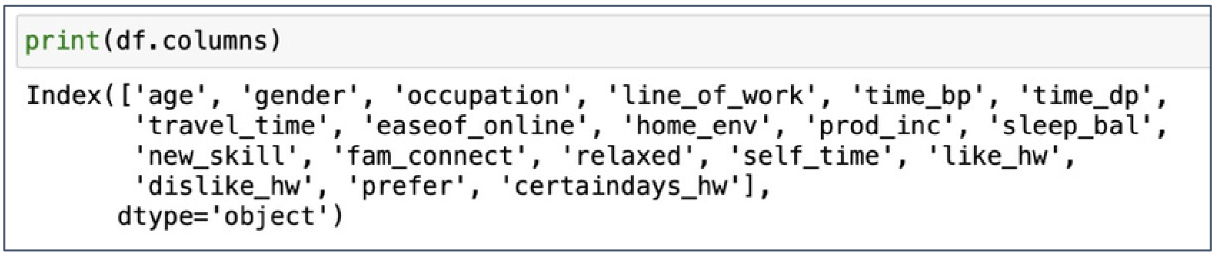
Table of Contents

Descriptive Analytics

Objective:

To perform descriptive analytics on a dataset sourced from the 2020 Psychological Effects of COVID and visualize specific attributes to gain insights into the data distribution and relationships between variables.



```
print(df.columns)

Index(['age', 'gender', 'occupation', 'line_of_work', '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', 'prefer', 'certaindays_hw'],
      dtype='object')
```

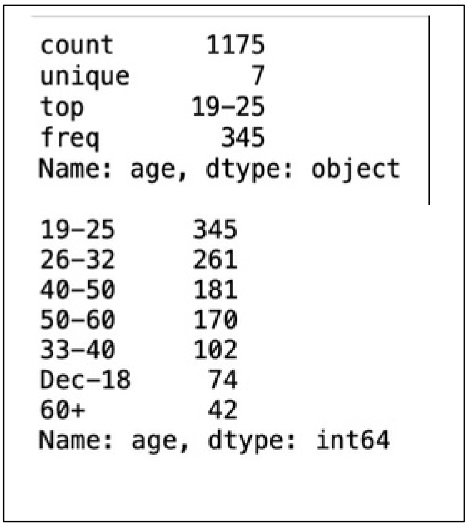
The dataset gives information about people's experiences in the period of COVID-19 pandemic, which was collected in the year 2020. This data mainly focuses on the psychological effects of the pandemic and the impact of lockdown.

The columns in the dataset says the different perspectives of individuals' lives in the pandemic. It includes demographic information of different age, gender, the professionals like occupation and sector, their work-life , both before and after pandemic.

It's got the nitty-gritty details, like how much time they clocked in working, how commuting changed, and their vibes about the whole ordeal. People spilled their feelings in ratings – from how cool it is to work from home to family time vibes. The juiciest part is all about the work-from-home scene – what they liked, hated, and preferred compared to the office grind. Digging into this data goldmine helps us decode the mental rollercoaster of the pandemic, showing how different folks coped. It's like a manual for dealing with crazy times, giving us clues for better crisis plans down the road.

Basic Statistics :

Age

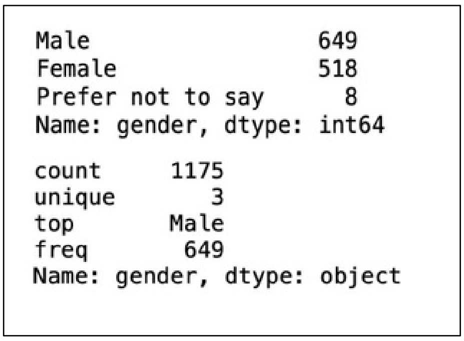


```
count      1175
unique        7
top       19-25
freq        345
Name: age, dtype: object

19-25      345
26-32      261
40-50      181
50-60      170
33-40      102
Dec-18      74
60+         42
Name: age, dtype: int64
```

The dataset contains information of individuals' ages, with a total of 1,175 entries. The age distribution is as follows:

- Majority are 19-25 years old (345 entries).
- Next is 26-32 age group (261 occurrences).
- 40-50 and 50-60 have 181 and 170 entries.
- 33-40 age group has 102 entries.
- "Dec-18" group has 74 entries.
- 60 and above has 42 entries..

Gender

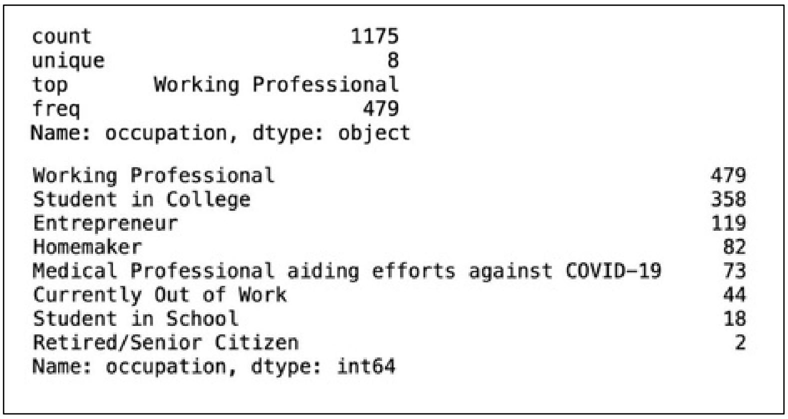


```
Male                 649
Female               518
Prefer not to say      8
Name: gender, dtype: int64

count      1175
unique        3
top        Male
freq        649
Name: gender, dtype: object
```

- There are three unique gender categories: Male, Female, and Prefer not to say.
- The most frequently occurring gender is Male, with a count of 649.
- The Male gender appears most frequently, occurring 649 times out of the total 1,175 records.

There are also entries for individuals who identify as Female (518), and a small number who prefer not to specify their gender (8).

Occupation

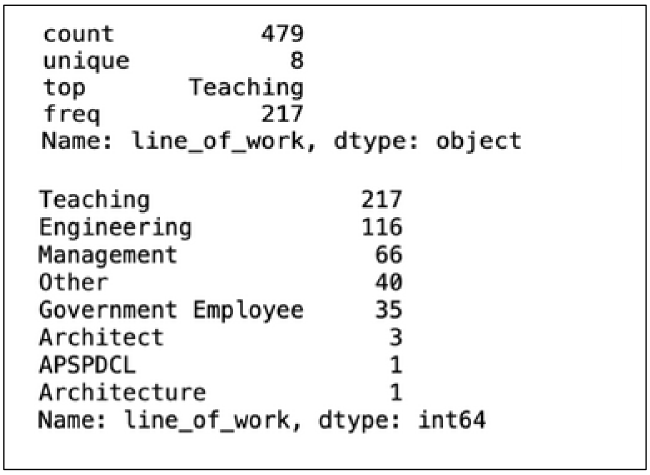


```
count                      1175
unique                        8
top       Working Professional
freq                        479
Name: occupation, dtype: object

Working Professional                                   479
Student in College                                     358
Entrepreneur                                           119
Homemaker                                               82
Medical Professional aiding efforts against COVID-19    73
Currently Out of Work                                   44
Student in School                                       18
Retired/Senior Citizen                                   2
Name: occupation, dtype: int64
```

- The dataset includes information about individuals' occupations.
- There are eight unique occupation categories.
- The most frequently occurring occupation is "Working Professional," which appears 479 times in the dataset.
- The least common occupation is Retired/Senior Citizen, with 2 people

Other occupations include "Student in College" (358 occurrences), "Entrepreneur" (119 occurrences), "Homemaker" (82 occurrences), "Medical Professional aiding efforts against COVID-19" (73 occurrences), "Currently Out of Work" (44 occurrences), "Student in School" (18 occurrences), and "Retired/Senior Citizen" (2 occurrences).

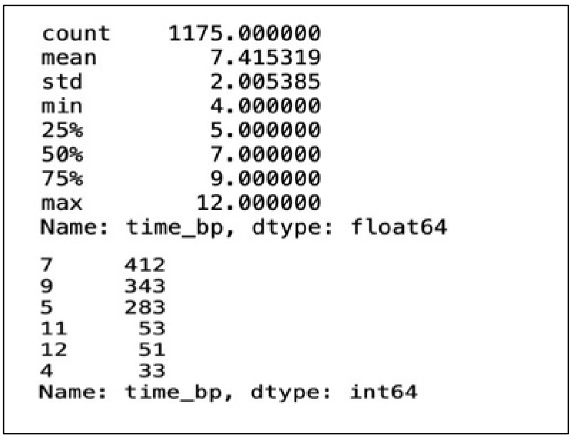
line_of_work



```
count          479
unique           8
top       Teaching
freq           217
Name: line_of_work, dtype: object

Teaching               217
Engineering            116
Management              66
Other                   40
Government Employee     35
Architect                3
APSPDCL                  1
Architecture             1
Name: line_of_work, dtype: int64
```

- In the 'line_of_work' column, we have 479 entries in total.
- Teaching is the most common occupation, with 217 individuals.
- Engineering has 116 people.
- Management roles include 66 individuals.
- 40 people fall into the 'Other' category.
- 35 individuals work for the government.
- There are 3 architects.
- 1 person is associated with APSPDCL.
- Another individual works in the architecture field .

time_bp

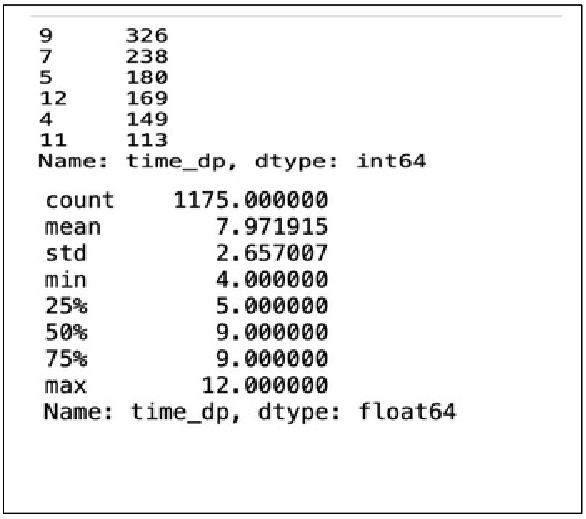


```
count    1175.000000
mean        7.415319
std         2.005385
min         4.000000
25%         5.000000
50%         7.000000
75%         9.000000
max        12.000000
Name: time_bp, dtype: float64

7     412
9     343
5     283
11     53
12     51
4      33
Name: time_bp, dtype: int64
```

- Average Time: People, on average, spent about 7.42 units on work before the pandemic.
- Variability: The time spent on work fluctuates around this average by roughly 2.01 units.
- Minimum Time: The least time spent on work pre-pandemic is 4 units.
- 25th Percentile (Q1): A quarter of individuals devoted 5 units or less to work.

- Median (50th Percentile): Half of people spent 7 units or less on work before the pandemic.
- 75th Percentile (Q3): About 75% of individuals limited their work time to 9 units or less.
- Maximum Time: The highest recorded time spent on work is 12 units.
- Range: The range covers a span from 4 to 12 units.

