PROGRAMMING FOR DATA ANALYSIS AND AI

COURSEWORK 1 : “Psychological Effects of COVID”

Up2224635

Table of Contents

Psychological Effects of COVID

Task 1: Descriptive Analytics

a. Basic Statistics
b. Data Visualisations

Task 2: Data Preparation

a. Handling of missing values
b. Removing outliers
c. Addition of new attributes
d. Normalisation

Task 3: Classification

Correlation Analysis

a. Decision Tree
b. K Nearest Neighbors
c. Adaptive Boosting

Task 4: Regression

a. Random Forest Regression
b. Ridge Regression

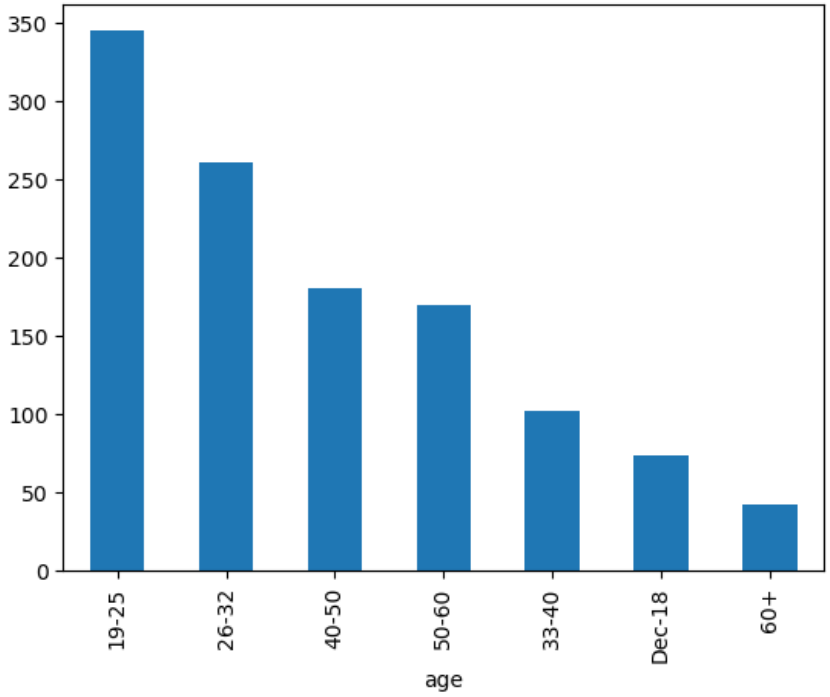
Task 5: Clustering

a. K Modes Clustering

Description for extra process used

References

Task 1:
Descriptive analytics:
A. Basic Statistics
1. Age :

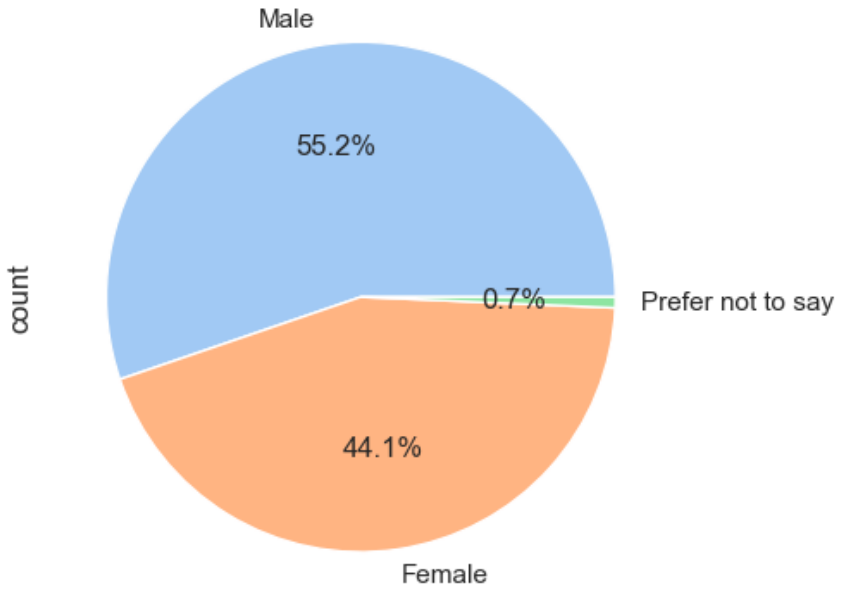


Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses
Age	Nominal	- 19-25	- 60+

The “Dec-18” category in the age is actually a mistake and it should be from 0 - 18 and this change has been made in the data pre-processing section.

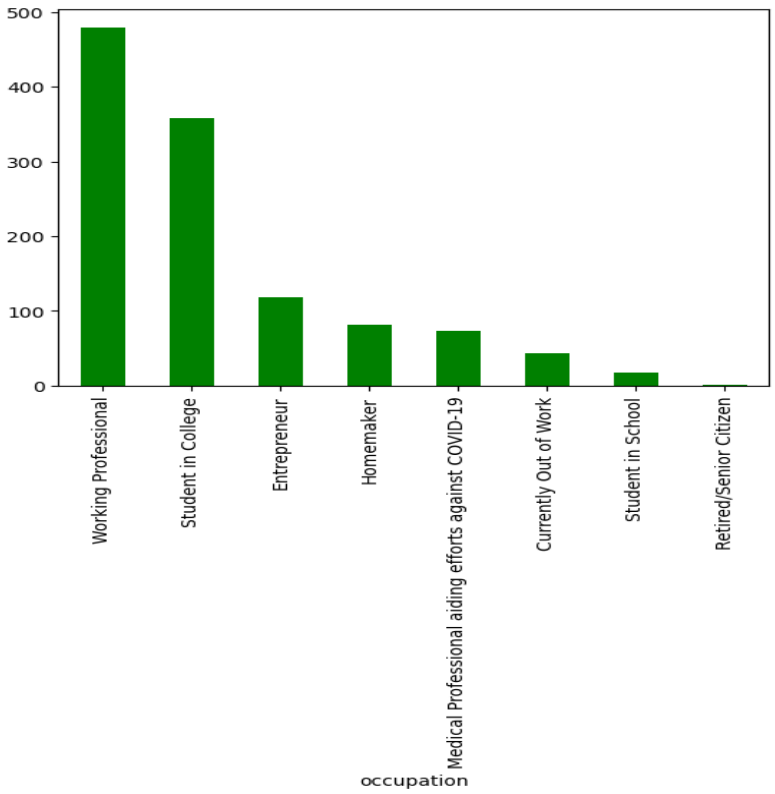
Majority of the respondents were youth candidates aged 19 to 32.

2. **Gender :**



The respondents consisted of the majority of the males leading by margin of 11.1%. The distribution of males and females being almost equal will give good results in further calculations. Only 0.7% of the respondents preferred not to reveal their age and hence they can be neglected in further calculations.

3. **Occupation:**

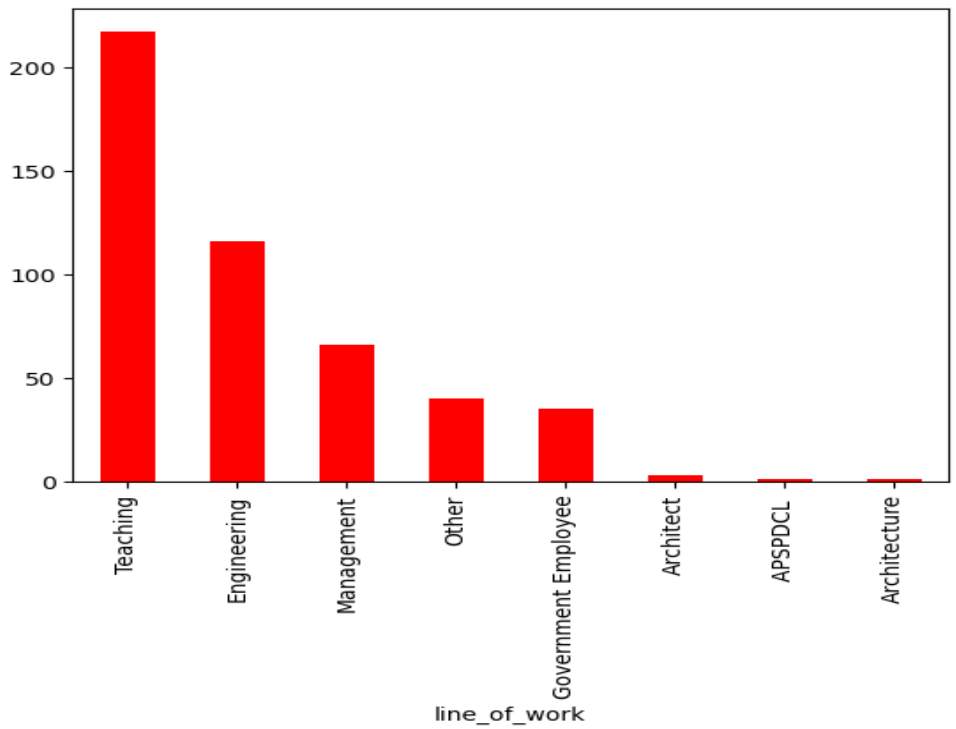


Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses
Occupation	Nominal	- Working professional	- Retired Citizens

Majority of the section is working professionals as most of the respondents are in the working age followed by students. Unique category to be seen was medical professionals working on aids of Covid and its involvement in good scale in the occupation list.

4. **Line of work :**

```
In [19]: data['line_of_work'].isnull().sum()
Out[19]: 696
```

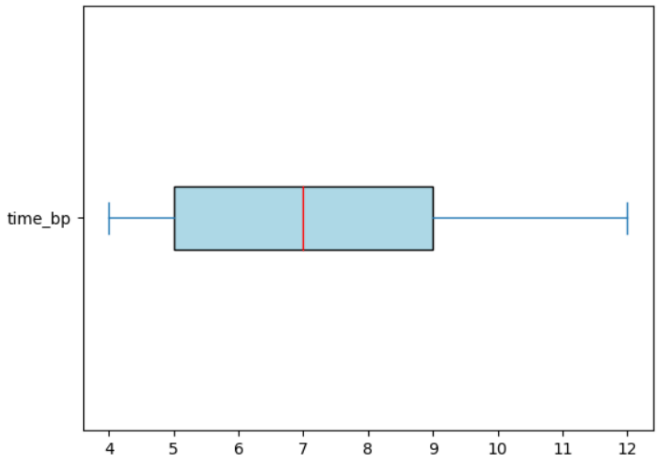


Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses	Missing values
Line of work	Nominal	- Teaching	- Architecture	696 missing values in total.

Teaching is the most preferred line of work followed by engineering. Government employees and management are also preferred but architecture and APSPDCL are the least preferred of all which justifies the least interest of the people in the field of arts and design.

5. **Time spent on work before pandemic:**

```
In [25]: data['time_bp'].describe()

Out[25]: 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
```



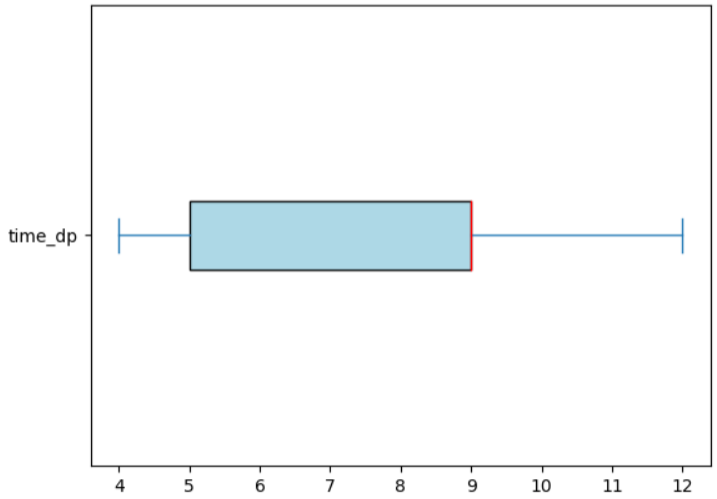
The average time spent on work before the pandemic was 7 hours 41 minutes. The least time spent on it was 4 hours a day and maximum was 12 hours. Most of the section was concentrated on 7 hours a day with no outliers in the category.

```
In [24]: from scipy.stats import skew

         # Calculate skewness
         skewness = skew(data['time_bp'])

         print(f'Skewness: {skewness}')

         Skewness: 0.35811914139791756
```

Most of the counts are on the lower side (around 4, 5, 7).
There are fewer occurrences of higher counts (9, 11, 12), contributing to the right skewness.

6. **Time spent on work during pandemic:**

```
In [29]: data['time_dp'].describe()

Out[29]: 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
```



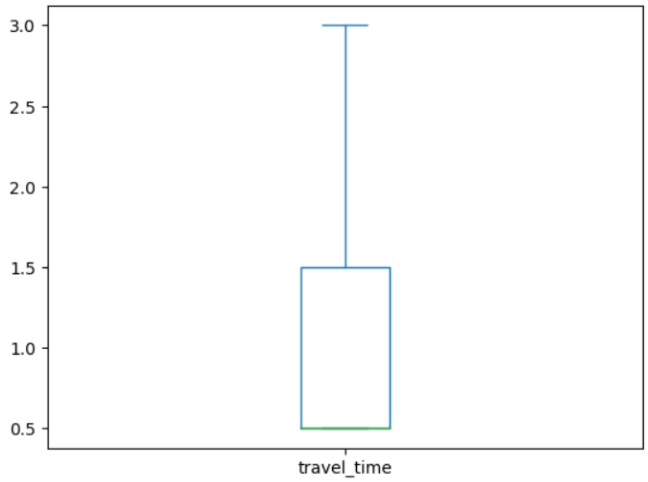
The average time spent on work during pandemic was 7.97 hours which is a growth than in terms of before pandemic. The minimum and maximum hours spent seem to remain the same as before the pandemic situation. The main thing to notice is that most of the values fall in the 9 hours of working category implying increase in work and studies during the pandemic.

```
In [31]: from scipy.stats import skew

         # Calculate skewness
         skewness = skew(data['time_dp'])

         print(f'Skewness: {skewness}')

         Skewness: 0.01169085673066641
```

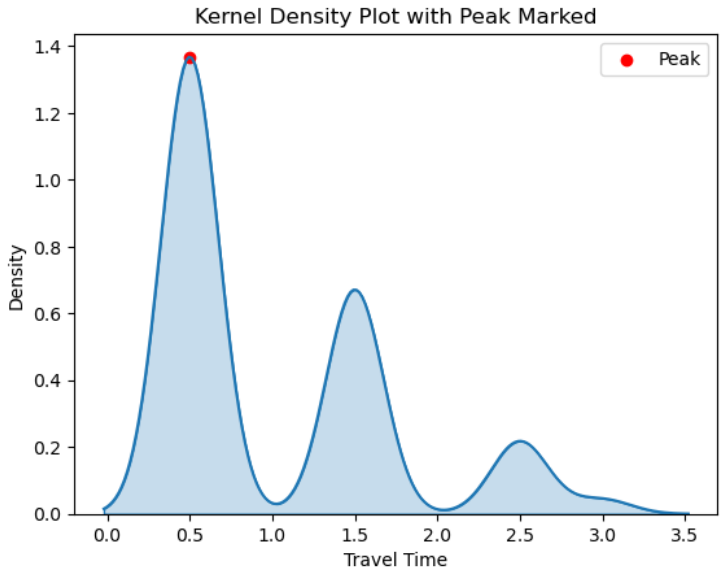
The majority of counts are concentrated around 4, 5, 7, and 9, with no strong indication of a tail on either side.
The positive skewness might be attributed to a subtle elongation of the right tail, but this effect is minimal.

