```
#Amir Tawfiq and Hanya Zamir

#Description:This data provides information the Human Resources
attrition and performances in a fictional office place
#while also talking about how each genders faces attrition, and how it
may affect different departments inside the office
#in terms of job involvement, job satisfaction, etc. We were able to
find this dataset on
#Kaggle in
"https://www.kaggle.com/patelprashant/employee-attrition/code"

#Abstract:In this notebook we will be analyzing this data, we were
able to extract information on how marital satus affects
#job involvment, and how different departments go through attrition.
We were also able to see which departments were bigger
#in numbers than others, which departments travel the most, if
distance from home affected the monthly income the rate of
#environment satisfaction in this workplace and so much more.
```

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

df=pd.read_csv("Assignment 1.csv")
df
```

```
      Age Attrition     BusinessTravel  DailyRate
Department  \
0      41       Yes      Travel_Rarely       1102
Sales
1      49        No  Travel_Frequently        279  Research &
Development
2      37       Yes      Travel_Rarely       1373  Research &
Development
3      33        No  Travel_Frequently       1392  Research &
Development
4      27        No      Travel_Rarely        591  Research &
Development
...   ...       ...                ...        ...
...
1465   36        No  Travel_Frequently        884  Research &
Development
1466   39        No      Travel_Rarely        613  Research &
Development
1467   27        No      Travel_Rarely        155  Research &
Development
1468   49        No  Travel_Frequently       1023
Sales
1469   34        No      Travel_Rarely        628  Research &
Development
```

```
      DistanceFromHome  Education EducationField  EmployeeCount  \
0                    1          2  Life Sciences              1
1                    8          1  Life Sciences              1
2                    2          2          Other              1
3                    3          4  Life Sciences              1
4                    2          1        Medical              1
...                ...        ...            ...            ...
1465                23          2        Medical              1
1466                 6          1        Medical              1
1467                 4          3  Life Sciences              1
1468                 2          3        Medical              1
1469                 8          3        Medical              1

      EmployeeNumber  ...  RelationshipSatisfaction StandardHours  \
0                  1  ...                         1            80
1                  2  ...                         4            80
2                  4  ...                         2            80
3                  5  ...                         3            80
4                  7  ...                         4            80
...              ...  ...                       ...           ...
1465            2061  ...                         3            80
1466            2062  ...                         1            80
1467            2064  ...                         2            80
1468            2065  ...                         4            80
1469            2068  ...                         1            80

      StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
0                    0                  8                      0
1                    1                 10                      3
2                    0                  7                      3
3                    0                  8                      3
4                    1                  6                      3
...                ...                ...                    ...
1465                 1                 17                      3
1466                 1                  9                      5
1467                 1                  6                      0
1468                 0                 17                      3
1469                 0                  6                      3

     WorkLifeBalance  YearsAtCompany YearsInCurrentRole  \
0                  1               6                  4
1                  3              10                  7
2                  3               0                  0
3                  3               8                  7
4                  3               2                  2
...              ...             ...                ...
1465               3               5                  2
1466               3               7                  7
1467               3               6                  2
```

```
1468                2               9                   6
1469                4               4                   3

      YearsSinceLastPromotion  YearsWithCurrManager
0                           0                     5
1                           1                     7
2                           0                     0
3                           3                     0
4                           2                     2
...                       ...                   ...
1465                        0                     3
1466                        1                     7
1467                        0                     3
1468                        0                     8
1469                        1                     2

[1470 rows x 35 columns]
```

```
#Head and Tail
print(df.head())
print(df.tail())
```

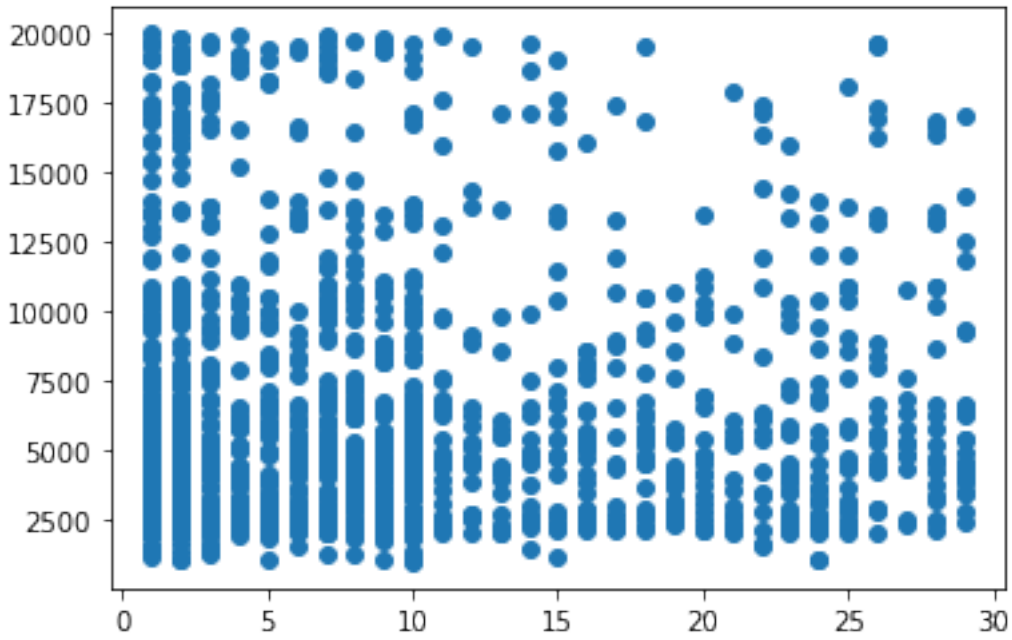
```
   Age Attrition     BusinessTravel  DailyRate              Department
\
0   41       Yes      Travel_Rarely       1102                   Sales

