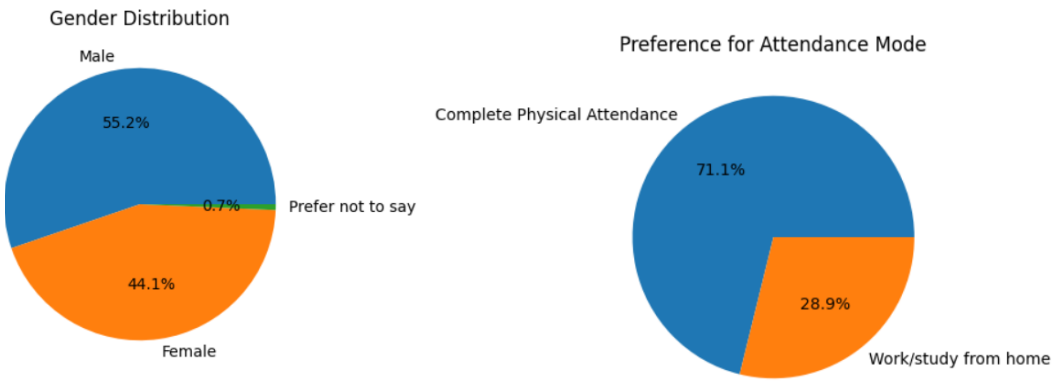
Data Collection and Preprocessing Phase

Date	21 June 2024
Team ID	739769
Project Title	Life Style Change Due To Covid Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

The data exploration and preprocessing phase is crucial in preparing the dataset for developing a predictive model that forecasts lifestyle changes due to COVID-19. This report outlines the steps taken to understand and preprocess the data to ensure its quality, relevance, and suitability for analysis and modeling.

Section	Description
Data Overview	Descriptive statistics: `data.describe()` (see table below)
Univariate Analysis	

	time_bp	time_dp	travel_time	easeof_online	home_env	prod_inc	sleep_bal	new_skill	fam_connect	relaxed	self_time	like_hw	dislike_hw	Unnamed: 19	t
count	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	1175.000000	0.0	11
mean	7.415319	7.971915	1.027660	2.533617	2.752340	0.008936	-0.108936	0.146809	0.260426	0.035745	0.082979	734.840851	651.067234	NaN	
std	2.005385	2.657007	0.713314	1.267609	1.235799	0.615083	0.621215	0.643686	0.686825	0.626637	0.541434	468.000935	502.319310	NaN	
min	4.000000	4.000000	0.500000	1.000000	1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	1.000000	1.000000	NaN	
25%	5.000000	5.000000	0.500000	1.000000	2.000000	-0.500000	-0.500000	-0.500000	0.000000	-0.500000	-0.500000	100.000000	101.000000	NaN	
50%	7.000000	9.000000	0.500000	2.000000	3.000000	0.000000	0.000000	0.500000	0.500000	0.000000	0.000000	1001.000000	1000.000000	NaN	
75%	9.000000	9.000000	1.500000	4.000000	4.000000	0.500000	0.500000	0.500000	1.000000	0.500000	0.500000	1100.000000	1101.000000	NaN	
max	12.000000	12.000000	3.000000	5.000000	5.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1111.000000	1111.000000	NaN	

Bivariate Analysis	-
Multivariate Analysis	-

Outliers and Anomalies	-
Data Preprocessing Code Screenshots	

Loading Data

	age	gender	occupation	line_of_work	time_bp	time_dp	travel_time	easeof_online	home_env	prod_inc	...	fam_connect	relaxed	self_time	like_hw	dislike_hw	prefer	certain
0	19-25	Male	Student in College	NaN	7	5	0.5	3	3	0.0	...	1.0	-0.5	-0.5	100	1	Complete Physical Attendance	
1	Dec-18	Male	Student in School	NaN	7	11	0.5	4	2	-0.5	...	1.0	1.0	1.0	1111	1110	Complete Physical Attendance	
2	19-25	Male	Student in College	NaN	7	7	1.5	2	2	1.0	...	0.5	0.5	0.5	1100	111	Complete Physical Attendance	
3	19-25	Male	Student in College	NaN	7	7	1.5	3	1	0.0	...	0.0	-1.0	-0.5	100	1111	Complete Physical Attendance	
4	19-25	Female	Student in College	NaN	7	7	1.5	2	2	0.0	...	0.0	0.5	0.0	1010	1000	Complete Physical Attendance	

5 rows × 22 columns

Handling Missing Data

```
data.isnull().sum()
```

```
age                0
gender             0
occupation         0
line_of_work     696
time_bp            0
time_dp            0
travel_time        0
easeof_online      0
home_env           0
prod_inc           0
sleep_bal          0
new_skill          0
fam_connect        0
relaxed            0
self_time          0
like_hw            0
dislike_hw         0
prefer             0
certaindays_hw     0
Unnamed: 19     1175
time_bp.1          0
travel+work     1175
dtype: int64
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1175 entries, 0 to 1174
Data columns (total 22 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             1175 non-null   object
 1   gender          1175 non-null   object
 2   occupation      1175 non-null   object
 3   line_of_work    479 non-null    object
 4   time_bp         1175 non-null   int64
 5   time_dp         1175 non-null   int64
 6   travel_time     1175 non-null   float64
 7   easeof_online   1175 non-null   int64
 8   home_env        1175 non-null   int64
 9   prod_inc        1175 non-null   float64
 10  sleep_bal       1175 non-null   float64
 11  new_skill       1175 non-null   float64
 12  fam_connect     1175 non-null   float64
 13  relaxed         1175 non-null   float64
 14  self_time       1175 non-null   float64
 15  like_hw         1175 non-null   int64
 16  dislike_hw      1175 non-null   int64
 17  prefer          1175 non-null   object
 18  certaindays_hw  1175 non-null   object
 19  Unnamed: 19     0 non-null      float64
 20  time_bp.1       1175 non-null   int64
 21  travel+work     0 non-null      float64
dtypes: float64(9), int64(7), object(6)
memory usage: 202.1+ KB
```

Data Transformation

```
le_age.fit_transform(data['age'])
le_gender.fit_transform(data['gender'])
le_occupation.fit_transform(data['occupation'])
le_line_of_work.fit_transform(data['line_of_work'])
le_prefer.fit_transform(data['prefer'])
le_certaindays_hw.fit_transform(data['certaindays_hw'])
```

```
array([2, 1, 2, ..., 0, 2, 2])
```

Feature Engineering

Attached the codes in final submission.

Save Processed Data

-