EDA on HR Data

In [1]:
```
# importing packages

import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
# pulling the data set

df = pd.read_csv(r"C:\Users\bader\OneDrive\Desktop\HR data\HRDataset_v14.csv")
```

In [3]:
```
df.info
```

Out[3]:
```
<bound method DataFrame.info of                Employee_Name  EmpID  MarriedID  MaritalStatusID  GenderID  \
0          Adinolfi, Wilson  K  10026          0                0         1
1    Ait Sidi, Karthikeyan     10084          1                1         1
2           Akinkuolie, Sarah  10196          1                1         0
3                Alagbe,Trina  10088          1                1         0
4             Anderson, Carol   10069          0                2         0
..                         ...    ...        ...              ...       ...
306            Woodson, Jason  10135          0                0         1
307        Ybarra, Catherine   10301          0                0         0
308          Zamora, Jennifer  10010          0                0         0
309               Zhou, Julia  10043          0                0         0
310             Zima, Colleen  10271          0                4         0

     EmpStatusID  DeptID  PerfScoreID  FromDiversityJobFairID  Salary  ...  \
0              1       5            4                       0   62506  ...
1              5       3            3                       0  104437  ...
2              5       5            3                       0   64955  ...
3              1       5            3                       0   64991  ...
4              5       5            3                       0   50825  ...
..           ...     ...          ...                     ...     ...  ...
306            1       5            3                       0   65893  ...
307            5       5            1                       0   48513  ...
308            1       3            4                       0  220450  ...
309            1       3            3                       0   89292  ...
310            1       5            3                       0   45046  ...

         ManagerName  ManagerID  RecruitmentSource PerformanceScore  \
0    Michael Albert       22.0          LinkedIn          Exceeds
1        Simon Roup        4.0            Indeed      Fully Meets
2    Kissy Sullivan       20.0          LinkedIn      Fully Meets
3      Elijiah Gray       16.0            Indeed      Fully Meets
4    Webster Butler       39.0     Google Search      Fully Meets
..              ...        ...               ...              ...
306  Kissy Sullivan       20.0          LinkedIn      Fully Meets
307  Brannon Miller       12.0     Google Search              PIP
308      Janet King        2.0  Employee Referral          Exceeds
309      Simon Roup        4.0  Employee Referral      Fully Meets
310   David Stanley       14.0          LinkedIn      Fully Meets

     EngagementSurvey EmpSatisfaction SpecialProjectsCount  \
0                4.60              5                    0
1                4.96              3                    6
2                3.02              3                    0
3                4.84              5                    0
4                5.00              4                    0
..                ...            ...                  ...
306              4.07              4                    0
307              3.20              2                    0
308              4.60              5                    6
309              5.00              3                    5
310              4.50              5                    0

    LastPerformanceReview_Date DaysLateLast30 Absences
0                    1/17/2019             0        1
1                    2/24/2016             0       17
2                    5/15/2012             0        3
3                     1/3/2019             0       15
4                     2/1/2016             0        2
..                         ...           ...      ...
306                  2/28/2019             0       13
307                   9/2/2015             5        4
308                  2/21/2019             0       16
309                   2/1/2019             0       11
310                  1/30/2019             0        2

[311 rows x 36 columns]>
```

In [4]:
```
df.columns
```

Out[4]:
```
Index(['Employee_Name', 'EmpID', 'MarriedID', 'MaritalStatusID', 'GenderID',
       'EmpStatusID', 'DeptID', 'PerfScoreID', 'FromDiversityJobFairID',
       'Salary', 'Termd', 'PositionID', 'Position', 'State', 'Zip', 'DOB',
       'Sex', 'MaritalDesc', 'CitizenDesc', 'HispanicLatino', 'RaceDesc',
       'DateofHire', 'DateofTermination', 'TermReason', 'EmploymentStatus',
       'Department', 'ManagerName', 'ManagerID', 'RecruitmentSource',
       'PerformanceScore', 'EngagementSurvey', 'EmpSatisfaction',
       'SpecialProjectsCount', 'LastPerformanceReview_Date', 'DaysLateLast30',
       'Absences'],
      dtype='object')
```