7. **Travel time:**

```
In [33]: data['travel_time'].describe()

Out[33]: 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
```

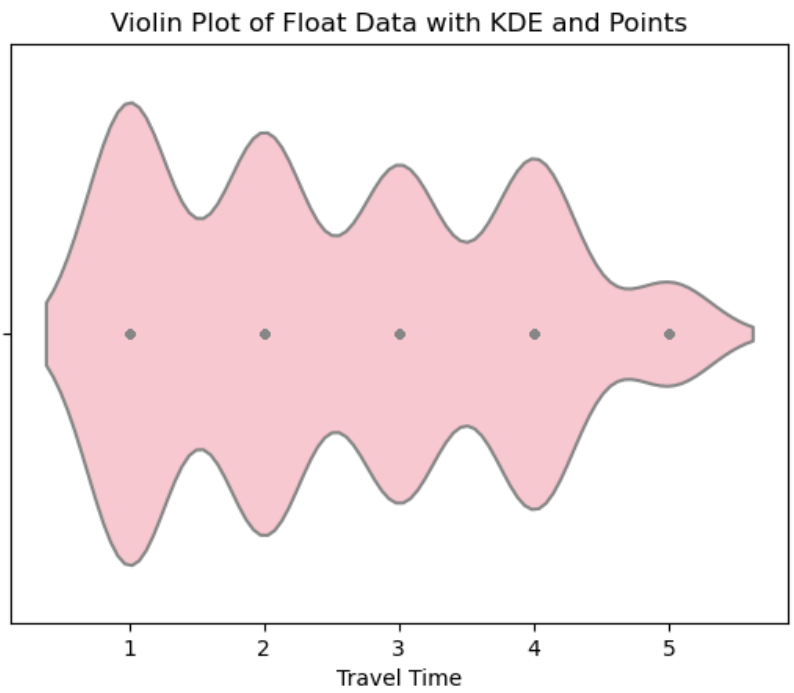


Average time spent on travel was 1 hour and the least was half hour. For the majority of the section it took only half an hour for travelling for work which is visible from the kernel density plot while the maximum anyone could take was 3 hours.The distribution of peak hours falled in mostly 0.5 , 1.5 and 2.5 hours.



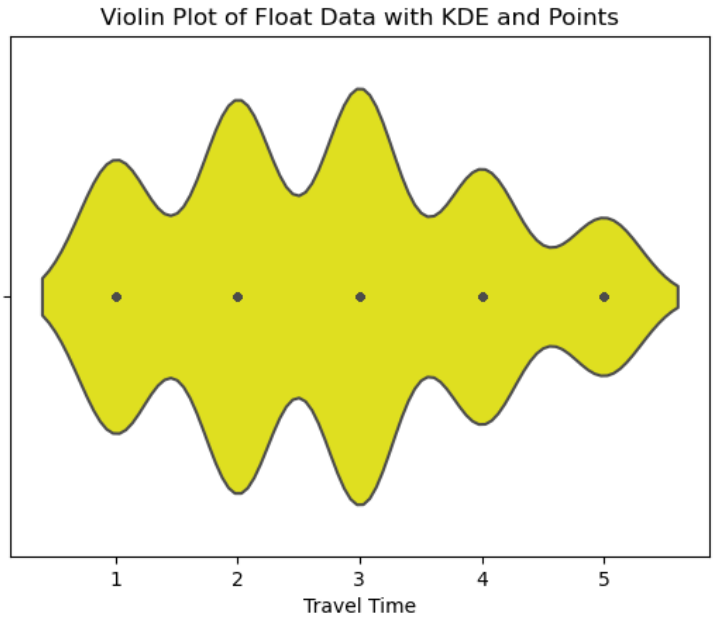
8. **Ease of Online :**

People preferred to have online work and liked the online work lifestyle better than the offline. From the violin chart it is clear that the majority of them supported online work while only a few did not prefer online mode.



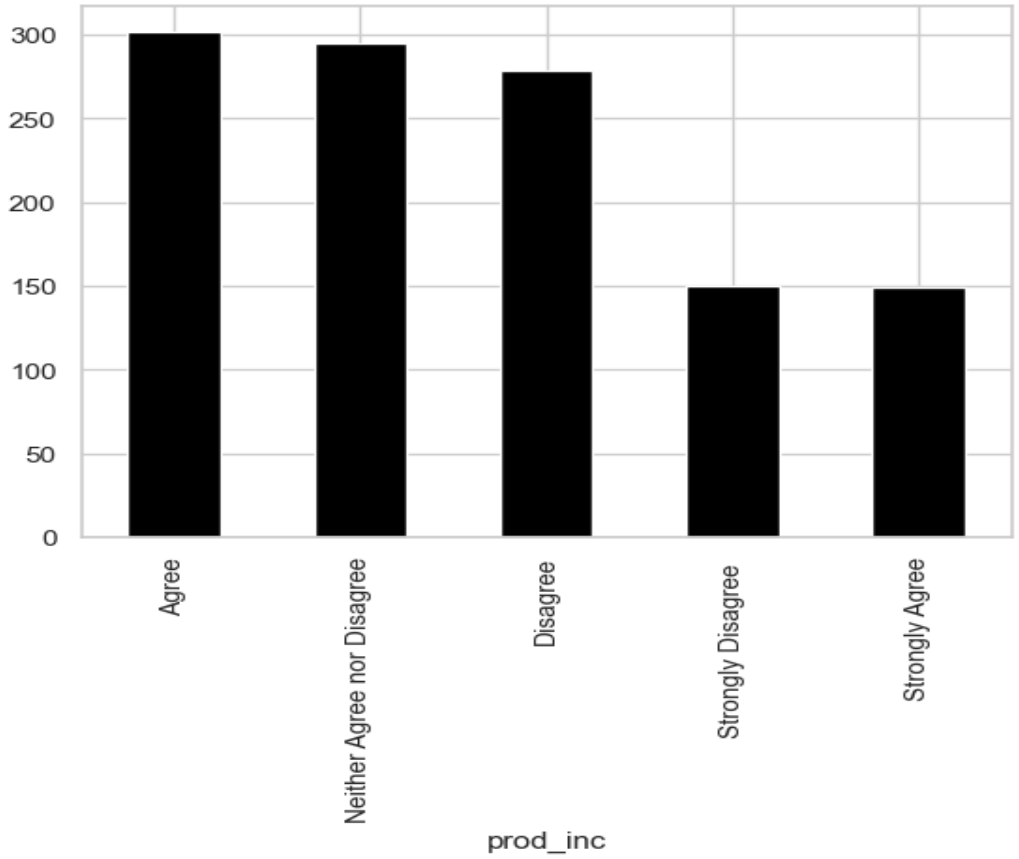
Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses
Ease of Online	Ordinal	1	5

9. Home Environment:



Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses
Home Environment	Ordinal	3 - Majority of the people took a neutral approach towards the liking of the home environment.	5 - Only a few didn't like the home environment and the possible reason for this might be the over time spent in the house.

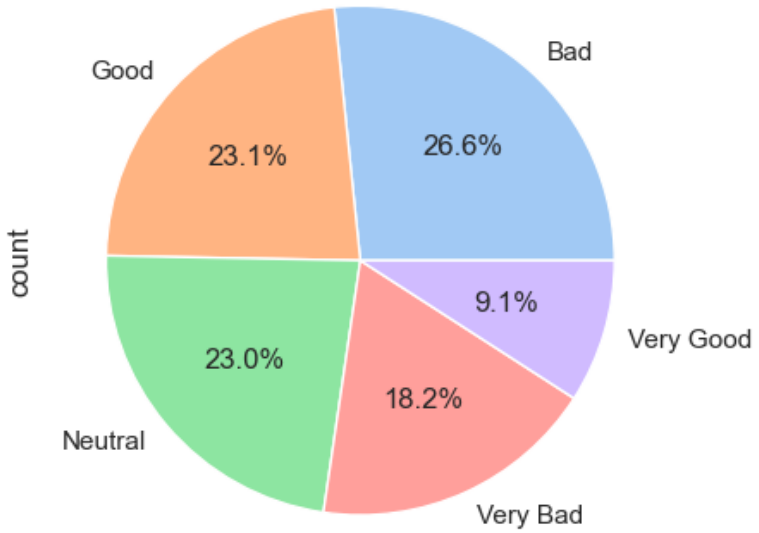
10. Productivity Increase :



Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses

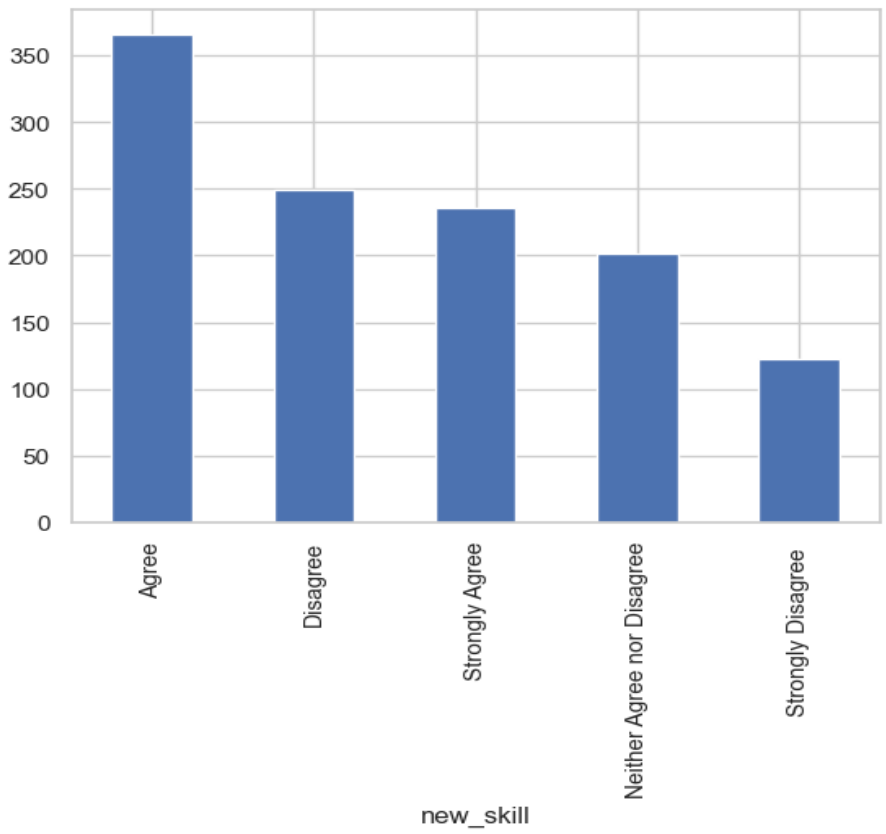
Productivity Increase	Ordinal	0.5 (Agree) - Majority of the people believed that their productivity increased in the pandemic possibly because of low stress and home environment.	1 (Strongly Agree) - Few responses strongly contradict the fact that people didn't state confidently whether their productivity increased in a pandemic.

11. Sleep balance :



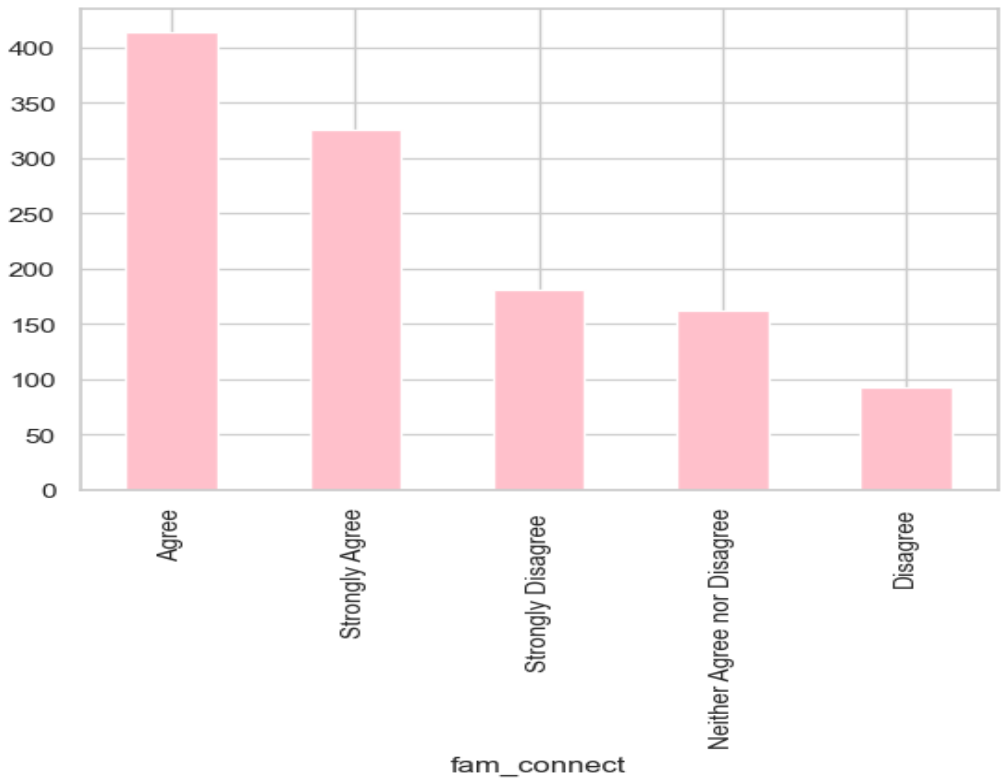
The sleep balance was almost in equal proportions distributed but majority of the people experienced a bad sleep cycle during covid affecting the mental and physical health.Only a few people could say that they had very good sleep while rest experienced neutral or bad sleep.

12. New skill learned :



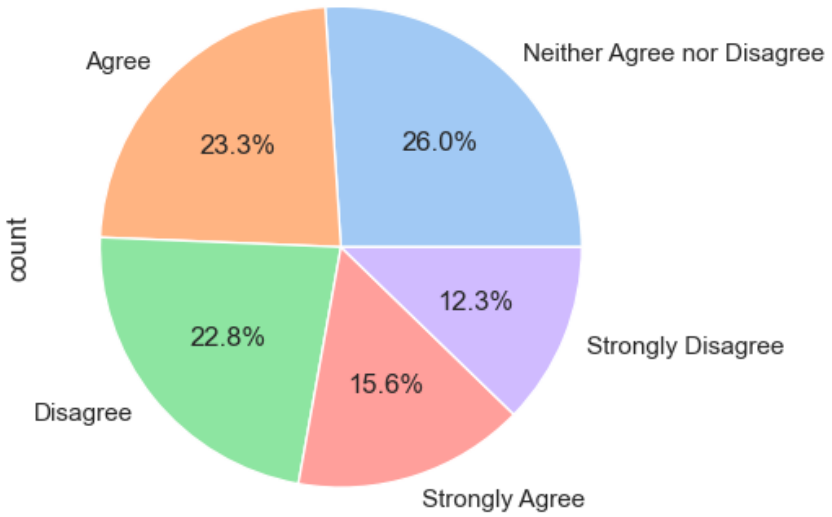
Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses
New Skill Learned	Ordinal	0.5 (Agree) - Majority of the respondents believed that they learned new skills in the pandemic showing a rising trend in productivity.	-0.5 (Strongly Disagree) - Very few people disagreed on the learning of new skills.