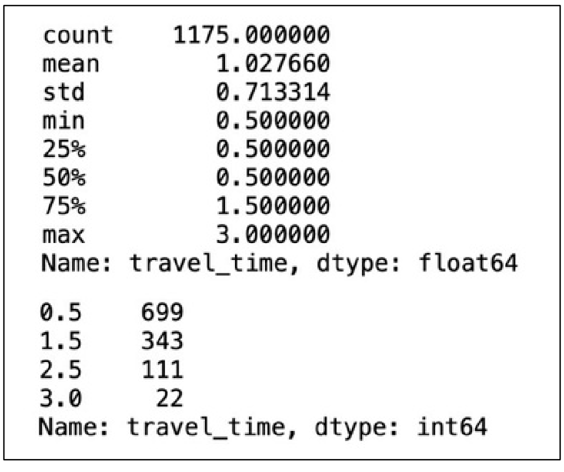
time_dp



```
9     326
7     238
5     180
12    169
4     149
11    113
Name: time_dp, dtype: int64

count    1175.000000
mean        7.971915
std         2.657007
min         4.000000
25%         5.000000
50%         9.000000
75%         9.000000
max        12.000000
Name: time_dp, dtype: float64
```

- Average Time: People, on average, spent about 7.42 units on work before the pandemic.
- Variability: The time spent on work fluctuates around this average by roughly 2.01 units.
- Minimum Time: The least time spent on work pre-pandemic is 4 units.
- 25th Percentile (Q1): A quarter of individuals devoted 5 units or less to work.
- Median (50th Percentile): Half of people spent 7 units or less on work before the pandemic.
- 75th Percentile (Q3): About 75% of individuals limited their work time to 9 units or less.
- Maximum Time: The highest recorded time spent on work is 12 units.
- Range: The range covers a span from 4 to 12 units.

travel_time

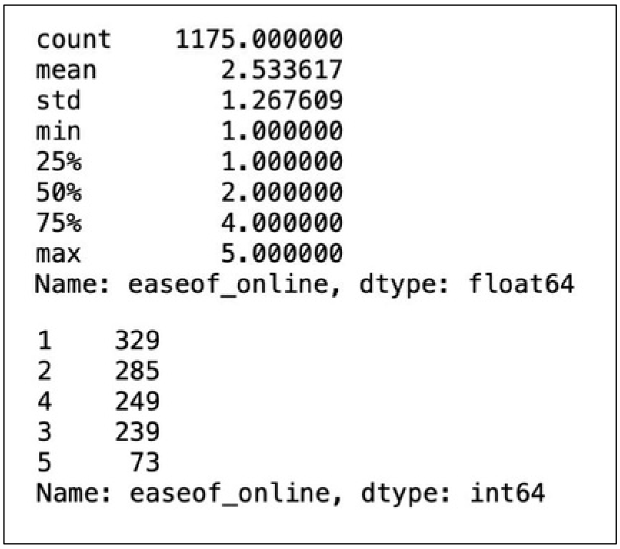


```
count    1175.000000
mean        1.027660
std         0.713314
min         0.500000
25%         0.500000
50%         0.500000
75%         1.500000
max         3.000000
Name: travel_time, dtype: float64

0.5    699
1.5    343
2.5    111
3.0     22
Name: travel_time, dtype: int64
```

The 'travel_time' column, which represents the travel time spent.

- Mean(Average) Travel Time : On average, people spent approximately 1.03 units of ti me on travel.
- Standard Deviation : The travel time varies around the average by approximately 0.71 units.
- Minimum Travel Time : The minimum travel time is 0.5 units.
- 25th Percentile(Q1) : 25% of individuals spent 0.5 units of time or less on travel.
- Median (50th Percentile) : 50% of individuals spent 0.5 units of time or less on travel.
- 75th Percentile(Q3) : 75% of individuals spent 1.5 units of time or less on travel.
- Maximum Travel Time : The maximum travel time is 3 units.

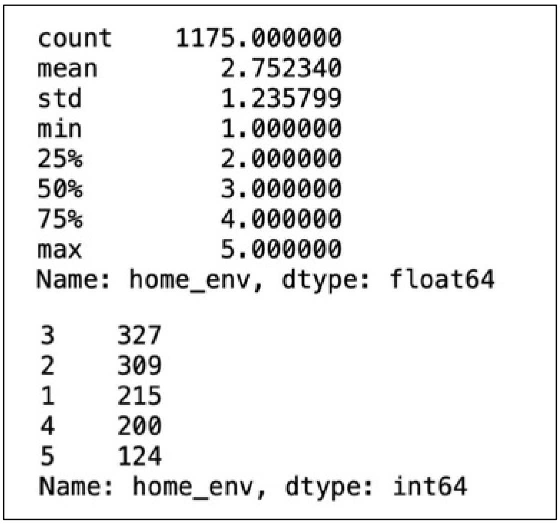
easeof_online



```
count    1175.000000
mean        2.533617
std         1.267609
min         1.000000
25%         1.000000
50%         2.000000
75%         4.000000
max         5.000000
Name: easeof_online, dtype: float64

1    329
2    285
4    249
3    239
5     73
Name: easeof_online, dtype: int64
```

- Most people rated working online as not very easy (rating 1).
- Ratings 2, 3, and 4 vary, indicating different levels of ease.
- Fewer people found it very easy (rating 5).
- On average, it's moderately easy with a rating of approximately 2.53.
- Ratings vary around this average by about 1.27.
- The minimum rating is 1, and the maximum is 5.
- 25% rated it 1 or lower, 50% rated it 2 or lower, and 75% rated it 4 or lower.

home_env

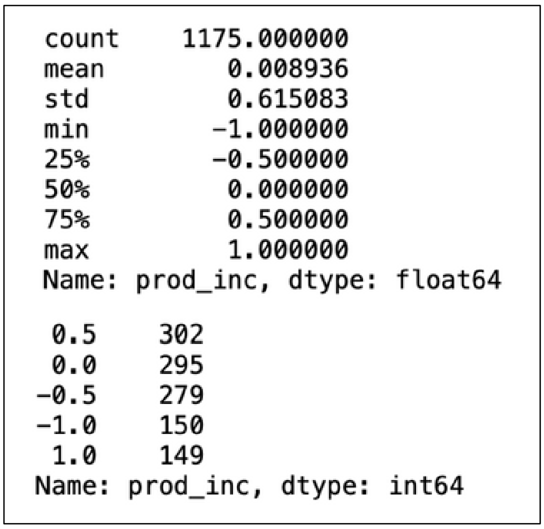


```
count    1175.000000
mean        2.752340
std         1.235799
min         1.000000
25%         2.000000
50%         3.000000
75%         4.000000
max         5.000000
Name: home_env, dtype: float64

3    327
2    309
1    215
4    200
5    124
Name: home_env, dtype: int64
```

The 'home_env' column, representing the liking of the home environment:

- People generally rate their liking of home environment as around 2.75 out of 5.
- The liking varies by about 1.24 units from this average.
- The lowest liking recorded is 1.
- 25% of people rated their liking at 2 or lower.
- Half of the individuals rated their liking at 3 or lower.
- 75% of people rated their liking at 4 or lower.
- The highest liking reported is 5.

prod_inc

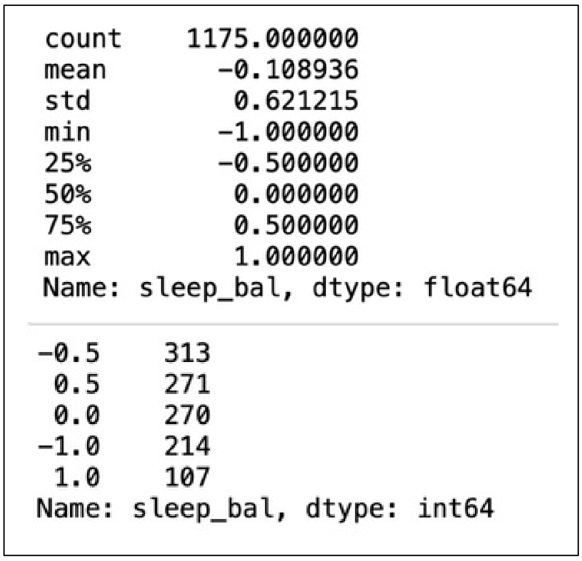


```
count    1175.000000
mean        0.008936
std         0.615083
min        -1.000000
25%        -0.500000
50%         0.000000
75%         0.500000
max         1.000000
Name: prod_inc, dtype: float64

 0.5    302
 0.0    295
-0.5    279
-1.0    150
 1.0    149
Name: prod_inc, dtype: int64
```

The 'prod_inc' column, which represents the rating of productivity increase.

- On average, people rated productivity increase at about 0.009.
- Ratings vary around this average by approximately 0.615.
- The lowest rating is -1.0, indicating a decrease in productivity.
- 25% of individuals have a rating of -0.5 or lower.
- Half of the people (50%) have a rating of 0.0 or lower.
- 75% of individuals have a rating of 0.5 or lower.
- The highest rating for productivity increase is 1.0, showing a significant improvement

sleep_bal

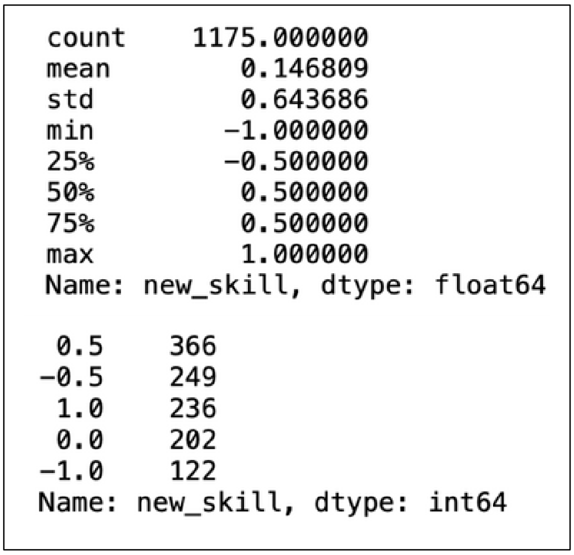


```
count    1175.000000
mean       -0.108936
std         0.621215
min        -1.000000
25%        -0.500000
50%         0.000000
75%         0.500000
max         1.000000
Name: sleep_bal, dtype: float64

-0.5    313
 0.5    271
 0.0    270
-1.0    214
 1.0    107
Name: sleep_bal, dtype: int64
```

The 'sleep_bal' column, which represents the rating of the sleep cycle.

- Average Rating: On average, people's sleep cycle is rated around -0.109.
- Rating Variation: The ratings vary by approximately 0.621 from the average.
- Lowest Rating: The worst sleep cycle rating recorded is -1.0.
- Q1 (25th Percentile): A quarter of people have a rating of -0.5 or lower.
- Median Rating (50th Percentile): Half the individuals have a rating of 0.0 or lower.
- Q3 (75th Percentile): Three-quarters of people have a rating of 0.5 or lower.
- Highest Rating: The best sleep cycle rating observed is 1.0.

new_skill

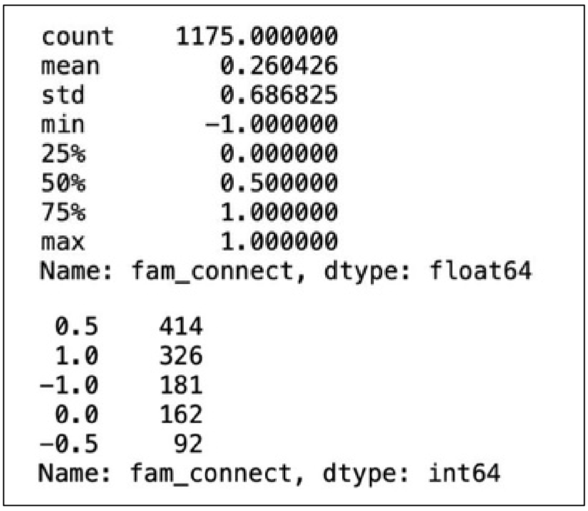


```
count    1175.000000
mean        0.146809
std         0.643686
min        -1.000000
25%        -0.500000
50%         0.500000
75%         0.500000
max         1.000000
Name: new_skill, dtype: float64

 0.5    366
-0.5    249
 1.0    236
 0.0    202
-1.0    122
Name: new_skill, dtype: int64
```

The 'new_skill' column, which represents whether any new skill was learned.