In [5]:
```
df.head(10)
```

Out[5]:

Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	Salary	...	ManagerName	Manag

0	Adinolfi, Wilson K	10026	0	0	1	1	5	4	0	62506	...	Michael Albert
1	Ait Sidi, Karthikeyan	10084	1	1	1	5	3	3	0	104437	...	Simon Roup
2	Akinkuolie, Sarah	10196	1	1	0	5	5	3	0	64955	...	Kissy Sullivan
3	Alagbe,Trina	10088	1	1	0	1	5	3	0	64991	...	Elijiah Gray
4	Anderson, Carol	10069	0	2	0	5	5	3	0	50825	...	Webster Butler
5	Anderson, Linda	10002	0	0	0	1	5	4	0	57568	...	Amy Dunn
6	Andreola, Colby	10194	0	0	0	1	4	3	0	95660	...	Alex Sweetwater
7	Athwal, Sam	10062	0	4	1	1	5	3	0	59365	...	Ketsia Liebig
8	Bachiochi, Linda	10114	0	0	0	3	5	3	1	47837	...	Brannon Miller
9	Bacong, Alejandro	10250	0	2	1	1	3	3	0	50178	...	Peter Monroe

10 rows × 36 columns

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
Employee_Name                0
EmpID                        0
MarriedID                    0
MaritalStatusID              0
GenderID                     0
EmpStatusID                  0
DeptID                       0
PerfScoreID                  0
FromDiversityJobFairID       0
Salary                       0
Termd                        0
PositionID                   0
Position                     0
State                        0
Zip                          0
DOB                          0
Sex                          0
MaritalDesc                  0
CitizenDesc                  0
HispanicLatino               0
RaceDesc                     0
DateofHire                   0
DateofTermination          207
TermReason                   0
EmploymentStatus             0
Department                   0
ManagerName                  0
ManagerID                    8
RecruitmentSource            0
PerformanceScore             0
EngagementSurvey             0
EmpSatisfaction              0
SpecialProjectsCount         0
LastPerformanceReview_Date   0
DaysLateLast30               0
Absences                     0
dtype: int64
```

In [7]:
```
df.nunique()
```

Out[7]:
```
Employee_Name              311
EmpID                      311
MarriedID                    2
MaritalStatusID              5
GenderID                     2
EmpStatusID                  5
DeptID                       6
PerfScoreID                  4
FromDiversityJobFairID       2
Salary                     308
Termd                        2
PositionID                  30
Position                    32
State                       28
Zip                        158
DOB                        307
Sex                          2
MaritalDesc                  5
CitizenDesc                  3
HispanicLatino               4
RaceDesc                     6
DateofHire                 101
DateofTermination           96
TermReason                  18
EmploymentStatus             3
Department                   6
ManagerName                 21
ManagerID                   23
RecruitmentSource            9
PerformanceScore             4
EngagementSurvey           119
EmpSatisfaction              5
SpecialProjectsCount         9
LastPerformanceReview_Date 137
DaysLateLast30               7
Absences                    20
dtype: int64
```

In [18]:
```
df.describe()
```

Out[18]:

	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	Salary	Termd	Positio
count	311.000000	311.000000	311.000000	311.000000	311.000000	311.000000	311.000000	311.000000	311.000000	311.000000	311.000

mean	10156.000000	0.398714	0.810289	0.434084	2.392283	4.610932	2.977492	0.093248	69020.684887	0.334405	16.845
std	89.922189	0.490423	0.943239	0.496435	1.794383	1.083487	0.587072	0.291248	25156.636930	0.472542	6.223
min	10001.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	0.000000	45046.000000	0.000000	1.000
25%	10078.500000	0.000000	0.000000	0.000000	1.000000	5.000000	3.000000	0.000000	55501.500000	0.000000	18.000
50%	10156.000000	0.000000	1.000000	0.000000	1.000000	5.000000	3.000000	0.000000	62810.000000	0.000000	19.000
75%	10233.500000	1.000000	1.000000	1.000000	5.000000	5.000000	3.000000	0.000000	72036.000000	1.000000	20.000
max	10311.000000	1.000000	4.000000	1.000000	5.000000	6.000000	4.000000	1.000000	250000.000000	1.000000	30.000

In [19]:
```
df.corr()
```