1   49        No  Travel_Frequently        279  Research & Development

2   37       Yes      Travel_Rarely       1373  Research & Development

3   33        No  Travel_Frequently       1392  Research & Development

4   27        No      Travel_Rarely        591  Research & Development


   DistanceFromHome  Education EducationField  EmployeeCount
EmployeeNumber  \
0                 1          2  Life Sciences              1
1
1                 8          1  Life Sciences              1
2
2                 2          2          Other              1
4
3                 3          4  Life Sciences              1
5
4                 2          1        Medical              1
7

   ...  RelationshipSatisfaction StandardHours  StockOptionLevel  \
0  ...                         1            80                 0
```

```
1  ...                        4            80                 1
2  ...                        2            80                 0
3  ...                        3            80                 0
4  ...                        4            80                 1

   TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance
YearsAtCompany  \
0                  8                      0               1
6
1                 10                      3               3
10
2                  7                      3               3
0
3                  8                      3               3
8
4                  6                      3               3
2

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                  4                       0                     5
1                  7                       1                     7
2                  0                       0                     0
3                  7                       3                     0
4                  2                       2                     2

[5 rows x 35 columns]
      Age Attrition     BusinessTravel  DailyRate
Department  \
1465   36        No  Travel_Frequently        884  Research &
Development
1466   39        No      Travel_Rarely        613  Research &
Development
1467   27        No      Travel_Rarely        155  Research &
Development
1468   49        No  Travel_Frequently       1023
Sales
1469   34        No      Travel_Rarely        628  Research &
Development

      DistanceFromHome  Education EducationField  EmployeeCount  \
1465                23          2        Medical              1
1466                 6          1        Medical              1
1467                 4          3  Life Sciences              1
1468                 2          3        Medical              1
1469                 8          3        Medical              1

      EmployeeNumber  ...  RelationshipSatisfaction StandardHours  \
1465            2061  ...                         3            80
1466            2062  ...                         1            80
1467            2064  ...                         2            80
```

```
1468            2065  ...                          4            80
1469            2068  ...                          1            80

      StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
1465                 1                 17                      3
1466                 1                  9                      5
1467                 1                  6                      0
1468                 0                 17                      3
1469                 0                  6                      3

     WorkLifeBalance  YearsAtCompany YearsInCurrentRole  \
1465               3               5                  2
1466               3               7                  7
1467               3               6                  2
1468               2               9                  6
1469               4               4                  3

      YearsSinceLastPromotion  YearsWithCurrManager
1465                        0                     3
1466                        1                     7
1467                        0                     3
1468                        0                     8
1469                        1                     2

[5 rows x 35 columns]
```

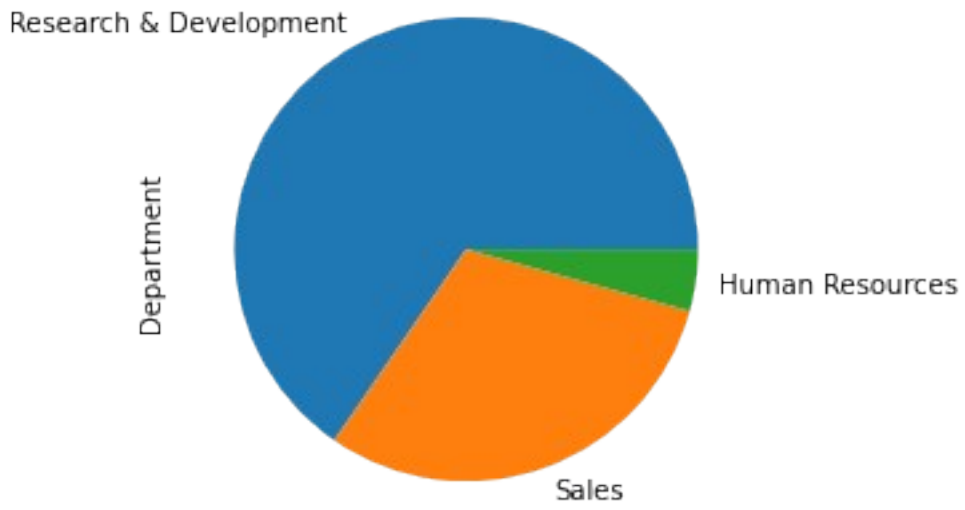
```
plt.scatter(df['DistanceFromHome'],df['MonthlyIncome'])
```

