final PROJECT

李嘉玲　林家輝

LIST OF CONTENTS

01 Introduction

02 Data Preprocessing

03 Data Visualization

04 Correlation Analysis

05 Model Building

06 Performance Analysis

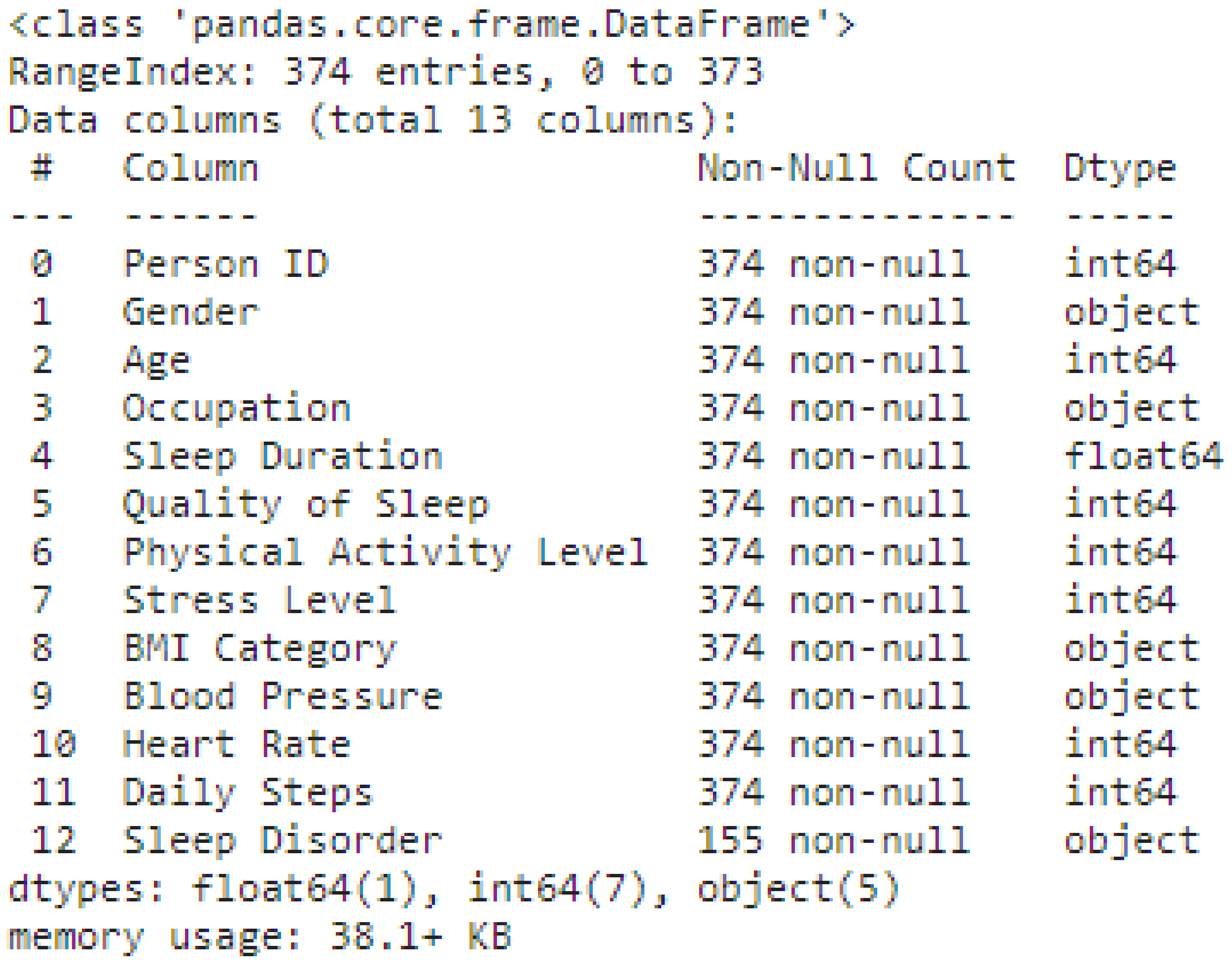
07 Conclusion

introduction

the overview of this final project and the dataset that we used through the project

Sleep Health and Lifestyle Dataset

Kaggle



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Person ID                374 non-null    int64  
 1   Gender                   374 non-null    object 
 2   Age                      374 non-null    int64  
 3   Occupation               374 non-null    object 
 4   Sleep Duration           374 non-null    float64
 5   Quality of Sleep         374 non-null    int64  
 6   Physical Activity Level  374 non-null    int64  
 7   Stress Level             374 non-null    int64  
 8   BMI Category             374 non-null    object 
 9   Blood Pressure           374 non-null    object 
 10  Heart Rate               374 non-null    int64  
 11  Daily Steps              374 non-null    int64  
 12  Sleep Disorder           155 non-null    object 
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
```

13

columns

5

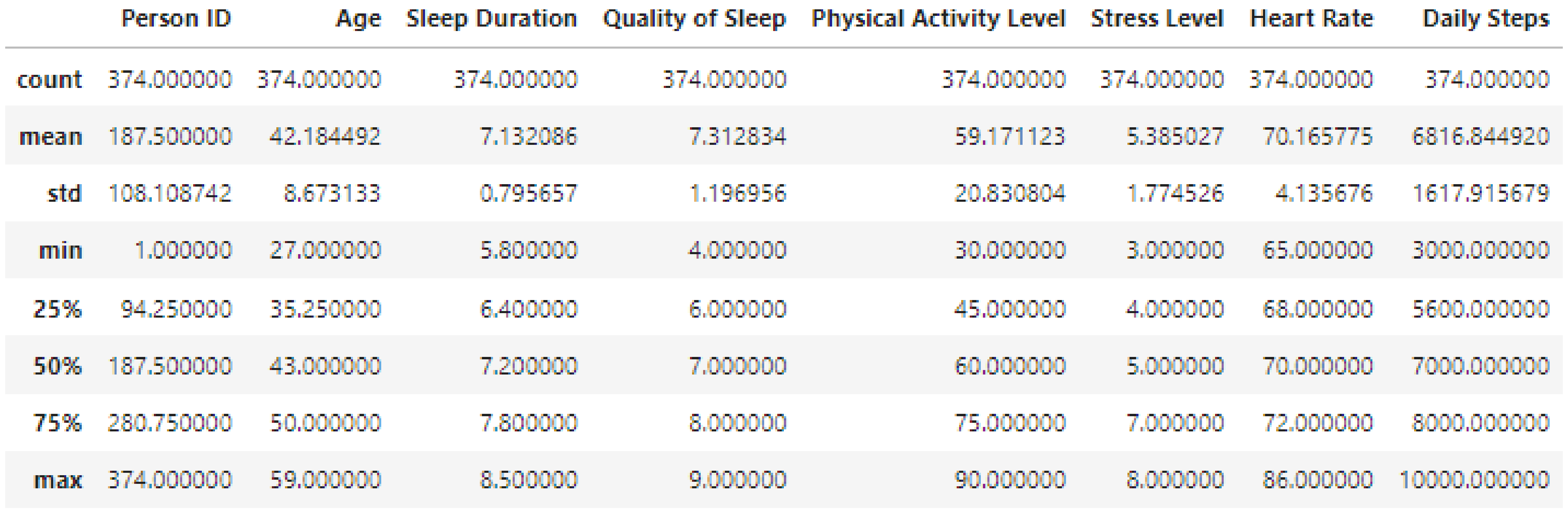
categorical

features

374

values

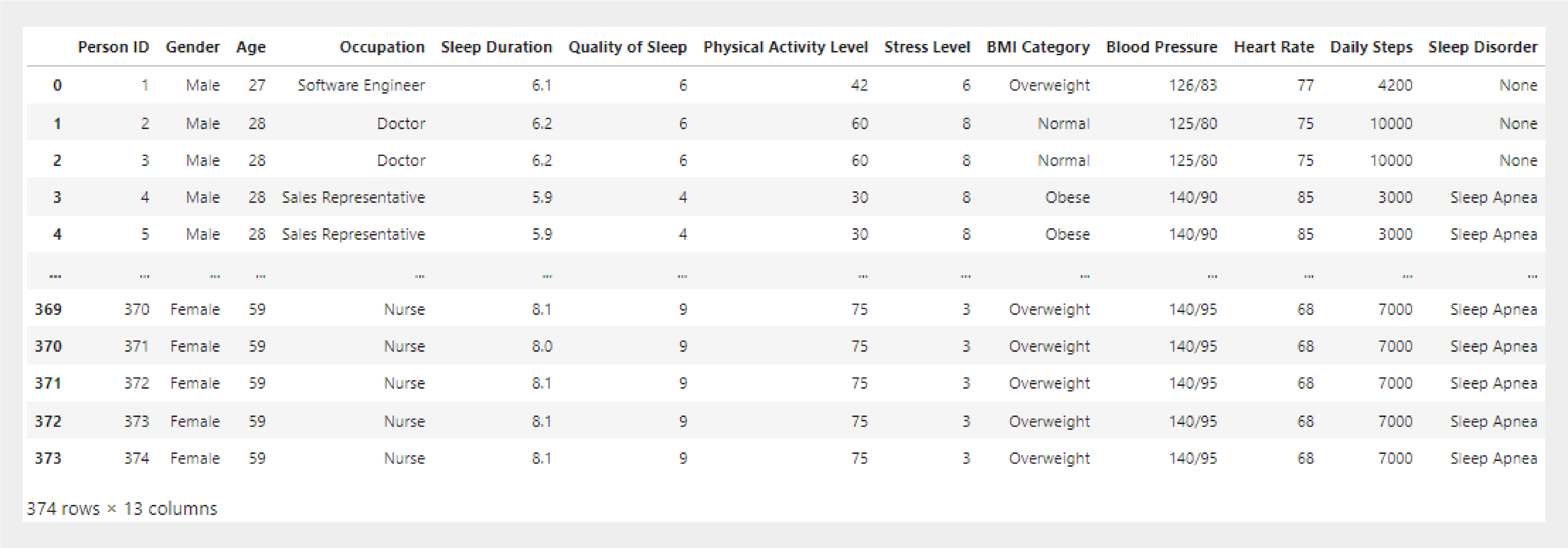
- duplicate value
- unique value
- BMI Category:
  - Overweight
  - Normal
  - Obese
  - Normal Weight
- Sleep Disorder
  - None
  - Sleep Apnea*
  - Insomnia**
- missing value



	Person ID	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate	Daily Steps
count	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000
mean	187.500000	42.184492	7.132086	7.312834	59.171123	5.385027	70.165775	6816.844920
std	108.108742	8.673133	0.795657	1.196956	20.830804	1.774526	4.135676	1617.915679
min	1.000000	27.000000	5.800000	4.000000	30.000000	3.000000	65.000000	3000.000000
25%	94.250000	35.250000	6.400000	6.000000	45.000000	4.000000	68.000000	5600.000000
50%	187.500000	43.000000	7.200000	7.000000	60.000000	5.000000	70.000000	7000.000000
75%	280.750000	50.000000	7.800000	8.000000	75.000000	7.000000	72.000000	8000.000000
max	374.000000	59.000000	8.500000	9.000000	90.000000	8.000000	86.000000	10000.000000

*Sleep Apnea: breathing stops and starts while sleeping

**Insomnia: habitual sleeplessness; inability to sleep

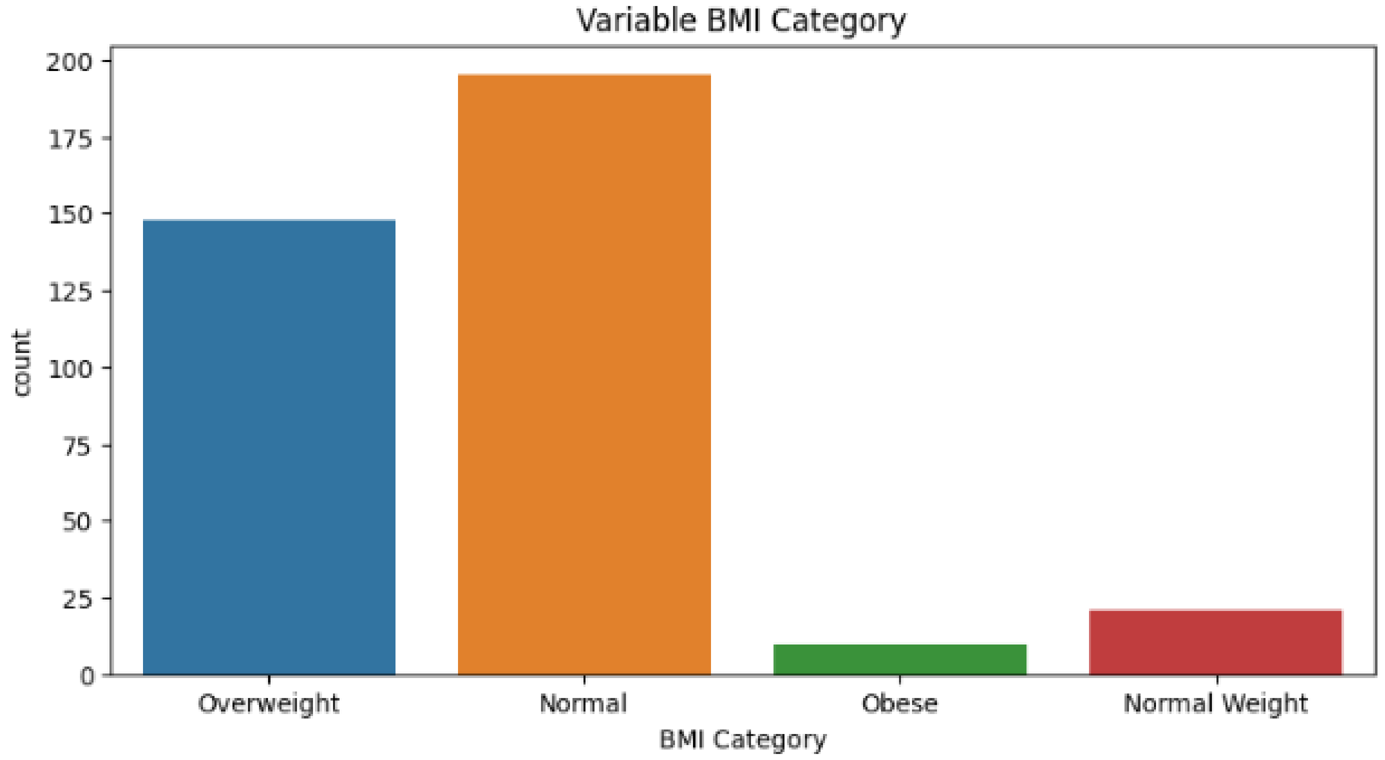


	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
...	...	...	...	...	...	...	...	...	...	...	...	...	...
369	370	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
370	371	Female	59	Nurse	8.0	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
371	372	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
372	373	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
373	374	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea

374 rows × 13 columns

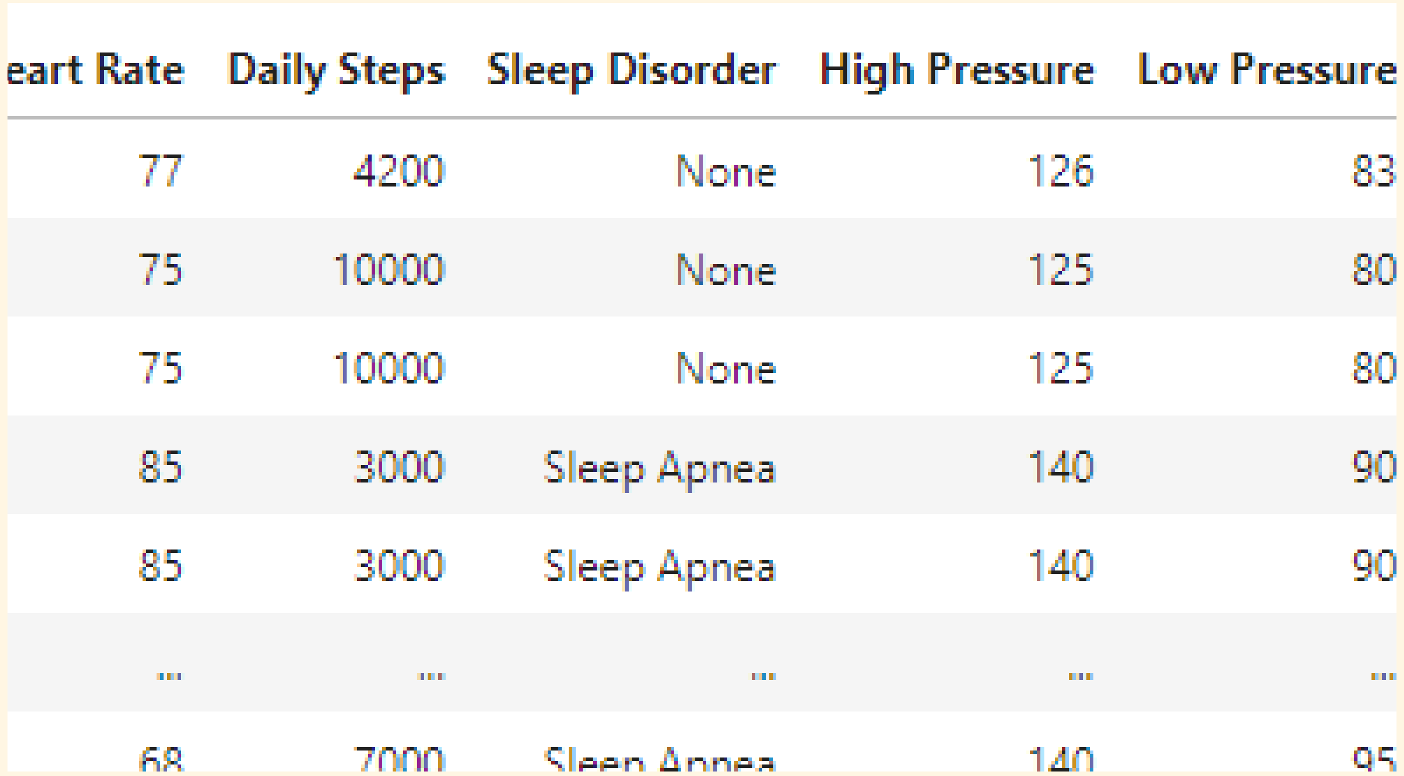
DATA *preprocessing*

01 HANDLING INCONSISTENT CATEGORICAL DATA



Normal Weight == Normal
Obese == Overweight

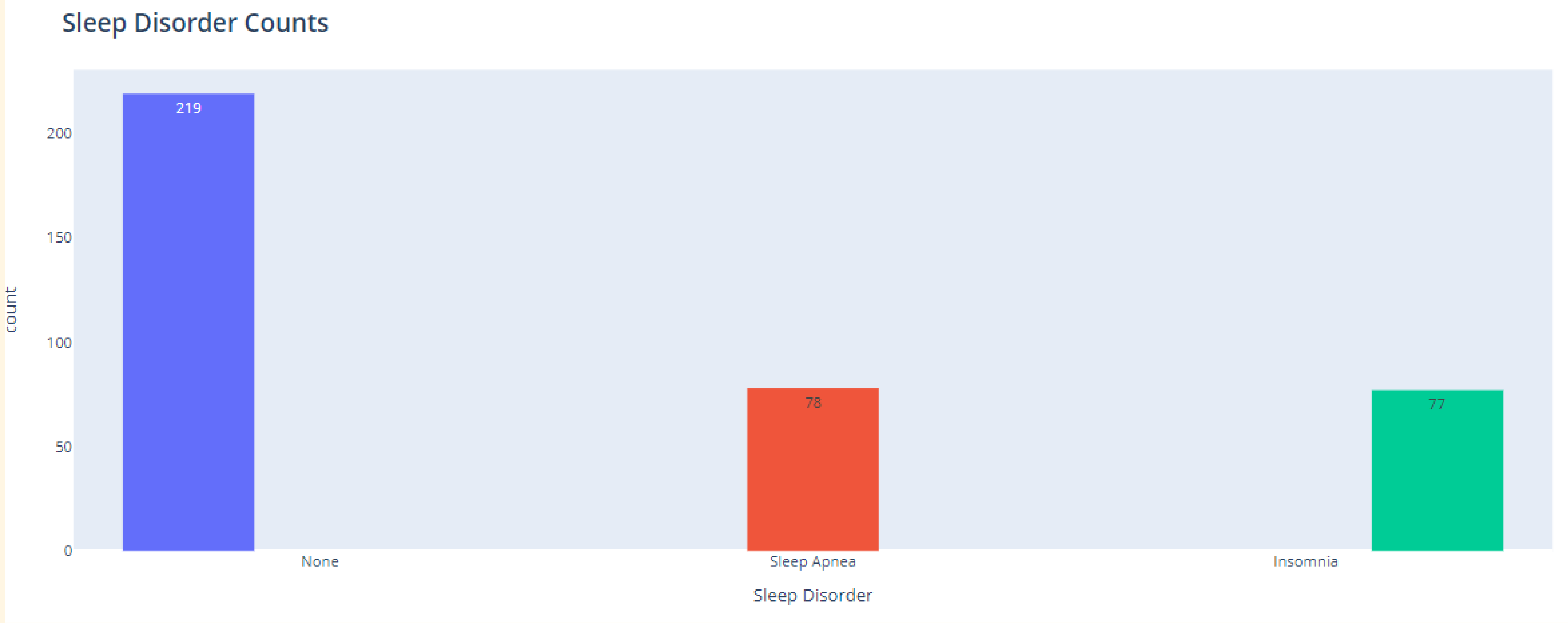
02 CATEGORICAL TO NUMERICAL FEATURES

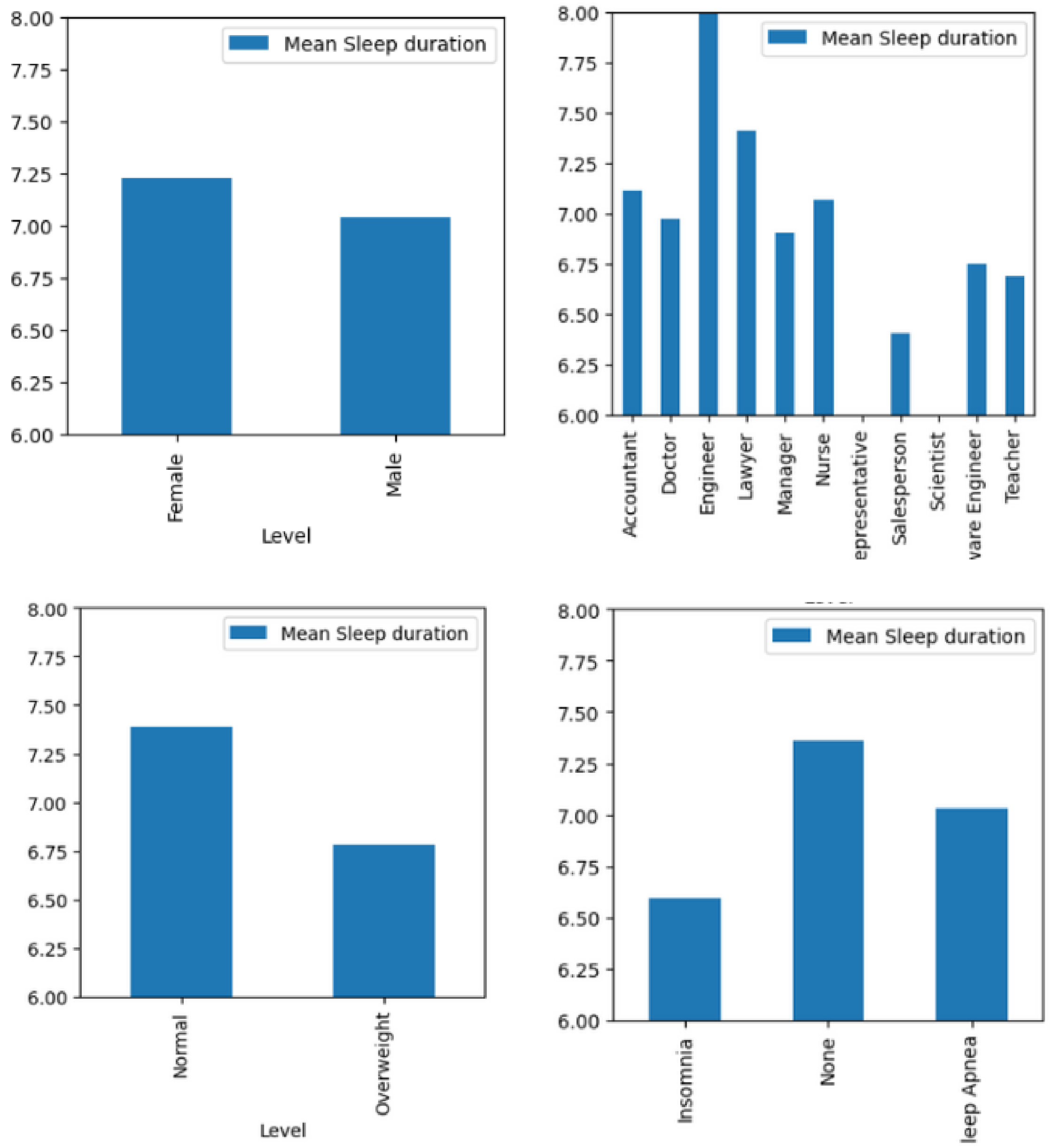


eart Rate	Daily Steps	Sleep Disorder	High Pressure	Low Pressure
77	4200	None	126	83
75	10000	None	125	80
75	10000	None	125	80
85	3000	Sleep Apnea	140	90
85	3000	Sleep Apnea	140	90
...	...	...	...	...
68	7000	Sleep Apnea	140	95

Blood Pressure to
High and Low Pressure

DATA *visualization*

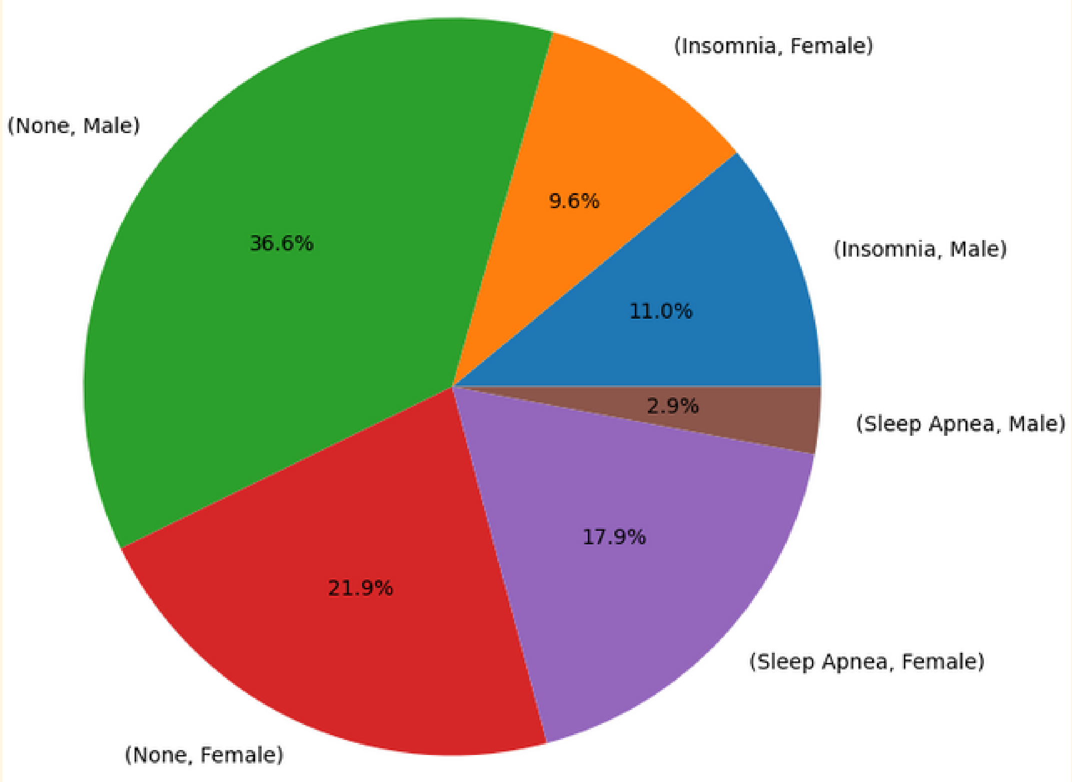




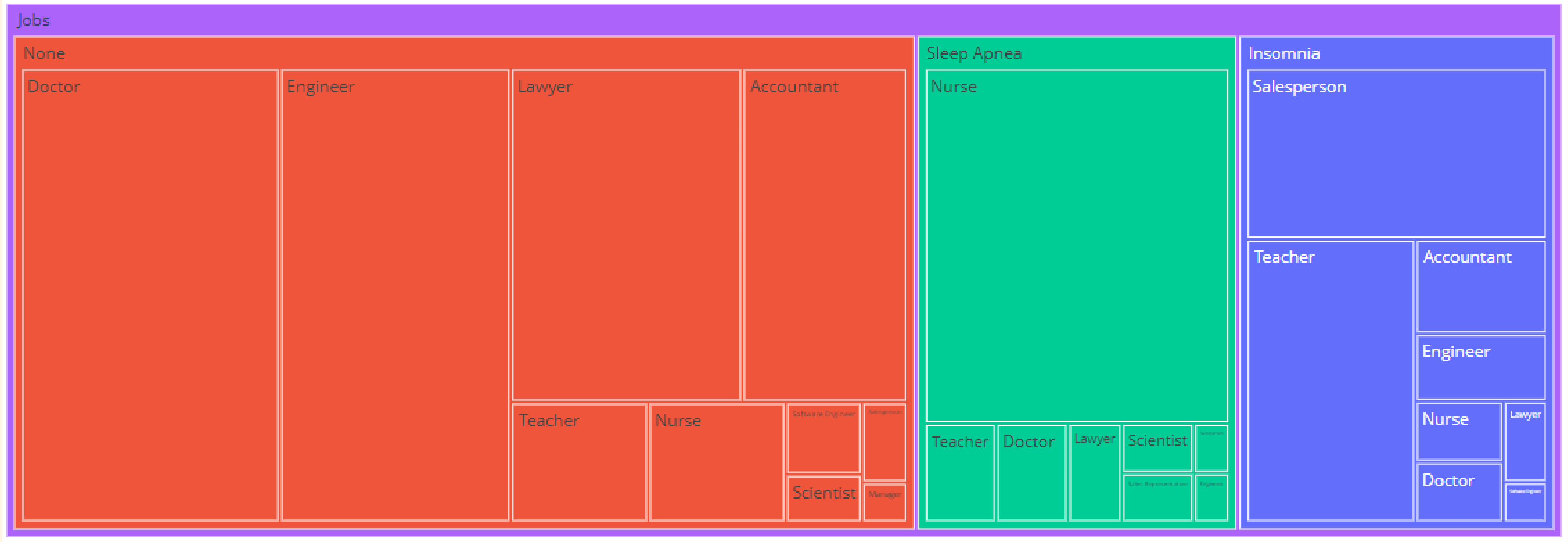
mainly focused on the analysis of the relationship between sleep disorder and other features..

categorical FEATURES

- Almost half of the people have sleep disorders
- **Sleep Apnea** is **more prevalent among women** than men
- **Insomnia** affects **more men** than women
- **Sleep Apnea nurses** took three quarters of the affected
    - thus, more common among women



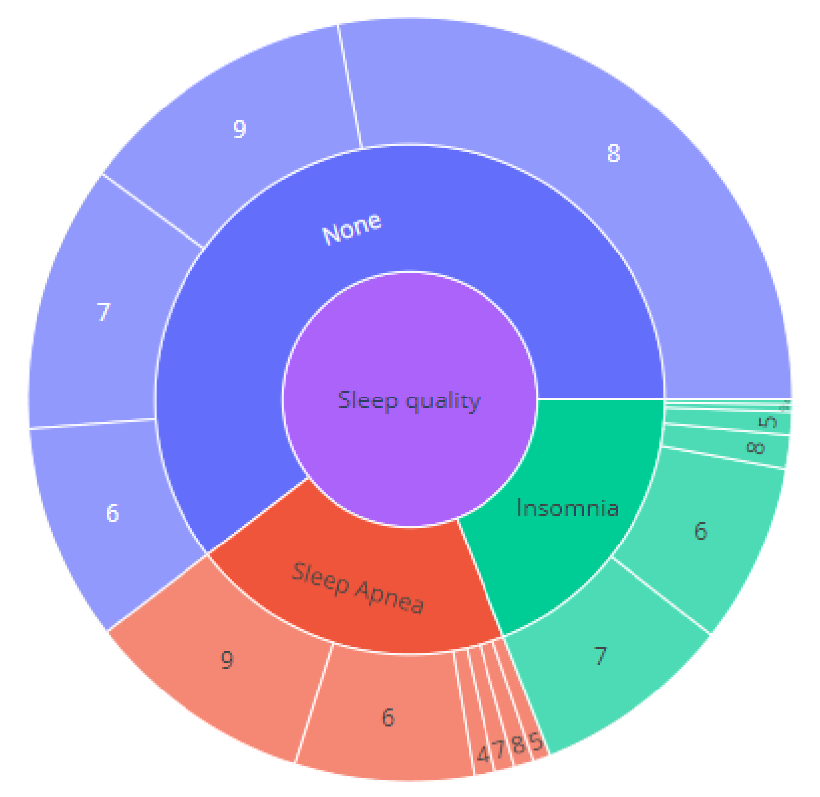
by Gender



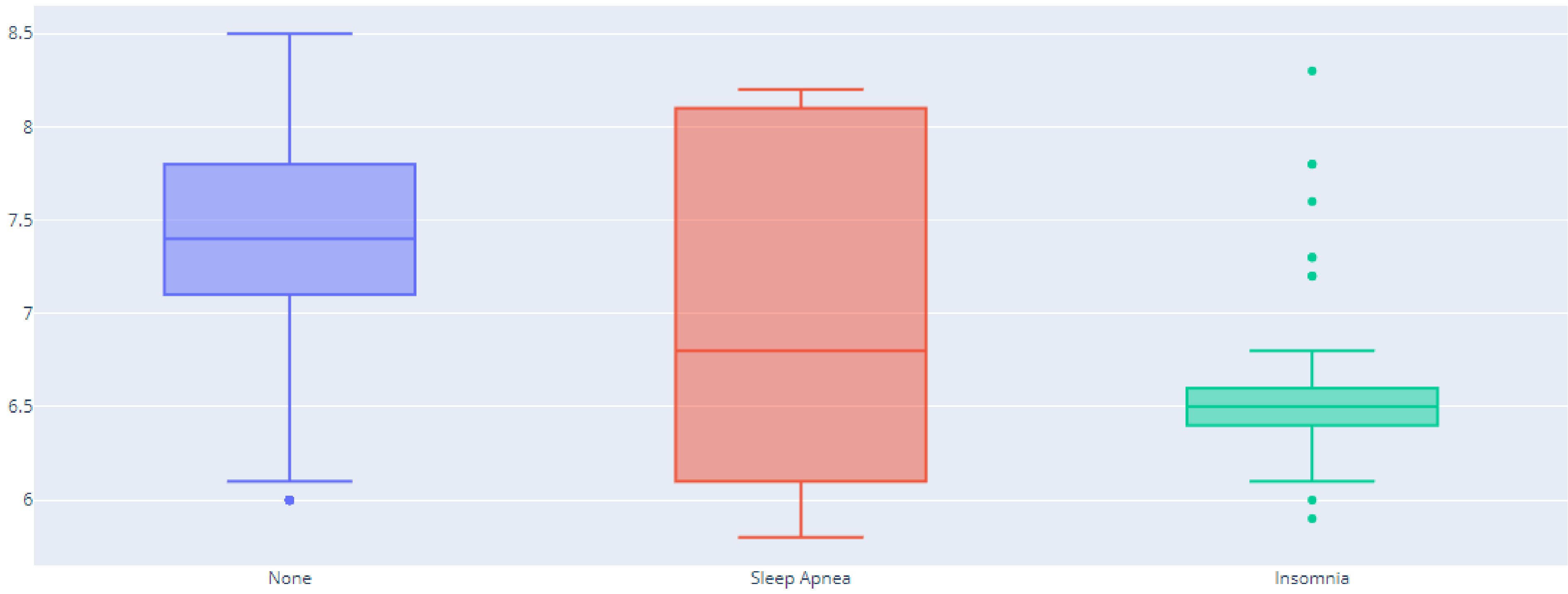
by Occupation

numerical FEATURES

- People without sleep disorders have high sleep quality
- **Insomnia** people have averagely **lower sleep quality**
- Average **sleeping hours for Insomnia** people is way lower
- **Interquartile range sleeping duration** for Sleep Apnea people is large

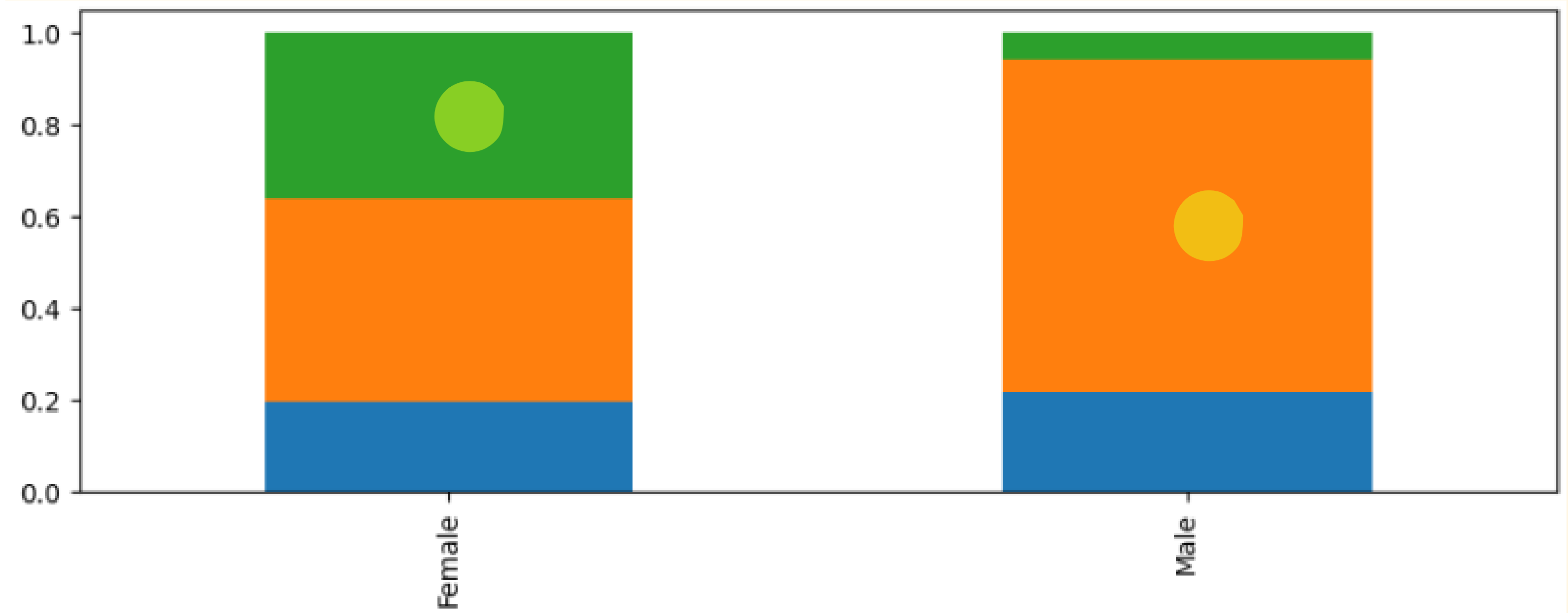


by Quality of Sleep

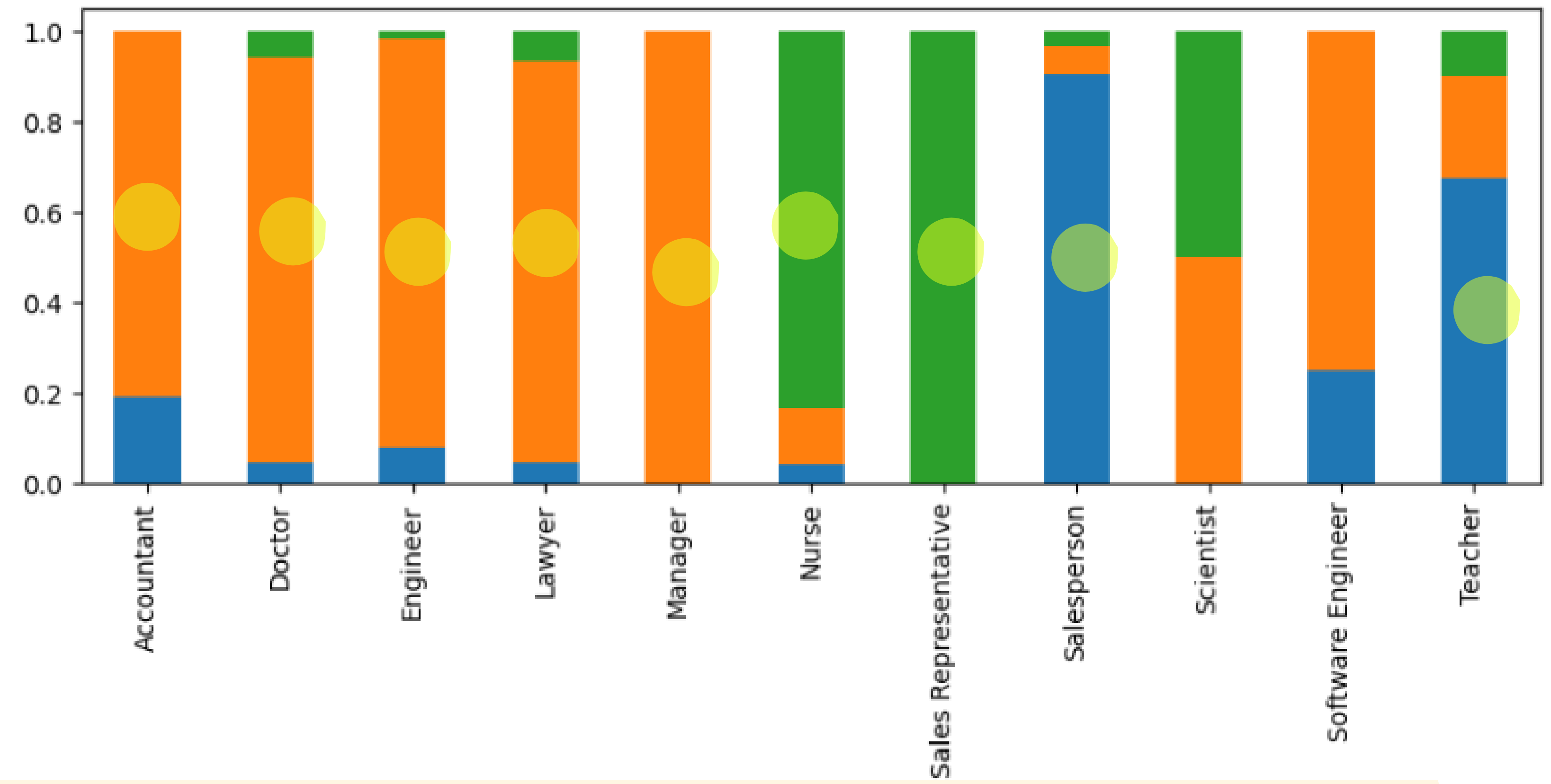


by Sleep Duration

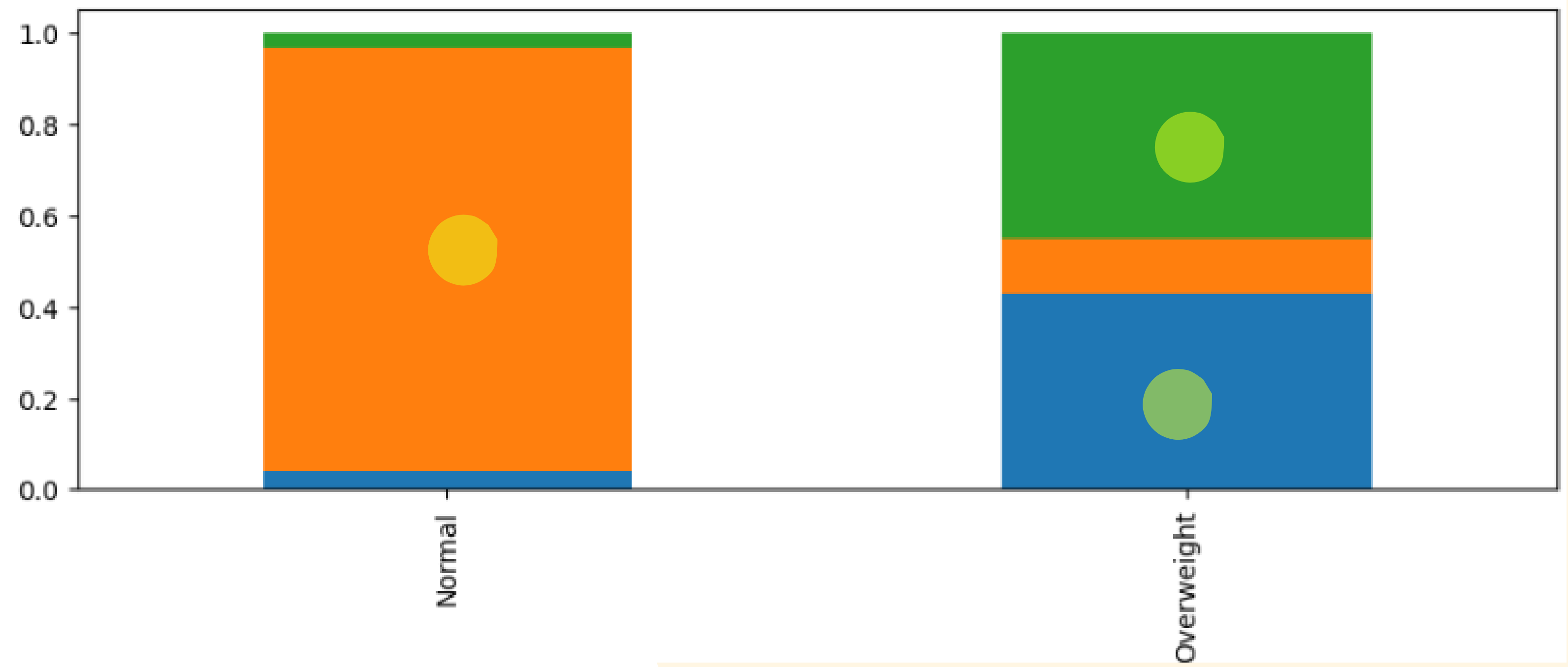
Gender



Occupation



BMI Category



Sleep Disorder

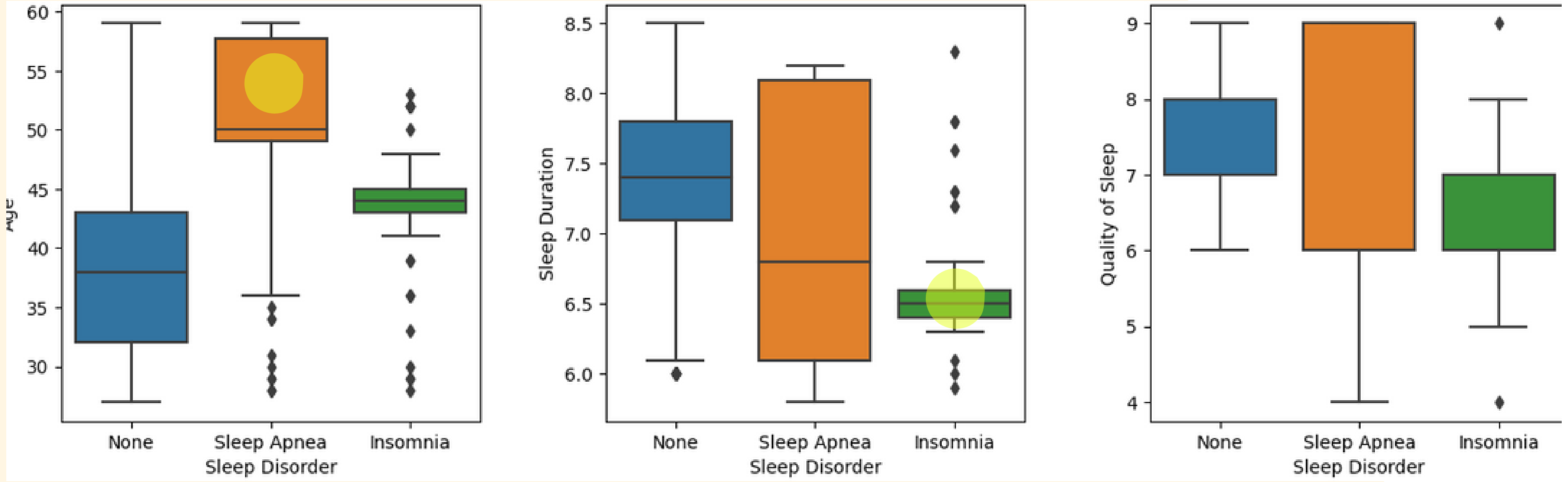
🟦 Insomnia

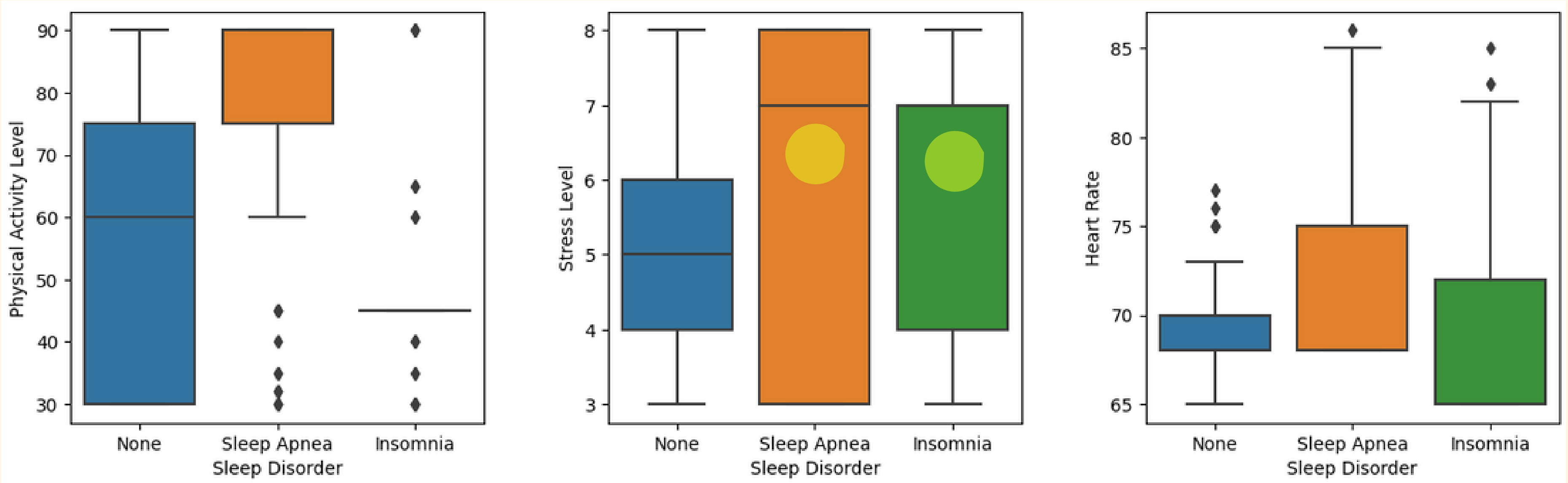
🟧 None

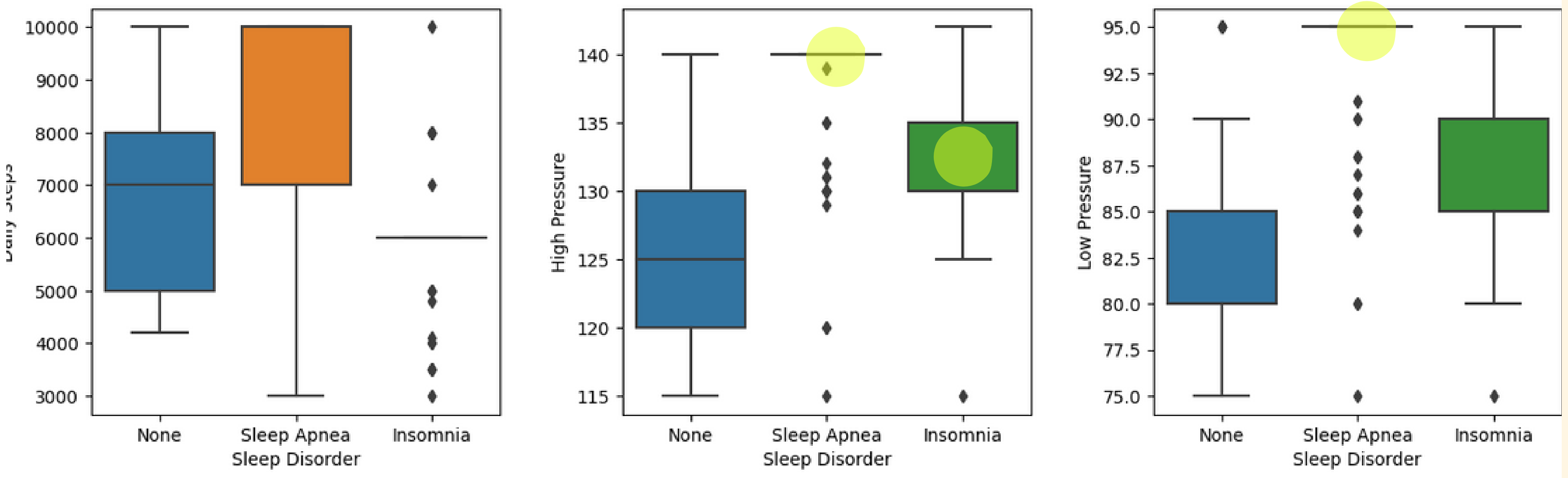
🟩 Sleep Apnea

Stacked Bar Charts Visualization of Sleep Disorder

BY CATEGORICAL FEATURES



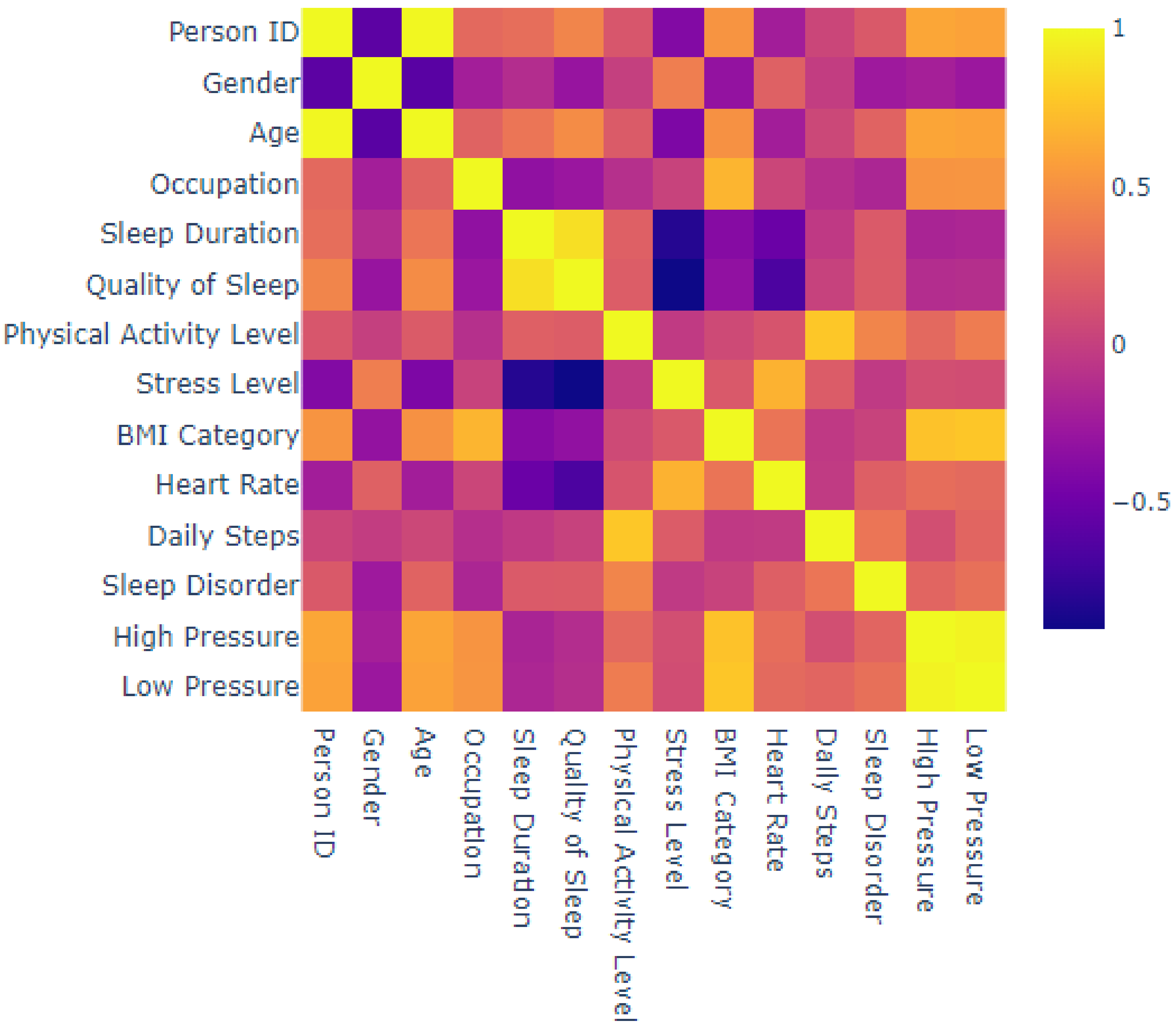




Grid of Box Plots Visualizations of Sleep Disorder

BY NUMERICAL FEATURES

matrix CORRELATION



Model BUILDING

01 Decision Tree

Grid Search: hyperparameter
K-fold: Training for 5 fold

02 Random Forest

Grid Search: hyperparameter
K-fold: Training for 5 fold

03 Logistic Regression

Grid Search: hyperparameter
K-fold: Training for 4 fold

04 SVM

Grid Search: hyperparameter
K-fold: Training for 2 fold

05 KNN

Grid Search: hyperparameter
K-fold: Training for 5 fold

06 Gradient Boosting

Grid Search: hyperparameter
K-fold: Training for 5 fold

Performance ANALYSIS

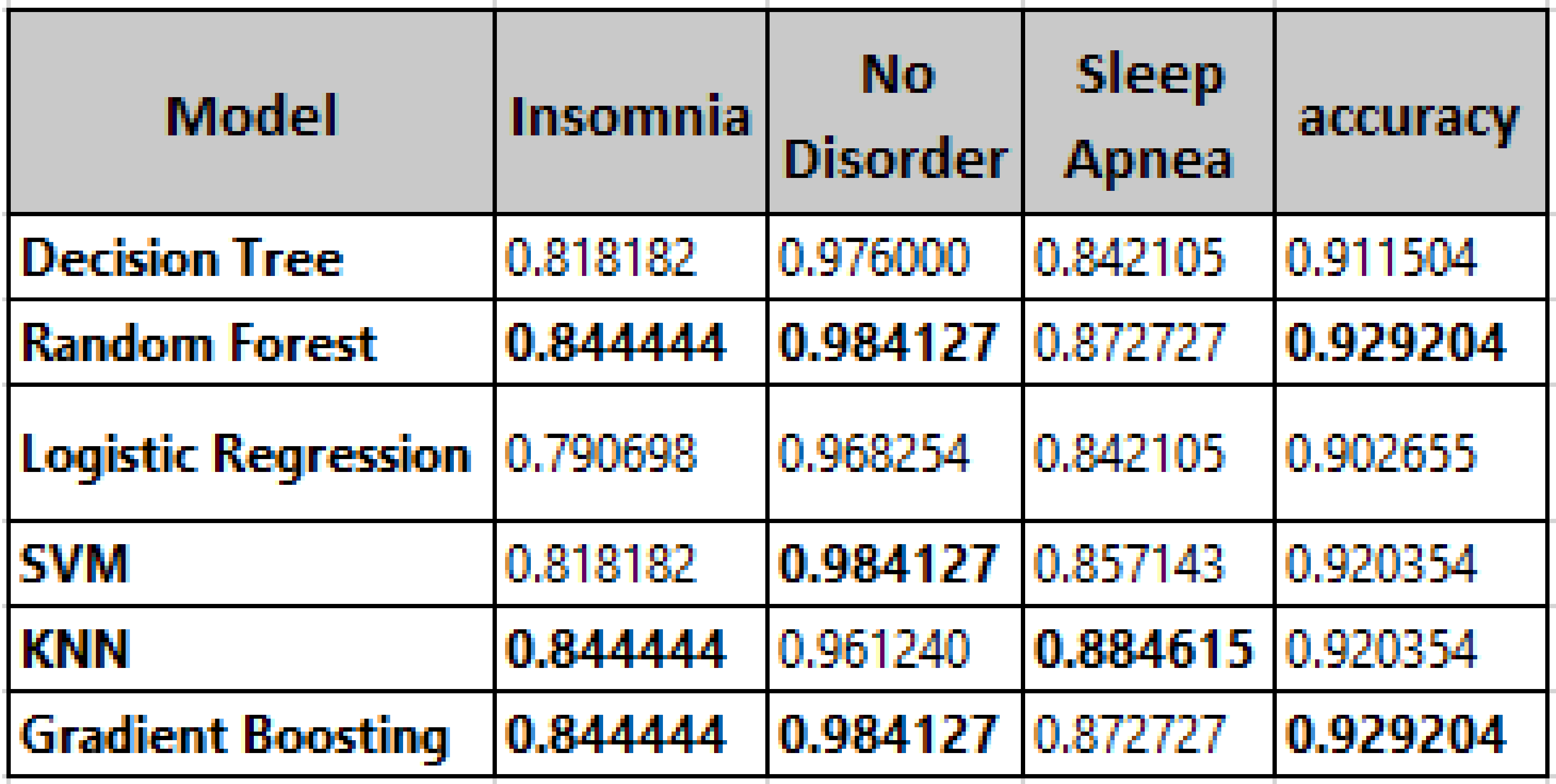
perform well on test data
all with accuracy > 0.9

Best F1-score for Insomnia:
Random Forest
KNN
Gradient Boosting

Best F1-score for No Disorder
Random Forest
SVM
Gradient Boosting

Best F1-score for Sleep Apnea
KNN

Best Overall Accuracy
Random Forest
Gradient Boosting

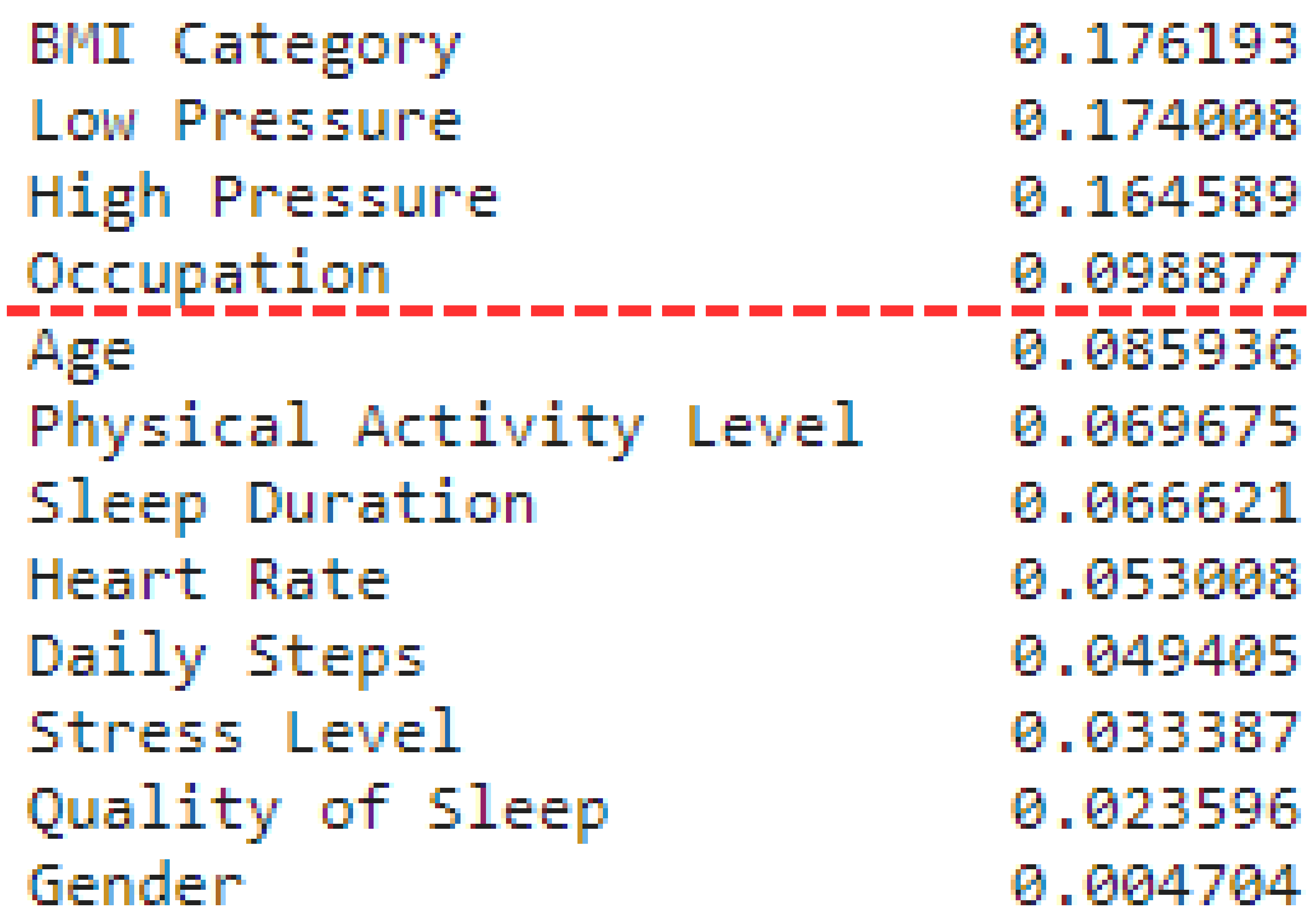


Model	Insomnia	No Disorder	Sleep Apnea	accuracy
Decision Tree	0.818182	0.976000	0.842105	0.911504
Random Forest	**0.844444**	**0.984127**	0.872727	**0.929204**
Logistic Regression	0.790698	0.968254	0.842105	0.902655
SVM	0.818182	**0.984127**	0.857143	0.920354
KNN	**0.844444**	0.961240	**0.884615**	0.920354
Gradient Boosting	**0.844444**	**0.984127**	0.872727	**0.929204**

F1 Score of the models

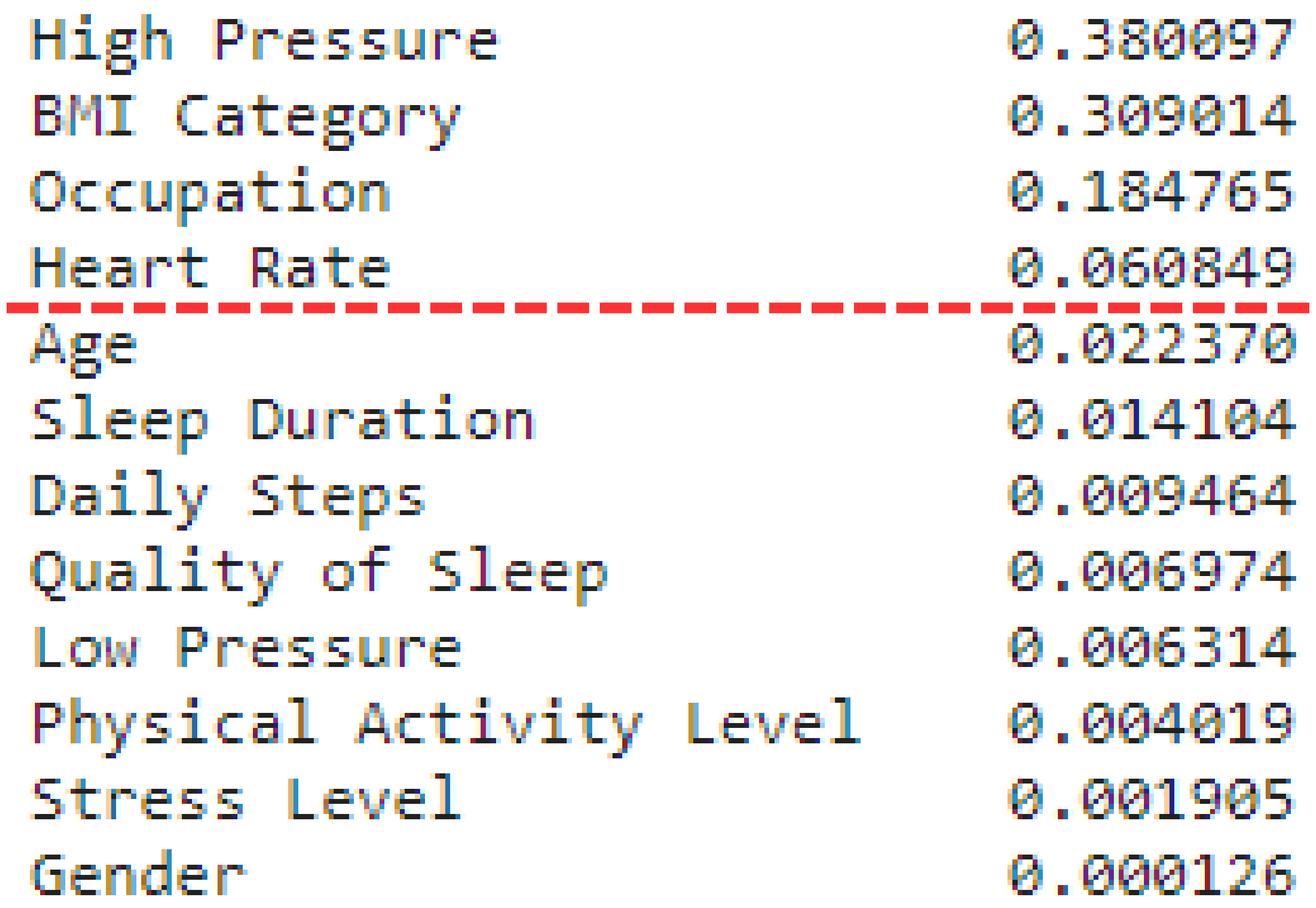
FEATURE IMPORTANCE

Random Forest



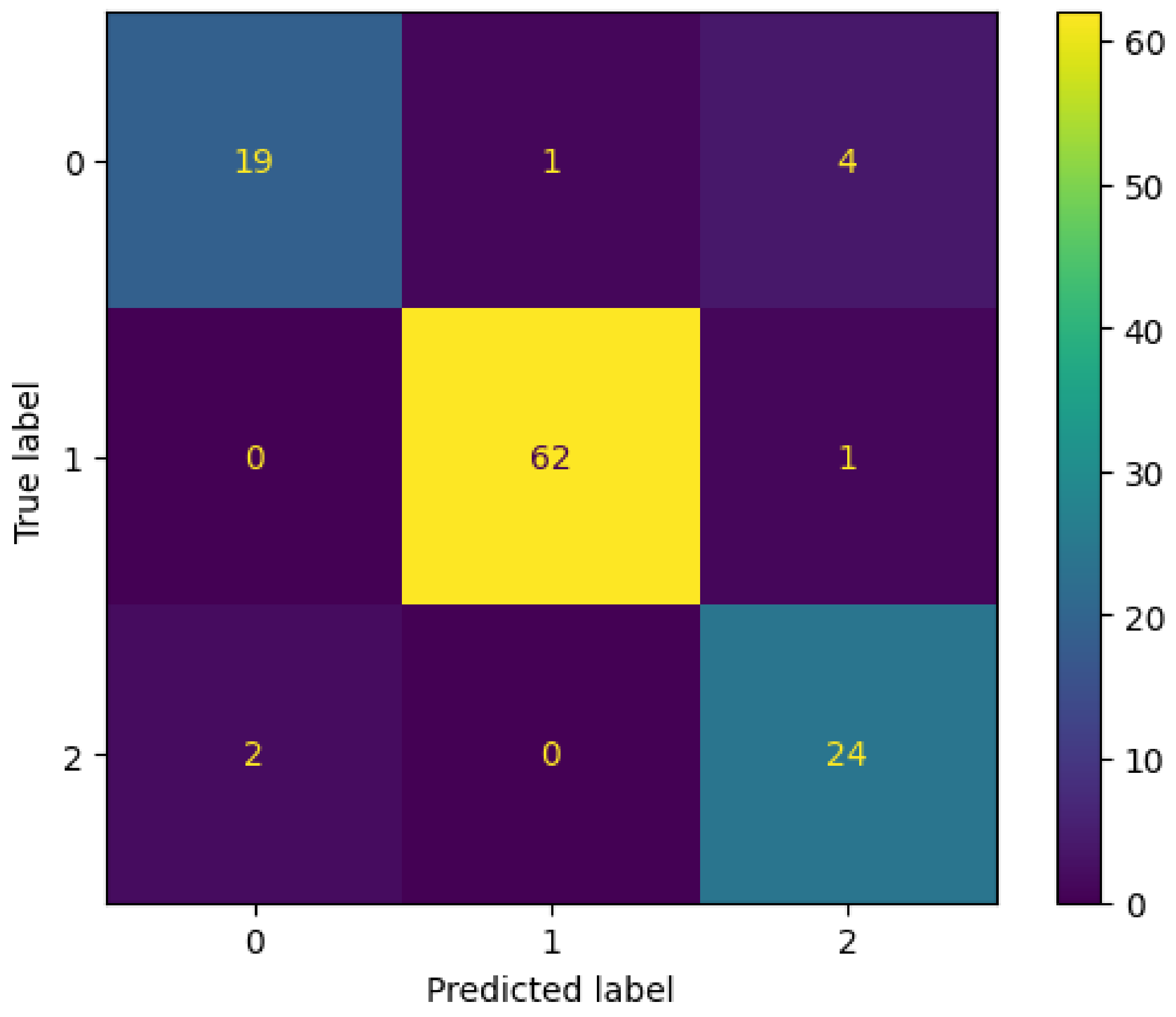
```
BMI Category                0.176193
Low Pressure                0.174008
High Pressure               0.164589
Occupation                  0.098877
Age                         0.085936
Physical Activity Level     0.069675
Sleep Duration              0.066621
Heart Rate                  0.053008
Daily Steps                 0.049405
Stress Level                0.033387
Quality of Sleep            0.023596
Gender                      0.004704
```

Gradient Boosting



```
High Pressure               0.380097
BMI Category                0.309014
Occupation                  0.184765
Heart Rate                  0.060849
Age                         0.022370
Sleep Duration              0.014104
Daily Steps                 0.009464
Quality of Sleep            0.006974
Low Pressure                0.006314
Physical Activity Level     0.004019
Stress Level                0.001905
Gender                      0.000126
```

Confusion Matrix of Random Forest and Gradient Boosting



CONCLUSION

Considering overall accuracy, Random Forest and Gradient Boosting proved to be the most effective models across all sleep disorders with 93% accuracy. From the visualization and the model, there are three dominant features that is important on determining sleep disorder: Blood Pressure, BMI and Occupation.

THANKS FOR LISTENING

109550186 李嘉玲
109550201 林家輝