13. Family Connection :



Huge proportion of the respondents agreed to the fact that the family connection and time spent has increased in the pandemic as compared to usual days. The work from home culture has given both parents and children the opportunity to be with the close ones and spend quality time.

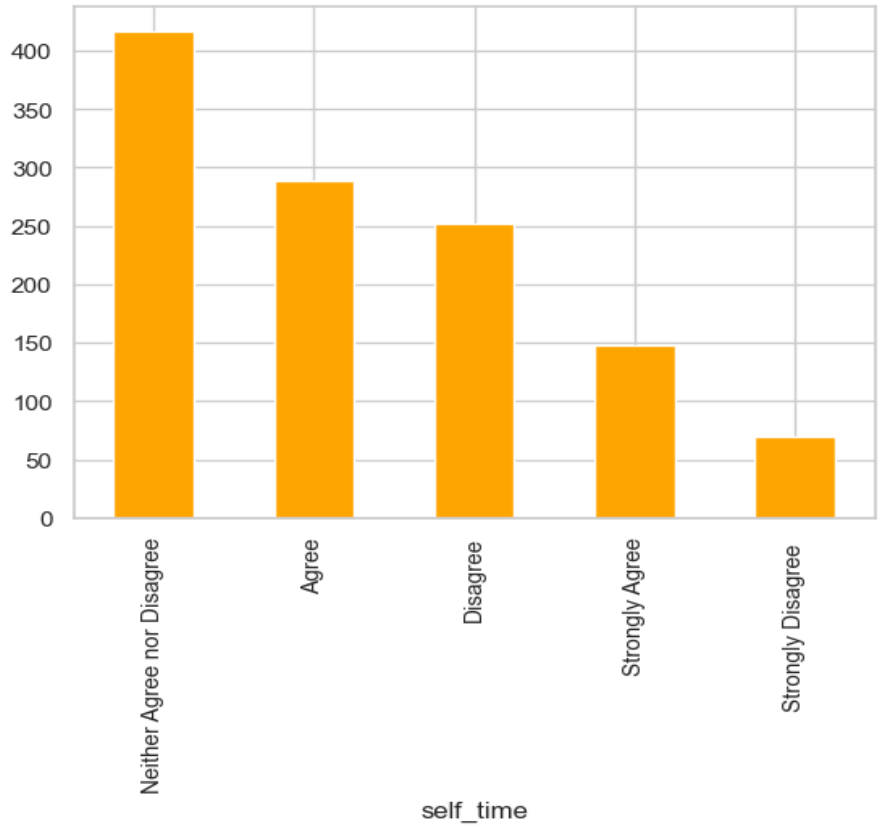
14. Relaxed :



Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses

Relaxed	Ordinal	0 (Neither Agree nor Disagree) - People didn't agree or disagree on them being relaxed during the pandemic mainly due to improper sleep cycle or lack of productivity.	-1 (Strongly Disagree) - Only 12.3% of the respondents strongly disagreed with the fact that they were able to relax during pandemic.

15. Self Time :

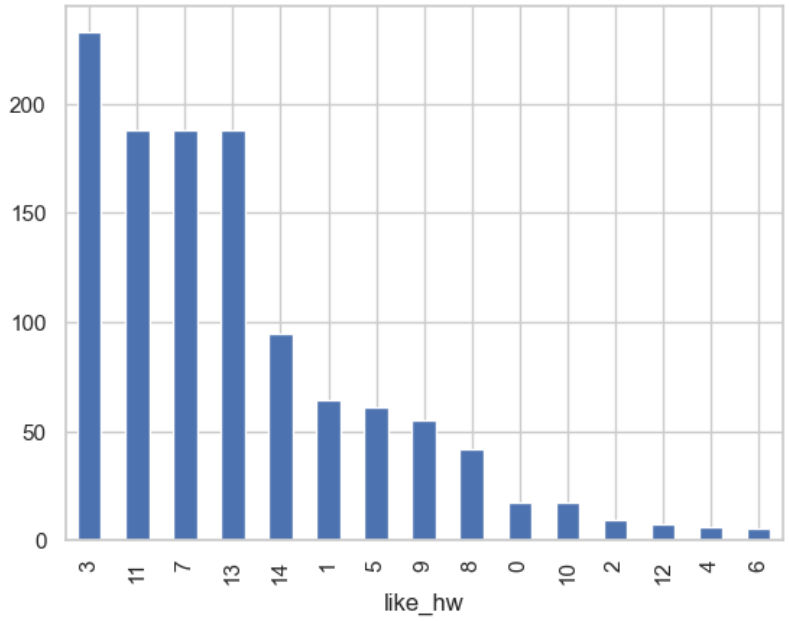


Attribute Name	**Type**	**Maximum Responses(Mode)**	**Minimum Responses**
Self Time	Ordinal	0 (Neither Agree nor Disagree) - People were not sure about the fact if they have been able to give themselves self time in the pandemic.	-1 (Strongly Disagree) - Only a few strongly disagreed stating the fact that these might be the same respondents as in the above category who didn't feel relaxed.

16. Like working from home :

The binary form was transformed into normal numerical form and converted into a ordinal likert scale from 0 to 14 with 7 as the neutral variable.

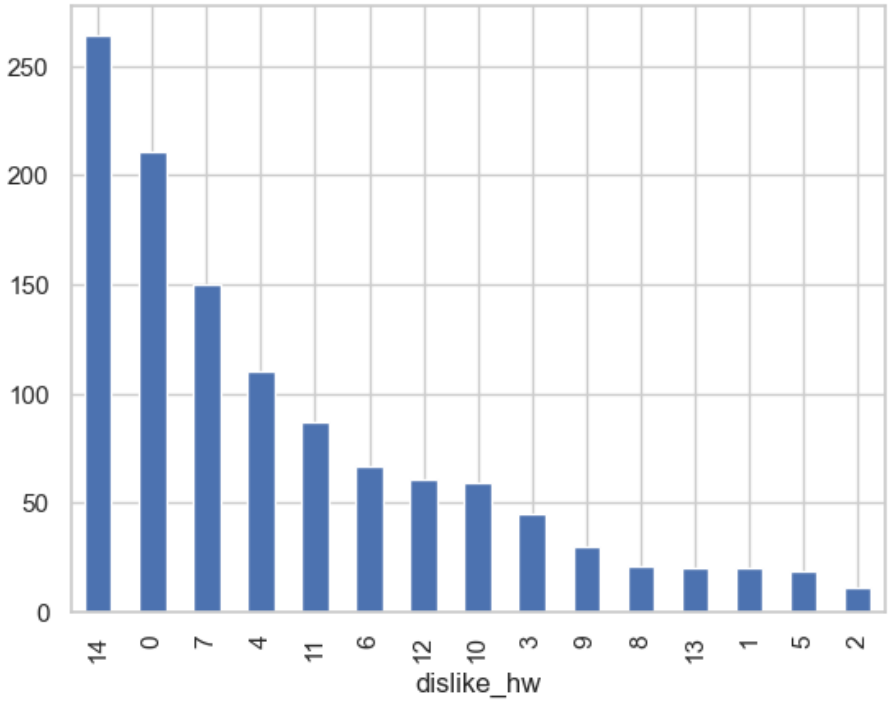
Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses
like_hw	Ordinal	3 - The maximum response being 3 implies that not many people liked the concept of working from home.	6 - The minimum response was 6 which states that few of them had a neutral approach.



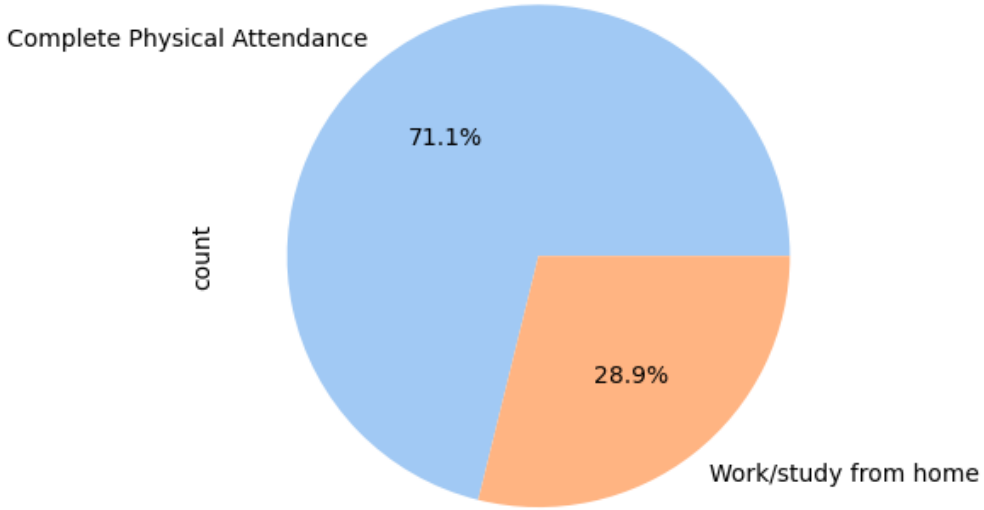
17. Dislike working from home :

The binary form was transformed into normal numerical form and converted into a ordinal likert scale from 0 to 14 with 7 as the neutral variable.

Attribute Name	Type	Maximum Responses(Mode)	Minimum Responses
dislike_hw	Ordinal	14 - The maximum response being 14 implies that many people disliked the concept of working from home.	2 - The minimum response was 2 which states that only a few of them were not ok with disliking the work from home concept.

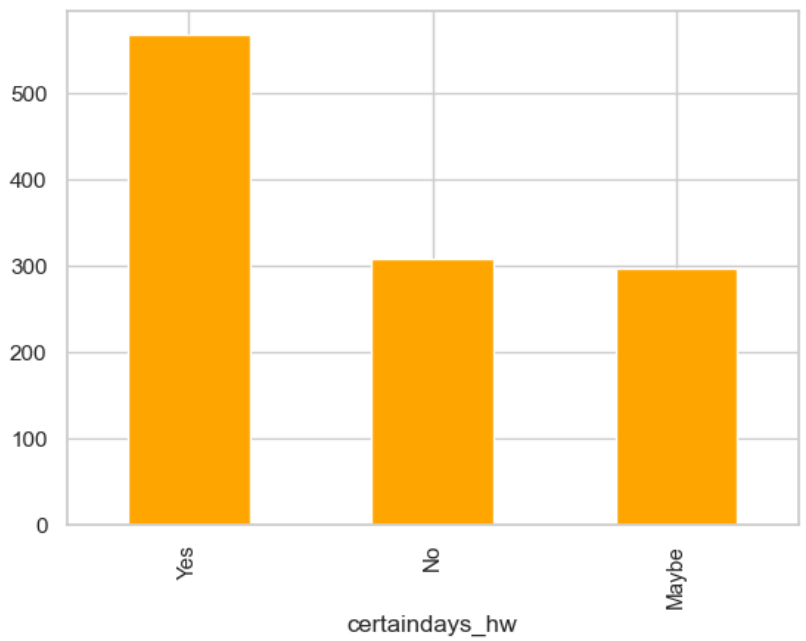


18. Prefer:



Majority of the students and working professionals found that working offline is better than online. There might be several factors contributing to this like lack of internet, miscommunications, explanation gap and much more.

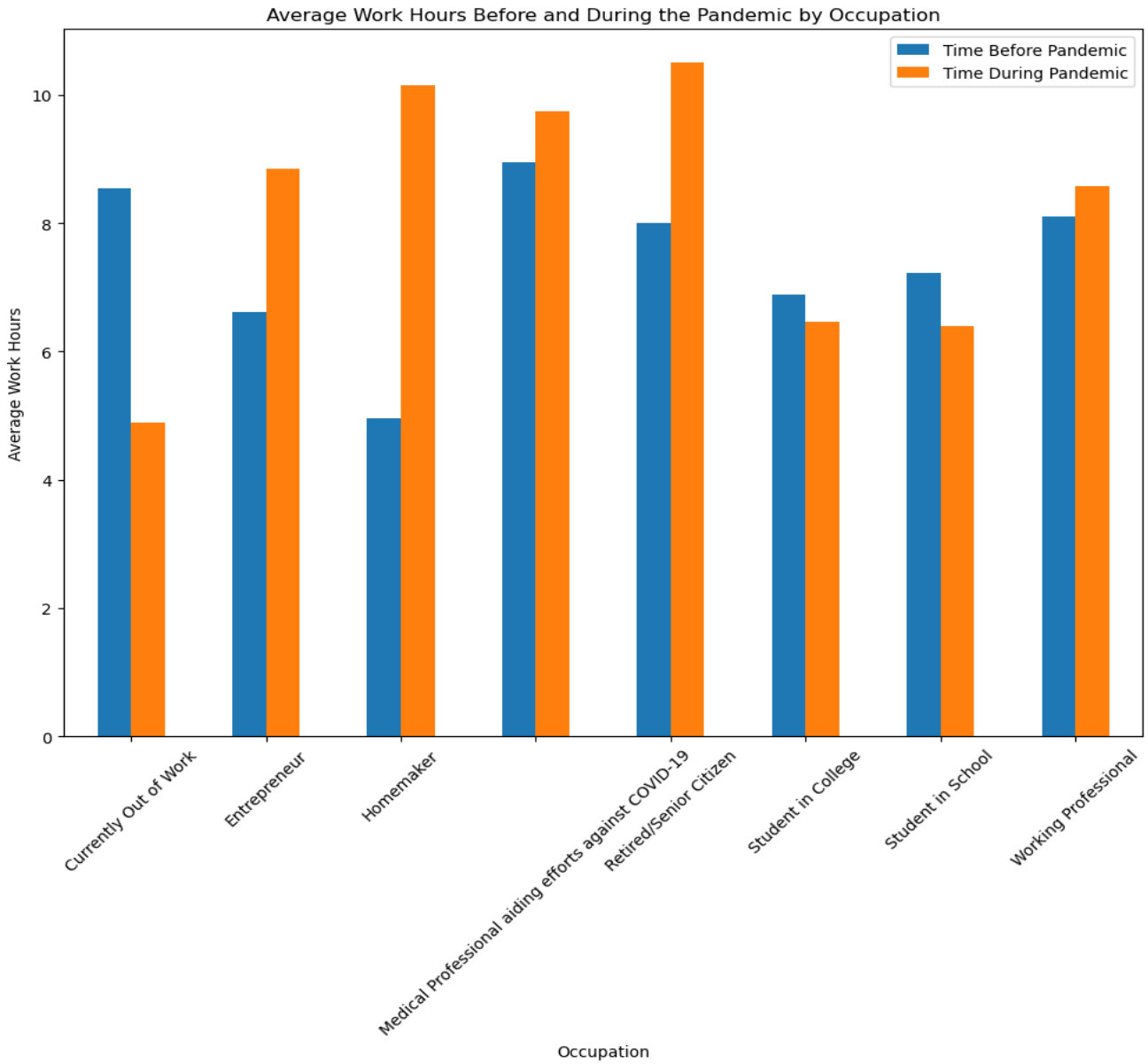
19. Certain days working from home:



Majority of the people preferred to have some days working from home while some were confused and didn't agree for the same.

