BD3P3 ML-II Lab Test

20BDA21 - Joel Bharat Monis

20BDA23 - Manu Tom

20BDA35 - Nidhi Teresa George

Background

The COVID-19 pandemic is among the deadliest infectious diseases to have emerged in recent history. As with all past pandemics, the specific mechanism of its emergence in humans remains unknown. Following a surge in cases of COVID-19 in April 2020, India became the third-worst affected country worldwide. This necessitated countrywide lockdown and resulted in mass exodus of migrants. Numerous professionals who relied on IT backed infrastructure and procedures for their work were all ensconced in their houses and Work From Home (WFH) began for them. As the lockdown dragged on, several researchers and organisations took this opportunity to assess the effect and efficacy of the lockdown on work related issues. Several papers have been published, which study the effect of WFH.

However, WFH is not a new phenomenon. It is attributed to a NASA scientist in the early 1973, who coined the term telecommuting to avoid the Los Angeles traffic. The idea of telecommuting was very different from what it is today, especially when there was no internet. Since then telecommuting has come a long way. The phone conferences and the, once novel, video calls have become the norm and permeate every aspect of professional sphere of almost every large company worth its salt – not to mention the myriad smaller concerns who have adopted this technology, much to their benefit. The constant technology based advancement of long distance office communication has made WFH a run-of-the-mill convenience adopted by numerous professionals and households. Availability of Internet was never appreciated more.

However, WFH has not been a smooth transition for everybody. Some have loved the transition, enjoyed the time with family, benefited from savings incurred due to lesser travel and generally been more productive. Some others have found the transition challenging, due to work cultures of groups they work with, assumption by the bosses that they are always available at their beck and call for work – now that they are at home etc.

There are several papers which bring out the pros and cons of WFH. Several of them are listed in the references.

In 2010-11, Ctrip, the largest travel agency in China experimented with the concept of work from home with a set of employees for a period of 9 months, under the watchful eye of Stanford University researchers. The research led to the conclusion that there was a general increase of productivity of WFH workers to the tune of 13%. On completion of the experiment, WFH was rolled out to all employees which led to the gains from WFH to rise to 22%. This was a scientific approach to ascertaining the efficacy of WFH. However, the company deliberately tried out this concept without any external pressure, which gave them enough time to prepare and orient the volunteers to the new regimen.

This luxury was not available when Covid-19 hit. In just a few weeks thousands of IT employees were isolated in their houses/homes in an unprecedented way. There were challenges faced by both the employees and the management to structure the new arrangement to make it efficient., There is another paper which brings out several hurdles faced by the individuals while trying to adjust to the new routine. Microsoft has had a favourable experience with WFH. Being an IT behemoth, it has had the advantage of researching the state of its employees and has instituted changes every week to make the experience better for all the affected professionals.

Comprehensively, the papers bring out a mixed bag of findings which has been used to tweak the WFH system to reduce the pressure on the employees and compensate for the social disconnect that everyone experienced.

Data

The data used by many companies is extensive and has been processed over a span of several weeks during which their corrective measures were also implemented. The data used in this study has been gleaned from the data and observations available online.

Problem Statement

Work from home: employee productivity and preference?

How working from home affects employee productivity?

What do employees prefer and why?

Task

Task

- To ascertain the factors which affect employee productivity and preference for WFH.
- To ascertain whether those working from home experienced an increase in productivity.
- To ascertain whether the employees working from home prefer the arrangement.

Hypothesis

Null Hypothesis: The outcome variable is independent of the independent variables.

Alternate Hypothesis: The outcome variable is not independent of the independent variables.

Approach

- Procure data (Employee Productivity and Preference Data)
- Carry out EDA
- Test the hypothesis. Carry out statistical tests to ascertain the features which affect the outcome variable
- Build Machine Learning Pipeline

Importing packages relevant for analysis and modelling

In [1]:

```
import numpy as np # linear algebra
import pandas as pd # data processing
import matplotlib.pyplot as plt # for data visualization purposes
import seaborn as sns # for statistical data visualization
import sklearn.preprocessing as pre
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report
from sklearn.feature_selection import chi2
import scipy.stats as stats
from scipy.stats import chi2_contingency
%matplotlib inline
```

I - Productivity

Data

Dataset, Data Description and Data information

In [2]:

```
pd.set_option('display.max_columns', None)
productivity = pd.read_csv("/Users/manuair/Study/Productivity.csv")
productivity.head()
```

Out[2]:

	Gender	Age	Marital Status	Children	No of Children	Experience (in months)	Efficacy of team meetings	Social interaction	Overwork	Stress	Motivation	Innovative Climate	Work Environment	No of working hrs per week
0	Female	23	Single	N	0	18	4	3	1	2	4	5	3	45
1	Male	24	Single	N	0	19	4	3	1	2	5	4	3	44
2	Male	24	Single	N	0	17	4	4	1	3	3	4	4	44
3	Male	22	Single	N	0	17	2	3	3	5	2	2	2	48
4	Male	25	Married	Y	1	24	4	5	3	1	4	3	5	40

Data Description

- Gender - Male/Female
- Age

- Marital Status - Whether married or not - Single / Married
- Children - Whether the individual has children or not - Y/N
- No of children
- Experience (in months) - Work experience of the individual in months
- Efficacy of team meetings - The individual's perception of the efficacy of the meetings attended by him/her - expressed on a Likert scale from 1 to 5
- Social Interaction - The individual's perception of social interaction while working from home - expressed on a Likert scale from 1 to 5
- Overwork - The individual's perception of overwork during WFH - expressed on a Likert scale from 1 to 5
- Stress - The individual's perception of stress during WFH - expressed on a Likert scale from 1 to 5
- Motivation - The individual's perception of own motivation - expressed on a Likert scale from 1 to 5
- Innovative Climate - The individual's perception of Innovative Climate in the company - expressed on a Likert scale from 1 to 5
- Work Environment - The individual's perception of Work Environment in the company - expressed on a Likert scale from 1 to 5
- No of working hrs per week - The no of working hrs put in by the individual, per week
- Uninterrupted Working hrs per week - The no of uninterrupted Working hrs per week
- Number of stretches of uninterrupted working (min 1 hr) - The number of stretches of uninterrupted working of minimum 1 hr durations.
- Average no of meetings attended per week - The average no of meetings attended per week by the individual
- Hrs spent in meetings per week - The number of hours spent in meetings per week by the individual
- Salary (Monthly) - The monthly salary
- EP Outcome - What is the preference of the individual towards own productivity - 0(not as productive)/1(as much or more productive)

In [3]:

```
productivity.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 20 columns):
 #   Column                                                  Non-Null Count  Dtype 
---  ------                                                  --------------  ----- 
 0   Gender                                                  250 non-null    object
 1   Age                                                     250 non-null    int64 
 2   Marital Status                                          250 non-null    object
 3   Children                                                250 non-null    object
 4   No of Children                                          250 non-null    int64 
 5   Experience (in months)                                  250 non-null    int64 
 6   Efficacy of team meetings                               250 non-null    int64 
 7   Social interaction                                      250 non-null    int64 
 8   Overwork                                                250 non-null    int64 
 9   Stress                                                  250 non-null    int64 
 10  Motivation                                              250 non-null    int64 
 11  Innovative Climate                                      250 non-null    int64 
 12  Work Environment                                        250 non-null    int64 
 13  No of working hrs per week                              250 non-null    int64 
 14  Uninterrupted Working hrs per week                      250 non-null    int64 
 15  Number of stretches of uninterrupted working (min 1 hr)  250 non-null    int64 
 16  Average no of meetings attended per week                250 non-null    int64 
 17  Hrs spent in meetings per week                          250 non-null    int64 
 18  Salary (Monthly)                                        250 non-null    int64 
 19  EP Outcome                                              250 non-null    int64 
dtypes: int64(17), object(3)
memory usage: 39.2+ KB
```

In [4]:

```
productivity.isnull().sum()
# No Null values
```

Out[4]:

```
Gender                                                     0
Age                                                        0
Marital Status                                             0
Children                                                   0
No of Children                                             0
Experience (in months)                                     0
Efficacy of team meetings                                  0
Social interaction                                         0
Overwork                                                   0
Stress                                                     0
Motivation                                                 0
Innovative Climate                                         0
Work Environment                                           0
No of working hrs per week                                 0
Uninterrupted Working hrs per week                         0
Number of stretches of uninterrupted working (min 1 hr)    0
Average no of meetings attended per week                   0
Hrs spent in meetings per week                             0
Salary (Monthly)                                           0
EP Outcome                                                 0
dtype: int64
```

In [5]:

```
productivity.duplicated().sum()
# No duplicate records
```

Out[5]:

```
0
```

In [6]:

```
productivity.describe().T
```

Out[6]:

	count	mean	std	min	25%	50%	75%	max
Age	250.0	23.960	0.976840	22.0	23.0	24.0	25.00	25.0
No of Children	250.0	0.260	0.608177	0.0	0.0	0.0	0.00	2.0
Experience (in months)	250.0	19.564	5.049743	12.0	16.0	19.0	23.00	30.0
Efficacy of team meetings	250.0	3.436	1.221422	1.0	3.0	3.0	4.75	5.0
Social interaction	250.0	3.432	1.256859	1.0	3.0	3.0	5.00	5.0
Overwork	250.0	2.600	1.241841	1.0	2.0	3.0	3.00	5.0
Stress	250.0	2.768	1.268310	1.0	2.0	3.0	4.00	5.0
Motivation	250.0	3.404	1.338638	1.0	3.0	4.0	5.00	5.0
Innovative Climate	250.0	3.188	1.209341	1.0	2.0	3.0	4.00	5.0
Work Environment	250.0	3.308	1.278937	1.0	3.0	3.0	4.00	5.0
No of working hrs per week	250.0	43.416	2.237626	40.0	42.0	43.0	45.00	48.0
Uninterrupted Working hrs per week	250.0	30.260	3.651539	23.0	28.0	29.0	33.00	40.0
Number of stretches of uninterrupted working (min 1 hr)	250.0	12.228	1.254716	10.0	11.0	12.0	13.00	14.0
Average no of meetings attended per week	250.0	8.280	1.643290	6.0	7.0	8.0	9.00	12.0
Hrs spent in meetings per week	250.0	5.428	1.146330	4.0	5.0	5.0	6.00	8.0
Salary (Monthly)	250.0	44006.620	5862.431710	35036.0	38972.5	42935.5	49021.25	54980.0
EP Outcome	250.0	0.808	0.394663	0.0	1.0	1.0	1.00	1.0

In [7]:

```
productivity.describe(include =['O']).T
```

Out[7]:

	count	unique	top	freq
Gender	250	2	Male	130
Marital Status	250	2	Single	163
Children	250	2	N	197

Exploratory Data Analysis on Productivity of Employee

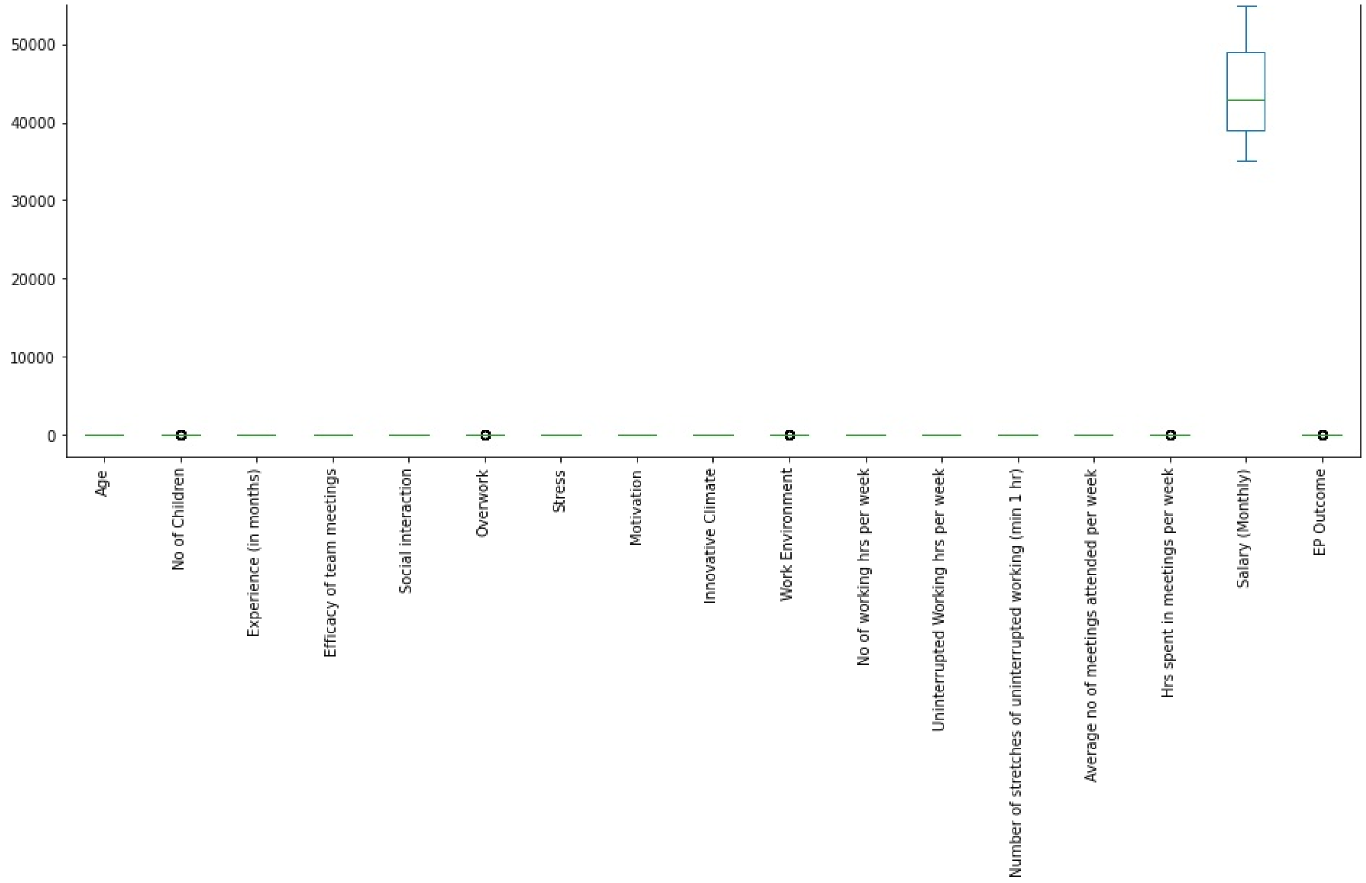
1. Outlier analysis

In [8]:

```
productivity.plot.box(figsize = (16,6))
plt.xticks(rotation = 90)
```

Out[8]:

```
(array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17]),
 [Text(1, 0, 'Age'),
  Text(2, 0, 'No of Children'),
  Text(3, 0, 'Experience (in months)'),
  Text(4, 0, 'Efficacy of team meetings'),
  Text(5, 0, 'Social interaction'),
  Text(6, 0, 'Overwork'),
  Text(7, 0, 'Stress'),
  Text(8, 0, 'Motivation'),
  Text(9, 0, 'Innovative Climate'),
  Text(10, 0, 'Work Environment'),
  Text(11, 0, 'No of working hrs per week'),
  Text(12, 0, 'Uninterrupted Working hrs per week'),
  Text(13, 0, 'Number of stretches of uninterrupted working (min 1 hr)'),
  Text(14, 0, 'Average no of meetings attended per week'),
  Text(15, 0, 'Hrs spent in meetings per week'),
  Text(16, 0, 'Salary (Monthly)'),
  Text(17, 0, 'EP Outcome')])
```



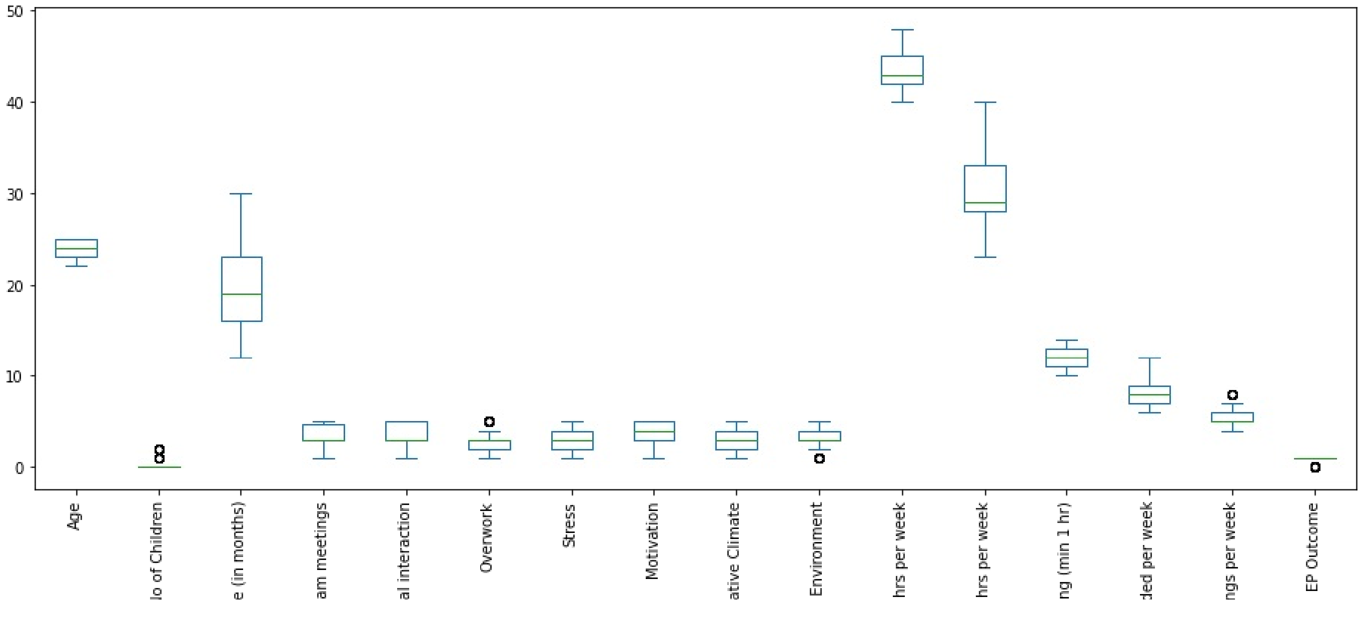
In the above plot, there are no outliers in Salary. The scale of representation distorts the depiction of data of the other features. Hence, it is removed and plotted again in the next plot for better visibilty of the other features.

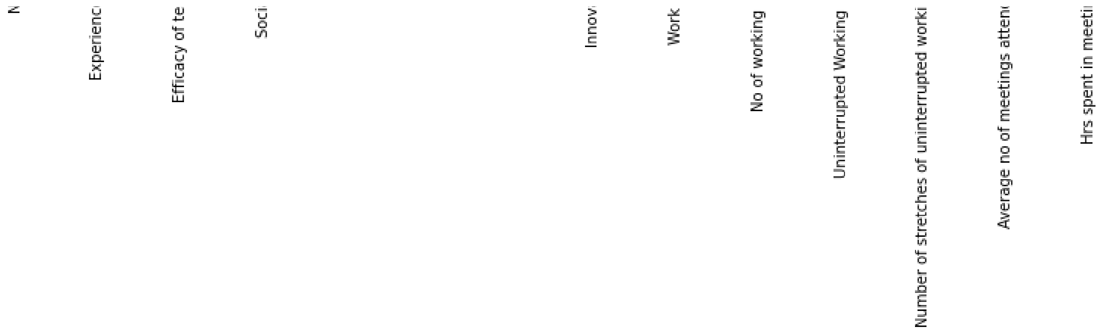
In [9]:

```
dfp = productivity.drop(["Salary (Monthly)"], axis = 1)
dfp.plot.box(figsize = (16,6))
plt.xticks(rotation = 90)
```

Out[9]:

```
(array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16]),
 [Text(1, 0, 'Age'),
  Text(2, 0, 'No of Children'),
  Text(3, 0, 'Experience (in months)'),
  Text(4, 0, 'Efficacy of team meetings'),
  Text(5, 0, 'Social interaction'),
  Text(6, 0, 'Overwork'),
  Text(7, 0, 'Stress'),
  Text(8, 0, 'Motivation'),
  Text(9, 0, 'Innovative Climate'),
  Text(10, 0, 'Work Environment'),
  Text(11, 0, 'No of working hrs per week'),
  Text(12, 0, 'Uninterrupted Working hrs per week'),
  Text(13, 0, 'Number of stretches of uninterrupted working (min 1 hr)'),
  Text(14, 0, 'Average no of meetings attended per week'),
  Text(15, 0, 'Hrs spent in meetings per week'),
  Text(16, 0, 'EP Outcome')])
```





In the above plot we can see that there are very few outliers and they are not severe in nature. In this study presence of outliers is important to capture the full range of perspectives.

2. Data Visualisation

In [10]:

```
#function for crosstabs
def cross_tab_pp(x,y):
    crtabp = pd.crosstab(productivity[x], productivity[y])
    return crtabp
```

In [11]:

```
#Productivity perspective of the people

p = productivity['EP Outcome'].value_counts()
print(p)
```

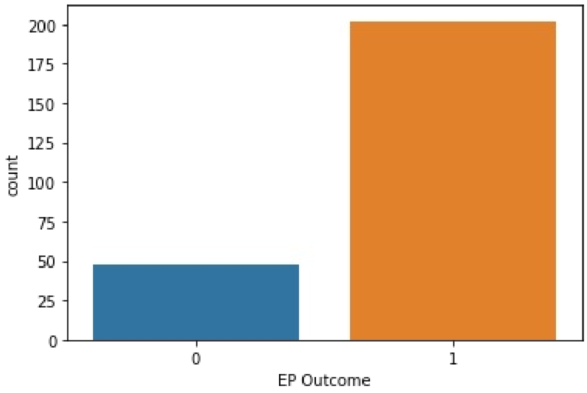
```
1    202
0     48
Name: EP Outcome, dtype: int64
```

In [12]:

```
sns.countplot(x=productivity['EP Outcome'])
```

Out[12]:

```
<AxesSubplot:xlabel='EP Outcome', ylabel='count'>
```



The data leans towards class 1. There is imbalance in the data which will have to be addressed during modelling.

In [13]:

```
#Age v/s EP Outcome

cross_tab_pp('Age','EP Outcome')
```

Out[13]:

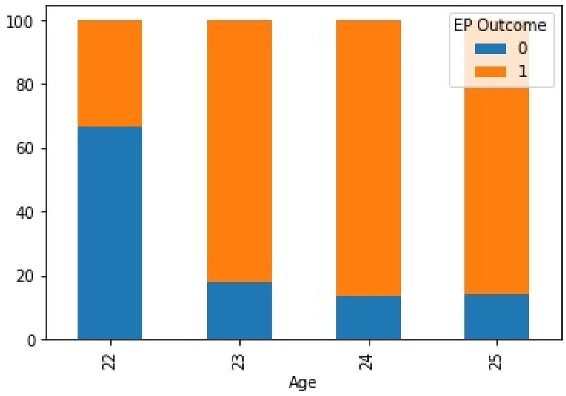
EP Outcome	0	1
Age		
22	14	7
23	11	50
24	10	65
25	13	80

In [14]:

```
table=pd.crosstab(productivity['Age'],productivity['EP Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[14]:

```
<AxesSubplot:xlabel='Age'>
```



Majority of the employees of age 22 feel that their productivity has deteriorated, while an overwhelming majority of the remaining age categories feel that their productivity is either the same or improved.

In [15]:

```
#Gender v/s EP Outcome

cross_tab_pp('Gender','EP Outcome')
```

Out[15]:

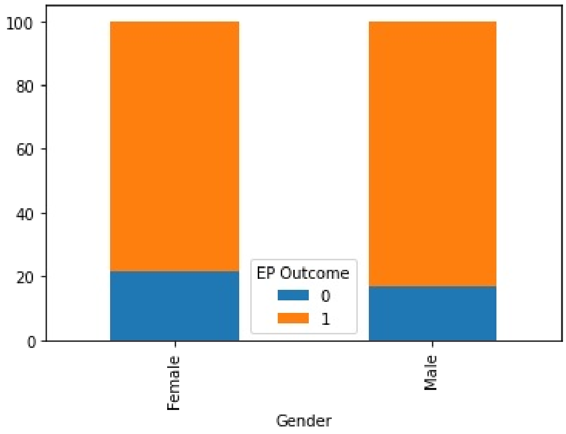
EP Outcome	0	1
Gender		
Female	26	94
Male	22	108

In [16]:

```
table=pd.crosstab(productivity['Gender'],productivity['EP Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[16]:

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



Almost 80% of both males and females working from home feel their productivity has improved or is the same.

In [17]:

```
#Marital Status v/s EP Outcome

cross_tab_pp('Marital Status','EP Outcome')
```

Out[17]:

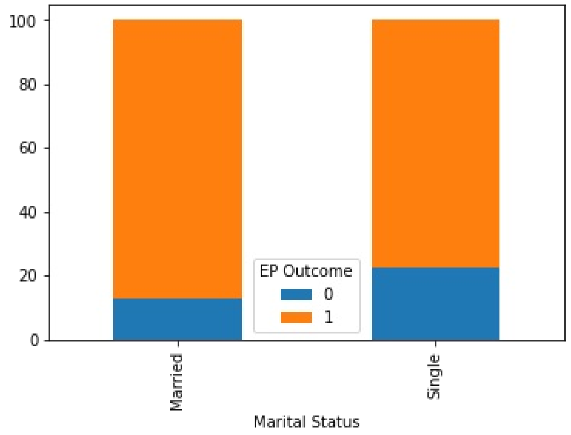
EP Outcome	0	1
Marital Status		
Married	11	76
Single	37	126

In [18]:

```
table=pd.crosstab(productivity['Marital Status'],productivity['EP Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[18]:

```
<AxesSubplot:xlabel='Marital Status'>
```



Almost 4/5th of the employees, whether married or single felt that their productivity had improved or remained the same.

In [19]:

```
#Children v/s EP Outcome

cross_tab_pp('Children','EP Outcome')
```

Out[19]:

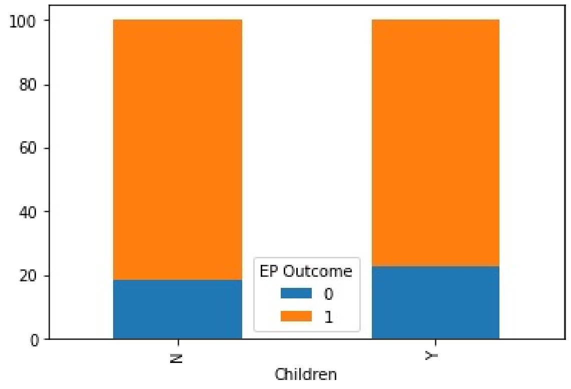
EP Outcome	0	1
Children		
N	36	161
Y	12	41

In [20]:

```
table=pd.crosstab(productivity['Children'],productivity['EP Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[20]:

```
<AxesSubplot:xlabel='Children'>
```



ALmost 80% of employees in both categories of having children and not having children felt that their productivity has improved or remained the same.

In [21]:

```
#No of Children v/s EP Outcome

cross_tab_pp('No of Children','EP Outcome')
```

Out[21]:

EP Outcome	0	1
No of Children		

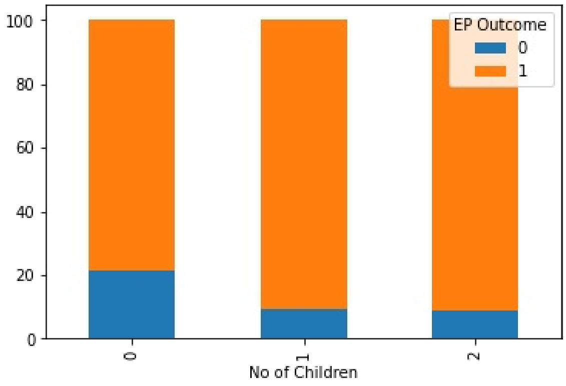
0	44	163
1	2	19
2	2	20

In [22]:

```
table=pd.crosstab(productivity['No of Children'],productivity['EP Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[22]:

```
<AxesSubplot:xlabel='No of Children'>
```



The above plot corroborates the previous observation of higher or same level of productivity among employees having children.

In [23]:

```
#Work Experience of the Employee

productivity['Experience (in months)'].describe().T
```

Out[23]:

```
count    250.000000
mean      19.564000
std        5.049743
min       12.000000
25%       16.000000
50%       19.000000
75%       23.000000
max       30.000000
Name: Experience (in months), dtype: float64
```

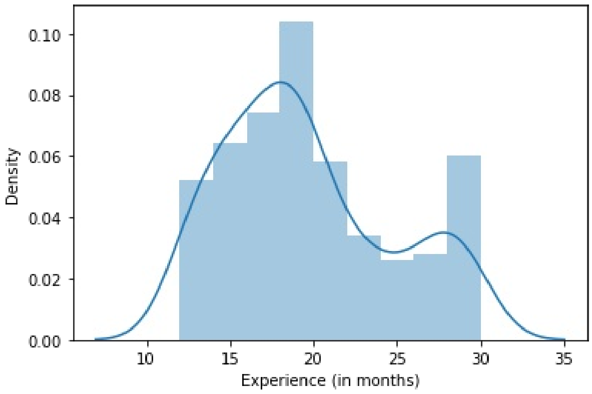
In [24]:

```
sns.distplot(productivity['Experience (in months)'])
```

```
/Users/manuair/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot`
is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot`
(a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[24]:

```
<AxesSubplot:xlabel='Experience (in months)', ylabel='Density'>
```

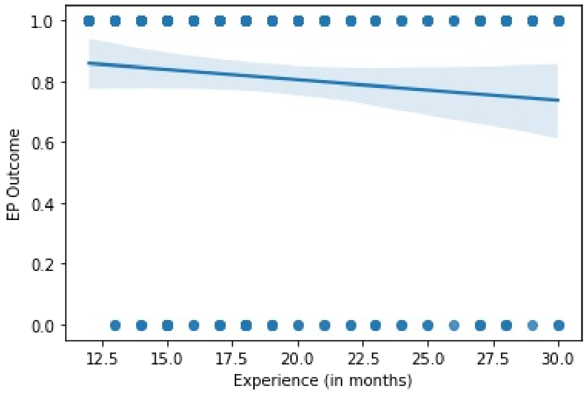


In [25]:

```
sns.regplot(x="Experience (in months)", y="EP Outcome", data=productivity)
```

Out[25]:

```
<AxesSubplot:xlabel='Experience (in months)', ylabel='EP Outcome'>
```



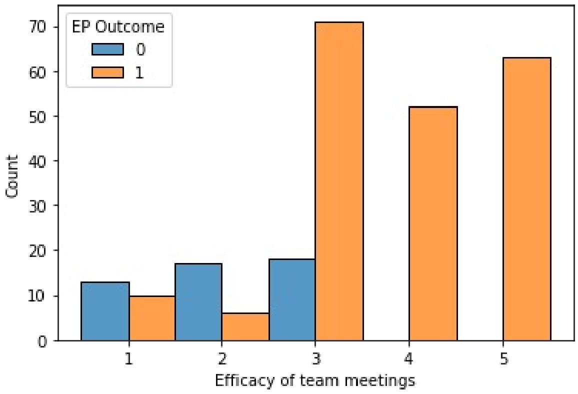
The segment of employees having 18-20 months of work experience are more in number than any other segment.