```
<matplotlib.collections.PathCollection at 0x159c0e37730>
```



```
df['Department'].value_counts().plot(kind='pie', title='')
```

```
<AxesSubplot:ylabel='Department'>
```

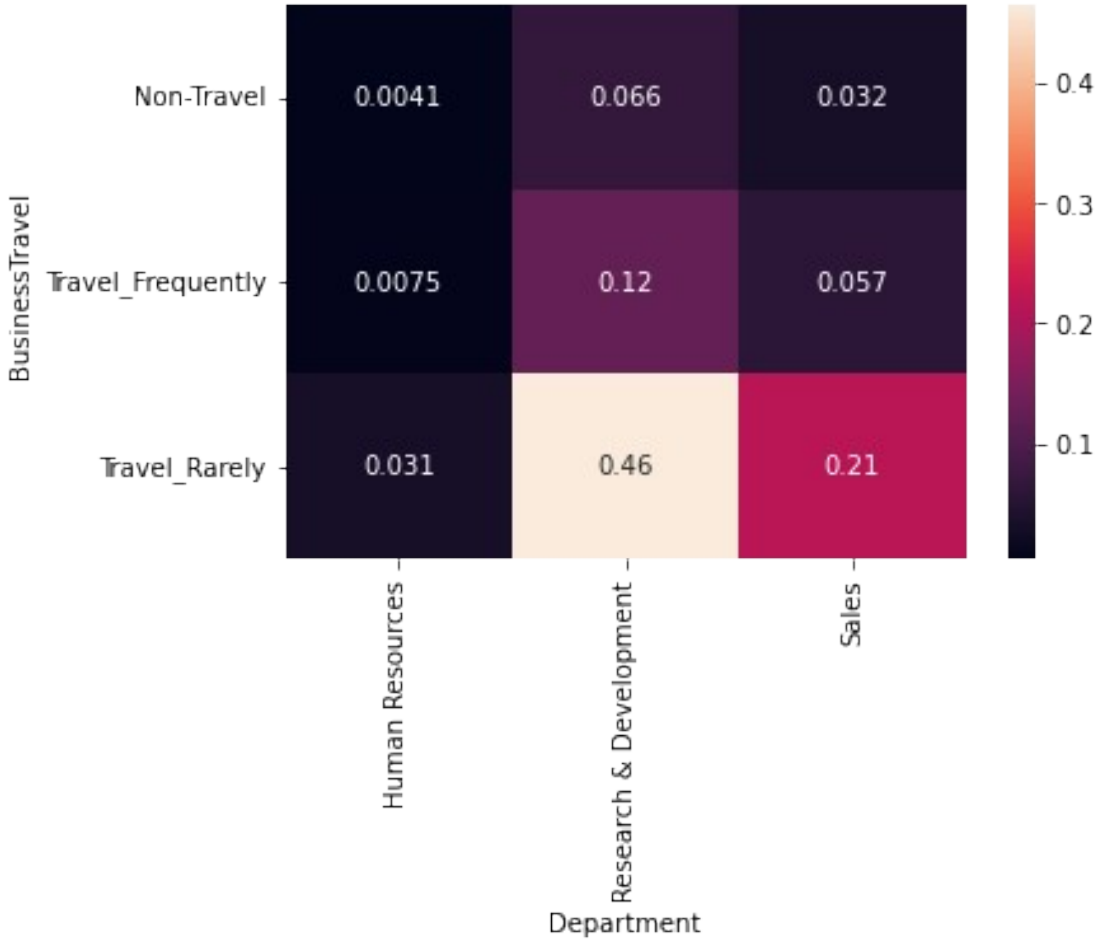



```
x=pd.crosstab(df['BusinessTravel'],df['Department'],normalize=True)
print(x)
```

```
Department         Human Resources  Research & Development     Sales
BusinessTravel
Non-Travel                0.004082                0.065986  0.031973
Travel_Frequently         0.007483                0.123810  0.057143
Travel_Rarely             0.031293                0.463946  0.214286
```

```
sns.heatmap(x, annot=True)
```

```
<AxesSubplot:xlabel='Department', ylabel='BusinessTravel'>
```




```
print(" In the heat map above, the research and development department has the most umbers in all 3 categories because it is the largest department so we need to measure them based on their percentages. ")
```