```
C:\Users\bader\AppData\Local\Temp\ipykernel_13496\1134722465.py:1: FutureWarning: The default value of numeric_only in DataFram
e.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_
only to silence this warning.
  df.corr()
```

Out[19]:

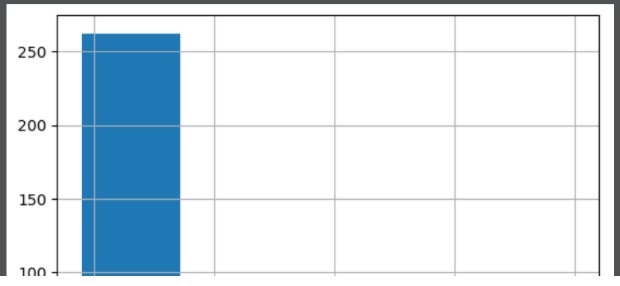
	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	Salary	Termd	Posi
EmpID	1.000000	0.048058	-0.043851	0.035914	0.073750	0.107406	-0.691348	0.046805	-0.115319	0.092389	-0.0
MarriedID	0.048058	1.000000	0.164044	-0.024199	0.085619	-0.119932	-0.058362	-0.012708	0.026165	0.077028	-0.0
MaritalStatusID	-0.043851	0.164044	1.000000	-0.030236	0.114630	0.012768	0.044693	0.041117	-0.070291	0.099367	0.0
GenderID	0.035914	-0.024199	-0.030236	1.000000	-0.032440	-0.038838	-0.054915	0.031493	0.056097	-0.015741	-0.0
EmpStatusID	0.073750	0.085619	0.114630	-0.032440	1.000000	0.088711	-0.071208	0.189025	-0.110912	0.948058	0.2
DeptID	0.107406	-0.119932	0.012768	-0.038838	0.088711	1.000000	-0.084811	-0.129998	-0.448132	0.065922	0.0
PerfScoreID	-0.691348	-0.058362	0.044693	-0.054915	-0.071208	-0.084811	1.000000	0.012315	0.130903	-0.089061	0.0
FromDiversityJobFairID	0.046805	-0.012708	0.041117	0.031493	0.189025	-0.129998	0.012315	1.000000	0.041248	0.147717	0.0
Salary	-0.115319	0.026165	-0.070291	0.056097	-0.110912	-0.448132	0.130903	0.041248	1.000000	-0.093994	-0.1
Termd	0.092389	0.077028	0.099367	-0.015741	0.948058	0.065922	-0.089061	0.147717	-0.093994	1.000000	0.1
PositionID	-0.036488	-0.027334	0.021923	-0.081612	0.221221	0.030294	0.005227	0.015085	-0.130563	0.147042	1.0
Zip	0.026858	-0.041147	0.010620	0.048539	-0.150527	0.290023	-0.058350	-0.028314	-0.037242	-0.139006	-0.5
ManagerID	0.090236	-0.094002	0.023065	-0.043218	0.234222	0.550240	-0.060552	0.007570	-0.435406	0.209113	0.0
EngagementSurvey	-0.589664	-0.091178	0.033249	-0.036276	0.024305	-0.094940	0.544927	-0.013040	0.064966	-0.017543	0.0
EmpSatisfaction	-0.146967	-0.126191	0.002068	-0.044603	0.010553	0.031997	0.303579	-0.034468	0.062718	-0.004732	-0.0
SpecialProjectsCount	-0.043730	0.061278	-0.051093	0.087073	-0.166560	-0.785101	0.045677	0.031393	0.508333	-0.147429	-0.1
DaysLateLast30	0.495513	0.002875	-0.096500	0.080329	0.078318	0.124630	-0.734728	0.042532	-0.069443	0.136379	-0.0
Absences	-0.025278	0.096086	0.018722	-0.004577	0.091834	0.053308	0.046629	0.062640	0.082382	0.098274	-0.0

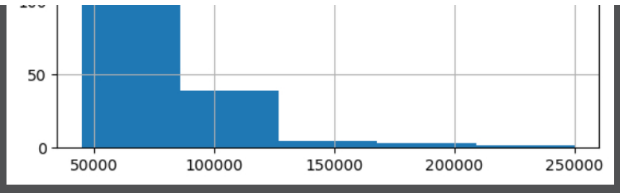
In [8]:
```
# value counts for all columns


columns = ['Employee_Name', 'EmpID', 'MarriedID', 'MaritalStatusID', 'GenderID',
       'EmpStatusID', 'DeptID', 'PerfScoreID', 'FromDiversityJobFairID',
       'Salary', 'Termd', 'PositionID', 'Position', 'State', 'Zip', 'DOB',
       'Sex', 'MaritalDesc', 'CitizenDesc', 'HispanicLatino', 'RaceDesc',
       'DateofHire', 'DateofTermination', 'TermReason', 'EmploymentStatus',
       'Department', 'ManagerName', 'ManagerID', 'RecruitmentSource',
       'PerformanceScore', 'EngagementSurvey', 'EmpSatisfaction',
       'SpecialProjectsCount', 'LastPerformanceReview_Date', 'DaysLateLast30',
       'Absences']

for i in columns:
    print(df[i].value_counts())
```

```
Yes     27
no       1
yes      1
Name: HispanicLatino, dtype: int64
White                                187
Black or African American             80
Asian                                 29
Two or more races                     11
American Indian or Alaska Native       3
Hispanic                               1
Name: RaceDesc, dtype: int64
1/10/2011    14
3/30/2015    12
1/5/2015     11
9/29/2014    11
7/5/2011     10
             ..
3/7/2011      1
7/9/2012      1
1/5/2016      1
```

In [9]:
```
# salaries distribution

df['Salary'].hist(bins = 5)
```

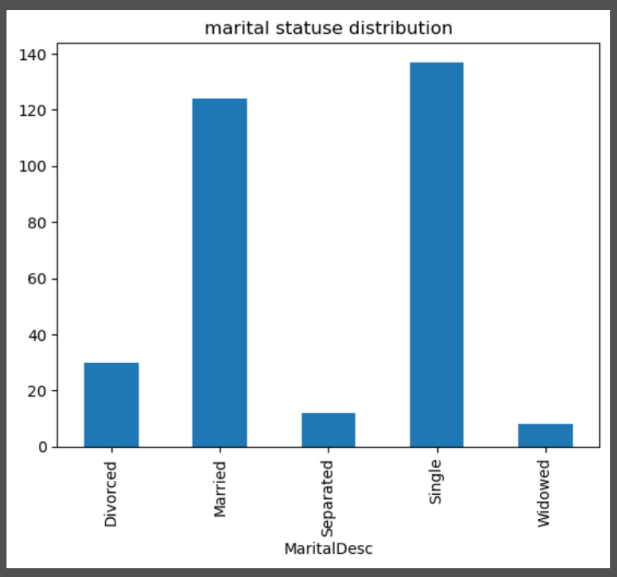
Out[9]: <Axes: >





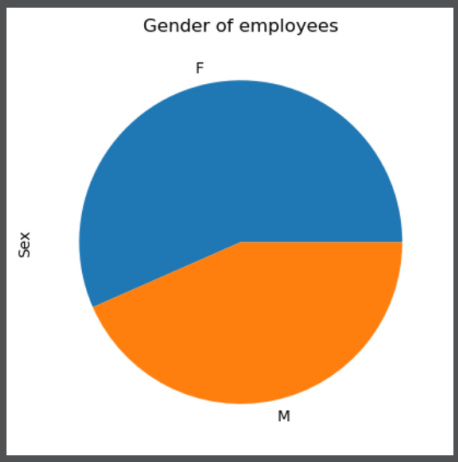
In [10]:
```
# distribution of employees across different marital statuses

barc = df.groupby('MaritalDesc')['MaritalDesc'].count()

barc.plot(kind = 'bar', title = 'marital statuse distribution')
```

Out[10]: <Axes: title={'center': 'marital statuse distribution'}, xlabel='MaritalDesc'>



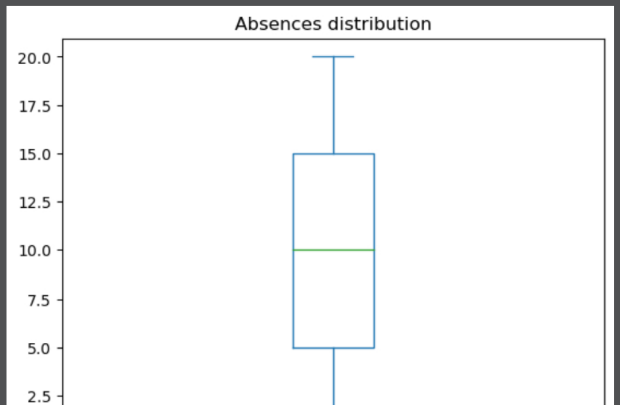
In [12]:
```
# gender distribution of employees

genderPlot = df.groupby('Sex')['Sex'].count()

genderPlot.plot(kind = 'pie', title = 'Gender of employees')
```

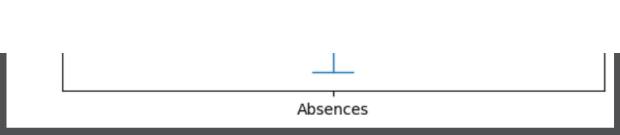
Out[12]: <Axes: title={'center': 'Gender of employees'}, ylabel='Sex'>



In [13]:
```
df['Absences'].plot(kind = 'box', title = 'Absences distribution')
```

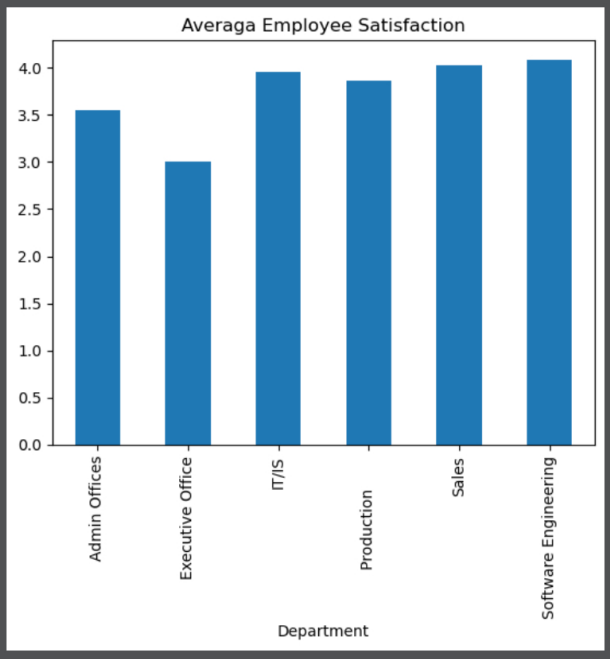
Out[13]: <Axes: title={'center': 'Absences distribution'}>





In [16]:
```
# averaga Employee satisfaction by department

avgSat = df.groupby('Department')['EmpSatisfaction'].mean()

avgSat.plot(kind = 'bar', title = 'Averaga Employee Satisfaction')
```

Out[16]:
```
<Axes: title={'center': 'Averaga Employee Satisfaction'}, xlabel='Department'>
```



In [17]:
```
# Performances by recruitment source

perBySource = df.groupby(['RecruitmentSource','PerformanceScore'])['PerformanceScore'].count()

perBySource.plot.barh(title = 'Performances by Recruitment Source')
```

Out[17]:
```
<Axes: title={'center': 'Performances by Recruitment Source'}, ylabel='RecruitmentSource,PerformanceScore'>
```