- Most people reported positively, with an average value of around 0.147 for learning new skills.
- Responses vary, indicating some diversity, with a standard deviation of approximately 0.644.
- The lowest reported value is -1.0, meaning some individuals did not learn new skills.
- About 25% of respondents reported a negative response (-0.5 or lower) for learning new skills.
- Half of the respondents reported a positive response (0.5 or lower).
- 75% of respondents reported a positive response (0.5 or higher).
- The highest reported value is 1.0, indicating some learned new skills.

fam_connect

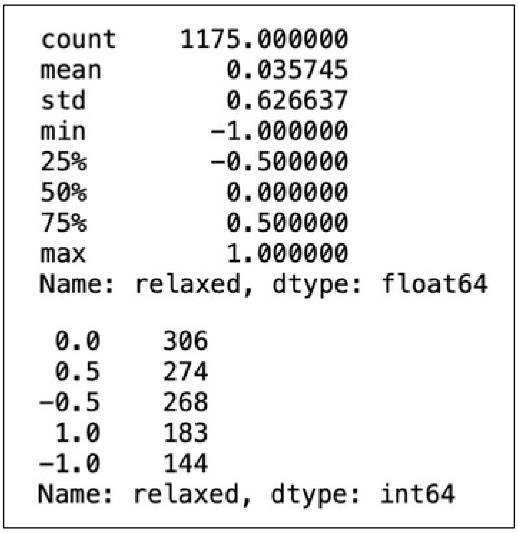


```
count    1175.000000
mean        0.260426
std         0.686825
min        -1.000000
25%         0.000000
50%         0.500000
75%         1.000000
max         1.000000
Name: fam_connect, dtype: float64

 0.5    414
 1.0    326
-1.0    181
 0.0    162
-0.5     92
Name: fam_connect, dtype: int64
```

The 'fam_connect' column, which represents the rating of how well the person connected with their family:

- People, on average, rate their family connection at about 0.036.
- Family connection ratings vary around this average by roughly 0.627.
- The lowest rating reported is -1.0, indicating some had negative family experiences.
- 25% of individuals reported a rating of -0.5 or lower, suggesting negative family connections.
- 50% reported a rating of 0.0 or lower, indicating a neutral family experience.
- 75% reported a rating of 0.5 or higher, suggesting positive family connections.
- The highest rating reported is 1.0, indicating highly positive family experiences.

relaxed

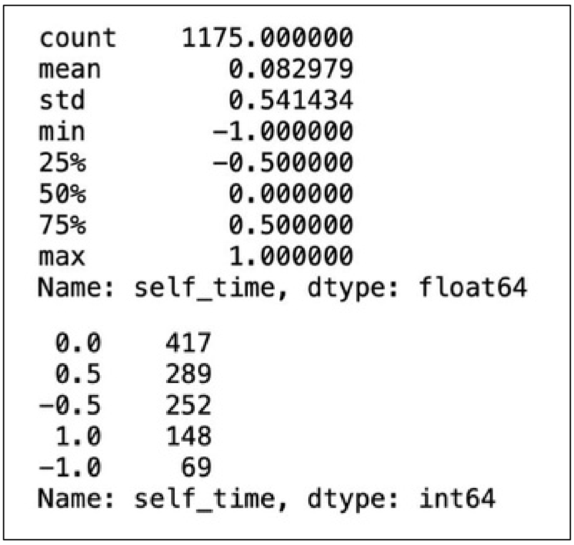


```
count    1175.000000
mean        0.035745
std         0.626637
min        -1.000000
25%        -0.500000
50%         0.000000
75%         0.500000
max         1.000000
Name: relaxed, dtype: float64

 0.0    306
 0.5    274
-0.5    268
 1.0    183
-1.0    144
Name: relaxed, dtype: int64
```

The 'relaxed' column, which represents the rating of how relaxed the person is feeling.

- On average, people feel somewhat relaxed, with a mean rating of 0.036.
- Relaxation ratings vary around this average by approximately 0.627.
- Some individuals reported feeling not relaxed at all, with a minimum rating of -1.0.
- 25% of people feel less relaxed, reporting a rating of -0.5 or lower.
- 50% report a neutral state of relaxation, with a median rating of 0.0.
- 75% feel more relaxed, reporting a rating of 0.5 or higher.
- The maximum relaxation rating is 1.0, indicating some individuals feel highly relaxed.

self_time

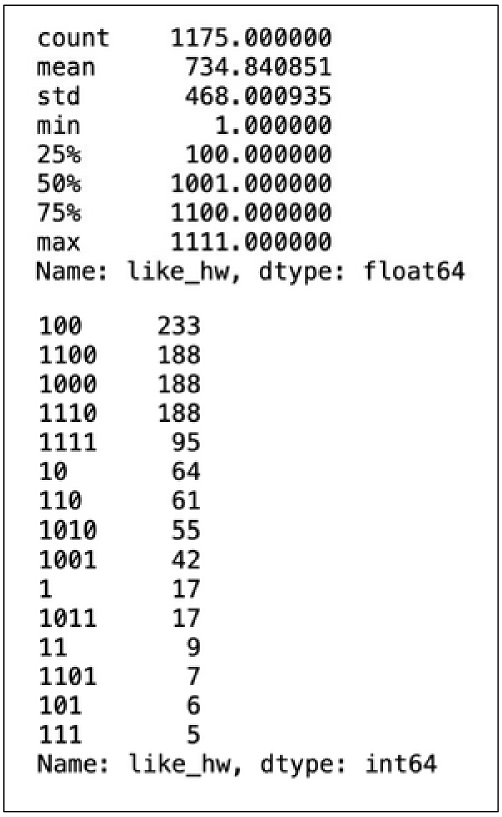


```
count    1175.000000
mean        0.082979
std         0.541434
min        -1.000000
25%        -0.500000
50%         0.000000
75%         0.500000
max         1.000000
Name: self_time, dtype: float64

 0.0    417
 0.5    289
-0.5    252
 1.0    148
-1.0     69
Name: self_time, dtype: int64
```

The 'self_time' column, which represents the rating of how much self-time was procured.

- On average, people rate their self-time at 0.083, indicating some self-care.
- Ratings vary around this average by about 0.541, showcasing diverse self-time experiences.
- The minimum rating is -1.0, suggesting some reported no self-time.
- 25% rated -0.5 or lower, indicating less self-time for a quarter of individuals.
- Median is 0.0, showing a neutral self-time state for half of the respondents.
- 75% rated 0.5 or higher, implying a significant self-time for most.
- Maximum rating is 1.0, denoting substantial self-time reported by some.

like_hw

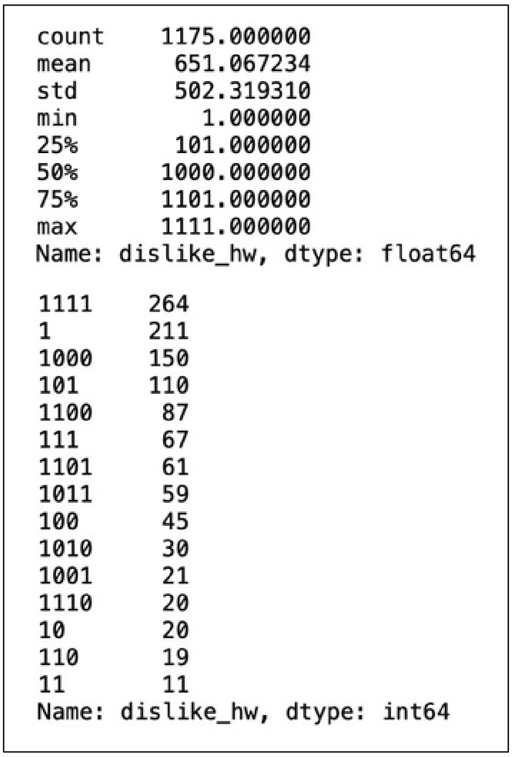


```
count    1175.000000
mean      734.840851
std       468.000935
min         1.000000
25%       100.000000
50%      1001.000000
75%      1100.000000
max      1111.000000
Name: like_hw, dtype: float64

100     233
1100    188
1000    188
1110    188
1111     95
10       64
110      61
1010     55
1001     42
1        17
1011     17
11        9
1101      7
101       6
111       5
Name: like_hw, dtype: int64
```

The 'like_hw' column, information on how individuals feel about working from home.

- On average, people like working from home, with a mean liking score of 734.84.
- Individual likings vary, as shown by the standard deviation of 468.00.
- The lowest liking score is 1, indicating someone least positive about remote work.
- 25% of people have a liking score of 100 or lower, suggesting some have lower preferences.
- The median liking score is 1001, meaning half of respondents like it less, and half like it more.
- 75% of people have a liking score of 1100 or lower, showing a majority with relatively lower preferences.
- The highest liking score is 1111, reflecting at least one person highly favors working from home.

dislike_hw

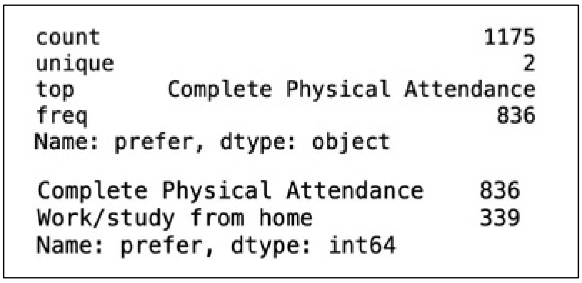


```
count    1175.000000
mean      651.067234
std       502.319310
min         1.000000
25%       101.000000
50%      1000.000000
75%      1101.000000
max      1111.000000
Name: dislike_hw, dtype: float64

1111    264
1       211
1000    150
101     110
1100     87
111      67
1101     61
1011     59
100      45
1010     30
1001     21
1110     20
10       20
110      19
11       11
Name: dislike_hw, dtype: int64
```

The distribution of dislike levels for working from home in your dataset, including the averag esentiment, variability, and the range of opinions among individuals.

- Average Dislike Level: Most people, on average, feel a dislike level of 651.07 towards working from home.
- Variability in Opinions: There's a wide range of opinions, with a standard deviation of 502.32, showing that people's feelings about remote work vary significantly.
- Minimum Dislike Level: At least one person expressed the lowest level of dislike, with a rating of 1.
- 25th Percentile (Q1): 25% of individuals have a dislike level of 101 or lower.
- Median Dislike Level (50th Percentile): The middle point is 1000, indicating that half dislike it less and half more.
- 75th Percentile (Q3): 75% of people have a dislike level of 1101 or lower.
- Maximum Dislike Level: At least one person expressed the highest level of dislike, wit h a rating of 1111.

- **Prefer**



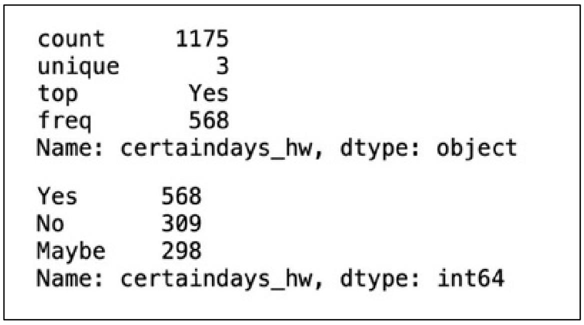
```
count                         1175
unique                           2
top       Complete Physical Attendance
freq                           836
Name: prefer, dtype: object

Complete Physical Attendance    836
Work/study from home            339
Name: prefer, dtype: int64
```

The 'prefer' column, information on how individuals preference of working from home or offi ce.

- Unique Values : There are two unique values, indicating that individuals in your data set have expressed two distinct preferences. (Complete Physical Attendance, Work/study from home)
- Top Value : The most frequently expressed preference is "Complete Physical Attendance," which occurs 836 times in the dataset.

The distribution of preferences for working arrangements in the dataset. The majority of individuals prefer complete physical attendance, while a smaller group prefers to work or study from home.

- **certaindays_hw**



```
count      1175
unique        3
top         Yes
freq        568
Name: certaindays_hw, dtype: object

Yes       568
No        309
Maybe     298
Name: certaindays_hw, dtype: int64
```

The distribution of preferences for working from home on certain days in the dataset.

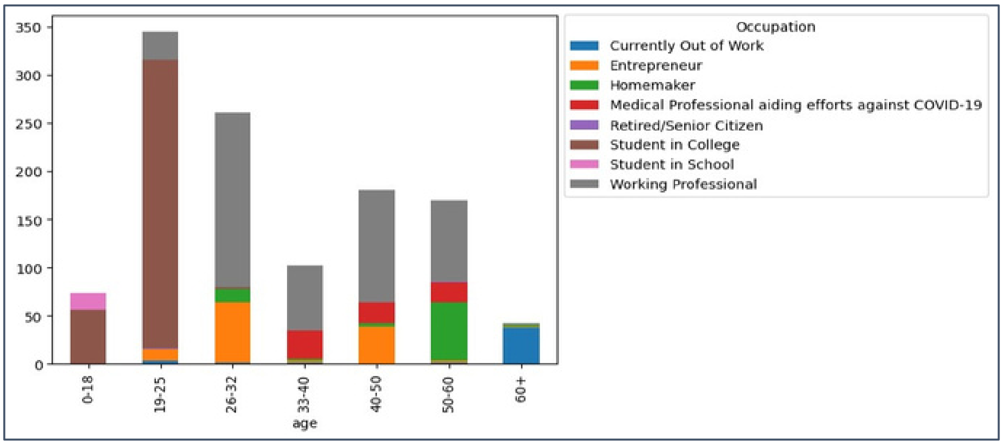
- Unique Value : There are three unique values: 'Yes', 'No', and 'Maybe'. This indicates the different responses individuals have regarding the need for working from home on certain days.
- Top value : The most common response is 'Yes', occurring 568 times. This means that a significant number of individuals express a preference for working from home on specific days.

Data Visualisation :

Using the **matplotlib** and **seaborn** libraries, multiple visualizations were generated to better understand the data's distribution and relationships:

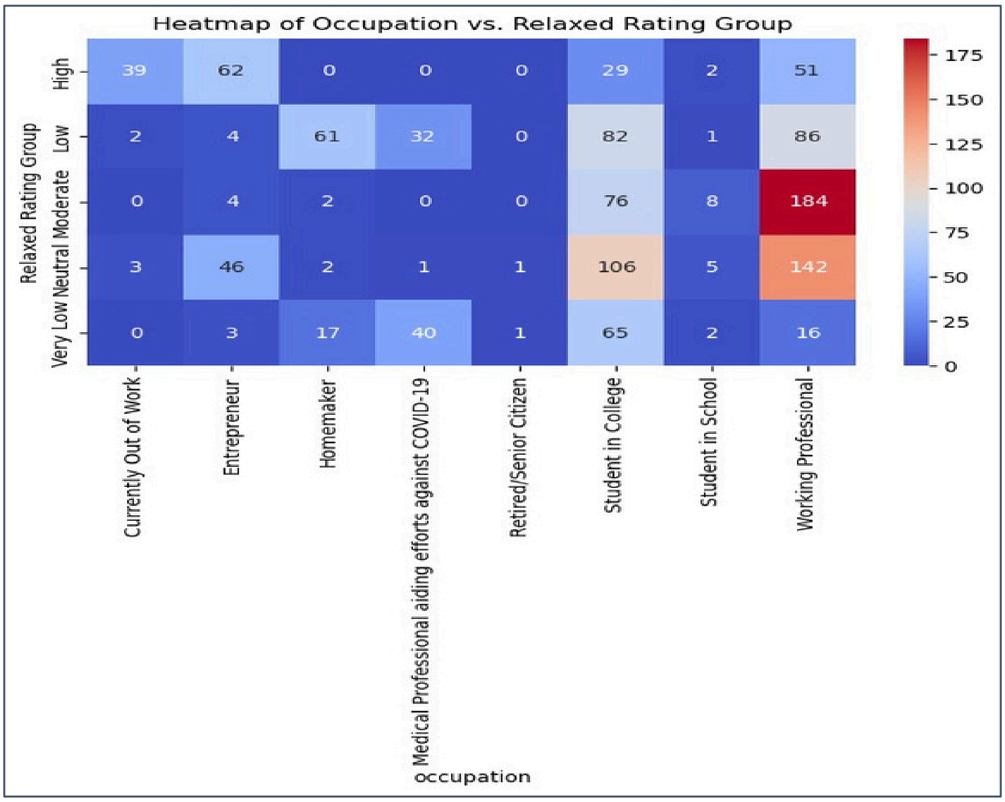
1.

Bar Graph for "Age Vs Occupation"



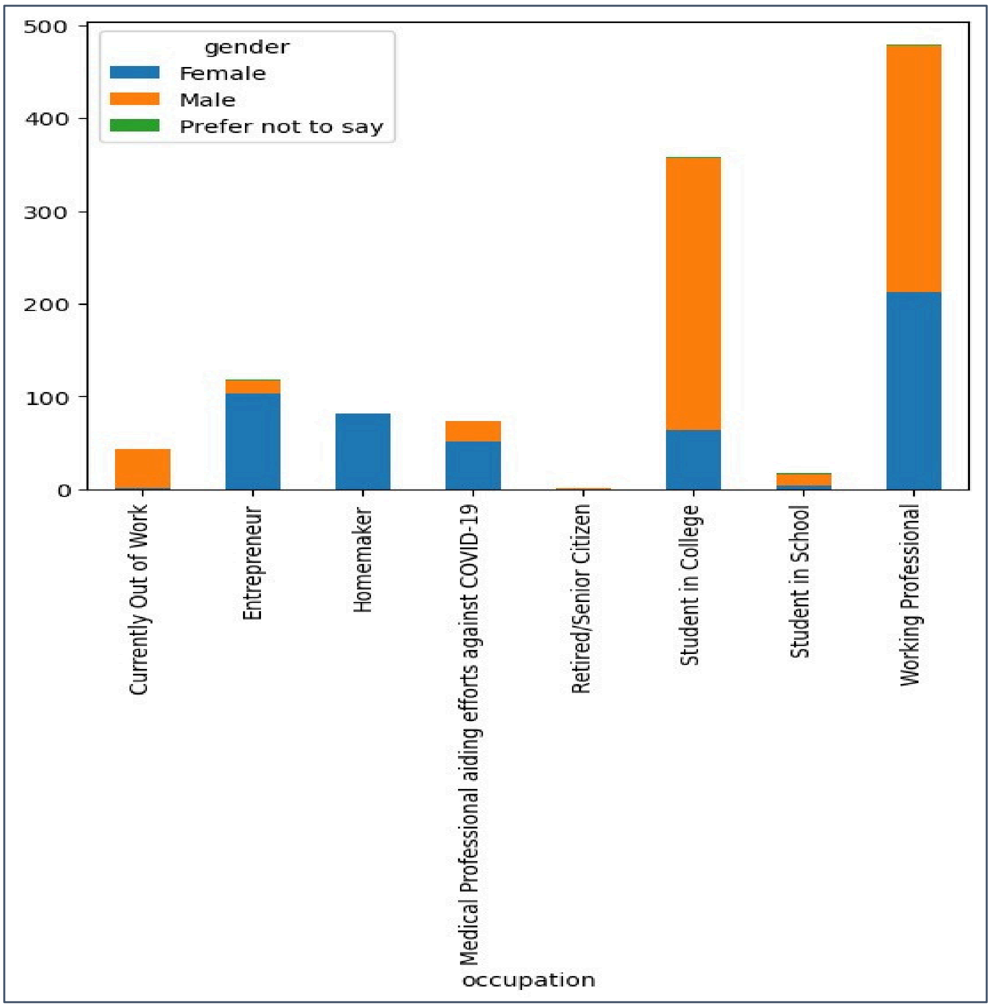
The bar graph illustrates the relationship between different occupations and age groups. The age group of 19-25 has the highest number of individuals identified as ' student in college'. On the other hand, the individuals who are 26 years of age and older are not mainly referred to as 'student in school' nor 'student in college'. It is seen that age group of 26–32 displays the most of them are 'working professionals' and 'entrepreneurs'. The most who are 60 years of age above are "currently out of work," with the least number fall into the category of 'currently out of work'. The age group '50-60' has the largest percentage of 'homemakers.'

٢. Trends between occupations and relaxed rating.



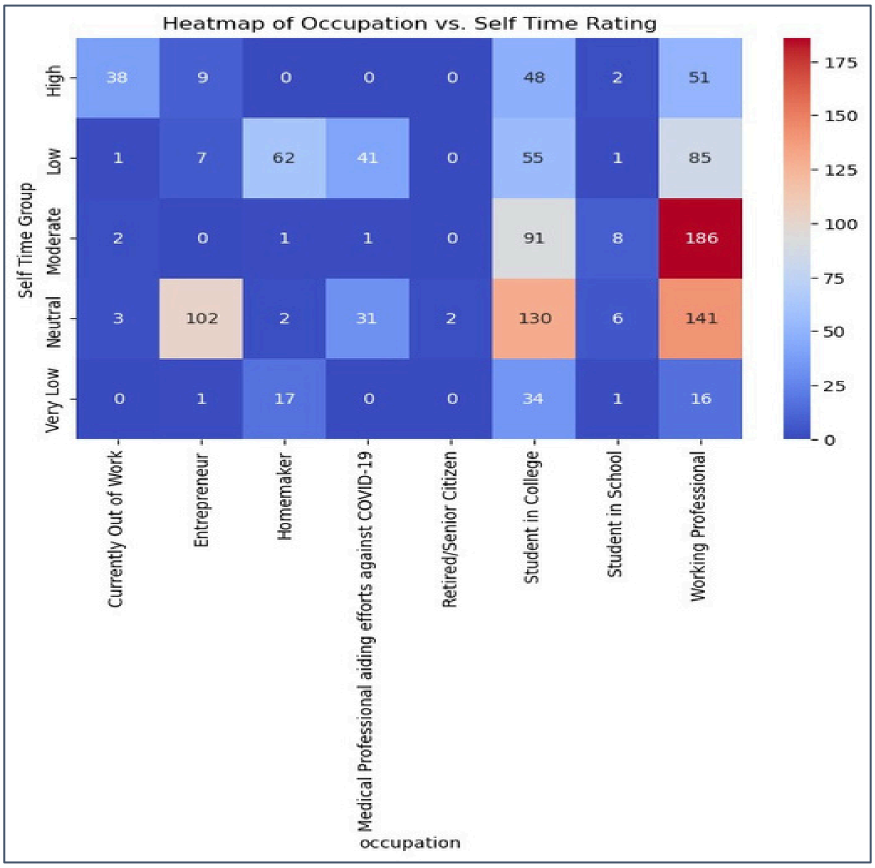
The heatmap illustrates the count of rating reflecting how relaxed the person is feeling based on occupation group using a colour gradient, with lighter colours indicating higher values. The ‘working professionals’ individuals have highest number of moderate rating followed by neutral rating. Meanwhile, individuals who are ‘Entrepreneur’ has rated highest number of high rating followed ‘working professionals’. Overall, the heatmap shows that there are some general trends in the data. For example, people who are currently out of work tend to have lower relaxed rating scores than people who are working.

3.
Bar Graph for "Gender Vs Occupation"



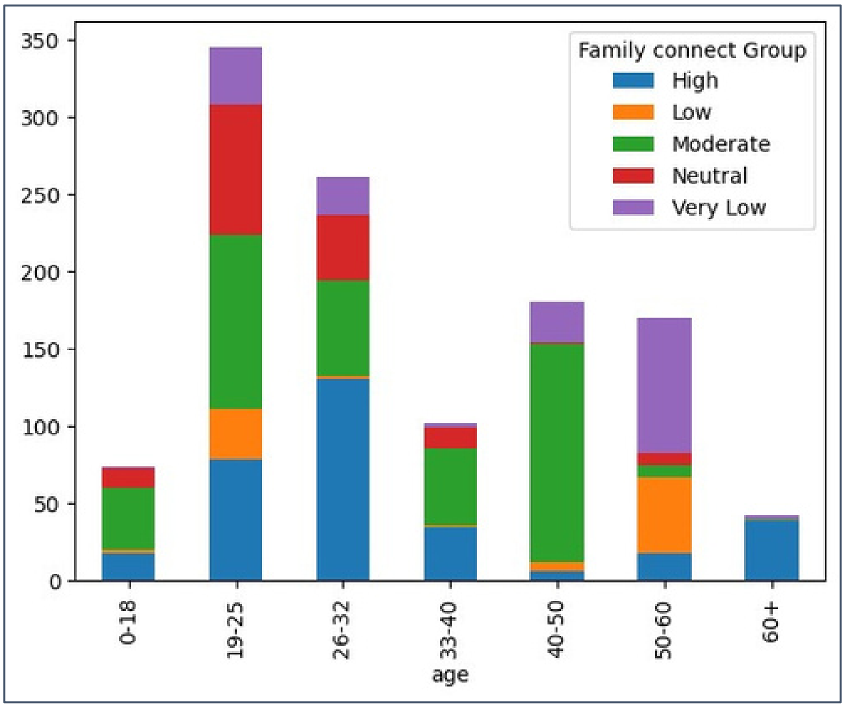
The bar graph depicts the relationship between different 'occupation' and'gender'. The occupation group of 'working professional' has the almost equal numberofindividuals identified as 'male' and 'female'. On the other hand, the individuals whoare'student in college' are 'male'. It is observed that there are 81 individuals categorizedas'homemakers,' and all of them are females. The most entrepreneurs are females, and theleast are those who prefer not to say . The lowest number of males and females are found amongretired/senior citizens.

4.
Trends between occupations and Self Time rating



The heatmap illustrates the count of rating reflecting the amount of self-time the individuals have procured in different occupations using a colour gradient, with lighter colours indicating higher values. The ‘working professionals’ individuals have highest number of moderate rating followed by neutral rating. The individuals who are ‘student in college’ has rated highest number of neutral rating followed ‘entrepreneur’. Overall, the heatmap shows that there are some general trends in the data. For example, people who are currently out of work/medical professionals/ retired or senior citizens tend to have lower self-time rating scores than ‘student in college’.

5. Bar Graph for “Age group Vs rating of connected with family”

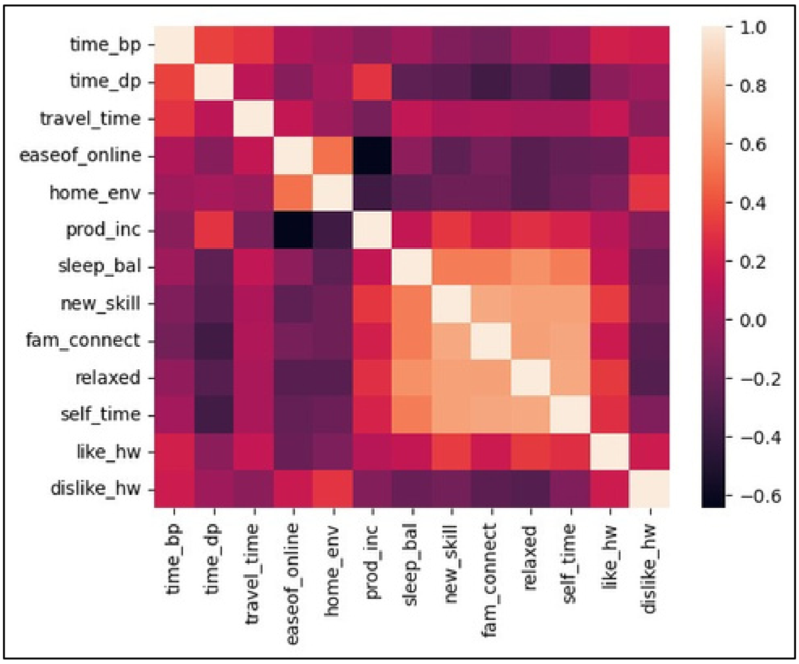


The bar graph illustrates the relationship between different agegroups and the ratings indicating how well individuals are connected with their families.The age group of 40-50 has one of the highest counts of individuals with moderate ratings.Onthe other hand, individuals in the '26-32' age group have the highest numberofcounts with 'high rating’. It is observed that the count of individuals in the age group of 50-60has a 'very low' rating. The most number of counts for ‘neutral’ rating is seen in 19-25 agegroup. The age group of 60 and above has the highest number of high ratings.

Data Preparation

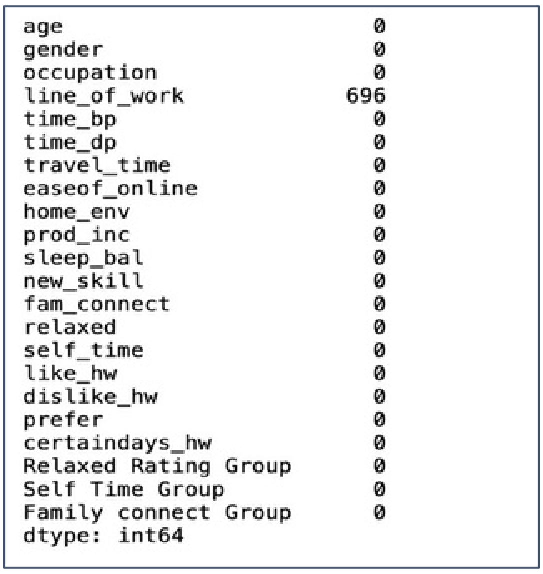
Correlation Analysis

Finding the correlation among the categorical features



The heatmap visually illustrates the relationships among categorical features. Lighter shades suggest features positively linked, while darker shades imply negative connections. Notably, "Relaxed feature" and "Self-Time" show a strong positive link, indicating those feeling more relaxed tend to spend more time on themselves. Additionally, "new_skill" and "fam_connect" exhibit positive correlation, suggesting those with stronger family connections are more inclined to acquire new skills. Conversely, "easeof_online" and "prod_inc" demonstrate a moderate negative correlation, revealing challenges in adapting to a purely online work environment, impacting productivity negatively.

Handling Missing Values



```
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
Relaxed Rating Group       0
Self Time Group            0
Family connect Group       0
dtype: int64
```

line_of_work	Teaching, Engineering, Management, Other, Government, Employee, Ar chitect, APSPDCL, Architecture
Occupation	Working Professional, Student in College, Entrepreneur, Homemaker, Medical Professional aiding efforts against COVID-19, Currently Out of Work, Student in School, Retired/Senior Citizen

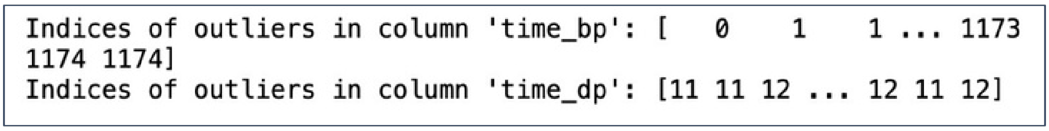
In the dataset the column” line_of_work” has 696 missing values. The "Occupation" and "line_of_work" columns consists of similar attributes, both indicating people’s profession. As the column "line_of_work" contains numerous missing values, dropping this column.

Dropping Column

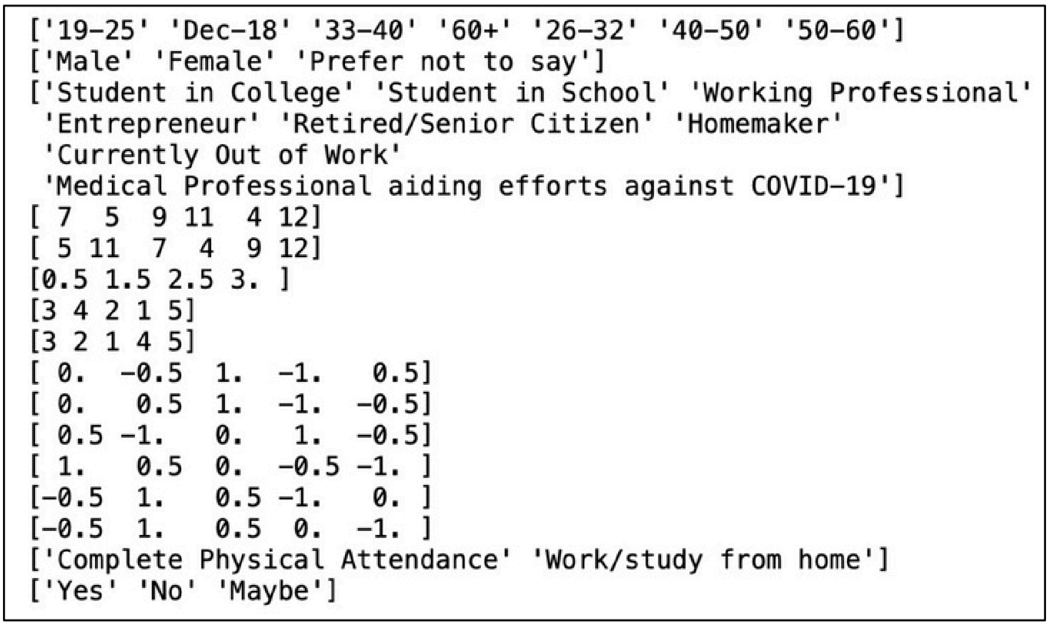
like_hw	100, 1100, 1000, 1110, 1111 , 10 ,110 ,1001 ,1 ,1011, 11,1101,101 ,111
dislike_hw	1111 ,1, 101,1100 ,111 ,1101,1011 ,100 ,1010 ,1001 ,1110 ,10,110 , 11

Given that the "like_hw" and "dislike_hw" features lack meaningful information, it would be prudent to remove these features from the dataframe before using it in machine learning models.

Handling outliers column



```
Indices of outliers in column 'time_bp': [   0    1    1 ... 1173
1174 1174]
Indices of outliers in column 'time_dp': [11 11 12 ... 12 11 12]
```

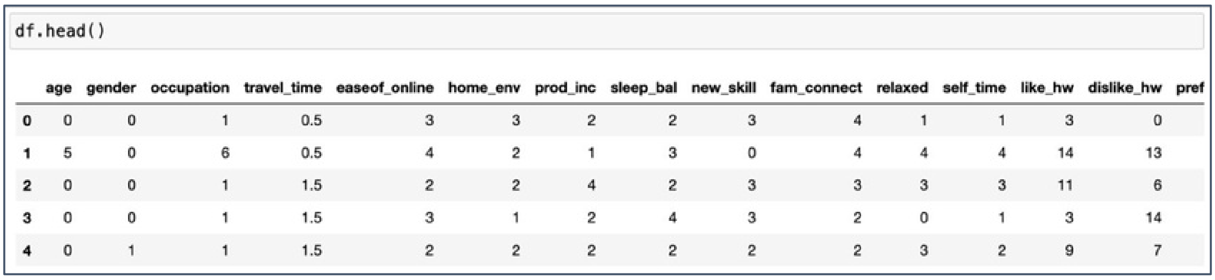


```
['19-25' 'Dec-18' '33-40' '60+' '26-32' '40-50' '50-60']
['Male' 'Female' 'Prefer not to say']
['Student in College' 'Student in School' 'Working Professional'
 'Entrepreneur' 'Retired/Senior Citizen' 'Homemaker'
 'Currently Out of Work'
 'Medical Professional aiding efforts against COVID-19']
[ 7  5  9 11  4 12]
[ 5 11  7  4  9 12]
[0.5 1.5 2.5 3. ]
[3 4 2 1 5]
[3 2 1 4 5]
[ 0.  -0.5  1.  -1.   0.5]
[ 0.   0.5  1.  -1.  -0.5]
[ 0.5 -1.   0.   1.  -0.5]
[ 1.   0.5  0.  -0.5 -1. ]
[-0.5  1.   0.5 -1.   0. ]
[-0.5  1.   0.5  0.  -1. ]
['Complete Physical Attendance' 'Work/study from home']
['Yes' 'No' 'Maybe']
```

The presence of outliers in the dataset, 'time_bp' and 'time_dp' columns. Data points that

significantly contrast with the majority of the data are referred to as outliers. The decision has been made based on the dataset, which comprises only three numerical columns: 'time_bp,' 'time_dp,' and 'travel_time.' Consequently, we can disregard the outliers as part of the analysis. This ensures a dataset analysis which is more reliable and robust.

Conversion of categorical columns into numerical form Numerical representations for the categorical values in the dataset are generated using a transformation process. For this, replacement function is applied to provide separate categories inside the category columns unique number codes. By checking that the category data is correctly encoded, this step enables it simpler to use the data in analysis and predictive modeling. **Apply label encoding to each categorical column** The code uses LabelEncoder from scikit-learn to convert categorical columns in a dataset (df) into numerical format, ensuring proper representation of categorical data for analysis and modeling in the dataset.



```
df.head()
```

	age	gender	occupation	travel_time	easeof_online	home_env	prod_inc	sleep_bal	new_skill	fam_connect	relaxed	self_time	like_hw	dislike_hw	pref
0	0	0	1	0.5	3	3	2	2	3	4	1	1	3	0	
1	5	0	6	0.5	4	2	1	3	0	4	4	4	14	13	
2	0	0	1	1.5	2	2	4	2	3	3	3	3	11	6	
3	0	0	1	1.5	3	1	2	4	3	2	0	1	3	14	
4	0	1	1	1.5	2	2	2	2	2	2	3	2	9	7	

Classification

Classification Report : “Gender”

Elimination of Numerical Attribute for improved classification

The column "travel_time" contains numerical values, it has been dropped from the dataset. Numerical attributes can interfere with the desired modeling approach, thus eliminating them is essential for classification tasks which attempt predict category data.

Splitting the Dataset

The dataset was split into a training set (70%) and a testing set (30%) to evaluate the performance of the models.

Classification Algorithms

Three classification algorithms were selected for this task:

a. Logistic Regression

b. Decision Tree

c. Random Forest

d. K- Nearest Neighbors

Model Training and Evaluation

Each algorithm was trained on the training data and evaluated on the testing data using the following evaluation metrics:

- Accuracy
- Precision (weighted)
- Recall (weighted)
- F1-score (weighted)

Results

The following are the results of the classification task for each algorithm:

a. Logistic Regression

Evaluation Result:

Accuracy	٧٩٪
Precision	٧٩٪
Recall	٨٠٪
F١-Score	٧٩٪

Confusion Matrix:

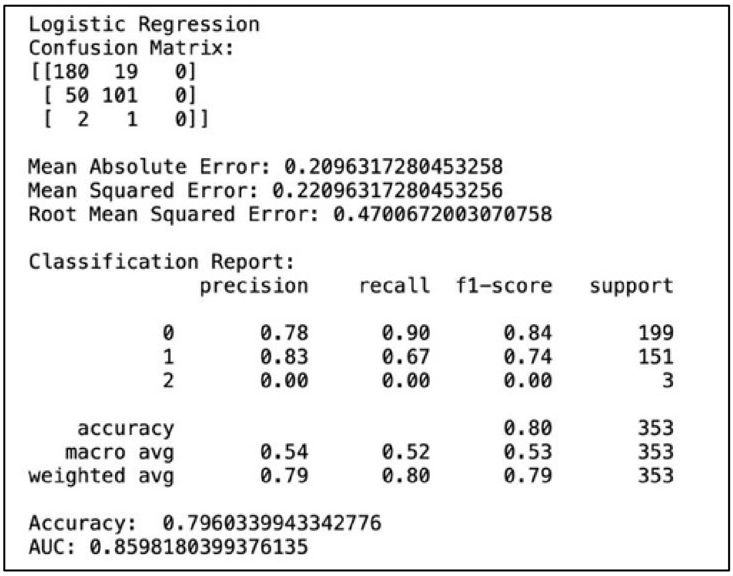
Class	True Positive	False Positive	True Negative	False Negative
Class 0	180	0	0 0 2	19
Class 1	101	0		50
Class 2	0	0		1

Cross-validation Results:

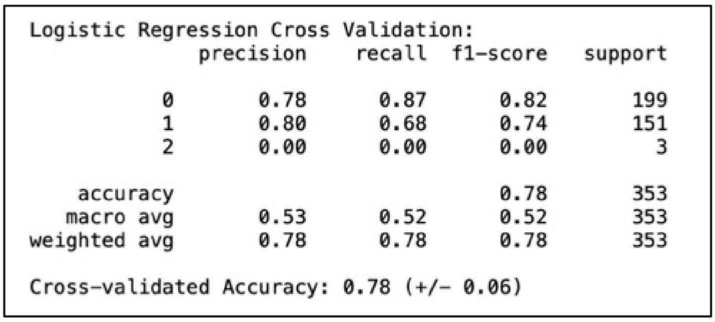
Mean	0.78
Standard Deviation	0.06

The Logistic Regression classification report breaks down how well the model predicted different gender categories. It did a good job with "Male (precision 78%, recall 90%, F1-score 84%) and "Female" (precision 83%, recall 67%, F1-score 74%) predictions but struggled with "Not Preferred to Say," possibly because there were very few instances of it. Overall, the model got things right about 79.6% of the time, which is decent. The ROC curve, a measure of overall performance, also looks good at 85.98%. When we tested the model in different scenarios (cross-validation), it stayed consistently accurate around 78%, but sometimes it varied by about 6%. So, in simple terms, the model is quite good, especially for predicting common gender categories, but it may not be as reliable for the less common "Not Preferred to Say" cases.

Output:



```
Logistic Regression
Confusion Matrix:
[[180  19   0]
 [ 50 101   0]
 [  2   1   0]]

Mean Absolute Error: 0.2096317280453258
Mean Squared Error: 0.22096317280453256
Root Mean Squared Error: 0.4700672003070758

Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.90      0.84       199
           1       0.83      0.67      0.74       151
           2       0.00      0.00      0.00         3

    accuracy                           0.80       353
   macro avg       0.54      0.52      0.53       353
weighted avg       0.79      0.80      0.79       353

Accuracy:  0.7960339943342776
AUC: 0.8598180399376135
```



```
Logistic Regression Cross Validation:
              precision    recall  f1-score   support

           0       0.78      0.87      0.82       199
           1       0.80      0.68      0.74       151
           2       0.00      0.00      0.00         3

    accuracy                           0.78       353
   macro avg       0.53      0.52      0.52       353
weighted avg       0.78      0.78      0.78       353

Cross-validated Accuracy: 0.78 (+/- 0.06)
```

b. Decision Tree

Evaluation Result:

Accuracy	٨٦٪
Precision	٨٦٪
Recall	٨٧٪
F١-Score	٨٦٪

Confusion Matrix:

Class	True Positive	False Positive	True Negative	False Negative
Class 0	178	25	128 180	3
Class 1	128	21	192	21
Class 2	0	1		23

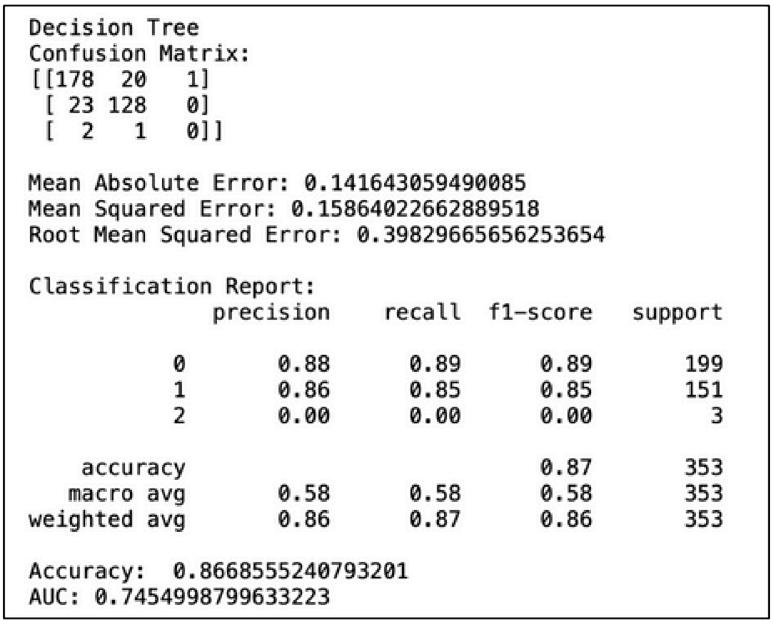
Cross-validation Results:

Mean	0.84
Standard Deviation	0.03

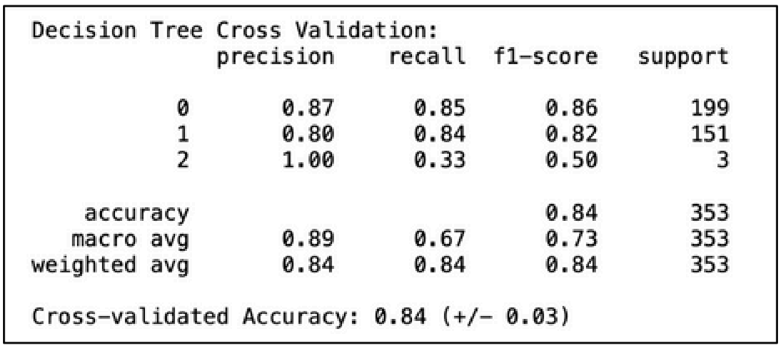
The model did a pretty good job overall, especially for predicting classes 'Male' and 'Female'. However, it struggled a bit with class 'Preferred not to say'. The accuracy of the model on the test set is about 86.6%, which means it got things right most of the time. The AUC score, which measures how well the model can tell the classes apart, is 74.5%. The cross-validated accuracy(84%) shows that the model performs well across different parts of the data.
To sum it up, the Decision Tree model is decent at guessing people's gender, but it has some trouble when individuals choose not to disclose their gender. The numbers give us a good idea of how well the model is doing on the test set and across different data subsets in cross-validation.

Output:

]]>



```
Decision Tree
Confusion Matrix:
[[178  20   1]
 [ 23 128   0]
 [  2   1   0]]

Mean Absolute Error: 0.141643059490085
Mean Squared Error: 0.15864022662889518
Root Mean Squared Error: 0.39829665656253654

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.89      0.89       199
           1       0.86      0.85      0.85       151
           2       0.00      0.00      0.00         3

    accuracy                           0.87       353
   macro avg       0.58      0.58      0.58       353
weighted avg       0.86      0.87      0.86       353

Accuracy:  0.8668555240793201
AUC: 0.7454998799633223
```



```
Decision Tree Cross Validation:
              precision    recall  f1-score   support

           0       0.87      0.85      0.86       199
           1       0.80      0.84      0.82       151
           2       1.00      0.33      0.50         3

    accuracy                           0.84       353
   macro avg       0.89      0.67      0.73       353
weighted avg       0.84      0.84      0.84       353

Cross-validated Accuracy: 0.84 (+/- 0.03)
```

c. Random Forest

Evaluation Result:

Accuracy	٨٩٪
Precision	٩٠٪
Recall	٩٠٪
F١-Score	٨٩٪

Confusion Matrix:

Class	**True Positive**	**False Positive**	**True Negative**	**False Negative**
Class 0	198	0	122	1
Class 1	119	0	201	32
Class 2	0	0	317	3

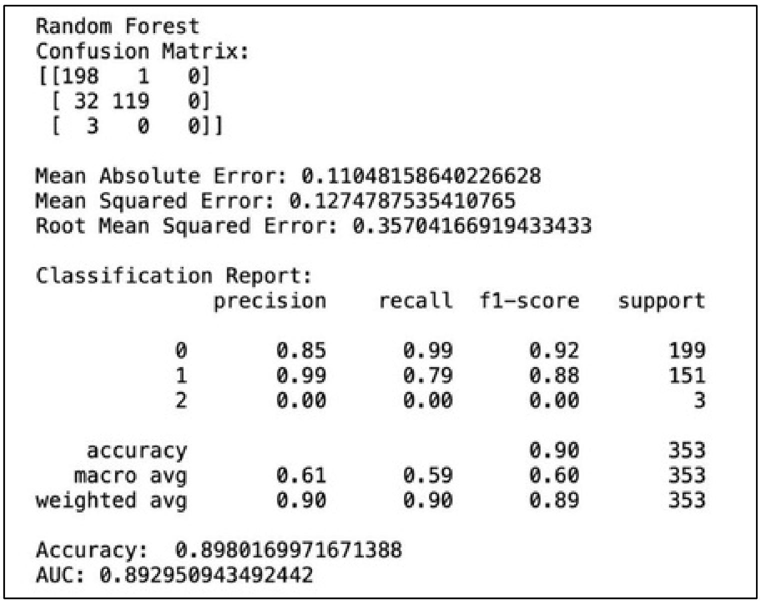
Cross-validation Results:

Mean	0.88
Standard Deviation	0.03

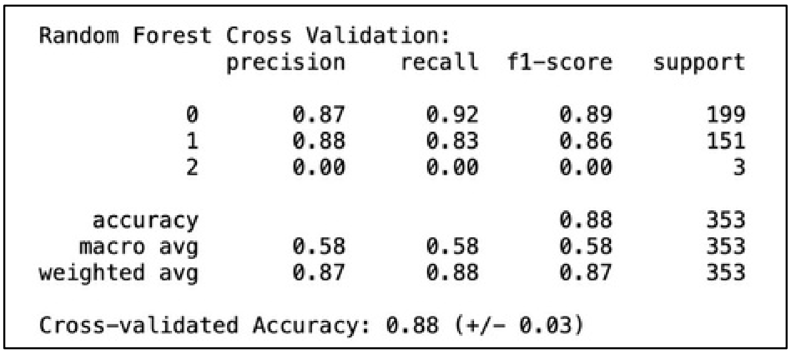
The model seems pretty good at predicting whether someone is male (class 0) or female (class 1), but it struggles when it comes to figuring out if someone prefers not to say (class 2). It gets the first two right most of the time, but it misses all the instances of people who prefer not to disclose their gender.

Overall, when we look at some fancy metrics like precision, recall, and F1-score, it's doing great for the male and female groups but not so hot for the prefer-not-to-say group. The accuracy score, which is like an overall grade, is 89.8%, and the AUC score, which is another performance measure, is 89.2%. These scores tell us that the model is generally good at its job. They also tested different models, and the cross-validation accuracy of Random Forest did well too, with an accuracy of 88% and a standard deviation of 3%. This means it's consistently reliable. However, the takeaway here is that while the model is good overall, it might need some improvement in understanding or predicting the prefer-not-to-say group.

Output:



```
Random Forest
Confusion Matrix:
[[198   1   0]
 [ 32 119   0]
 [  3   0   0]]

Mean Absolute Error: 0.11048158640226628
Mean Squared Error: 0.1274787535410765
Root Mean Squared Error: 0.35704166919433433

Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.99      0.92       199
           1       0.99      0.79      0.88       151
           2       0.00      0.00      0.00         3

    accuracy                           0.90       353
   macro avg       0.61      0.59      0.60       353
weighted avg       0.90      0.90      0.89       353

Accuracy:  0.8980169971671388
AUC: 0.892950943492442
```



```
Random Forest Cross Validation:
              precision    recall  f1-score   support

           0       0.87      0.92      0.89       199
           1       0.88      0.83      0.86       151
           2       0.00      0.00      0.00         3

    accuracy                           0.88       353
   macro avg       0.58      0.58      0.58       353
weighted avg       0.87      0.88      0.87       353

Cross-validated Accuracy: 0.88 (+/- 0.03)
```

d. K- Nearest Neighbors

Evaluation Result:

Accuracy	٨٤٪
Precision	٨٤٪
Recall	٨٥٪
F١-Score	٨٥٪

Confusion Matrix：

Class	True Positive	False Positive	True Negative	False Negative
Class 0	179	20	124 179	32 32 3
Class 1	121	21	350	
Class 2	0	0		

Cross-validation Results：

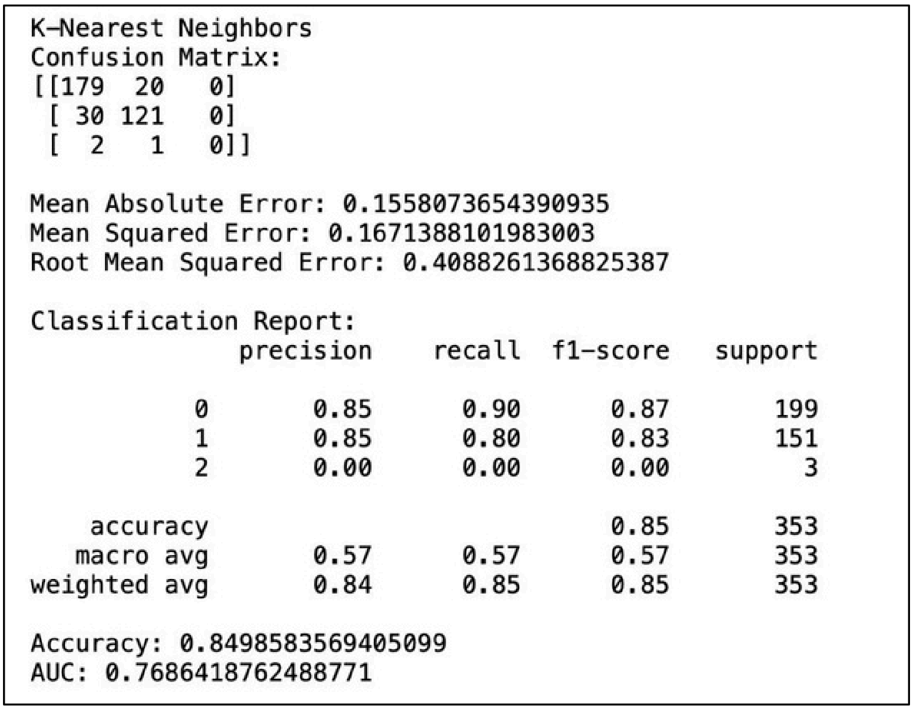
Mean	0.80
Standard Deviation	0.02

The classification report gives us more details about how well the model is doing for each gender category. The report talks about how well our model does in predicting whether someone is male, female, or prefers not to say. For male and female, it's doing pretty good—getting things right and not missing much. But for those who prefer not to say, the model struggles a bit, probably because there aren't many examples to learn from.

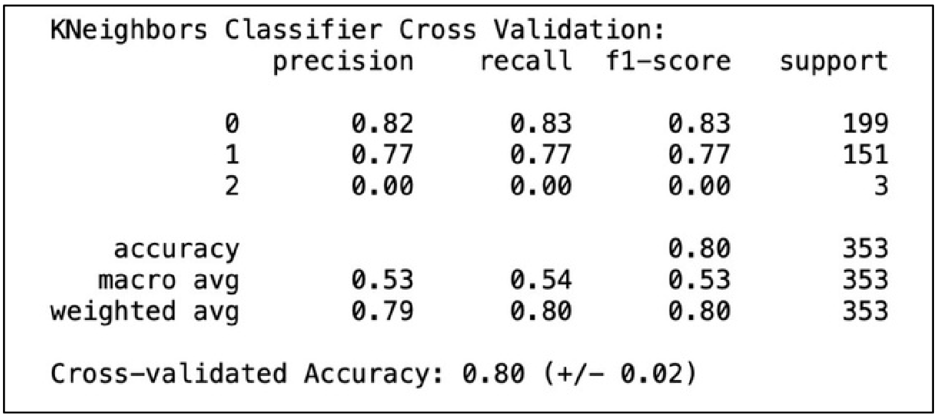
There's also a part about different models being tested, and one called K-Nearest Neighbors (KNN) is getting things right about 80% of the time, which is okay. It seems to work well with new data, but there's still room for improvement, especially in predicting the less common category.

They use something called AUC to measure how well the KNN model can tell the different groups apart. It's around 0.77, which is decent but not perfect. Overall, the report suggests the model is doing alright, but it's not great at predicting those who prefer not to say.

Output:



```
K-Nearest Neighbors
Confusion Matrix:
[[179  20   0]
 [ 30 121   0]
 [  2   1   0]]

Mean Absolute Error: 0.1558073654390935
Mean Squared Error: 0.1671388101983003
Root Mean Squared Error: 0.4088261368825387

Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.90      0.87       199
           1       0.85      0.80      0.83       151
           2       0.00      0.00      0.00         3

    accuracy                           0.85       353
   macro avg       0.57      0.57      0.57       353
weighted avg       0.84      0.85      0.85       353

Accuracy: 0.8498583569405099
AUC: 0.7686418762488771
```



```
KNeighbors Classifier Cross Validation:
              precision    recall  f1-score   support

           0       0.82      0.83      0.83       199
           1       0.77      0.77      0.77       151
           2       0.00      0.00      0.00         3

    accuracy                           0.80       353
   macro avg       0.53      0.54      0.53       353
weighted avg       0.79      0.80      0.80       353

Cross-validated Accuracy: 0.80 (+/- 0.02)
```

Performance evaluation of the models

Classification model	Accuracy Score	Precision	Recall	F-1 score
	79%	0.79	0.80	0.79
Logistic Regression	86%	0.86	0.87	0.86
Decision Tree	89%	0.90	0.89	0.88
Random Forest K-	84%	0.84	0.85	0.85
NearestNeighbors				

- **Logistic Regression** achieved an accuracy of 79%, making it a reasonable choice for this classification task. It is a simple and interpretable model.
- **Random Forest** achieved the highest accuracy among all the algorithms, indicating its ability to handle complex classification tasks.
- **Decision Tree** performs well, it's crucial to evaluate computational efficiency, and achieved an accuracy of 86.69% but had a lower AUC of 74.55%. Cross-validated accuracy was 85.00% with low variability (standard deviation of 2.00%). **K-Nearest**
- **Neighbors** performed similarly to Decision Tree but decision tree has better metrics.
-

Regression

The goal is to predict "self_time," which rates how much self-time is procured.

Two Regression algorithms were selected for this task:

a. Linear Regression

b. Random Forest Regression

Model Training and Evaluation

Both algorithms were trained on the training data and evaluated on the testing data. The following evaluation metrics were used:

- Root Mean Squared Error (RMSE)
- R-squared (R2) Score
- Mean Absolute Error

Results

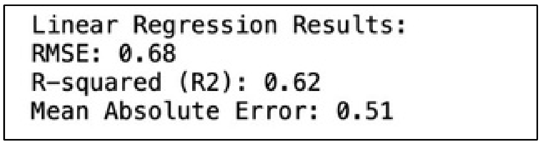
a. Linear Regression:

RMSE	0.68
R-squared	0.62
Means Absolute Error	0.51

- The Root Mean Squared Error : The model's predictions, on average, deviate by 0.68 units from the actual values.
- R-squared (R2): Approximately 62% of the variability in the "self_time" ratings is explained by the model, indicating moderate predictive power.
- Mean Absolute Error (MAE): On average, the model's predictions differ by 0.51 units from the true values.

The Linear Regression model, trained on features from the dataset, provides insights into predicting the "self_time" ratings. The evaluation metrics (RMSE, R2, MAE) offer a nuanced understanding of how well the model performs in capturing and explaining the variation in self-time ratings.

Output:



```
Linear Regression Results:
RMSE: 0.68
R-squared (R2): 0.62
Mean Absolute Error: 0.51
```

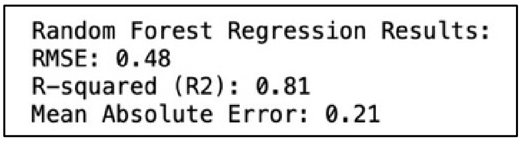
b. Random Forest Regression:

RMSE	0.48
R-squared	0.81
Means Absolute Error	0.21

- RMSE (Root Mean Squared Error): 0.48 Indicates the average prediction error, with lower values being better.
- R-squared (R2): 0.81 Represents the proportion of variance in the "self_time" that the model captures.
- Mean Absolute Error (MAE): 0.21 measures the average absolute difference between predicted and actual "self_time" values.

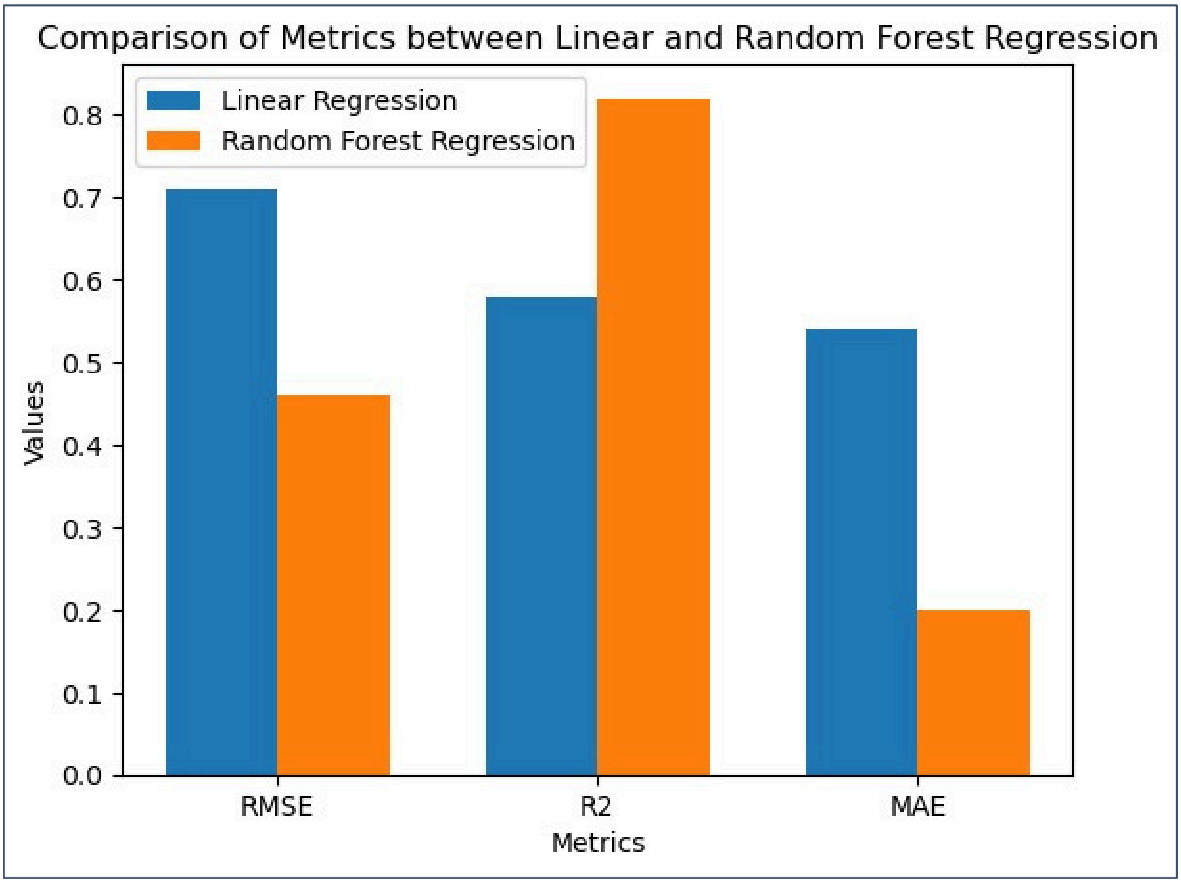
The Random Forest Regression model performs well, with low errors (RMSE and MAE) and a highR-squared value (81%). This suggests that the model accurately predicts the "self_time" ratings, capturing a significant portion of the variance in the data.

Output:



```
Random Forest Regression Results:
RMSE: 0.48
R-squared (R2): 0.81
Mean Absolute Error: 0.21
```

Performance Comparison of Linear Regression and Random Forest Regression



- RMSE (Root Mean Squared Error): Compared to linear regression, random forest generates predictions that are more precise as a result of its lower RMSE.
- With a higher R2 (Coefficient of Determination) than linear regression, random forest has the capacity to explain a greater percentage of the variability in the data.
- In comparison to linear regression, random forest produces predictions with less errors because to the smaller mean absolute error (MAE).

In summary, the graph clearly shows that random forest regression is more suitable than linear regression for this specific regression issue based on these criteria.

Clustering

K-**means clustering**

Silhouette Score:

The obtained silhouette score for K-means clustering : 0.19, which a moderate level of separa tion and coherence among the clusters. While not exceptionally high, it still signifies a reason able degree of cluster distinction.

Data Visualization for K-Means Clustering :
Figure ١ :

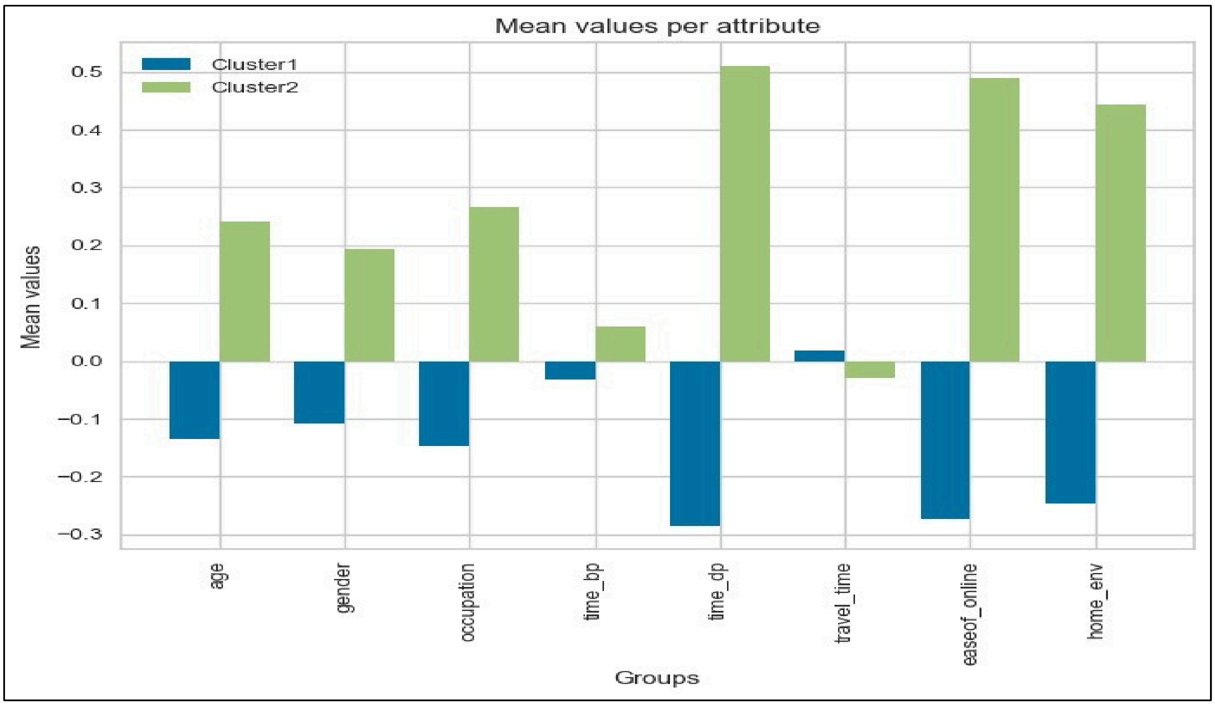
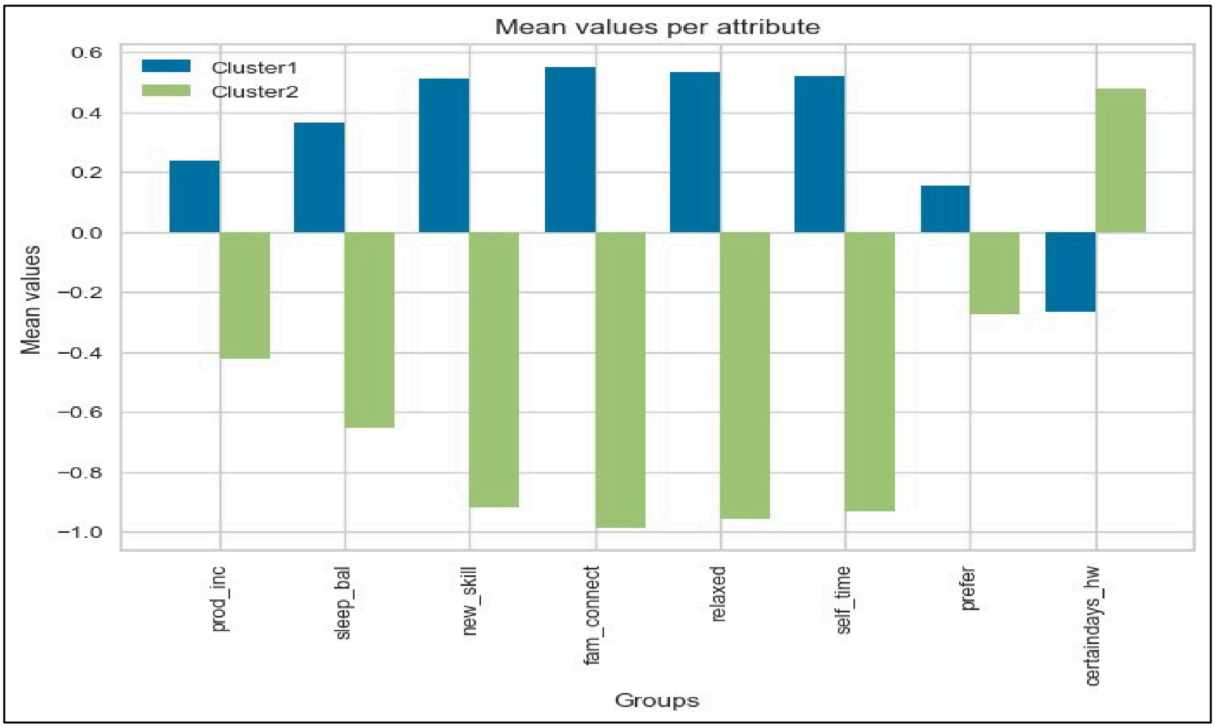


Figure ٢ :



K-Means Clustering:

In this section, the silhouette score for the K-Means clustering algorithm is computed to assess the quality of the clustering results. A silhouette score of 0.19 in K-Means clustering means the clusters are somewhat distinct but could be better. It's like organizing items on shelves—some are in the right place, but there's room for improvement. Adjusting how many shelves (clusters) or other factors might make things neater. It's a balance of keeping similar things together and making sure each group is clearly different from the others.

- The K-Means algorithm grouped the individuals into two clusters (Cluster 1 and Cluster 2).
- Figure 1 reveals key patterns in various features for the identified clusters. In Cluster 2, there are higher positive values for factors like ease of online work (easeof_online), time spent on work during the pandemic (time_dp), and liking of the home environment (home_env). Conversely, time spent on work before the pandemic (time_bp) displays lower positive values in Cluster 2, and the travel time (travel_time) in Cluster 1 is the lowest positive value. For these features, Cluster 2 also exhibits higher negative values. On the other hand, age group, occupation group, and gender group show lower positive and negative values, indicating less variability across the clusters. These findings suggest distinct trends and preferences associated with the identified clusters, providing valuable insights into the dataset.

- Figure 2 suggests two distinct clusters. In Cluster 2, factors like new_skill, fam_connect, relaxed, sleep_bal, and self_time have notably higher negative values. However, 'prefer' and 'prod_inc' exhibit low negative values in Cluster 2, while 'certaindays_hw' has low negative values in Cluster 1. Cluster 1 generally shows average positive values for new_skill, fam_connect, relaxed, sleep_bal, and self_time, whereas 'prefer' has the lowest positive value in Cluster 1.

Reference

- https://www.kaggle.com/code/shahkan/text-classification-using-logistic-regression
- https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html