```
 In the heat map above, the research and development department has the most umbers in all 3 categories because it is the largest department so we need to measure them based on their percentages.
```

```
print(" After converting it to percentages, we can see that r which represents the research and development department has the most employees that travel frequently while human resources has the least.")
```

```
 After converting it to percentages, we can see that r which represents the research and development department has the most employees that travel frequently while human resources has the least.
```

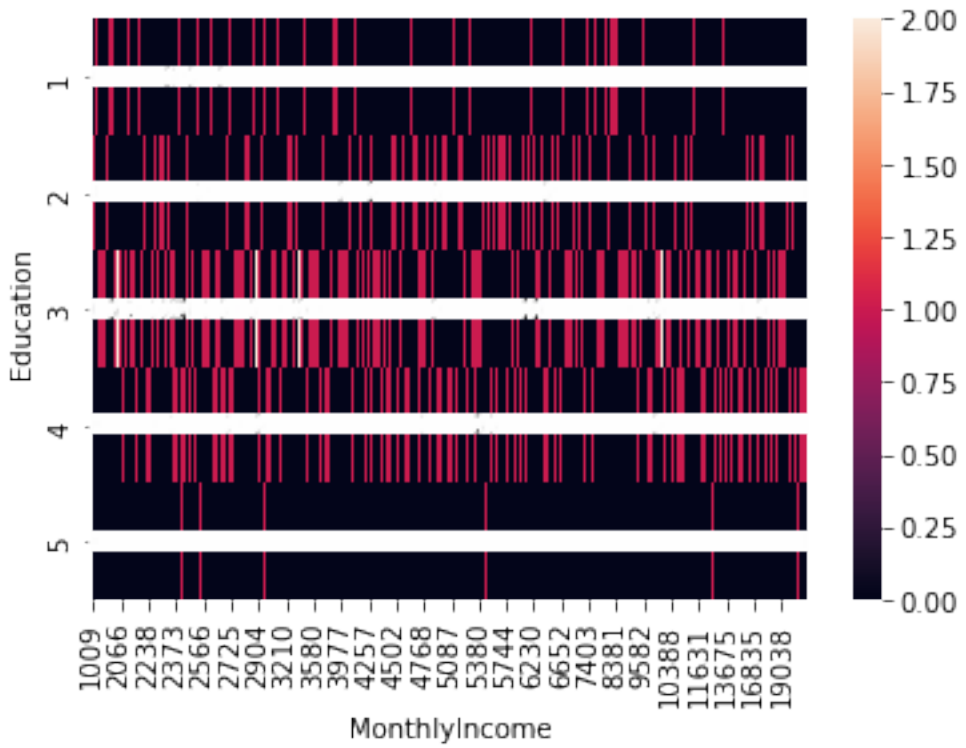
```
d=pd.crosstab(df['Education'],df['MonthlyIncome'])
print(d)
```

```
MonthlyIncome  1009   1051   1052   1081   1091   1102   1118   1129   1200   \
```

```
Education

1                   1      0      1      0      0      0      1      1
0
2                   0      1      0      0      0      0      0      0
0
3                   0      0      0      1      1      1      0      0
1
4                   0      0      0      0      0      0      0      0
0
5                   0      0      0      0      0      0      0      0
0

MonthlyIncome  1223   ...  19717  19740  19833  19845  19847  19859
19926  \
Education             ...

1                   0  ...      0      0      0      1      0      0
0
2                   1  ...      0      0      0      0      0      0
0
3                   0  ...      0      1      0      0      1      1
1
4                   0  ...      1      0      1      0      0      0
0
5                   0  ...      0      0      0      0      0      0
0

MonthlyIncome  19943  19973  19999
Education
1                   0      1      0
2                   0      0      0
3                   0      0      0
4                   1      0      1
5                   0      0      0

[5 rows x 1349 columns]
```

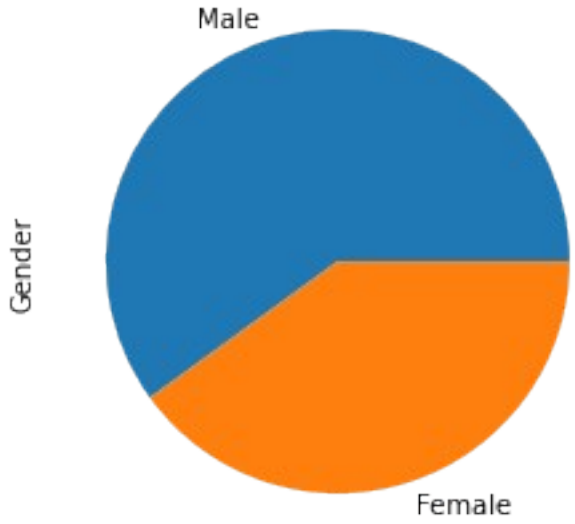
```
sns.heatmap(d, annot=True)
```

