# Statistic Analysis

**Here, we try to do statistical analysis on how Job Satisfaction level depends on various factors or predictors which we were taking by Primary Data Collection.**

# Importing Relevant Libraries

In [3]:
```
import pandas as pd
import numpy as np
df=pd.read_csv('file:///C:/Users/angsh/OneDrive/Desktop/PRAXIS/Own%20Projects/ML/Statistics%20Project/JOB%20SATISFACT
```

# Data Preprocessing

In [4]:
```
df
```

Out[4]:

| | Timestamp | AGE | GENDER | WORK CITY | IS WORK CITY DIFFERENT FROM YOUR HOME CITY? | DOMAIN/INDUSTRY | JOB ROLE/DESIGNATION | CURRENT WORKING MODE | PREFERRED WORKING MODE | WORKING EXP (YEARS) | ... | PLE/ CURI SATI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 3/17/2023 12:25:52 | 28 | MALE | Kolkata | No | IT | Advance Analytics | HYBRID | HYBRID | 4 | ... | |
| **1** | 3/17/2023 12:27:06 | 48 | FEMALE | Kolkata | No | EDUCATION | Assistant Manager, Content Marketing | WORK FROM OFFICE | WORK FROM OFFICE | 18+ | ... | |
| **2** | 3/17/2023 12:34:54 | 34 | MALE | Kolkata | No | Construction | Senior project Engineer | WORK FROM OFFICE | WORK FROM OFFICE | 14 | ... | |
| **3** | 3/17/2023 12:36:01 | 30 | MALE | Mumbai | No | FINANCE/BANKING | Assistant Manager - IT | WORK FROM HOME | WORK FROM HOME | 5 | ... | |
| **4** | 3/17/2023 12:54:00 | 28 | MALE | Kolkata | Yes | Building material construction | Business development associate | WORK FROM OFFICE | WORK FROM OFFICE | 4.5 | ... | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **100** | 3/22/2023 10:23:34 | 32 | MALE | Kolkata | No | EDUCATION | Guest lecturer | WORK FROM OFFICE | WORK FROM OFFICE | 3years | ... | |
| **101** | 3/22/2023 20:05:48 | 27 | MALE | Farakka | Yes | FINANCE/BANKING | Assistant Manager | WORK FROM OFFICE | WORK FROM OFFICE | 2 | ... | |
| **102** | 3/28/2023 23:12:55 | 25 | MALE | Kolkata | No | ECOMMERCE | Business development executive | WORK FROM OFFICE | WORK FROM OFFICE | 2 | ... | |
| **103** | 3/29/2023 23:08:23 | 24 | MALE | Sambalpur | Yes | GOVT.SECTOR | Management Trainee | WORK FROM OFFICE | WORK FROM HOME | 0.125 | ... | |
| **104** | 3/17/2023 12:29:27 | 30 | MALE | Hyderabad | Yes | FINANCE/BANKING | Chief Manager | WORK FROM OFFICE | WORK FROM OFFICE | 5 | ... | |

105 rows × 33 columns

In [5]:
```
df.dtypes
```

Out[5]:
```
Timestamp                                                                    object
AGE                                                                          object
GENDER                                                                       object
WORK CITY                                                                    object
IS WORK CITY DIFFERENT FROM YOUR HOME CITY?                                  object
DOMAIN/INDUSTRY                                                              object
JOB ROLE/DESIGNATION                                                         object
CURRENT WORKING MODE                                                         object
PREFERRED WORKING MODE                                                       object
WORKING EXP (YEARS)                                                          object
AVG WORKING HOURS                                                            object
INCOME                                                                      float64
MARITAL STATUS                                                               object
TIME TAKEN TO REACH OFFICE                                                  float64
RAPPORT WITH COLLEAGUES/TEAM                                                float64
SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?    float64
HAVING CHILDREN BELOW AGE OF 12 ?                                            object
NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?                                 object
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL                             float64
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [1st choice]        object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [2nd choice]        object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [3rd choice]        object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [4th Choice]        object
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL                             float64
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [1st Choice]       object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [2nd Choice]       object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [3rd Choice]       object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [4th Choice]       object
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL .1                          float64
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [1st Choice]         object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [2nd Choice]         object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [3rd Choice]         object
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [4th Choice]         object
dtype: object
```

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

Out[6]:
```
Timestamp                                                                    0
AGE                                                                          0
GENDER                                                                       0
WORK CITY                                                                    0
IS WORK CITY DIFFERENT FROM YOUR HOME CITY?                                  0
DOMAIN/INDUSTRY                                                              0
JOB ROLE/DESIGNATION                                                         0
CURRENT WORKING MODE                                                         0
PREFERRED WORKING MODE                                                       0
WORKING EXP (YEARS)                                                          0
AVG WORKING HOURS                                                            0
INCOME                                                                       0
MARITAL STATUS                                                               0
TIME TAKEN TO REACH OFFICE                                                   2
RAPPORT WITH COLLEAGUES/TEAM                                                 0
SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?    0
HAVING CHILDREN BELOW AGE OF 12 ?                                            0
NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?                                 0
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL                             84
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [1st choice]       84
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [2nd choice]       84
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [3rd choice]       84
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [4th Choice]       84
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL                              55
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [1st Choice]      55
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [2nd Choice]      55
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [3rd Choice]      55
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [4th Choice]      55
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL .1                           71
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [1st Choice]        71
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [2nd Choice]        71
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [3rd Choice]        71
REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [4th Choice]        71
dtype: int64
```

In [7]:
```
df.columns
```

Out[7]:
```
Index(['Timestamp', 'AGE', 'GENDER', 'WORK CITY ',
       'IS WORK CITY DIFFERENT FROM YOUR HOME CITY?', 'DOMAIN/INDUSTRY',
       'JOB ROLE/DESIGNATION', 'CURRENT WORKING MODE',
       'PREFERRED WORKING MODE', 'WORKING EXP (YEARS)', 'AVG WORKING HOURS',
       'INCOME', 'MARITAL STATUS', 'TIME TAKEN TO REACH OFFICE',
       'RAPPORT WITH COLLEAGUES/TEAM',
       'SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?',
       'HAVING CHILDREN BELOW AGE OF 12 ?',
       'NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?',
       'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [1st choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [2nd choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [3rd choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [4th Choice]',
       'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL ',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [1st Choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [2nd Choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [3rd Choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [4th Choice]',
       'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL .1',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [1st Choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [2nd Choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [3rd Choice]',
       'REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [4th Choice]'],
      dtype='object')
```

In [8]:
```
#Merging Job Rating Satisfaction level to one single column
df['PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW']=df['PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL'].filln
```

In [9]:
```
#Merging different choices to their respective choice columns
df['1st_Choice']=df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [1st choice]'].fillna("")+df['REASONS
df['2nd_Choice']=df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [2nd choice]'].fillna("")+df['REASONS
df['3rd_Choice']=df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [3rd choice]'].fillna("")+df['REASONS
df['4th_Choice']=df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [4th Choice]'].fillna("")+df['REASONS
```

In [10]:
```
#Droping redundant, unecessary and errorenous columns and rows
df.drop(['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [1st choice]','REASONS BEHIND SELECTING YOUR "PRE
df.drop(df[df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=='Hybrid '].index,inplace=True)
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
AGE                                                                                  0
GENDER                                                                               0
WORK CITY                                                                            0
IS WORK CITY DIFFERENT FROM YOUR HOME CITY?                                          0
DOMAIN/INDUSTRY                                                                      0
JOB ROLE/DESIGNATION                                                                 0
CURRENT WORKING MODE                                                                 0
PREFERRED WORKING MODE                                                               0
WORKING EXP (YEARS)                                                                  0
AVG WORKING HOURS                                                                    0
INCOME                                                                               0
MARITAL STATUS                                                                       0
TIME TAKEN TO REACH OFFICE                                                           1
RAPPORT WITH COLLEAGUES/TEAM                                                         0
SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?    0
HAVING CHILDREN BELOW AGE OF 12 ?                                                    0
NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?                                         0
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW                                  0
1st_Choice                                                                           0
2nd_Choice                                                                           0
3rd_Choice                                                                           0
4th_Choice                                                                           0
dtype: int64
```

# Data Cleaning

In [12]:
```
#DATA CLEANING
#AGE
df.loc[64,'AGE']='23'
df['AGE']=df['AGE'].astype('int64')
#WORK CITY
df['WORK CITY ']=df['WORK CITY '].str.replace(' ','')
df.loc[29,'WORK CITY ']='AHMEDABAD'
df['WORK CITY ']=df['WORK CITY '].str.replace('Ahemdabad','AHMEDABAD')
df['WORK CITY ']=df['WORK CITY '].str.replace('BLR','BANGALORE')
df['WORK CITY ']=df['WORK CITY '].str.replace('Sanandahmedabad','AHMEDABAD')
df['WORK CITY ']=df['WORK CITY '].str.replace('Newtown','Kolkata')
df['WORK CITY ']=df['WORK CITY '].str.replace('Sanand','AHMEDABAD')
df['WORK CITY ']=df['WORK CITY '].str.upper()
#DOMAIN/INDUSTRY
df['DOMAIN/INDUSTRY']=df['DOMAIN/INDUSTRY'].str.upper()
df['DOMAIN/INDUSTRY']=df['DOMAIN/INDUSTRY'].str.replace('BUILDING MATERIAL CONSTRUCTION','CONSTRUCTION ')
#JOB ROLE/DESIGNATION
df['JOB ROLE/DESIGNATION']=df['JOB ROLE/DESIGNATION'].str.upper()
#WORKING EXP (YEARS)
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace(' ','')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('years','')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('year','')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('yr','')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('yrs','')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('0.125','0')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('1yr5months','1.5')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('Fewmonths','.5')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('3s','3')
df.loc[1,'WORKING EXP (YEARS)']='19'
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('Months','')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('months','')
df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].astype('float64')
#AVG WORKING HOURS
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace(' ','')
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('hours','')
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('8-9','8.5')
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('hr','')
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('s','')
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('7to13','10')
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('Hours','')
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('Hour','')
```

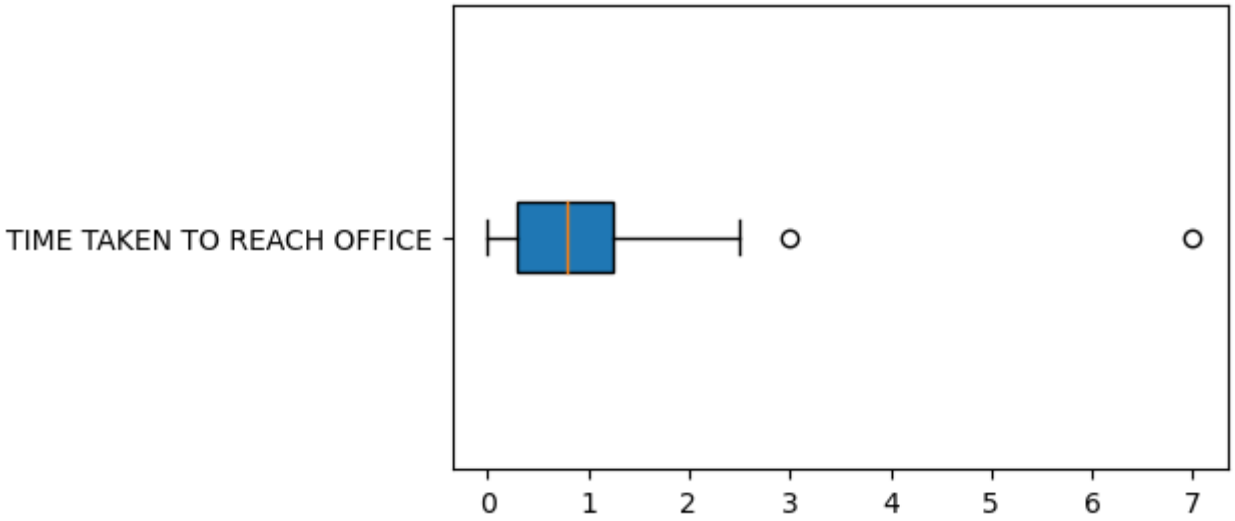
```
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('hrs.','')
df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].astype('float64')
#TIME TAKEN TO REACH OFFICE
df['TIME TAKEN TO REACH OFFICE']=df['TIME TAKEN TO REACH OFFICE'].astype('float64')
#NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace(' '
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('No
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('Ze
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('me
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('Me
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('No
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('Nu
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].astype('float64
```

```
C:\Users\angsh\AppData\Local\Temp\ipykernel_12204\832836406.py:25: FutureWarning: The default value of regex will ch
ange from True to False in a future version.
  df['WORKING EXP (YEARS)']=df['WORKING EXP (YEARS)'].str.replace('0.125','0')
C:\Users\angsh\AppData\Local\Temp\ipykernel_12204\832836406.py:42: FutureWarning: The default value of regex will ch
ange from True to False in a future version.
  df['AVG WORKING HOURS']=df['AVG WORKING HOURS'].str.replace('hrs.','')
```

# Treating Missing Values

In [13]:
```
#Checking for outliers
df1=df.dropna()
import matplotlib.pyplot as plt
plt.figure(figsize=(5,3))
plt.boxplot(df1['TIME TAKEN TO REACH OFFICE'],vert=False, labels=['TIME TAKEN TO REACH OFFICE'],patch_artist=True)
```

Out[13]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x257aab212b0>,
  <matplotlib.lines.Line2D at 0x257aab21580>],
 'caps': [<matplotlib.lines.Line2D at 0x257aab21850>,
  <matplotlib.lines.Line2D at 0x257aab21b20>],
 'boxes': [<matplotlib.patches.PathPatch at 0x257aab06eb0>],
 'medians': [<matplotlib.lines.Line2D at 0x257aab21df0>],
 'fliers': [<matplotlib.lines.Line2D at 0x257aab34100>],
 'means': []}
```




Since 'TIME TAKEN TO REACH OFFICE (HH:MM)" have outliers as inferred from above boxplot, We will replace the missing value of 'TIME TAKEN TO REACH OFFICE (HH:MM)' with its median, as median is roboust to Outliers

In [14]:
```
Med_Tmtkn=np.median(np.array(df['TIME TAKEN TO REACH OFFICE']))
df.loc[25,'TIME TAKEN TO REACH OFFICE']=Med_Tmtkn
```

In [15]:
```
df.describe()
```

Out[15]:

| | AGE | WORKING EXP (YEARS) | AVG WORKING HOURS | INCOME | TIME TAKEN TO REACH OFFICE | RAPPORT WITH COLLEAGUES/TEAM | SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ? | NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ? | PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 104.000000 | 104.000000 | 104.000000 | 104.000000 | 103.000000 | 104.000000 | 104.000000 | 104.000000 | 104.000000 |
| **mean** | 28.567308 | 4.975000 | 8.201923 | 6.513462 | 0.942718 | 3.629808 | 3.485577 | 2.826923 | 3.227885 |
| **std** | 5.957396 | 5.954663 | 1.589589 | 4.288256 | 0.886345 | 0.866404 | 1.097253 | 1.887323 | 1.036186 |
| **min** | 21.000000 | 0.000000 | 2.000000 | 0.900000 | 0.000000 | 1.000000 | 1.300000 | 0.000000 | 0.400000 |
| **25%** | 25.000000 | 2.000000 | 8.000000 | 3.000000 | 0.300000 | 3.300000 | 2.600000 | 2.000000 | 2.600000 |
| **50%** | 27.000000 | 3.000000 | 8.000000 | 5.400000 | 0.800000 | 3.700000 | 3.700000 | 3.000000 | 3.500000 |
| **75%** | 31.000000 | 5.000000 | 9.000000 | 8.475000 | 1.250000 | 4.400000 | 4.500000 | 4.000000 | 3.900000 |
| **max** | 60.000000 | 40.000000 | 15.000000 | 17.700000 | 7.000000 | 5.000000 | 5.000000 | 9.000000 | 5.000000 |

In [16]:
```
df.describe(exclude='number')
```

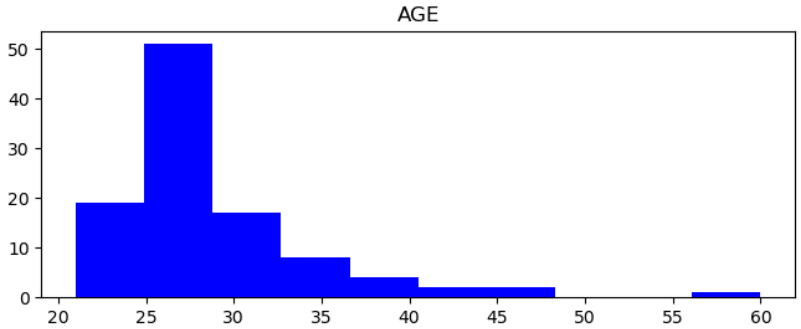
Out[16]:

| | GENDER | WORK CITY | IS WORK CITY DIFFERENT FROM YOUR HOME CITY? | DOMAIN/INDUSTRY | JOB ROLE/DESIGNATION | CURRENT WORKING MODE | PREFERRED WORKING MODE | MARITAL STATUS | HAVING CHILDREN BELOW AGE OF 12 ? | 1st_Choice |
|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 104 | 104 | 104 | 104 | 104 | 104 | 104 | 104 | 104 | 104 |
| **unique** | 2 | 24 | 2 | 19 | 86 | 3 | 3 | 3 | 2 | 12 |
| **top** | MALE | KOLKATA | No | IT | TEACHER | WORK FROM OFFICE | WORK FROM OFFICE | SINGLE | No | Better Working environment |
| **freq** | 63 | 55 | 57 | 29 | 5 | 81 | 49 | 74 | 93 | 23 |

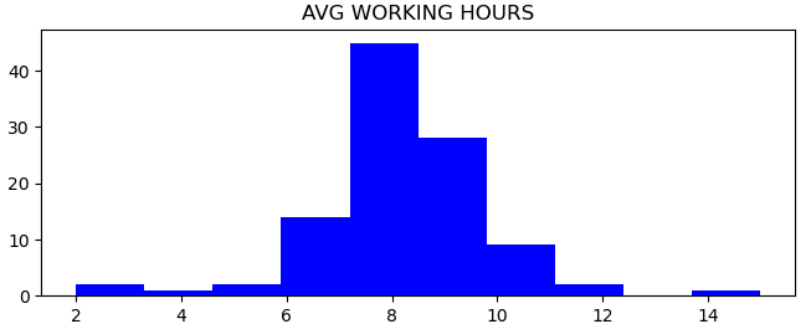
# Univariate Analysis

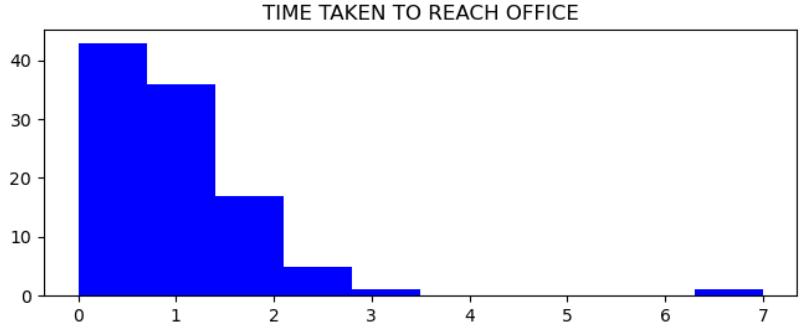
In [17]:
```
df.hist(
['AGE','WORKING EXP (YEARS)',
 'AVG WORKING HOURS',
 'TIME TAKEN TO REACH OFFICE',
 'NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'],figsize=(18,10),color='blue',grid=False)
```

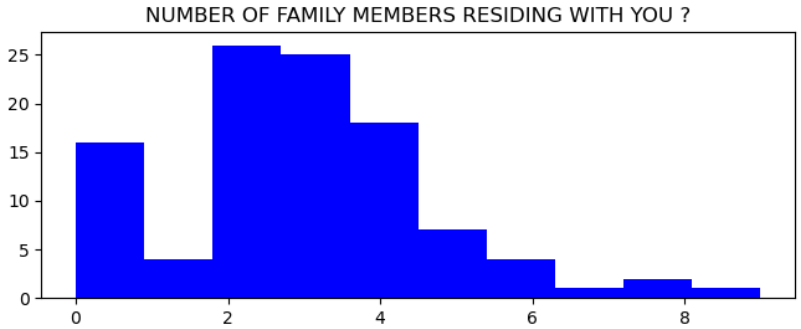
Out[17]:
```
array([[<AxesSubplot:title={'center':'AGE'}>,
        <AxesSubplot:title={'center':'WORKING EXP (YEARS)'}>],
       [<AxesSubplot:title={'center':'AVG WORKING HOURS'}>,
        <AxesSubplot:title={'center':'TIME TAKEN TO REACH OFFICE'}>],
       [<AxesSubplot:title={'center':'NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'}>,
        <AxesSubplot:>]], dtype=object)
```


AGE
50
40
30
20
10
0
20
25
30
35
40
45
50
55
60


WORKING EXP (YEARS)
50
40
30
20
10
0
0
5
10
15
20
25
30
35
40

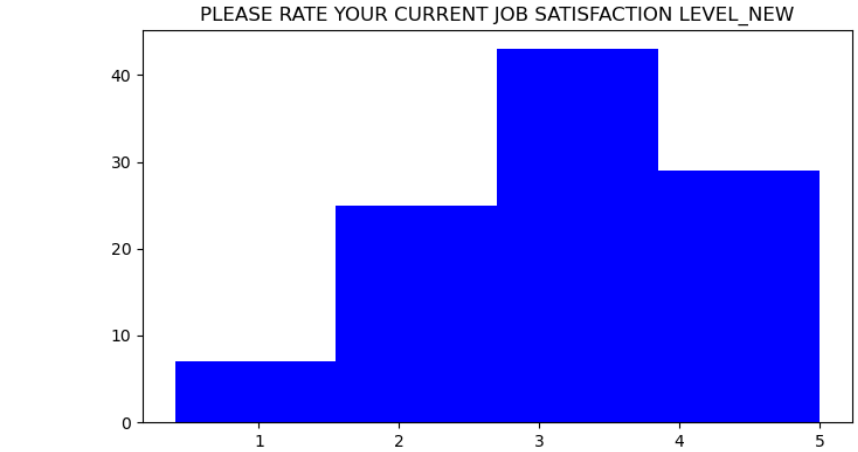

AVG WORKING HOURS
40
30
20
10
0
2
4
6
8
10
12
14

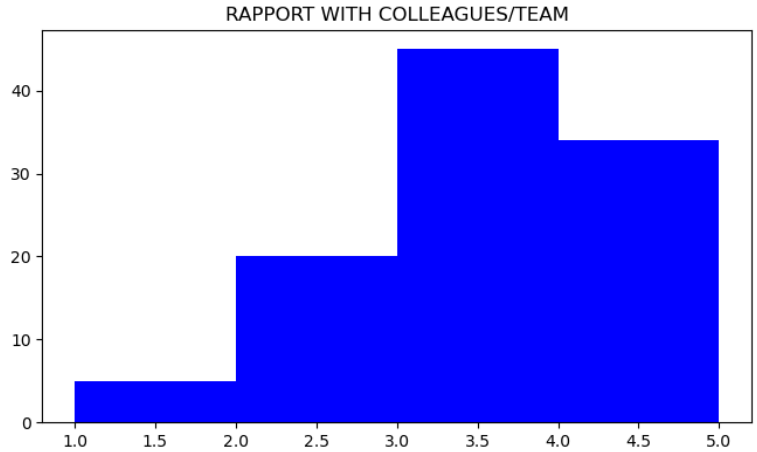

TIME TAKEN TO REACH OFFICE
40
30
20
10
0
0
1
2
3
4
5
6
7

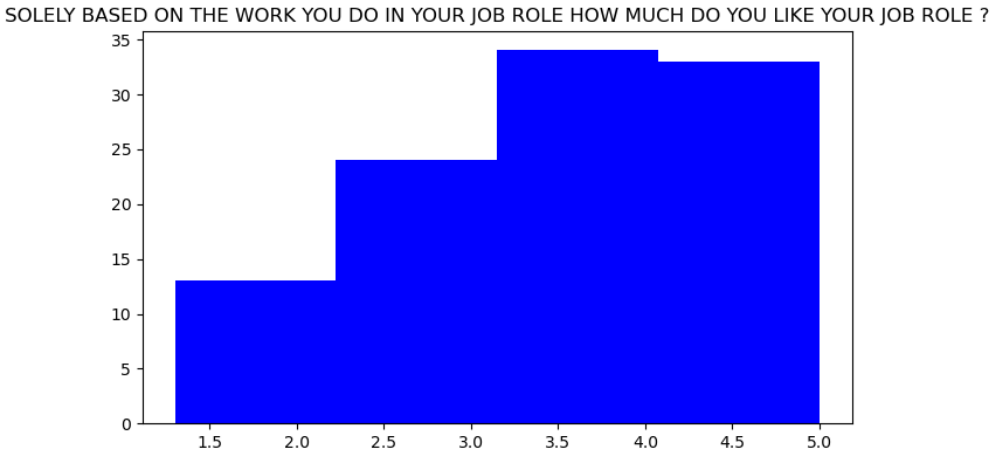

NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?
25
20
15
10
5
0
0
2
4
6
8

In [18]:
```
df.hist(
[ 'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW',
 'RAPPORT WITH COLLEAGUES/TEAM',
 'SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?'],figsize=(18,10),color='blue
```

Out[18]:
```
array([[<AxesSubplot:title={'center':'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW'}>,
        <AxesSubplot:title={'center':'RAPPORT WITH COLLEAGUES/TEAM'}>],
       [<AxesSubplot:title={'center':'SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB
ROLE ?'}>,
        <AxesSubplot:>]], dtype=object)
```



In [19]:
```
list_numerical=list(df[df.dtypes[df.dtypes=='float64'].index].columns)+list(df[df.dtypes[df.dtypes=='int64'].index].c
```

In [20]:
```
list_numerical
```

Out[20]:
```
['WORKING EXP (YEARS)',
 'AVG WORKING HOURS',
 'INCOME',
 'TIME TAKEN TO REACH OFFICE',
 'RAPPORT WITH COLLEAGUES/TEAM',
 'SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?',
 'NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?',
 'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW',
 'AGE']
```

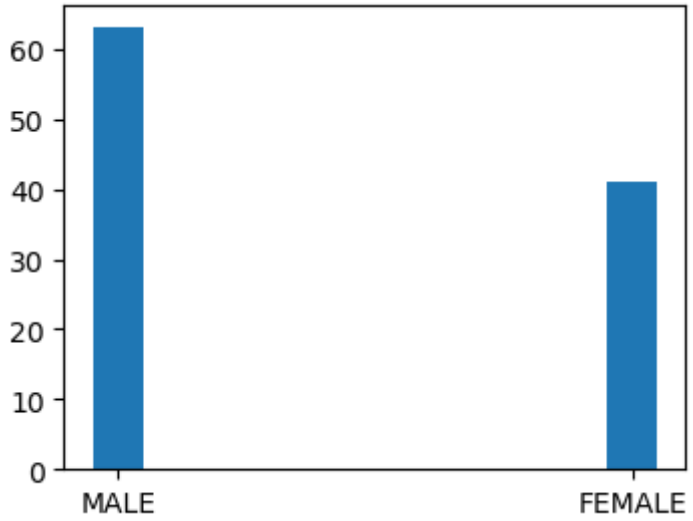
In [21]:
```
Categorical=list(df[df.dtypes[df.dtypes=='object'].index].columns)
```

In [22]:
```
Categorical
```

Out[22]:
```
['GENDER',
 'WORK CITY ',
 'IS WORK CITY DIFFERENT FROM YOUR HOME CITY?',
 'DOMAIN/INDUSTRY',
 'JOB ROLE/DESIGNATION',
 'CURRENT WORKING MODE',
 'PREFERRED WORKING MODE',
 'MARITAL STATUS',
 'HAVING CHILDREN BELOW AGE OF 12 ?',
 '1st_Choice',
 '2nd_Choice',
 '3rd_Choice',
 '4th_Choice']
```

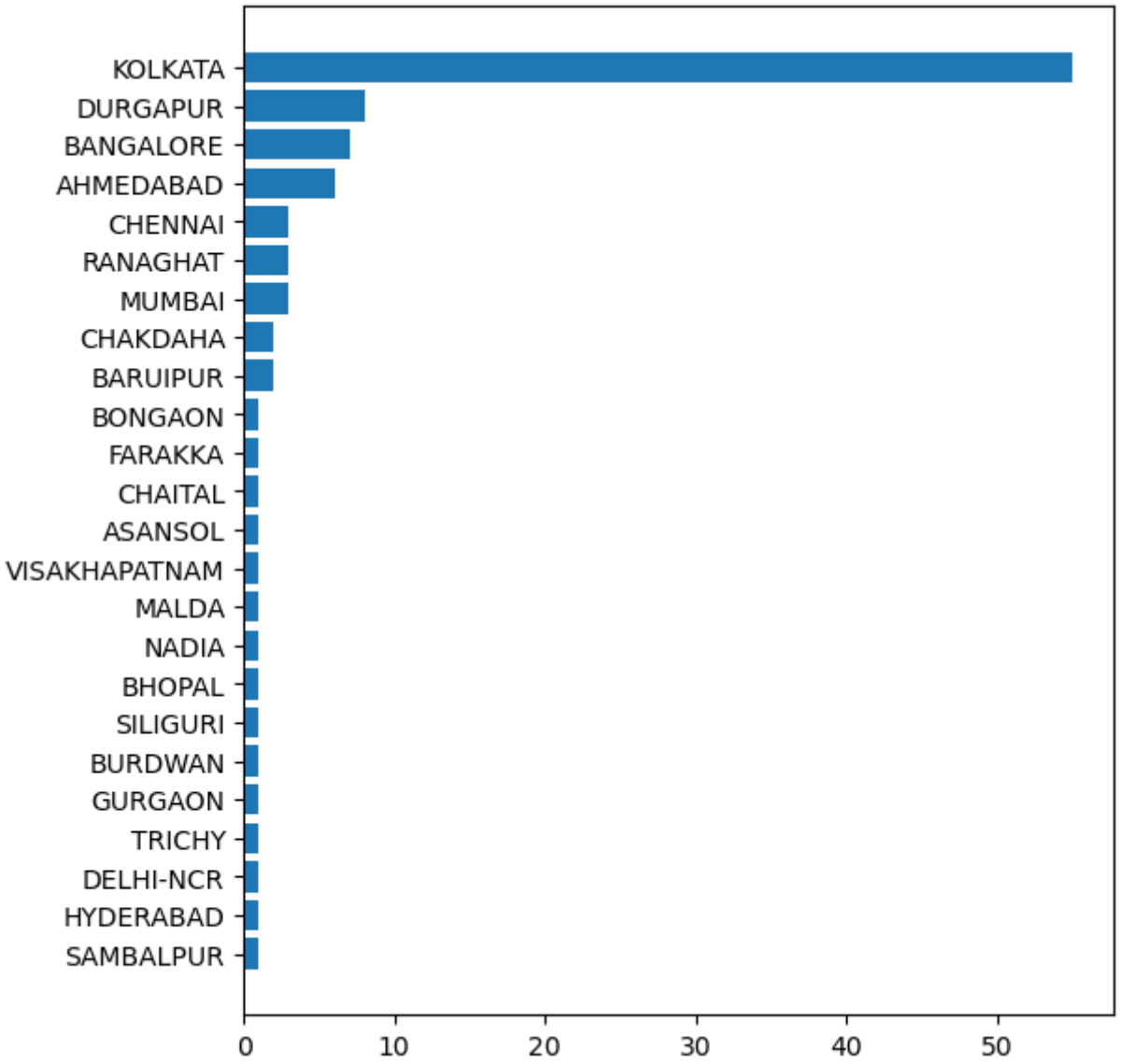
In [23]:
```
import matplotlib.pyplot as plt
plt.figure(figsize=(4,3))
plt.bar(list(df['GENDER'].value_counts().index),list(df['GENDER'].value_counts()),width=0.1)
```

Out[23]:
```
<BarContainer object of 2 artists>
```



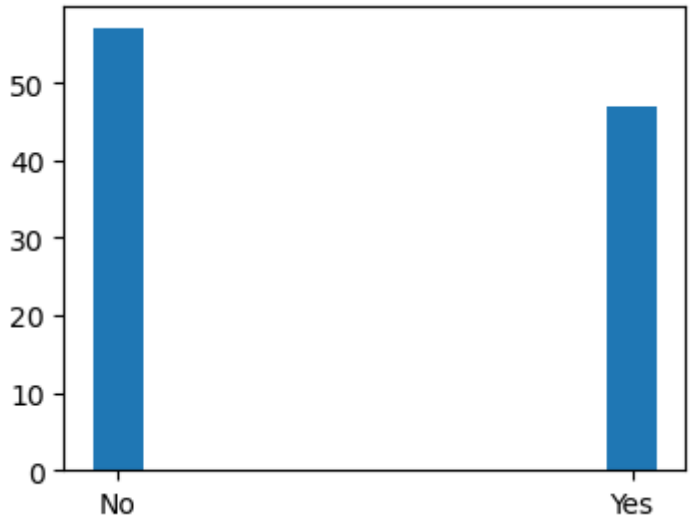
In [24]:
```
plt.figure(figsize=(6,7))
plt.barh(df['WORK CITY '].value_counts().index.tolist()[::-1],list(df['WORK CITY '].value_counts())[::-1])
```

Out[24]: <BarContainer object of 24 artists>



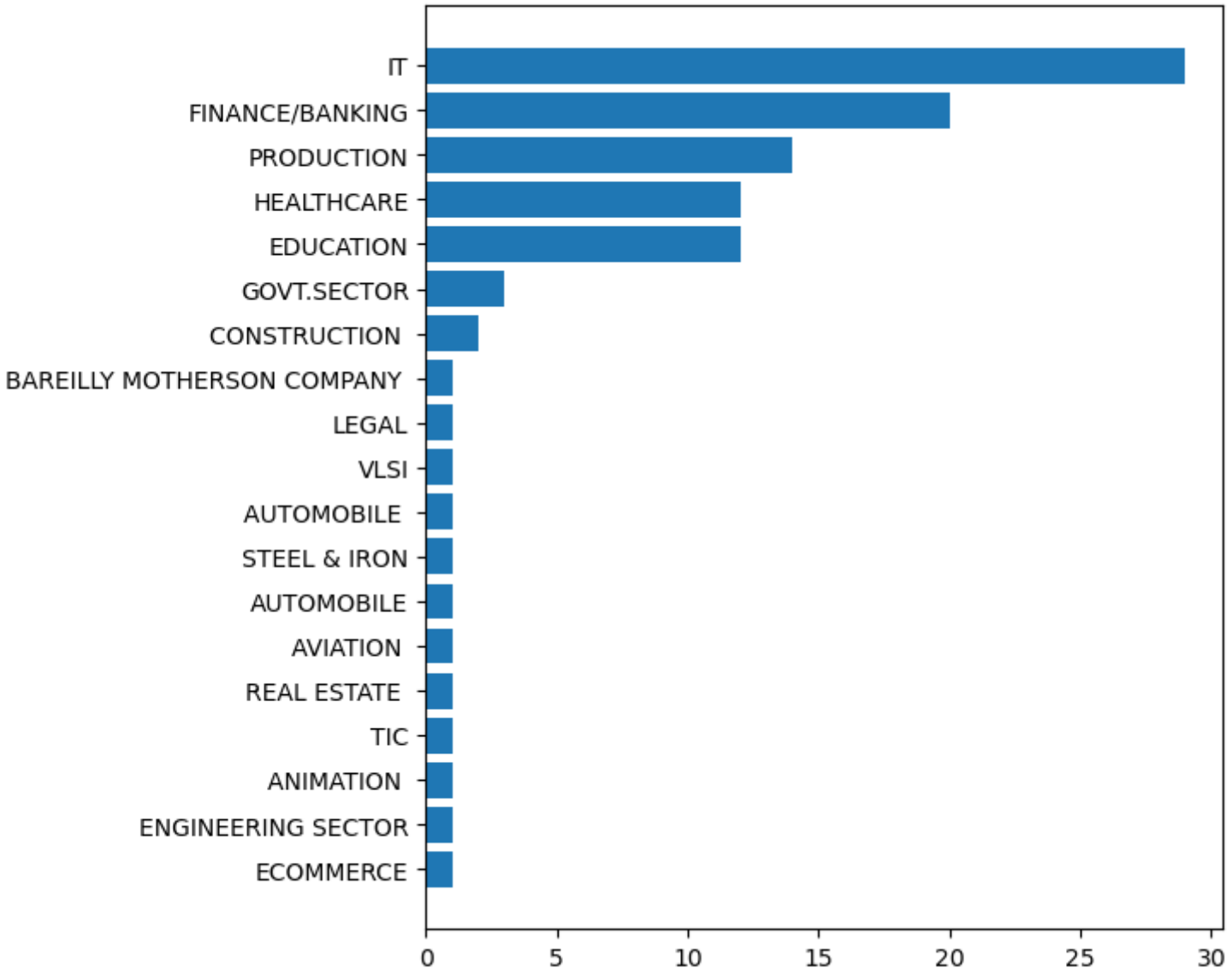
In [25]:
```
plt.figure(figsize=(4,3))
plt.bar(list(df['IS WORK CITY DIFFERENT FROM YOUR HOME CITY?'].value_counts().index),list(df['IS WORK CITY DIFFERENT
```

Out[25]:
```
<BarContainer object of 2 artists>
```



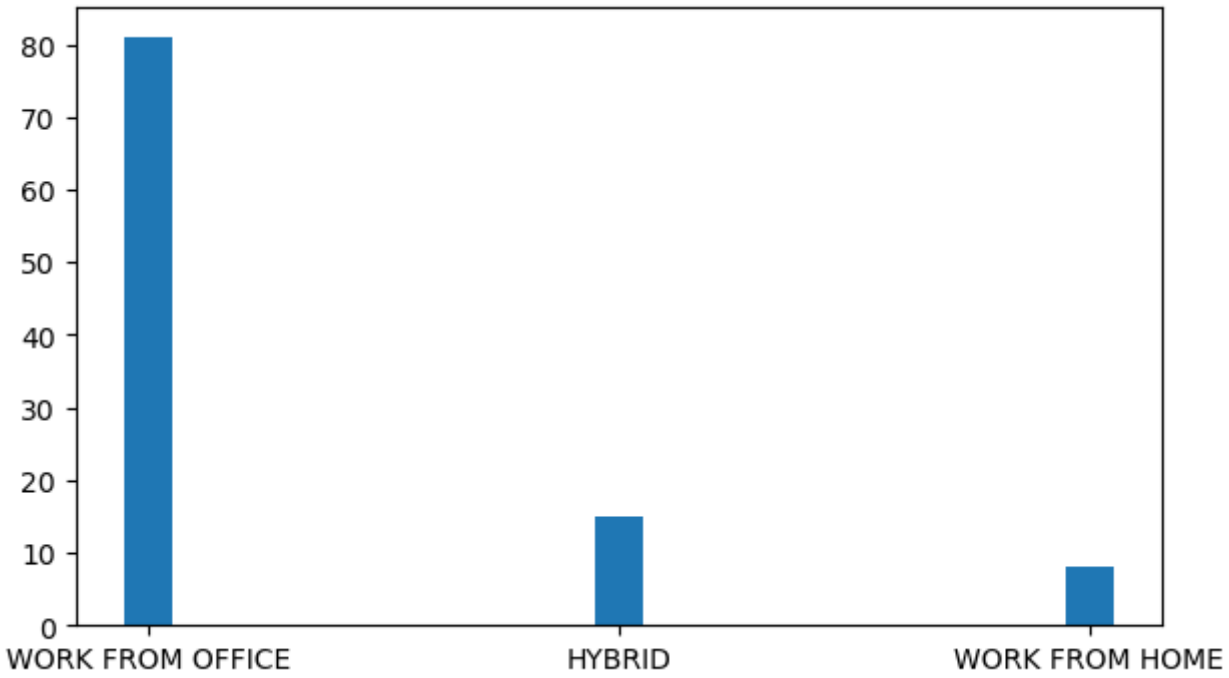
In [26]:
```
plt.figure(figsize=(6,7))
plt.barh(df['DOMAIN/INDUSTRY'].value_counts().index.tolist()[::-1],list(df['DOMAIN/INDUSTRY'].value_counts())[::-1])
```

Out[26]: <BarContainer object of 19 artists>



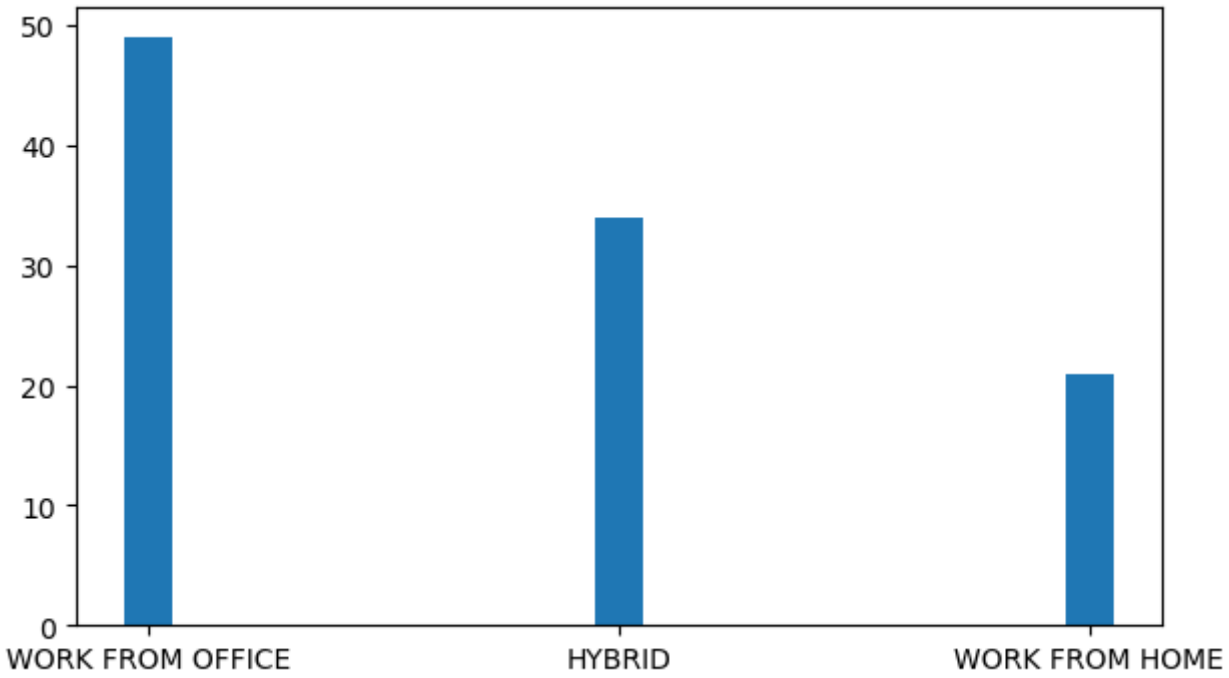
In [27]:
```
plt.figure(figsize=(7,4))
plt.bar(df['CURRENT WORKING MODE'].value_counts().index.tolist(),list(df['CURRENT WORKING MODE'].value_counts()),widt
```

Out[27]:
```
<BarContainer object of 3 artists>
```



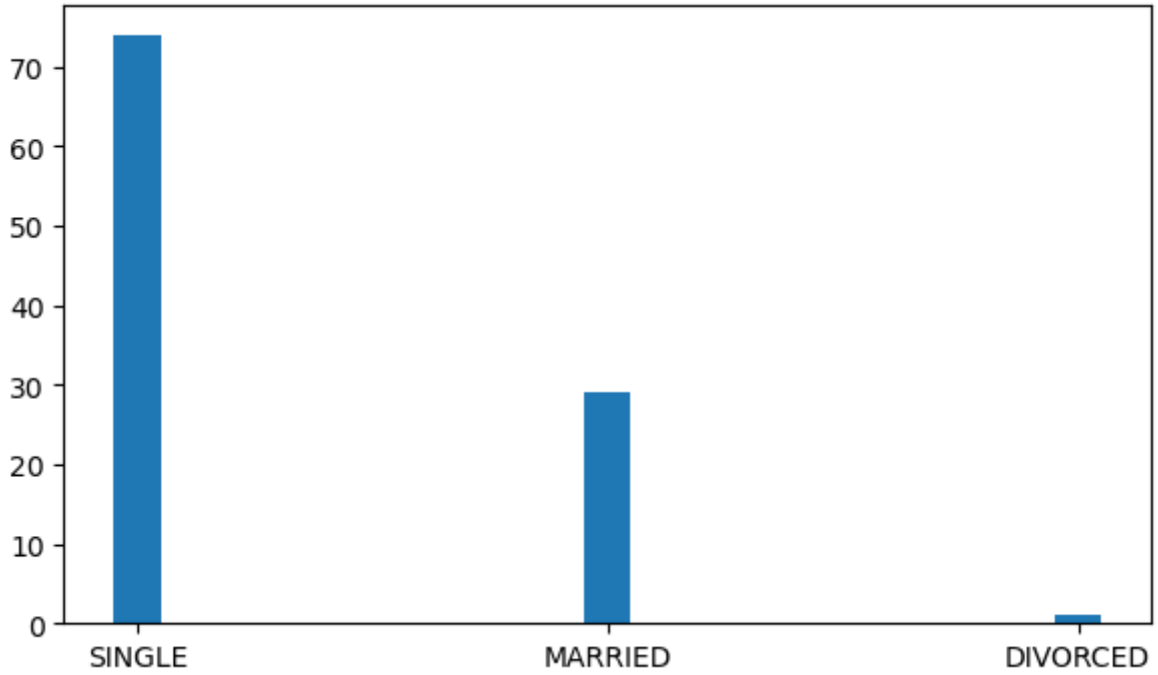
In [28]:
```
plt.figure(figsize=(7,4))
plt.bar(df['PREFERRED WORKING MODE'].value_counts().index.tolist(),list(df['PREFERRED WORKING MODE'].value_counts()),
```

Out[28]:
```
<BarContainer object of 3 artists>
```



In [29]:
```
plt.figure(figsize=(7,4))
plt.bar(df['MARITAL STATUS'].value_counts().index.tolist(),list(df['MARITAL STATUS'].value_counts()),width=0.1)
```

Out[29]: <BarContainer object of 3 artists>

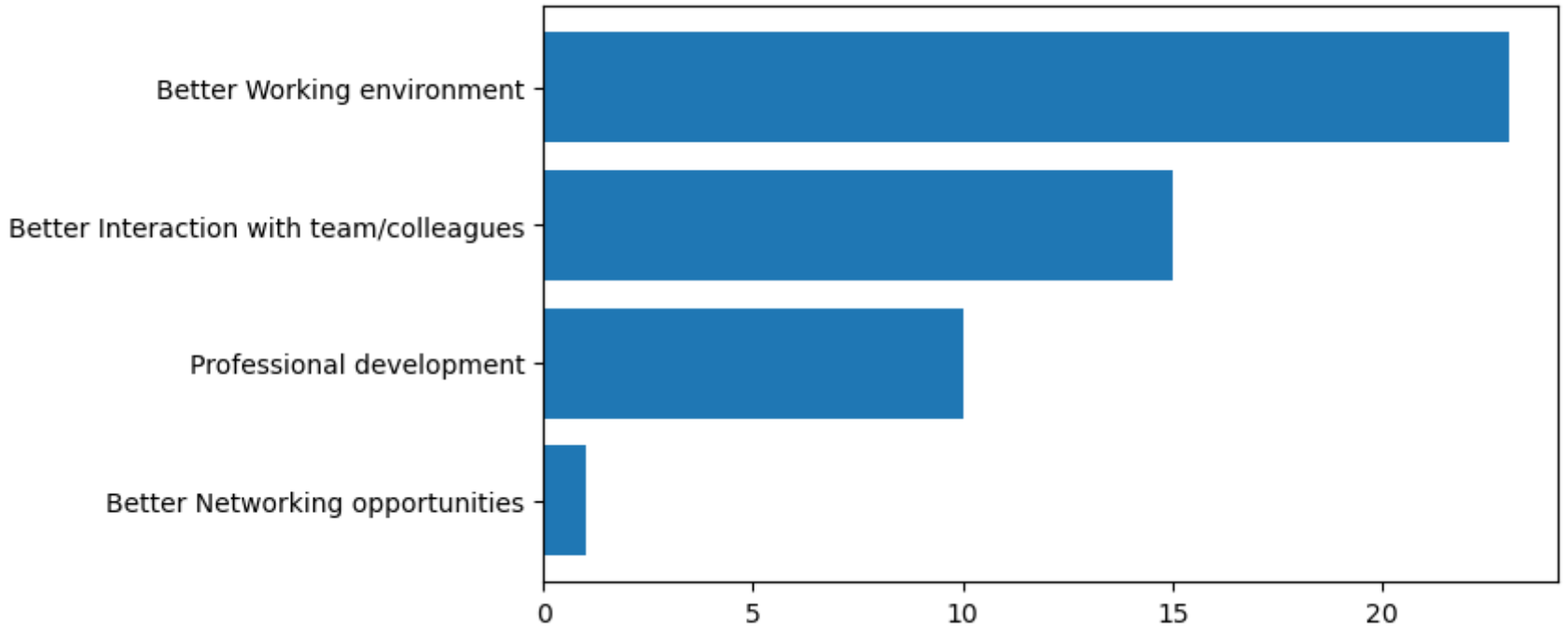


In [30]:
```
plt.figure(figsize=(7,4))
plt.bar(df['HAVING CHILDREN BELOW AGE OF 12 ?'].value_counts().index.tolist(),list(df['HAVING CHILDREN BELOW AGE OF 1
```

Out[30]:
```
<BarContainer object of 2 artists>
```

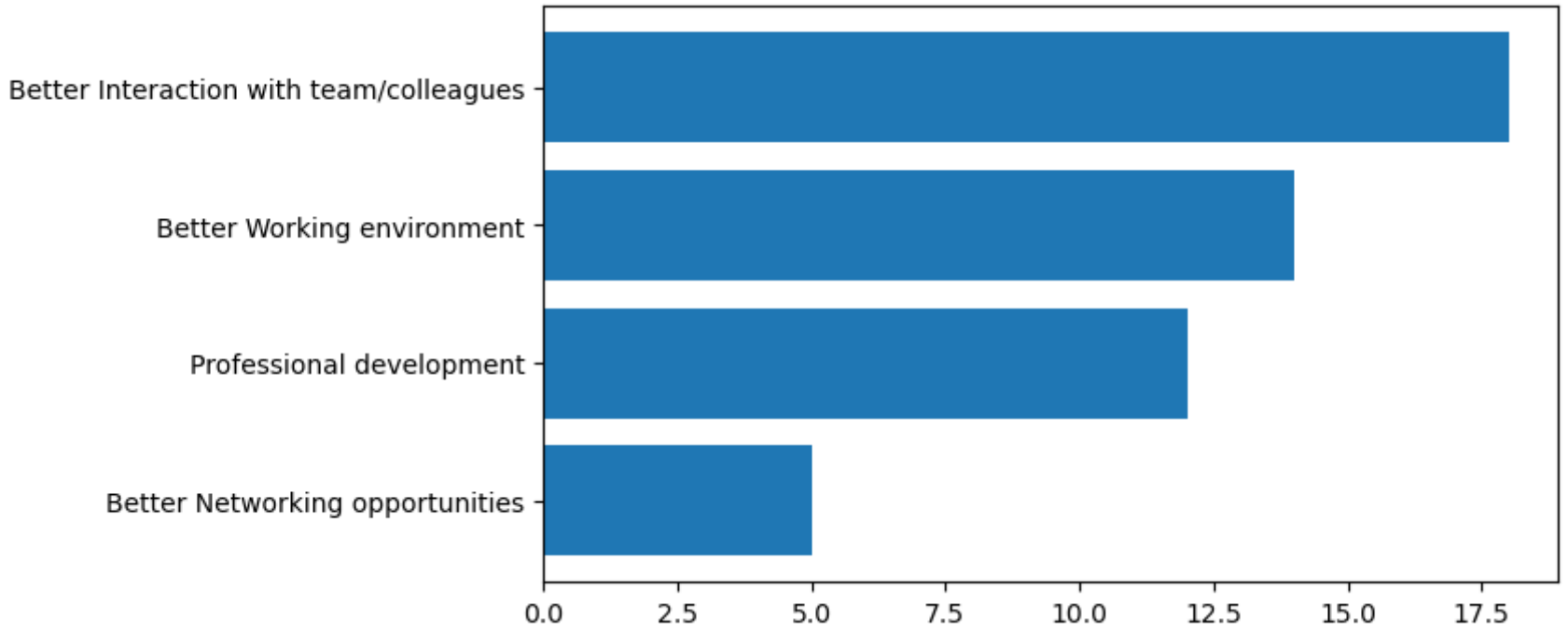
In [31]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='WORK FROM OFFICE']['1st_Choice'].value_counts().index.tolist()[::-1],list(
```

Out[31]: <BarContainer object of 4 artists>



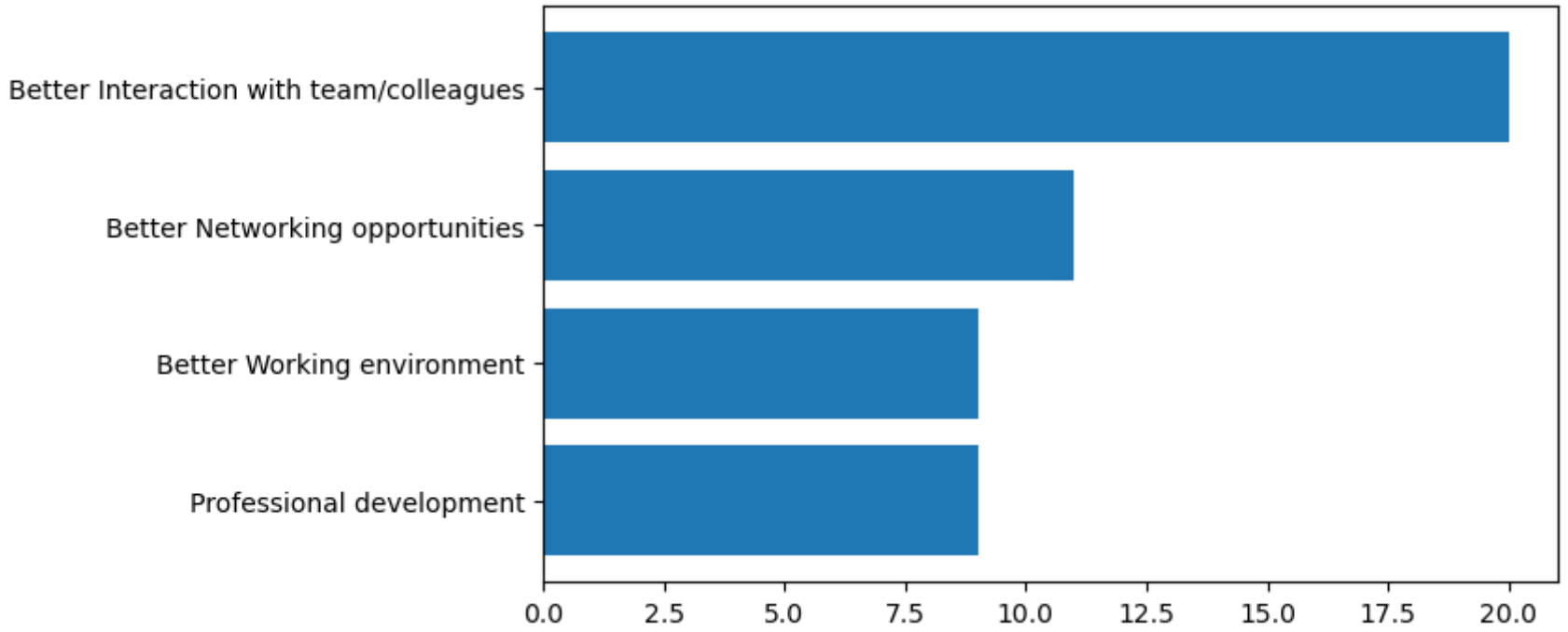
In [32]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='WORK FROM OFFICE']['2nd_Choice'].value_counts().index.tolist()[::-1],list(
```

Out[32]: <BarContainer object of 4 artists>



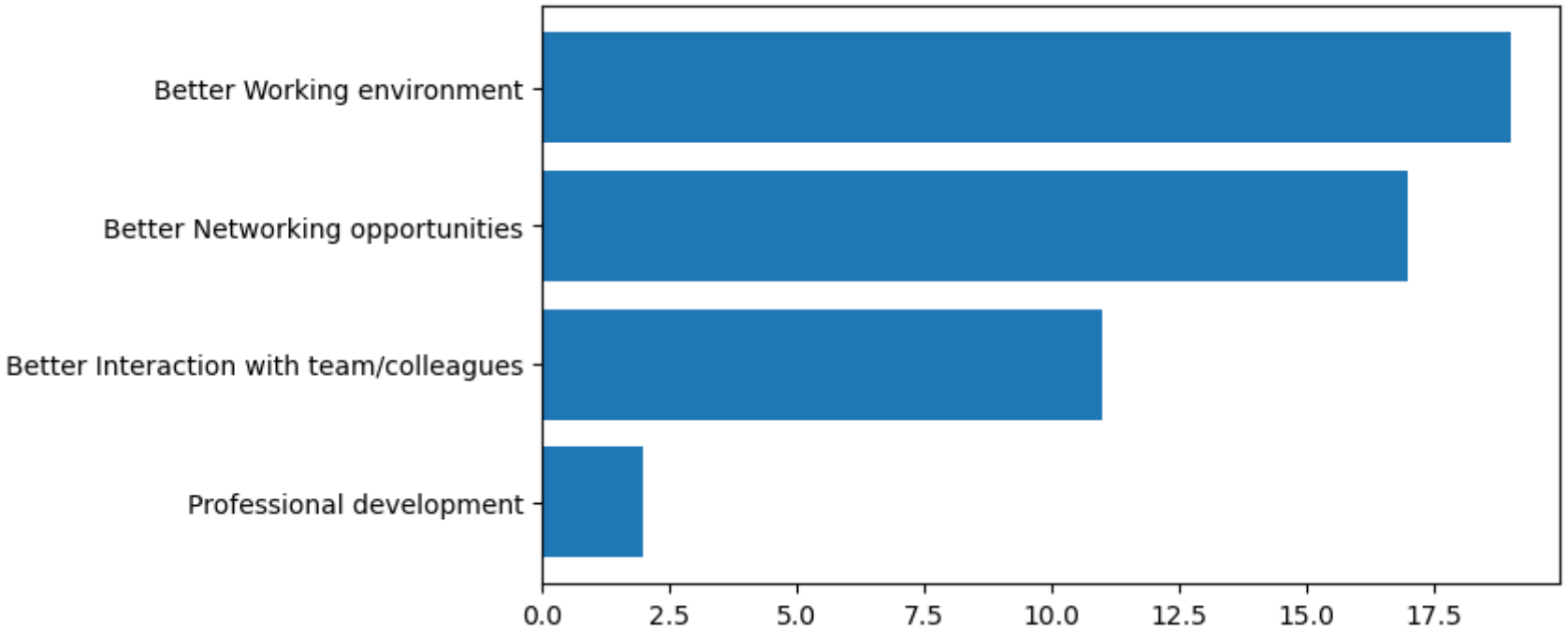
In [33]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='WORK FROM OFFICE']['3rd_Choice'].value_counts().index.tolist()[::-1],list(
```

Out[33]:
```
<BarContainer object of 4 artists>
```



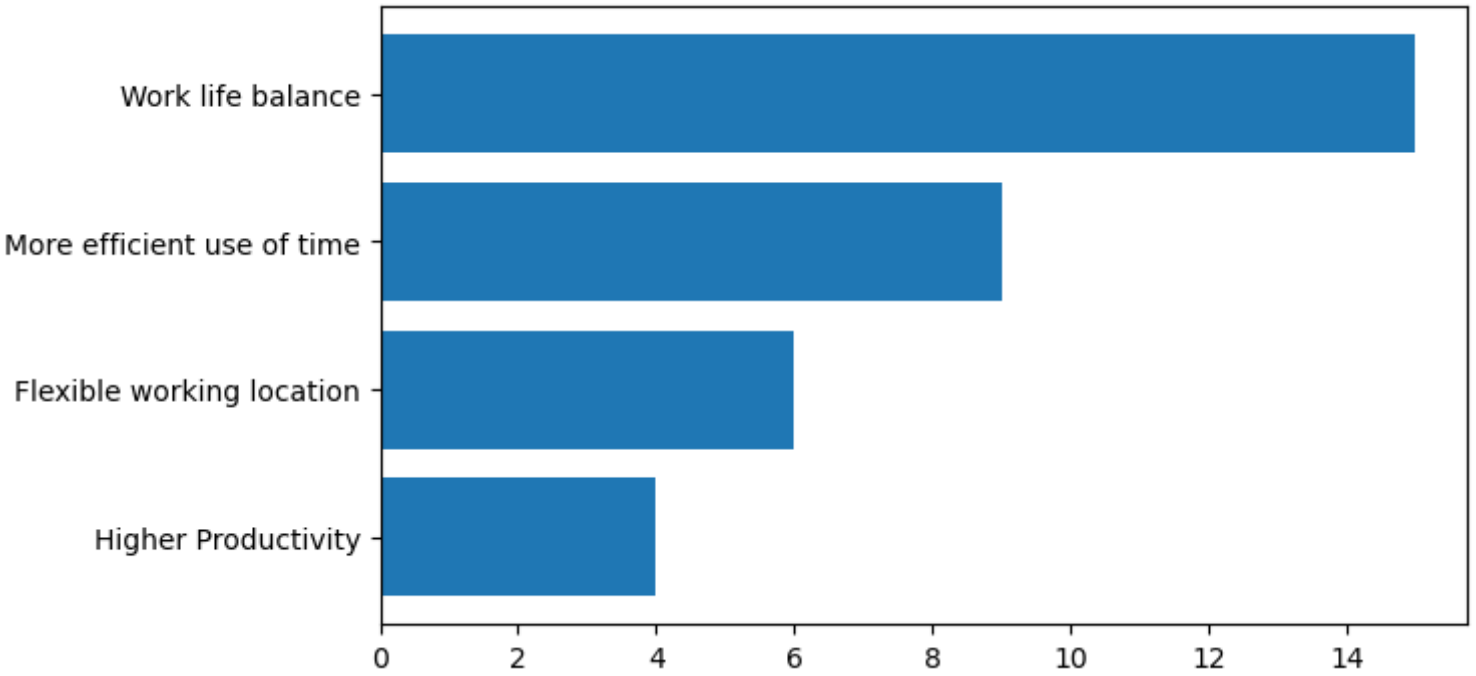
In [34]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='WORK FROM OFFICE']['4th_Choice'].value_counts().index.tolist()[::-1],list(
```

Out[34]: <BarContainer object of 4 artists>



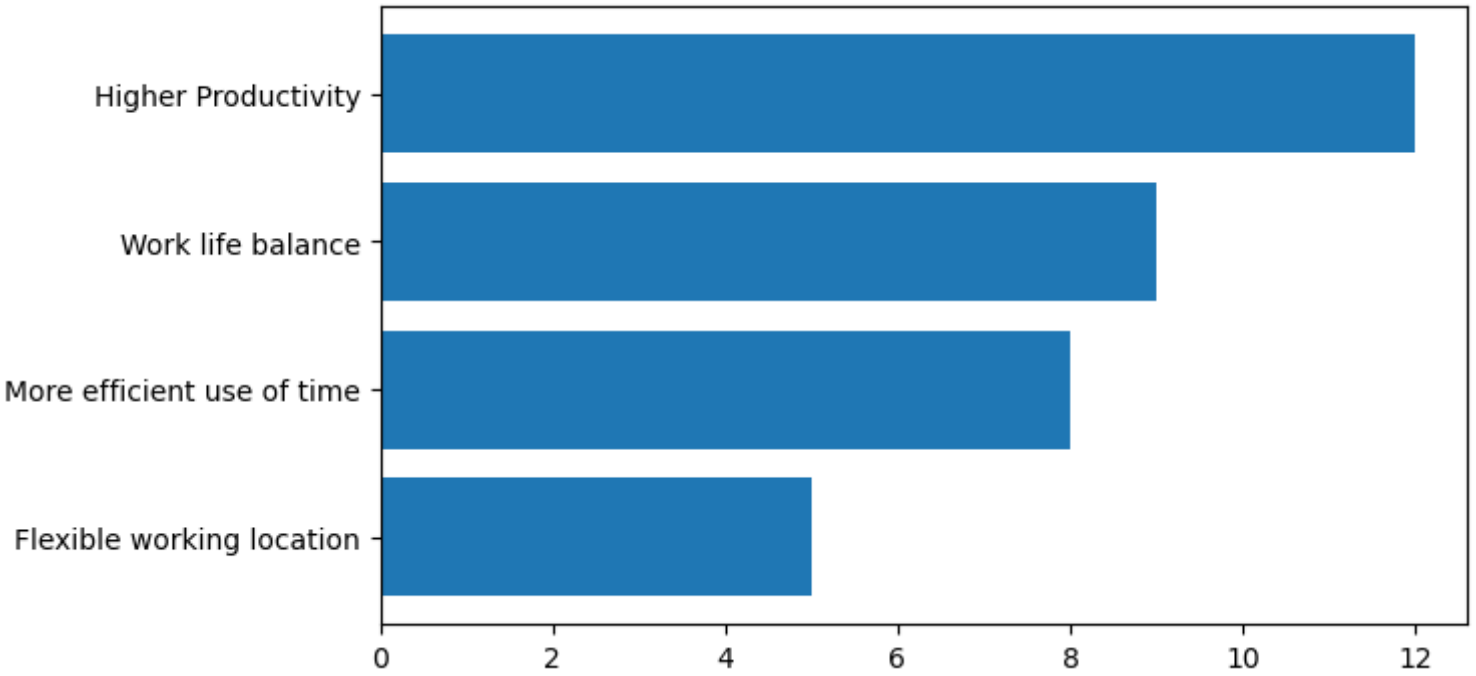
In [35]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='HYBRID']['1st_Choice'].value_counts().index.tolist()[::-1],list(df[df['PRE
```

Out[35]:
```
<BarContainer object of 4 artists>
```



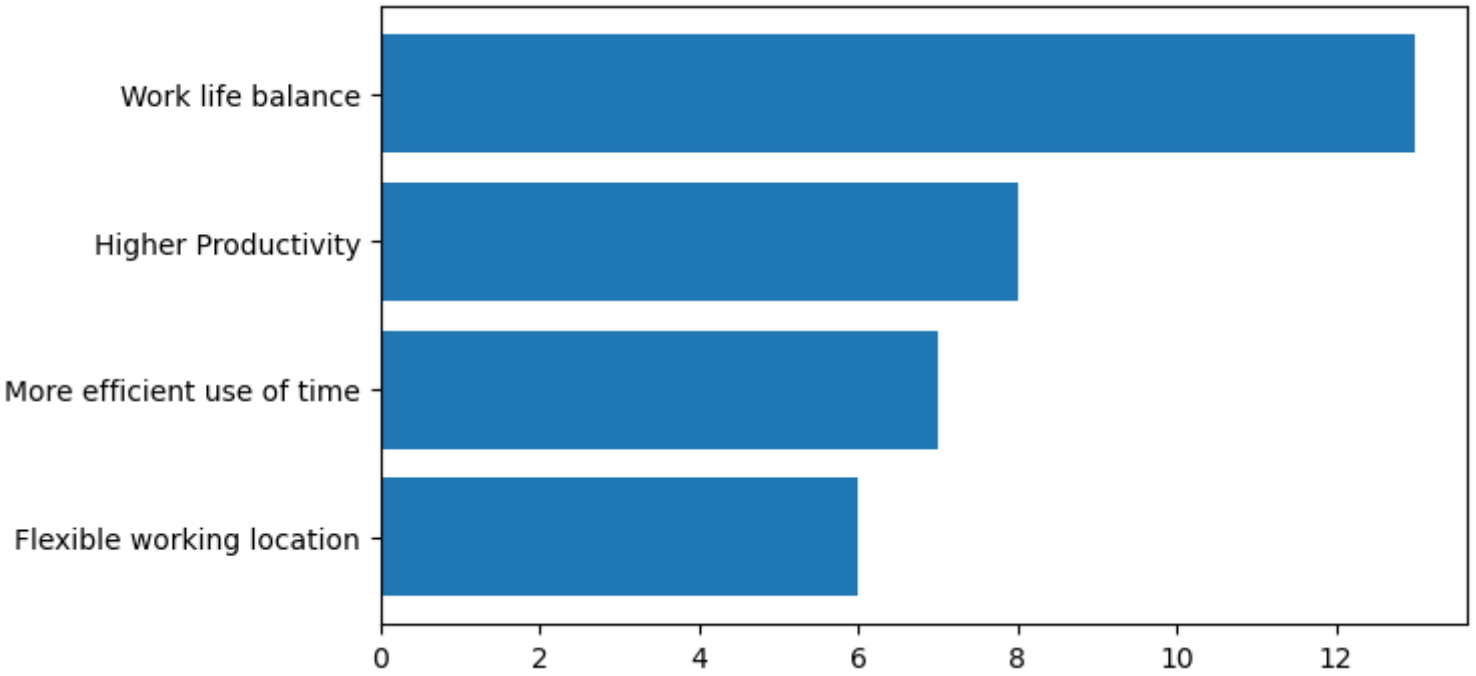
In [36]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='HYBRID']['2nd_Choice'].value_counts().index.tolist()[::-1],list(df[df['PRE
```

Out[36]: <BarContainer object of 4 artists>



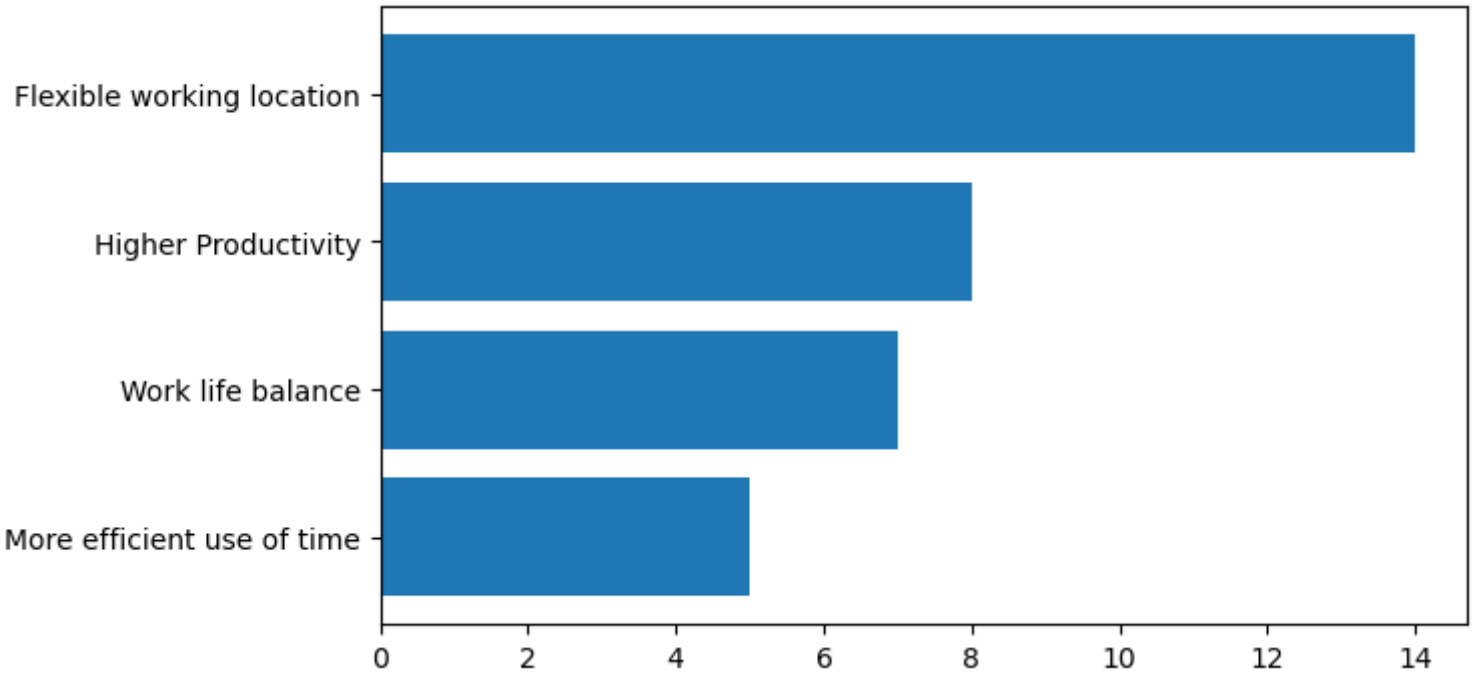
In [37]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='HYBRID']['3rd_Choice'].value_counts().index.tolist()[::-1],list(df[df['PRE
```

Out[37]: <BarContainer object of 4 artists>



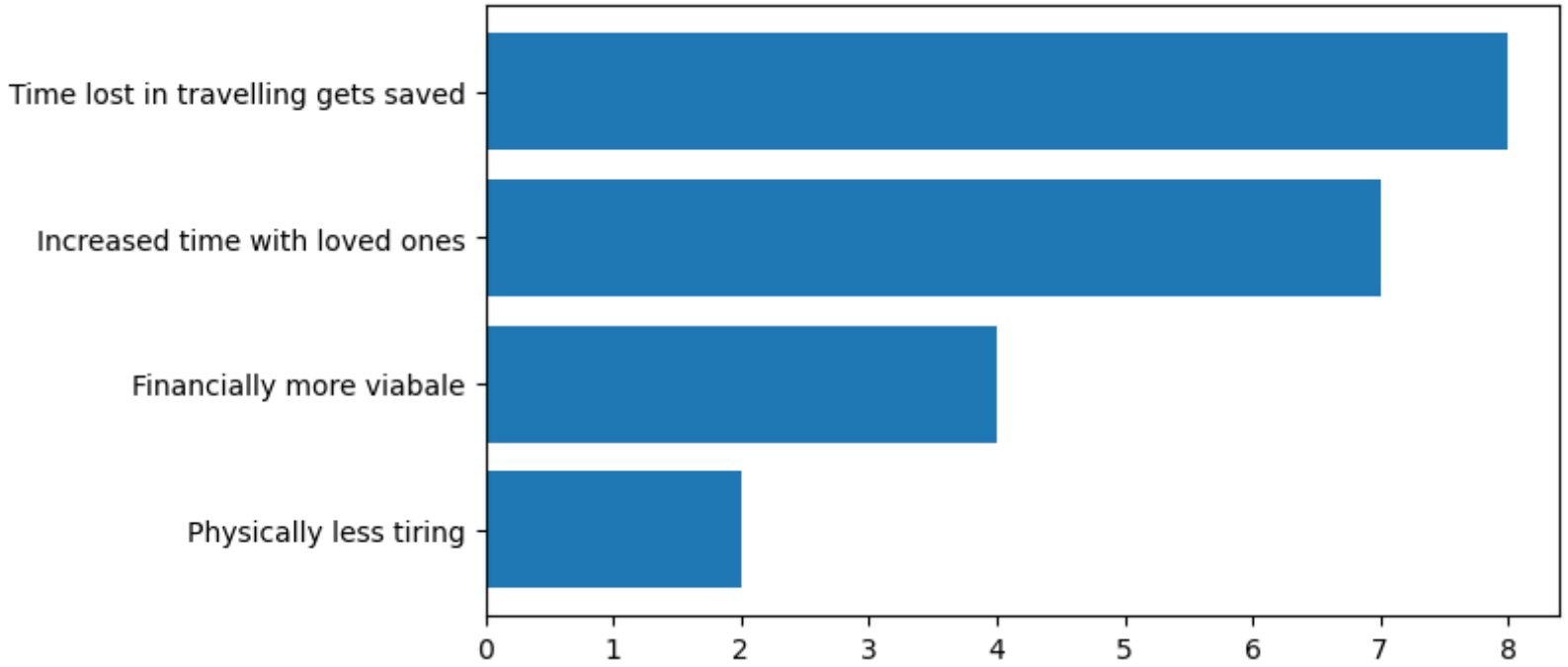
In [38]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='HYBRID']['4th_Choice'].value_counts().index.tolist()[::-1],list(df[df['PRE
```

Out[38]: <BarContainer object of 4 artists>



In [39]:
```
plt.figure(figsize=(7,4))
plt.barh(df[df['PREFERRED WORKING MODE']=='WORK FROM HOME']['1st_Choice'].value_counts().index.tolist()[::-1],list(df
```

Out[39]: <BarContainer object of 4 artists>

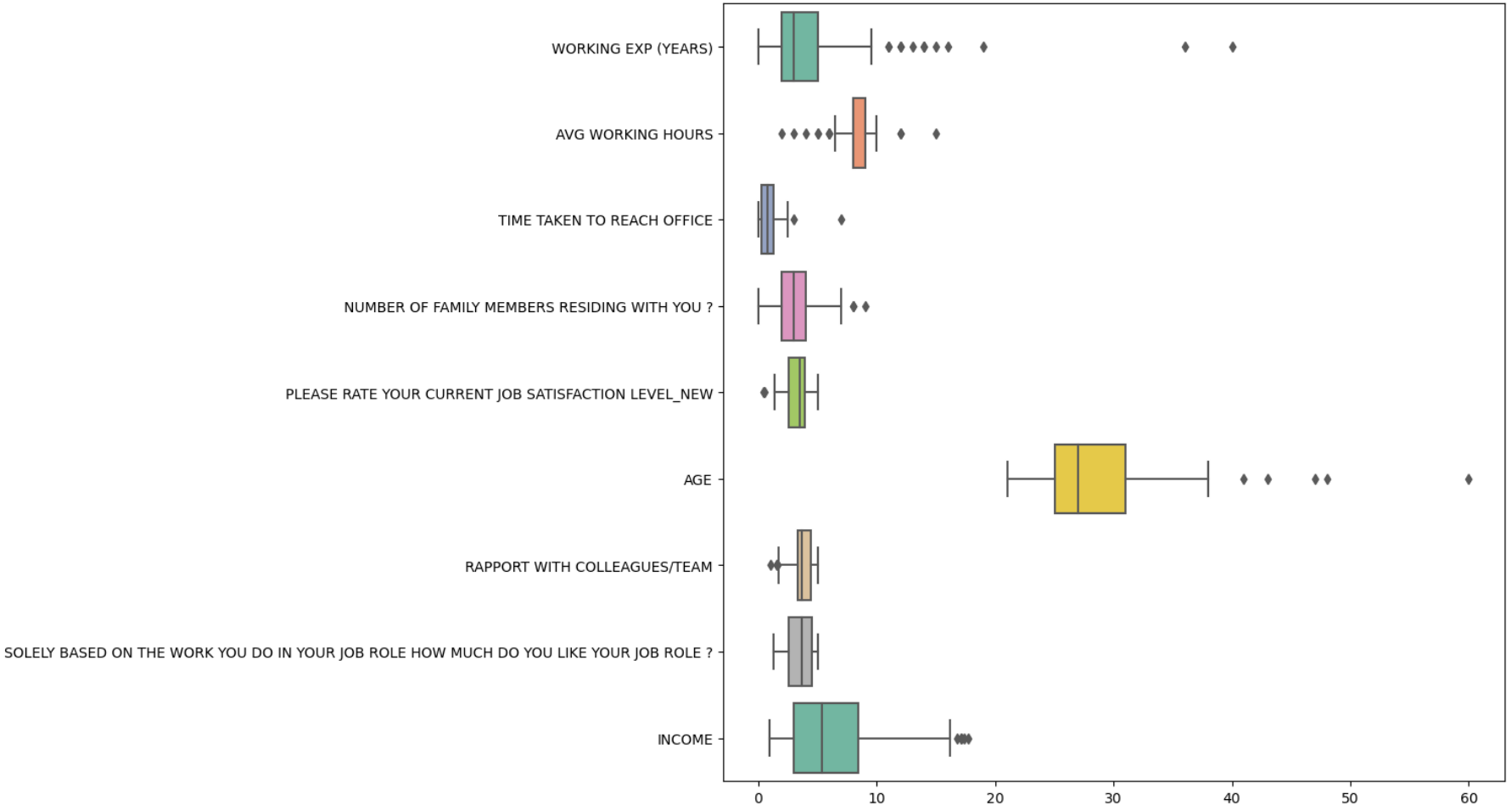



# Treating Outliers

In [40]:
```
list1=df[['WORKING EXP (YEARS)',
 'AVG WORKING HOURS',
 'TIME TAKEN TO REACH OFFICE',
 'NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?',
 'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW','AGE',
 'RAPPORT WITH COLLEAGUES/TEAM',
 'SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?','INCOME']]
```

In [41]:
```
plt.figure(figsize=(10,10))
import seaborn as sns
sns.boxplot(data=list1,orient="h", palette="Set2")
```

Out[41]: <AxesSubplot:>

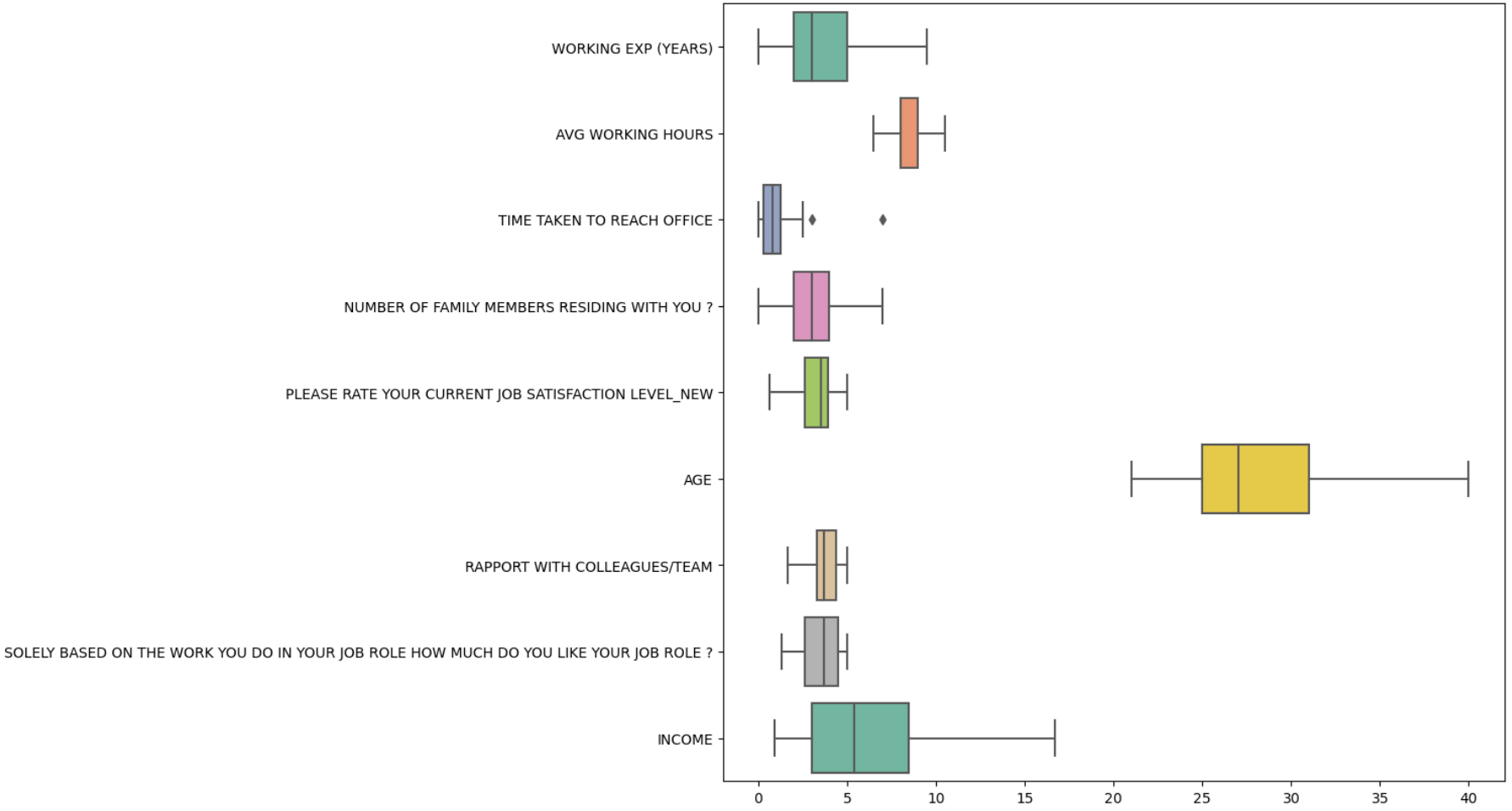


In [42]:
```
new_df_cap = df.copy()
for i in list1:
    q1 = np.percentile(df[i], 25)
    q3 = np.percentile(df[i], 75)
    iqr = q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    new_df_cap[i] = np.where(new_df_cap[i] > upper_bound, upper_bound,np.where(new_df_cap[i] < lower_bound,lower_boun
```

In [43]:
```
list2=new_df_cap[['WORKING EXP (YEARS)',
 'AVG WORKING HOURS',
 'TIME TAKEN TO REACH OFFICE',
 'NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?',
 'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW','AGE',
 'RAPPORT WITH COLLEAGUES/TEAM',
 'SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?','INCOME']]
```

In [44]:
```
plt.figure(figsize=(10,10))
import seaborn as sns
sns.boxplot(data=list2,orient="h", palette="Set2")
```

Out[44]: <AxesSubplot:>




# Bivariate Analysis

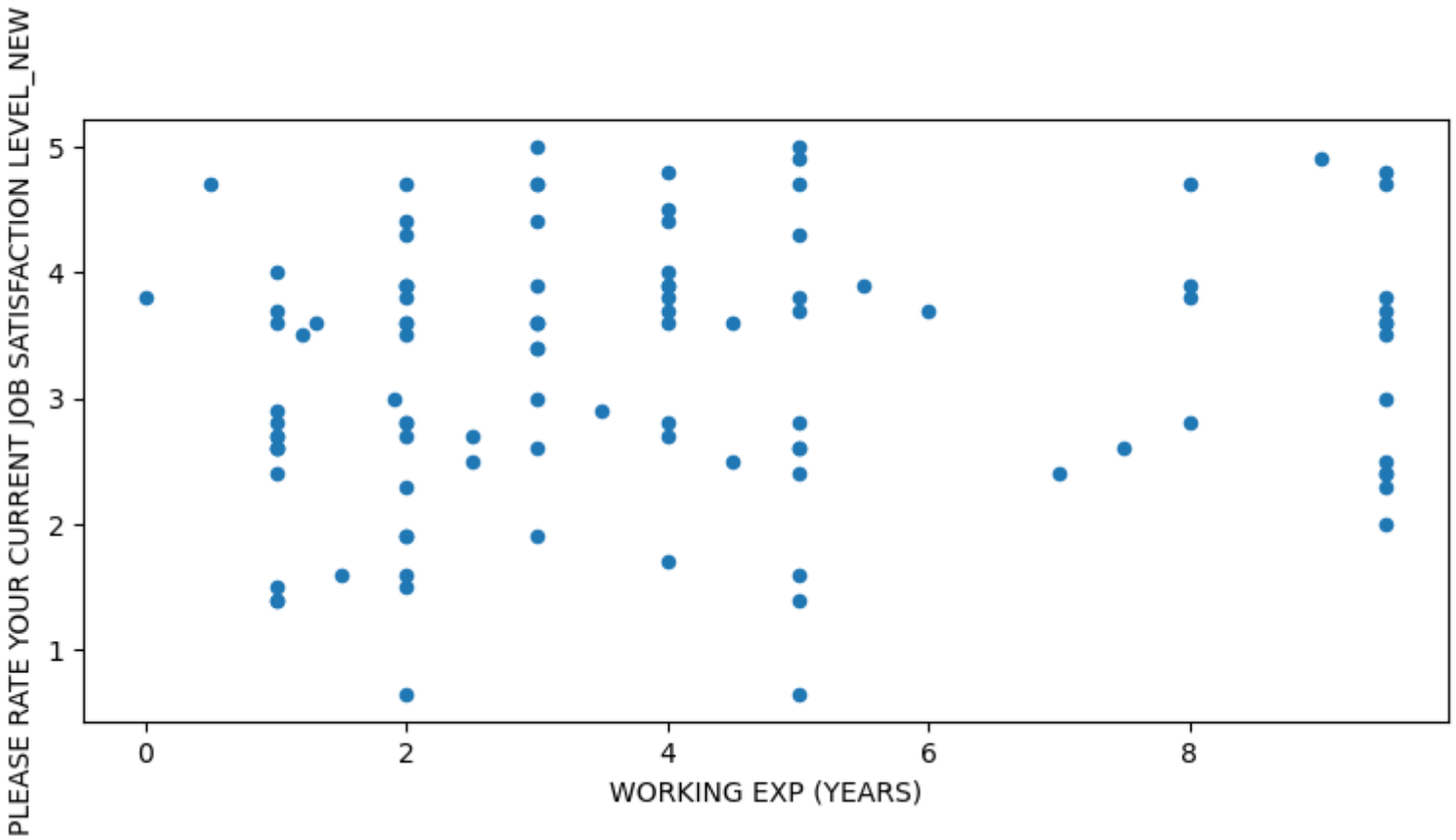
Finding correlation coefficients with the target variable "PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW".

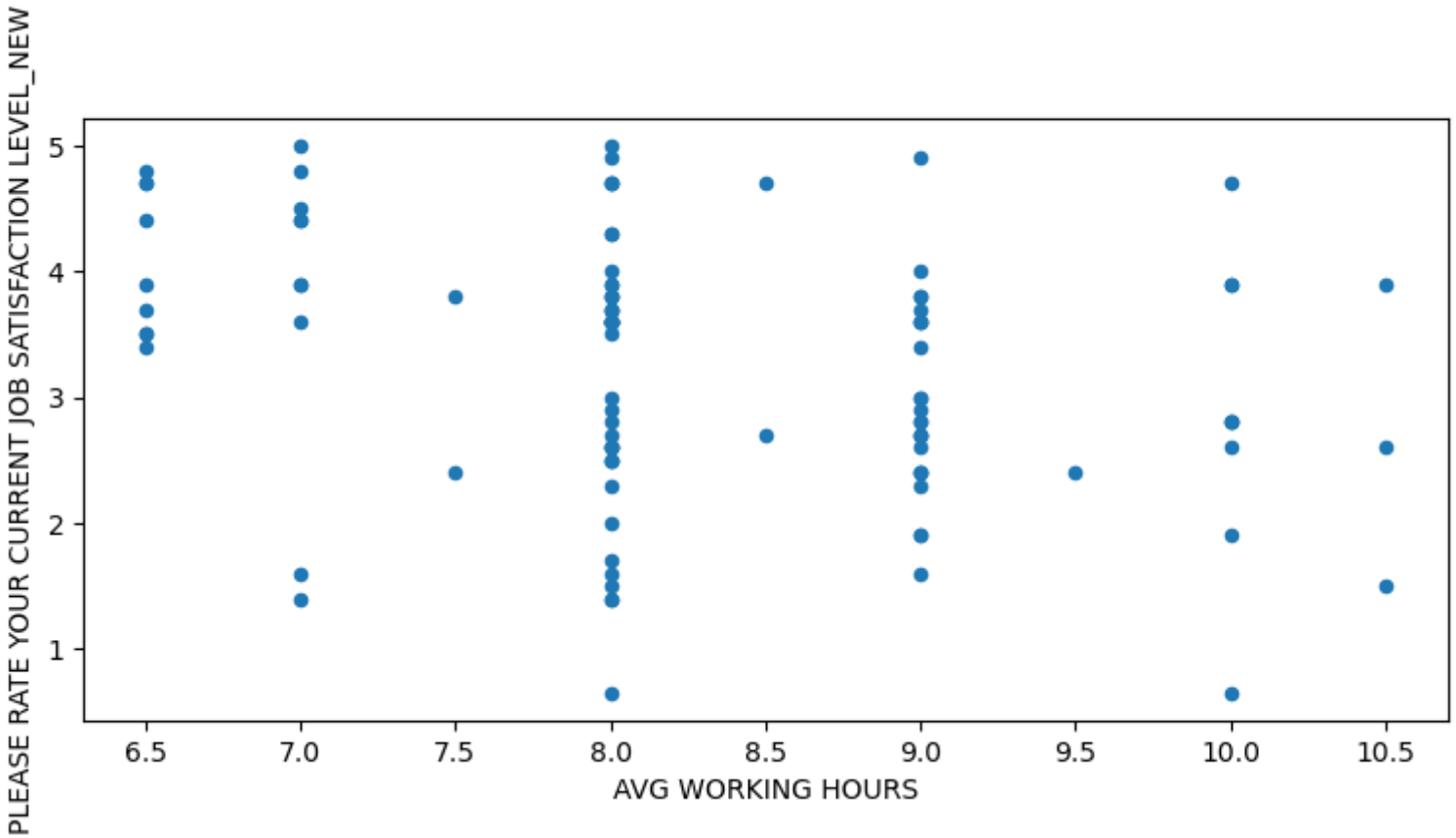
In [45]:
```
pd.DataFrame(new_df_cap.corr()['PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW'])
```

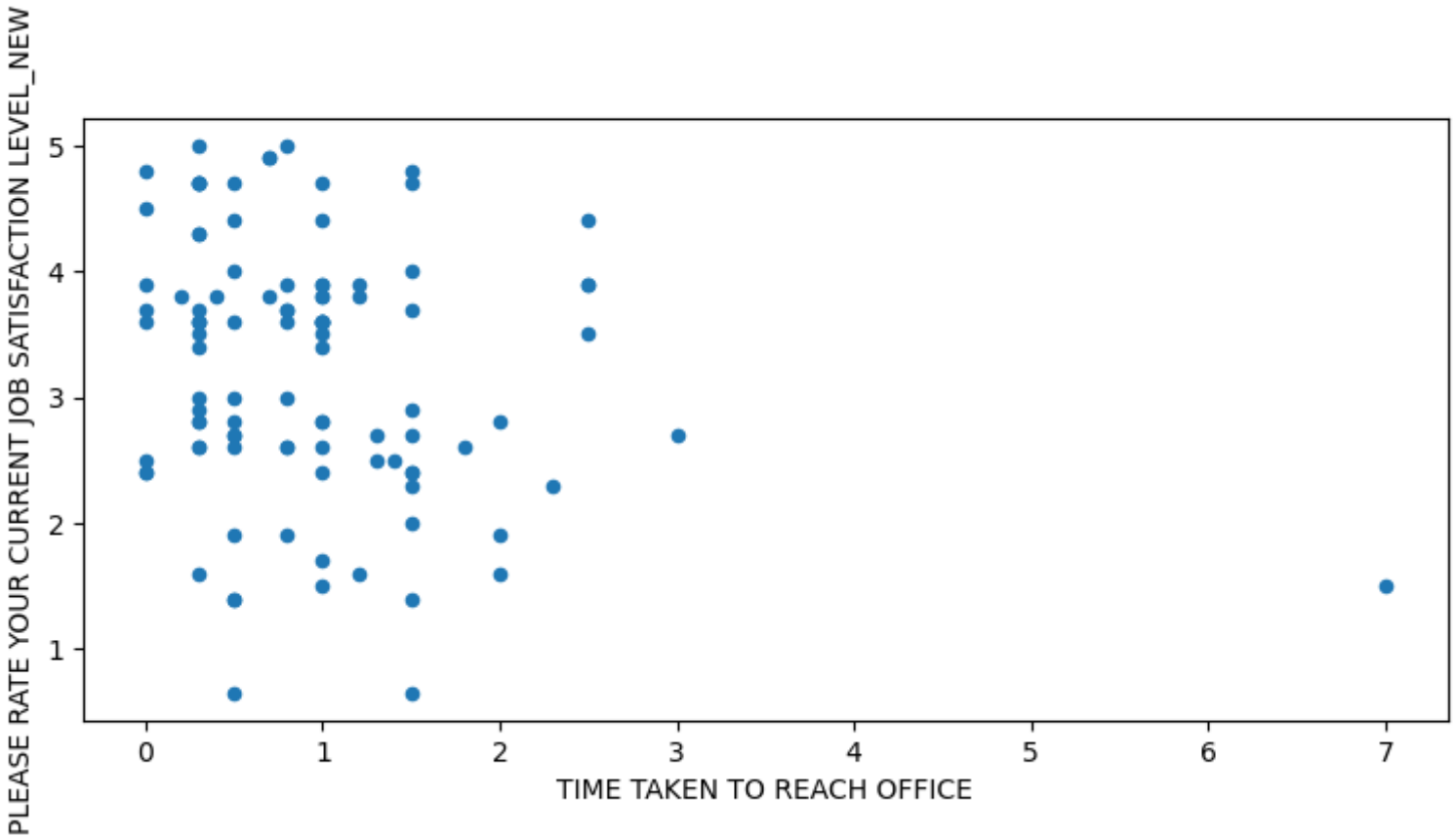
Out[45]:

| | PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW |
|---|---|
| **AGE** | 0.191451 |
| **WORKING EXP (YEARS)** | 0.131042 |
| **AVG WORKING HOURS** | -0.315201 |
| **INCOME** | 0.167951 |
| **TIME TAKEN TO REACH OFFICE** | -0.234898 |
| **RAPPORT WITH COLLEAGUES/TEAM** | 0.409684 |
| **SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?** | 0.647641 |
| **NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?** | 0.103754 |
| **PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW** | 1.000000 |

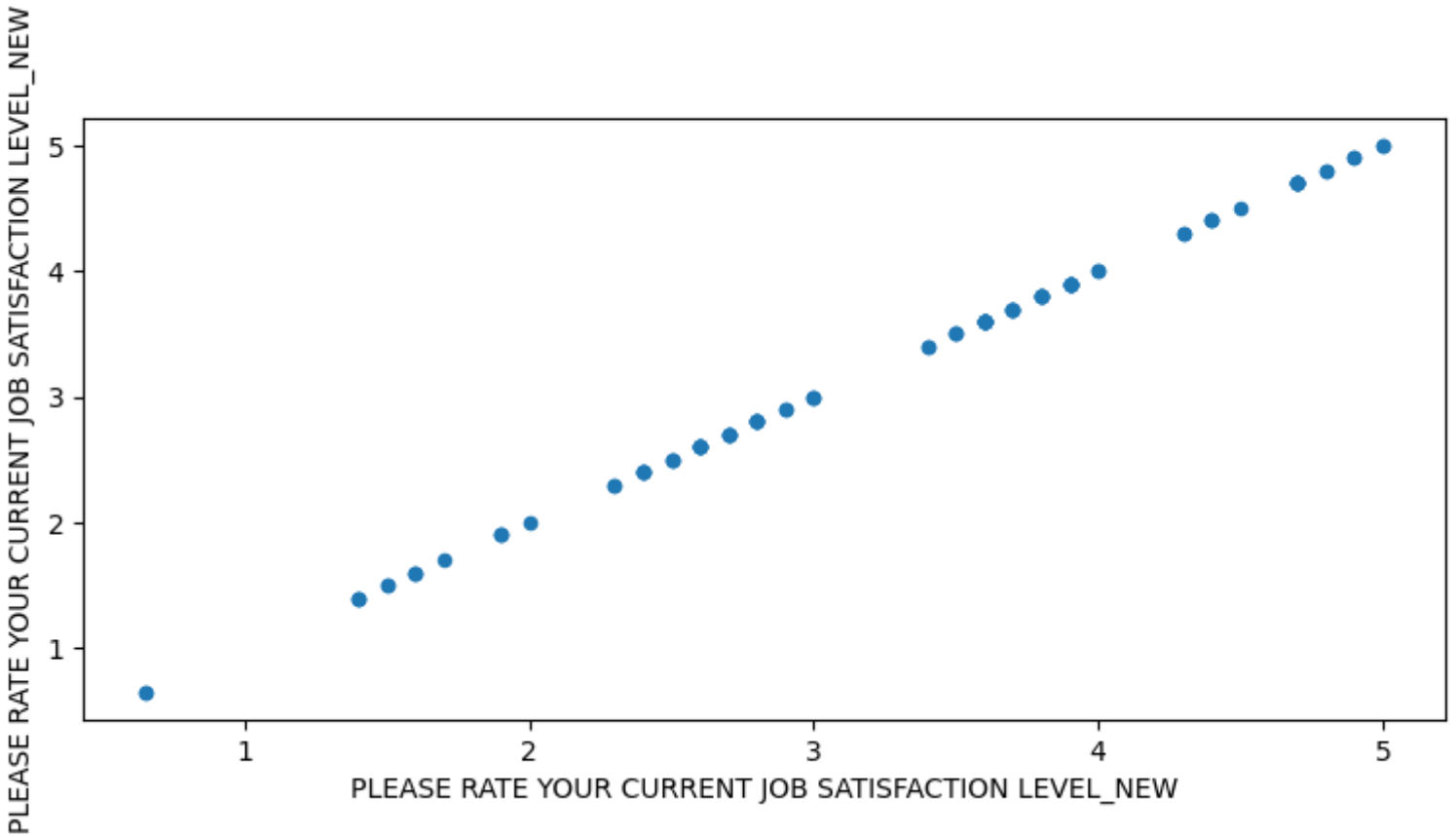
In [46]:
```
for i in list2:
    new_df_cap.plot.scatter(x=i,y='PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW',figsize=(9,4))
```

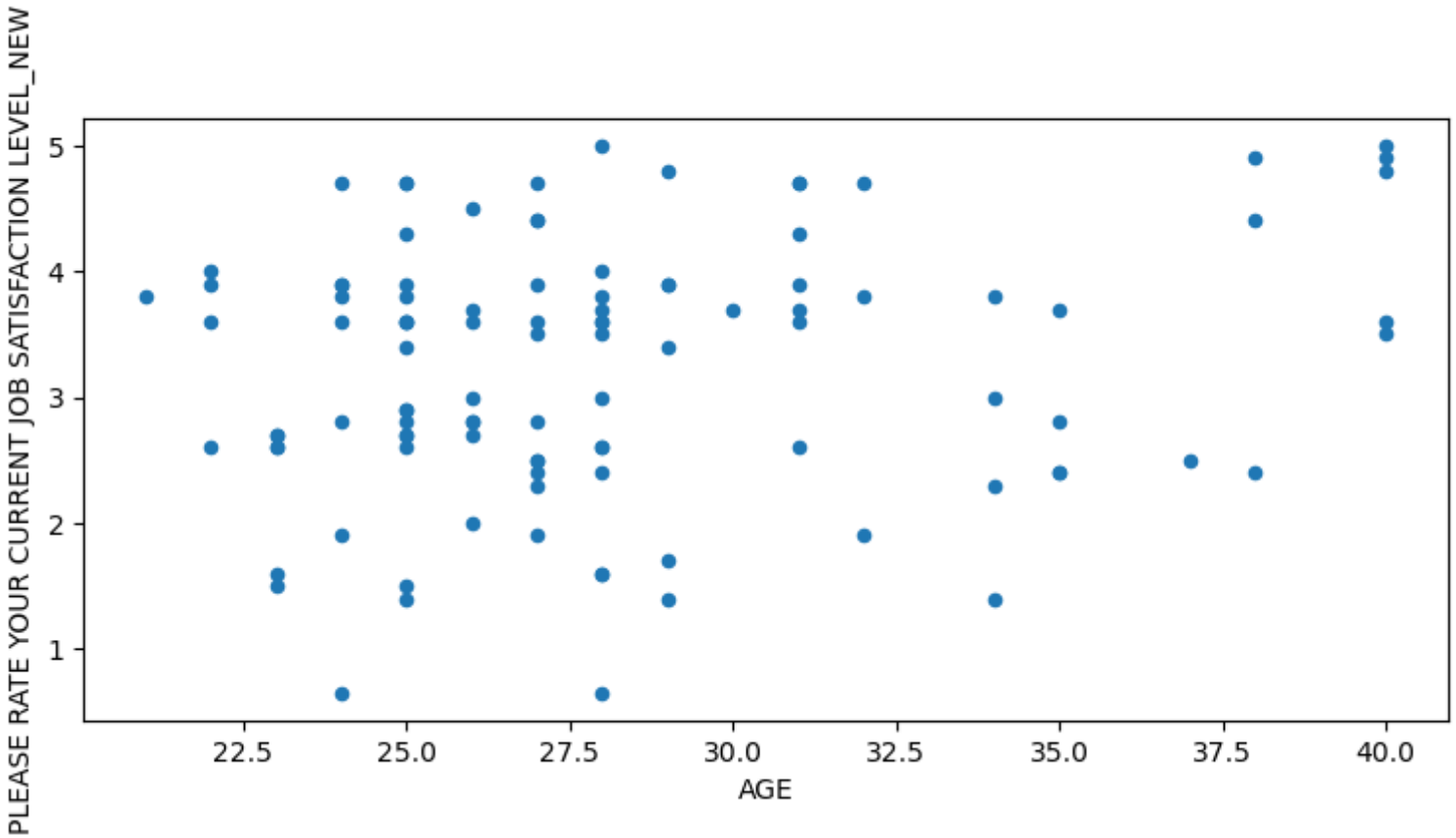


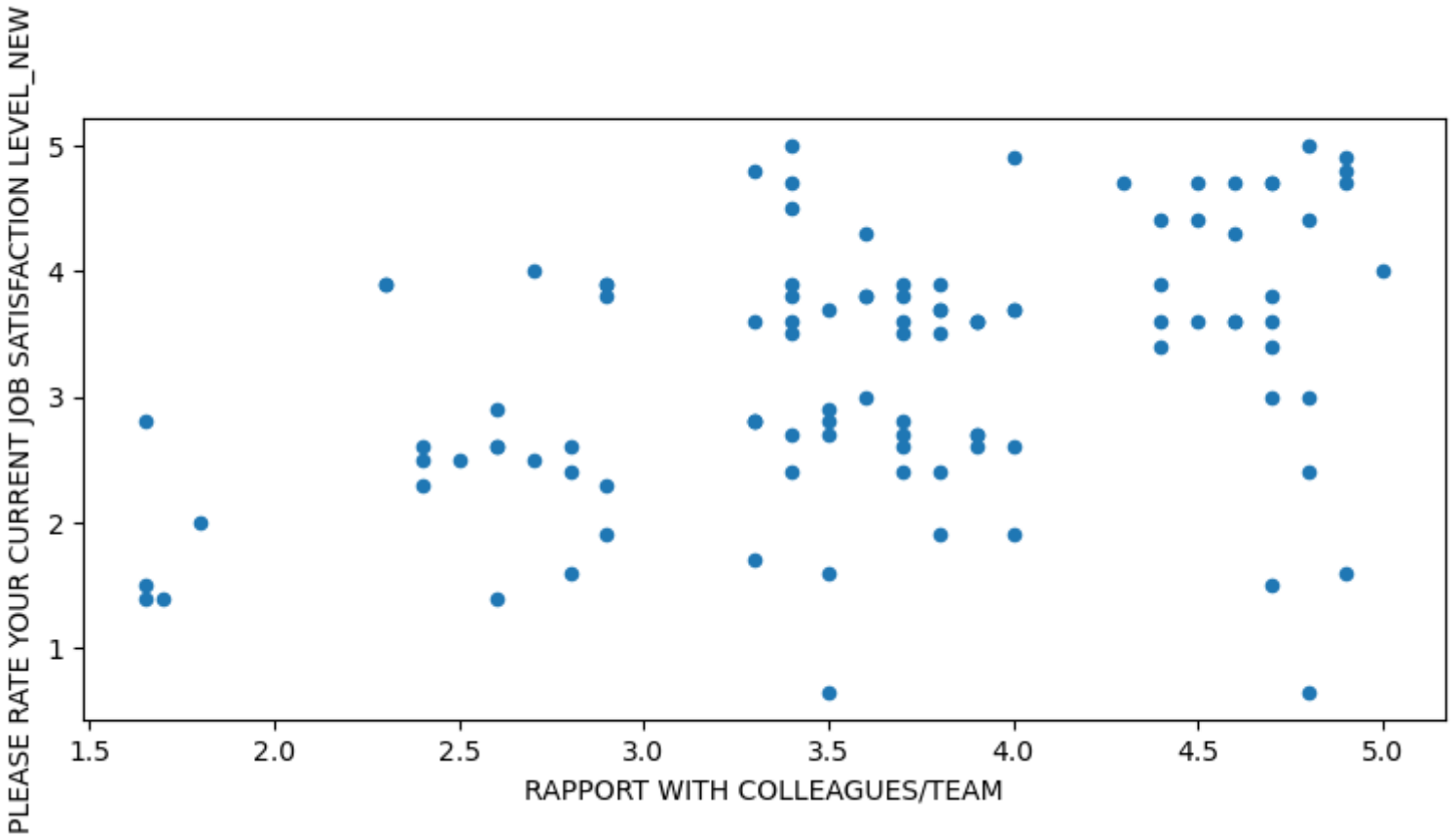

PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW
5
4
3
2
1
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
10.5
AVG WORKING HOURS

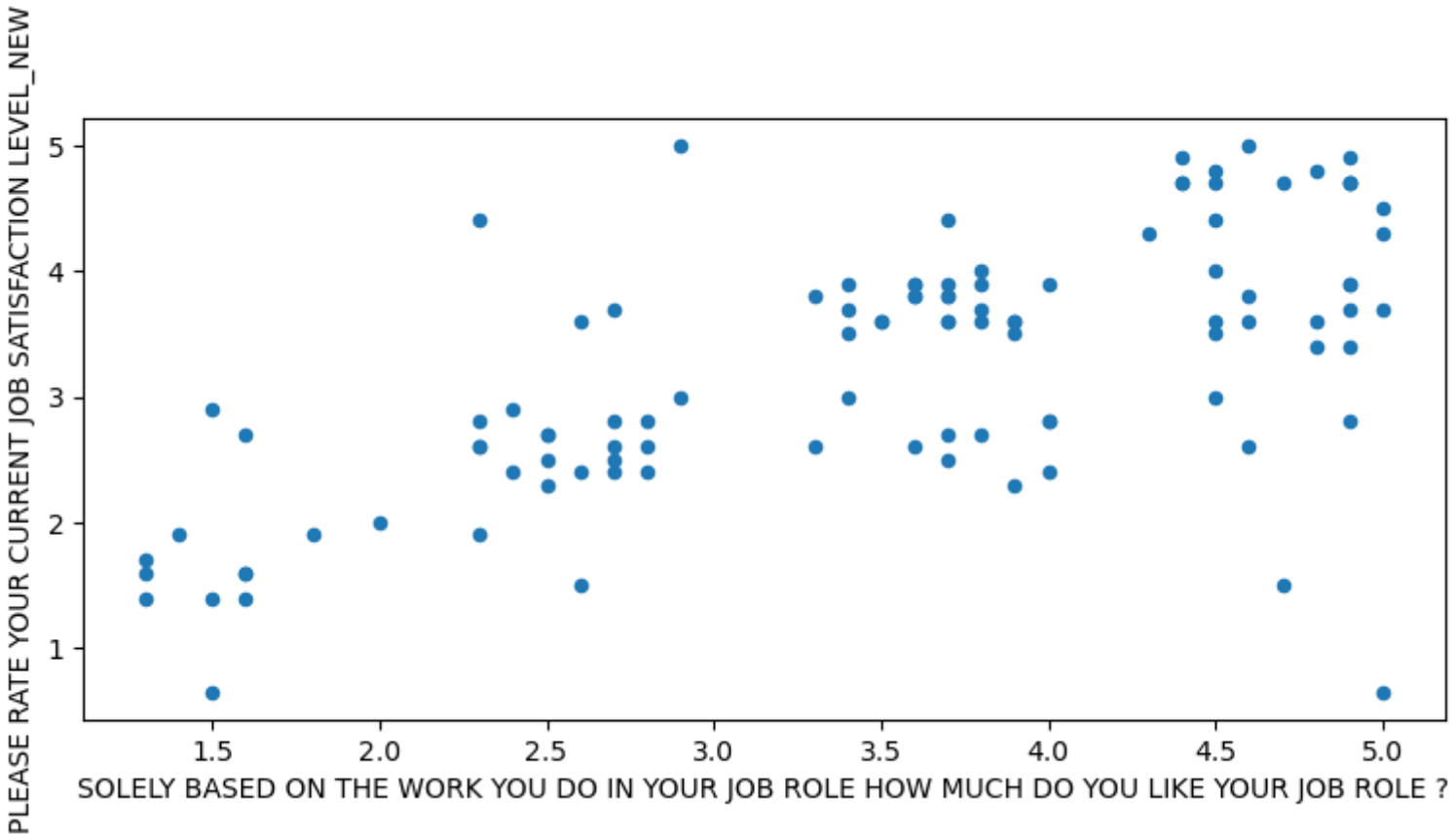

PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW
TIME TAKEN TO REACH OFFICE
1
2
3
4
5
0
1
2
3
4
5
6
7

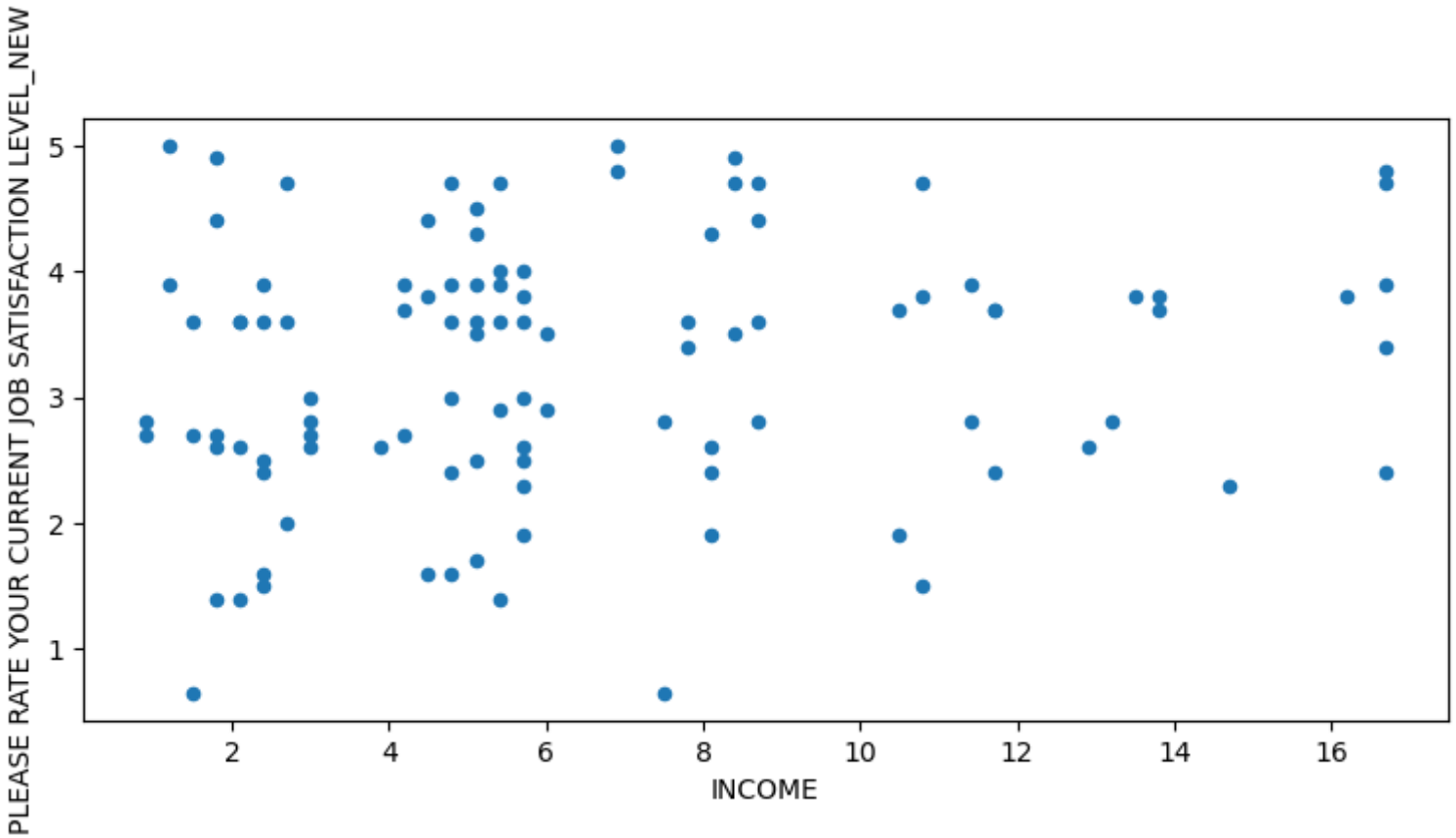

PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW
NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?
1
2
3
4
5
0
1
2
3
4
5
6
7


PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW
1
2
3
4
5
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW


PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW
1
2
3
4
5
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0
AGE


PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW
RAPPORT WITH COLLEAGUES/TEAM
1
2
3
4
5
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW
1
2
3
4
5
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?




# Hypothesis Testing

## 1) Independent sample t-test

H0: Mean Job satisfaction level of Female <= Mean Job satisfaction level of Male.
H1: Mean Job satisfaction level of Female > Mean Job satisfaction level of Male.

In [47]:
```
from scipy import stats

group1 = new_df_cap[new_df_cap['GENDER']=='FEMALE']['PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW']
group2 = new_df_cap[new_df_cap['GENDER']=='MALE']['PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW']

t_statistic, p_value = stats.ttest_ind(group1, group2, equal_var=False, alternative='greater')

print("T-statistic:", t_statistic)
print("P-value:", p_value)
```

```
T-statistic: 0.6219735065417906
P-value: 0.26796657744128805
```

From the above p-value of hypothesis testing, we are not able to reject Null Hypothesis at 0.05 significance level and conclude that the mean job satisfaction level of women are less than or equal to mean job satisfaction level of male.

## 2) One-way ANOVA

H0: Mean Job satisfaction level of people working from Home = Mean Job satisfaction level of people working from Office = Mean Job satisfaction level of people having Hybrid working mode.
H1: Atleast one pair of group have significant difference in mean job satisfaction level.

In [48]:
```
group1 = new_df_cap[new_df_cap['CURRENT WORKING MODE']=='WORK FROM HOME']['PLEASE RATE YOUR CURRENT JOB SATISFACTION
group2 = new_df_cap[new_df_cap['CURRENT WORKING MODE']=='WORK FROM OFFICE']['PLEASE RATE YOUR CURRENT JOB SATISFACTIO
group3 = new_df_cap[new_df_cap['CURRENT WORKING MODE']=='HYBRID']['PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NE

data = [group1, group2, group3]

f_statistic, p_value = stats.f_oneway(*data)

print("F-statistic:", f_statistic)
print("P-value:", p_value)
```

```
F-statistic: 0.514494948239316
P-value: 0.5993606519291936
```

From the above p-value of hypothesis testing, we are not able to reject Null Hypothesis at 0.05 significance level and conclude that there is no significant difference in mean job satisfaction level between any pair of current working mode group.

## 3) Chi-square test of independence

H0: There is no association between Gender and Preferred Working Mode.
H1: There is association between Gender and Preferred Working Mode.

In [49]:
```
category1 = new_df_cap['GENDER']
category2 = new_df_cap['PREFERRED WORKING MODE']

contingency_table = pd.crosstab(category1, category2)

print(contingency_table,"\n")

# Perform Chi-square test of independence
chi2_statistic, p_value, dof, expected = stats.chi2_contingency(contingency_table)

# Print the results
print("Chi-square statistic:", chi2_statistic,"\n")
print("P-value:", p_value,"\n")
print("Degrees of freedom:", dof,"\n")
print("Expected frequencies:\n", expected)
```

```
PREFERRED WORKING MODE  HYBRID  WORK FROM HOME  WORK FROM OFFICE
GENDER
FEMALE                      13               8                20
MALE                        21              13                29

Chi-square statistic: 0.07541909524165817

P-value: 0.9629926037677479

Degrees of freedom: 2

Expected frequencies:
 [[13.40384615  8.27884615 19.31730769]
 [20.59615385 12.72115385 29.68269231]]
```

From the above p-value of hypothesis testing, we are not able to reject Null Hypothesis at 0.05 significance level and conclude that there is no association between Gender and Preferred Working Mode.

## Few More Insights

***1> Most prefereed working mode of people is "work from office" and least preffered mode of people is "work from home".***

***2> Top two reasons with people having preffered working modes "Work form office" are "Better working enviornment" and "Better interaction with teams/ collegues"***

***3> Top two reasons with peolple having preffered working mode as "work form hybrid" are "work life balance" and "Higher productivity".***

***4> Top two reasons with people having preffered working mode as "Work from home" are "Time lost in travelling gets used" and "Financially more viable".***

***5> Top three factors that influnce the job satisfaction level of an individual are-***

i) Job role of an individual
ii) Rappart with colleauges.
iii) Avg. Working Hours.

In [ ]: