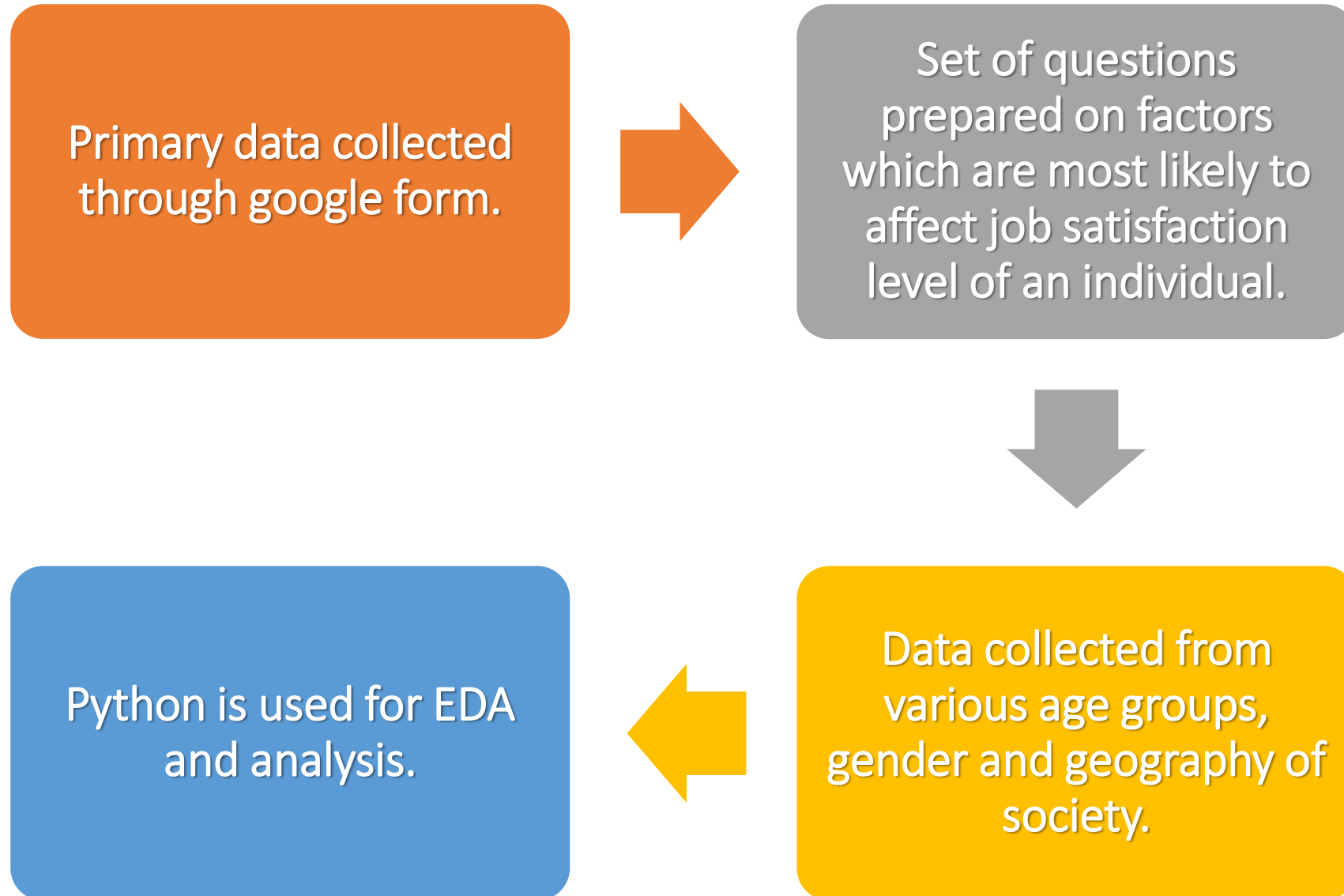**GROUP MEMBERS:**

1. C23036-TANMOY CHANDRA
2. C23010-DIPAM SARKAR
3. C23005-ANGSHUMAN PANDEY

INTRODUCTION

Job satisfaction is defined as the level of contentment employees feel with their job. This goes beyond their daily duties to cover satisfaction with team members and the impact of their job on employees' personal lives. This project provides insights on various factors which decide job satisfaction level of an individual. The insights will help an individual to focus/improve his/her job satisfaction level.
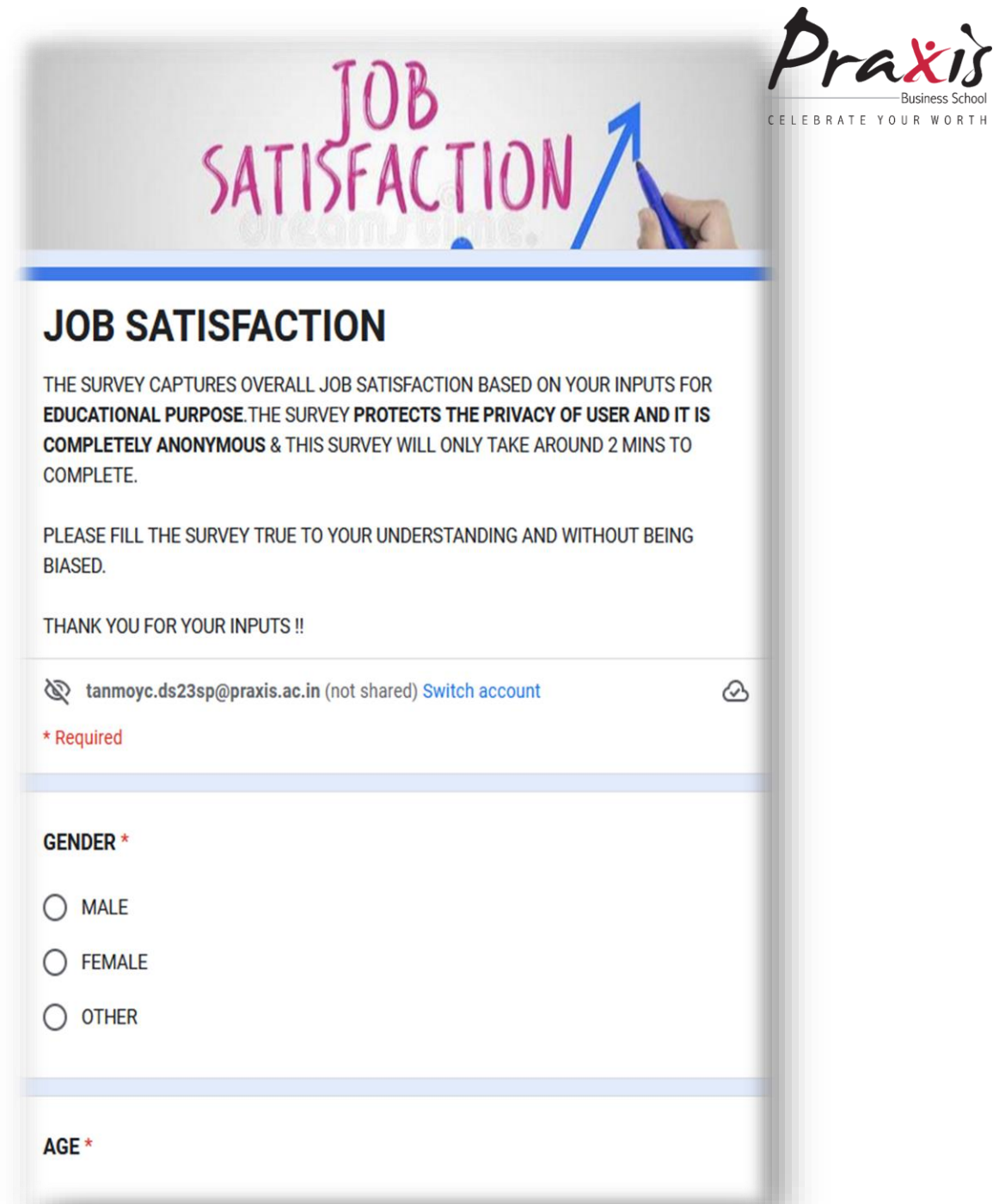
DATA COLLECTION AND METHODOLOGY

Primary data collected through google form.

Set of questions prepared on factors which are most likely to affect job satisfaction level of an individual.

Data collected from various age groups, gender and geography of society.

Python is used for EDA and analysis.

Snapshot of our google form used for primary data collection. Alternatively the whole form can be viewed using the following hyperlink:-

JOB SATISFACTION FORM

JOB SATISFACTION

THE SURVEY CAPTURES OVERALL JOB SATISFACTION BASED ON YOUR INPUTS FOR **EDUCATIONAL PURPOSE**.THE SURVEY **PROTECTS THE PRIVACY OF USER AND IT IS COMPLETELY ANONYMOUS** & THIS SURVEY WILL ONLY TAKE AROUND 2 MINS TO COMPLETE.

PLEASE FILL THE SURVEY TRUE TO YOUR UNDERSTANDING AND WITHOUT BEING BIASED.

THANK YOU FOR YOUR INPUTS !!

tanmoyc.ds23sp@praxis.ac.in (not shared) Switch account

* Required

GENDER *

○ MALE

○ FEMALE

○ OTHER

AGE *

DATA DESCRIPTION
QUESTIONS ASKED DURING SURVEY –

GENDER-

Data Type – Categorical (Nominal)

Answer type - Short answer text

AGE-

Data Type – Numerical (Discrete)

Answer type - Short answer text

WORK CITY –

Data Type – Categorical (Nominal)

Answer type - Short answer text

IS WORK CITY DIFFERENT FROM YOUR HOME CITY?

Data Type – Categorical (Nominal)

Answer type - Short answer text

DOMAIN/INDUSTRY

Data Type – Categorical (Nominal)

Answer type - Short answer text

JOB ROLE/DESIGNATION

Data Type – Categorical (Nominal)

Answer type - Short answer text

WORKING EXP (YEARS)

Data Type – Numerical (Continuous)

Answer type - Short answer text

AVG WORKING HOURS

Data Type – Numerical (Continuous)

Answer type - Short answer text

DATA DESCRIPTION:
QUESTIONS ASKED DURING SURVEY –

Praxis
Business School
CELEBRATE YOUR WORTH

SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?

(HIGHEST=5 LOWEST=1)

Data Type – Numerical (Continuous)
Answer type - Short answer text

INCOME (in LPA)

Data Type – Numerical (Continuous)
Answer type - Short answer text

RAPPORT WITH COLLEAGUES/TEAM?

Data Type – Numerical (Continuous)
Answer type - Short answer text

HAVING CHILDREN BELOW AGE OF 12 ?

Data Type – Categorical (Nominal)
Answer type - Short answer text

DOMAIN/INDUSTRY

Data Type – Categorical (Nominal)
Answer type - Short answer text

JOB ROLE/DESIGNATION

Data Type – Categorical (Nominal)
Answer type - Short answer text

WORKING EXP (YEARS)

Data Type – Numerical (Continuous)
Answer type - Short answer text

AVG WORKING HOURS

Data Type – Numerical (Continuous)
Answer type - Short answer text

DATA DESCRIPTION
QUESTIONS ASKED DURING SURVEY –

MARITAL STATUS –

      Data Type – Categorical (Nominal)

      Answer type - Short answer text

NUMBER OF FAMILY MEMBERS RESIDING WITH YOU?

      Data Type – Numerical (Discrete)

      Answer type - Short answer text

TIME TAKEN TO REACH OFFICE (HH:MM)

      Data Type – Categorical (Nominal)

      Answer type - Short answer text

CURRENT WORKING MODE

      Data Type – Categorical (Nominal)

      Answer type - Short answer text

Exploratory Data Analysis

Exploratory Data Analysis or EDA is a process where we analyze the data characteristic, data behavior & relation between the features and the target variable. Therefore we are going to perform each and every step of EDA and find out what data tells us.

Lets find out !!!

1) Finding data types of all the columns:-

- There are total 26 features having object data type which generally tells us categorical information about the dataset and 7 features having float data type, which are numerical and measurable.

- The shape of dataset is (105,33) which means 105 records with 33 variables.

```
df.dtypes

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
```
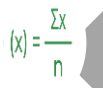
```
df.shape

(105, 33)
```

2) Describe(Quantitative Features) is used to calculate the descriptive statistical analysis both for the continuous and categorical variables. For numerical variables, it summarizes the measures of central tendency and measures of distribution of each numeric column of the dataset, which includes count, mean, standard deviation, min, 25 percentile, 50 percentile/median, 75 percentile and max of each column and for categorical features, count unique, top and freq is calculated.

- Measures calculated for Numerical variables:-

Count - Count of total number of records present under each numerical column.

$(x) = \dfrac{\Sigma x}{n}$ Mean - Calculates the average value of each numerical column.

Std - Standard Deviation shows how much the data points are dispersed from the mean. In other words if the standard deviation is on the lower side that defines the lesser no of extreme values present in the feature.

Min and Max – Measures the minimum and maximum value present for each numerical column in the dataset. Large difference between and min and max value shows the variability is very high for that feature

25%, 50%, 75% - 25% or Q1 tells that 25 % of the datapoints for that variable are below that value when ordered ascendingly. Similar goes for 50% or Q2 and 75% or Q3 . Also 50% gives the median value of that variable.

The following table shows the count, mean , standard deviation, min, 25% percentile, 50% percentile, 75% percentile and max value of each numerical column.

	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 ?	PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL	PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL	PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL .1
count	105.000000	103.000000	105.000000	105.000000	21.000000	50.000000	34.000000
mean	6.578095	0.942718	3.640000	3.479048	2.857143	3.506000	3.097059
std	4.318675	0.886345	0.868531	1.094013	1.000785	1.080893	0.933707
min	0.900000	0.000000	1.000000	1.300000	1.400000	0.400000	0.500000
25%	3.000000	0.300000	3.300000	2.600000	2.300000	2.725000	2.525000
50%	5.400000	0.800000	3.700000	3.700000	2.800000	3.600000	3.150000
75%	8.700000	1.250000	4.400000	4.500000	3.700000	4.400000	3.775000
max	17.700000	7.000000	5.000000	5.000000	4.900000	5.000000	4.800000

Here the table shows the count, unique, top and freq of each categorical column.

Count- Count of records present for the feature.

Unique- Number of unique values present in that feature.

Top- Most number of occurred value present in the feature or represents the mode of the feature.

Freq - No of times top value occurred in the that feature.

	Timestamp	AGE	GENDER	WORK CITY	IS WORK CITY DIFFERENT FROM YOUR HOME CITY?	DOMAIN/IN DUSTRY	JOB ROLE/DESIG NATION	CURRENT WORKING MODE	PREFERRED WORKING MODE	WORKING EXP (YEARS)	AVG WORKING HOURS	MARITAL STATUS	HAVING CHILDREN BELOW AGE OF 12 ?	NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?
count	105	105	105	105	105	105	105	105	105	105	105	105	105	105
unique	105	22	2	42	2	20	88	3	3	44	30	3	2	17
top	3/17/2023 12:25:52	25	MALE	Kolkata	No	IT	Teacher	WORK FROM OFFICE	WORK FROM OFFICE	2	8	SINGLE	No	3
freq	1	17	64	26	57	29	5	82	50	16	36	74	94	25

	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]	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]	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]
count	21	21	21	21	50	50	50	50	34	34	34	34
unique	4	4	4	4	4	4	4	4	4	4	4	4
top	Time lost in travelling gets saved	Financially more viabale	Physically less tiring	Increased time with loved ones	Better Working environment	Better Interaction with team/colleagues	Better Interaction with team/colleagues	Better Working environment	Work life balance	Higher Productivity	Work life balance	Flexible working location
freq	8	8	8	7	23	18	20	19	15	12	13	14

3) Data Cleaning

In this dataset we have 26 categorical features, so our first goal is to reduce the dimension for further machine learning process and analysis of data. Here we have 5 different columns ('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]' and 'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL') for each preferred working mode as we have collected data of the above columns for each preferred working mode. So we have merged all columns in their respective headers which are '1st_Choice', '2nd_Choice', '3rd_Choice', '4th_Choice' and 'PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW' column.

```python
In [6]: #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'].fillna(0.)+
        df['PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL '].fillna(0.)+
        df['PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL .1'].fillna(0.)
```

```python
In [7]: #Merging different choices to their respective choice columns
        df['1st_Choice']=df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [1st choice]'].fillna("")+
        df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [1st Choice]'].fillna("")+
        df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [1st Choice]'].fillna("")
        df['2nd_Choice']=df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [2nd choice]'].fillna("")+
        df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [2nd Choice]'].fillna("")+
        df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [2nd Choice]'].fillna("")
        df['3rd_Choice']=df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [3rd choice]'].fillna("")+
        df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [3rd Choice]'].fillna("")+
        df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [3rd Choice]'].fillna("")
        df['4th_Choice']=df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?  [4th Choice]'].fillna("")+
        df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ?   [4th Choice]'].fillna("")+
        df['REASONS BEHIND SELECTING YOUR "PREFERRED WORKING MODE" ? [4th Choice]'].fillna("")
```

- After merging the features , existing features were removed.
- We changed the datatypes of features like age, working hours , number of family members, time taken to reach office having object data type to integer/float as they were numerical data.
- While changing the datatype we replaced few irrelevant values with relevant values like changing "1yr 5months" to 1.5 as integer data type does not contain any strings in it.
- Removed extra white spaces in data values.
- Replaced different spelling and cases of same type with same single value, like replacing "Ahmedabad", "Ahemdabad" and "Sanandahmedabad" with "AHMEDABAD".
- In working hours feature we converted string values into int and where values were like "8-9", we replaced it with 8.5.
- Dropping a record with erroneous data where NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ? was entered as "Hybrid".
- Finally after cleaning these irrelevant values we changed the datatype of few features.

```python
#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('None','0')
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('Zero','0')
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('members','')
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('Meandspouse')
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('Notcapableto')
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].str.replace('Null','0')
df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?'].astype('float64')

df.drop(df[df['NUMBER OF FAMILY MEMBERS RESIDING WITH YOU ?']=='Hybrid '].index,inplace=True)
```

After Data Cleaning :-

- There are total 13 features having object data type 8 features having float data type and 1 feature having int data type.
- The shape of dataset is now (104,22) which means 104 records with 22 variables.

```
df.dtypes
```

AGE	int64
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)	float64
AVG WORKING HOURS	float64
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 ?	float64
PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW	float64
1st_Choice	object
2nd_Choice	object
3rd_Choice	object
4th_Choice	object

Count Plot/Bar Plot

Count plot is used to count number of records in each category of categorical variables. The main purpose of a count plot is to provide a visual representation of the distribution of a categorical variable.

- For Domain/Industry feature we can analyse out of 104 records almost 30 people are working in IT industry, the 2nd highest category is Finance/Banking sector, while the lowest values recorded for ecommerce sector.
- For working city, Kolkata is having most number of records and that is near about 50 while the lowest is Hyderabad and Sambalpur.

INDUSTRY

CITY
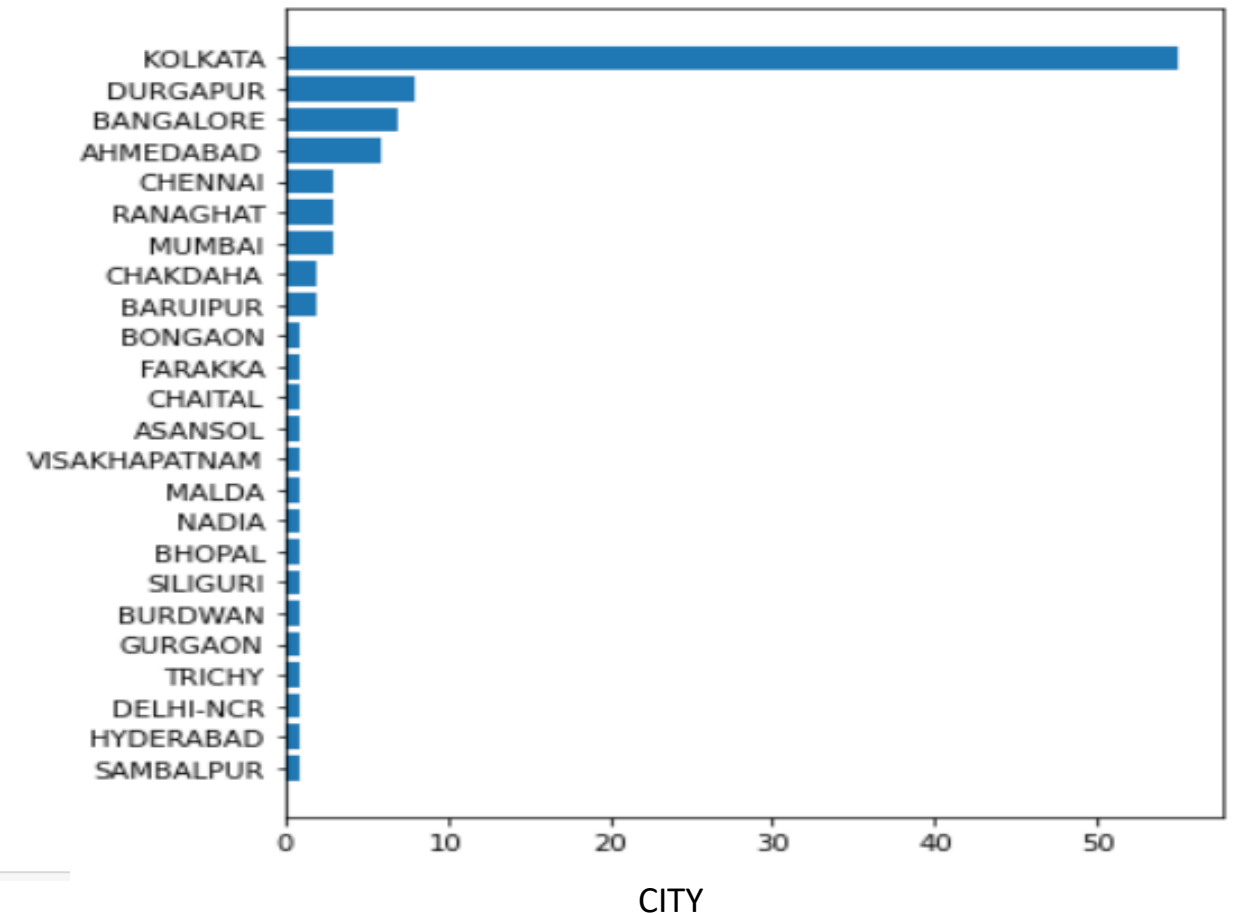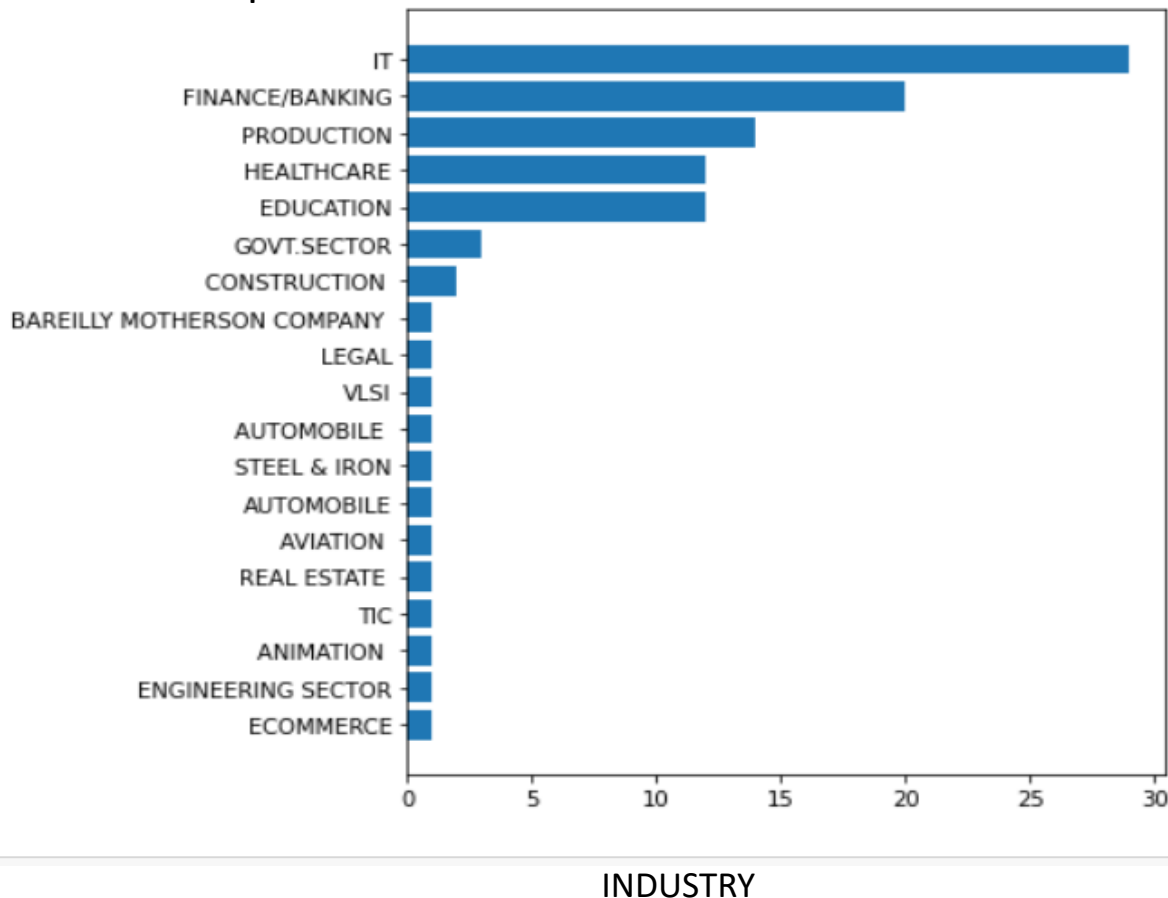
- For current working mode we can see that the work from office count is almost 80% of overall data present in the dataset. As the data is recorded in recent times so we can analyse that the companies are now reducing their work from home facility .
- For Gender no of male present in the dataset is 60 and female is 45.
- Surprisingly, most preferred working mode of people is "work from office" and least preferred working mode of people is "work from home" and Hybrid working mode is in the 2nd most preferred working mode.
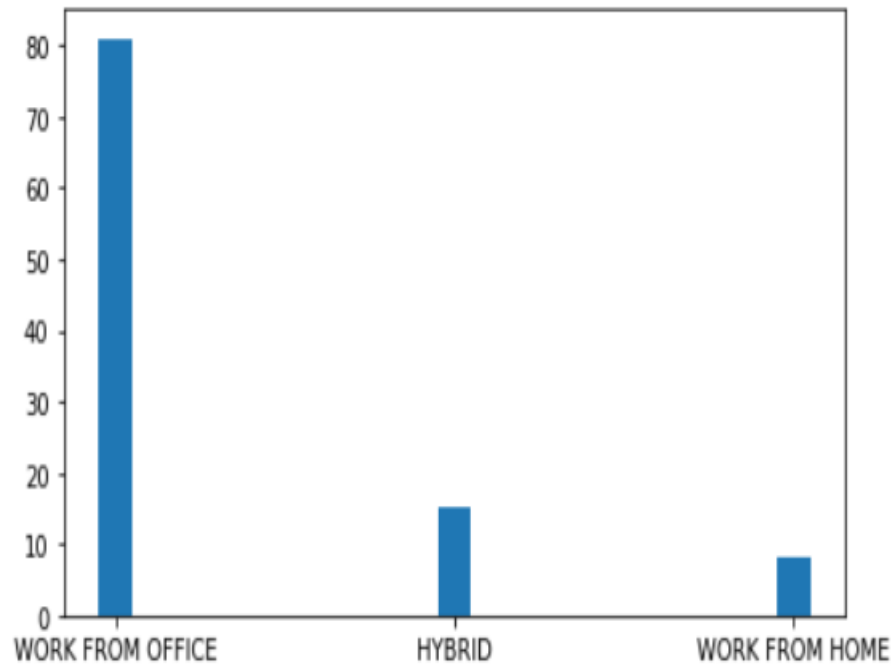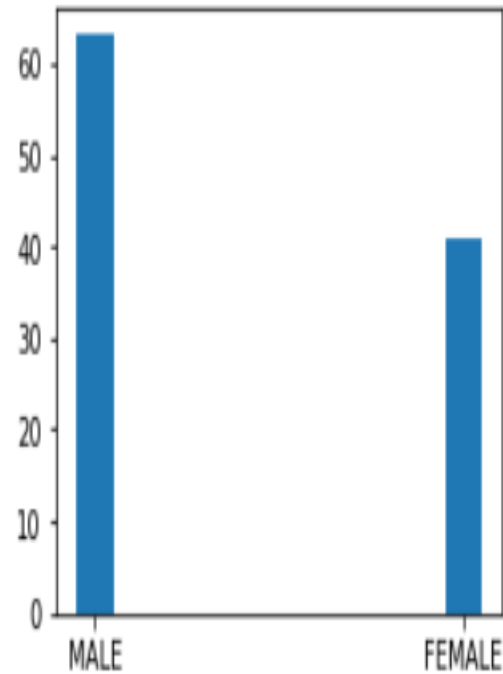
CURRENT WORKING MODE   GENDER   PREFERRED WORKING MODE
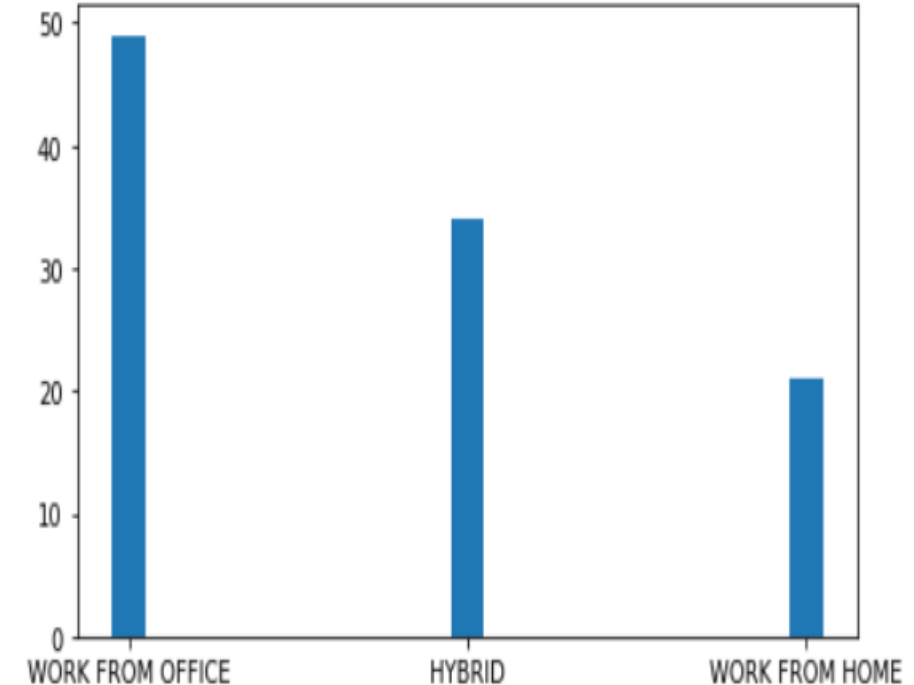
Here are the top two reasons for each preferred working mode based on response.

Preferred working mode = "Work from Office" :-

▪ Earlier we have seen that most of user's preferable working mode is work from office and the reason topping the 1st choice is "Better Working Environment" and the reason topping the 2nd choice is "Better interaction with team/colleagues".
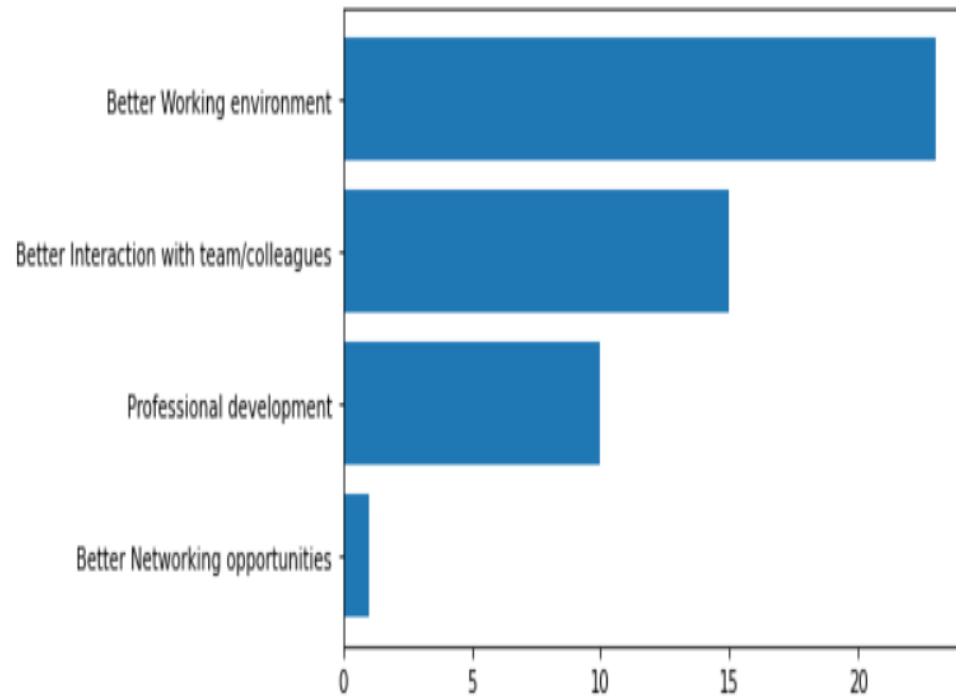
WORK FROM OFFICE - 1st Choice
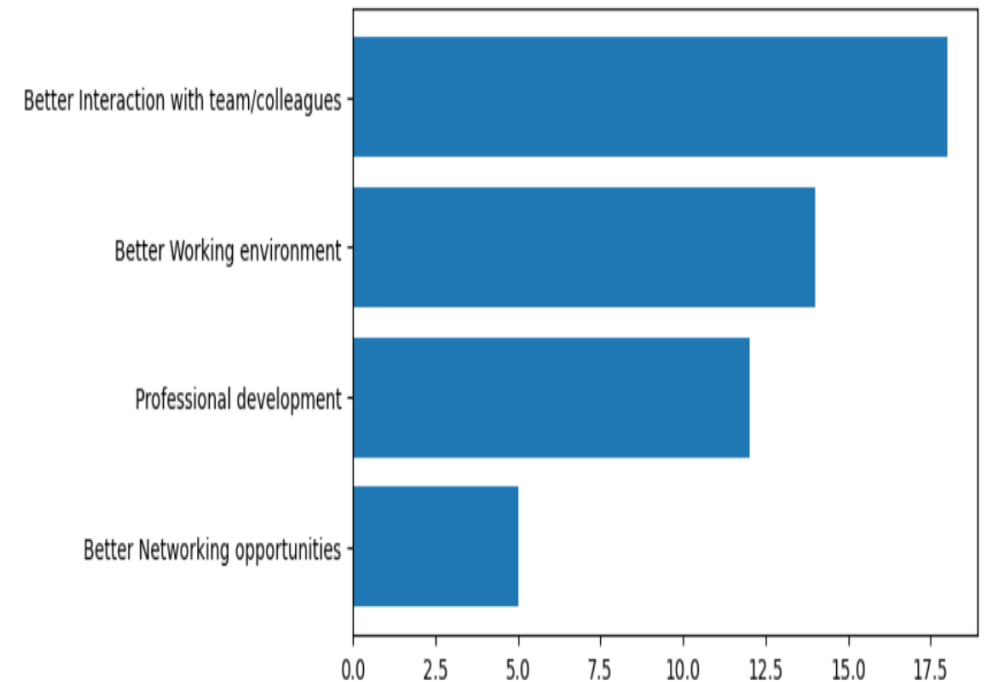
WORK FROM OFFICE – 2nd Choice

Preferred working mode = "Work from Hybrid" :-

- Earlier we have seen that most of user's 2nd most preferable working mode is Hybrid work mode and the reason topping the 1st choice is "Work life balance" and the reason topping the 2nd choice is "Higher Productivity".

Preferred working mode = "Work from Home" :-

- Earlier we have seen that least preferable working mode is work from home and the reason topping the 1st choice of the respondents is "Time lost in travelling gets used" and the reason topping the 2nd choice is "Financially more viable".
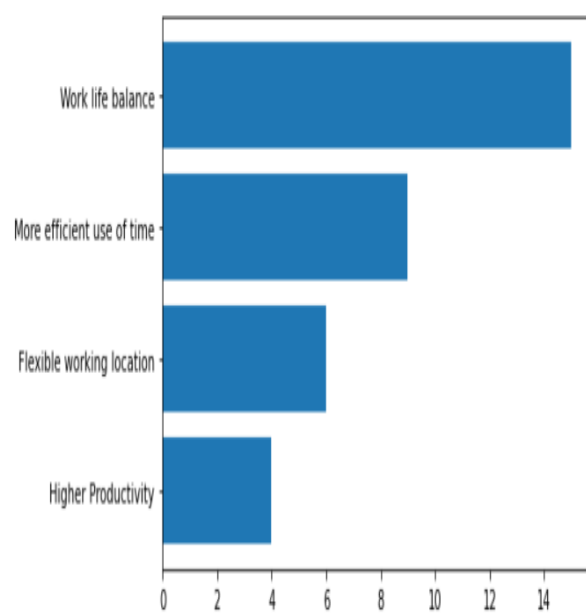
WORK FROM HYBRID - 1st Choice
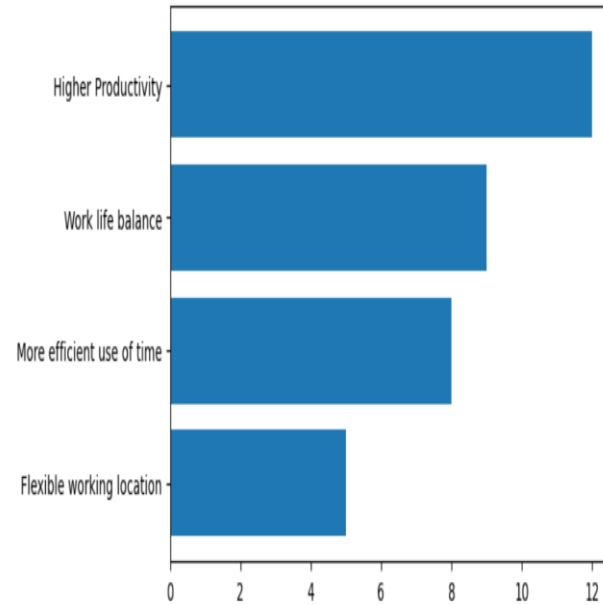
WORK FROM HYBRID – 2nd Choice
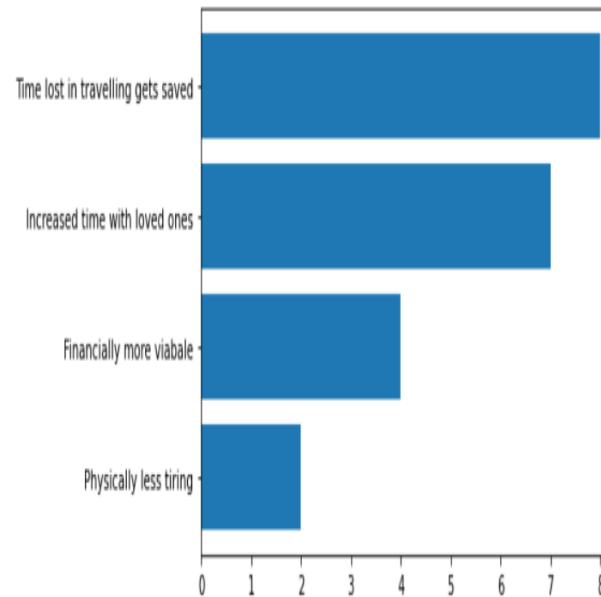
WORK FROM HOME - 1st Choice
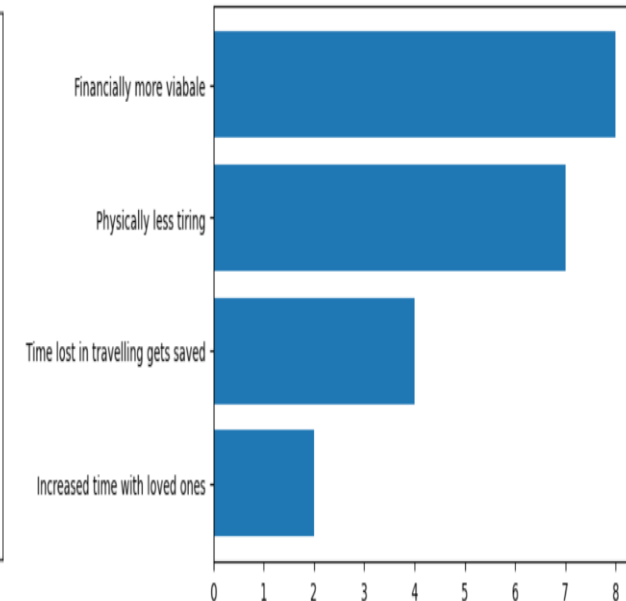
WORK FROM HOME – 2nd Choice
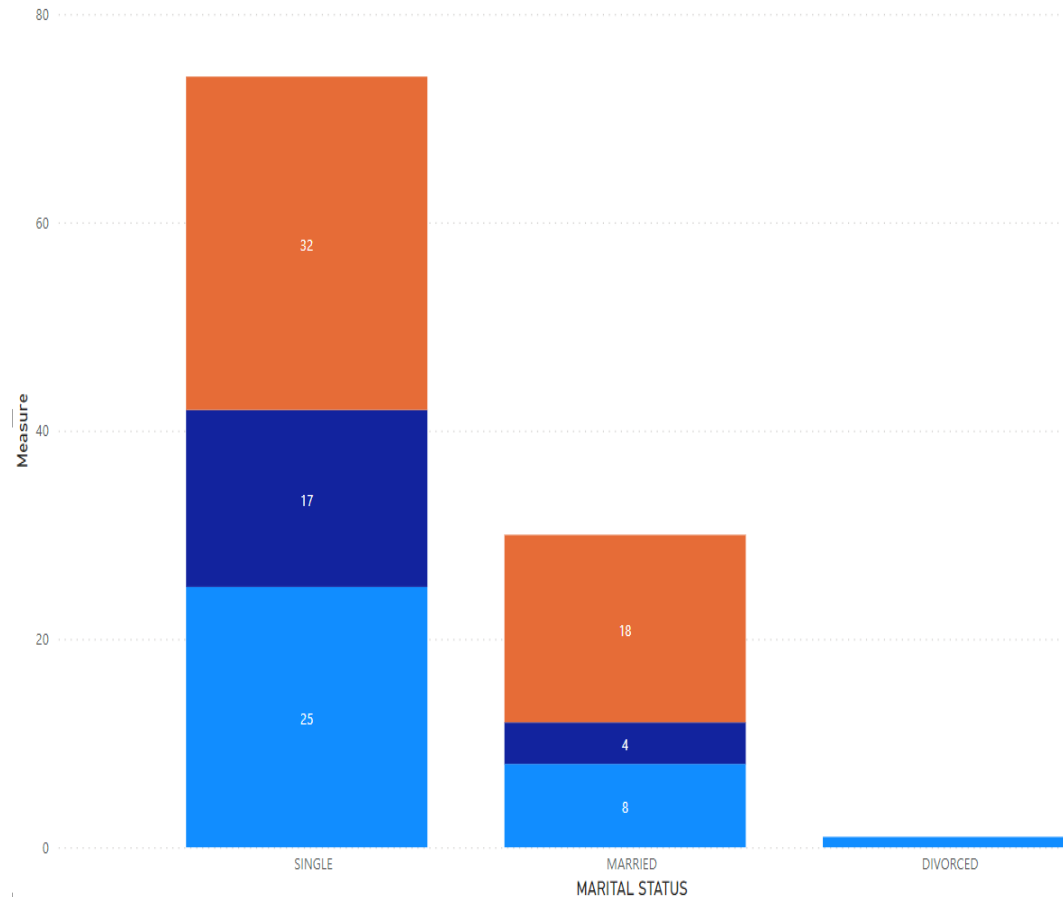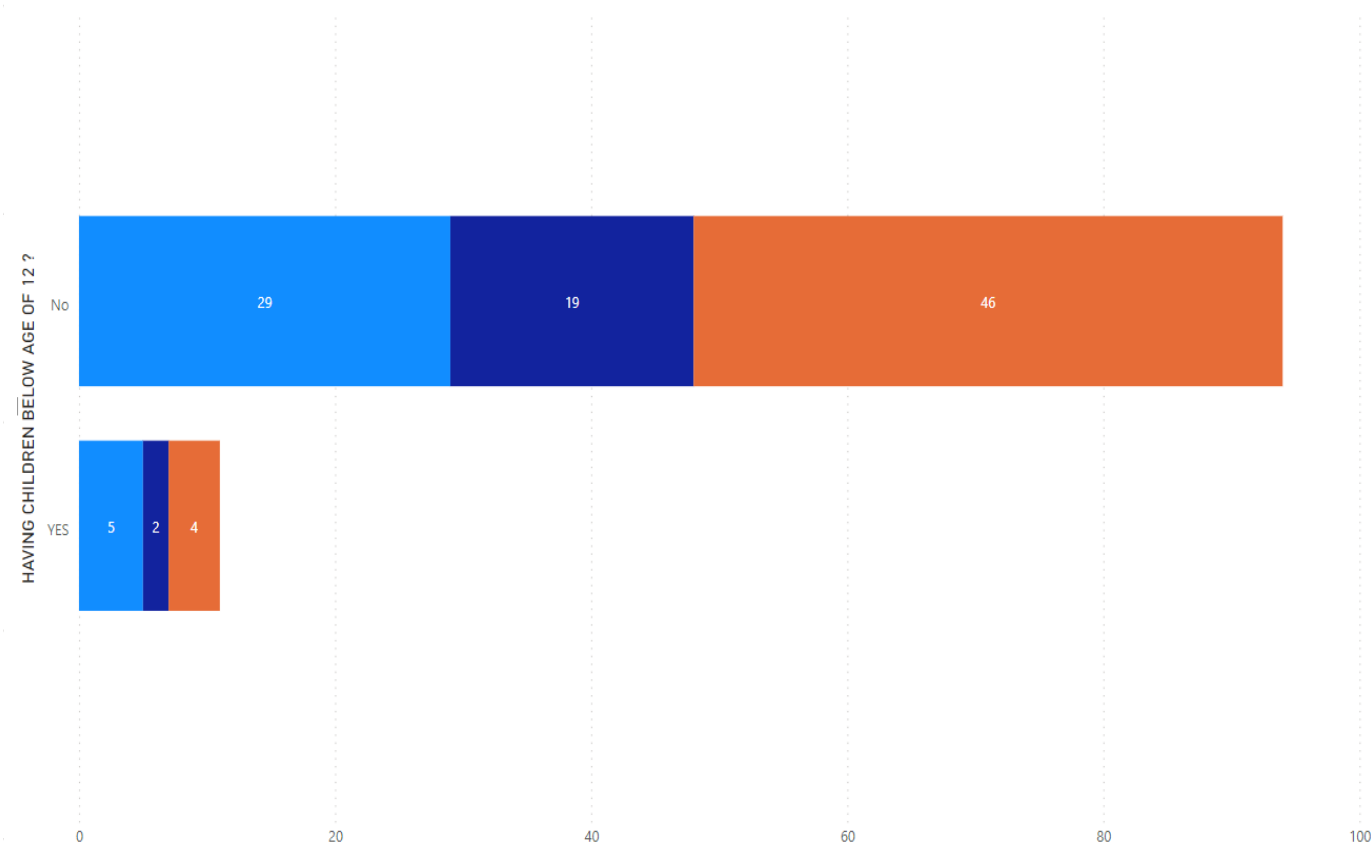
Missing Values

Missing values are always problematic for model creation and prediction, if missing values are present in the dataset it usually creates biased results or wrong prediction result. So to get rid of missing values there are various techniques with which we can treat missing values. Here "Time taken to reach office" have one missing value, so either we can delete it or impute it with median as this variable have outliers. Following are the methods that can be used to treat missing values.

- Deletion: If 5% or less of the overall dataset are missing values, we can delete those rows/columns, but it may delete some valuable information.
- Imputation: Imputing with Mean, Median or Mode imputation generally used widely.
- Prediction Based: Prediction based algorithm to predict the missing values.
- K means imputation: Using k means imputation to know the nearest value and replace it with.

We have imputed the missing value of "Time taken to reach office" with median value of "Time taken to reach office".

1 Missing Value Imputation

```
3]: df.isnull().sum()
```

```
3]: 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
```

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

Box Plot

Box Plot usually gives us the information of distribution of data of a variable. In box plot, Q1 is the lower side of rectangle box and represents 25 % percentile of the data, middle line of the box is 50 % percentile and represents median of the data and the Q3 is the higher side of the rectangle box and represents 75 % percentile of the data. Length of the rectangle box is considered as IQR or Inter Quartile Range (Q3-Q1). Any data points beyond maximum whisker or minimum whisker is considered as outlier.

Here for each variables we can see that the few data points are beyond maximum whisker as well as minimum whisker, those data points are considered as outliers.
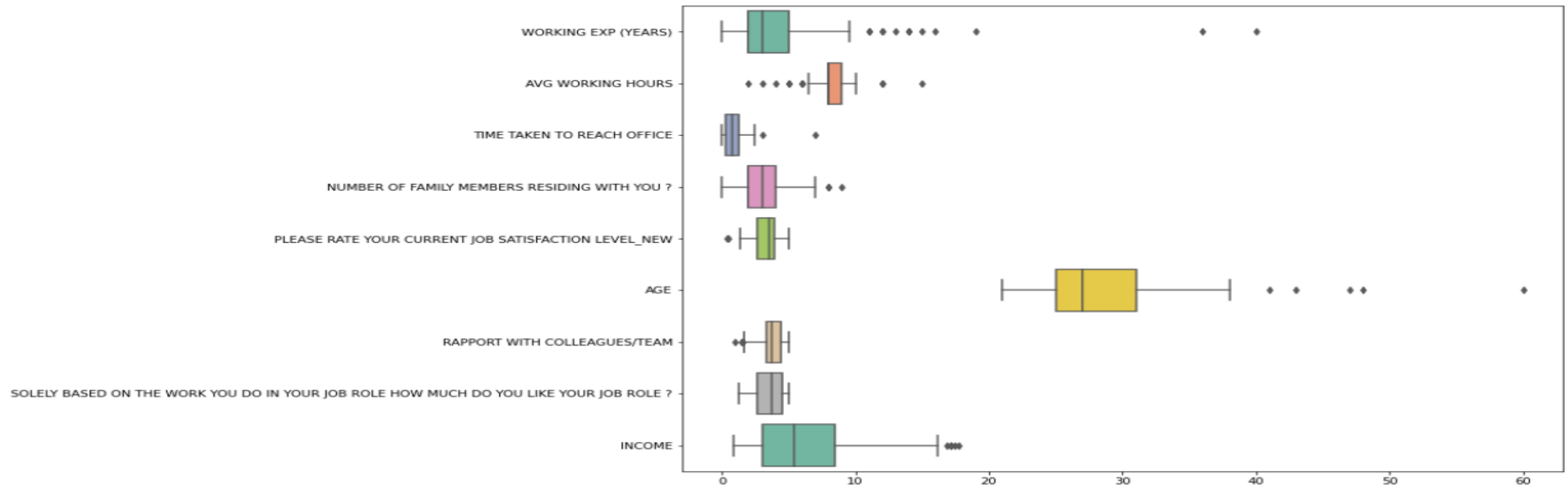
Outliers

One essential part of the EDA is the detection of outliers. Simply said, outliers are observations that are far away from the other data points in a random sample of a population. When identified, outliers should be dealt with proper methods as outliers may affect the results of analysis.

Here we have removed the outliers to penalize the effect of outliers, also there are imputation, Winsorization techniques to treat outliers. From this box plot we can see that no data points are present beyond any whiskers.
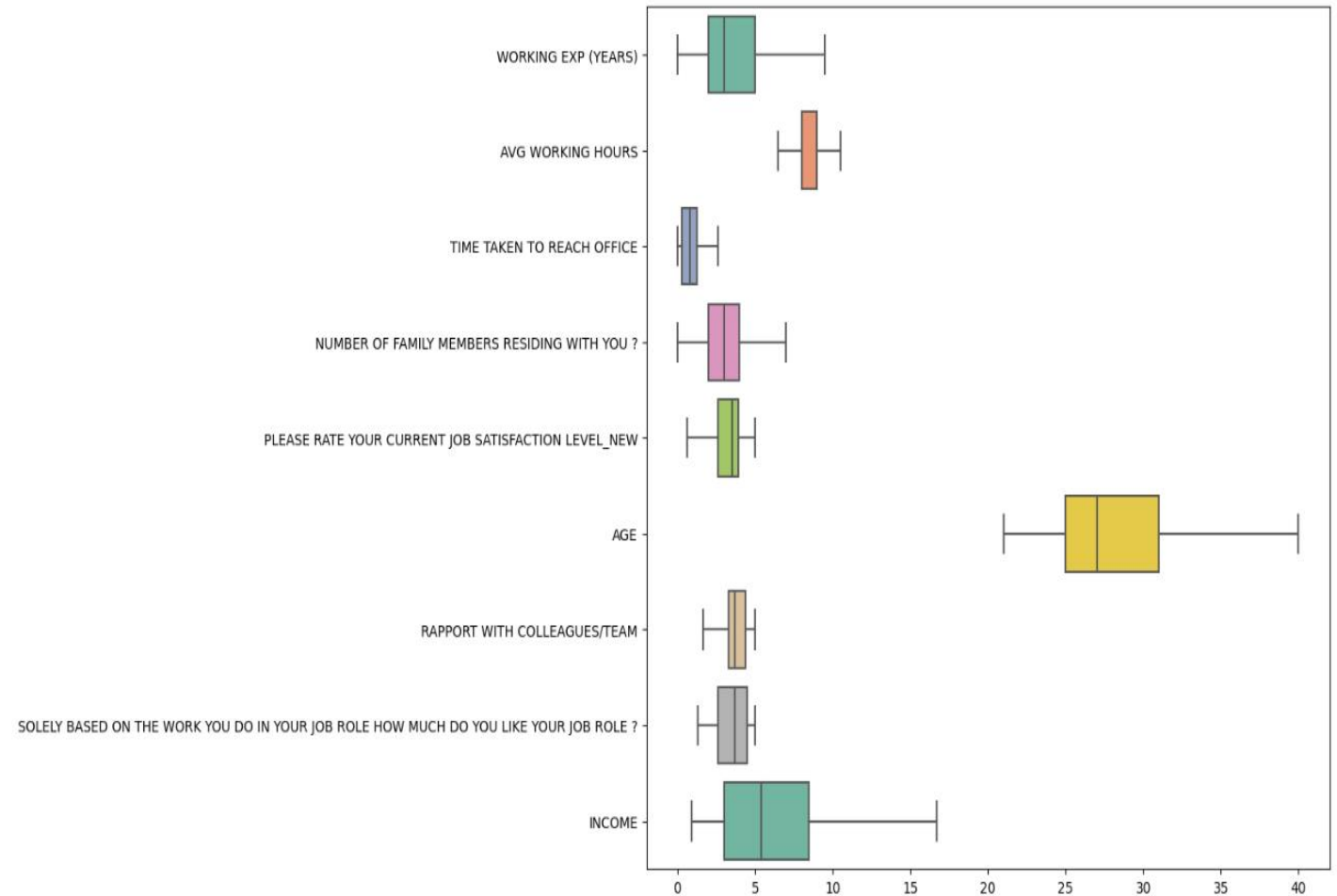
Bivariate Analysis

- Bivariate analysis is a type of quantitative analysis which determines how two variables are related and scatter plot is one of the important measure of bivariate analysis.

- Here our target variable is "Please rate your current job satisfaction level new" with which we have tried to analyze relation with "Age" –

- The scatter plot between "Please rate your current job satisfaction level new" and "Age", somewhat shows that when the age is respectively lower the job Satisfaction level vary low to high but with increasing age the variation of job satisfaction tends to decrease and it normally between medium to higher side of the job satisfaction level. Here we can say that age is not a proportional relationship with job satisfaction level but at higher age , job satisfaction level of an individual is also on higher side.
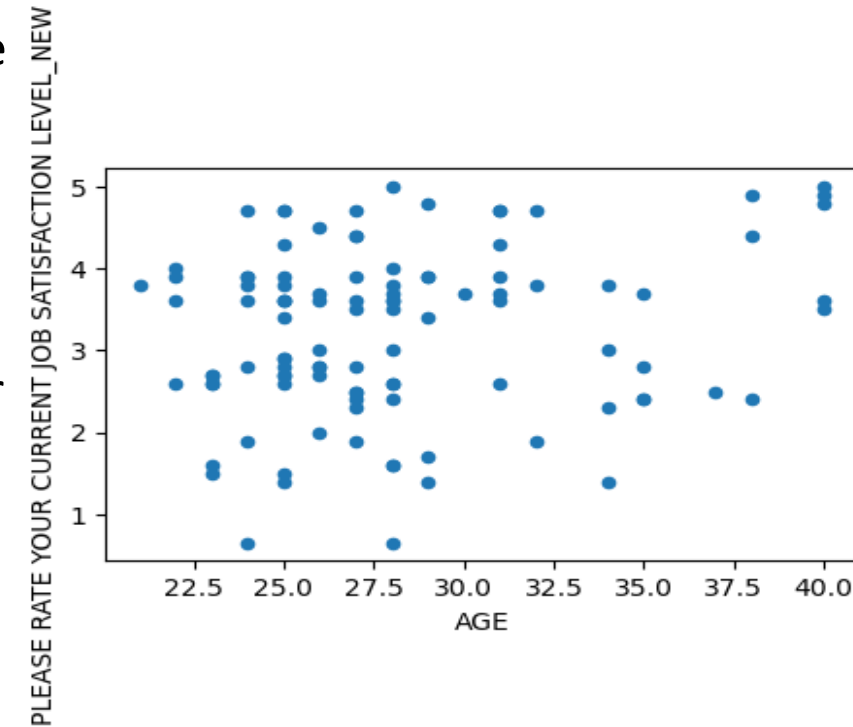
Next we compare "Please rate your current job satisfaction level new" with "AVG Working HOURS" –

The scatter plot between "Please rate your current job satisfaction level new" and "AVG WORKING HOURS", somewhat shows that the Current job satisfaction level is normally on higher side when the average working hours is low. But as average working hours increases , the job satisfaction level varies between average to lower level of job satisfaction.
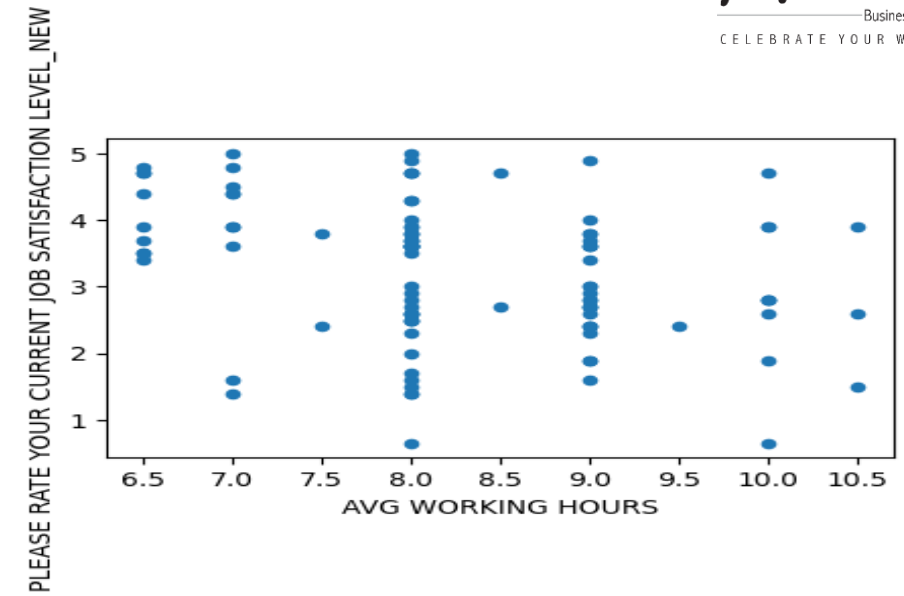
Next we compare "Please rate your current job satisfaction level new" with "Working exp" –

The scatter plot between "Please rate your current job satisfaction level new" and "Working Experience ", shows when the working experience is respectively lower the job Satisfaction level varies between all along low to high but as working exp is getting higher the variation in Job satisfaction level decrease and tends to move from medium to high job satisfaction level.
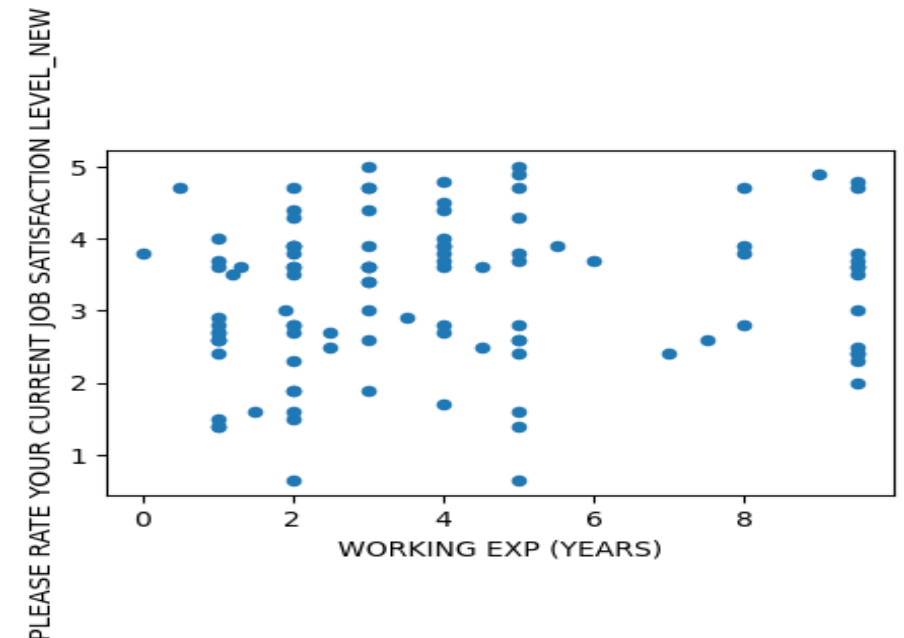
Next we compare "Please rate your current job satisfaction level new" with "Income" –

The scatter plot between "Please rate your current job satisfaction level new" and "Income", shows an important factor as in general low income may lead to low job satisfaction level but the plot shows that even at low income levels quite a number of respondents are having high job satisfaction level, though with increasing income, the job satisfaction level tends to increase and varies from average to high job satisfaction level.
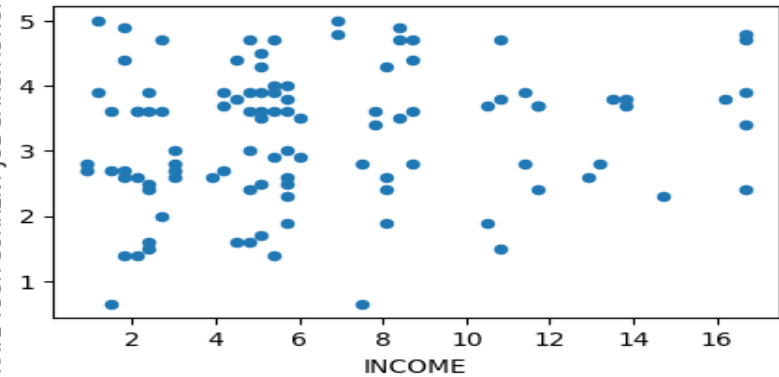
Next we compare "Please rate your current job satisfaction level new" with "Time taken to reach office" –

The scatter plot between "Please rate your current job satisfaction level new" and "Time taken to reach office", somewhat shows that with low travel time taken to reach office, job satisfaction level concentrates mostly between average to high level of job satisfaction and with increasing travel time taken to reach office job satisfaction level concentrates mostly between average to low level of job satisfaction. Though not much can be inferred from the relation between "Please rate your current job satisfaction level new" and "Time taken to reach office".
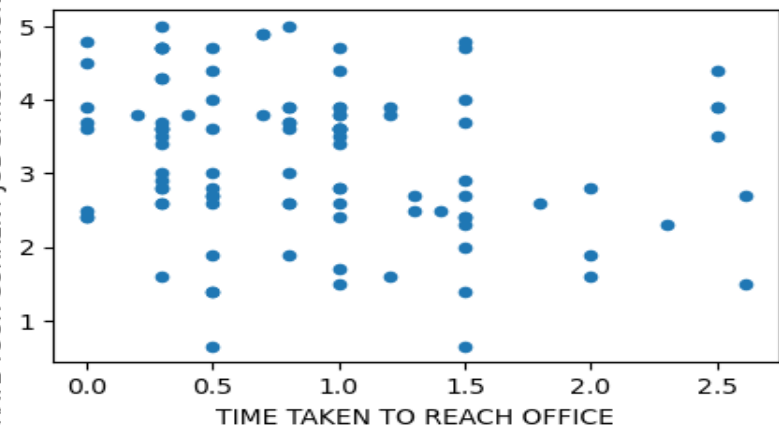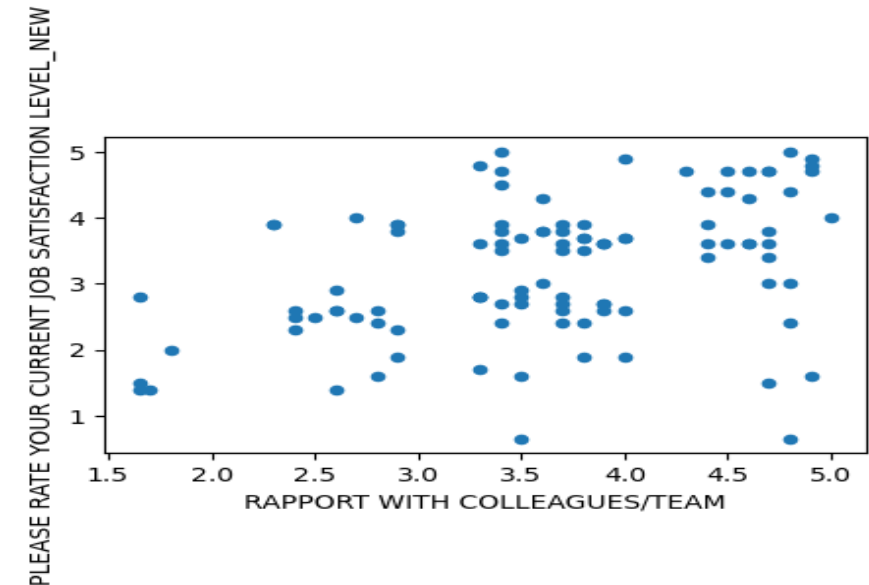
Next we compare "Please rate your current job satisfaction level new" with "Number of family members residing with you" –

The scatter plot between "Please rate your current job satisfaction level new" and "Number of family members residing with you", somewhat shows that with high no. of family members job satisfaction level is somewhat on the average side of job satisfaction level. Though not much can be inferred from the relation between "Please rate your current job satisfaction level new" and "Number of family members residing with you".

Next we compare "Please rate your current job satisfaction level new" with "Rapport with Colleagues/Team" –

The scatter plot between "Please rate your current job satisfaction level new" and "Rapport with Colleagues/Team", clearly shows that with low rapport with colleagues/ team members the job satisfaction level is generally on the lower side and with increasing rapport with colleagues / team members also increase the job satisfaction of an individual. The relation is somewhat proportional to each other.

Next we compare "Please rate your current job satisfaction level new" with "SOLELY BASED ON THE WORK YOU DO IN YOUR JOB ROLE HOW MUCH DO YOU LIKE YOUR JOB ROLE ?" –

The scatter plot between these two variables clearly shows that with low liking of job role of an individual, the job satisfaction level is also on the lower side and with increasing liking of job role of an individual also increases the job satisfaction of an individual to a significant level. The relation is quite proportional to each other.
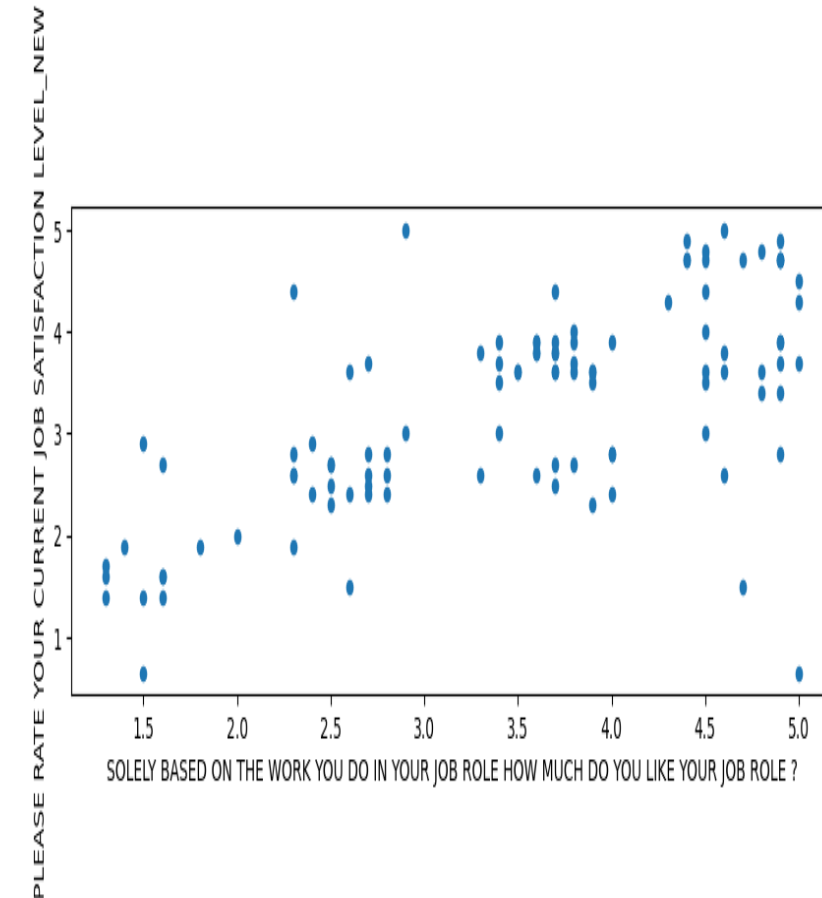
From the scatter charts we know the feature is correlated with target variable but we don't know how strongly or weakly correlated with target variable. Here correlation coefficient measures the magnitude of correlation between features and target variable. Generally +1 indicates strong relation and -1 weak relation.

- Job Role has the most strong correlation (0.64) with Target variable and rapport with colleagues 2nd most strong correlation (0.41) with target variable, this defines that the people who love their job role and have good rapport with colleagues and team members are more satisfied in their work places.
- Next influential factors are Age and Income of an individual.

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

]:

	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	-0.203419
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

Linear Regression –

We also found out how "Rapport with Colleagues/team" is linearly related with the target variable "Current job satisfaction level of an Individual". And the linear equation between these two variables is y=0.497x +1.42 where y is the target variable "Current job satisfaction level of an Individual" and x is the dependent variable "Rapport with Colleagues/team".

```python
from sklearn import linear_model
regModel = linear_model.LinearRegression()
regModel.fit(np.array(new_df_cap['RAPPORT WITH COLLEAGUES/TEAM']).reshape(-1,1), np.array(new_df_cap['PLEASE RATE YOUR CURREN
print("Coefficient: \n", regModel.coef_)
print("Intercept: \n", regModel.intercept_)
```

```
Coefficient:
 [[0.49763156]]
Intercept:
 [1.42135673]
```

```python
#Regression Equation between RAPPORT WITH COLLEAGUES/TEAM (x) and PLEASE RATE YOUR CURRENT JOB SATISFACTION LEVEL_NEW (y)
y=0.497x + 1.42
```

CONCLUSION

The top 3 factors are:-

1) Job Role of an Individual
2) Rapport with colleagues
3) Age of an Individual.

The top 2 reasons why people prefer work from home are :-
1) Time saved from travelling.
2) Financially more viable.

The top 2 reasons people prefer work from office are :-
1) Better working environment.
2) Better interaction with team/colleagues

The top 2 reasons people prefer Hybrid working mode are :-
1) Work life balance.
2) Higher Productivity.

Looking at the top three factors individual needs to focus on building rapport with colleagues and look for better job roles and jobs with better pay. Though there might be few external factors that might influence the job satisfaction level of an individual which is out of scope of our study. But focusing on improving above three aspects will help increase job satisfaction level of an individual by significant amount.