B. Data Visualisations

1. **Occupation vs Time before and during pandemic at work :**

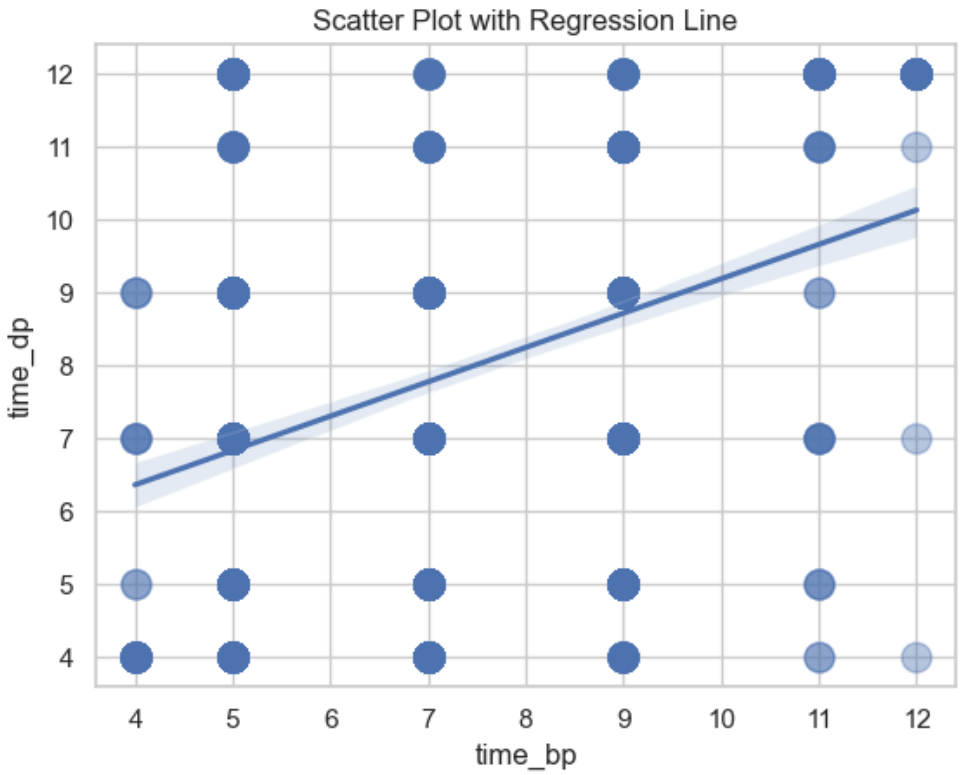


- Medical Professionals , Homemakers and working professionals have relatively higher time of work during the pandemic.
- Students in school and college show a decrease in average time during the pandemic.
- The occupation of homemakers shows a significant increase in the work during the pandemic, explaining the time spent more by people in the houses during pandemic.

2. Gender vs home environment:

- Males have a spread of scores across 3 indicating that they were not much affected by liking the home environment.
- For females the scores were more accumulated at 2 followed by 4 implying that majority of the females liked the home environment and only some didn't like as we can see that the violin chart gets smaller for values above 3.
- The people who didn't specify their gender can be seen to have a neutral approach slightly bulged toward liking the home environment as seen from the chart.

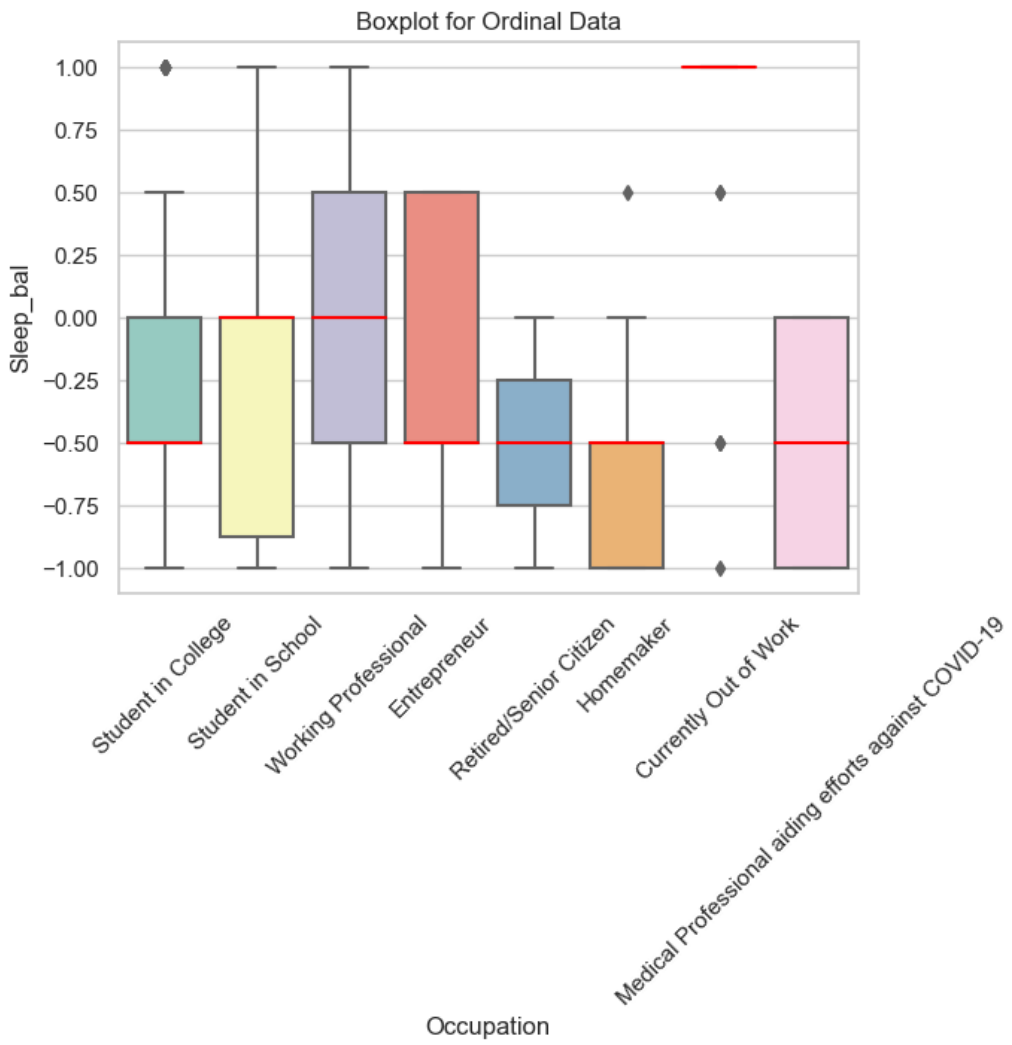
3. **Time Before pandemic vs Time During Pandemic :**
    - The diagonal values from bottom left to right in the chart appear to be more dominant implying that the significant number of individuals have the same time before the pandemic of work and time during the pandemic of work.
    - The lightness in the dots for 12 hours before pandemic and during pandemic indicates that the time has significantly reduced during pandemic.
    - The regression line shows us that there is a positive relation between time before and time during pandemic.



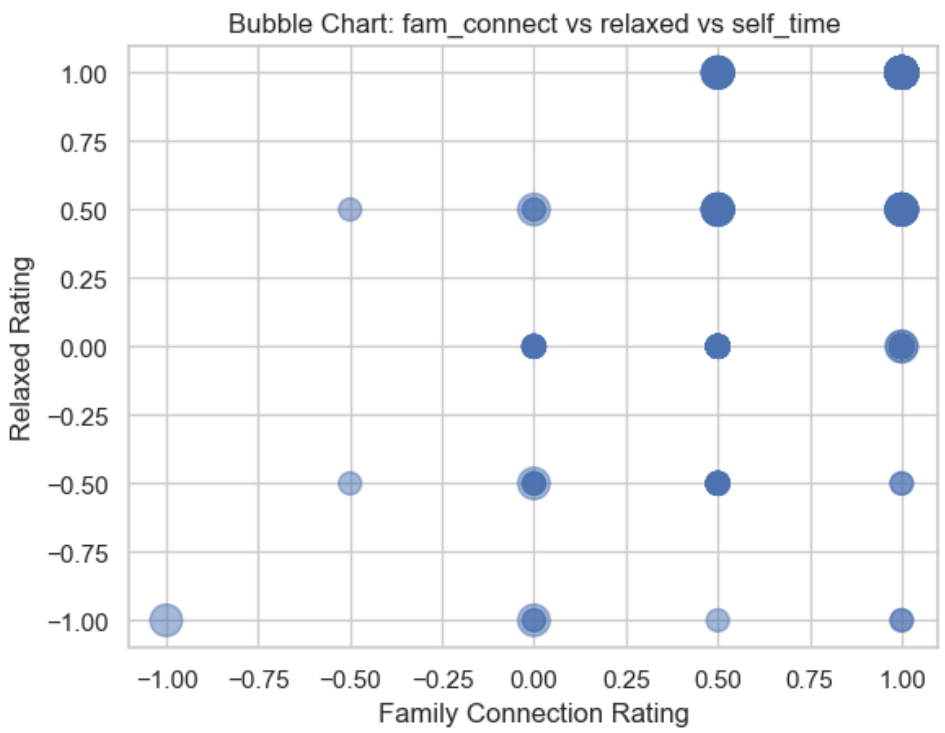
4. **Occupation vs Sleep balance :**

	Q1	**Q2**	**Q3**	**Minimum**	**Maximum**	**Outlier Interpretation**
Sleep balance vs student in college	About 25% of the respondents going to college experience a bad sleep.	About 50% of the respondents going to college experience bad sleep.	About 75% of the respondents going to college have a neutral sleep.	Very bad sleep	Good sleep	Only few college going students experience a very good sleep meaning the study pressure is too high for the students.
Sleep balance vs student in school	About 25% of the respondents going to school experience very bad sleep.	About 50% of the respondents going to school experience neutral sleep.	About 75% of the respondents going to school have a neutral sleep.	Very bad sleep	Very good sleep	
Sleep	About 25% of	About 50%	About	Very bad	Very	

balance vs working professional	the working professionals experience bad sleep.	of the working professionals experience neutral sleep.	75% of the working professionals experience good sleep.	sleep	good sleep	
Sleep balance vs entrepreneur	About 25% of the entrepreneurs experience bad sleep.	About 50% of the entrepreneurs experience bad sleep.	About 75% of the entrepreneurs experience good sleep.	Very bad sleep	Good sleep	
Sleep balance vs retired	About 25% of the retirees experience between bad and very bad sleep.	About 50% of the retirees experience bad sleep.	About 75% of the retirees experience between bad and neutral sleep.	Very bad sleep	Neutral sleep	
Sleep balance vs homemaker	About 25% of the homemakers experience very bad sleep.	About 50% of the homemakers experience bad sleep.	About 75% of the homemakers experience bad sleep.	Very bad sleep	Neutral sleep	Only a few homemakers experience a good sleep stating the intensity of a homemaker job.
Sleep balance vs medical professionals	About 25% of the medical professionals experience very bad sleep.	About 50% of the medical professionals experience bad sleep.	About 50% of the medical professionals experience neutral sleep.	Very bad sleep	Neutral sleep	



5. **Family connection vs self time vs relaxed :**
   Here the x axis represents family connection rating , y axis represents the relaxed rating and the size of the bubble is determined by the self time rating based on both axes.The size of the dots represents the value of self time in that combined category and the transparency depicts the correlation.
   - People who strongly disagreed(-1) upon no family connection and no relaxation time during the pandemic had given priority to more self time.
   - People who disagreed(-0.5) upon no family connection and no relaxation time had also disagreed on the fact of self time.
   - People who believed in having a very good(1) and good(0.5) family time and relaxation time also had a good self time in a pandemic.


Bubble Chart: fam_connect vs relaxed vs self_time
Relaxed Rating
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Family Connection Rating

Data Pre Processing

1. Handling of Missing values :

```
In [22]: data.isnull().sum()

Out[22]: 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
         dtype: int64
```

The data consisted of a total of 696 missing values in the line of work attribute.
Since line of work and occupation go hand in hand we compare both and using the combined knowledge of both we see for which occupation which line of work is missing and fill the values accordingly.
Except for the working professionals category of occupation , all other categories had missing values in respect to line of work.

```
In [35]: data.loc[data['occupation'] == 'Student in College', 'line_of_work'] = 'No Job'

In [36]: data.loc[data['occupation'] == 'Entrepreneur', 'line_of_work'] = 'Entrepreneur'

In [37]: data.loc[data['occupation'] == 'Homemaker', 'line_of_work'] = 'No Job'

In [38]: data.loc[data['occupation'] == 'Medical Professional aiding efforts against COVID-19', 'line_of_work'] = 'Medical professional'

In [39]: data.loc[data['occupation'] == 'Currently Out of Work', 'line_of_work'] = 'No Job'

In [40]: data.loc[data['occupation'] == 'Student in School', 'line_of_work'] = 'No Job'

In [41]: data.loc[data['occupation'] == 'Retired/Senior Citizen', 'line_of_work'] = 'Retired'

In [42]: data.isnull().sum()

Out[42]: age              0
         gender           0
         occupation       0
         line_of_work     0
         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
```

We replace the missing values with the respective line of work category based on occupation and now our data has no missing values.

2. **Removing Outliers :**

```
In [343]: # For gender we see the following
          data['gender'].value_counts()

Out[343]: gender
          Male                649
          Female              518
          Prefer not to say     8
          Name: count, dtype: int64
```

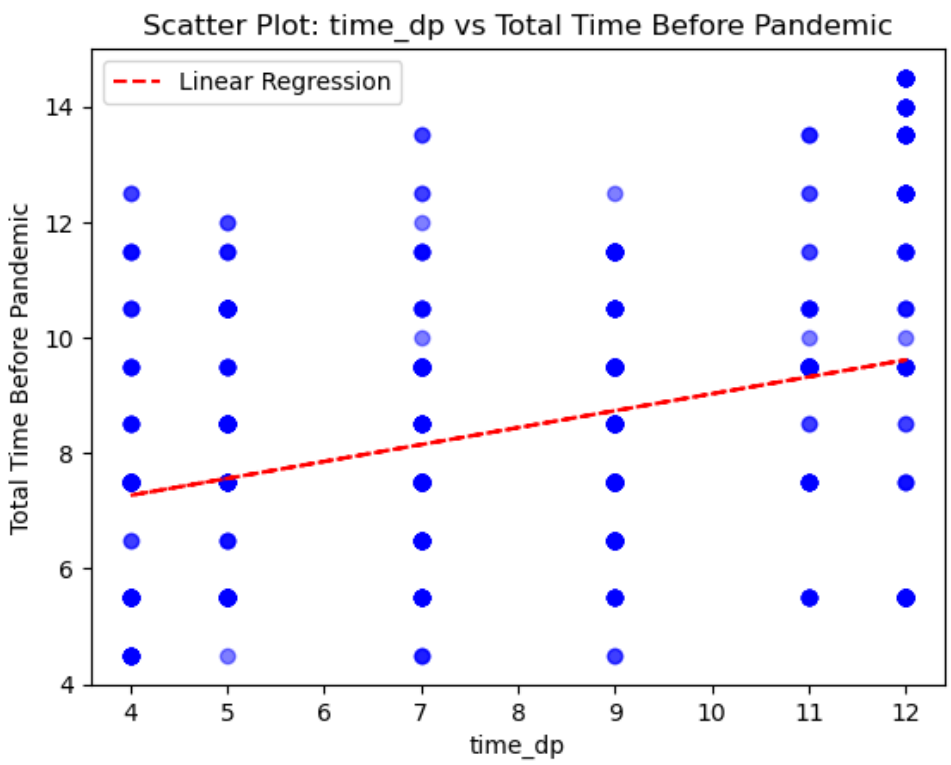
For gender attributes we see that for the prefer not to say category there are only 8 responses. In the context of data privacy individuals may have intentionally withheld their gender from revealing and hence taking out these values makes sense. Removing this category ensures that gender-related analyses are based only on explicitly provided information, reducing potential inaccuracies and misconceptions. Also these few observations may lead to data inconsistency in further calculations and it is best to drop these values.

```
In [45]: # there are only 8 responses other than male and female. This can increase t
         # Hence we drop these values
         data.drop(data[data['gender'] == 'Prefer not to say'].index, inplace=True)

In [46]: data.shape

Out[46]: (1167, 19)
```

3. **Addition of new attribute :**
The attributes of time before and during the pandemic give us a distinguished idea about how the time of work changed because of the pandemic. But there is another big factor that played an important role before the pandemic and that was the travel time involved in going to school,college or work.Hence we create a new column with attribute 'total_time_before_pandemic' and compare it with time during pandemic to get a proper idea to what extent travel contributes in work life.

```
In [47]: data['total_time_before_pandemic'] = data['time_bp'] + data['travel_time']
```

- time_dp vs total_time_before_pandemic:



The above diagram shows us that time during pandemic and total time before pandemic including travel time have a positive correlation. There has not been a much difference in the work time due to pandemic but yes for some specific hours like people who worked 9 hours during pandemic had to work in total of more than 12 hours before pandemic.Travel time was an essential factor as it could make huge difference in specific occupations and line of works.

4. Normalisation :

In [616]:
```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler(feature_range=(0, 1))

# List of columns to transform
columns_to_transform = ['prod_inc', 'sleep_bal', 'new_skill', 'fam_connect', 'relaxed', 'self_time']

# Apply MinMax scaling to each specified column
for column in columns_to_transform:
    data2[column + '_num'] = scaler.fit_transform(data2[[column]])

# Drop the original columns
data2.drop(columns=columns_to_transform, inplace=True)
```

For normalisation we normalise the ordinal attributes having a likert scale between -1 to 1.This is because the negative values may skew the data and hence we convert them using min max scaler method into the range of 0 to 1.

We do not convert the other ordinal attributes like ease of online and home environment because they are already in a numerical form. The attributes like_hw and dislike_hw can be just transformed into their binary forms.

```
mapping10 = {1:'0', 10: '1', 11: '2', 100: '3', 101: '4',110:'5',111:'6',1000:'7',1001:'8',1010:'9',1011:'10',1100:'11',1101:'12'
data2['like_hw'] = data2['like_hw'].replace(mapping10)
```

```
mapping12 = {1:'0', 10: '1', 11: '2', 100: '3', 101: '4',110:'5',111:'6',1000:'7',1001:'8',1010:'9',1011:'10',1100:'11',1101:'12'
data2['dislike_hw'] = data2['dislike_hw'].replace(mapping12)
```

Classification :

Target Attribute : Gender

For every model the train test was split into half to avoid any unbiased skewness or result interpretations.
No sample set from the original dataset was taken as the length of the dataset was already not that big.

Correlation analysis for classification:

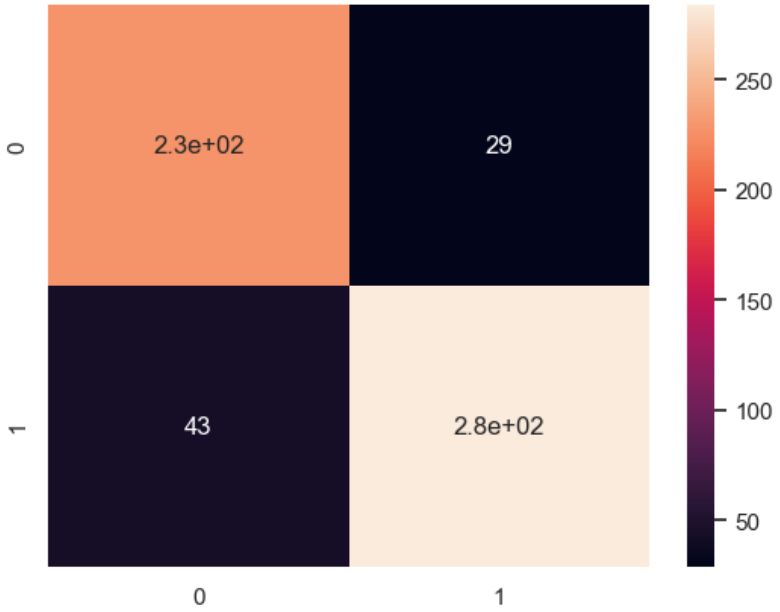
The target variable for classification is gender and except it there are 19 more variables. It might not be totally efficient for a model to use all the 19 attributes as there may be many variables having low correlation with target attribute and hence affecting the accuracy. Hence we perform correlation using chi square test and Cramer's V test and find the best 10 correlated attributes to the target.
On performing the test we get the following as our top 10 highly correlated attributes with gender: line_of_work , occupation , age , total_time_before_pandemic , time_dp , new_skill_num , dislike_hw , like_hw , home_env and prod_inc_num.
We use these 10 attributes for the models to have better accuracy and precision and the model to be more accurate to use in real life.

```python
import pandas as pd
from scipy.stats import chi2_contingency
from sklearn.preprocessing import LabelEncoder

categorical_variable = data_corr['gender']

all_variables = ['age','occupation','line_of_work','time_bp','time_dp','travel_time','easeof_online','home_env','like_hw','dislik
                 'prod_inc_num','sleep_bal_num','new_skill_num','fam_connect_num','relaxed_num','self_time_num']

label_encoder = LabelEncoder()
for variable in all_variables:
    if data_corr[variable].dtype == 'O':
        data_corr[variable] = label_encoder.fit_transform(data_corr[variable])

association_results = []


for variable in all_variables:
    # Create a contingency table
    contingency_table = pd.crosstab(categorical_variable, data_corr[variable])

    # Perform chi-square test
    chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

    # Calculate Cramér's V as a measure of association
    n = contingency_table.sum().sum()
    cramers_v = (chi2_stat / (n * (min(contingency_table.shape) - 1)))**0.5

    association_results.append({
        'Variable': variable,
        'Chi2 Statistic': chi2_stat,
        'P-Value': p_value,
        'Cramér\'s V': cramers_v
    })

association_df = pd.DataFrame(association_results)
print(association_df.sort_values(by='Cramér\'s V', ascending=False))
```

	Variable	Chi2 Statistic	P-Value	Cramér's V
2	line_of_work	490.709725	3.068756e-98	0.648451
1	occupation	345.875965	9.406441e-71	0.544408
0	age	310.662479	4.242082e-64	0.515952
12	total_time_before_pandemic	243.911990	1.174764e-44	0.457174
4	time_dp	216.866106	6.969851e-45	0.431083
15	new_skill_num	180.683399	5.318234e-38	0.393481
9	dislike_hw	166.761569	3.085022e-28	0.378018
8	like_hw	141.576318	3.446584e-23	0.348305
7	home_env	128.893616	6.714218e-27	0.332338
13	prod_inc_num	124.784211	5.075607e-26	0.326997
17	relaxed_num	99.316056	1.375437e-20	0.291725
3	time_bp	71.300844	5.493143e-14	0.247179
10	prefer	59.657532	1.128864e-14	0.226098
14	sleep_bal_num	43.547330	7.966184e-09	0.193173
6	easeof_online	40.457867	3.480278e-08	0.186194
18	self_time_num	34.466868	5.976998e-07	0.171856
16	fam_connect_num	23.518001	9.975798e-05	0.141960
5	travel_time	21.370200	8.819036e-05	0.135322
11	certaindays_hw	1.455814	4.829188e-01	0.035320

Decision Tree :

Train Test model:
Confusion Matrix :



[[228 29]
 [43 284]]

True Positive	True Negative	False Positive (Type 1 Error)	False Negative (Type 2 Error)
284 The instances which were actually positive and the model predicted positive which is a high number and good result.	228 The instances which were actually negative and the model predicted negative which is a high number and good result.	29 The type 1 error value count is low and good for the model.	43 The type 2 error value count is low and good for the model.

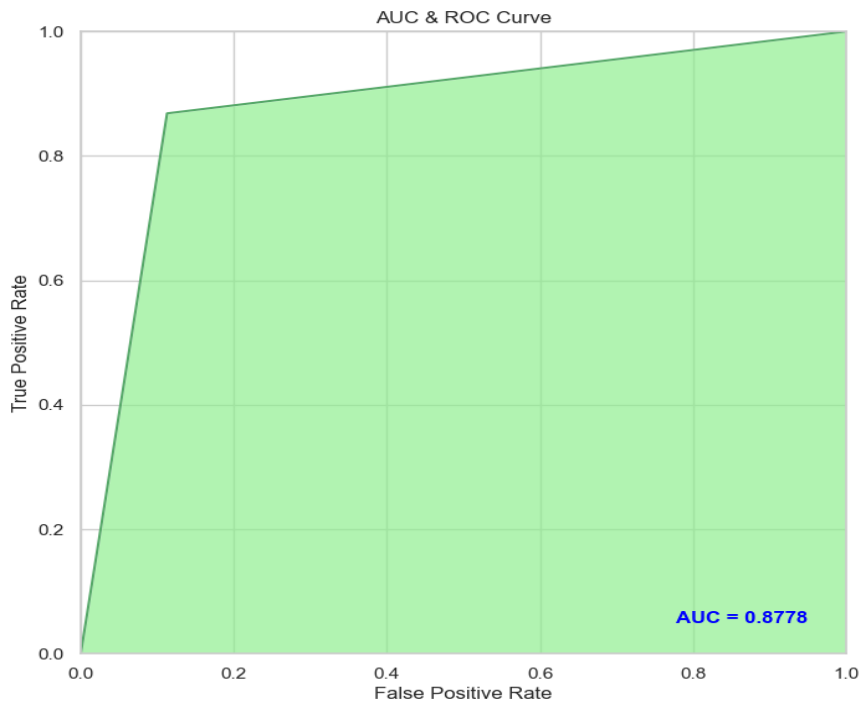
Result Interpretation :

```
[[228  29]
 [ 43 284]]

Mean Absolute Error: 0.1232876712328767
Mean Squared Error: 0.1232876712328767
Root Mean Squared Error: 0.3511234415883917

              precision    recall  f1-score   support

           0       0.84      0.89      0.86       257
           1       0.91      0.87      0.89       327

    accuracy                           0.88       584
   macro avg       0.87      0.88      0.88       584
weighted avg       0.88      0.88      0.88       584

Accuracy: 0.8767123287671232
```

1. The precision values are 84% and 91% indicating that the model has not labelled a positive sample as a negative 84% of times and vice versa 91% of times which is very good for the model.
2. The recall value for positive is 89% implying a high true positive rate and for the negative it is 87% implying a high true negative rate.This values tell us about the sensitivity of the model and the correctness of the model and in this case the model seems to perform good for both positive and negative values.
3. The f1 score being the harmonic mean of precision and recall and having high percentage indicates that the model is a good fit to the data.
4. Accuracy shows the overall correctness of the model and in this case it is 87% which is a good accuracy.
5. In summary the confusion matrix indicates that the model is performing well with high precision,recall and f1 scores.The relatively low number of false positives and false negatives suggests that the model is making good predictions for both classes.
6. This model is a good fit and can be used in practice.

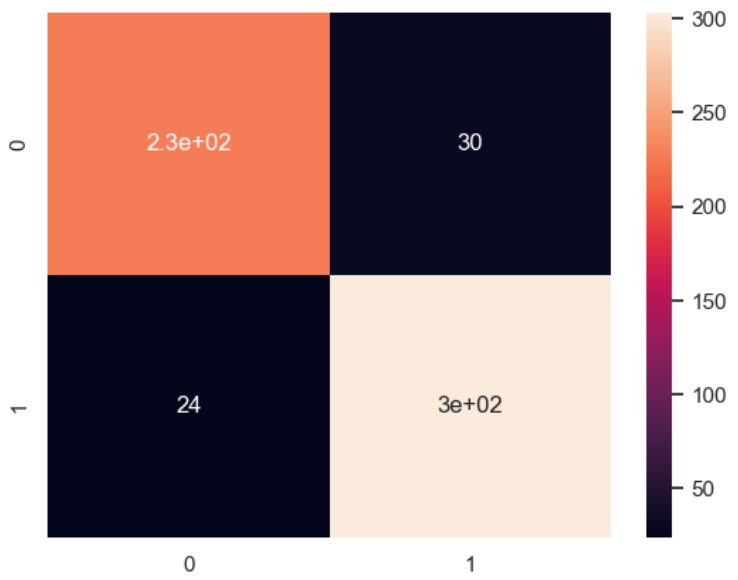
ROC AUC Curve :

Performance Category	**AUC Score**	**Interpretation**
0.8 to 0.9: Excellent discrimination	0.8778	The score is relatively high suggesting that the model has good discriminatory power. It performs well in distinguishing between the positive and negative values across different thresholds.



K Nearest Neighbours :

Train Test Model:
Confusion Matrix :



[[227 30]
 [24 303]]

True Positive	True Negative	False Positive (Type 1 Error)	False Negative (Type 2 Error)
303 The instances which were actually positive and the model predicted positive which is a high number and good result.	227 The instances which were actually negative and the model predicted negative which is a high number and good result.	30 The type 1 error value count is low and good for the model.	24 The type 2 error value count is low and good for the model.

