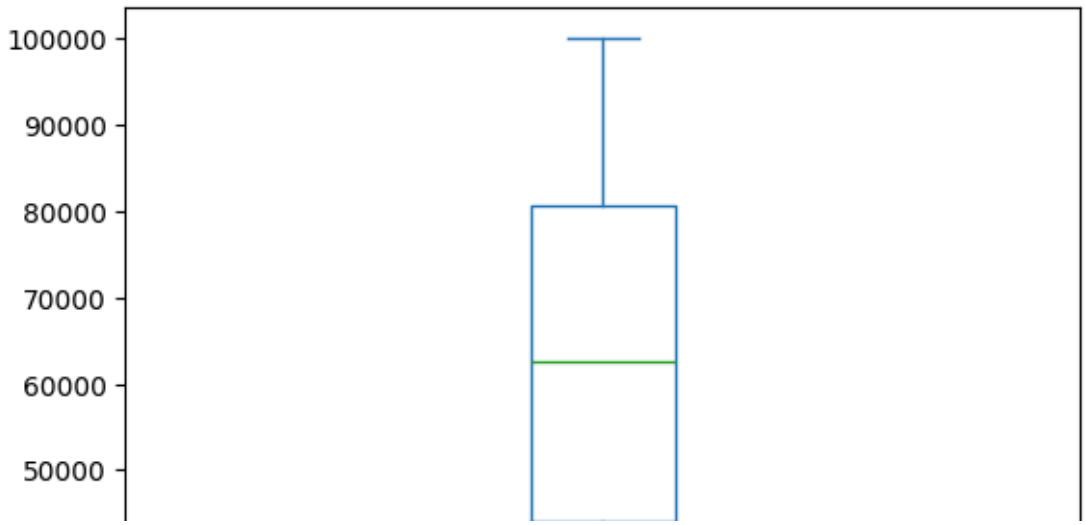
In [72]:
```
data.plot(x='Age', y='Department', kind='scatter')  # scatter plot
```

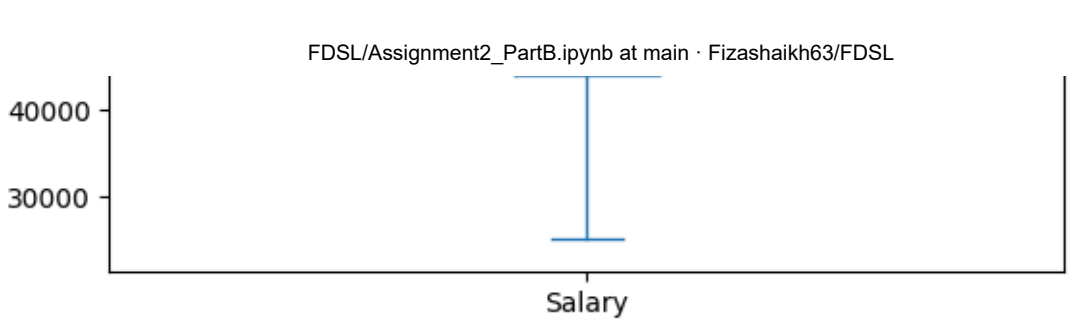
Out[72]: <Axes: xlabel='Age', ylabel='Department'>



In [67]:
```
data['Salary'].plot(kind='box')      # boxplot
```

Out[67]: <Axes: >





In [ ]: