Sleep Health and Lifestyle Banner

# Sleep Health and Lifestyle

## STAT 315 Final Project: Nicholas Johannessen, Saketh Marrapu, Alyssa Nugent, Iraam Rabbani

This synthetic dataset contains sleep and cardiovascular metrics as well as lifestyle factors of close to 400 fictive persons.

The workspace is set up with one CSV file, `data.csv`, with the following columns:

- `Person ID`
- `Gender`
- `Age`
- `Occupation`
- `Sleep Duration`: Average number of hours of sleep per day
- `Quality of Sleep`: A subjective rating on a 1-10 scale
- `Physical Activity Level`: Average number of minutes the person engages in physical activity daily
- `Stress Level`: A subjective rating on a 1-10 scale
- `BMI Category`
- `Blood Pressure`: Indicated as systolic pressure over diastolic pressure
- `Heart Rate`: In beats per minute
- `Daily Steps`
- `Sleep Disorder`: One of `None`, `Insomnia` or `Sleep Apnea`

Source: Kaggle

# Defining the Questions

Our research addresses several significant issues concerning the complexities of sleep quality across various health and demographic domains. By utilizing sophisticapted vizualization methods and machine learning models, we investigate if we can predict an individuals occupation based on certian cardiovascular and lifestyle factors, as well as explore whether different ages groups and BMI catergories are linked to quality of sleep. Moreover, the variety of vizualization methods we apply helps showcase how discrepencies in personal external factors are associated and to solidify the dynamic relationship between an individuall's quality of sleep and the presence or absence of certain sleep disorders.

# Data Collection

This is a simulated dataset containing data on lifestyle and cardiovascular factors (listed above) of 400 fabricated individuals. Since this data is simulated and not surveyed from a real population, some results of this project may not coincide with popular beliefs or assumptions.

# Data Cleaning and Preparation

We are going to check if our data has any null values to see if we can use it in our analysis. We are also checking to make sure no data has been misinput. By checking the minimum and maximum values of each category, we can see if there are any unreasonable values, such as minimum values below 0, or a sleep duration greater than 24, etc.

```
Number of null values:  0
Minimum age:  27
Maximum age:  59
Minimum sleep duration:  5.8
Maximum sleep duration:  8.5
Minimum sleep quality:  4
Maximum sleep quality:  9
Minimum physical activity level:  30
Maximum physical activity level:  90
Minimum stress level:  3
Maximum stress level:  8
Minimum heart rate:  65
Maximum heart rate:  86
Minimum daily steps:  3000
Maximum daily steps:  10000
```

There are no null values. Through examining the minimum and maximum values of our data, we can see that there are no values that are unreasonable, or outside the bounds of what would be possible.

## Cleaning our non numeric variables.

```
Genders:  ['Male' 'Female']
Occupations:  ['Software Engineer' 'Doctor' 'Sales Representative' 'Teacher' 'Nurse'
 'Engineer' 'Accountant' 'Scientist' 'Lawyer' 'Salesperson' 'Manager']
BMI Categories:  ['Overweight' 'Normal' 'Obese' 'Normal Weight']
Blood Pressures:  ['126/83' '125/80' '140/90' '120/80' '132/87' '130/86' '117/76' '118/7
6'
 '128/85' '131/86' '128/84' '115/75' '135/88' '129/84' '130/85' '115/78'
 '119/77' '121/79' '125/82' '135/90' '122/80' '142/92' '140/95' '139/91'
 '118/75']
Sleep Disorders  ['None' 'Sleep Apnea' 'Insomnia']
```

From this, we can see that there are some categories that are repetitive, such as Normal and Normal weight in the BMI categories, and Sales Representative and Sales Person in the Occupations category. To fix this, we will merge the smaller of the two categories to be in the larger of the two. So if someone is classified as Normal weight, we will now classify them as Normal.

```
Number of normal BMI:  195
Number or normal weight BMI:  21
Number of salespeople:  32
Number or sales representatives:  2
```

As seen above, the normal BMI Category is larger, so we will change all people labeled as normal weight to just normal.

Again, we can see above that the salesperson category is larger so we will relabel sales represenatives as salespeople.

Out[5]:

| | Person ID | Gender | Age | Occupation | Sleep Duration | Quality of Sleep | Physical Activity Level | Stress Level | BMI Category | Blood Pressure | Heart Rate | Daily Steps |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | Male | 27 | Software Engineer | 6.1 | 6 | 42 | 6 | Overweight | 126/83 | 77 | 420 |
| **1** | 2 | Male | 28 | Doctor | 6.2 | 6 | 60 | 8 | Normal | 125/80 | 75 | 1000 |
| **2** | 3 | Male | 28 | Doctor | 6.2 | 6 | 60 | 8 | Normal | 125/80 | 75 | 1000 |
| **3** | 4 | Male | 28 | Salesperson | 5.9 | 4 | 30 | 8 | Obese | 140/90 | 85 | 300 |
| **4** | 5 | Male | 28 | Salesperson | 5.9 | 4 | 30 | 8 | Obese | 140/90 | 85 | 300 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | . |
| **369** | 370 | Female | 59 | Nurse | 8.1 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |
| **370** | 371 | Female | 59 | Nurse | 8.0 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |
| **371** | 372 | Female | 59 | Nurse | 8.1 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |
| **372** | 373 | Female | 59 | Nurse | 8.1 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |
| **373** | 374 | Female | 59 | Nurse | 8.1 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |

374 rows × 13 columns

One flaw with this data set is that blood pressure is entered as a string. To fix this we are going to create a new column called pulse pressure, which is the difference between the numerator and denominator of the blood pressure variable.

We have ensured we have no repeating categorical variables and that our numerical variables are within the bounds of what is reasonable. We created a new column called pulse pressure which is a numeric way of viewing the blood pressure column, which is entered as a string. We are ready to prepare the data for binary classification and then perform classification tasks and modeling.

# Preparing Data for Binary Classification

We are hoping to classify someone as having a disorder or not based on other traits they have. To do this, we need to encode the presence or absence of a sleep disorder with a binary variable. To do this we add in a column called Disorder Present. It is a binary variable used to designate if a person is suffering from a sleep disorder or not. 1 represents they have a disorder, 0 represents that they do not. We base this off the already built in column called Sleep Disorder, which is a string variable with options 'None', 'Sleep Apnea', or 'Insomnia'. If Sleep Disorder column has value 'None', we code in a 0 for Disorder Present. Otherwise, Disorder Present has value 1. We also create binary variables called Apnea and Insomnia that designate if a person is suffering from Apnea or Insomnia respectively. These columns are constructed in a similar manner.

Out[7]:

| | Person ID | Gender | Age | Occupation | Sleep Duration | Quality of Sleep | Physical Activity Level | Stress Level | BMI Category | Blood Pressure | Heart Rate | Daily Step |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | Male | 27 | Software Engineer | 6.1 | 6 | 42 | 6 | Overweight | 126/83 | 77 | 420 |
| **1** | 2 | Male | 28 | Doctor | 6.2 | 6 | 60 | 8 | Normal | 125/80 | 75 | 1000 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 3 | Male | 28 | Doctor | 6.2 | 6 | 60 | 8 | Normal | 125/80 | 75 | 1000 |
| **3** | 4 | Male | 28 | Salesperson | 5.9 | 4 | 30 | 8 | Obese | 140/90 | 85 | 300 |
| **4** | 5 | Male | 28 | Salesperson | 5.9 | 4 | 30 | 8 | Obese | 140/90 | 85 | 300 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | . |
| **369** | 370 | Female | 59 | Nurse | 8.1 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |
| **370** | 371 | Female | 59 | Nurse | 8.0 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |
| **371** | 372 | Female | 59 | Nurse | 8.1 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |
| **372** | 373 | Female | 59 | Nurse | 8.1 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |
| **373** | 374 | Female | 59 | Nurse | 8.1 | 9 | 75 | 3 | Overweight | 140/95 | 68 | 700 |

374 rows × 19 columns

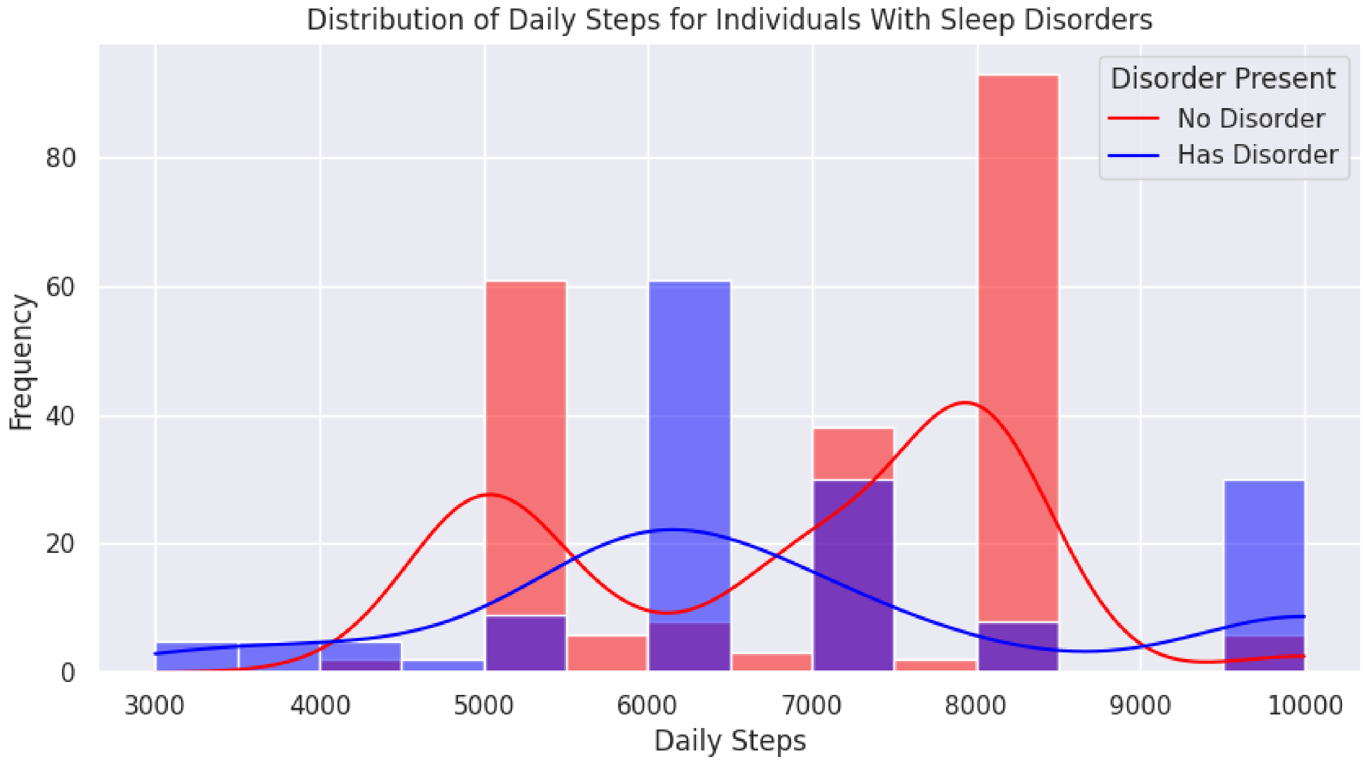
## Splitting Data

We create four dataframes that split our original data based on the sleep disorder or lack thereof that someone is suffering from. Thus, we have separate dataframes that contain all the data for each category of Sleep disorder, and a fourth dataframe that contains data for anyone with a sleep disorder, whether it is sleep apnea or insomnia. This will make working with only data for people from a certain category much easier, as well as examining data for people with and without sleep disorders.

# Can we predict if a person has a sleep disorder based on cardiovascular and lifestyle factors?

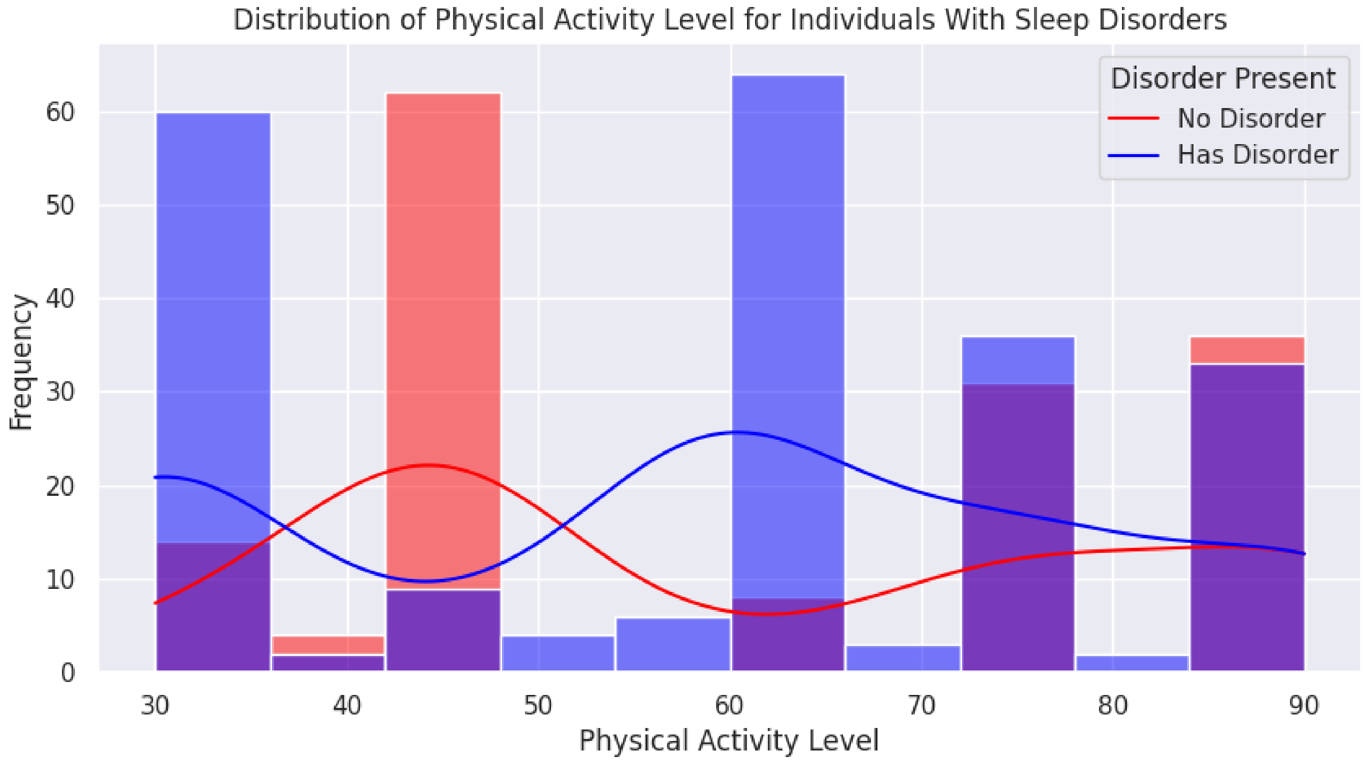
## Taking a look at data for people with sleep disorders and people without sleep disorders

We will create histograms to examine the differences in cardiovascular and lifestyle factors for people with and without sleep disorders. We will also examine summary statistics to see if there is a discernible difference between the groups.



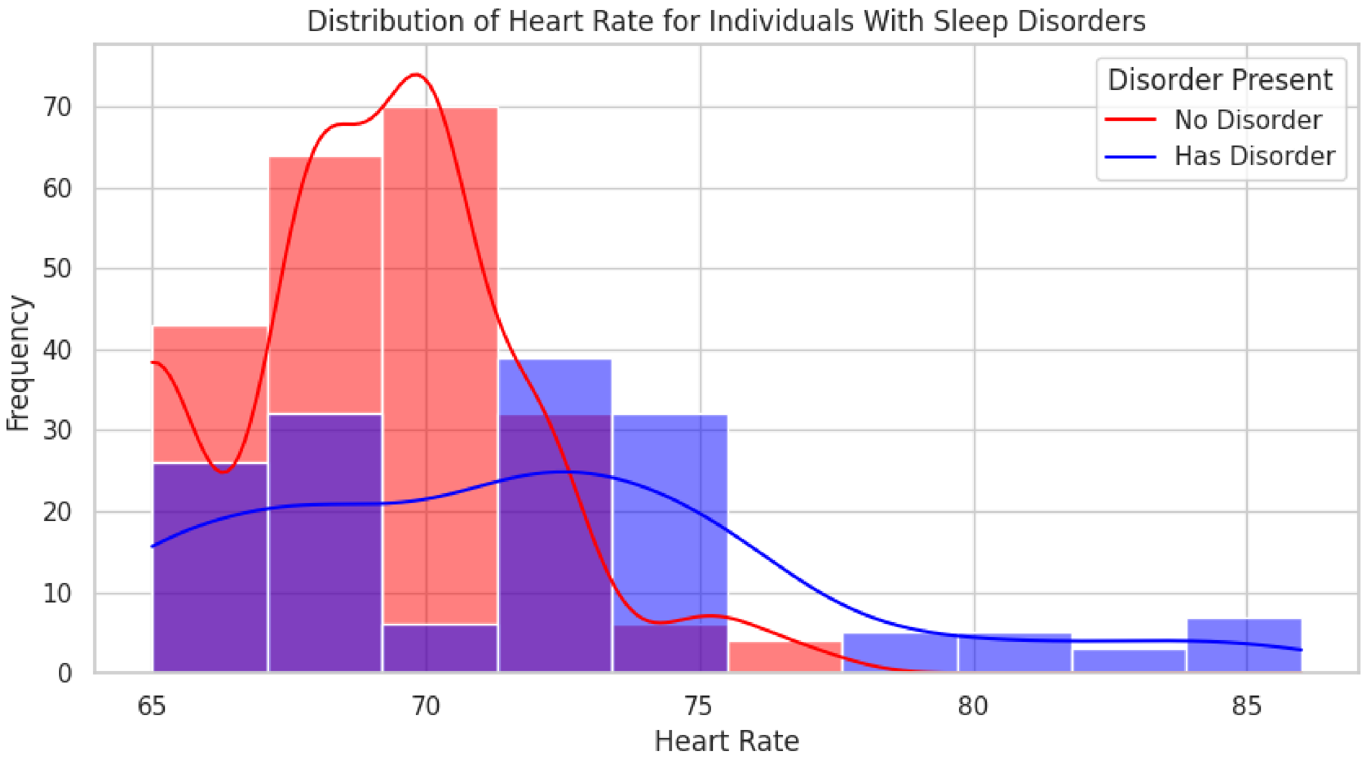
```
Average daily steps for people without sleep disorders:  6852.96803652968
Average daily steps for people with sleep disorders:  6765.806451612903
Standard deviation of daily steps for people without sleep disorders:  1393.473599780626
3
Standard deviation of daily steps for people with sleep disorders:  1893.9218812205268
```

From the above graph, we can see that there is no clear distribution for daily steps for either cateogry. No pattern is present However, we can see that the distribution for people with sleep disorders has a wider spread, which is also seen in its standard deviation being larger than that of the standard deviation of daily steps for people with sleep disorders. However, the average daily steps for both categories is about the same. We will further investigate with the heart rate variable.



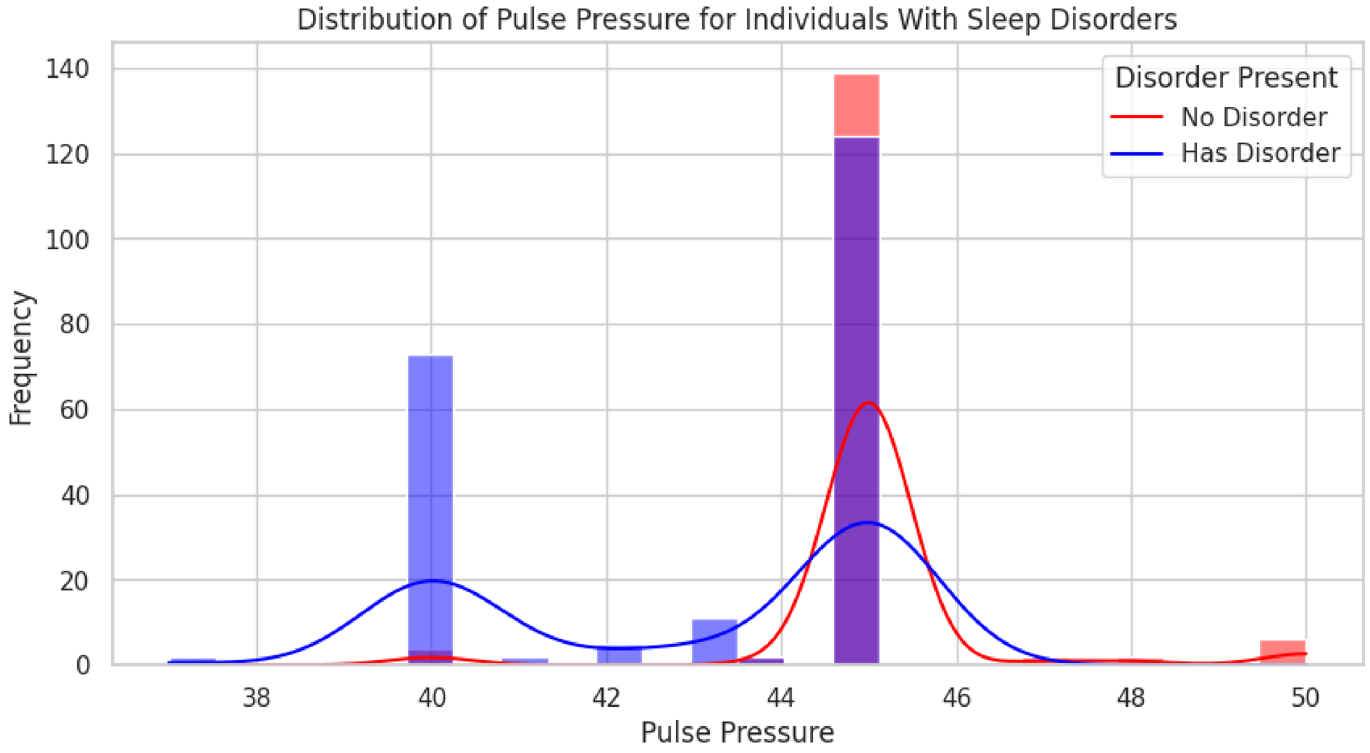
```
Average physical activity level for people without sleep disorders:  57.949771689497716
Average physical activity level for people with sleep disorders:  60.89677419354839
Standard deviation of physical activity level for people without sleep disorders:  20.92
9813629640574
Standard deviation of physical activity level for people with sleep disorders:  20.63415
847617995
```

Here we see that physical activity level has a large standard deviation which means it is less reliable and indicates low association. The averages are near one another but basically alludes to nothing since the standard deviation variance already shows this data is irrelevant as an indicator of sleep disorder prevalence.



```
Average heart rate for people without sleep disorders:  69.01826484018265
Average heart rate for people with sleep disorders:  71.78709677419354
Standard deviation of heart rate for people without sleep disorders:  2.6577970292714204
Standard deviation of heart rate for people with sleep disorders:  5.187381163432063
```

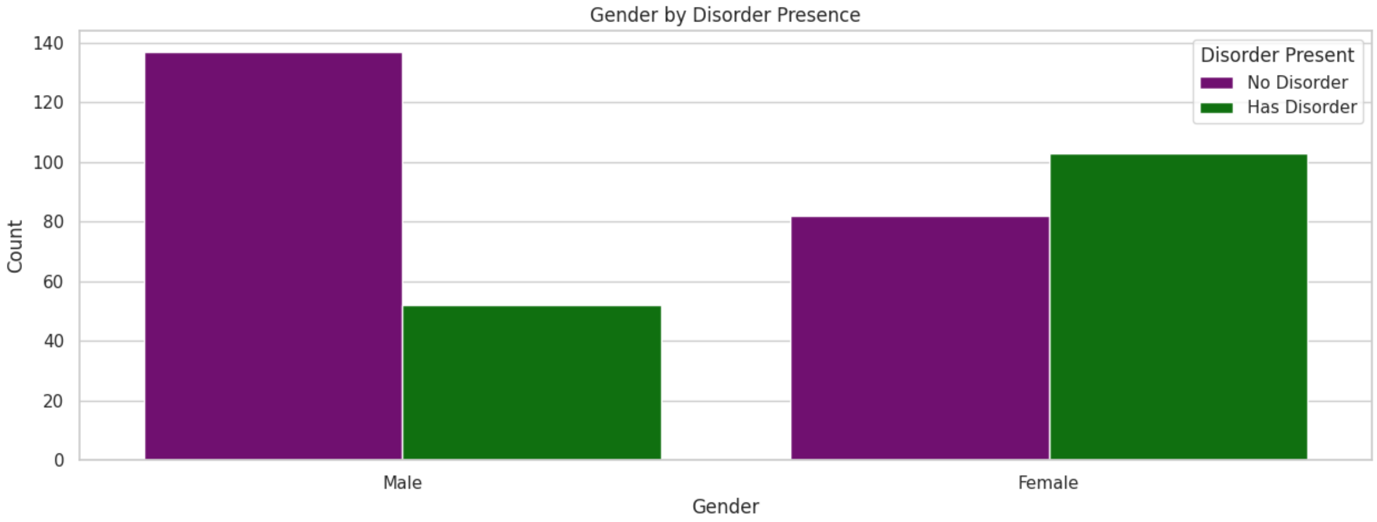
Now we look into cardiovascular health. From this, we can again see that there is no direct pattern in the histograms for either variable. We can see however, that people with sleep disorders seems to have a larger spread, which is supported by the summary statistics. We can also see that people with sleep disorders have a higher heart rate on average.

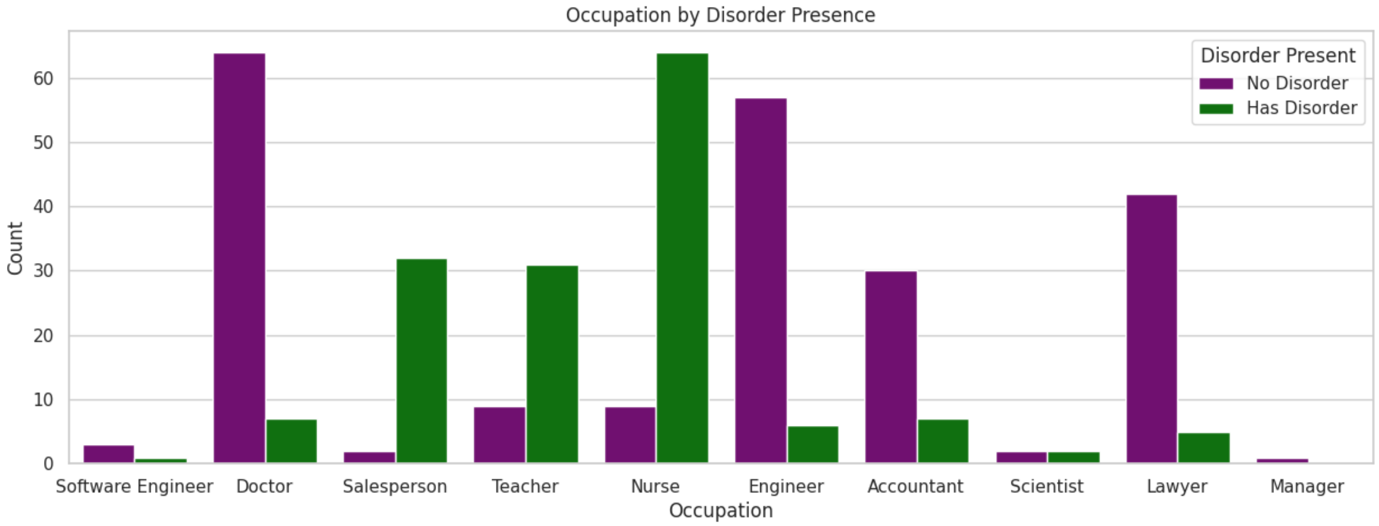


```
Average pulse pressure for people without sleep disorders:  43.045662100456624
Average pulse pressure for people with sleep disorders:  45.116129032258065
Standard deviation of pulse pressure for people without sleep disorders:  2.384529052607
9457
Standard deviation of pulse pressure for people with sleep disorders:  1.338514580400177
5
```

Through this graph we can see how pulse pressure has a normal distribution for those with no sleep disorder, whereas there is a bimodal distribution for those with a sleep disorder. The two peaks appear at the human average (same as for people without sleep disorders) and one around 40, much lower than normal. It is evident there is some association between pulse pressure and prevalence of sleep disorders but this does not prove causation.

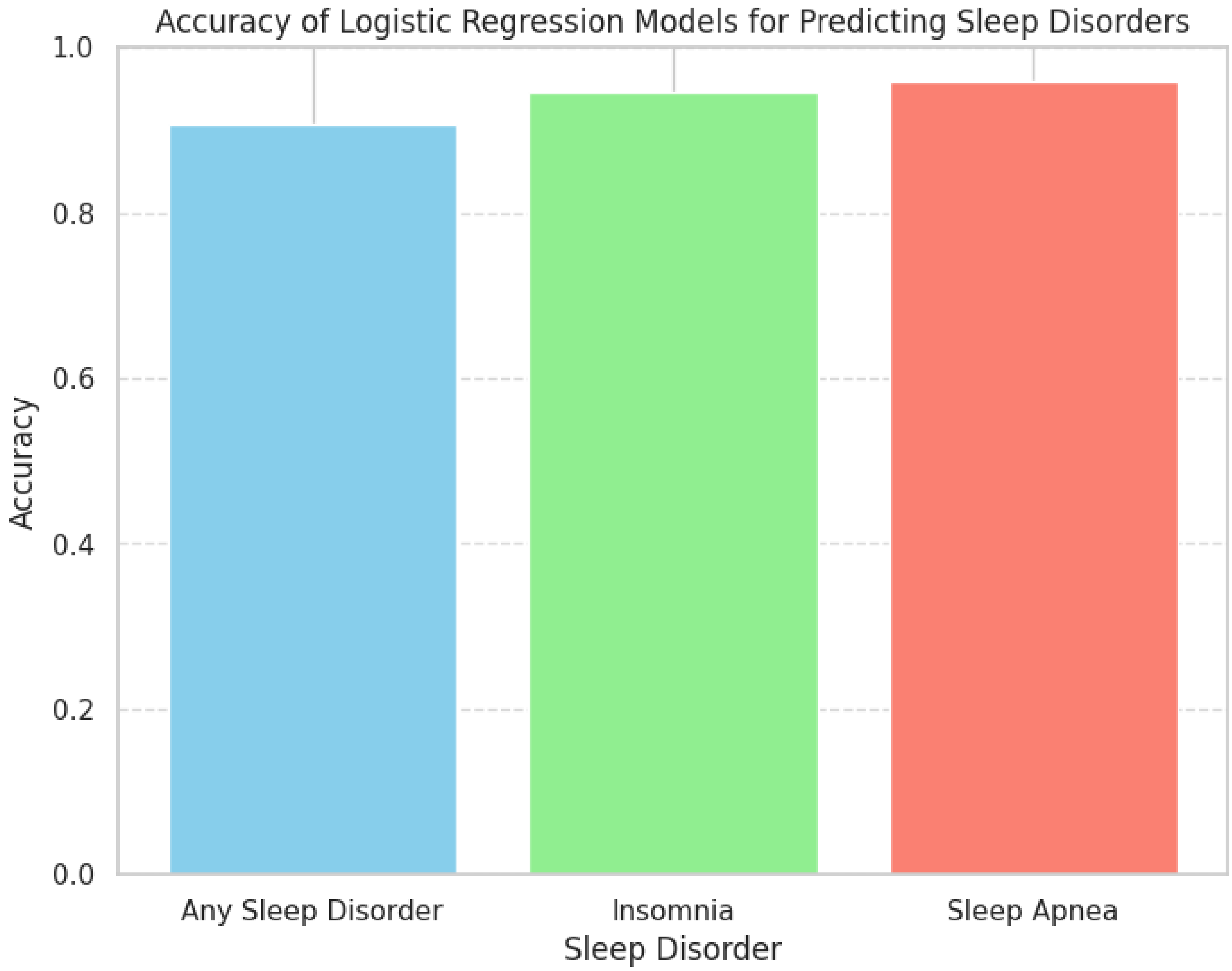
## Are Gender and Occupation Revelant?





The above two graphs can help us analyze such categorical information about the indivuduals, but first we must look into logistical regression analysis to solidify any claims.

# Creating a machine learning model to determine if someone has a disorder or not based on all factors. Then seeing if we can classify for sleep apnea and insomnia separately. Then, we will examine it for only lifestyle and health factors. We will use a logistic regression model since this is a classification task.

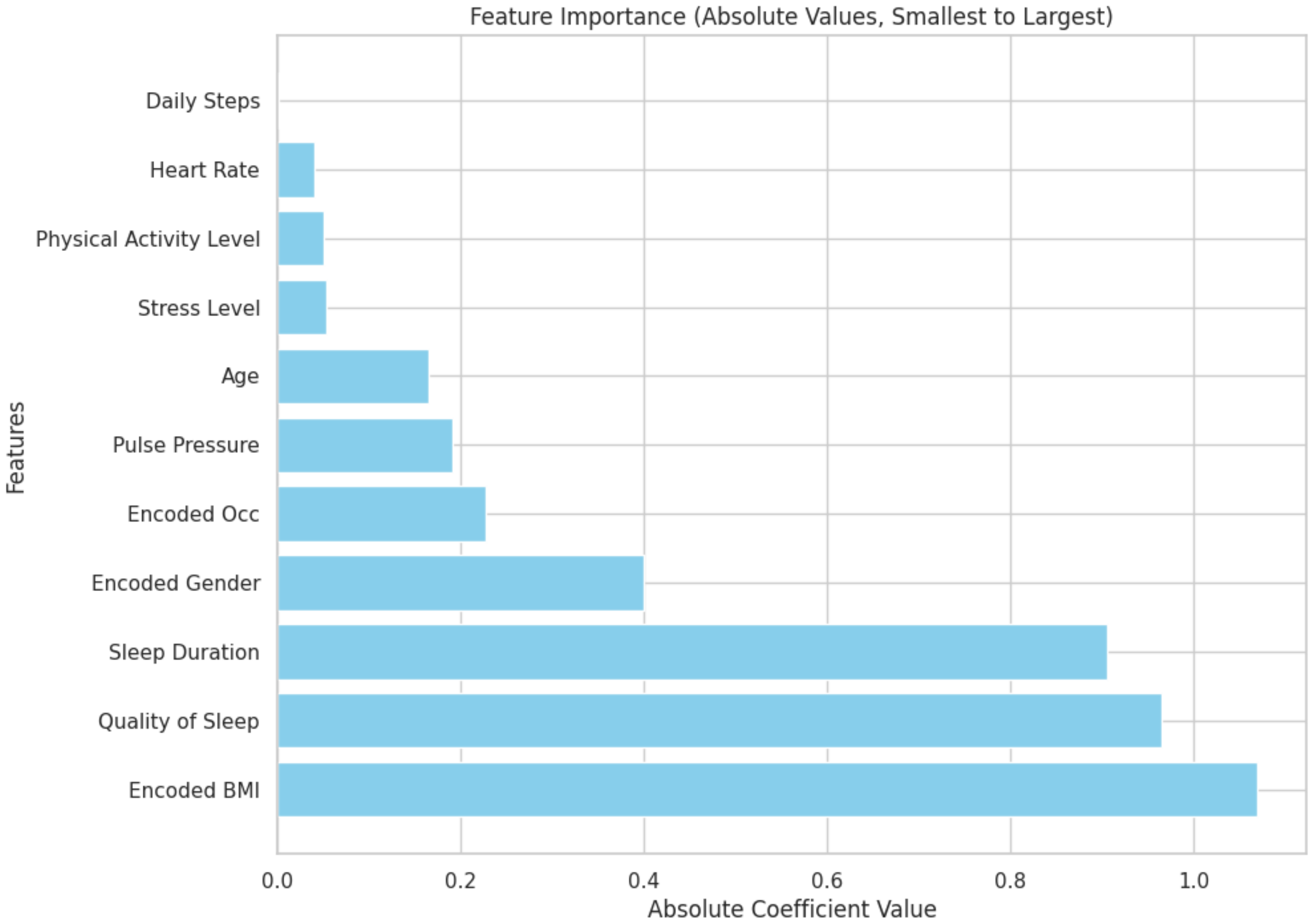


```
Accuracy for predicting any disorder:  0.9066666666666666
Accuracy for predicting any insomnia:  0.9466666666666667
Accuracy for predicting any disorder:  0.96
```

The above bar graph shows the accuracy of the logistic regression model determining the presence of sleep disorders. Predicting the presence of any sleep disorder has a 92% accuracy, more specifically, insomnia is 93%, and sleep apnea is 96%. I chose to use a bar graph because the simple graph was appropriate to show the discrepencies between this simple set of data.

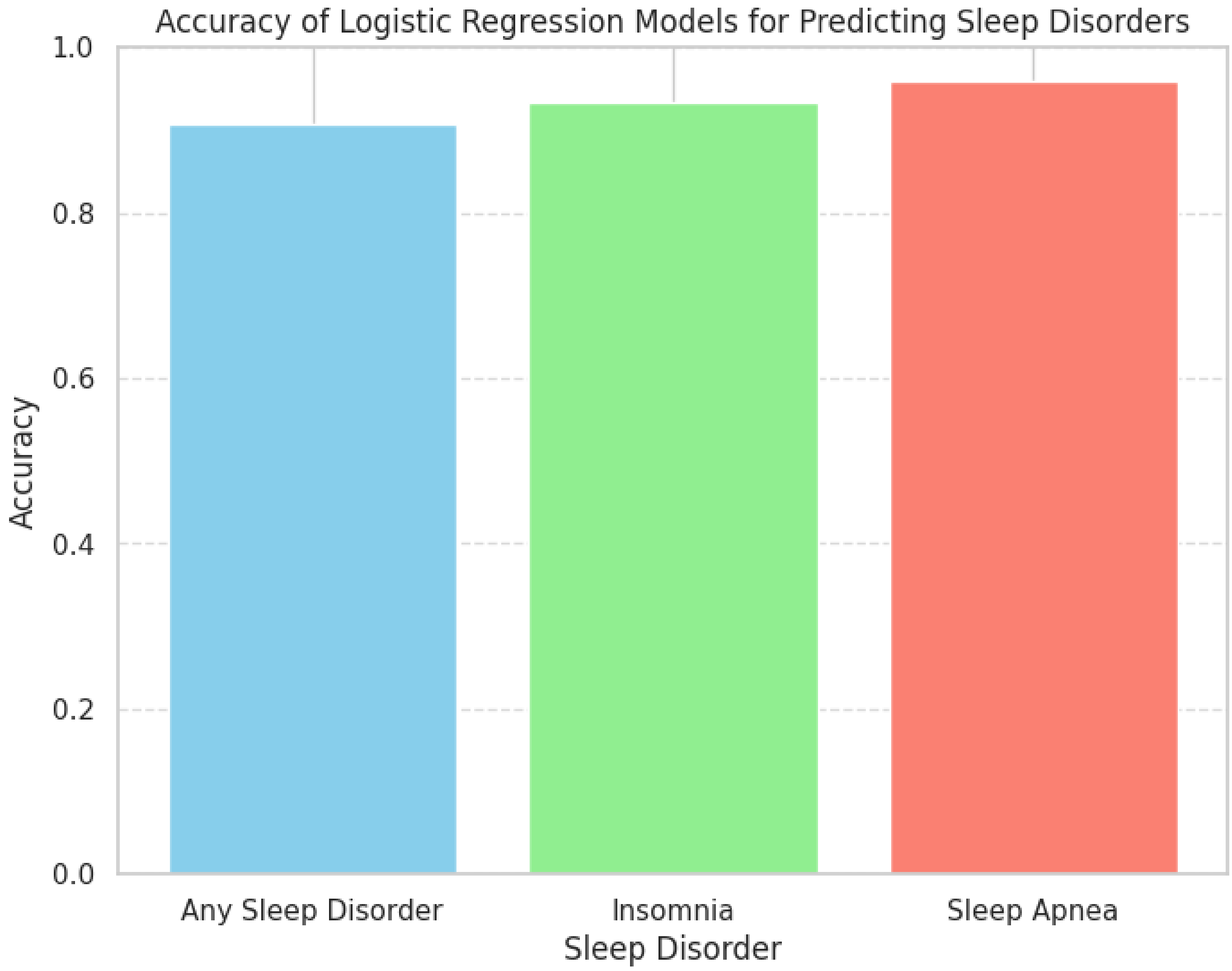
It is interesting to see that the accuracy for each separate order is higher than for the combined sleep disorders. This is an outcome I would not have expected to see. I would think that if the model could predict each separate disorder, that predicting if someone has *any* disorder would be easier. However, this accuracy above 90% is very high and shows that we can predict if someone has a sleep disorder based off the given factors very well.

```
('Encoded Gender', -0.39942147703849445)
('Age', 0.1648212929470206)
('Encoded Occ', 0.22667054470334969)
('Sleep Duration', -0.9059350514408948)
('Quality of Sleep', -0.9647433701156014)
('Physical Activity Level', 0.050379984596678334)
('Stress Level', -0.052758537928022715)
('Encoded BMI', 1.0696827677002165)
('Pulse Pressure', 0.19042790976563062)
('Heart Rate', -0.03953913314401261)
('Daily Steps', -0.0005202763566918665)
```



When we examine the coefficient values of each feature, we can see that quality of sleep, sleep duration, encoded BMI, and encoded gender were the most important features. Encoded BMI seems to be the most important feature, which is great since we want to examine based only on health and activity features.

## Refining the model now to only include our health and activity level factors

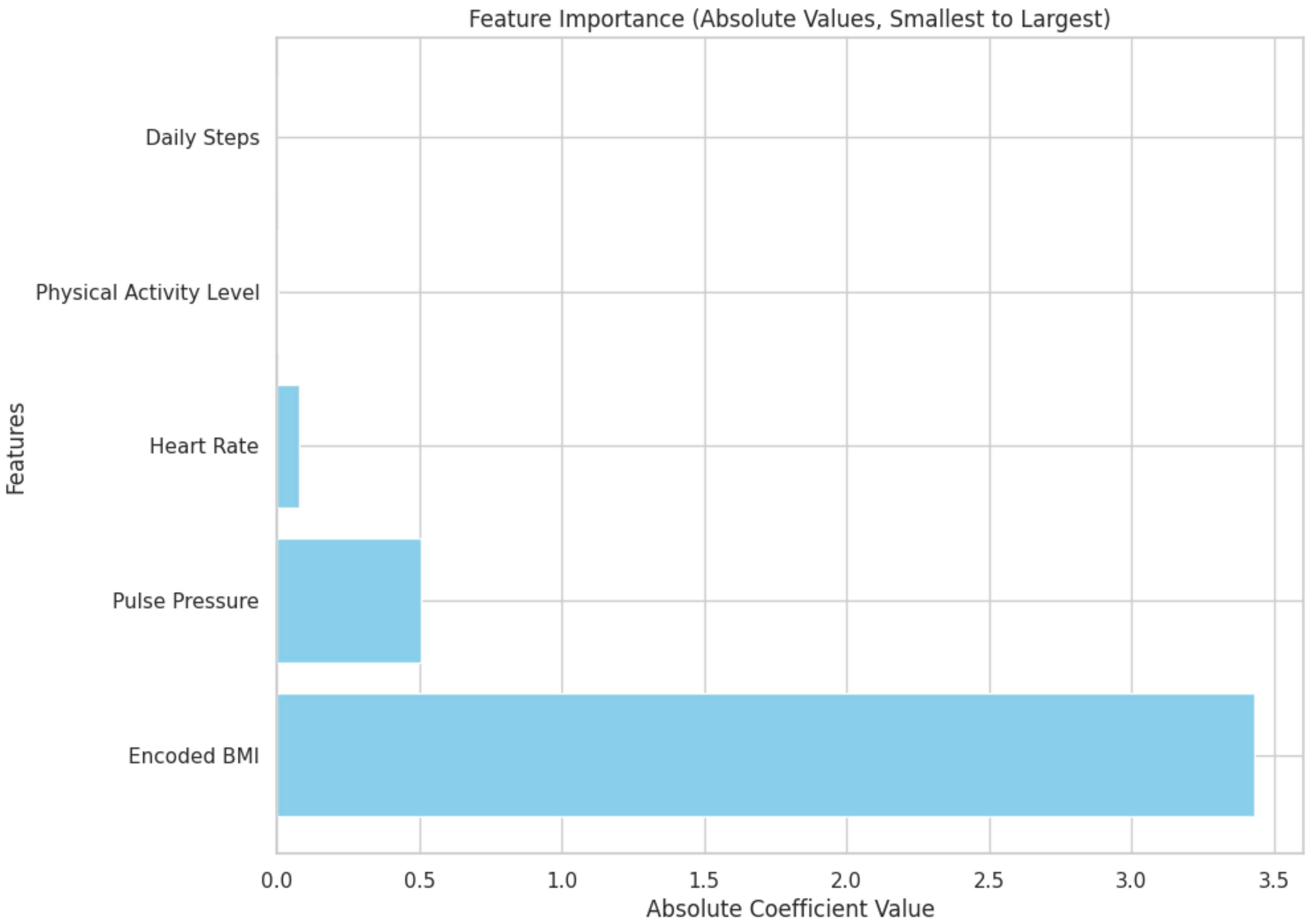


```
Accuracy for predicting any disorder:  0.9066666666666666
Accuracy for predicting any insomnia:  0.9333333333333333
Accuracy for predicting any disorder:  0.96
```

The above bar graph shows the accuracy of the logistic regression model determining the presence of sleep disorders. Predicting the presence of any sleep disorder has a 90.6% accuracy, more specifically, insomnia is 92%, and sleep apnea is 94.6%. I chose to use a bar graph because the simple graph was appropriate to show the discrepencies between this simple set of data.

When looking at only health and activity factors, we see that accuracy decreases a bit. This is interesting to me, as it shows that there are things besides health and activity level that affect the presence or absence of a sleep disorder. However, we still have very high accuracy in our predictions. Again, it is interesting to see that the accuracy for each separate order is higher than for the combined sleep disorders.

```
('Physical Activity Level', -0.002697702328277484)
('Encoded BMI', 3.433524872536418)
('Pulse Pressure', 0.506836441292141)
('Heart Rate', 0.07687601686663037)
('Daily Steps', 0.00010883907155239811)
```

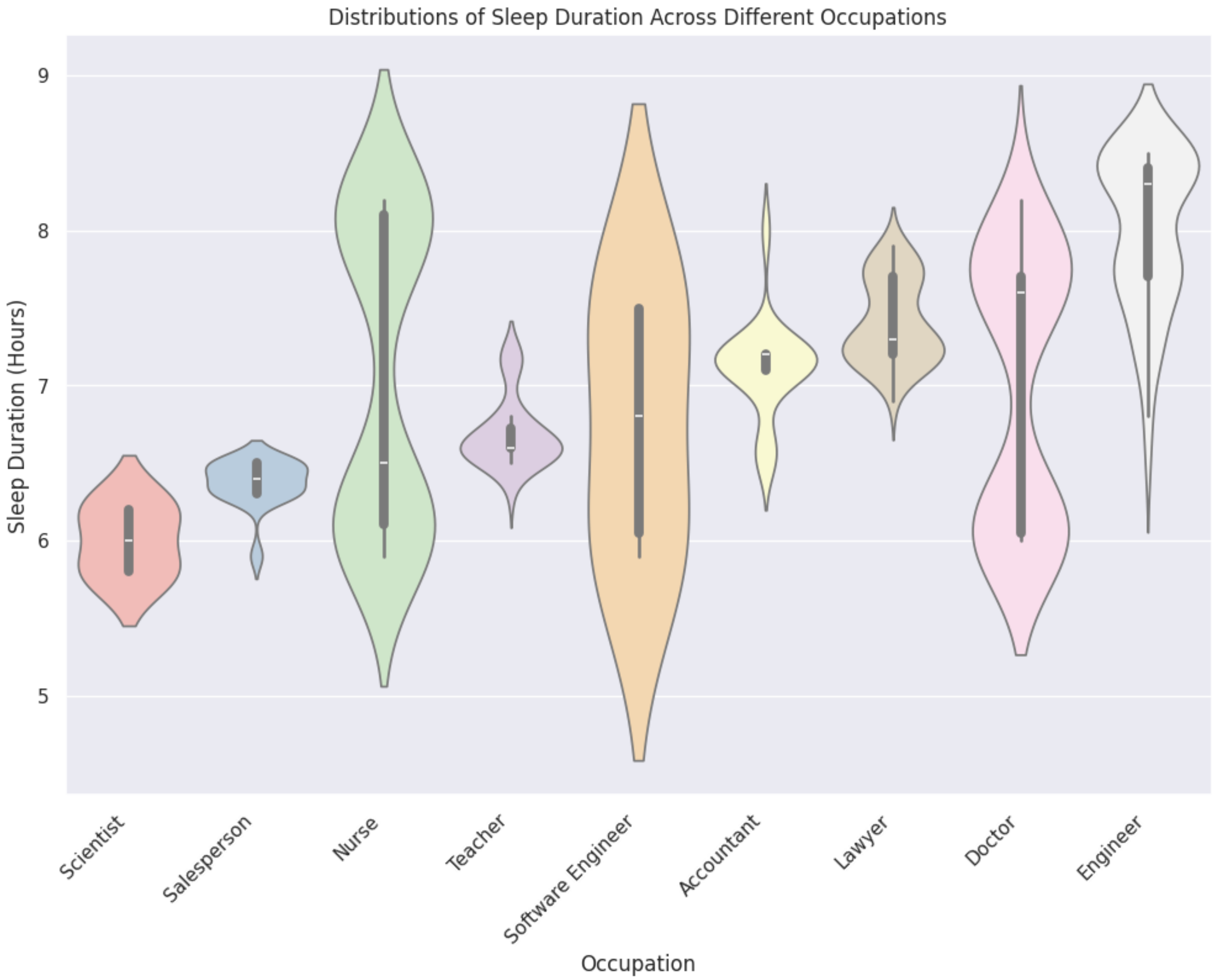


Doing a coefficient value plot of the features of our refined model shows that encoded BMI, or the BMI status of a person, was the most important feature in determining if someone had a sleep disorder or not. None of the other features compare in comparison to BMI values in terms of their feature importance. Looking at this, we can say that people with higher BMIs are much more likely to have sleep disorders.

# Can we predict a person's occupation based on their quality of sleep, duration of sleep, stress levels, etc.?

## Exploratory Data Analysis

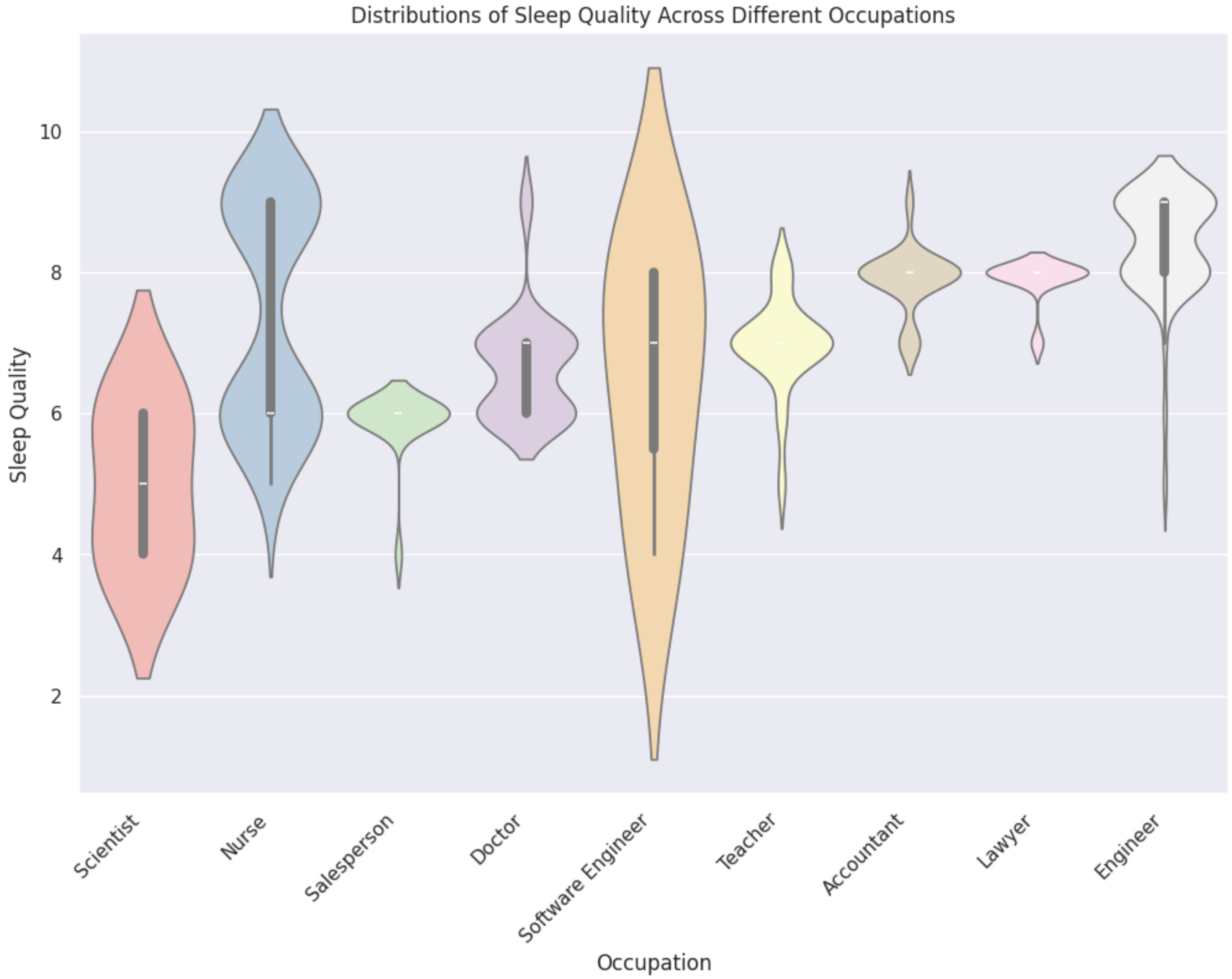
```
/tmp/ipykernel_46/3041315041.py:13: FutureWarning: 

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. 
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.violinplot(x='Occupation', y='Sleep Duration', data=filtered_sleep_data, palette
='Pastel1', order=ordered_occupations)
```



I created a violin plot to analyze the distributions of sleep duration across the different occupations. Violin plots include a box plot, which lets us recognize sumamry statistics such as maximum, minimum, IQR, and median. These plots also include the pdf of the data (probability density funciton), which allows us to recognize the distibutions. From this plot we can discern that scientists have the shortest avergae sleep duration and engineers have the longest average sleep duration. Nurses, software engineers, and doctors have the widest range of sleep durations with nurses and doctors having bimodal data sets. This is an indication that nurses and doctors might have 2 subpopulations, the subpopulations also look to resemble each other (one centered around 6 hours, and another centered around 8 hours).Software engineers have a more uniform data set and accountants look to have a fairly normal data set of sleep duration.

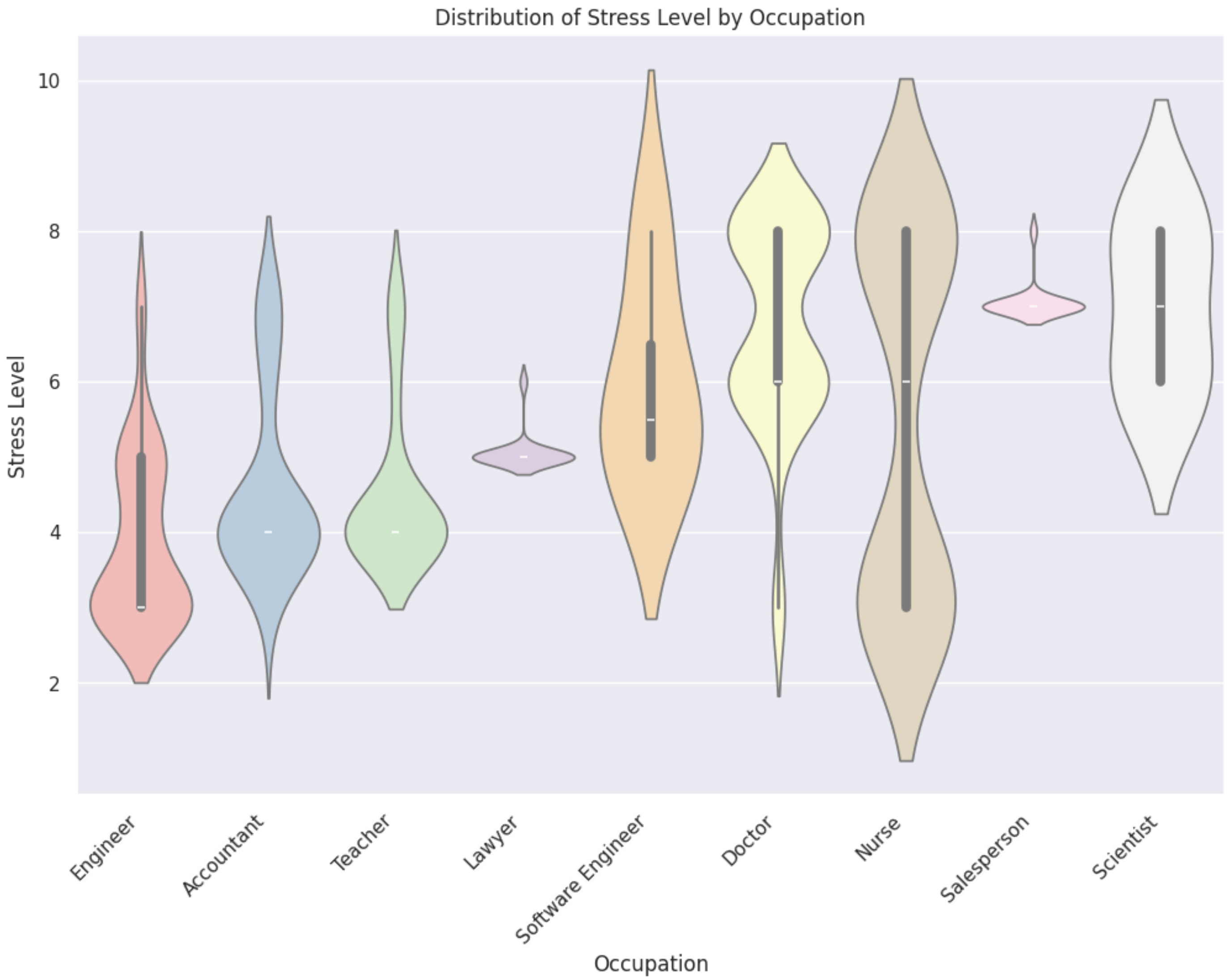
Manager and sales representative were excluded from the analysis as there are not enough observations for these occupations for a meaningful analysis to be made.

```
/tmp/ipykernel_46/1382236515.py:8: FutureWarning: 

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.violinplot(x='Occupation', y='Quality of Sleep', data=filtered_sleep_data, palette
='Pastel1', order = ordered_occ)
```

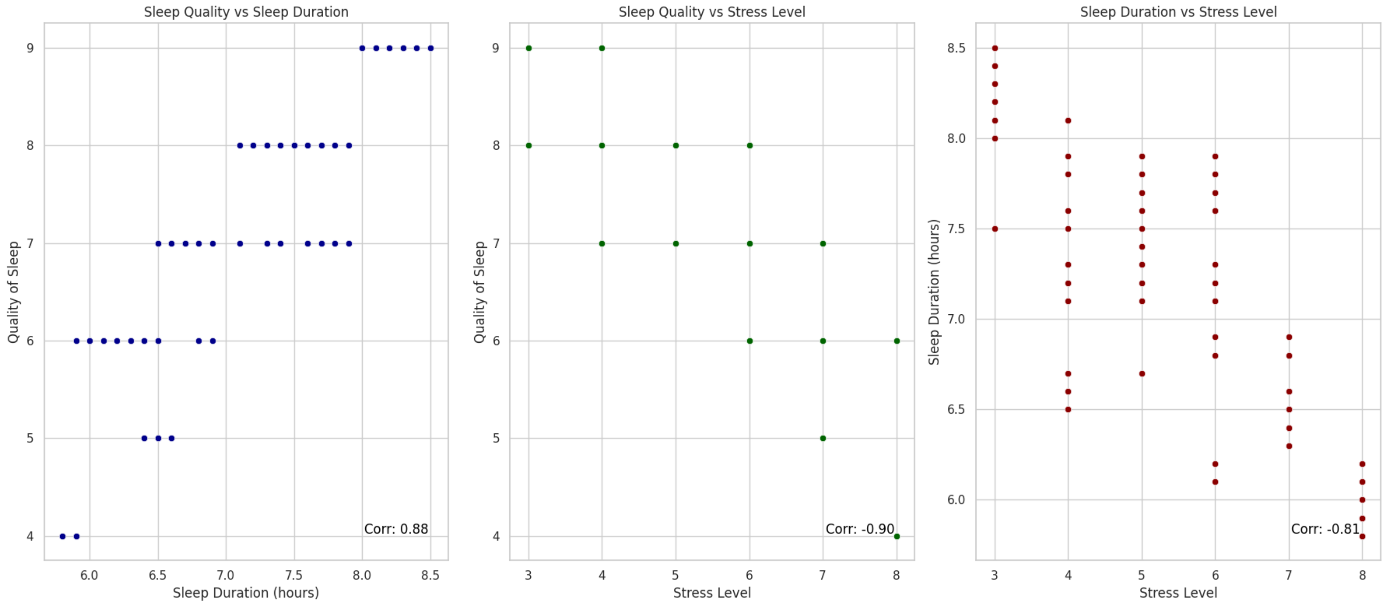


We used another violin plot to show the difference in distributions between sleep quality across different occupations. Scientists and engineers again have the lowest and highest medians. Scientists have a uniform distribution, nurses have a bimodal distribution, and accountants and lawyers have fairly normal distributions. From this plot we can also see that all salespeople reported a quality of sleep score of 6.

```
/tmp/ipykernel_46/4196359679.py:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.
Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.violinplot(x='Occupation', y='Stress Level', data=filtered_sleep_data, palette='Pa
stel1', order = ordered_stress)
```



This violin plot looks at the relationship betwen occupation and stress level. Engineers have the lowest median stress level and scientists ahve the highest reported stress level, this is the opposite of what was found for sleep quality and sleep duration. It may be interesting to look at if stress levels and sleep quality/duration are negatively correlated. Accountants and teachers seem to have skewed normal distributions, both centered around a reported stress level of 4/10. Nurses and doctors have bimodal distributions again, again a possible indication of subpoopulations in the data. Nurses looking to have the widest spread of data. It also can be seen that all salespeople reported a stress level of 7/10.



These three scatter plots examine the relationships between Sleep Quality vs. Sleep Duration, Sleep Quality vs. Stress Level, and Sleep Duration vs. Stress Level. For the Sleep Quality vs. Sleep Duration the plot as well as a correlation coefficient of 0.88 indicates a strong positive correlation. The plot Sleep Quality vs. Stress Level and its correlation coefficient of -0.90 indicates a strong negative correlation. And the plot Sleep Duration vs. Stress Level and its correlation coefficient of -0.81 indicates a strong negative correlation. This means as sleep duration increases, quality of sleep is expected to deacrese; and as stress level increases, quality of sleep and sleep duration are expected to decrease. This confirms suspicious we had about these relationships from the violin plots.

Out[27]:

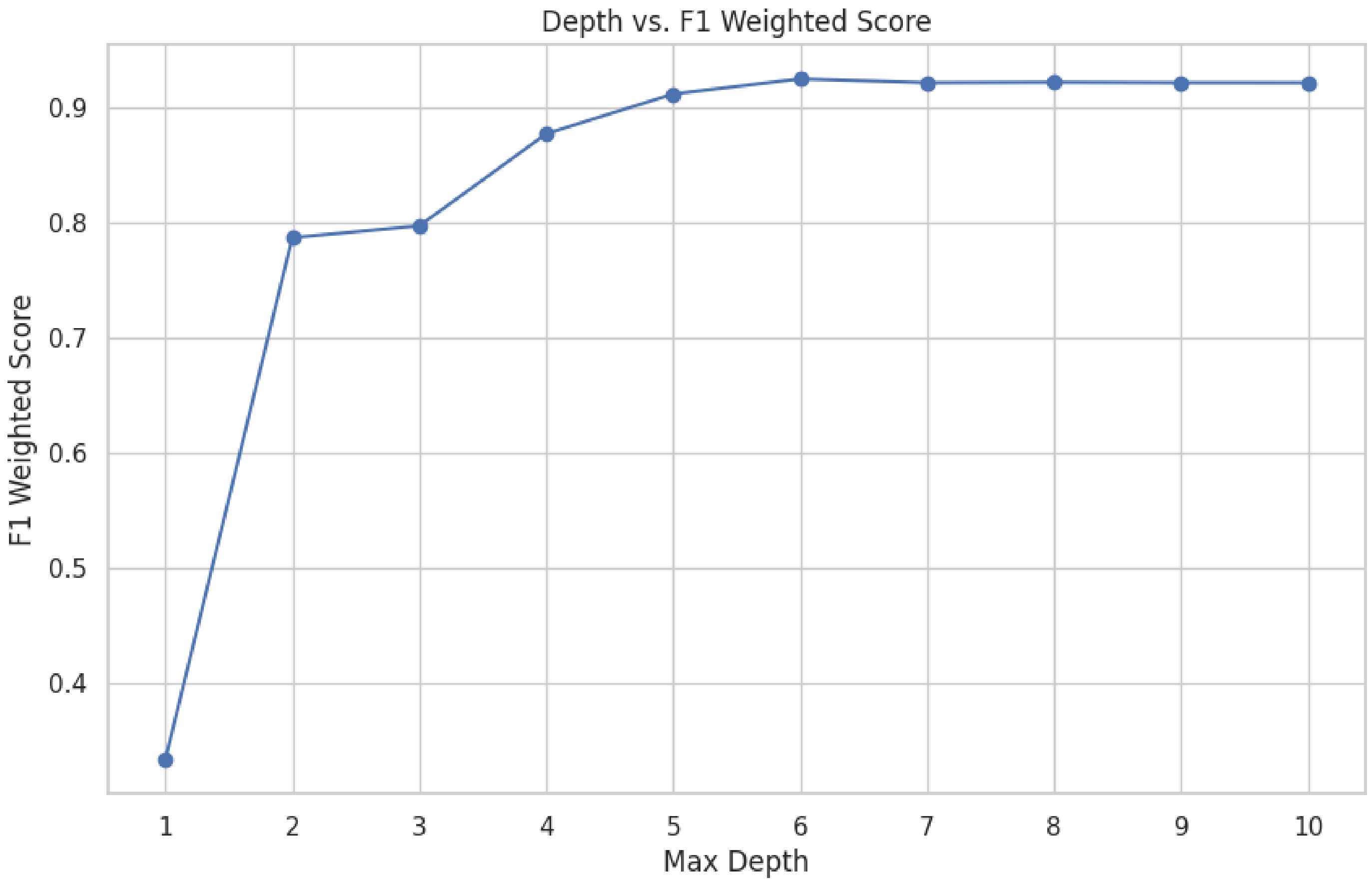
| | Occupation | Average Sleep Duration | Average Quality of Sleep | Average Age | Average Stress Level | Average Pulse Pressure | Average Heart Rate | Average Physical Activity Level | Average BMI | Average Daily Steps |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Accountant | 7.11 | 7.89 | 39.62 | 4.59 | 40.81 | 68.86 | 58.11 | 0.16 | 6881.08 |
| **1** | Doctor | 6.97 | 6.65 | 32.68 | 6.73 | 42.49 | 71.52 | 55.35 | 0.11 | 6808.45 |
| **2** | Engineer | 7.99 | 8.41 | 46.59 | 3.89 | 44.52 | 67.19 | 51.86 | 0.05 | 5980.95 |
| **3** | Lawyer | 7.41 | 7.89 | 39.43 | 5.06 | 44.96 | 69.64 | 70.43 | 0.13 | 7661.70 |
| **4** | Nurse | 7.06 | 7.37 | 51.79 | 5.55 | 44.79 | 72.00 | 78.59 | 0.90 | 8057.53 |
| **5** | Salesperson | 6.37 | 5.88 | 42.62 | 7.06 | 45.29 | 72.76 | 44.12 | 1.06 | 5823.53 |
| **6** | Scientist | 6.00 | 5.00 | 33.50 | 7.00 | 44.00 | 78.50 | 41.00 | 1.00 | 5350.00 |
| **7** | Software Engineer | 6.75 | 6.50 | 31.25 | 6.00 | 43.25 | 75.50 | 48.00 | 0.75 | 5800.00 |
| **8** | Teacher | 6.69 | 6.98 | 41.72 | 4.53 | 44.32 | 67.22 | 45.62 | 0.88 | 5957.50 |

The table above shows all average statistics across the occupations. Again, manager and sales representative were excluded from the analysis as there are not enough observations for these occupations for a meaningful analysis to be made.

# We will now create a machine learning model to predict a person's occupation

I chose to use a Random Forest to model the relationship between occupation and the features of this dataset. Random forests are a collection of decision trees, and decision trees are a type of machine learning model made to imitate human thinking which makes them easy to interpret. Decision trees make predictions based on answers to a specified set of questions. While decsion trees are good for multiclassification problems, such as predicting occupations based on certain features, they are prone to overfitting. We use a collection of decision trees, a random forest, becuase the predictions in these models will generalize to the data much better than a single 'tree'.

```
/usr/local/lib/python3.8/site-packages/sklearn/model_selection/_split.py:700: UserWarnin
g: The least populated class in y has only 1 members, which is less than n_splits=5.
  warnings.warn(
Best Parameters: {'max_depth': 6, 'n_estimators': 100, 'random_state': 42}
Best F1 Weighted Score: 0.9255981649141173
```



```
[0.33296039 0.78732005 0.79762318 0.87778766 0.91248366 0.92559816
 0.92231853 0.92267911 0.92231853 0.922259  ]
```

The code above preforms hyperparameter tuning using grid search and cross validation to select the optimal depth of our random forest model. Hyperparameter tuning is the process of finding the best set of hypeparameters (external configuration settings that cannot be learned from the data) for a machine learning model. These parameters control aspects of the learnign process such as complexity, capacity, or optimization strategy. Maximum depth of a decision tree is an example of a hyperparameter.

Grid search is a technique used for hyperparameter tuning that iterates through all combinations of hyperparameters defined in the parameter grid and selects the best combination based on the specified scoring metric ('f1_weighted') anc cross validation ('cv=5').

A cross validation of 5 means that the dataset is divided into 5 equal-sized folds. The model was trained 5 times, each time using 4 folds for training and 1 fold for validation. Cross validation helps to estimate the model's performance more accurately that a single train-test split by reducing the variability in performance metrics. 'f1_weighted' refers to the weighted average of F1 score for each occupation in the multi-class classification. F1 score is the harmonic mean of precision and recall, the weighted version takes into account class imbalances by computing the average weighted by the number of samples for each occupation.

I split the data into training and testing sets using train_test_split from sklearn.model_selection. The testing set size is set to 20%, and a random state of 42 is fixed for reproducibility.

The plot above provides insights into how the performance of the random forest classifier changes with different depths. The mean test scores obtained during cross-validation for each depth were also printed. From this we can see that the model's performance imporves with increasing depth up to a depth of 6 and the starts to decline after that. Because of this, we will choose a maximum depth of 6 for our random forest model in order to maximise the F1 score on the validation set while avoiding overfitting.

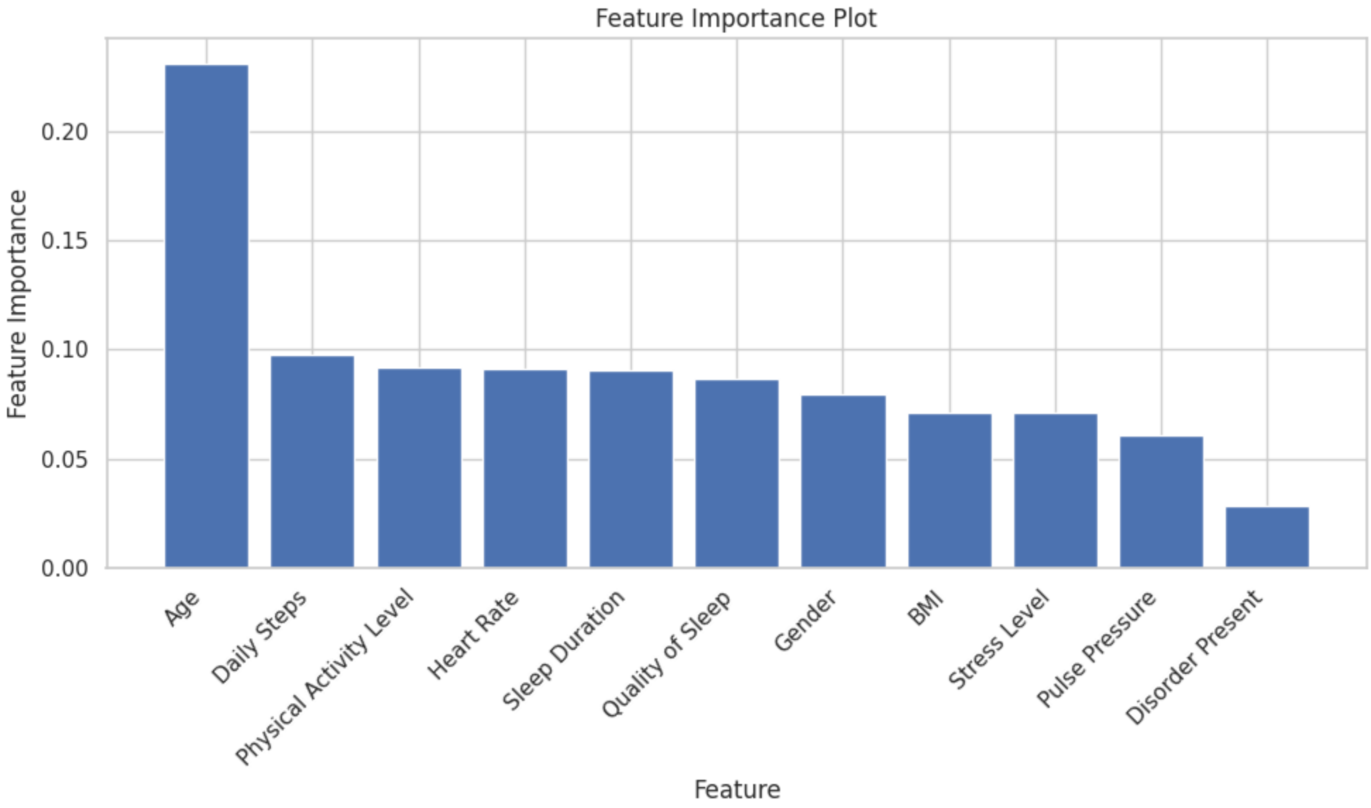
```
Accuracy:  0.92

Classification Report:
                   precision    recall  f1-score   support

       Accountant       0.80      0.80      0.80         5
           Doctor       0.95      1.00      0.97        18
         Engineer       1.00      1.00      1.00        10
           Lawyer       1.00      0.82      0.90        11
            Nurse       1.00      1.00      1.00        12
      Salesperson       0.92      1.00      0.96        11
        Scientist       0.00      0.00      0.00         0
Software Engineer       0.00      0.00      0.00         2
          Teacher       0.71      0.83      0.77         6

         accuracy                           0.92        75
        macro avg       0.71      0.72      0.71        75
     weighted avg       0.91      0.92      0.91        75


Confusion Matrix:
[[ 4  0  0  0  0  0  0  0  1]
 [ 0 18  0  0  0  0  0  0  0]
 [ 0  0 10  0  0  0  0  0  0]
 [ 0  1  0  9  0  0  0  0  1]
 [ 0  0  0  0 12  0  0  0  0]
 [ 0  0  0  0  0 11  0  0  0]
 [ 0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  1  1  0  0]
 [ 1  0  0  0  0  0  0  0  5]]
```

```
/usr/local/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1344: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels wit
h no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1344: Undefine
dMetricWarning: Recall and F-score are ill-defined and being set to 0.0 in labels with n
o true samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1344: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels wit
h no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1344: Undefine
dMetricWarning: Recall and F-score are ill-defined and being set to 0.0 in labels with n
o true samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1344: Undefine
dMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels wit
h no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.8/site-packages/sklearn/metrics/_classification.py:1344: Undefine
dMetricWarning: Recall and F-score are ill-defined and being set to 0.0 in labels with n
o true samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

In the code above, we have refined our random forest classifier model by setting the maximum depth of the decision tress to 6. We installed the random forest classifier with 'rf_model' this random forest has 100 trees. We then trained the model on the training data using the 'fit' method. The trained model was used to make predictions ('rf_y_pred') on the testing data.

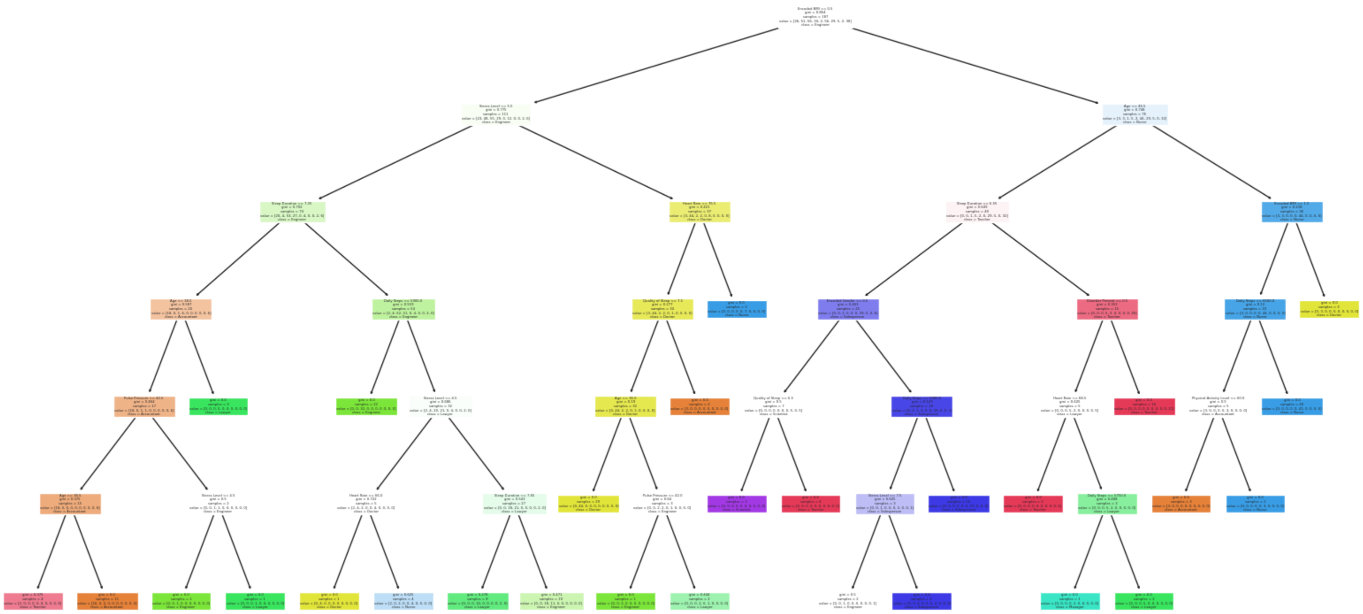
The accuracy of the model, classification report, and confusion matrix all provide insights into the model's performance (precision, recall, F1 score, and confusion between predicted and true labels). Precision is the probability that an observation classified as a 1 is actually a 1. Recall is the probability of an observation in class 1 actually being predicted as class 1. F1 score is the harmonic mean of precision and recall.

The model has an accuracy of 92% which is great! The classificaiton report indicates that the model preforms well for most classes, with high (and in some cases perfect) precision, recall, and F1 scores for several occupations. However, there are some classes with lower performance such as 'Sales Representative' and 'Software Engineer', this may be due to a small number of data points of these classifications in this test set.



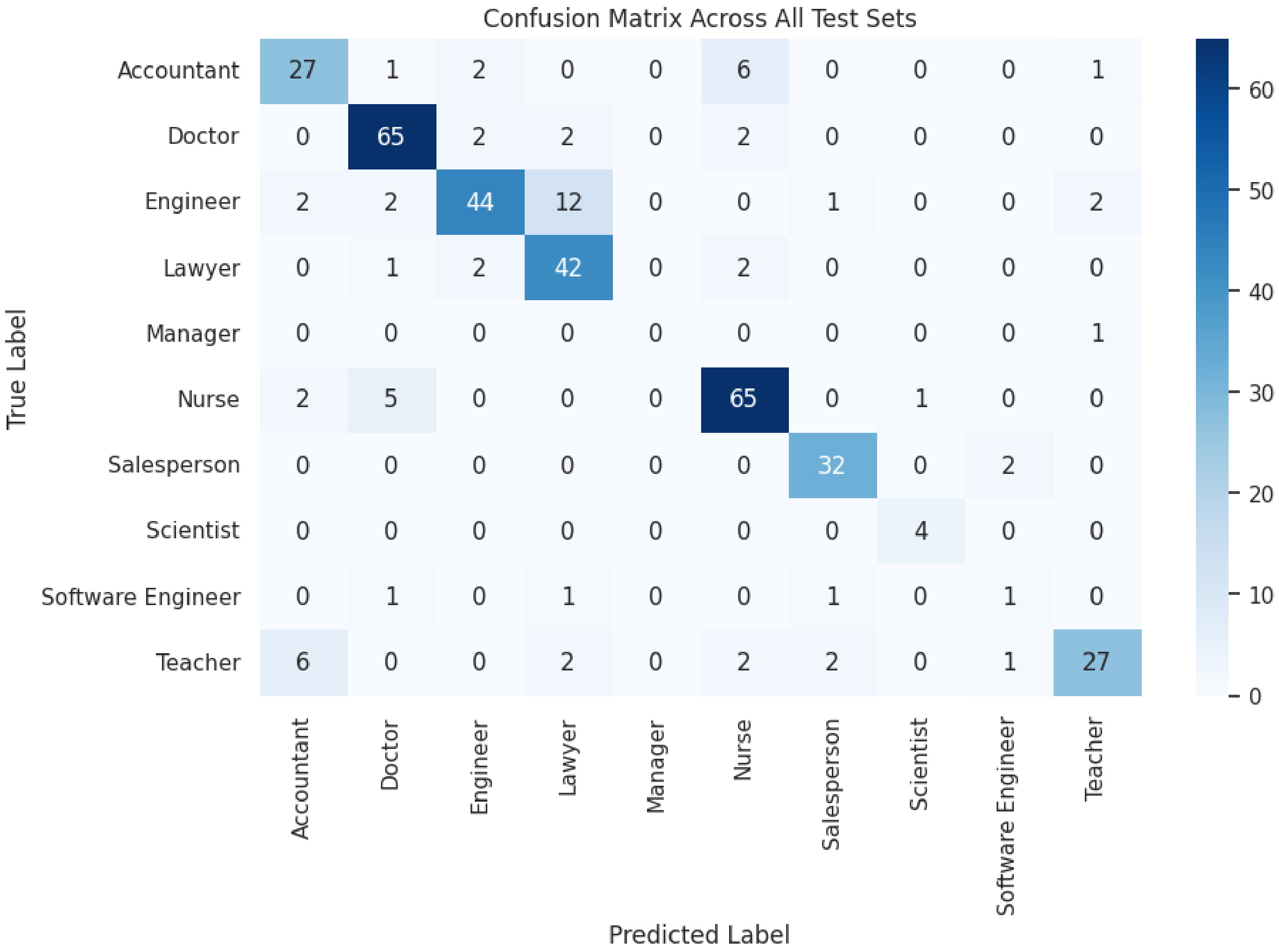
An importance score refers to the relative contribution of each feature (gender, age, sleep duration, quality of sleep, etc.) to the model's ability to predict the target variable, which in this case is occupation. In the sense of a Random Forest model, importance scores are found based on how much each feature decreases the impurity at each node of the decision trees in the forest. Gini Impurity is used in decision tree models to determine how the features should split nodes in order to form the trees. The Gini Impurity is a number 0-0.5, this indicates the likelihood of a randomly chosen element from the dataset would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

In other words, importance scores generated by the Random Forest model quantify the influence of each feature on predicting occupations based on sleep-related factors. From this plot we can see that age seems play a large role in determining a person's occupation.



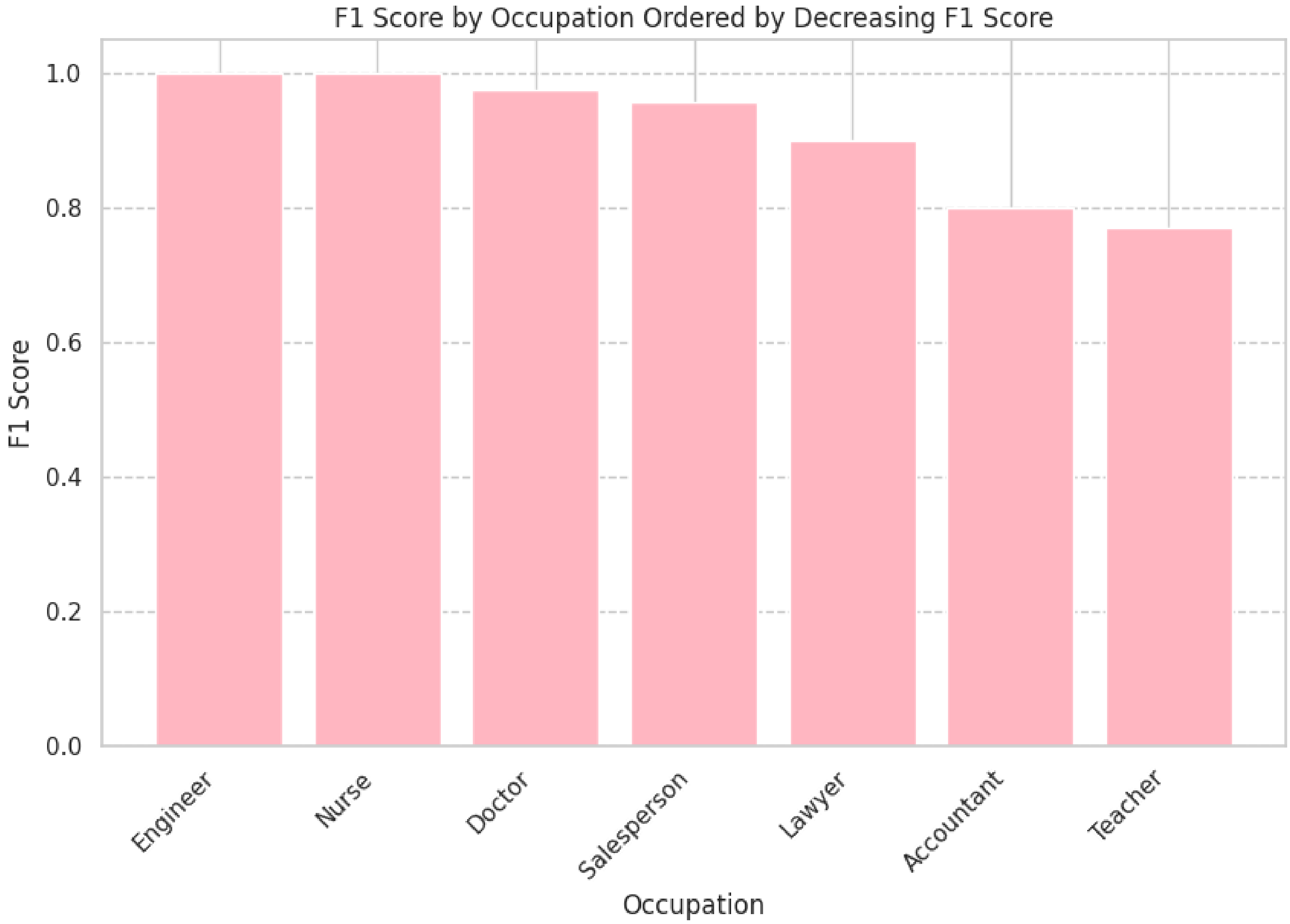
The code block above visualizes a single decision tree from our random forest classifier. This allows for a deeper understanding of the desicion making process within the collection of trees. Zooming in on this diagram, we can see that features with high feature importance (from the previous feature importance plot) such as age, quality of sleep, pulse pressure, stress level, and sleep duration are highly involved in this tree's decision making process.

```
/usr/local/lib/python3.8/site-packages/sklearn/model_selection/_split.py:700: UserWarnin
g: The least populated class in y has only 1 members, which is less than n_splits=5.
  warnings.warn(
```



| True Label \ Predicted Label | Accountant | Doctor | Engineer | Lawyer | Manager | Nurse | Salesperson | Scientist | Software Engineer | Teacher |
|---|---|---|---|---|---|---|---|---|---|---|
| Accountant | 27 | 1 | 2 | 0 | 0 | 6 | 0 | 0 | 0 | 1 |
| Doctor | 0 | 65 | 2 | 2 | 0 | 2 | 0 | 0 | 0 | 0 |
| Engineer | 2 | 2 | 44 | 12 | 0 | 0 | 1 | 0 | 0 | 2 |
| Lawyer | 0 | 1 | 2 | 42 | 0 | 2 | 0 | 0 | 0 | 0 |
| Manager | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Nurse | 2 | 5 | 0 | 0 | 0 | 65 | 0 | 1 | 0 | 0 |
| Salesperson | 0 | 0 | 0 | 0 | 0 | 0 | 32 | 0 | 2 | 0 |
| Scientist | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 |
| Software Engineer | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| Teacher | 6 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 1 | 27 |

The code block above visualizes the confusion matrix across all folds of cross-validation for our random forest model. This allows us to assess the model's performance across different subsets of the data. The diagonal elements of the confusion matrix represent correct predictions, where the true occupation matches to the predicted occupation. The value of 25 in cell (1,1) means that there were 25 instances where an accountant was correctly predicted as an accountant, 65 in cell (2,2) means that there were 65 instances where a doctor was correctly predicted to be a doctor, and so on. The off-diagonal elements represent false predictions. For example, there were 12 instances (cell (4,3)) where an engineer was wrongly predicted to be a lawyer.



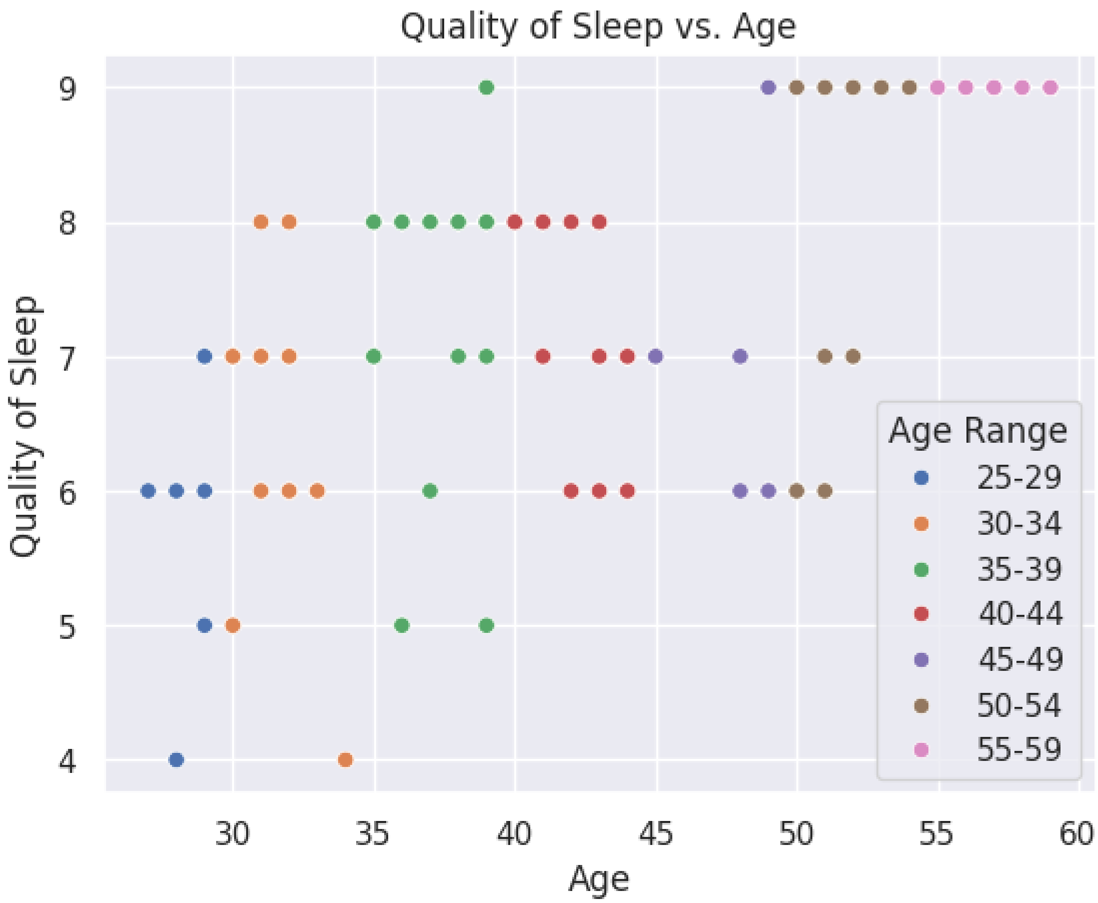
An F1 score is a measure that balances precision and recall, this provides a single measure of a model's accuracy. The model is doing a great job at predicting most occupations, specifically engineers, nurses, and sales people as these all have an F1 score of 1. The model is struggling to predict sales represntatives the most. These F1 scores may be influenced by the imbalance of persons in each occupation class.

Overall, the random forest model does a good job classifing a person's occupation based on their cardiovascular and lifestyle factors.

# Does Duration or Quality of sleep correlate with age?

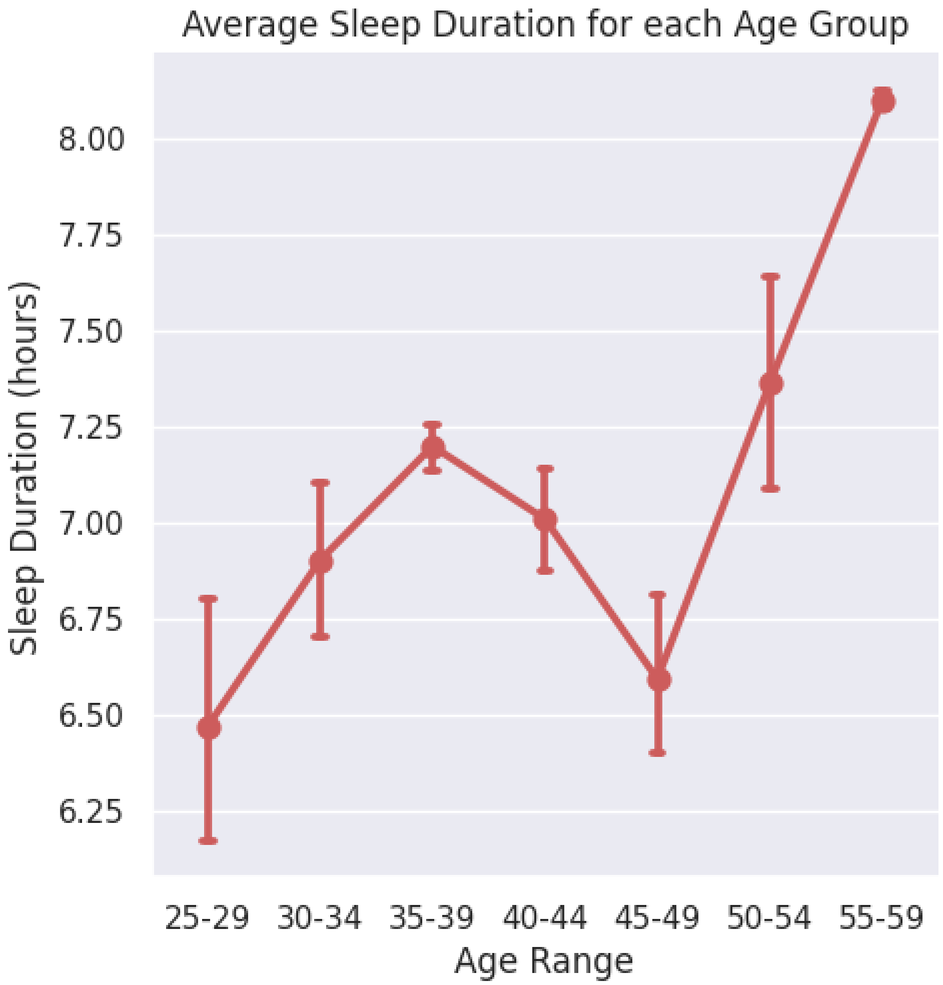
As noted earlier, age is one of the most important features when determining quality of sleep, so we want to look into that relationship further.

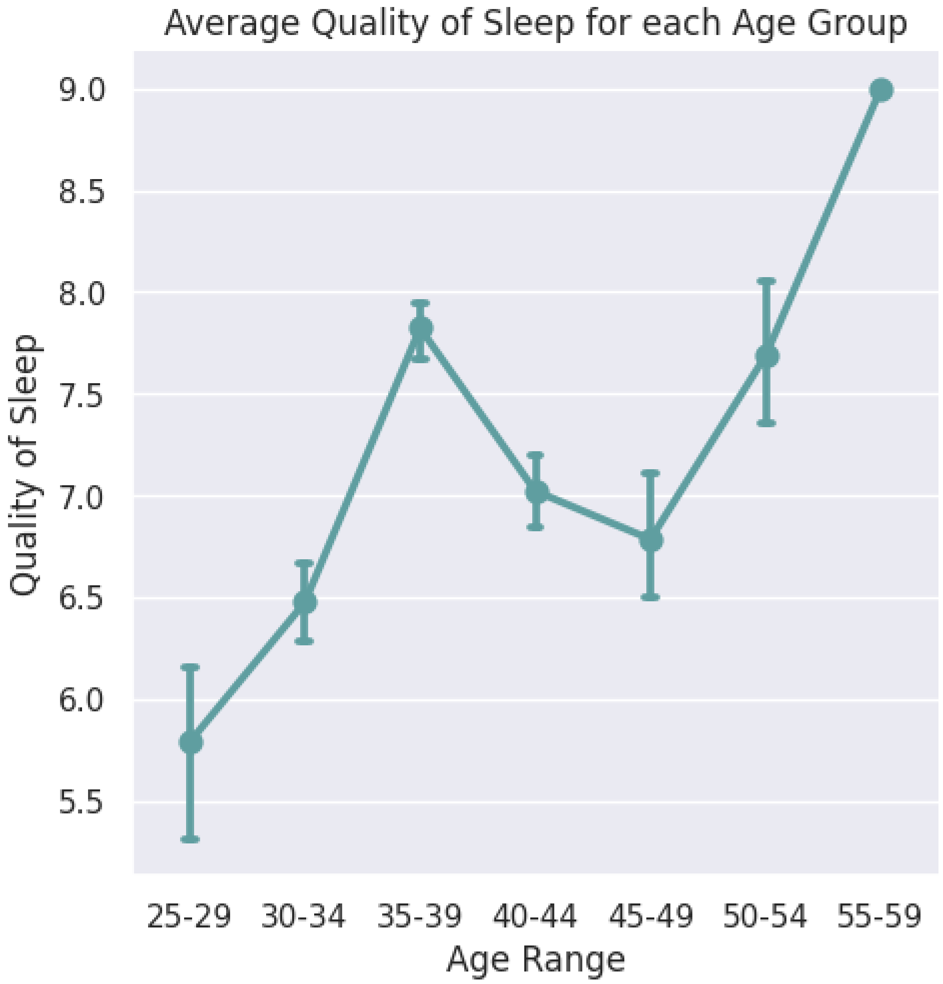
```
Out[34]: [Text(0.5, 1.0, 'Quality of Sleep vs. Age')]
```



As we can see there is a weak positive correlation between Age and Quality of Sleep in this dataset. Notably, all the pink points (age range 55-59) have quality of sleep values of 9. In addition the minimum quality of sleep values are in the age ranges of 25-35. This suggests that, while weakly, that as age increases so too does quality of sleep.

```
Out[35]: <seaborn.axisgrid.FacetGrid at 0x7f66bc888bb0>
```

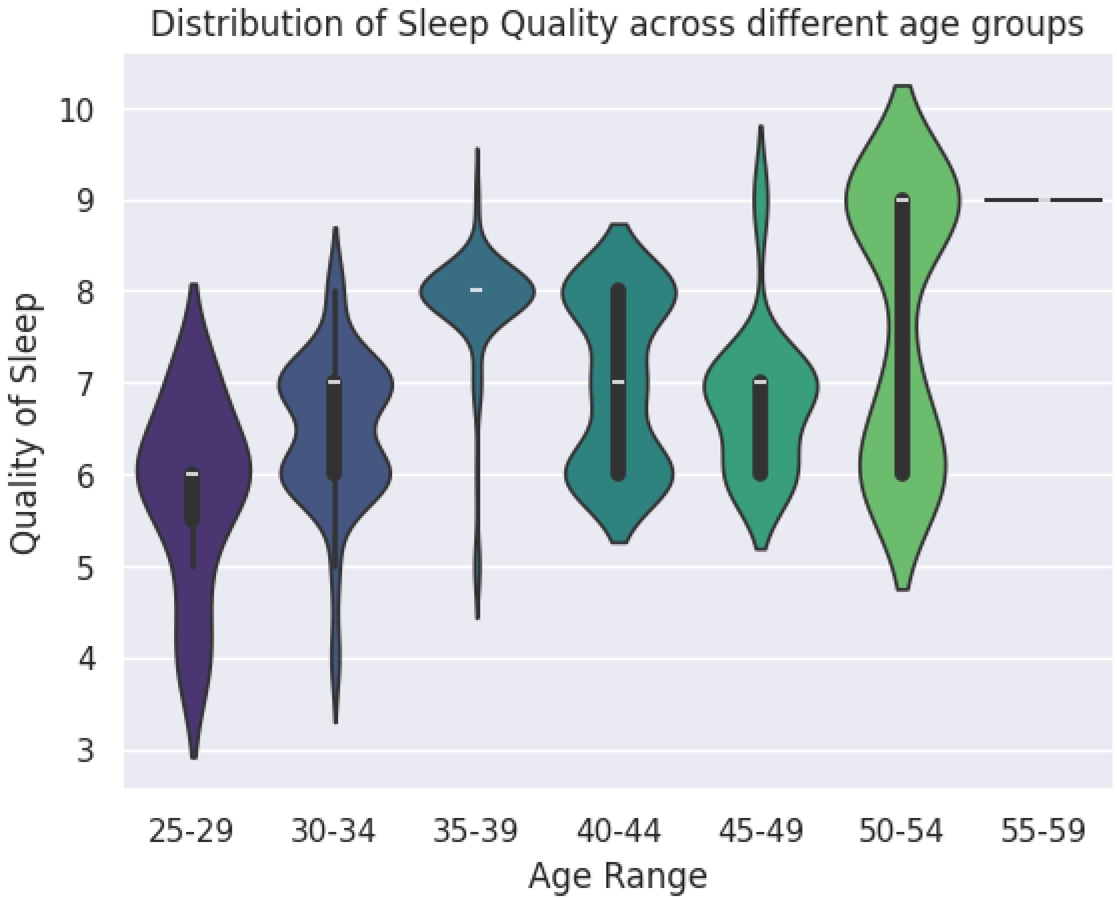




In order to create these graphs, I sorted the age data into bins with 5 year ranges, from 25 to 59. Then I used point plots to create a good way to compare the mean of each of these age groups' quality of sleep and average sleep duration. Both graphs follow a similar trend, where the lowest quality of sleep and skeep duration is generally on the younger side while the highest is on the older side. Curiously, while these graphs follow similar shapes, despite 30-34 year olds having a higher duration of sleep on average than 45-49 year olds, they have on average a lower quality of sleep by almost 0.3. While these means are interesting, we can use other visualizations to see more about these distributions beyond the mean.

```
/tmp/ipykernel_46/3977048551.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  k = sns.violinplot(y=sleep_data_age['Quality of Sleep'],x=sleep_data_age['Age_Range'], palette='viridis')
```

```
Out[36]: [Text(0.5, 1.0, 'Distribution of Sleep Quality across different age groups'),
          Text(0.5, 0, 'Age Range'),
          Text(0, 0.5, 'Quality of Sleep')]
```

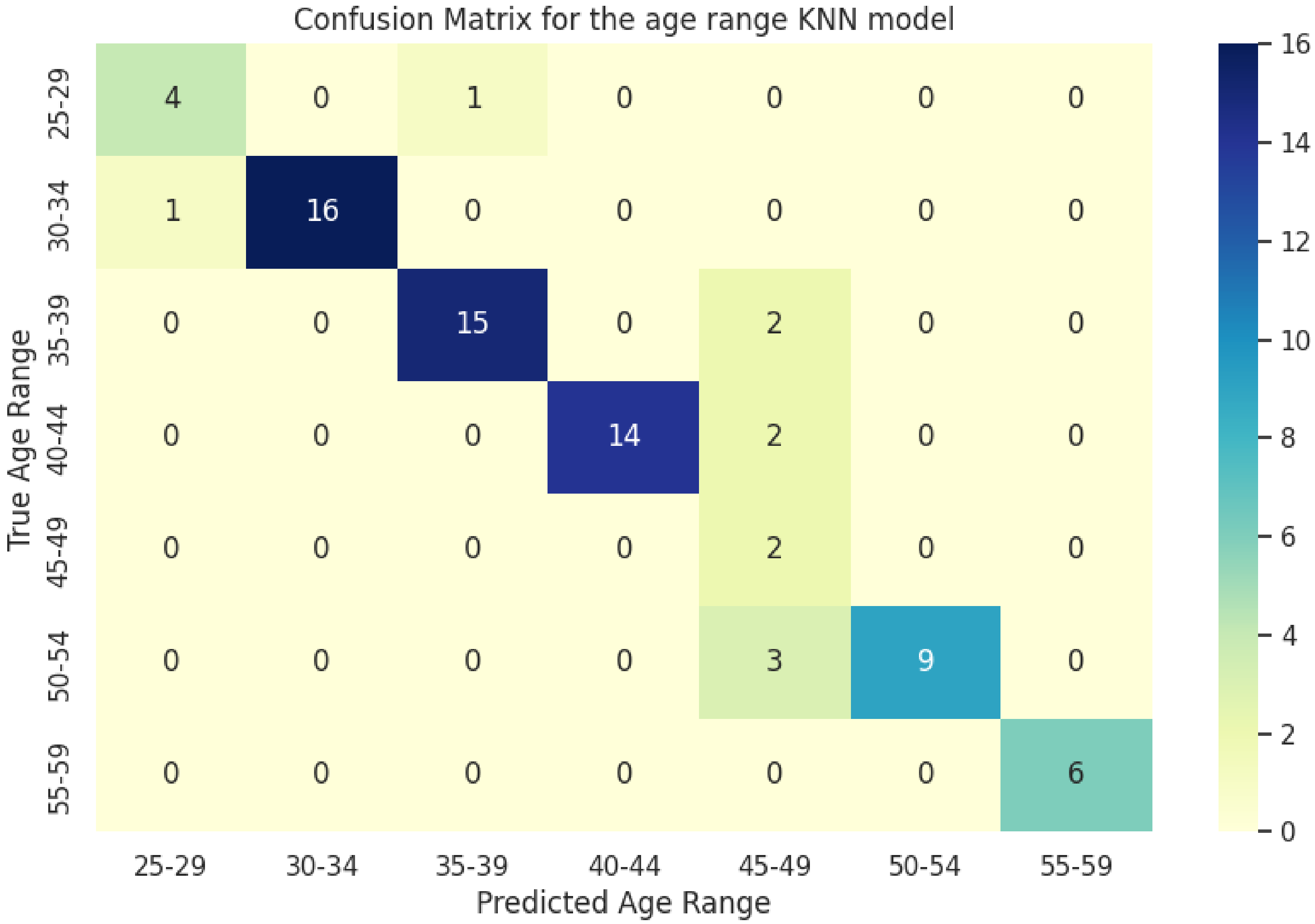


Looking at these plots, it seems that the 50-54 age range has the largest range in values, although 40-44 has the highest IQR. In addition, the 50-54 and 40-44 age ranges look almost bimodal. This implies that a different factor could be influencing the quality of sleep within each age group. For instance, being an accountant may make 40-44 year olds have lower quality of sleep. Another notable thing from this graph is that 55-59 year olds did not have any variance, they all put a 9 as their quality of sleep.

In order to predict the age of someone, I decided to use the K nearest neighbors method. This is because it is reasonable to assume that people of similar ages have similar health and lifestyles. I decided not to use random forest models or gaussian naive bayes because it isn't reasonable to assume every single variable (for example step count and BMI) is independent. Finally, logistic regression would not be effective either not only because the independent variables could be correlated, but also because the response is not binary. For this k nearest neighbors analysis, I had to change the independent categorical variables gender, occupation, BMI. In addition I combined the pulses into the pulse pressure variable. I did this because the k nearest neighbors needs quantitative variables to predict boundaries between each prediction region. I decided to test k values from 1 to 20, and had 5 folds for the cross validation, as is standard for k nearest neighbors.

```
Best n_neighbors: 1
Best score: 0.8664971751412 43
```

As we can see, the k nearest neighbors model suggest 1 nearest neighbor. The low number of neighbors indicates a high level variance in predictions. This is because a lower number of neighbors means the predictive model is sensitive to very small changes in data values. In the real world, this means that each person in each age group holds different values for the independent variables, since it means there are no larger trends within age groups to create decision boundaries around. In essence, there may be a low correlation between health/sleep and age. The accuracy for this model is 86%, which is good, however we can look further into this accuracy, as well as questions raised by the low number of neighbors, using a confusion matrix.



This confusion matrix shows the difference between predicted and true values in the test set for the age range model. The diagonal is the correct predictions, for instance 30-34 was correctly predicted 16 times. From this confusion matrix we can see that 45-49 was predicted much more by the predictive model than in the test set. This could be because of a small sample size for 45-49 year olds in original data. In addition, the fact that predicted ages from 35 to 54 to all be 45-59 suggest that these age ranges hold similarities in terms of health and sleep attributes. This would imply that there may be a low correlation between age and health/sleep in this age range. However, in contrast the model rarely predicted incorrectly for ages from 25 to 34 or 55-59. This suggests that middle aged people specifically may share quite similar health and sleep conditions, however outside of that age range, one's age is tied to health/sleep. This would explain the errors in the confusion matrix, low number of neighbors in the knn model while also accounting for the large feature importance age had in the previous predictive model for quality of sleep.

```
              precision    recall  f1-score   support

           0       0.80      0.80      0.80         5
           1       1.00      0.94      0.97        17
           2       0.94      0.88      0.91        17
           3       1.00      0.88      0.93        16
           4       0.22      1.00      0.36         2
           5       1.00      0.75      0.86        12
           6       1.00      1.00      1.00         6

    accuracy                           0.88        75
   macro avg       0.85      0.89      0.83        75
weighted avg       0.95      0.88      0.91        75
```

As we can see in this classification report, the fourth response group (45-49) has the lowest f score of 0.36, which is significantly lower than any other response group. Notably, both those on the youngest end (response groups 1-2) and the oldest end (5-6) had incredibly high f scores. This aligns with what we see in the confusion matrix, supporting our idea that middle aged folk specifically share similar health conditions while those on the extremes of the age range have health traits that correlate to their age in specific. It should also be noted that the 45-49 age group has significantly fewer data points in the test group, which could also explain the high variability in predictions.

# Conclusion

From our exploration of the Sleep Health and Lifestyle dataset, as well as the machine learning models we built using logistic regression, random forest, and k-nearest-neighbors models, we were able to come to a variety of conclusions. The first being that a logistic regression model is able to predict if someone has a sleep disorder, as well as specifically if someone has sleep apnea or insomnia, with high accuracy. BMI was found to play a large role in determining a if a person has a sleep disorder. The random forest model was able to predict a person's occupation based on cardiovascular and lifestyle factors with high accuracy as well. Age, quality of sleep, and pulse pressure were significant features in making this classification. Building on the relationship between age and quality of sleep, we analyzed the relationship between the two. We found that while weak, a correlation does exist, with quality of sleep increasing with age, however quality of sleep was quite similar for middle aged people. This was confirmed in the following classification model, as the most errors in classification occured in those middle ages. All in all, age was a factor in sleep and health in broad strokes, with only those on the younger and older extremas experiencing significantly different effects on sleep and health.

# Contribution Report

Overall, we believe our group worked well together to produce a cohesive and cohesive report. Nick Johannessen did a great job contributing to the team and communicated with the rest of the group effectively. His work on the data cleaning and preparation section was vital to the project as it allowed each of the other team members to create their visualizations and models with ease and without having to worry about dataset related errors. Nick's logistic regression to predict if someone has a sleeping disorder works smoothly and is a fantastic contribution to the project as a whole. Saketh worked on the section considering the relationship between Age and Health/Sleep. He worked to deepen the understanding in the relationship between these features, creating all the age related visualizations to analyze the trends and distributions for each age group. In addition, he created a K nearest neighbors model in order to see if health and sleep could predict age groups, and built on that using a confusion matrix and classification report. Finally, he created the docker image and github repository for the report. Alyssa worked on the section about if a person's occupation could be predicted based on their stress level, sleep quality, and sleep duration. She created amazing visualizations that helped tell the story and gave way to her machine learning model. Alyssa's random forest classification was effective in classifying people into their occupation based on our features. The work she did is phenomenal and was beneficial to our group and the project as a whole, and she did a good job working with the group, communicating, and asking Prof. Bruce for advice. Iraam worked on the introduction paragraph to set the scene for what our report would address. She also worked on the distribution histograms showcasing the associations between sleep disorders and cardiovascular health and lifestyle choices. The graphs of categorical factors also helped give a visual aide for comparison between variables and the prevalence of sleep disorders. Iraam's work on the logistic regression including the graphs and analysis helped to visualize what the data and calculations represented and make further conclusions on associations amongst the data.