Result Interpretation :

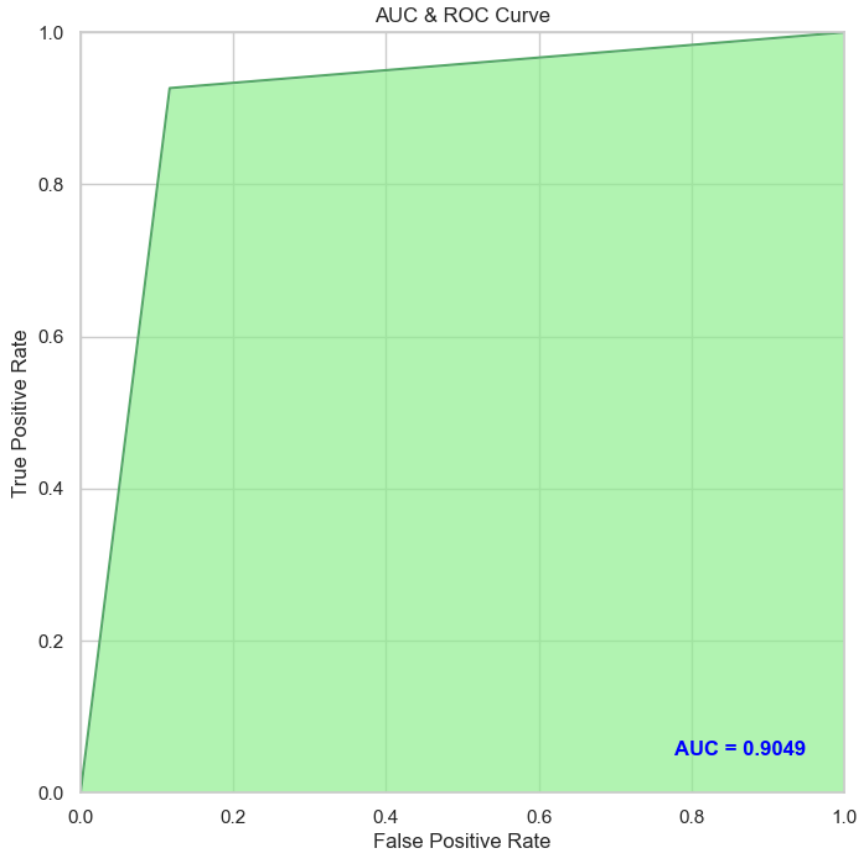
```
[[227  30]
 [ 24 303]]

Mean Absolute Error: 0.09246575342465753
Mean Squared Error: 0.09246575342465753
Root Mean Squared Error: 0.30408182027976866

              precision    recall  f1-score   support

           0       0.90      0.88      0.89       257
           1       0.91      0.93      0.92       327

    accuracy                           0.91       584
   macro avg       0.91      0.90      0.91       584
weighted avg       0.91      0.91      0.91       584

Accuracy: 0.9075342465753424
```

1. The precision values are 90% and 91% indicating that the model has not labelled a positive sample as a negative 90% of times and vice versa 91% of times which is excellent for the model output and is not an excellent result.
2. The recall value for positive is 88% implying a high true positive rate and for the negative it is 93% implying a high true negative rate.This values tell us about the sensitivity of the model and the correctness of the model and in this case the model seems to perform very good for both positive and negative values.
3. The f1 score being the harmonic mean of precision and recall and having high percentage indicates that the model is a good fit to the data.
4. Accuracy shows the overall correctness of the model and in this case it is 90.7% which is a very good accuracy.
5. In summary the confusion matrix indicates that the model is performing well with high precision,recall and f1 scores.The relatively low number of false positives and false negatives suggests that the model is making good predictions for both classes.

6. This model is a good fit and can be used in practice.

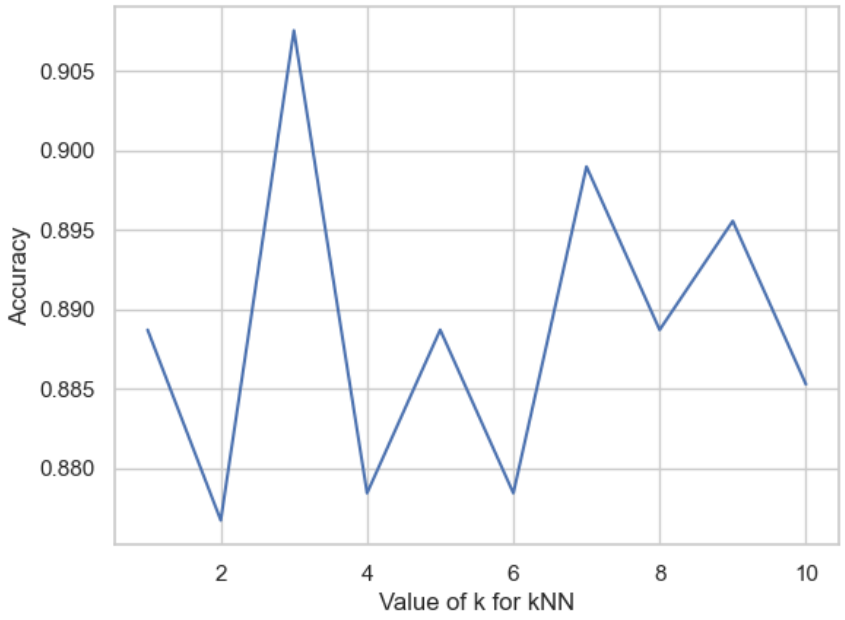
ROC AUC Curve :

Performance Category	**AUC Score**	**Interpretation**
0.9 to 1: Outstanding discrimination	0.9049	The score is relatively very high suggesting that the model has good discriminatory power. It performs well in distinguishing between the positive and negative values across different thresholds.



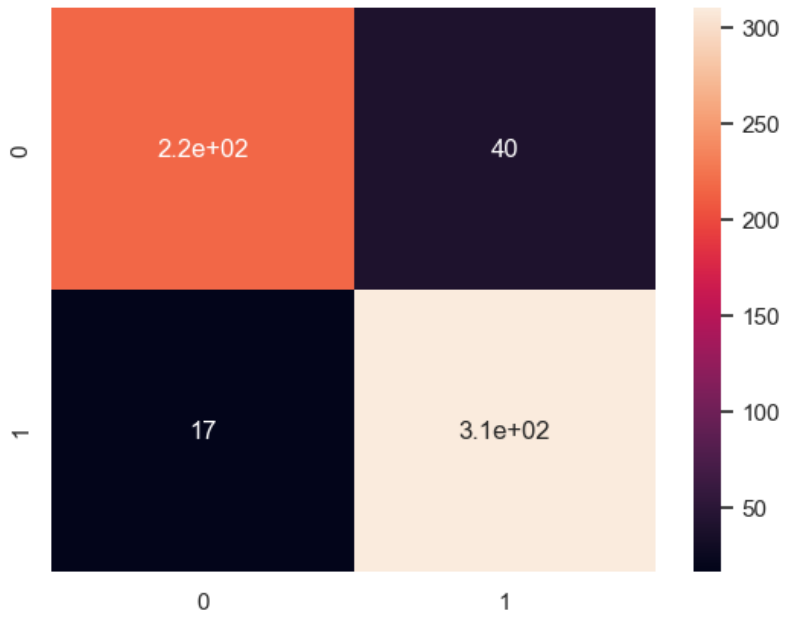
Models Performance for different values of K :

The model best performs when the value of n neighbours is set to 3 and we get the best accuracy. Testing for different values of k from 1 to 10 the second best model is given at k equals 7 followed by k equals 9. To use the model in practice we can use any of the top 3 values of k giving higher accuracy and better confusion matrix.



Adaptive Boosting (Ada Boost) :

Train Test Model :
Confusion matrix:



[[217 40]
 [17 310]]

True Positive	True Negative	False Positive (Type 1 Error)	False Negative (Type 2 Error)
310 The instances which were actually positive and the model predicted positive which is a high number and good result.	217 The instances which were actually negative and the model predicted negative which is a high number and good result.	40 The type 1 error value count is low and good for the model.	17 The type 2 error value count is low and good for the model.

Result Interpretation :

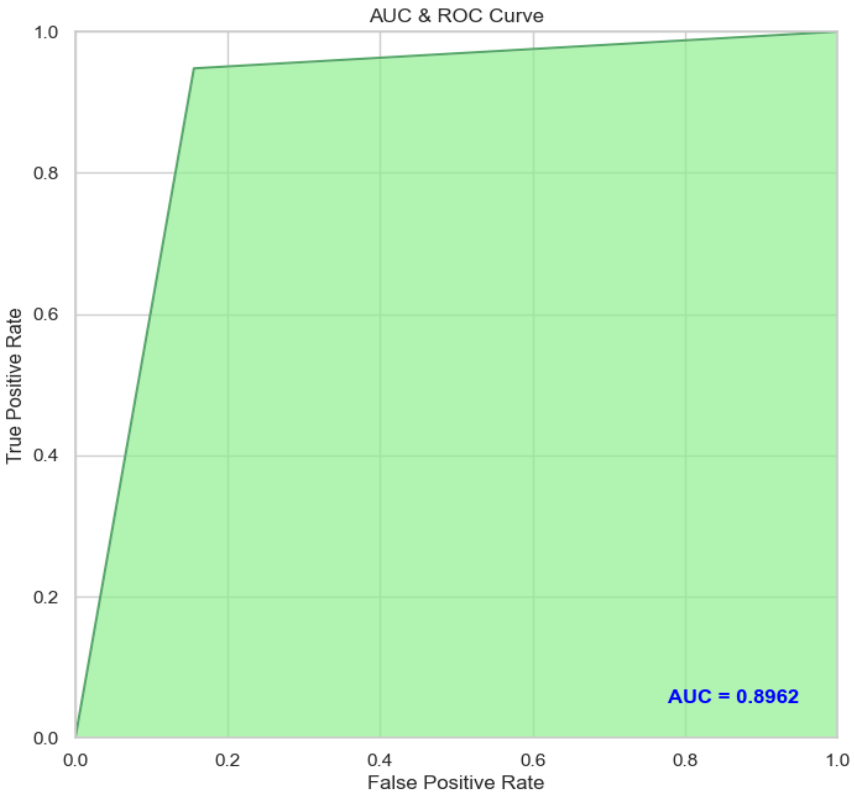
```
[[217  40]
 [ 17 310]]

Mean Absolute Error: 0.0976027397260274
Mean Squared Error: 0.0976027397260274
Root Mean Squared Error: 0.3124143718301503

              precision    recall  f1-score   support

           0       0.93      0.84      0.88       257
           1       0.89      0.95      0.92       327

    accuracy                           0.90       584
   macro avg       0.91      0.90      0.90       584
weighted avg       0.90      0.90      0.90       584

Accuracy: 0.9023972602739726
```

1. The precision values are 93% and 89% indicating that the model has not labelled a positive sample as a negative 93% of times and vice versa 89% of times which is very good for the model output and is a good result.
2. The recall value for positive is 84% implying a good true positive rate and for the negative it is 95% implying a high true negative rate.This values tell us about the sensitivity of the model and the correctness of the model and in this case the model seems to perform very good for both positive and negative values.
3. The f1 score being the harmonic mean of precision and recall and having high percentage indicates that the model is a good fit to the data.

4. Accuracy shows the overall correctness of the model and in this case it is 90.239% which is a very good accuracy.
5. In summary the confusion matrix indicates that the model is performing well with high precision,recall and f1 scores.The relatively low number of false positives and false negatives suggests that the model is making good predictions for both classes.
6. This model is a good fit and can be used in practice.

ROC AUC Curve :

Performance Category	**AUC Score**	**Interpretation**
0.8 to 0.9: Excellent discrimination	0.8962	The score is relatively high suggesting that the model has good discriminatory power. It performs well in distinguishing between the positive and negative values across different thresholds.



Comparison of the three classification models :

Algorithm Name	Algorithm type	Precision	Recall	F1 score	Accuracy	AUC Score
Decision Tree	Train/Test	> 84%	> 87%	> 86%	87.67%	0.8778
K - Nearest Neighbour	Train/Test	> 90%	> 88%	> 89%	90.75%	0.9049
Ada Boost	Train/Test	> 89%	> 84%	> 88%	90.23%	0.8962

- All three algorithms function well with accuracy scores lying in ranges of 87.67% to 90.75%.
- K- Nearest Neighbours(KNN) performs the best overall with highest accuracy and AUC score indicating that a KNN model with 3 neighbours is best fitted for this dataset.
- Ada boost also does well and closes the range of accuracy with KNN by a very small margin.
- The overall precision,recall and f1 scores of all three models being high indicates that all three can be used in practice.
- However the choice of selecting which algorithm to use can depend on various other factors like algorithm type , interpretability , computational efficiency and the involvement of attributes based on correlation analysis and hence based on that the decision of which model to be used should be made , in this case we use KNN with k = 3.

Regression

Target attribute : self_time

For the model the train test was split into half to avoid any unbiased skewness or result interpretations.
No sample set from the original dataset was taken as the length of the dataset was already not that big.

Correlation analysis for Regression :

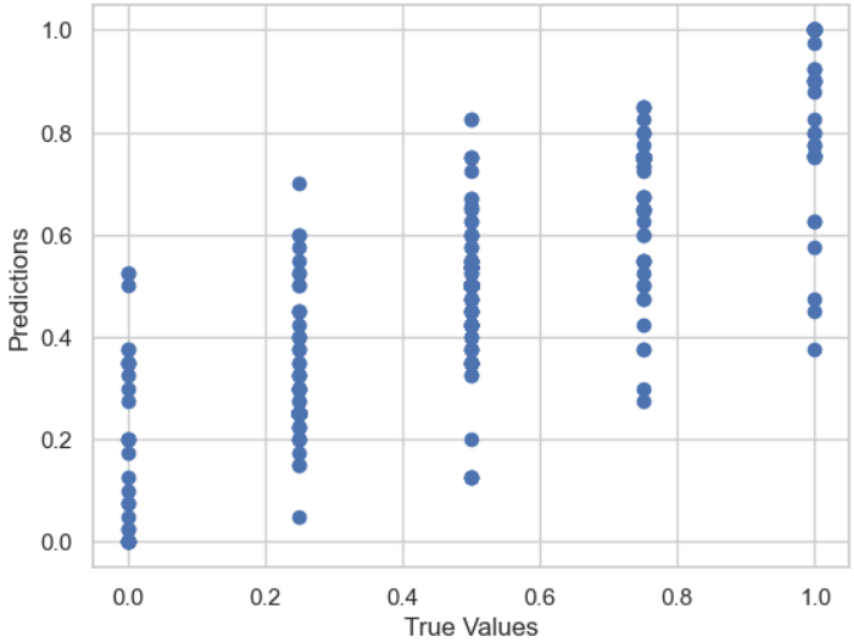
We perform correlation analysis on the self_time attribute with respect to other attributes so that we can take into account only those attributes that are highly correlated with the target attribute to give us a better regression result.On performing the test we get the following as our top 10 highly correlated attributes with self_time: relaxed_num, occupation , age , total_time_before_pandemic , time_dp , new_skill_num , sleep_balance_num , fam_connect_num, like_hw and prod_inc_num.
We use these 10 attributes for the models to have better accuracy and precision and the model to be more accurate to use in real life.

```python
import pandas as pd
from scipy.stats import chi2_contingency
from sklearn.preprocessing import LabelEncoder

categorical_variable = data_corr['self_time_num']

all_variables = ['age','occupation','line_of_work','time_bp','time_dp','travel_time','easeof_online','home_env','like_hw','dislik
                 'prod_inc_num','sleep_bal_num','new_skill_num','fam_connect_num','relaxed_num','gender']

label_encoder = LabelEncoder()
for variable in all_variables:
    if data_corr[variable].dtype == 'O':
        data_corr[variable] = label_encoder.fit_transform(data_corr[variable])

association_results = []


for variable in all_variables:
    # Create a contingency table
    contingency_table = pd.crosstab(categorical_variable, data_corr[variable])

    # Perform chi-square test
    chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

    # Calculate Cramér's V as a measure of association
    n = contingency_table.sum().sum()
    cramers_v = (chi2_stat / (n * (min(contingency_table.shape) - 1)))**0.5

    association_results.append({
        'Variable': variable,
        'Chi2 Statistic': chi2_stat,
        'P-Value': p_value,
        'Cramér\'s V': cramers_v
    })

association_df = pd.DataFrame(association_results)
print(association_df.sort_values(by='Cramér\'s V', ascending=False))
```

	Variable	Chi2 Statistic	P-Value	Cramér's V
17	relaxed_num	1392.923287	5.404828e-287	0.546258
15	new_skill_num	1146.278542	5.006052e-234	0.495541
16	fam_connect_num	1004.253459	1.375560e-203	0.463827
14	sleep_bal_num	983.290489	4.231854e-199	0.458961
12	total_time_before_pandemic	836.429899	5.528383e-142	0.423301
0	age	817.258766	4.675248e-157	0.418422
1	occupation	695.483206	1.723034e-128	0.385992
4	time_dp	654.853257	7.740185e-126	0.374547
13	prod_inc_num	632.885543	2.398664e-124	0.368211
8	like_hw	576.609261	1.630928e-87	0.351460
9	dislike_hw	555.818903	1.986402e-83	0.345065
2	line_of_work	528.268499	2.976829e-84	0.336405
3	time_bp	485.884717	2.615556e-90	0.322627
11	certaindays_hw	181.205907	5.746068e-35	0.278635
10	prefer	79.174278	2.605627e-16	0.260469
6	easeof_online	279.283524	4.888718e-50	0.244600
7	home_env	213.063770	1.790334e-36	0.213643
5	travel_time	149.237779	8.094272e-26	0.206464
18	gender	34.466868	5.976998e-07	0.171856

Random Forest Regression

```
Mean Absolute Error: 0.05524326985799589
Mean Squared Error: 0.014650016179320888
Root Mean Squared Error: 0.12103725120524213

R2 score: 0.8015617809263884
Adjusted R2 score: 0.7980986357418577
```

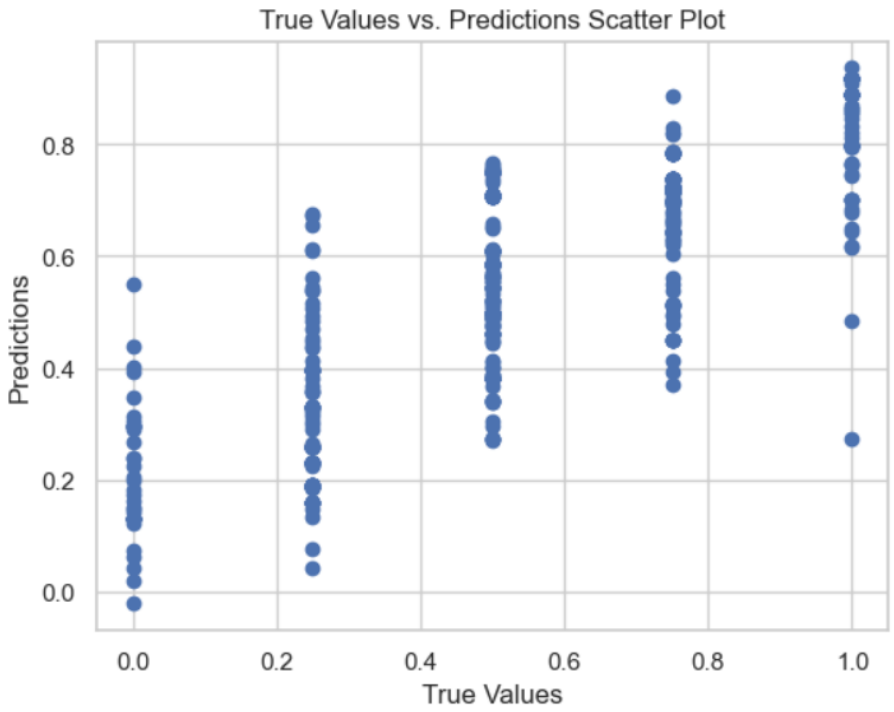


Result Interpretation :

1. The R2 score of 80.15% indicates a good fit of the model as it explains over 80 % of the variance in the dependent variable is because of the independent variable.
2. The adjusted R2 score being high is also a good sign and it also explains the complexity of the model. The model has adjusted the R2 score for the different predictors in the model and given a good value.
3. The MAE , MSE and RMSE are also quite low, explaining that the model's predicted values are close to the actual ones.
4. The graph shows the scatter plot between the predicted values by the model and the actual values. The graph has pretty much the same clusters of dots except few points lying above or below the line of clusters of the points. This tells us that the model has quite efficiently predicted the true values.

Ridge Regression

```
Mean Absolute Error: 0.11652874073282923
Mean Squared Error: 0.024073913729415437
Root Mean Squared Error: 0.15515770599430578

R2 score: 0.6739126764009885
Adjusted R2 score: 0.6682217981531873
```



Result Interpretation :

1. The R2 score of 67.39% indicates a good fit of the model as it explains over two-thirds of the variance in the dependent variable.
2. The adjusted R2 score being high is also a good sign and it also explains the complexity of the model. The model has adjusted the R2 score for the different predictors in the model and given a good value.
3. The MAE , MSE and RMSE are also quite low, explaining that the model's predicted values are close to the actual ones.
4. The graph shows the scatter plot between the predicted values by the model and the actual values. The graph has pretty much the same clusters of dots except few points lying above or below the line of clusters of the points. This tells us that the model has quite efficiently predicted the true values.
5. The regularisation strength is kept as 1 for penalties.

Comparison of the two Regression models :

Algorithm Name	Algorithm type	R2 score	Adjusted R2 score	MAE	MSE	RMSE
Random Forest Regression	Train/Test	80.15%	79.80%	0.0552	0.0146	0.121
Ridge Regression	Train/Test	67.39%	66.82%	0.116	0.024	0.155

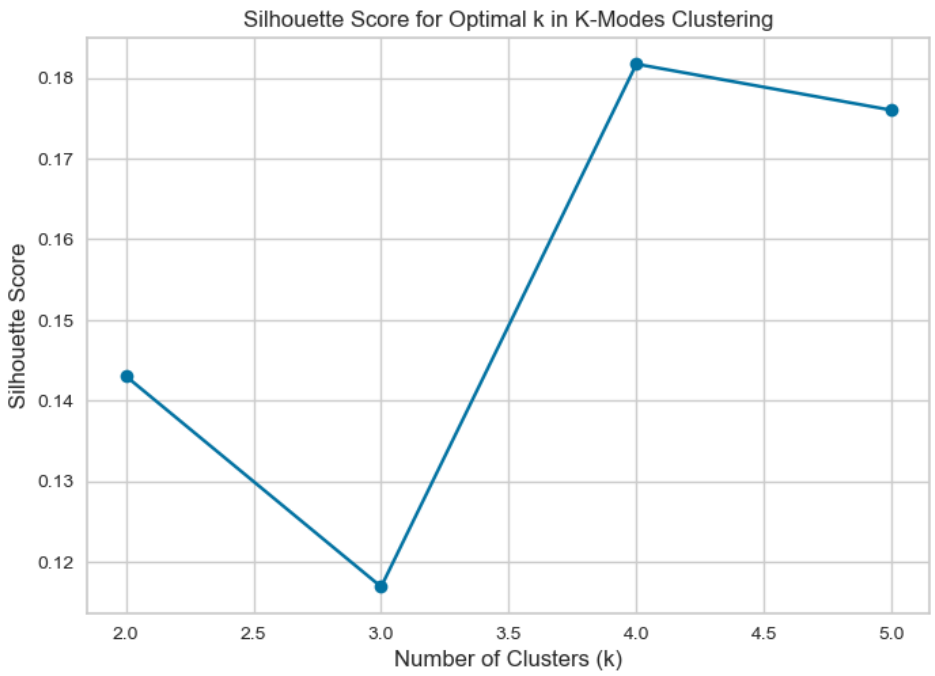
- The random forest regression model performs better in all the aspects and overall can be said to be a good fit to the dataset and can be applied in practice.
- The difference between the R2 scores of both the models makes it clear to go for the random forest model as the margin of difference is very high.
- The error terms also contribute and hence comparing them we see that the random forest model was able to predict more true values.
- The ridge model can vary if the regularisation term changes and more or less penalties are applied.
- Hence in conclusion both the models performed well but random forest can be used over the ridge model for this dataset.

Clustering

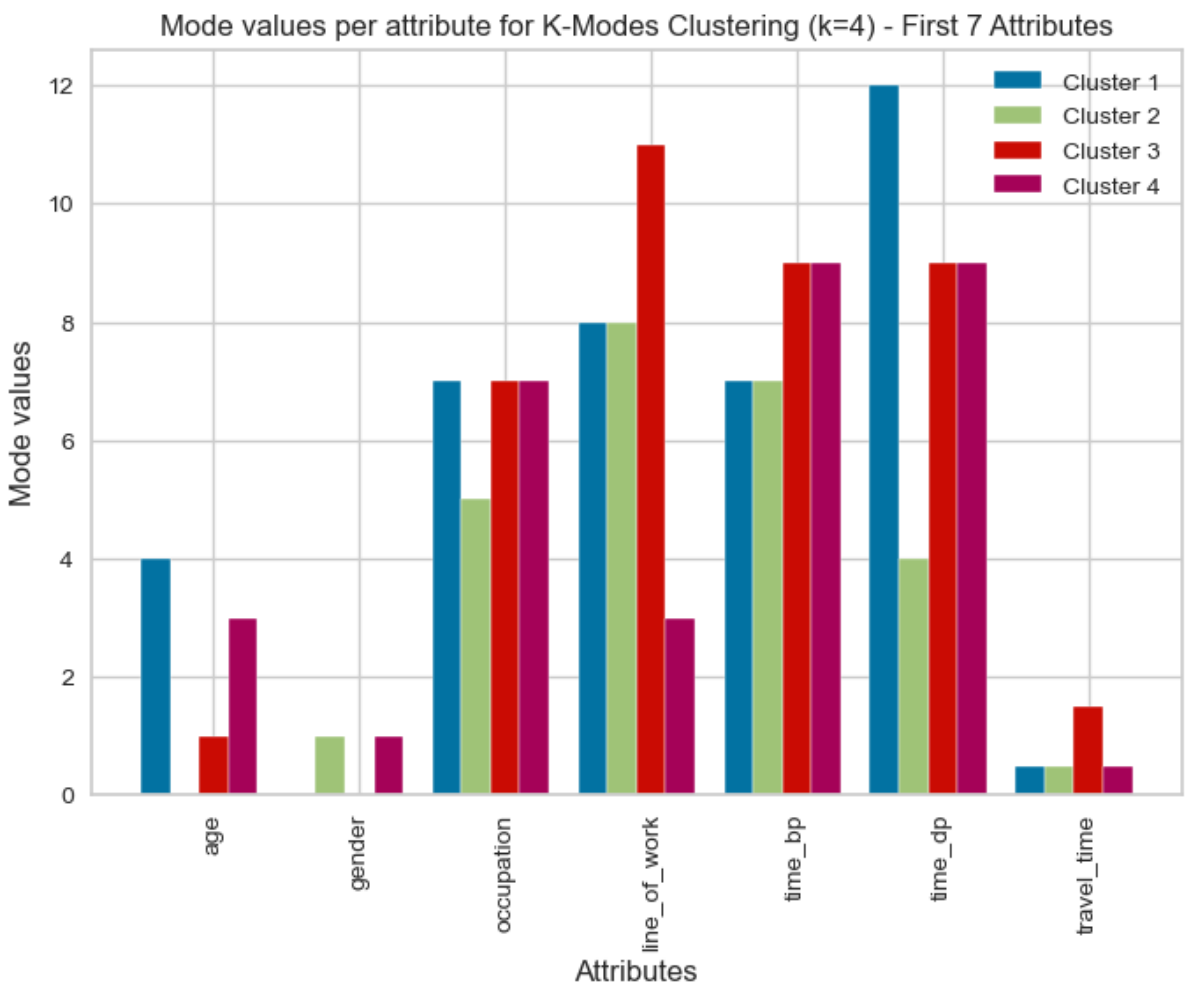
K- Modes

After inputting the data for clustering first we transform the categorical attributes into numerical ones. We do not normalise the data again as it has been already taken care of in the data pre processing.

After all the basic steps we perform the silhouette analysis to find the optimal number of clusters.

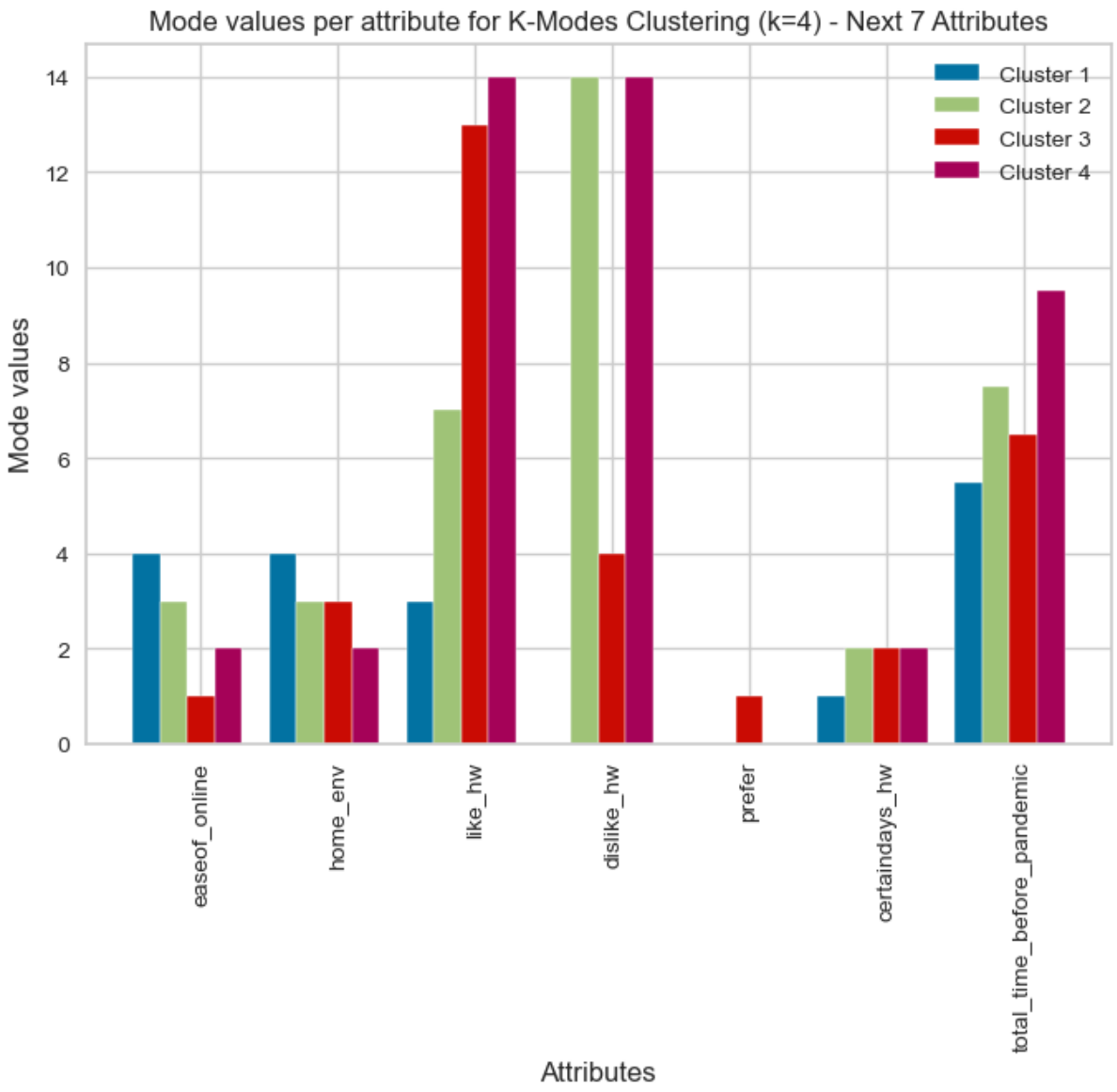


On performing the analysis we conclude that the optimal number of clusters will be 4 and hence we use 4 clusters.