```
<AxesSubplot:xlabel='MonthlyIncome', ylabel='Education'>
```



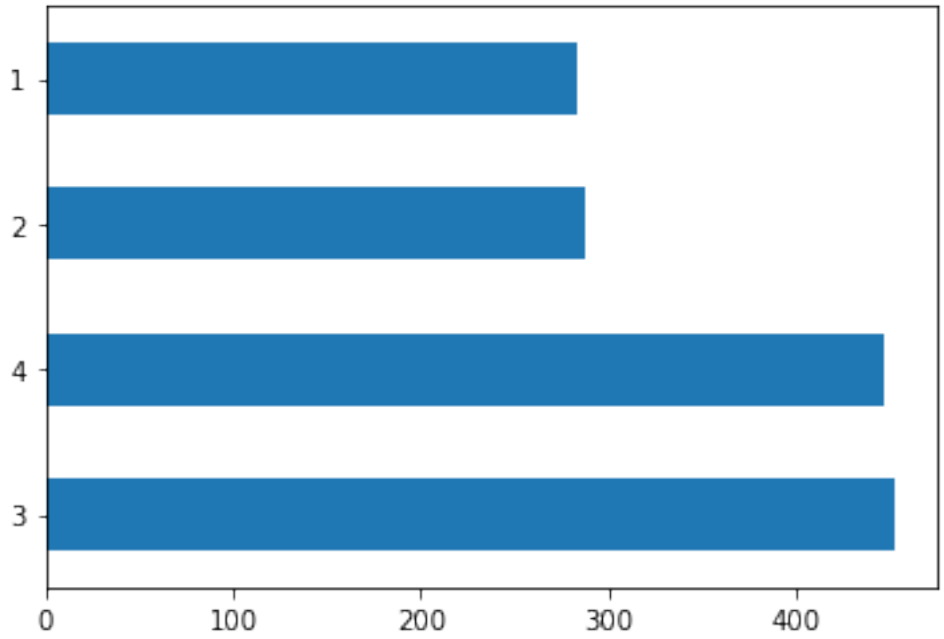

```
df['Gender'].value_counts().plot(kind='pie', title='')
```

```
<AxesSubplot:ylabel='Gender'>
```




```
df['EnvironmentSatisfaction'].value_counts().plot(kind='barh',
title='')
```

```
<AxesSubplot:>
```




```
df['Age'].describe()
```

```
count    1470.000000
mean       36.923810
std         9.135373
min        18.000000
25%        30.000000
50%        36.000000
75%        43.000000
max        60.000000
Name: Age, dtype: float64
```

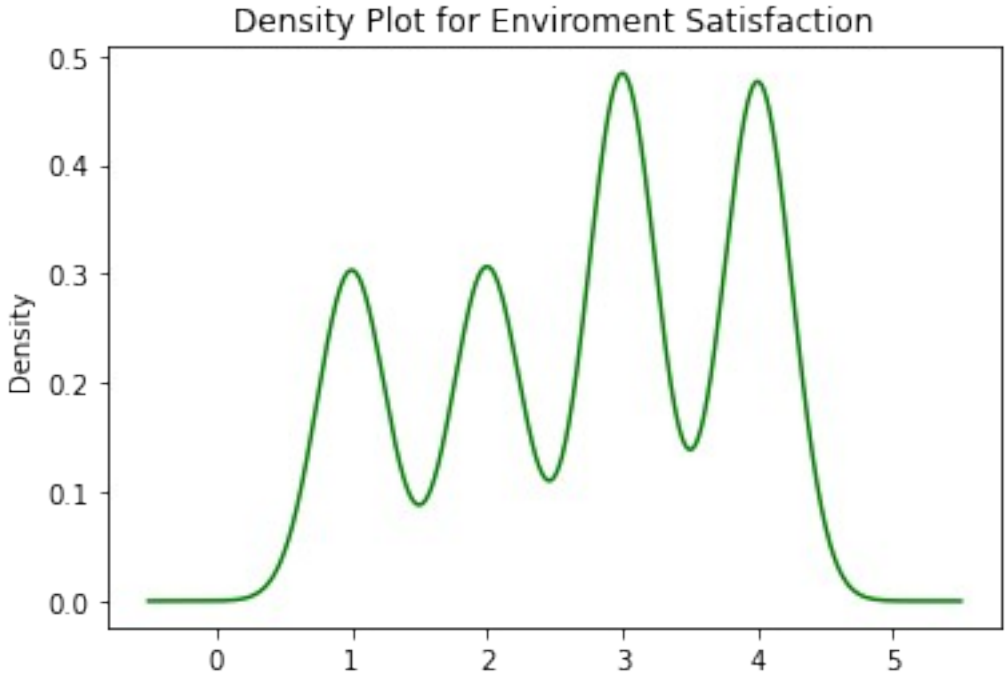
```
#Mean
print(df['DailyRate'].mean())
print(df['DistanceFromHome'].mean())
print(df['Education'].mean())
print(df['HourlyRate'].mean())
print(df['MonthlyIncome'].mean())
print(df['MonthlyRate'].mean())
print(df['NumCompaniesWorked'].mean())
print(df['TotalWorkingYears'].mean())
```

```
802.4857142857143
9.19251700680272
2.912925170068027
65.89115646258503
6502.931292517007
14313.103401360544
2.6931972789115646
11.279591836734694
```

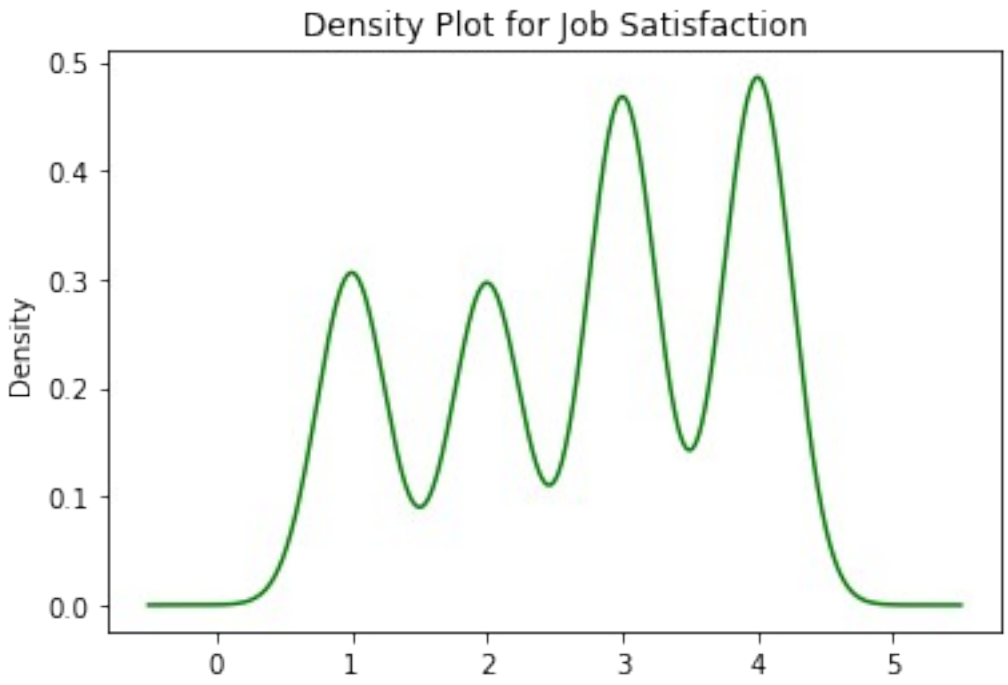
```
#median
print(df['DailyRate'].median())
print(df['DistanceFromHome'].median())
print(df['Education'].median())
print(df['HourlyRate'].median())
print(df['MonthlyIncome'].median())
print(df['MonthlyRate'].median())
print(df['NumCompaniesWorked'].median())
print(df['TotalWorkingYears'].median())
```

```
802.0
7.0
3.0
66.0
4919.0
14235.5
2.0
10.0
```

```
df.EnvironmentSatisfaction.plot.density(color='green')
plt.title('Density Plot for Enviroment Satisfaction')
plt.show()
```




```
df.JobSatisfaction.plot.density(color='green')
plt.title('Density Plot for Job Satisfaction')
plt.show()
```




```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
```

```
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

```
df.dtypes
```

```
Age                         int64
Attrition                  object
BusinessTravel             object
DailyRate                   int64
Department                 object
DistanceFromHome            int64
Education                   int64
EducationField             object
EmployeeCount               int64
EmployeeNumber              int64
EnvironmentSatisfaction     int64
Gender                     object
HourlyRate                  int64
JobInvolvement              int64
JobLevel                    int64
JobRole                    object
JobSatisfaction             int64
MaritalStatus              object
MonthlyIncome               int64
MonthlyRate                 int64
NumCompaniesWorked          int64
Over18                     object
OverTime                   object
PercentSalaryHike           int64
PerformanceRating           int64
RelationshipSatisfaction    int64
StandardHours               int64
StockOptionLevel            int64
TotalWorkingYears           int64
TrainingTimesLastYear       int64
WorkLifeBalance             int64
YearsAtCompany              int64
YearsInCurrentRole          int64
```