In [26]:

```
# Efficacy of team meetings v/s EP Outcome

sns.histplot(productivity, x="Efficacy of team meetings", hue="EP Outcome", discrete=True, multiple="dodge")
```

Out[26]:

```
<AxesSubplot:xlabel='Efficacy of team meetings', ylabel='Count'>
```

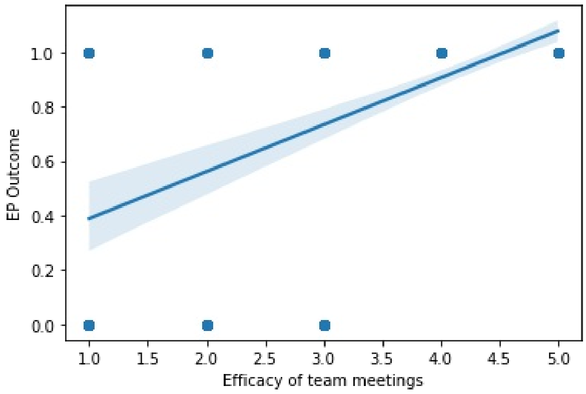


In [27]:

```
sns.regplot(x="Efficacy of team meetings", y="EP Outcome", data=productivity)
```

Out[27]:

```
<AxesSubplot:xlabel='Efficacy of team meetings', ylabel='EP Outcome'>
```



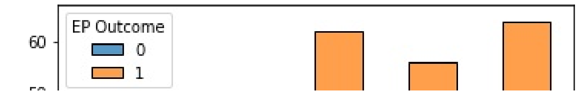
The employees who felt that they were as or more productive, found the team meetings more efficacious compared to those who felt that their productivity had dropped.

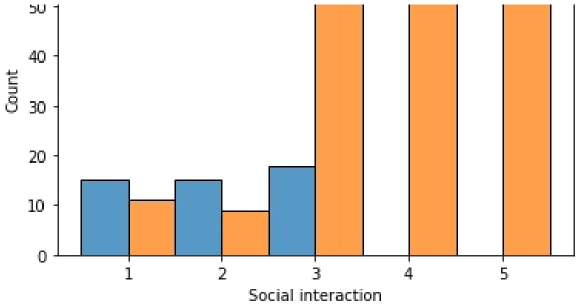
In [28]:

```
# Social interaction v/s EP Outcome

sns.histplot(productivity, x="Social interaction", hue="EP Outcome", discrete=True, multiple="dodge")
```

Out[28]:

```
<AxesSubplot:xlabel='Social interaction', ylabel='Count'>
```

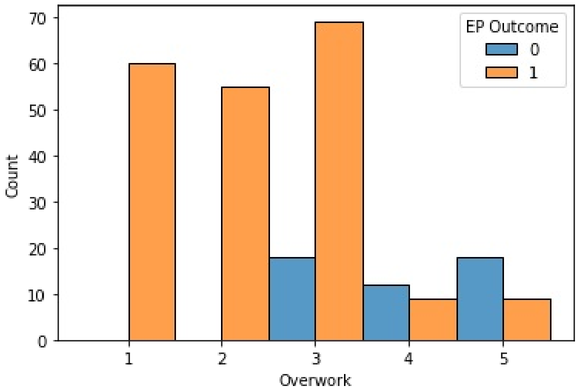




The employees who felt that they were as or more productive, found social interaction more satisfying compared to those who felt that their productivity had dropped.

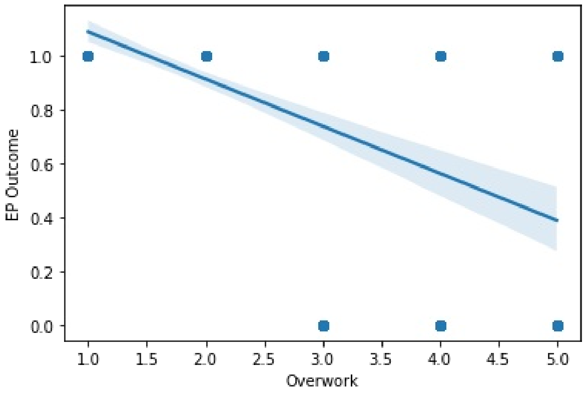
In [29]:
```
# Overwork v/s EP Outcome

sns.histplot(productivity, x="Overwork", hue="EP Outcome", discrete=True, multiple="dodge")
```

Out[29]:
```
<AxesSubplot:xlabel='Overwork', ylabel='Count'>
```



In [30]:
```
sns.regplot(x="Overwork", y="EP Outcome", data=productivity)
```

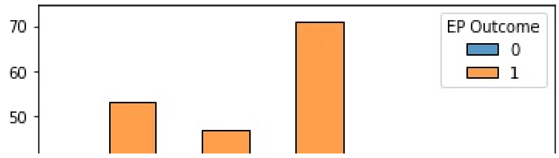
Out[30]:
```
<AxesSubplot:xlabel='Overwork', ylabel='EP Outcome'>
```

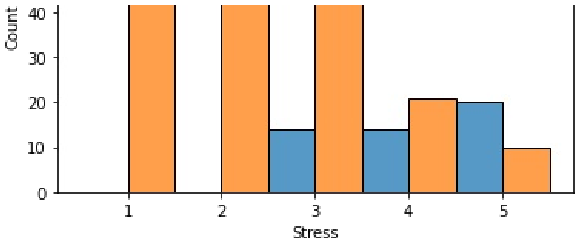


The employees who felt that they were lesser productive, complained of overwork compared to those who felt that their productivity was same or had improved.

In [31]:
```
# Stress v/s EP Outcome

sns.histplot(productivity, x="Stress", hue="EP Outcome", discrete=True, multiple="dodge")
```

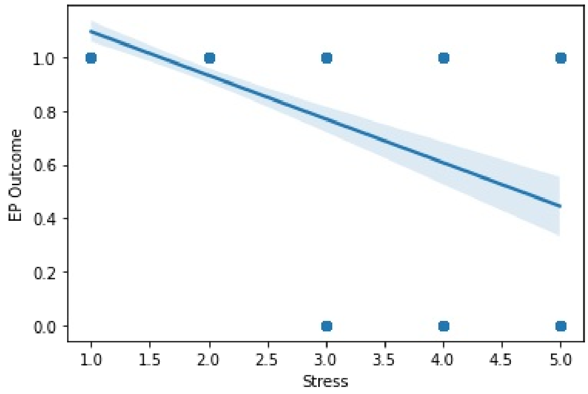
Out[31]:
```
<AxesSubplot:xlabel='Stress', ylabel='Count'>
```





In [32]:
```
sns.regplot(x="Stress", y="EP Outcome", data=productivity)
```

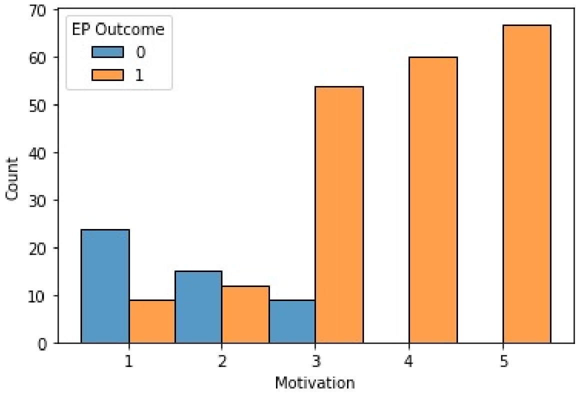
Out[32]:
```
<AxesSubplot:xlabel='Stress', ylabel='EP Outcome'>
```



The employees who felt that they were lesser productive, complained of more stress compared to those who felt that their productivity was same or had improved.

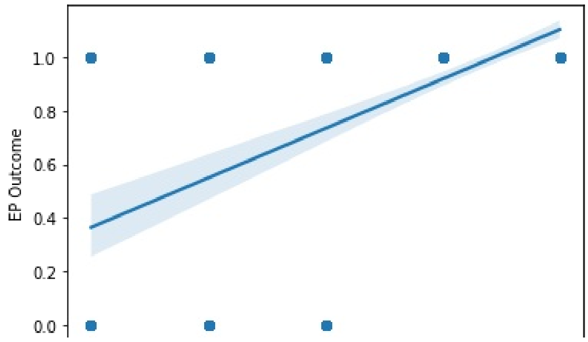
In [33]:
```
# Motivation v/s EP Outcome

sns.histplot(productivity, x="Motivation", hue="EP Outcome", discrete=True, multiple="dodge")
```

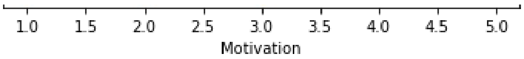
Out[33]:
```
<AxesSubplot:xlabel='Motivation', ylabel='Count'>
```



In [34]:
```
sns.regplot(x="Motivation", y="EP Outcome", data=productivity)
```

Out[34]:
```
<AxesSubplot:xlabel='Motivation', ylabel='EP Outcome'>
```





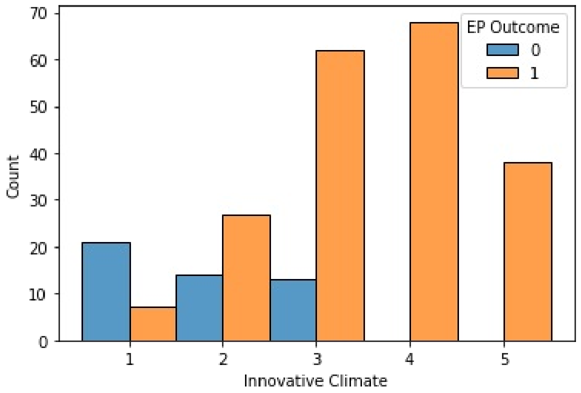
The employees who felt that they were as or more productive were more motivated compared to those who felt that their productivity had dropped.

In [35]:

```
# Innovative Climate v/s EP Outcome

sns.histplot(productivity, x="Innovative Climate", hue="EP Outcome", discrete=True, multiple="dodge")
```

Out[35]:

```
<AxesSubplot:xlabel='Innovative Climate', ylabel='Count'>
```

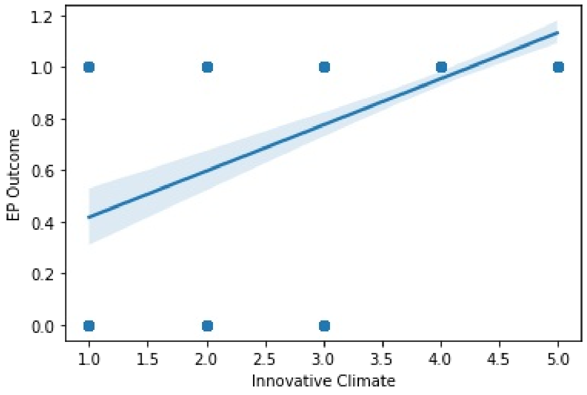


In [36]:

```
sns.regplot(x="Innovative Climate", y="EP Outcome", data=productivity)
```

Out[36]:

```
<AxesSubplot:xlabel='Innovative Climate', ylabel='EP Outcome'>
```



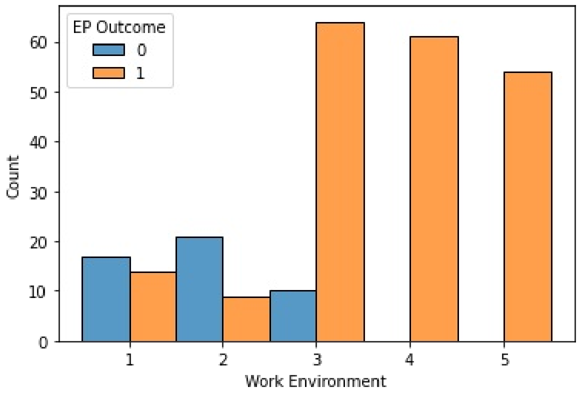
The employees who felt that they were as or more productive felt that the company climate was supported innovation compared to those who felt that their productivity had dropped.

In [37]:

```
# Work Environment v/s EP Outcome|

sns.histplot(productivity, x="Work Environment", hue="EP Outcome", discrete=True, multiple="dodge")
```

Out[37]:

```
<AxesSubplot:xlabel='Work Environment', ylabel='Count'>
```

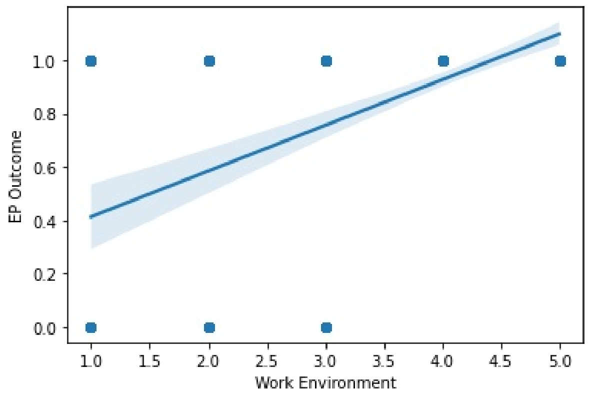


In [38]:

```
sns.regplot(x="Work Environment", y="EP Outcome", data=productivity)
```

Out[38]:

```
<AxesSubplot:xlabel='Work Environment', ylabel='EP Outcome'>
```



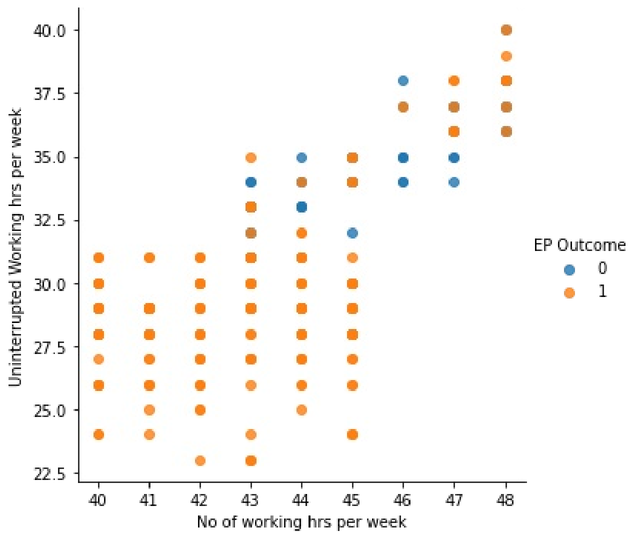
The employees who felt that they were as or more productive felt that the work environment was conducive and friendly compared to those who felt that their productivity had dropped.

In [39]:

```
# No of working hrs per week v/s Uninterrupted Working hrs per week hued with EP Outcome
sns.lmplot(x='No of working hrs per week', y='Uninterrupted Working hrs per week', hue='EP Outcome',
           data=productivity,
           fit_reg=False)
```

Out[39]:

```
<seaborn.axisgrid.FacetGrid at 0x7fa038c78af0>
```



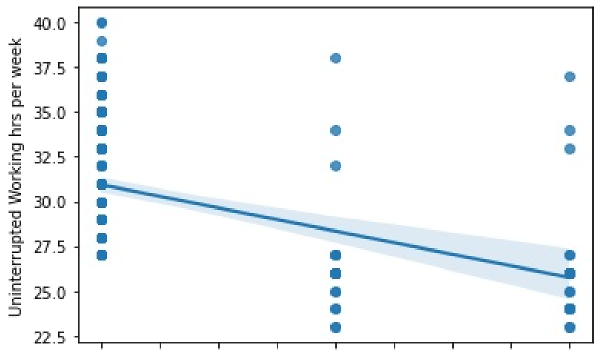
The working hours put in by the employees who feel that their productivity has reduced is in the higher regime compared to majority of those who feel that their productivity is the same or better. Also, the former have more uninterrupted working hours compared to the latter.

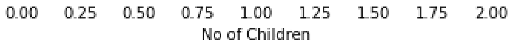
In [40]:

```
# No of Children v/s Uninterrupted Working hrs per week
sns.regplot(x="No of Children", y="Uninterrupted Working hrs per week", data=productivity)
```

Out[40]:

```
<AxesSubplot:xlabel='No of Children', ylabel='Uninterrupted Working hrs per week'>
```





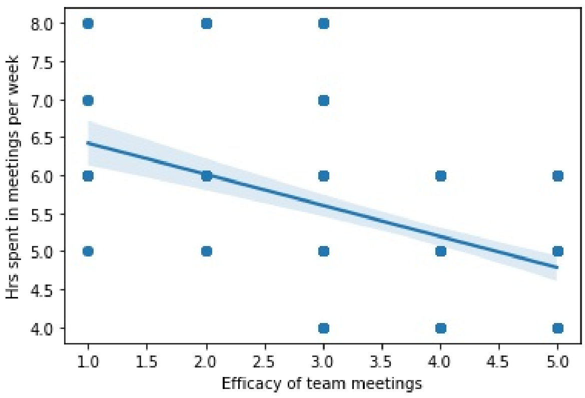
The employees having children have lesser uninterrupted working hours compared to those not having children.

In [41]:

```
# Efficacy of team meetings v/s Hrs spent in meetings per week
sns.regplot(x="Efficacy of team meetings", y="Hrs spent in meetings per week", data=productivity)
```

Out[41]:

```
<AxesSubplot:xlabel='Efficacy of team meetings', ylabel='Hrs spent in meetings per week'>
```



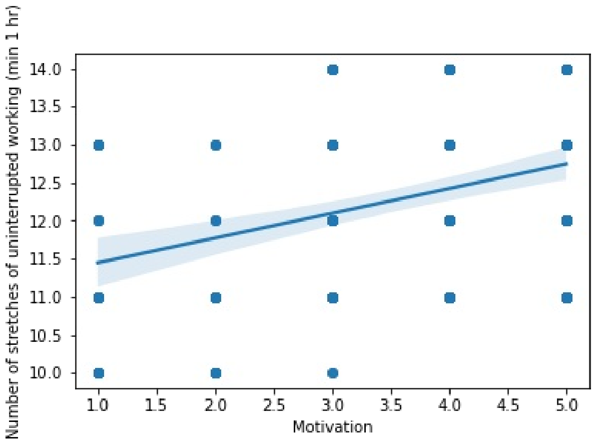
Those employees who spent more time in meetings felt that the meetings were less efficacious.

In [42]:

```
# Efficacy of Motivation v/s Number of stretches of uninterrupted working (min 1 hr)
sns.regplot(x="Motivation", y="Number of stretches of uninterrupted working (min 1 hr)", data=productivity)
```

Out[42]:

```
<AxesSubplot:xlabel='Motivation', ylabel='Number of stretches of uninterrupted working (min 1 hr)'>
```



Those employees who were more motivated reperted more number of stretches of uninterrupted working.

Statistical Analysis and Hypothesis Testing

1. Correlation

In [43]:

```
productivity.describe().T
```

Out[43]:

	count	mean	std	min	25%	50%	75%	max
Age	250.0	23.960	0.976840	22.0	23.0	24.0	25.00	25.0
No of Children	250.0	0.260	0.608177	0.0	0.0	0.0	0.00	2.0
Experience (in months)	250.0	19.564	5.049743	12.0	16.0	19.0	23.00	30.0
Efficacy of team meetings	250.0	3.436	1.221422	1.0	3.0	3.0	4.75	5.0
Social interaction	250.0	3.432	1.256859	1.0	3.0	3.0	5.00	5.0
Overwork	250.0	2.600	1.241841	1.0	2.0	3.0	3.00	5.0

Stress	250.0	2.768	1.268310	1.0	2.0	3.0	4.00	5.0
Motivation	250.0	3.404	1.338638	1.0	3.0	4.0	5.00	5.0
Innovative Climate	250.0	3.188	1.209341	1.0	2.0	3.0	4.00	5.0
Work Environment	250.0	3.308	1.278937	1.0	3.0	3.0	4.00	5.0
No of working hrs per week	250.0	43.416	2.237626	40.0	42.0	43.0	45.00	48.0
Uninterrupted Working hrs per week	250.0	30.260	3.651539	23.0	28.0	29.0	33.00	40.0
Number of stretches of uninterrupted working (min 1 hr)	250.0	12.228	1.254716	10.0	11.0	12.0	13.00	14.0
Average no of meetings attended per week	250.0	8.280	1.643290	6.0	7.0	8.0	9.00	12.0
Hrs spent in meetings per week	250.0	5.428	1.146330	4.0	5.0	5.0	6.00	8.0
Salary (Monthly)	250.0	44006.620	5862.431710	35036.0	38972.5	42935.5	49021.25	54980.0
EP Outcome	250.0	0.808	0.394663	0.0	1.0	1.0	1.00	1.0

In [44]:

```
pd.DataFrame(abs(productivity.corr()['EP Outcome']).sort_values(ascending = False))
```

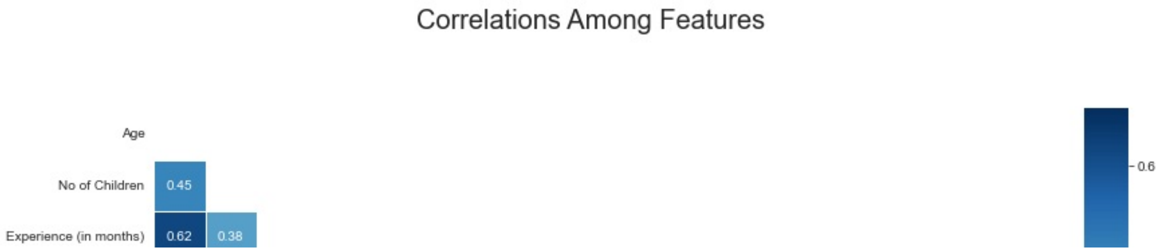
Out[44]:

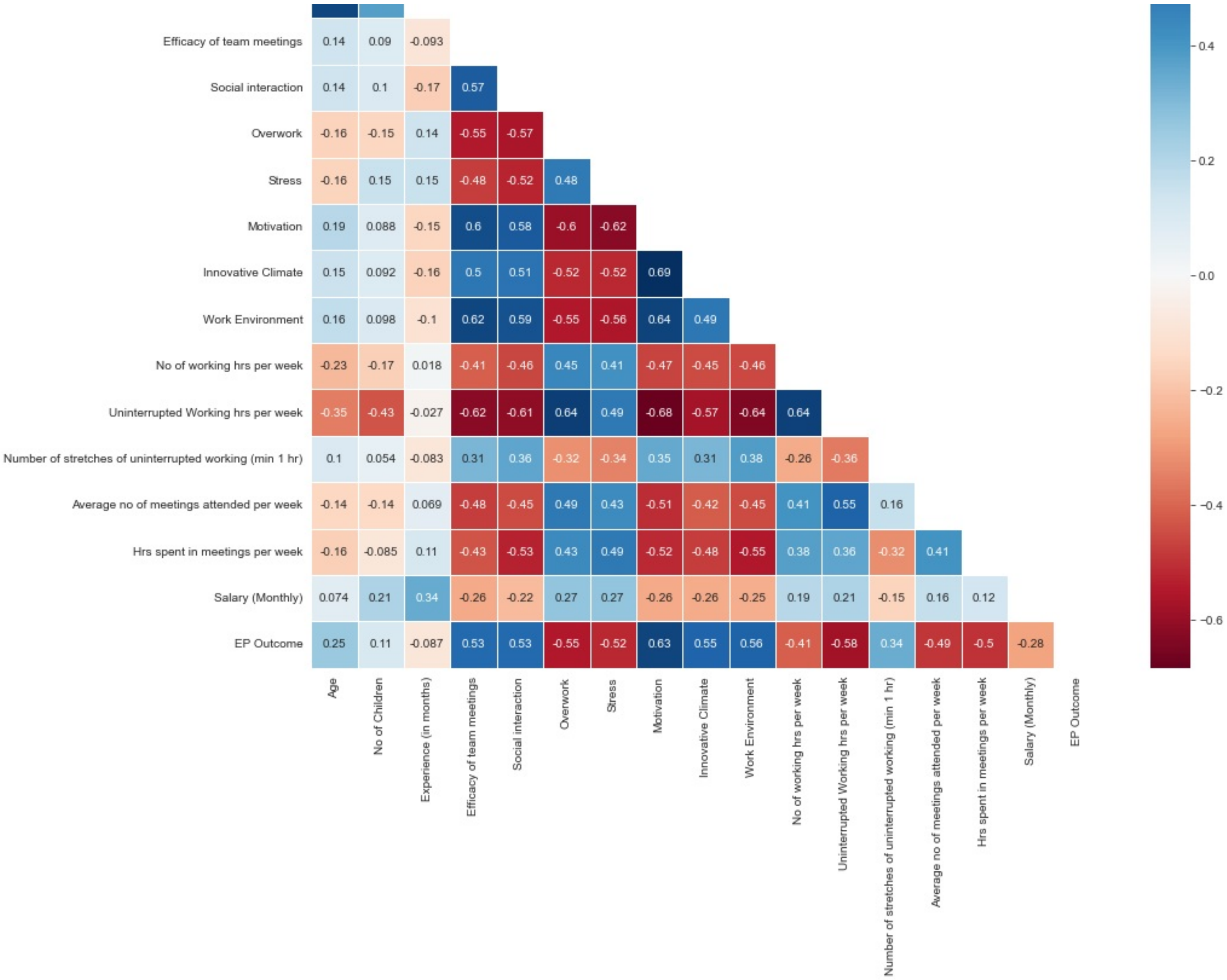
	EP Outcome
EP Outcome	1.000000
Motivation	0.626319
Uninterrupted Working hrs per week	0.578307
Work Environment	0.555240
Overwork	0.550652
Innovative Climate	0.547141
Efficacy of team meetings	0.532598
Social interaction	0.532219
Stress	0.522600
Hrs spent in meetings per week	0.501158
Average no of meetings attended per week	0.492668
No of working hrs per week	0.413981
Number of stretches of uninterrupted working (min 1 hr)	0.340172
Salary (Monthly)	0.279511
Age	0.250846
No of Children	0.108422
Experience (in months)	0.086506

In [45]:

```
mask = np.zeros_like(productivity.corr(), dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
sns.set_style('whitegrid')
plt.subplots(figsize = (15,12))
sns.heatmap(productivity.corr(),
            annot=True,
            mask = mask,
            cmap = 'RdBu', ## in order to reverse the bar replace "RdBu" with "RdBu_r"
            linewidths=.9,
            linecolor='white',
            fmt='.2g',
            center = 0,
            square=True)
plt.title("Correlations Among Features", y = 1.03,fontsize = 20, pad = 40);
```

```
/var/folders/2k/6r_xg74n77b1ytf4y98pm18w0000gn/T/ipykernel_1293/2572906870.py:1: DeprecationWarning: `np.bool` is
a deprecated alias for the builtin `bool`. To silence this warning, use `bool` by itself. Doing this will not mod
ify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.bool_` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#depr
ecations
  mask = np.zeros_like(productivity.corr(), dtype=np.bool)
```





2. Data Transformation

In [46]:

```
le=pre.LabelEncoder()
lt_2= productivity.select_dtypes(exclude = ['float','int']).columns.to_list()
```

In [47]:

```
for x in lt_2:
    productivity[x]=le.fit_transform(productivity[x])
productivity.head()
```

Out[47]:

	Gender	Age	Marital Status	Children	No of Children	Experience (in months)	Efficacy of team meetings	Social interaction	Overwork	Stress	Motivation	Innovative Climate	Work Environment	No of working hrs per week
0	0	23	1	0	0	18	4	3	1	2	4	5	3	45
1	1	24	1	0	0	19	4	3	1	2	5	4	3	44
2	1	24	1	0	0	17	4	4	1	3	3	4	4	44
3	1	22	1	0	0	17	2	3	3	5	2	2	2	48
4	1	25	0	1	1	24	4	5	3	1	4	3	5	40

3. Hypothesis Testing

In [48]:

```
xx = productivity.drop('EP Outcome',axis=1)
yy = productivity['EP Outcome']
```

In [49]:

```
chi_scores = chi2(xx,yy)
```

In [50]:

```
chi_scores
```

Out[50]:

```
(array([4.34437675e-01, 6.23984143e-01, 1.28664531e+00, 4.04632916e-01,
        4.16412795e+00, 2.42868426e+00, 3.06673664e+01, 3.24643469e+01,
        4.47829398e+01, 3.95205961e+01, 5.14195108e+01, 3.41960901e+01,
        3.79571521e+01, 4.92136711e+00, 3.66942821e+01, 3.70963911e+00,
        1.97109072e+01, 1.51400661e+01, 1.51926482e+04]),
 array([5.09819444e-01, 4.29570595e-01, 2.56667100e-01, 5.24706273e-01,
        4.12886622e-02, 1.19132855e-01, 3.06273030e-08, 1.21399399e-08,
        2.20132341e-11, 3.24617595e-10, 7.45935023e-13, 4.98289677e-09,
        7.23154865e-10, 2.65265187e-02, 1.38184151e-09, 5.40990870e-02,
        9.00857245e-06, 9.98221451e-05, 0.00000000e+00]))
```

In [51]:

```
ddd = pd.DataFrame(chi_scores, columns = xx.columns.to_list())
ddd
```

Out[51]:

	Gender	Age	Marital Status	Children	No of Children	Experience (in months)	Efficacy of team meetings	Social interaction	Overwork	Stress	Motivation	Inn
0	0.434438	0.623984	1.286645	0.404633	4.164128	2.428684	3.066737e+01	3.246435e+01	4.478294e+01	3.952060e+01	5.141951e+01	3.4196
1	0.509819	0.429571	0.256667	0.524706	0.041289	0.119133	3.062730e-08	1.213994e-08	2.201323e-11	3.246176e-10	7.459350e-13	4.982

In [52]:

```
p_values = pd.Series(chi_scores[1],index = xx.columns)
p_values.sort_values(ascending = False , inplace = True)
```

In [53]:

```
p_values
```

Out[53]:

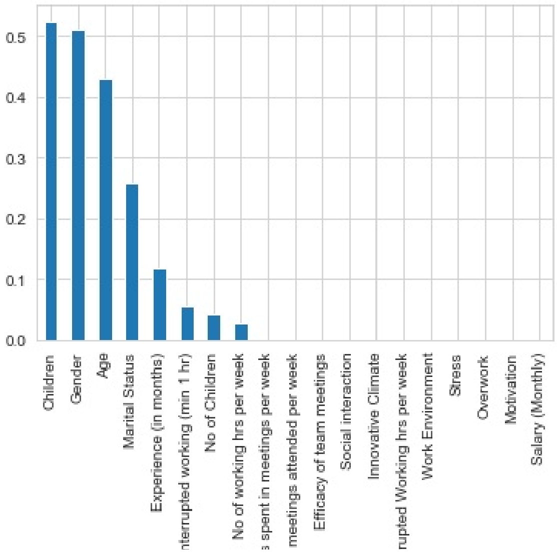
```
Children                                            5.247063e-01
Gender                                              5.098194e-01
Age                                                 4.295706e-01
Marital Status                                      2.566671e-01
Experience (in months)                              1.191329e-01
Number of stretches of uninterrupted working (min 1 hr)    5.409909e-02
No of Children                                      4.128866e-02
No of working hrs per week                          2.652652e-02
Hrs spent in meetings per week                      9.982215e-05
Average no of meetings attended per week            9.008572e-06
Efficacy of team meetings                           3.062730e-08
Social interaction                                  1.213994e-08
Innovative Climate                                  4.982897e-09
Uninterrupted Working hrs per week                  1.381842e-09
Work Environment                                    7.231549e-10
Stress                                              3.246176e-10
Overwork                                            2.201323e-11
Motivation                                          7.459350e-13
Salary (Monthly)                                    0.000000e+00
dtype: float64
```

In [54]:

```
p_values.plot.bar()
```

Out[54]:

```
<AxesSubplot:>
```



From the above results, it is clear that the factors which affect a person's productivity are

- No of Children
- No of working hrs per week
- Hrs spent in meetings per week
- Average no of meetings attended per week
- Efficacy of team meetings
- Social interaction
- Innovative Climate
- Uninterrupted Working hrs per week
- Work Environment
- Stress
- Overwork
- Motivation
- Salary (Monthly).

In conclusion, for the above features there is evidence to reject the null hypothesis that the outcome variable and each of these features are dependent. Thus, these features contribute to the productivity of an individual working from home.

Modelling

In [55]:
```
df_clean_scale_q=pre.minmax_scale(productivity)
```

In [56]:
```
df_clean_scale_q=pd.DataFrame(df_clean_scale_q,columns=productivity.columns.tolist())
```

In [57]:
```
X_1 = df_clean_scale_q.drop(['EP Outcome'], axis=1)

y_1 = productivity.iloc[:,19:20]
```

In [58]:
```
#SMOTE

smote = SMOTE()
x_smote1, y_smote1 = smote.fit_resample(X_1, y_1)
```

In [59]:
```
X_train1, X_test1, y_train1, y_test1 = train_test_split(x_smote1, y_smote1, test_size = 0.3, random_state = 0)
```

In [60]:
```
X_train1.shape, X_test1.shape
```

Out[60]:
```
((282, 19), (122, 19))
```

Decision Tree

In [61]:
```
dtc = DecisionTreeClassifier(criterion = 'entropy', max_depth = 5)
dtc.fit(X_train1, y_train1)
```

Out[61]:
```
DecisionTreeClassifier(criterion='entropy', max_depth=5)
```

In [62]:
```
y_pred_DT1 = dtc.predict(X_test1)
```

In [63]:
```
accuracy_score(y_test1,y_pred_DT1)
```

Out[63]:
```
0.9098360655737705
```

In [64]:

```
cmDT1 = confusion_matrix(y_test1, y_pred_DT1)

print('Confusion matrix\n\n', cmDT1)
```

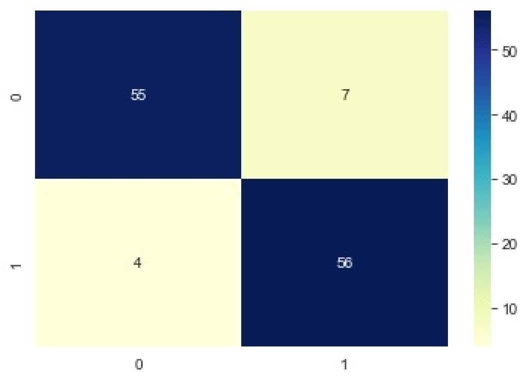
```
Confusion matrix

 [[55  7]
 [ 4 56]]
```

In [65]:

```
cm_matrix1 = pd.DataFrame(data=cmDT1)

sns.heatmap(cm_matrix1, annot=True, fmt='d', cmap='YlGnBu')
```

Out[65]:

```
<AxesSubplot:>
```



In [66]:

```
print(classification_report(y_test1, y_pred_DT1))
```

```
              precision    recall  f1-score   support

           0       0.93      0.89      0.91        62
           1       0.89      0.93      0.91        60

    accuracy                           0.91       122
   macro avg       0.91      0.91      0.91       122
weighted avg       0.91      0.91      0.91       122
```

The accuracy of prediction is 91%. The F1 score for class 0 and class 1 are both 91%.

II - Preference

Data

Dataset, Data Description and Data information

In [67]:

```
pd.set_option('display.max_columns', None)
preference = pd.read_csv("/Users/manuair/Study/Preference.csv")
preference.head()
```

Out[67]:

	Gender	Age	Marital Status	Children	No of Children	Distance	Cost of commute(per month)	Flexibility	Safety	Physical Exercise	Family Time	Mental Health	Care for elders	Pref Outcome
0	Female	23	Single	N	0	25	1276	5	4	5	4	5	3	1
1	Male	24	Single	N	0	16	817	5	3	4	4	5	4	1
2	Male	24	Single	N	0	9	459	4	4	3	3	4	5	1
3	Male	22	Single	N	0	30	1531	2	3	1	3	3	4	0
4	Male	25	Married	Y	1	22	1123	5	4	3	4	5	4	1

Data Description

- Gender - Male/Female
- Age
- Marital Status - Whether married or not - Single / Married
- Children - Whether the individual has children or not - Y/N
- No of children
- Distance - Distance of the individual's home from the workplace in km
- Cost of Commute - The average amount of money spent on commuting per month
- Flexibilty - Whether the individual finds flexibility in working from home - expressed on a Likert scale from 1 to 5
- Safety - Whether the individual feels safer while working from home - expressed on a Likert scale from 1 to 5
- Physical Exercise - Whether the individual finds more time to be physically fit - expressed on a Likert scale from 1 to 5
- Family Time - Whether the time spent with his/her family has changed - expressed on a Likert scale from 1 to 5
- Mental Health - Whether the individual finds any impact on mental health wrt WFH conditions - expressed on a Likert scale from 1 to 5
- Care for elders - Whether the individual finds more time to look after the elders in the family - expressed on a Likert scale from 1 to 5
- Pref Outcome - What is the preference of the individual towards WFH - 0(not preferred)/1(preferred)

In [68]:

```
preference.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 14 columns):
 #   Column                      Non-Null Count  Dtype 
---  ------                      --------------  ----- 
 0   Gender                      250 non-null    object
 1   Age                         250 non-null    int64 
 2   Marital Status              250 non-null    object
 3   Children                    250 non-null    object
 4   No of Children              250 non-null    int64 
 5   Distance                    250 non-null    int64 
 6   Cost of commute(per month)  250 non-null    int64 
 7   Flexibility                 250 non-null    int64 
 8   Safety                      250 non-null    int64 
 9   Physical Exercise           250 non-null    int64 
 10  Family Time                 250 non-null    int64 
 11  Mental Health               250 non-null    int64 
 12  Care for elders             250 non-null    int64 
 13  Pref Outcome                250 non-null    int64 
dtypes: int64(11), object(3)
memory usage: 27.5+ KB
```

In [69]:

```
preference.isnull().sum()
# No null values
```

Out[69]:

```
Gender                        0
Age                           0
Marital Status                0
Children                      0
No of Children                0
Distance                      0
Cost of commute(per month)    0
Flexibility                   0
Safety                        0
Physical Exercise             0
Family Time                   0
Mental Health                 0
Care for elders               0
Pref Outcome                  0
dtype: int64
```

In [70]:

```
preference.duplicated().sum()
# No duplicate records
```

Out[70]:

```
0
```

In [71]:

```
preference.describe().T
```

Out[71]:

	count	mean	std	min	25%	50%	75%	max
Age	250.0	23.960	0.976840	22.0	23.0	24.0	25.0	25.0
	250.0	0.260	0.608177	0.0	0.0	0.0	0.0	2.0

No of Children								
Distance	250.0	19.116	6.460767	8.0	14.0	19.0	25.0	30.0
Cost of commute(per month)	250.0	975.708	329.792034	408.0	715.0	970.0	1276.0	1531.0
Flexibility	250.0	3.556	1.143720	1.0	3.0	4.0	4.0	5.0
Safety	250.0	3.992	0.801565	3.0	3.0	4.0	5.0	5.0
Physical Exercise	250.0	3.584	1.131399	1.0	3.0	4.0	4.0	5.0
Family Time	250.0	3.636	0.877830	2.0	3.0	4.0	4.0	5.0
Mental Health	250.0	3.452	1.264155	1.0	3.0	3.0	5.0	5.0
Care for elders	250.0	3.708	0.960430	2.0	3.0	4.0	5.0	5.0
Pref Outcome	250.0	0.832	0.374616	0.0	1.0	1.0	1.0	1.0

In [72]:

```
preference.describe(include =['O']).T
```

Out[72]:

	count	unique	top	freq
Gender	250	2	Male	130
Marital Status	250	2	Single	163
Children	250	2	N	197

Exploratory Data Analysis on Preference of Employees

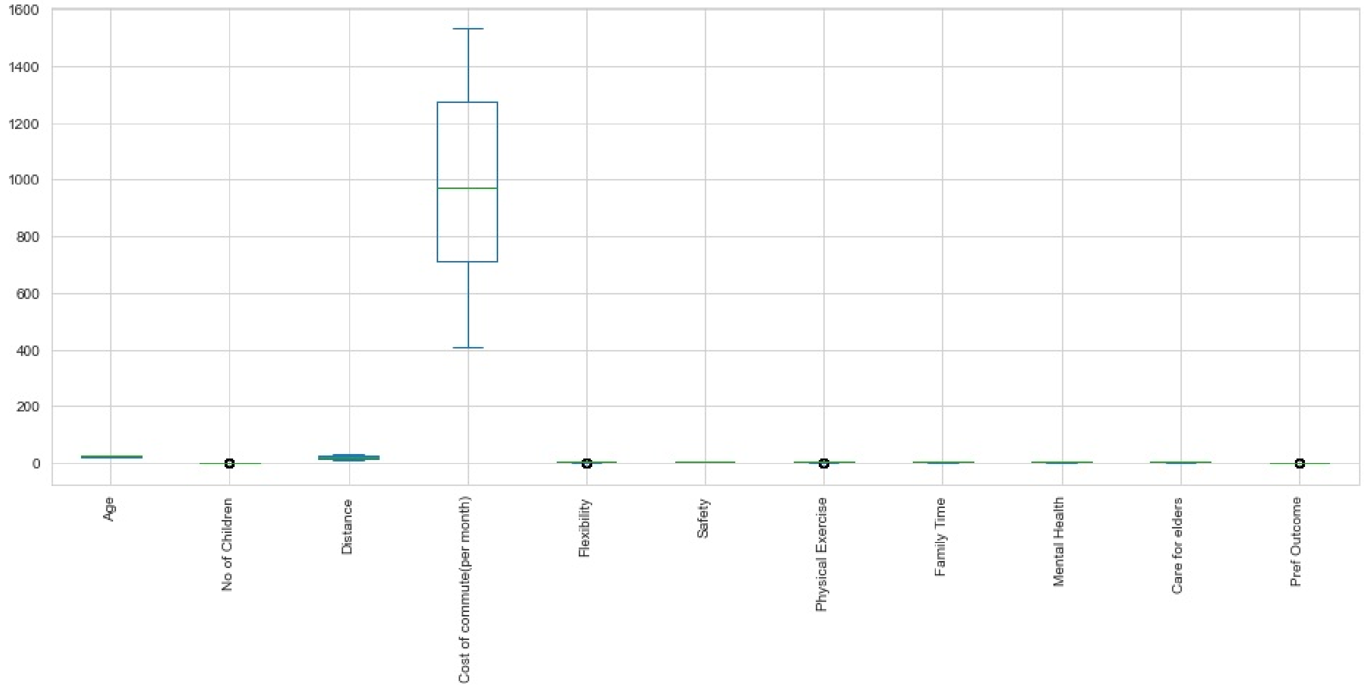
1. Outlier analysis

In [73]:

```
preference.plot.box(figsize = (16,6))
plt.xticks(rotation = 90)
```

Out[73]:

```
(array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(1, 0, 'Age'),
  Text(2, 0, 'No of Children'),
  Text(3, 0, 'Distance'),
  Text(4, 0, 'Cost of commute(per month)'),
  Text(5, 0, 'Flexibility'),
  Text(6, 0, 'Safety'),
  Text(7, 0, 'Physical Exercise'),
  Text(8, 0, 'Family Time'),
  Text(9, 0, 'Mental Health'),
  Text(10, 0, 'Care for elders'),
  Text(11, 0, 'Pref Outcome')])
```



The plot above does not show any significant outliers. However, the scale of representation is distorted by the plot of Cost of commute.

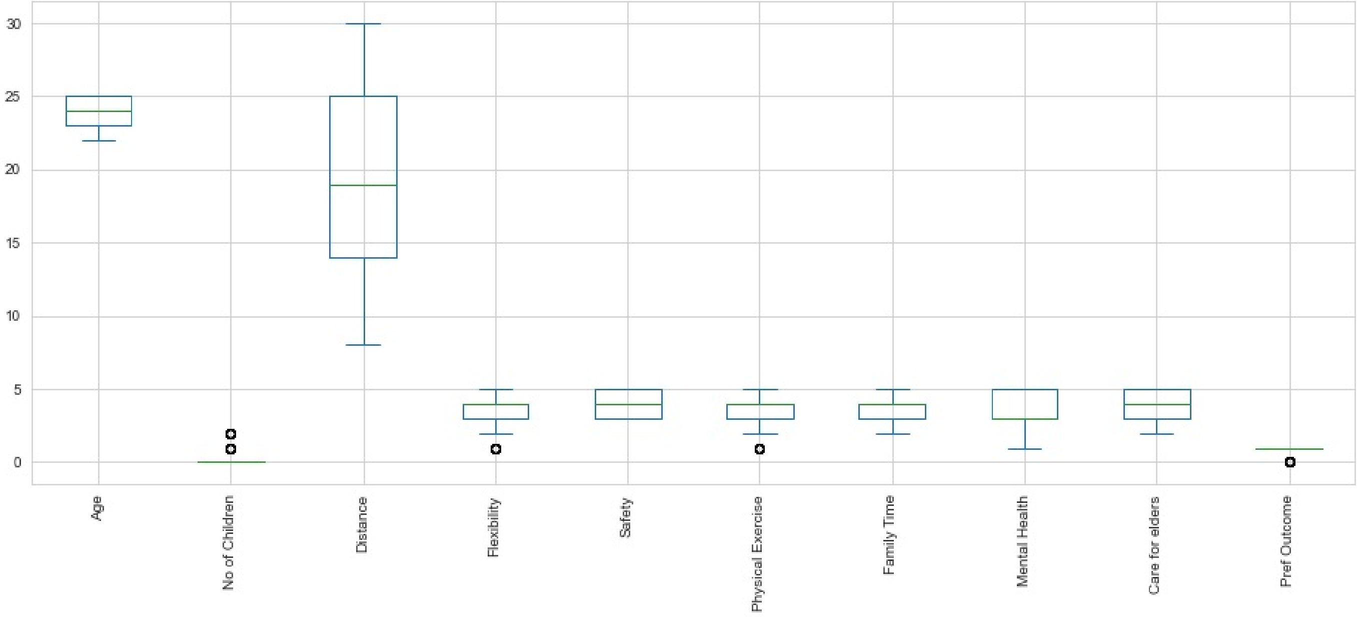
Therefore, it is removed and replotted to get a clearer picture of the outliers.

In [74]:

```
df1 = preference.drop(["Cost of commute(per month)"], axis = 1)
df1.plot.box(figsize = (16,6))
plt.xticks(rotation = 90)
```

Out[74]:

```
(array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10]),
 [Text(1, 0, 'Age'),
  Text(2, 0, 'No of Children'),
  Text(3, 0, 'Distance'),
  Text(4, 0, 'Flexibility'),
  Text(5, 0, 'Safety'),
  Text(6, 0, 'Physical Exercise'),
  Text(7, 0, 'Family Time'),
  Text(8, 0, 'Mental Health'),
  Text(9, 0, 'Care for elders'),
  Text(10, 0, 'Pref Outcome')])
```



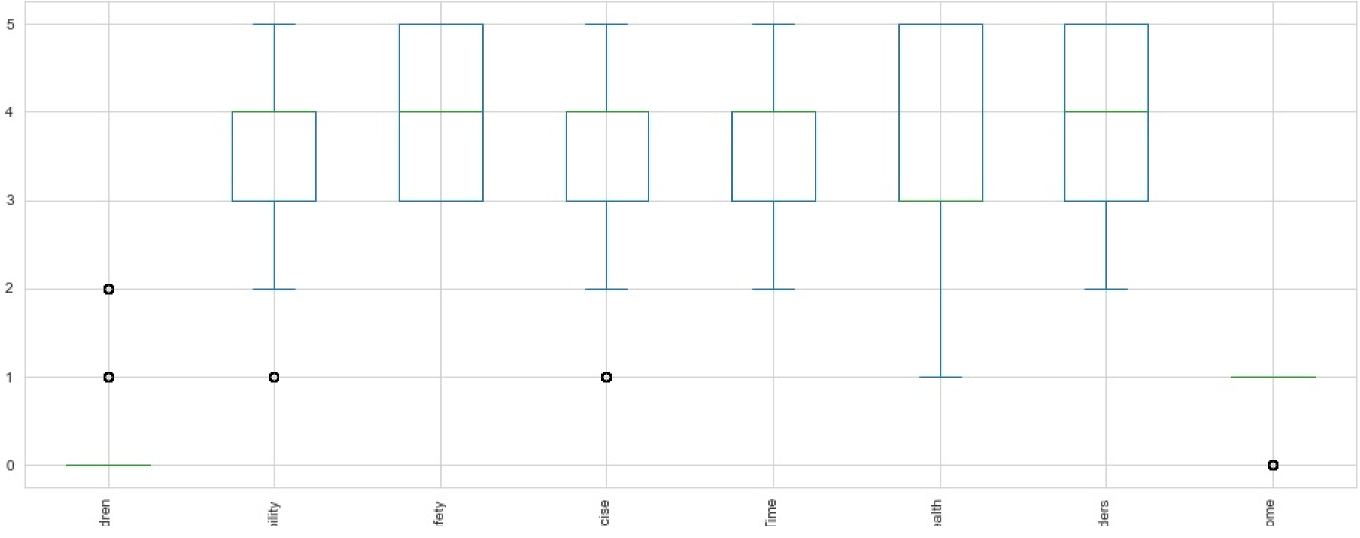
There are very few outliers and not severe in nature. However, since Distance again distorts the representation, the feature is removed and replotted in the next plot.

In [75]:

```
df2 = df1.drop(["Distance","Age"], axis = 1)
df2.plot.box(figsize = (16,6))
plt.xticks(rotation = 90)
```

Out[75]:

```
(array([1, 2, 3, 4, 5, 6, 7, 8]),
 [Text(1, 0, 'No of Children'),
  Text(2, 0, 'Flexibility'),
  Text(3, 0, 'Safety'),
  Text(4, 0, 'Physical Exercise'),
  Text(5, 0, 'Family Time'),
  Text(6, 0, 'Mental Health'),
  Text(7, 0, 'Care for elders'),
  Text(8, 0, 'Pref Outcome')])
```



There are very few outliers and not severe in nature. In this particular case study, the presence of outliers is important in the data to analyse the variety of responses.

2. Data Visualisation

In [76]:

```
#function for crosstabs
def cross_tab(x,y):
    crtab = pd.crosstab(preference[x], preference[y])
    return crtab
```

In [77]:

```
#Preference of the people

p = preference['Pref Outcome'].value_counts()
print(p)
```

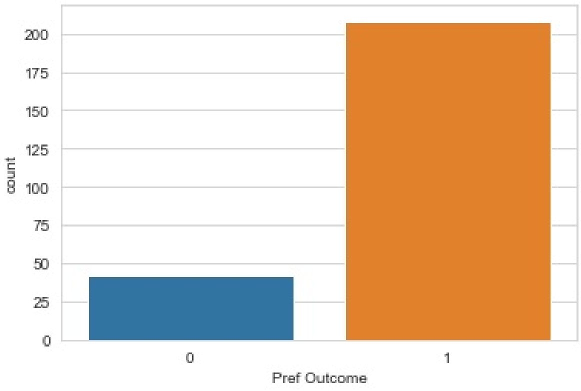
```
1    208
0     42
Name: Pref Outcome, dtype: int64
```

In [78]:

```
sns.countplot(x=preference['Pref Outcome'])
```

Out[78]:

```
<AxesSubplot:xlabel='Pref Outcome', ylabel='count'>
```



Majority of the employees prefer work from home.

In [79]:

```
#Age v/s Pref Outcome

cross_tab('Age','Pref Outcome')
```

Out[79]:

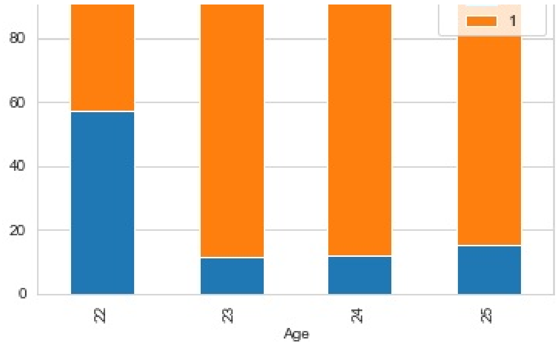
Pref Outcome	0	1
Age		
22	12	9
23	7	54
24	9	66
25	14	79

In [80]:

```
table=pd.crosstab(preference['Age'],preference['Pref Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[80]:

```
<AxesSubplot:xlabel='Age'>
```



The large majority in the age category of 23-25 prefer WFH, while most of the ones of age 22 prefer to work from office.

In [81]:

```
#Gender v/s Pref Outcome

cross_tab('Gender','Pref Outcome')
```

Out[81]:

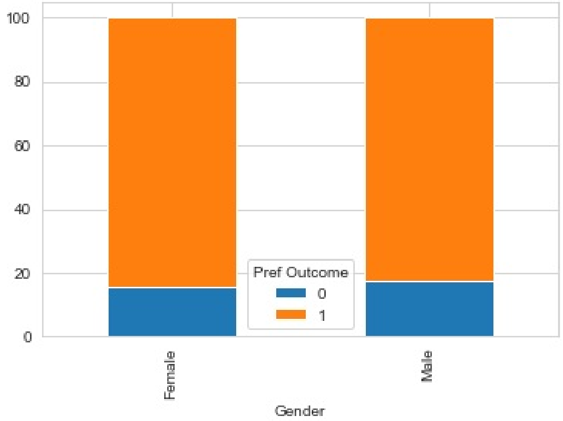
Pref Outcome	0	1
Gender		
Female	19	101
Male	23	107

In [82]:

```
table=pd.crosstab(preference['Gender'],preference['Pref Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[82]:

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



Almost 85% of both males and females prefer to work from home.

In [83]:

```
#Marital Status v/s Pref Outcome

cross_tab('Marital Status','Pref Outcome')
```

Out[83]:

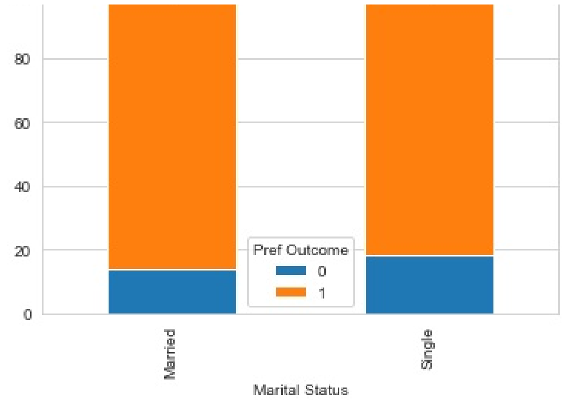
Pref Outcome	0	1
Marital Status		
Married	12	75
Single	30	133

In [84]:

```
table=pd.crosstab(preference['Marital Status'],preference['Pref Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[84]:

```
<AxesSubplot:xlabel='Marital Status'>
```



Almost 80% of those who are both married and unmarried prefer to work from home.

In [85]:

```
#Children v/s Pref Outcome

cross_tab('Children','Pref Outcome')
```

Out[85]:

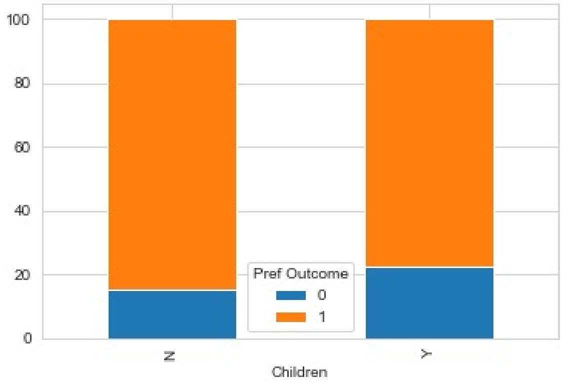
Pref Outcome	0	1
Children		
N	30	167
Y	12	41

In [86]:

```
table=pd.crosstab(preference['Children'],preference['Pref Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[86]:

```
<AxesSubplot:xlabel='Children'>
```



Almost 80% of all employees, both having children and without children, prefer to work from home.

In [87]:

```
#No of childern v/s Pref Outcome

cross_tab('No of Children','Pref Outcome')
```

Out[87]:

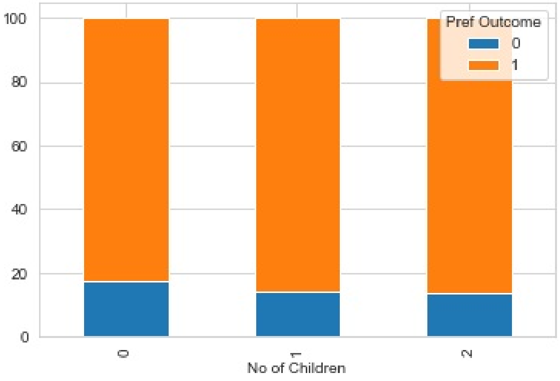
Pref Outcome	0	1
No of Children		
0	36	171
1	3	18
2	3	19

In [88]:

```
table=pd.crosstab(preference['No of Children'],preference['Pref Outcome'])
stacked_data = table.apply(lambda x: x*100/sum(x), axis=1)
stacked_data.plot(kind="bar", stacked=True)
```

Out[88]:

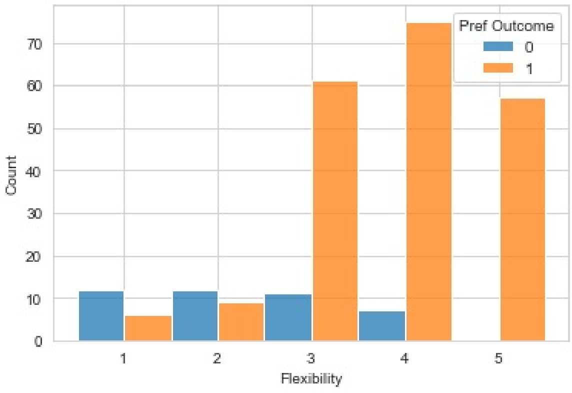
```
<AxesSubplot:xlabel='No of Children'>
```



The above plot corroborates the previous observation.

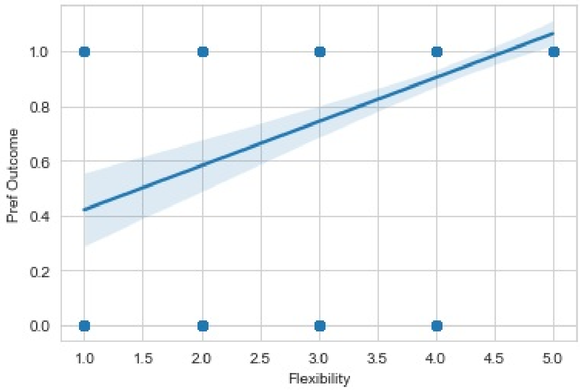
In [89]:
```
# Flexibility v/s Pref Outcome

sns.histplot(preference, x="Flexibility", hue="Pref Outcome", discrete=True, multiple="dodge")
```

Out[89]:
```
<AxesSubplot:xlabel='Flexibility', ylabel='Count'>
```



In [90]:
```
sns.regplot(x="Flexibility", y="Pref Outcome", data=preference)
```

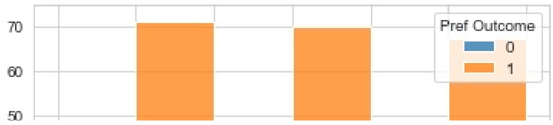
Out[90]:
```
<AxesSubplot:xlabel='Flexibility', ylabel='Pref Outcome'>
```

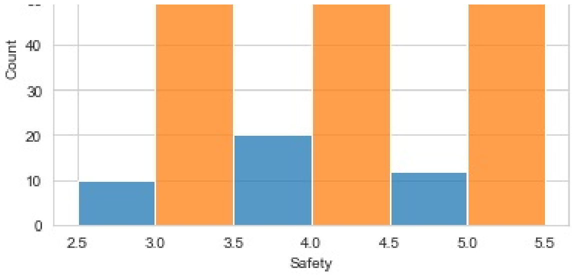


The majority of the employees who prefer WFH feel that flexibility is more in WFH.

In [91]:
```
# Safety v/s Pref Outcome

sns.histplot(preference, x="Safety", hue="Pref Outcome", discrete=True, multiple="dodge")
```

Out[91]:
```
<AxesSubplot:xlabel='Safety', ylabel='Count'>
```

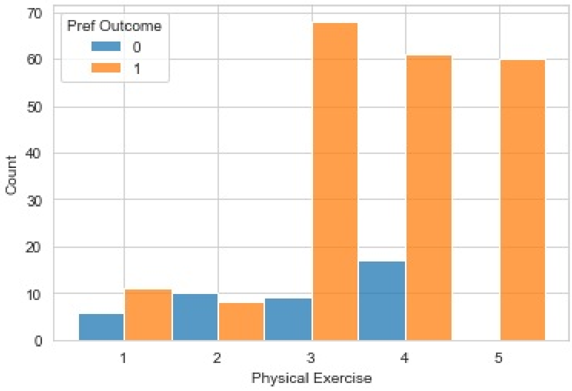




Majority of the employees feel that safety is more in a WFH scenario.

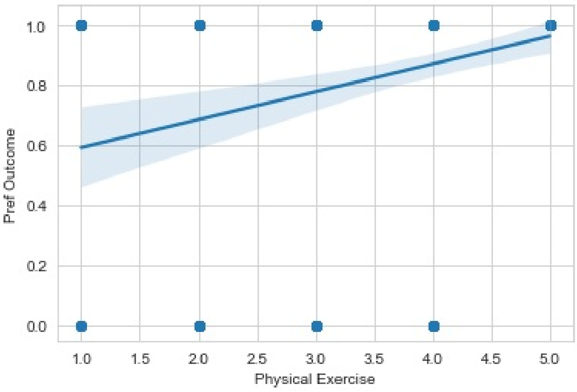
In [92]:
```
# Physical Exercise v/s Pref Outcome

sns.histplot(preference, x="Physical Exercise", hue="Pref Outcome",discrete=True, multiple="dodge")
```

Out[92]:
```
<AxesSubplot:xlabel='Physical Exercise', ylabel='Count'>
```



In [93]:
```
sns.regplot(x="Physical Exercise", y="Pref Outcome", data=preference)
```

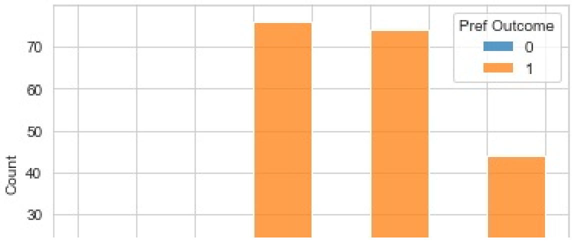
Out[93]:
```
<AxesSubplot:xlabel='Physical Exercise', ylabel='Pref Outcome'>
```

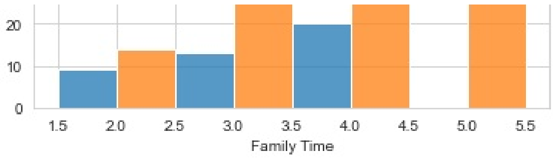


The employees who prefer work from home feel that they get more opportunities to be physically fit.

In [94]:
```
# Family Time v/s Pref Outcome

sns.histplot(preference, x="Family Time", hue="Pref Outcome",discrete=True,multiple="dodge")
```

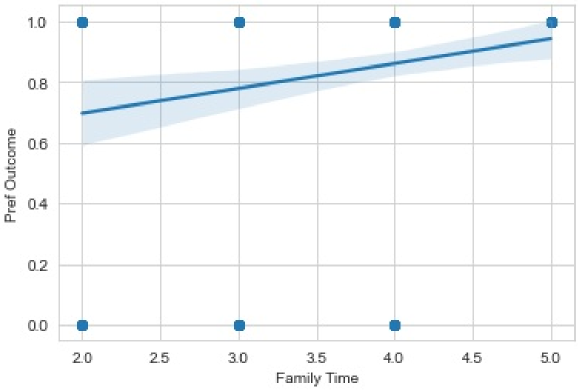
Out[94]:
```
<AxesSubplot:xlabel='Family Time', ylabel='Count'>
```





In [95]:
```
sns.regplot(x="Family Time", y="Pref Outcome", data=preference)
```

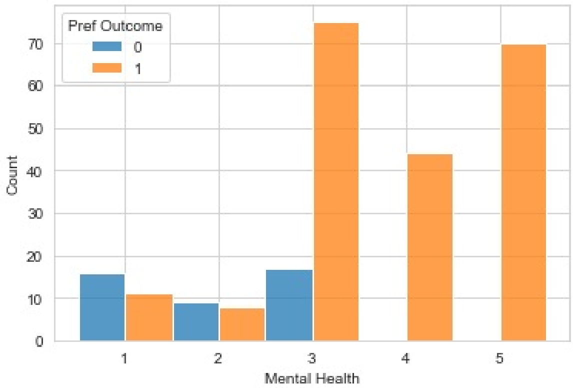
Out[95]:
```
<AxesSubplot:xlabel='Family Time', ylabel='Pref Outcome'>
```



Majority of the employees who want to WFH feel that they get more family time.

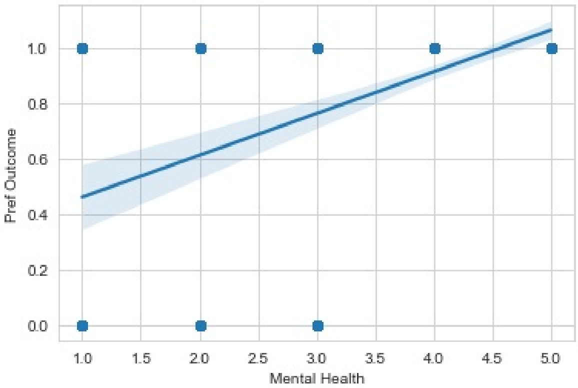
In [96]:
```
# Mental Health v/s Pref Outcome
sns.histplot(preference, x="Mental Health", hue="Pref Outcome",discrete=True,multiple="dodge")
```

Out[96]:
```
<AxesSubplot:xlabel='Mental Health', ylabel='Count'>
```



In [97]:
```
sns.regplot(x="Mental Health", y="Pref Outcome", data=preference)
```

Out[97]:
```
<AxesSubplot:xlabel='Mental Health', ylabel='Pref Outcome'>
```



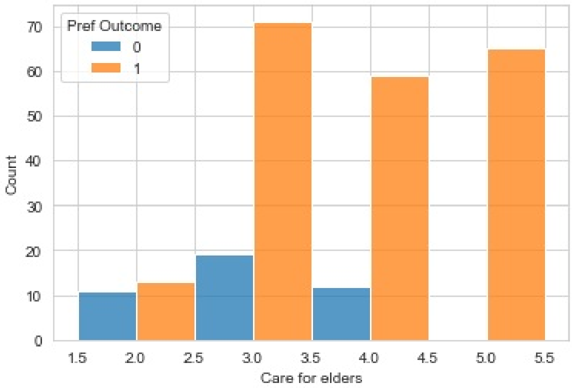
Majority of the employees who want to WFH feel that they are more mentally healthy.

In [98]:

```
# Care for elders v/s Pref Outcome

sns.histplot(preference, x="Care for elders", hue="Pref Outcome",discrete=True,multiple="dodge")
```

Out[98]:

```
<AxesSubplot:xlabel='Care for elders', ylabel='Count'>
```



Majority of the employees who want to WFH feel that they get more time to care for their elders.

Statistical Analysis and Hypothesis Testing

1. Correlation

In [99]:

```
preference.describe().T
```

Out[99]:

	count	mean	std	min	25%	50%	75%	max
Age	250.0	23.960	0.976840	22.0	23.0	24.0	25.0	25.0
No of Children	250.0	0.260	0.608177	0.0	0.0	0.0	0.0	2.0
Distance	250.0	19.116	6.460767	8.0	14.0	19.0	25.0	30.0
Cost of commute(per month)	250.0	975.708	329.792034	408.0	715.0	970.0	1276.0	1531.0
Flexibility	250.0	3.556	1.143720	1.0	3.0	4.0	4.0	5.0
Safety	250.0	3.992	0.801565	3.0	3.0	4.0	5.0	5.0
Physical Exercise	250.0	3.584	1.131399	1.0	3.0	4.0	4.0	5.0
Family Time	250.0	3.636	0.877830	2.0	3.0	4.0	4.0	5.0
Mental Health	250.0	3.452	1.264155	1.0	3.0	3.0	5.0	5.0
Care for elders	250.0	3.708	0.960430	2.0	3.0	4.0	5.0	5.0
Pref Outcome	250.0	0.832	0.374616	0.0	1.0	1.0	1.0	1.0

In [100…

```
# Checking the correlation between the features and the outcome variable.
pd.DataFrame(abs(preference.corr()['Pref Outcome']).sort_values(ascending = False))
```

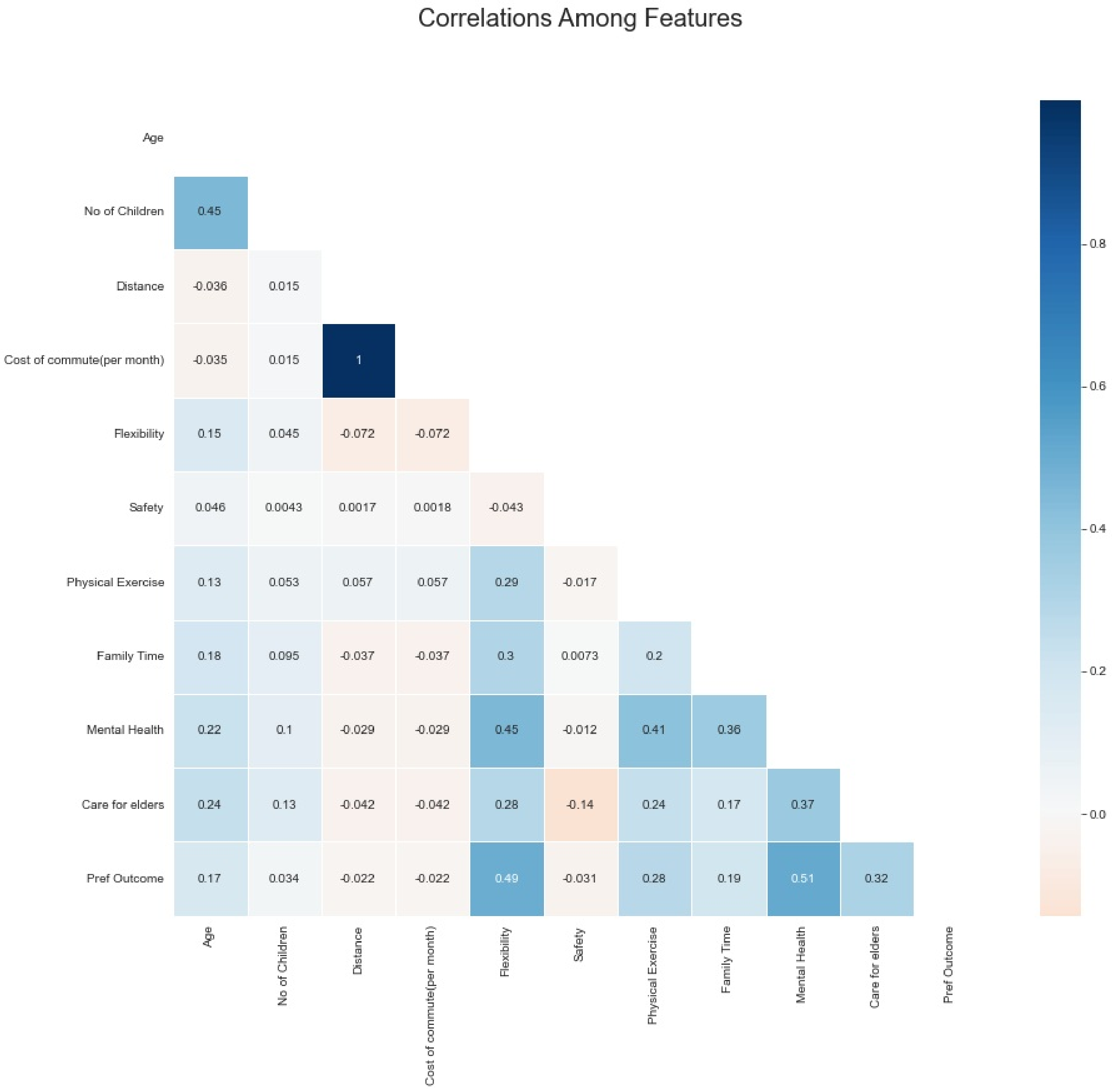
Out[100…

	Pref Outcome
Pref Outcome	1.000000
Mental Health	0.508685
Flexibility	0.490713
Care for elders	0.320756
Physical Exercise	0.279790
Family Time	0.191882
Age	0.168132
No of Children	0.033844
Safety	0.031243
Cost of commute(per month)	0.021853
Distance	0.021784

In [101…

```
mask = np.zeros_like(preference.corr(), dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
sns.set_style('whitegrid')
plt.subplots(figsize = (15,12))
sns.heatmap(preference.corr(),
            annot=True,
            mask = mask,
            cmap = 'RdBu', ## in order to reverse the bar replace "RdBu" with "RdBu_r"
            linewidths=.9,
            linecolor='white',
            fmt='.2g',
            center = 0,
            square=True)
plt.title("Correlations Among Features", y = 1.03,fontsize = 20, pad = 40);
```

```
/var/folders/2k/6r_xg74n77b1ytf4y98pm18w0000gn/T/ipykernel_1293/2789358653.py:1: DeprecationWarning: `np.bool` is
a deprecated alias for the builtin `bool`. To silence this warning, use `bool` by itself. Doing this will not mod
ify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.bool_` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#depr
ecations
  mask = np.zeros_like(preference.corr(), dtype=np.bool)
```



	Age	No of Children	Distance	Cost of commute(per month)	Flexibility	Safety	Physical Exercise	Family Time	Mental Health	Care for elders	Pref Outcome
Age											
No of Children	0.45										
Distance	-0.036	0.015									
Cost of commute(per month)	-0.035	0.015	1								
Flexibility	0.15	0.045	-0.072	-0.072							
Safety	0.046	0.0043	0.0017	0.0018	-0.043						
Physical Exercise	0.13	0.053	0.057	0.057	0.29	-0.017					
Family Time	0.18	0.095	-0.037	-0.037	0.3	0.0073	0.2				
Mental Health	0.22	0.1	-0.029	-0.029	0.45	-0.012	0.41	0.36			
Care for elders	0.24	0.13	-0.042	-0.042	0.28	-0.14	0.24	0.17	0.37		
Pref Outcome	0.17	0.034	-0.022	-0.022	0.49	-0.031	0.28	0.19	0.51	0.32	

2. Hypothesis Testing

H0 : The outcome variable is independent of the independent variables.

H1: The outcome variable is not independent of the independent variables.

To check these hypotheses, Chi Square test will be carried out.

In [102…

```
preference_chi=preference.copy()
```

In [103…

```
preference_chi['distance_bin']=pd.cut(preference_chi['Distance'], bins=[7,13,19,25,30],
                               labels=[1,2, 3, 4])
preference_chi['cost_of_commute_bin']=pd.cut(preference_chi['Cost of commute(per month)'], bins=[400,800,1200,160
                               labels=[1,2, 3])
preference_chi.drop(['Distance','Cost of commute(per month)'],axis=1, inplace=True)
```

In [104…

```
preference_chi.columns
```

Out[104…

```
Index(['Gender', 'Age', 'Marital Status', 'Children', 'No of Children',
       'Flexibility', 'Safety', 'Physical Exercise', 'Family Time',
       'Mental Health', 'Care for elders', 'Pref Outcome', 'distance_bin',
       'cost_of_commute_bin'],
      dtype='object')
```

In [105…

```
def crosstab_(feature):
    tab=pd.crosstab(preference_chi["Pref Outcome"], preference_chi[feature], margins = True,  margins_name="Total
    #print("Crosstabulation for ", feature, "\n", tab)
    # significance level
    alpha = 0.05
    # Calcualtion of Chisquare test statistics
    chi_square = 0
    rows = preference_chi["Pref Outcome"].unique()
    #print("rows",rows)
    columns = preference_chi[feature].unique()
    #print("columns", columns)
    for i in columns:
        for j in rows:
            O = tab[i][j]
            E = tab[i]['Total'] * tab['Total'][j] / tab['Total']['Total']
            chi_square += (O-E)**2/E
    # The p-value approach
    #print("For ", feature)
    p_value = 1 - stats.norm.cdf(chi_square, (len(rows)-1)*(len(columns)-1))
    conclusion = "Failed to reject the null hypothesis."
    if p_value <= alpha:
        conclusion = "Null Hypothesis is rejected."
    table.append([feature, p_value, conclusion])
    #print("chisquare-score is:", chi_square, " and p value is:", p_value)
    #print(conclusion)
```

In [106…

```
table=[]
cols=['Gender', 'Age', 'Marital Status', 'Children', 'No of Children',
       'Flexibility', 'Safety', 'Physical Exercise', 'Family Time',
       'Mental Health', 'Care for elders', 'distance_bin',
       'cost_of_commute_bin']
for x in cols:
    crosstab_(x)
chisq_pref = pd.DataFrame(table, columns=['features', 'p value', 'verdict'])
chisq_pref.sort_values(by=['p value'], inplace=True, ascending=False)
chisq_pref
```

Out[106…

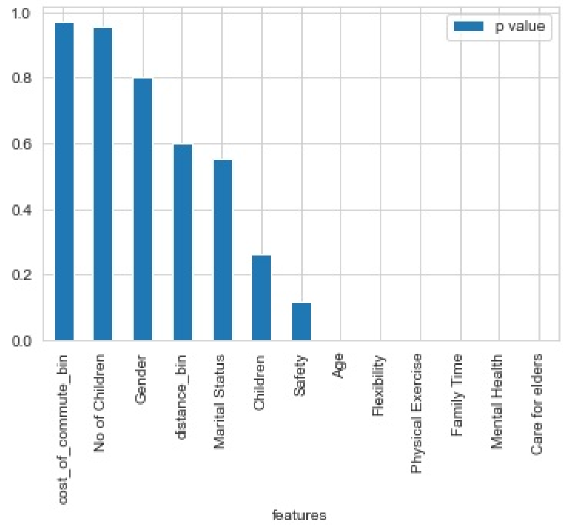
	features	p value	verdict
12	cost_of_commute_bin	0.970083	Failed to reject the null hypothesis.
4	No of Children	0.955030	Failed to reject the null hypothesis.
0	Gender	0.801147	Failed to reject the null hypothesis.
11	distance_bin	0.599311	Failed to reject the null hypothesis.
2	Marital Status	0.554433	Failed to reject the null hypothesis.
3	Children	0.260444	Failed to reject the null hypothesis.
6	Safety	0.117147	Failed to reject the null hypothesis.
1	Age	0.000000	Null Hypothesis is rejected.
5	Flexibility	0.000000	Null Hypothesis is rejected.
7	Physical Exercise	0.000000	Null Hypothesis is rejected.
8	Family Time	0.000000	Null Hypothesis is rejected.
9	Mental Health	0.000000	Null Hypothesis is rejected.
10	Care for elders	0.000000	Null Hypothesis is rejected.

In [107…

```python
chisq_pref.plot(x ='features', y='p value', kind = 'bar')
```

Out[107…

```
<AxesSubplot:xlabel='features'>
```



From the above results, it is clear that the factors which affect a person's preference are Age, Flexibolity, Physical Exercise, Family Time, Mental Health and Care for elders. For Cost of Commute, No of children, Gender, Distance, Safety, Marital Status and Children there is no evidence to reject the Null Hypothesis that the outcome variable and each of these independent variables are independent. In conclusion, for Age, Flexibolity, Physical Exercise, Family Time, Mental Health and Care for elders there is evidence to reject the null hypothesis that the outcome variable and each of these factors are dependent. Thus, these features contribute to the preference of an individual working from home.

Modelling

In [109…

```python
le=pre.LabelEncoder()
lt_1= preference.select_dtypes(exclude = ['float','int']).columns.to_list()
for x in lt_1:
    preference[x]=le.fit_transform(preference[x])
preference.head()
```

Out[109…

	Gender	Age	Marital Status	Children	No of Children	Distance	Cost of commute(per month)	Flexibility	Safety	Physical Exercise	Family Time	Mental Health	Care for elders	Pref Outcome
0	0	23	1	0	0	25	1276	5	4	5	4	5	3	1
1	1	24	1	0	0	16	817	5	3	4	4	5	4	1
2	1	24	1	0	0	9	459	4	4	3	3	4	5	1
3	1	22	1	0	0	30	1531	2	3	1	3	3	4	0
4	1	25	0	1	1	22	1123	5	4	3	4	5	4	1

In [110…

```python
features=preference.drop('Pref Outcome', axis=1)
outcome=preference['Pref Outcome']
df_pref=pre.minmax_scale(features)
```

In [111…

```python
#To balance the data

smote = SMOTE()

# fit predictor and target variable
x_smote, y_smote = smote.fit_resample(features, outcome)
```

In [112…

```python
X_train, X_test, y_train, y_test = train_test_split(x_smote, y_smote, test_size = 0.3, random_state = 0)
```

In [113…

```python
X_train.shape, X_test.shape
```

Out[113…

```
((291, 13), (125, 13))
```

Decision Tree

In [114…

```
dtc = DecisionTreeClassifier(criterion = 'entropy', max_depth = 5)
```

In [115…

```
dtc.fit(X_train, y_train)
```

Out[115…

```
DecisionTreeClassifier(criterion='entropy', max_depth=5)
```

In [116…

```
y_pred_DT = dtc.predict(X_test)
```

In [117…

```
accuracy_score(y_test,y_pred_DT)
```

Out[117…

```
0.952
```

In [118…

```
cmDT = confusion_matrix(y_test, y_pred_DT)

print('Confusion matrix\n\n', cmDT)
```

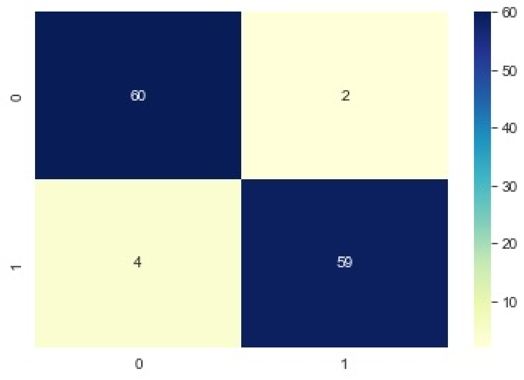
```
Confusion matrix

 [[60  2]
 [ 4 59]]
```

In [119…

```
cm_matrix = pd.DataFrame(data=cmDT)

sns.heatmap(cm_matrix, annot=True, fmt='d', cmap='YlGnBu')
```

Out[119…

```
<AxesSubplot:>
```



In [120…

```
print(classification_report(y_test, y_pred_DT))
```

```
              precision    recall  f1-score   support

           0       0.94      0.97      0.95        62
           1       0.97      0.94      0.95        63

    accuracy                           0.95       125
   macro avg       0.95      0.95      0.95       125
weighted avg       0.95      0.95      0.95       125
```

The accuracy of prediction is 95%. The F1 score for class 0 and class 1 are both 95%.

Conclusion

The objectives of this exercise were :-

- To find out whether working from home affects employee productivity.
- To find out what do employees prefer and why.

Taking the first point, from the study carried out above, it is amply clear that majority of the employees had improved productivity, which has also been brought out in the research papers that were referred. The reasons which contributed to employee productvity are :-

- No of children
- No of working hrs per week
- Hrs spent in meetings
- Average no of meetings attended per week
- Efficacy of team meetings
- Social interactions
- Innovative climate
- Uninterrupted working hrs per week
- Work environment
- Stress
- Overwork
- Motivation
- Salary

The pointers are mainly towards all factors which ensure a proper balance between work and home. No doubt, the commitment for work is there - which is indicated by the emphasis towards encouragement towards innovation, positive work culture, and undisturbed time towards work. At the same time no of children is a factor contributing to higher productivity, probably because presence of children makes it more imperative to retain the job.

For the second point, majority of the employees prefer WFH regimen. The reasons that came to fore are :-

- Age
- Flexibility
- Physical Exercise
- Family Time
- Mental Health
- Care for elders

Basically, the underlying factors points towards more availability of time which an individual is channeling towards issues of personal priority i,e, better work-life balance.

Business Context

WFH has come into sharp focus after hundreds of professionals were forced to adopt it full-time after Covid-19 hit. It has forced companies to adapt, improvise and innovate ways of working. However, WFH has been followed by many of these companies for years - been provided to persons based on distance from workplace, ability to monitor progress online, ability to participate in business meetings effectively etc. In the US itself, even before the pandemic, there were over 5 million employees working from home at least half the time. The bottom line for providing WFH has been performance, while balancing personal requirement, to whatever extent allowable by the company.

However, ever since Covid-19 struck operations, the companies had no option but to improvise on this arrangement. As a result, all operations shifted online. This brought forth advantages that accrued from this new modus operandi such as sharply reduced requirement of leased workspace. This has to be viewed in light of varying efficiencies of employees. Achieving same overall productivity is a team effort, and when people are not under the watchful eye of the supervisor, there can be trust issues within the team. One of the factors which was brought out in the papers was inability to know what other team members were working on due to lack of social interaction. Working from Home requires more self-discipline and it can impact team productivity.

Such and other factors brought focus on the managerial style of the manager / supervisor. A good manager would be able to harness the potential of the team while keeping track of factors affecting them inclementally. Thats why during this WFH stage, empathy among employees increased, deliberate informal sessions and No-meeting-Fridays were implemented in conjunction with other measures. Regular feedback from the employees was taken on their well being. As a result mental health came into prominence. Better employees could manage personal and social impact better.

Now that people were home, creating a concerted schedule to address work requirements, disciplining oneself to meet punctuality requirements, belying the illusion that the person is more available for home requiremnts during working hours, meeting business targets in a non-office environment etc were faced.

All these factors impinged on productivity of the teams and consequent business output of the company. Overall, it was found that for WFH to be beneficial to the company, apart from motivated employees it required effective substitute for the social disconnect which people started feeling a lack of, focussed meetings (it is easy to get carried away in discussions when not facing time constraints), peer interaction to counter loneliness, wellness interactions, periodic encouragement etc.

Change is often tough, but it can also be very rewarding. The increase in employees working from home will have impact, both good and bad, on individual employees, the organizations they work for and the larger economy. For the companies, the struggle can simply mean changing the process and putting the right collaboration/communication tools in place.

Companies such as Microsoft kept a track of its employees and were able to address their concerns and changing requirements effectively. As a result they were able to ensure consistent productivity. However, the same may not be possible by similarly impacted companies with the same effectiveness. Hence, companies need to take a hard and fast look at progress of each team and institute measures to tackle issues thrown up by WFH in the new scenario.

What worked before may not work anymore. Some organizations will eventually have their employees return to an office or building. Others will adapt to a new way of business with WFH employees. One is not necessarily better than the other. If we were to look for a silver lining, the pandemic gave many companies the opportunity to test the waters and discover new ways of doing business that may continue to work for them—and their employees—long into the future.

References

1. C. Miller, P. Rodeghero, M. -A. Storey, D. Ford and T. Zimmermann, "Survey Instruments for "How Was Your Weekend?" Software Development Teams Working from Home During COVID-19," 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), 2021, pp. 223-223, doi: 10.1109/ICSE-Companion52605.2021.00101.

2. K. D. Flack, S. R. Lenton, A. Murphy and A. Pilkington, "How we made homeworking work for us," IEE Colloquium on The Home as an Office, 1996, pp. 6/1-6/3, doi: 10.1049/ic:19960271.

3. M. L. Watkins, "Working from home," University as a Bridge from Technology to Society. IEEE International Symposium on Technology and Society (Cat. No.00CH37043), 2000, pp. 127-132, doi: 10.1109/ISTAS.2000.915590.

4. M. Blake, "Information for home-based teleworkers," IEE Colloquium on The Home as an Office, 1996, pp. 4/1-4/6, doi: 10.1049/ic:19960269.

5. S. Jaffe, "Work from home During and After COVID-19," 2021 IEEE/ACM 8th International Workshop on Software Engineering Research and Industrial Practice (SER&IP), 2021, pp. 28-28, doi: 10.1109/SER-IP52554.2021.00012.

6. E. Clark, "Telecommuting and working from home," IPCC 98. Contemporary Renaissance: Changing the Way we Communicate. Proceedings 1998 IEEE International Professional Communication Conference (Cat. No.98CH36332), 1998, pp. 21-25 vol.2, doi: 10.1109/IPCC.1998.722074.

7. J. Butler and S. Jaffe, "Challenges and Gratitude: A Diary Study of Software Engineers Working From Home During Covid-19 Pandemic," 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 2021, pp. 362-363, doi: 10.1109/ICSE-SEIP52600.2021.00047.

8. M. Maternaghan, "Workplace 2000," IEE Colloquium on The Home as an Office, 1996, pp. 7/1-7/5, doi: 10.1049/ic:19960272.

9. T. Golden, "Technology and the balance of work family conflict: an investigation into the role of telecommuting," IEMC '03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change, 2003, pp. 439-442, doi: 10.1109/IEMC.2003.1252310.

10. L. Ahuja, A. Rana and S. Gupta, "Security & Privacy Model for Work from Home Paradigm," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 1351-1355, doi: 10.1109/ICRITO48877.2020.9197773.

11. Longqi Yang, Sonia Jaffe, David Holtz, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, CJ Lee, Brent Hecht and Jaime Teevan, “How Work From Home Affects Collaboration: A Large-Scale Study of Information Workers in a Natural Experiment During COVID-19”, Microsoft Corporation

12. Denae Ford, Margaret-Anne Storey, Thomas Zimmermann, Christian Bird, Sonia Jaffe, Chandra Maddila, Jenna L. Butler, Brian Houck, Nachiappan Nagappan, A Tale of Two Cities: Software DevelopersWorking from Home During the COVID-19 Pandemic”, arXiv:2008.11147v3 [cs.SE] 10 Sep 2021

13. Esra Thorstensson, “The Influence of Working from Home on Employees' Productivity : Comparative document analysis between the years 2000 and 2019-2020”, Karlstad Business School publication

14. Prithwiraj (Raj) Choudhury, Cirrus Foroughi, Barbara Larson, “Work-From-Anywhere: The Productivity Effects of Geographic Flexibility”, Harvard Business School publication

15. Bloom, Nicholas, J. Joseph Beaulieu, James Liang, Donald John Roberts and Zhichun Jenny Ying. “Does Working from Home Work? Evidence from a Chinese Experiment.” Kauffman: Large Research Projects - NBER (Topic) (2013): n. pag. 54

16. Øystein Tønnessen, Amandeep Dhir, Bjørn-Tore Flåten, Digital knowledge sharing and creative performance: Work from home during the COVID-19 pandemic, Technological Forecasting and Social Change, Volume 170, 2021, 120866, ISSN 0040-1625,

https://doi.org/10.1016/j.techfore.2021.120866.

17. Kramer, Amit and Karen Z. Kramer. “The potential impact of the Covid-19 pandemic on occupational status, work from home, and occupational mobility.” Journal of Vocational Behavior 119 (2020): 103442 - 103442.

18. Galanti, Teresa MPsyc; Guidetti, Gloria PhD; Mazzei, Elisabetta MPsyc; Zappalà, Salvatore PhD; Toscano, Ferdinando MPsyc Work From Home During the COVID-19 Outbreak, Journal of Occupational and Environmental Medicine: July 2021 - Volume 63 - Issue 7 - p e426-e432 doi: 10.1097/JOM.0000000000002236

19. E. Thorstensson, ‘The Influence of Working from Home on Employees’ Productivity : Comparative document analysis between the years 2000 and 2019-2020’, Dissertation, 2020.

20. https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223

21. https://www.forbes.com/sites/shephyken/2021/02/28/the-impact-of-the-virtual-work-from-home-workforce/?sh=1febafc82873

22. Shrivastava A, Sharma MK, Marimuthu P. Internet use at workplaces and its effects on working style in indian context: An exploration. Indian J Occup Environ Med. 2016;20(2):88-94. doi:10.4103/0019-5278.197531

23. https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

24. https://support.optimizely.com/hc/en-us/articles/4410282998541-Design-an-effective-hypothesis

25. https://online.hbs.edu/blog/post/hypothesis-testing

26. Teodorovicz T, Sadun R, Kun AL, Shaer O. Working from Home during COVID19: Evidence from Time-Use Studies, 2021, Working Paper, Harvard Business School.

27. Gibbs M, Mengel F, and Siemroth C. Work from Home & Productivity: Evidence from Personnel & Analytics Data on IT Professionals, 2021, Working Paper, Becker Friedman Institute.