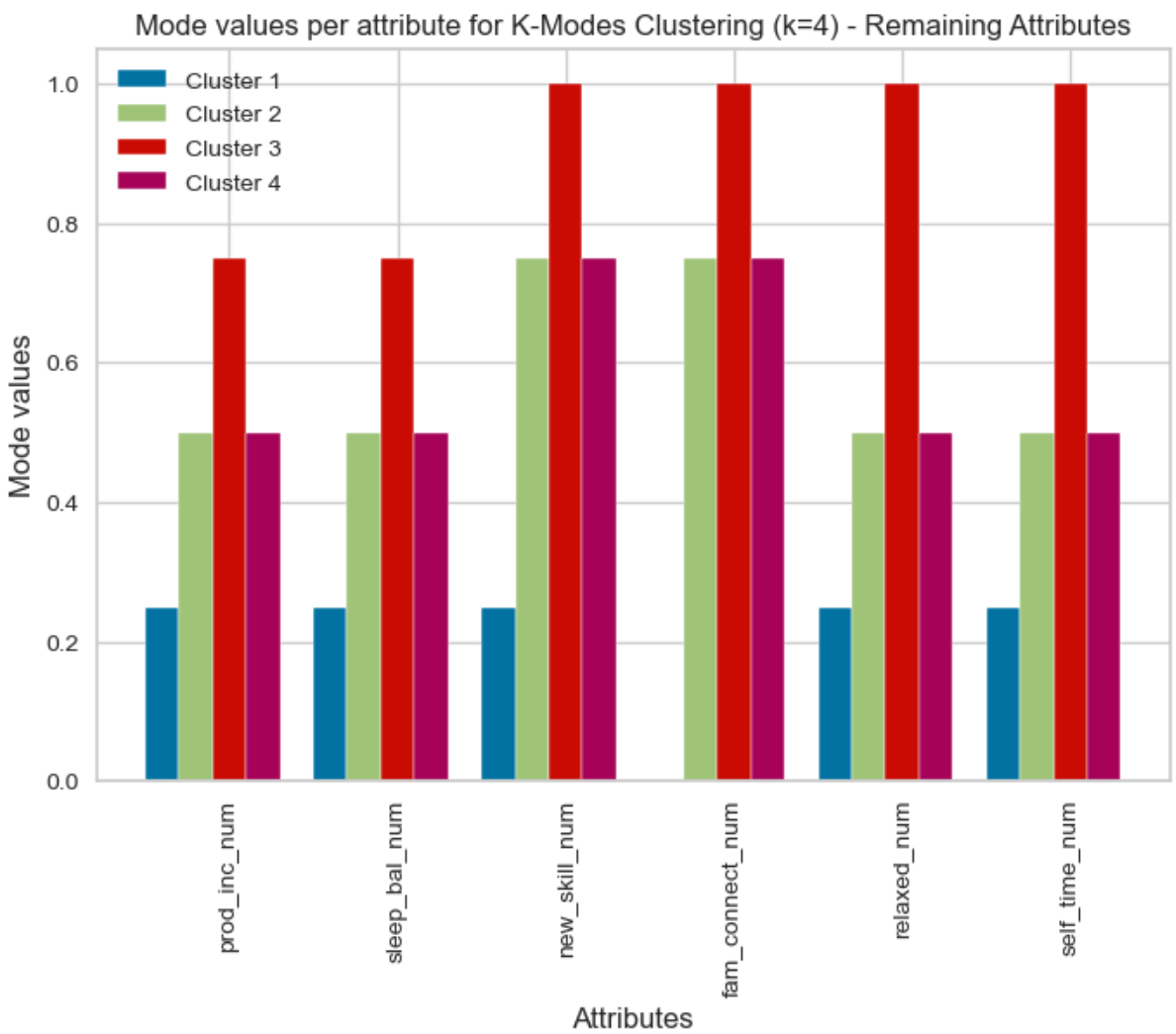


Attribute	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Age	This cluster dominates among all clusters and consists of age groups between 50-60.	No Cluster	This is the smallest cluster for this category with age groups between 26-32	The second largest cluster with age groups between 40-50
Gender	No Cluster	Largest cluster with Male as gender.	No cluster	Largest Cluster with Male as gender.
Occupation	Largest cluster for the category with working professionals as occupation.	Smallest cluster for the category with students in college as occupation.	Largest cluster for the category with working professionals as occupation.	Largest cluster for the category with working professionals as occupation.
line_of_work	Second Largest cluster for the attribute with No jobs as the line of work.	Second Largest cluster for the attribute with No jobs as the line of work.	Largest cluster for the attribute with Teaching as the line of work.	The Smallest cluster among all clusters with engineering as line of work.

time_bp	Smallest cluster for the attribute with 7 hours as time spent before pandemic.	Smallest cluster for the attribute with 7 hours as time spent before pandemic.	Largest cluster for the attribute with 9 hours as time spent before pandemic.	Largest cluster for the attribute with 9 hours as time spent before pandemic.
time_dp	Largest cluster among all with 12 hours as time spent during pandemic.	Smallest cluster among all with 4 hours spent during pandemic.	Second largest cluster with 9 hours spent during pandemic.	Second largest cluster with 9 hours spent during pandemic.
travel_time	Smallest cluster for the attribute with 0.5 hours as travel time.	Smallest cluster for the attribute with 0.5 hours as travel time.	Largest cluster for the attribute with 1.5 hours as travel time.	Smallest cluster for the attribute with 0.5 hours as travel time.



Attribute	Cluster 1	Cluster 2	Cluster 3	Cluster 4
easeof_online	Largest Cluster with a rating of 4 for ease of working online.	Second largest Cluster with a rating of 3 for ease of working online.	Smallest cluster among all with a rating of 1 for ease of working online.	Third largest cluster with a rationing of 2 for ease of working online
home_env	Largest cluster for attribute with a rating of 4 for liking the home environment.	Second largest cluster for attributes with a rating of 3 for liking the home environment.	Second largest cluster for attributes with a rating of 3 for liking the home environment.	Smallest cluster among all with a rating of 2 towards liking of home environment.
like_hw	Smallest cluster with a rating of 3 towards liking working from home.	Third largest cluster with a rating of 7 towards liking working from home.	Second largest cluster with a rating of 13 towards liking working from home.	Largest cluster with a rating of 14 towards liking working from home.
dislike_hw	No cluster	Largest cluster with a rating of 14 towards disliking working from home	Second largest cluster with a rating of 4 towards disliking working from home	Largest cluster with a rating of 14 towards disliking working from home
prefer	No cluster	No cluster	Only cluster for the attribute with preferring to work/study from home.	No cluster
certaindays_hw	Smallest cluster with not wanting to have certain days of work from home.	Largest cluster with wanting to have certain days of work from home.	Largest cluster with wanting to have certain days of work from home.	Largest cluster with wanting to have certain days of work from home.
total_time_before_pandemic	Smallest cluster with 5.5 total hours working before pandemic including travel.	Second largest cluster with 7.5 total hours working before pandemic including travel.	Third largest cluster with 6.5 total hours working before pandemic including travel.	Largest cluster with 9.5 total hours working before pandemic including travel.



Attribute	Cluster 1	Cluster 2	Cluster 3	Cluster 4
prod_inc_num	Smallest Cluster with bad rating of productivity increase.	Second largest Cluster with neutral rating of productivity increase.	Largest Cluster with good rating of productivity increase.	Second largest Cluster with neutral rating of productivity increase.
sleep_bal_num	Smallest Cluster with bad rating for sleep cycle.	Second largest Cluster with neutral rating for sleep cycle.	Largest Cluster with good rating for sleep cycle.	Second largest Cluster with neutral rating for sleep cycle.
new_skill_num	Smallest Cluster with bad rating for a new skill learned.	Second largest Cluster with good rating for a new skill learned.	Largest Cluster with Very good rating for a new skill learned.	Second largest Cluster with good rating for a new skill learned.
fam_connect_num	No cluster	Second largest Cluster with good rating for family connection.	Largest Cluster with Very good rating for family connection.	Second largest Cluster with good rating for family connection.
relaxed_num	Smallest Cluster with bad rating for feeling relaxed.	Second largest Cluster with neutral rating for feeling relaxed.	Largest Cluster with Very good rating for feeling relaxed.	Second largest Cluster with neutral rating for feeling relaxed.

self_time _num	Smallest Cluster with a bad rating for self time spent.	Second largest Cluster with neutral rating for self time spent.	Largest Cluster with Very good rating for self time spent.	Second largest Cluster with neutral rating for self time spent.

Cluster analysis on the various sizes of clusters :

From the sizes of the clusters we could see that the third cluster in most cases had the largest size indicating that the data points in that cluster share a similar value for that attribute.This indicates a high level of homogeneity within the cluster for that attribute. The first cluster seemed to have most small clusters with various attributes implying heterogeneity within the cluster. Most of the clusters based on their sizes seem to be significant but in some cases we have some overlapping clusters. The good homogeneity of clusters and less heterogeneity of clusters make them a set of good clusters for result interpretations.

Cluster Interpretation :

Cluster 1 :

Cluster 1 consisted of the people in the age group 50-60 who were working professionals with no job as occupation and line of work. Their total time of work before pandemic was 7 hours and during pandemic was 12 hours which shows a clear shift in the increase of work and job opportunities during pandemic.These people liked the ease of work from online and the home environment and rated it a 4 out of 5 . These cluster people said no to working certain days from home. The other factors such as productivity increase , sleep cycle , new skills learned , family connection , relaxation time and self time were rated as bad by these categories of people.

Cluster 2 :

Cluster 2 consisted of Males who were Students in college with no job as occupation and line of work. Their total time of work before pandemic was 7 hours and during pandemic was 4 hours which shows a clear shift in the decrease of work for college students as the classes were short and less hectic during pandemic.These people had a neutral response for liking the ease of work from online and the home environment and rated it a 3 out of 5 . These cluster people said yes to working certain days from home. The other factors such as productivity increase , sleep cycle,self time and relaxation time were rated neutral by them and new skills learned and family connection were rated as Good by these categories of people. These cluster people rated a high disliking for working from home.

Cluster 3 :

Cluster 3 consisted of the people in the age group 26-32 who were working professionals with teaching as occupation and line of work. Their total time of work before pandemic was 9 hours and during pandemic was 9 hours which shows that the teachers had no difference in working time before and during pandemic.These people didn't like the ease of work from

online and rated it 1 out of 5 and the home environment was rated 3 out of 5 . These cluster people said yes to working certain days from home. The other factors such as productivity increase and sleep cycle were rated good and new skills learned , family connection , relaxation time and self time were rated as Very good by these categories of people. These cluster people preferred and rated high liking towards working from home.

Cluster 4 :

Cluster 4 consisted of the people in the age group 40-50 who were male working professionals with Engineering as occupation and line of work. Their total time of work before the pandemic was 9 hours and during pandemic was 9 hours which shows that the time of working for the engineers was not highly affected by pandemic. These people didn' like the ease of work from online and the home environment that much and rated it a 2 out of 5 . These cluster people said yes to working certain days from home. The other factors such as productivity increase , sleep cycle , relaxation time and self time were rated as neutral and new skills learned and family connection were rated as good by these categories of people.

Description and Referencing for new algorithms and charts :

1. **Kernel density plot :**
   This plot provides a smooth explanation of the variations in the numerical variables and clearly shows the ups and downs of the dataset.

2. **Violin plot:**
   The violin plot is a combination of a box plot and a kernel density plot. It gives the smoothness and the variation across different parameters like Q1,Q2 and Q3 and explains which of them are affecting the most.

3. **Scatter plot with regression line:**
   It is a combination of scatter plot and regression line. This chart shows the connection between two attributes and the regression line tells us whether the connection is positive or negative.

4. **Bubble Chart:**
   This chart shows the three dimensional scatter along the x axis, the y axis and the bubble size and the bubble density shows the impact of the third attribute.Each bubble displays a data point and its size can be used to convey additional information.

5. **Adaptive Boosting (Ada Boost):**
   Ada boost is an ensemble method of classification. How this algorithm works is that it first trains a weak learner i.e. decision tree and it evaluates it and gives more weights to the misclassified data points.In next iteration the weights are increased and a new weak learner is created.this process is repeated for a number of times and each weak learner is given a weight on the basis of their performance.The final model is the weighted sum of the weak learners and the weights are based on individual performances.
   Advantages:
   It is less prone to overfitting.
   It can be used with various base learners.

6. **Ridge Regression:**
   In ridge regression the regularisation term penalises large coefficients discouraging or setting the model off from heavily relying on any single predictor. This proves to be beneficial as all the attributes involved are taken into account.the balance between training the data and keeping the model simple is maintained.
   Advantages:
   Prevents overfitting
   Handels multicollinearity
   Improves model generalisation

7. **K Modes Clustering:**
   K modes clustering in specific works best for categorical data. But the data adjustments techniques make it suitable for the numerical data as well. We take the mode values and hence no normalisation is necessary in this case.It uses dissimilarity matrix unlike the euclidean distance.

References

Kernel density estimation and its application
Stanisław Węglarczyk ITM Web Conf., 23 (2018) 00037 Published online: 2018-11-07
DOI: https://doi.org/10.1051/itmconf/20182300037
https://seaborn.pydata.org/generated/seaborn.kdeplot.html
Kernel density estimation and its application

Han S, Kwak IY. Mastering data visualisation with Python: practical tips for researchers. J Minim Invasive Surg. 2023 Dec 15;26(4):167-175. doi: 10.7602/jmis.2023.26.4.167. PMID: 38098348; PMCID: PMC10728683.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10728683/
https://seaborn.pydata.org/generated/seaborn.violinplot.html

Marena Manley, Torbjörn Lestander
First published: 22 September 2003 https://doi.org/10.1002/cem.814
https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.814
https://python-graph-gallery.com/scatterplot-with-regression-fit-in-matplotlib/

Fortino, V., Alenius, H. & Greco, D. BACA: bubble chArt to compare annotations. BMC Bioinformatics 16, 37 (2015). https://doi.org/10.1186/s12859-015-0477-4
https://python-graph-gallery.com/bubble-plot/

[1] Y. Freund, and R. Shapire, "A decision-theoretic generalization of on-line learning and an application to boosting", Proceedings of the Second European Conference on Computational Learning Theory, 1995, pp. 23–37. [2] Paul A. Viola, Michael J. Jones, "Robust Real-Time Face Detection", ICCV 2001, Vol. 2, pp. 747. [3] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, 2001.
https://www.cmi.ac.in/~madhavan/courses/dmml2022/literature/rojas-adaboost.pdf
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html#:~:text=An%20AdaBoost%20%5B1%5D%20classifier%20is,focus%20more%20on%20difficult%20cases.

Hoerl AE. Application of ridge analysis to regression problems. Chem Eng Prog 1962, 58: 54–59.
First published: 13 July 2009 https://doi.org/10.1002/wics.14
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

Chaturvedi, A., Green, P. & Caroll, J. K-modes Clustering. J. of Classification 18, 35–55 (2001). https://doi.org/10.1007/s00357-001-0004-3
https://www.geeksforgeeks.org/k-mode-clustering-in-python/