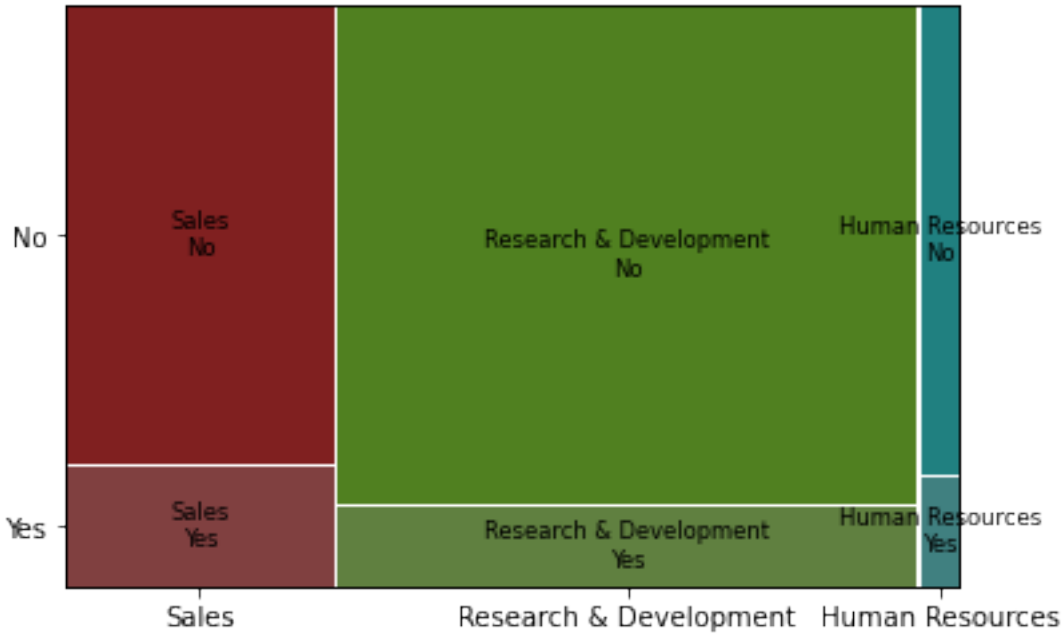
```
YearsSinceLastPromotion      int64
YearsWithCurrManager         int64
dtype: object
```

```
df_crosstab=pd.crosstab(df['Attrition'],df['Department'],margins=False
)
print(df_crosstab)
```

```
Department  Human Resources  Research & Development  Sales
Attrition
No                       51                     828    354
Yes                      12                     133     92
```

```
from statsmodels.graphics.mosaicplot import mosaic
import matplotlib.pyplot as plt
import pandas as pd
mosaic(df, index=['Department','Attrition'])
```

```
(<Figure size 432x288 with 3 Axes>,
 {('Sales', 'Yes'): (0.0, 0.0, 0.30039738667744326,
0.20559271784634178),
  ('Sales', 'No'): (0.0,
   0.2089149769825544,
   0.30039738667744326,
   0.7910850230174454),
  ('Research & Development', 'Yes'): (0.3053478817269482,
   0.0,
   0.6472688085135044,
   0.1379377102340101),
  ('Research & Development', 'No'): (0.3053478817269482,
   0.14125996937022273,
   0.6472688085135044,
   0.8587400306297772),
  ('Human Resources', 'Yes'): (0.9575671852899577,
   0.0,
   0.04243281471004239,
   0.18984337921214994),
  ('Human Resources', 'No'): (0.9575671852899577,
   0.19316563834836256,
   0.04243281471004239,
   0.8068343616516374)})
```




```
df_crosstab=pd.crosstab(df['Attrition'],df['Gender'], normalize=True)
print(df_crosstab)
```

```
Gender       Female      Male
Attrition
No         0.340816  0.497959
Yes        0.059184  0.102041
```

```
print(" A higher percentage of males than females are faced with
attrition in this workplace.")
```

```
pd.crosstab(df['WorkLifeBalance'],df['RelationshipSatisfaction'])
```

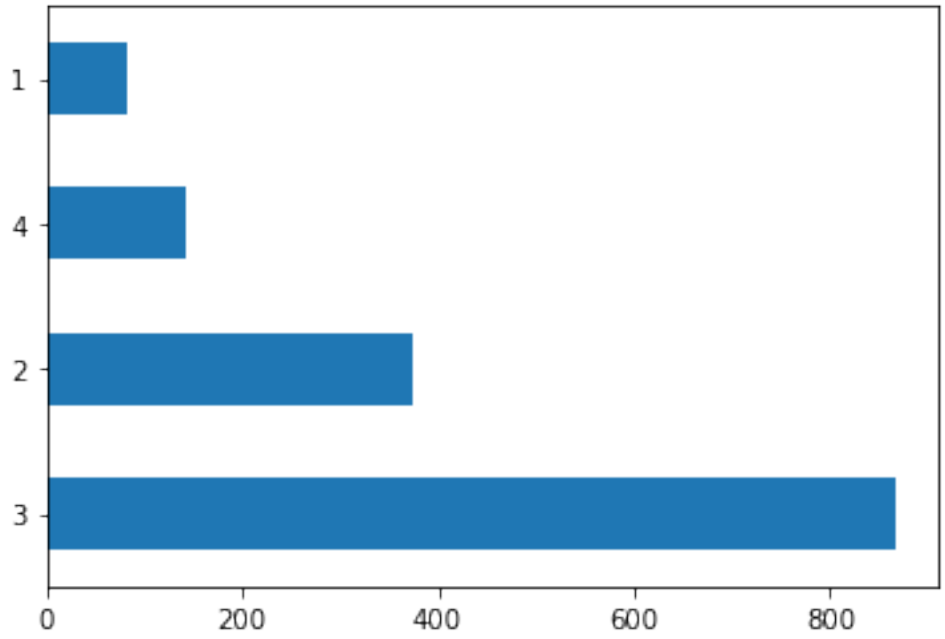
```
RelationshipSatisfaction    1    2    3    4
WorkLifeBalance
1                          20   13   26   21
2                          59   74  117   94
3                         171  181  272  269
4                          26   35   44   48
```

```
pd.crosstab(df['WorkLifeBalance'],df['Gender'],normalize=True)
```

```
Gender             Female      Male
WorkLifeBalance
1                0.020408  0.034014
2                0.092517  0.141497
3                0.248299  0.359184
4                0.038776  0.065306
```

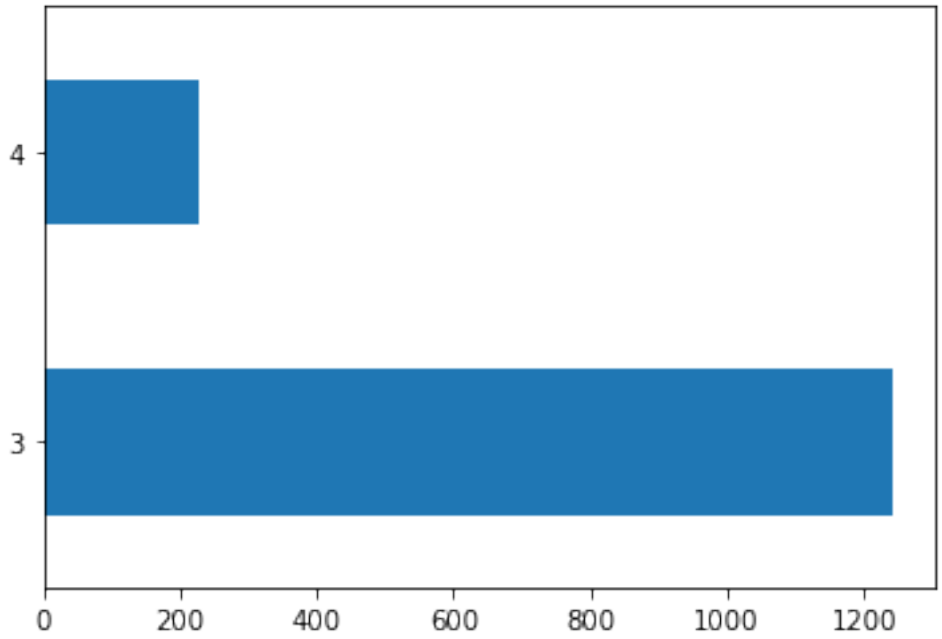
```
df['PercentSalaryHike'],df['JobInvolvement'].value_counts().plot(kind=
'barh', title='')
```

```
(0       11
 1       23
 2       15
 3       11
 4       12
         ..
 1465    17
 1466    15
 1467    20
 1468    14
 1469    12
 Name: PercentSalaryHike, Length: 1470, dtype: int64,
 <AxesSubplot:>)
```




```
df['MonthlyIncome'],df['PerformanceRating'].value_counts().plot(kind='
barh', title='')
```

```
(0       5993
 1       5130
 2       2090
 3       2909
 4       3468
         ...
 1465    2571
 1466    9991
 1467    6142
 1468    5390
 1469    4404
 Name: MonthlyIncome, Length: 1470, dtype: int64,
 <AxesSubplot:>)
```




```
pd.crosstab(df['WorkLifeBalance'],df['RelationshipSatisfaction'],norma
lize=True)
```

RelationshipSatisfaction	1	2	3	4
WorkLifeBalance				
1	0.013605	0.008844	0.017687	0.014286
2	0.040136	0.050340	0.079592	0.063946
3	0.116327	0.123129	0.185034	0.182993
4	0.017687	0.023810	0.029932	0.032653

```
pd.crosstab(df['JobInvolvement'],df['MaritalStatus'],normalize=True)
```

MaritalStatus	Divorced	Married	Single
JobInvolvement			
1	0.014966	0.022449	0.019048
2	0.048299	0.119048	0.087755
3	0.136054	0.265986	0.188435
4	0.023129	0.050340	0.024490