OESON Project 2: Exploratory Data Analysis (EDA)

By: Christopher Gonzalez

Introduction

- **Exploratory Data Analysis is the process of analyzing a given dataset for significant patterns/anomalies present.**

- **Initial investigation is done to allow data scientists to then visualize the data in a more comprehensive way.**

- **This presentation will demonstrate the significance of EDA using a real-world scenario.**

Background of the Dataset

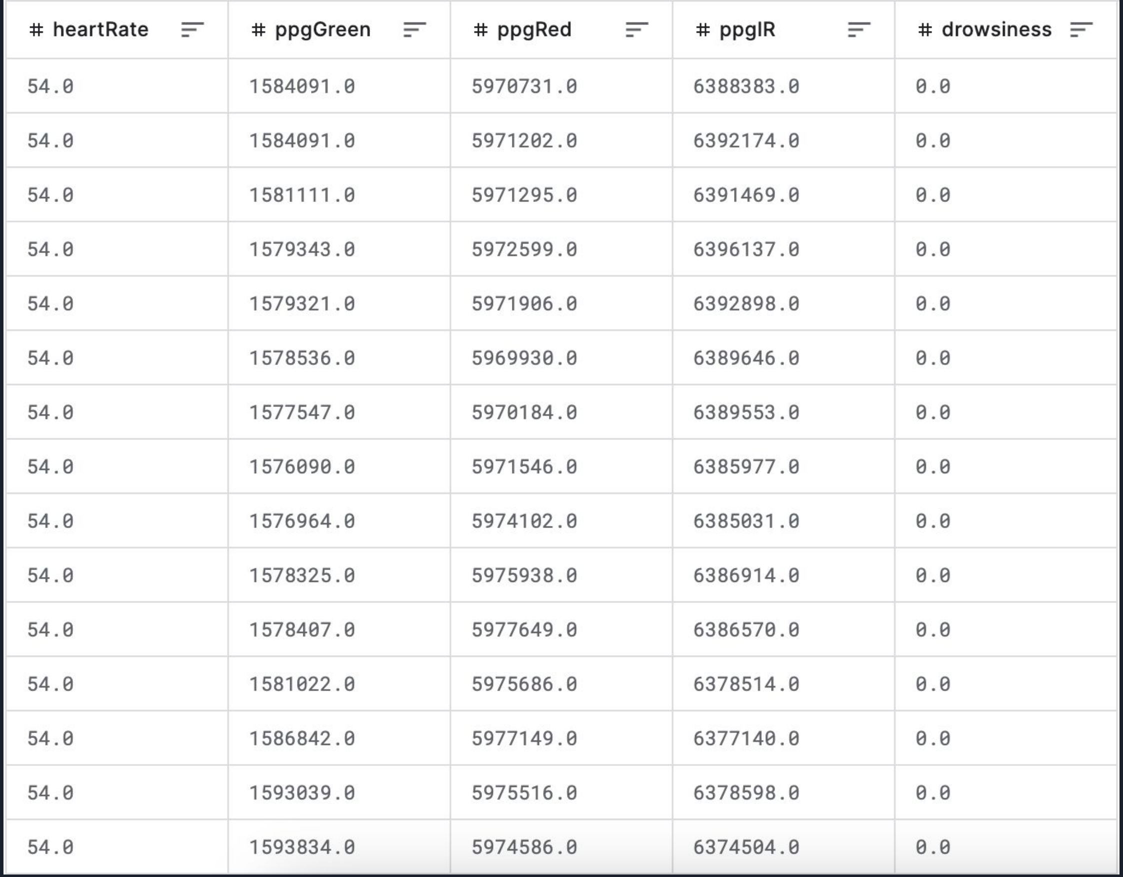
- **The dataset comes from a wearable technology company that produces smartwatches with vital sign sensors.**

- **These sensors monitor heart rate and Photoplethysmography (PPG) signals.**

- **The PPG signals include variations in green, red, and infrared light.**

- **A key feature of the smartwatches is its ability to detect/alert users of potential drowsiness based on this data.**

Portion of drowsiness_dataset.csv



# heartRate	# ppgGreen	# ppgRed	# ppgIR	# drowsiness
54.0	1584091.0	5970731.0	6388383.0	0.0
54.0	1584091.0	5971202.0	6392174.0	0.0
54.0	1581111.0	5971295.0	6391469.0	0.0
54.0	1579343.0	5972599.0	6396137.0	0.0
54.0	1579321.0	5971906.0	6392898.0	0.0
54.0	1578536.0	5969930.0	6389646.0	0.0
54.0	1577547.0	5970184.0	6389553.0	0.0
54.0	1576090.0	5971546.0	6385977.0	0.0
54.0	1576964.0	5974102.0	6385031.0	0.0
54.0	1578325.0	5975938.0	6386914.0	0.0
54.0	1578407.0	5977649.0	6386570.0	0.0
54.0	1581022.0	5975686.0	6378514.0	0.0
54.0	1586842.0	5977149.0	6377140.0	0.0
54.0	1593039.0	5975516.0	6378598.0	0.0
54.0	1593834.0	5974586.0	6374504.0	0.0

Additional Context for Dataset

- **The dataset contains a total of 4,890,260 entries collected from the smartwatches.**

- **Photoplethysmography (PPG) measures changes in blood volume per heartbeat.**

- **Level of drowsiness is based on an adapted Karolinska Sleepiness Scale (KSS).**

- **Drowsiness levels range from 0.0 - 2.0, where 0.0 represents alertness and 2.0 represents significant drowsiness.**

Understanding the dataset

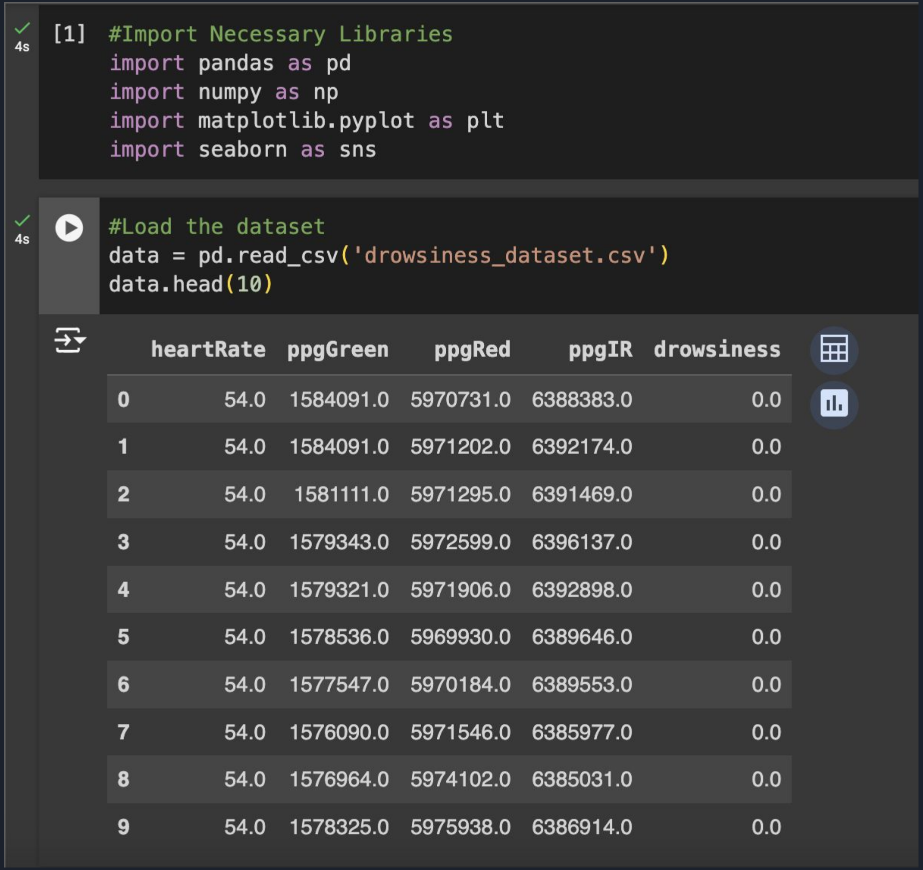
- There are a total of five columns present in the dataset (heartRate, ppgGreen, ppgRed, ppgIR, drowsiness)

- Recalling that the dataset has nearly five million entries, it is recommended to use python for additional information.

- To do this, we must import all necessary libraries and load the dataset first.

- The following slides exhibit python code used on the dataset and the additional information obtained.

Preparing the Data

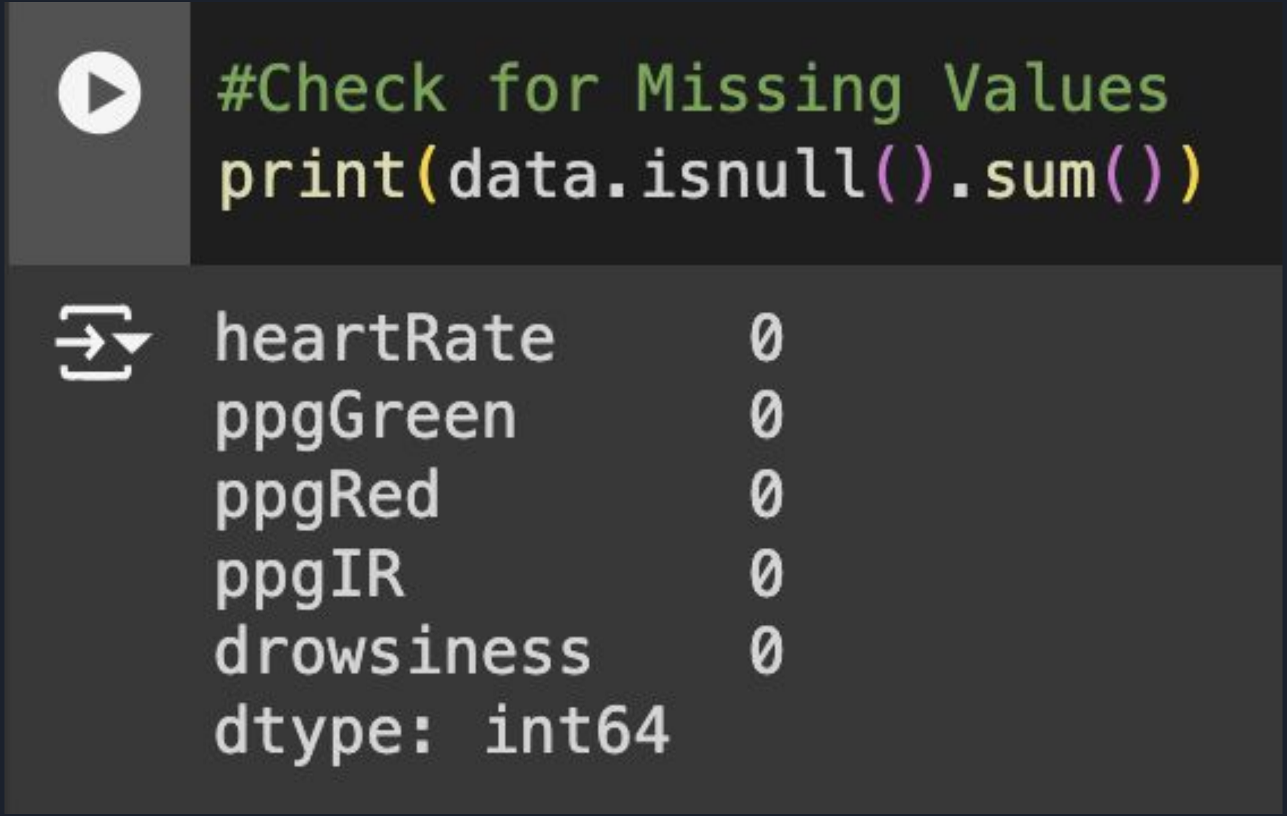


```
[1] #Import Necessary Libraries
    import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
#Load the dataset
data = pd.read_csv('drowsiness_dataset.csv')
data.head(10)
```

	heartRate	ppgGreen	ppgRed	ppgIR	drowsiness
0	54.0	1584091.0	5970731.0	6388383.0	0.0
1	54.0	1584091.0	5971202.0	6392174.0	0.0
2	54.0	1581111.0	5971295.0	6391469.0	0.0
3	54.0	1579343.0	5972599.0	6396137.0	0.0
4	54.0	1579321.0	5971906.0	6392898.0	0.0
5	54.0	1578536.0	5969930.0	6389646.0	0.0
6	54.0	1577547.0	5970184.0	6389553.0	0.0
7	54.0	1576090.0	5971546.0	6385977.0	0.0
8	54.0	1576964.0	5974102.0	6385031.0	0.0
9	54.0	1578325.0	5975938.0	6386914.0	0.0

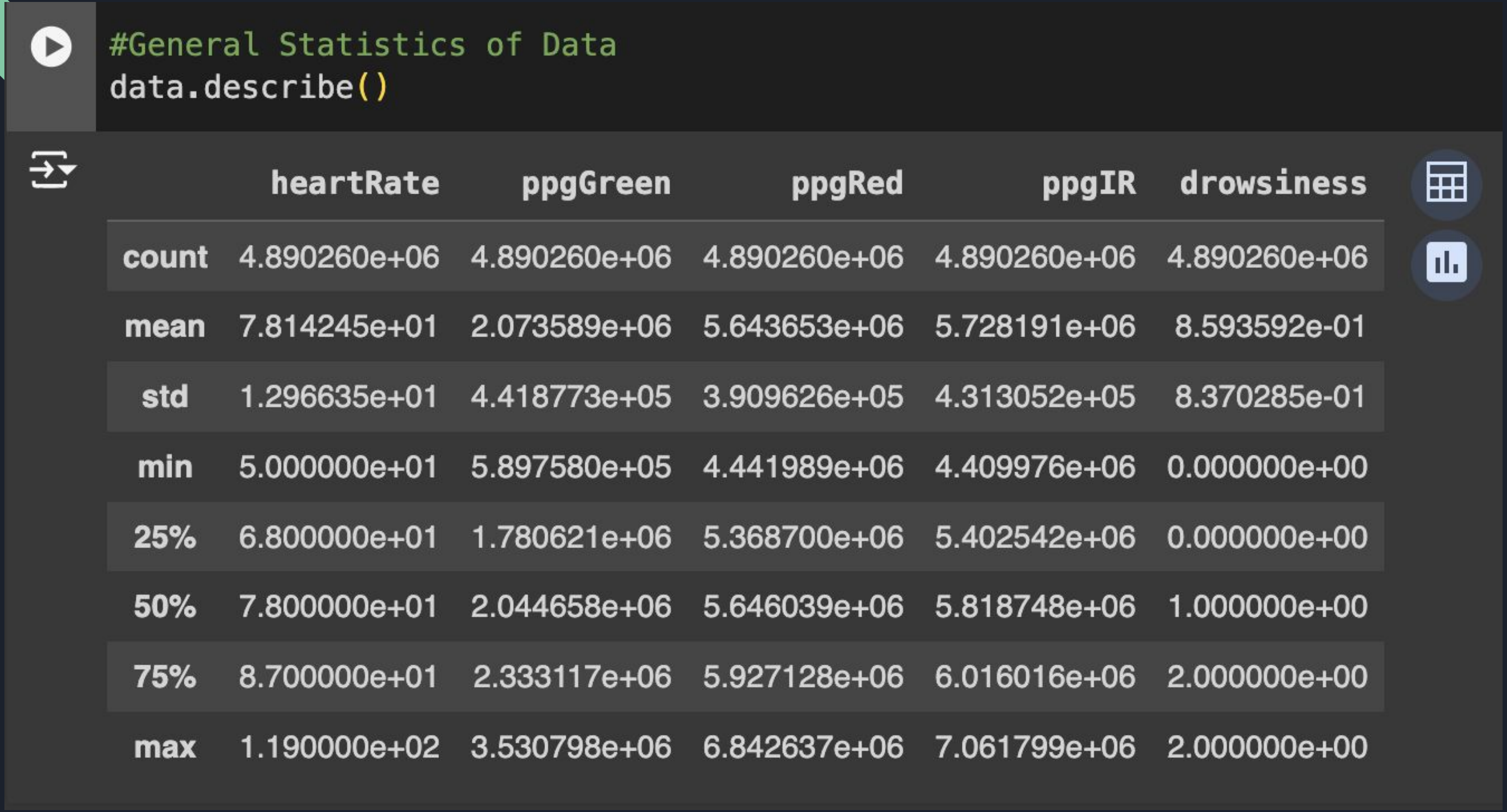
Checking for Missing Data



```
#Check for Missing Values
print(data.isnull().sum())
```

```
heartRate     0
ppgGreen      0
ppgRed        0
ppgIR         0
drowsiness    0
dtype: int64
```

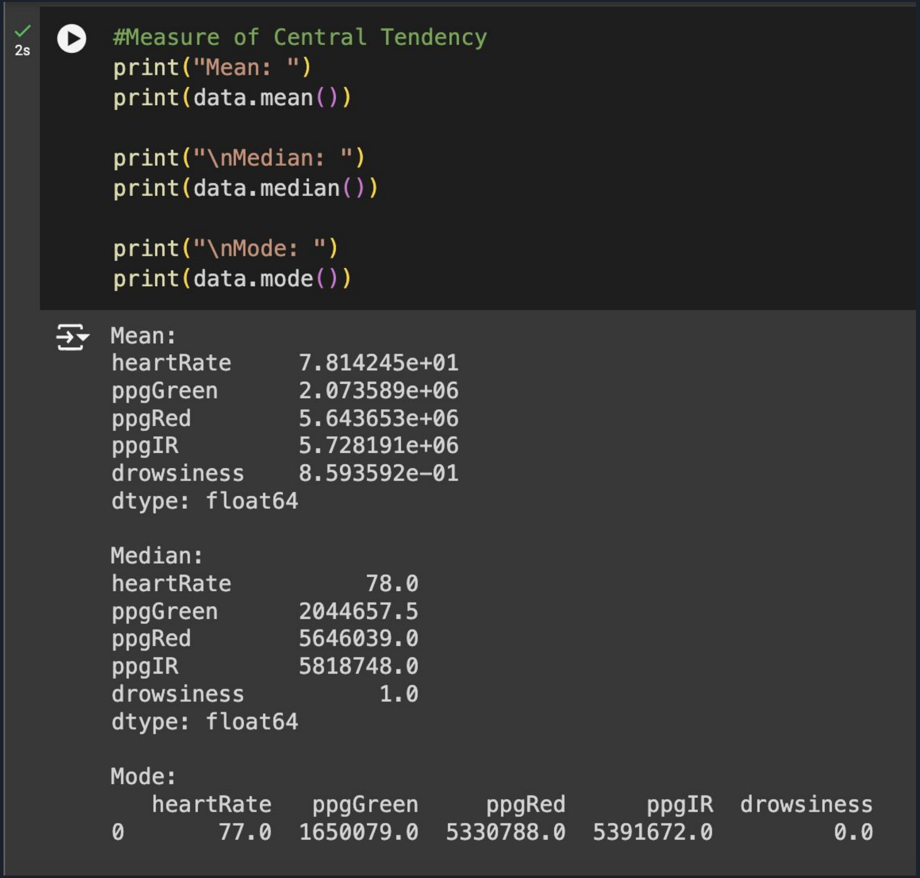
General Statistics of the Data



```
#General Statistics of Data
data.describe()
```

	heartRate	ppgGreen	ppgRed	ppgIR	drowsiness
count	4.890260e+06	4.890260e+06	4.890260e+06	4.890260e+06	4.890260e+06
mean	7.814245e+01	2.073589e+06	5.643653e+06	5.728191e+06	8.593592e-01
std	1.296635e+01	4.418773e+05	3.909626e+05	4.313052e+05	8.370285e-01
min	5.000000e+01	5.897580e+05	4.441989e+06	4.409976e+06	0.000000e+00
25%	6.800000e+01	1.780621e+06	5.368700e+06	5.402542e+06	0.000000e+00
50%	7.800000e+01	2.044658e+06	5.646039e+06	5.818748e+06	1.000000e+00
75%	8.700000e+01	2.333117e+06	5.927128e+06	6.016016e+06	2.000000e+00
max	1.190000e+02	3.530798e+06	6.842637e+06	7.061799e+06	2.000000e+00

Measure of Central Tendency



```
#Measure of Central Tendency
print("Mean: ")
print(data.mean())

print("\nMedian: ")
print(data.median())

print("\nMode: ")
print(data.mode())
```

```
Mean:
heartRate     7.814245e+01
ppgGreen      2.073589e+06
ppgRed        5.643653e+06
ppgIR         5.728191e+06
drowsiness    8.593592e-01
dtype: float64

Median:
heartRate          78.0
ppgGreen      2044657.5
ppgRed        5646039.0
ppgIR         5818748.0
drowsiness          1.0
dtype: float64

Mode:
   heartRate   ppgGreen     ppgRed      ppgIR  drowsiness
0       77.0  1650079.0  5330788.0  5391672.0         0.0
```

Significant Findings

- Values for heart rate in the dataset ranges from 50 - 119 Beats Per Minute (BPM).

- The mean heart rate is roughly 78 BPM, while the mean drowsiness level is approximately 0.85 (somewhat alert).

- Recordings of ppgRed and ppgIR values appear significantly higher than ppgGreen values.

- The mode heart rate is 77 BPM (which is considered a normal resting heart rate).

Visualizing the Data

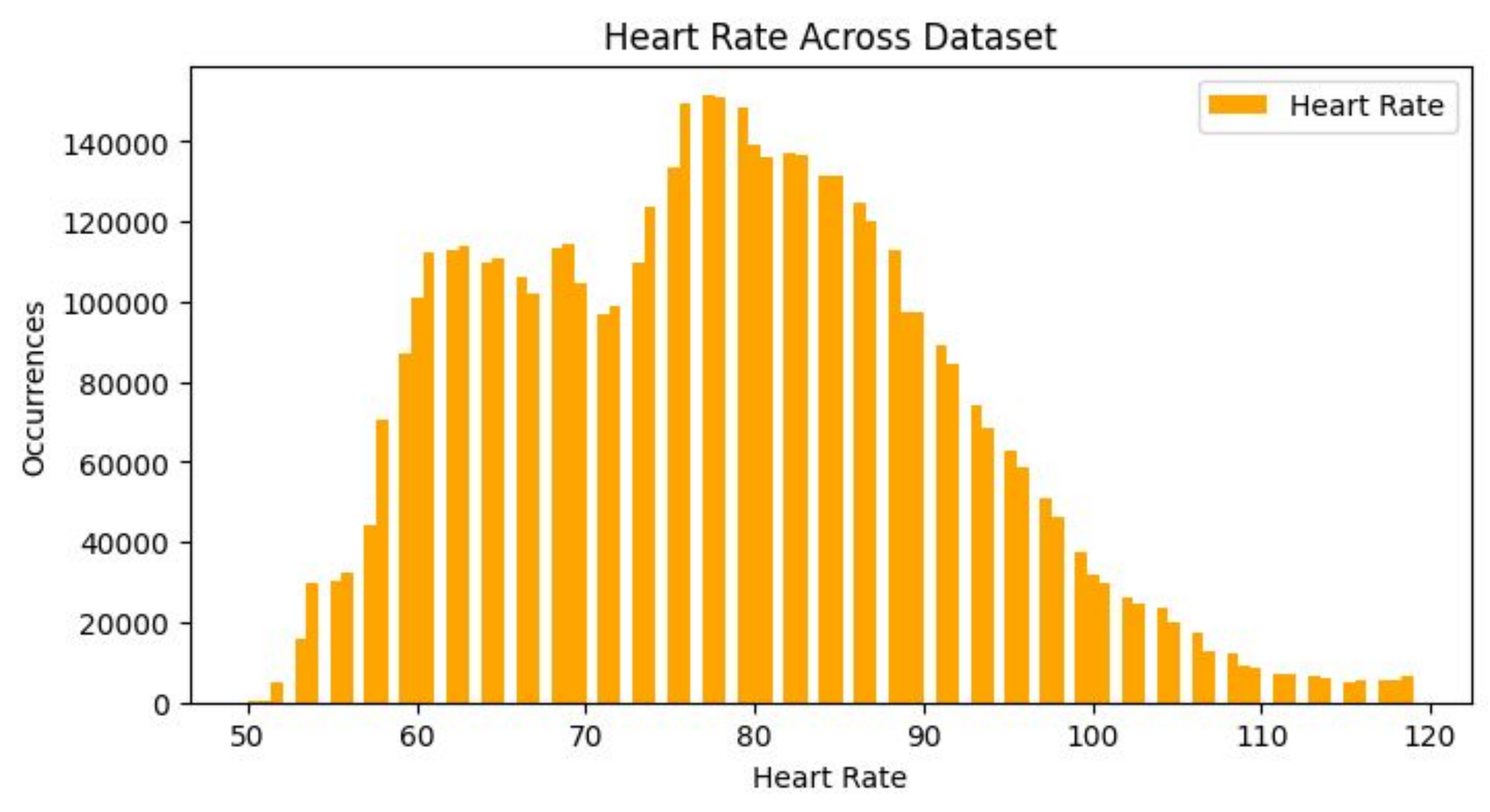
- The next step is to visualize the given data by creating histograms/distributions.

- First, histograms will be created for each variable present in the dataset.

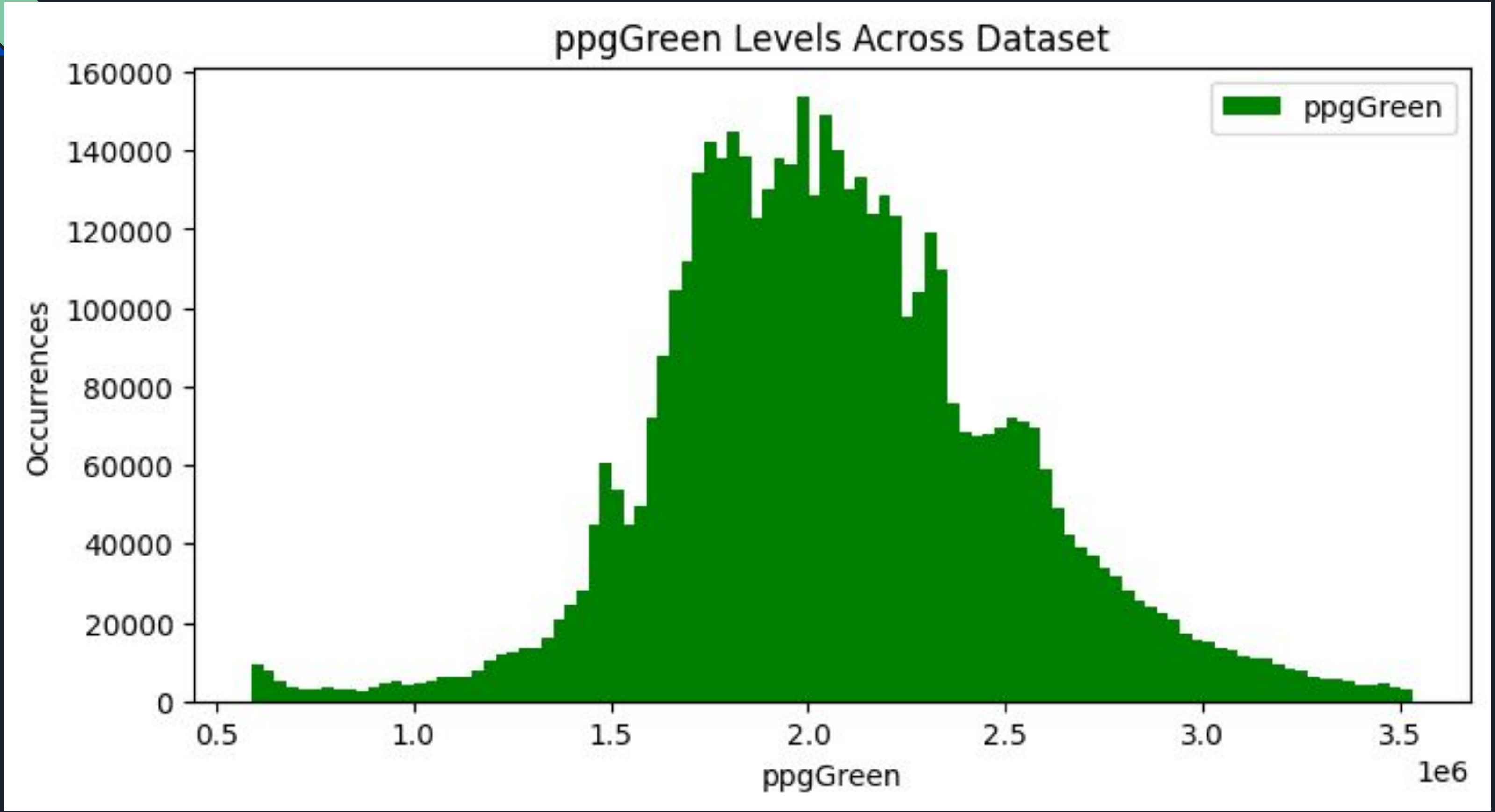
- Next, it will be observed how heart rate and PPG signals vary across different levels of drowsiness.

- The following graphs illustrate the information provided in drowsiness_dataset.csv.

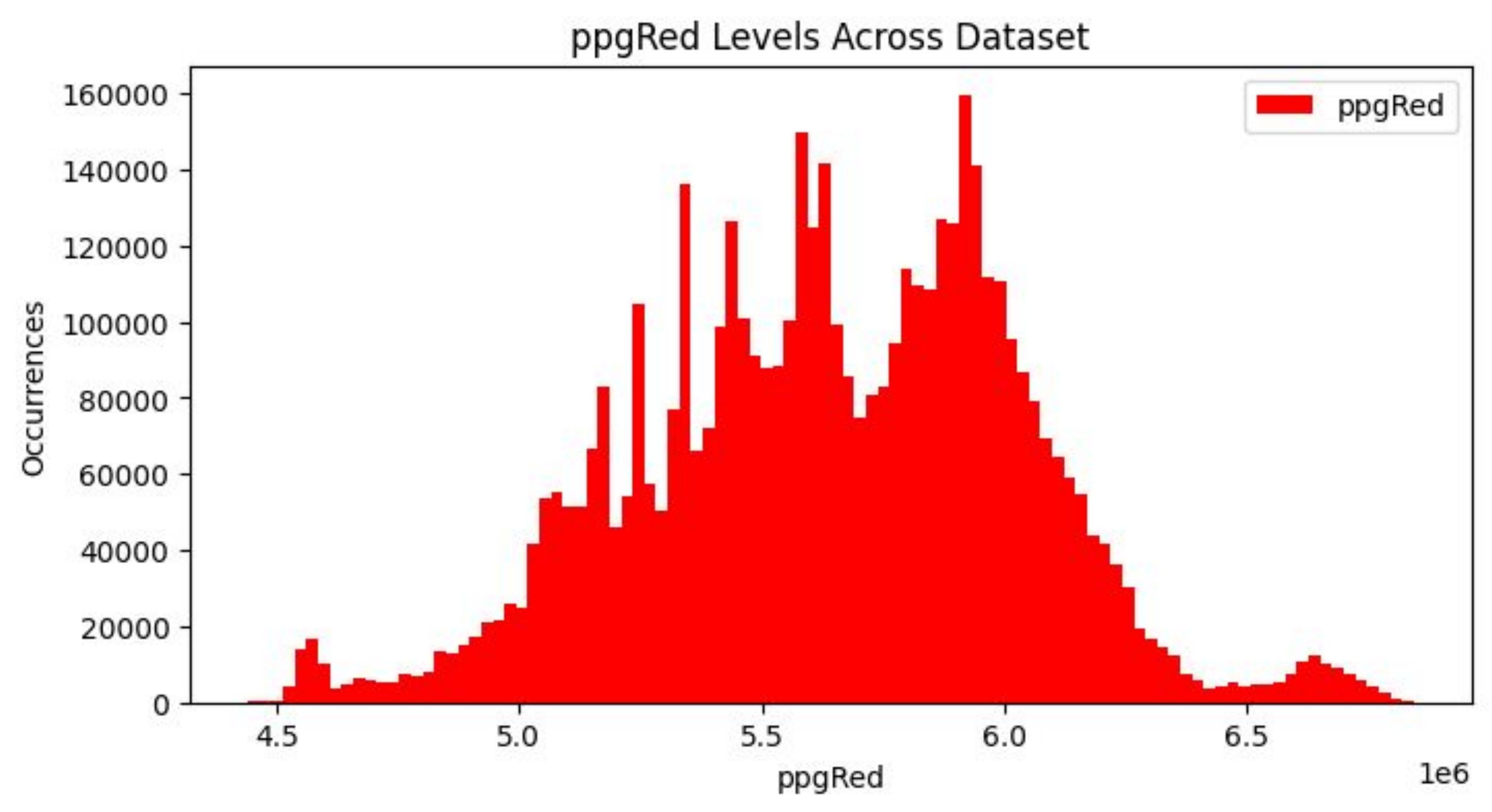
Heart Rate


Heart Rate Across Dataset
Heart Rate
Occurrences
0
20000
40000
60000
80000
100000
120000
140000
50
60
70
80
90
100
110
120
Heart Rate

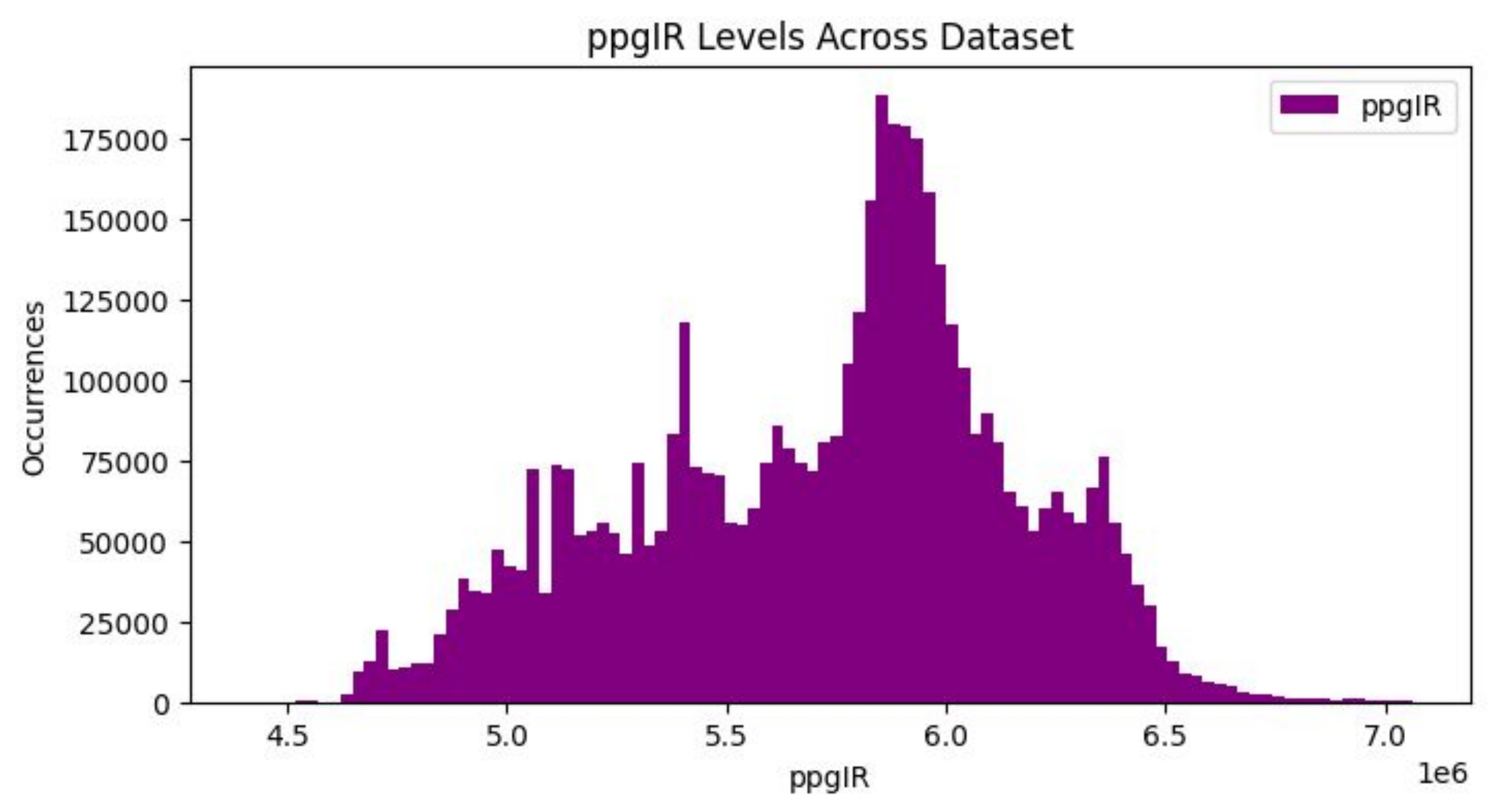
ppgGreen Levels



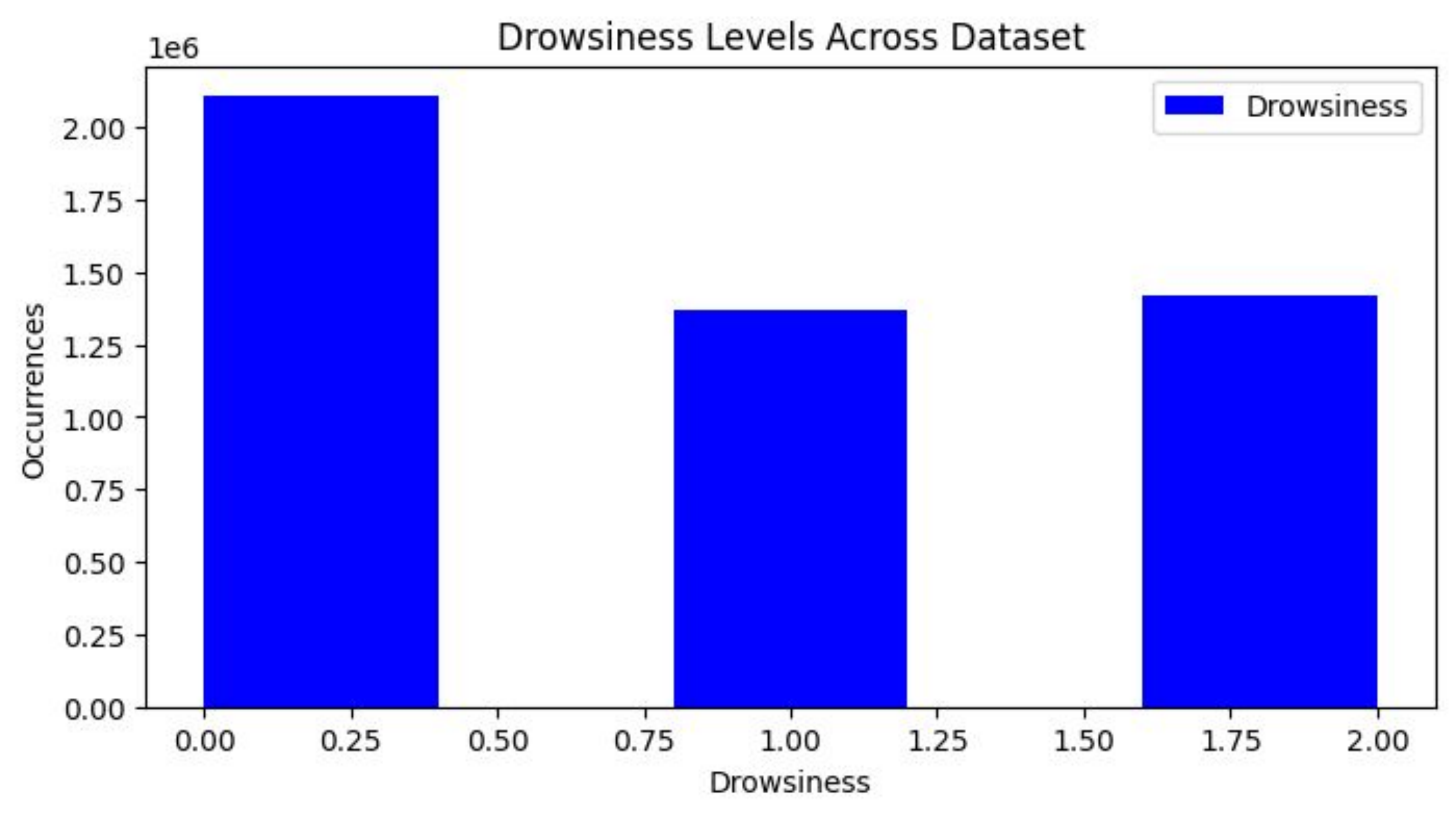
ppgRed Levels



ppgInfrared Levels



Drowsiness Levels



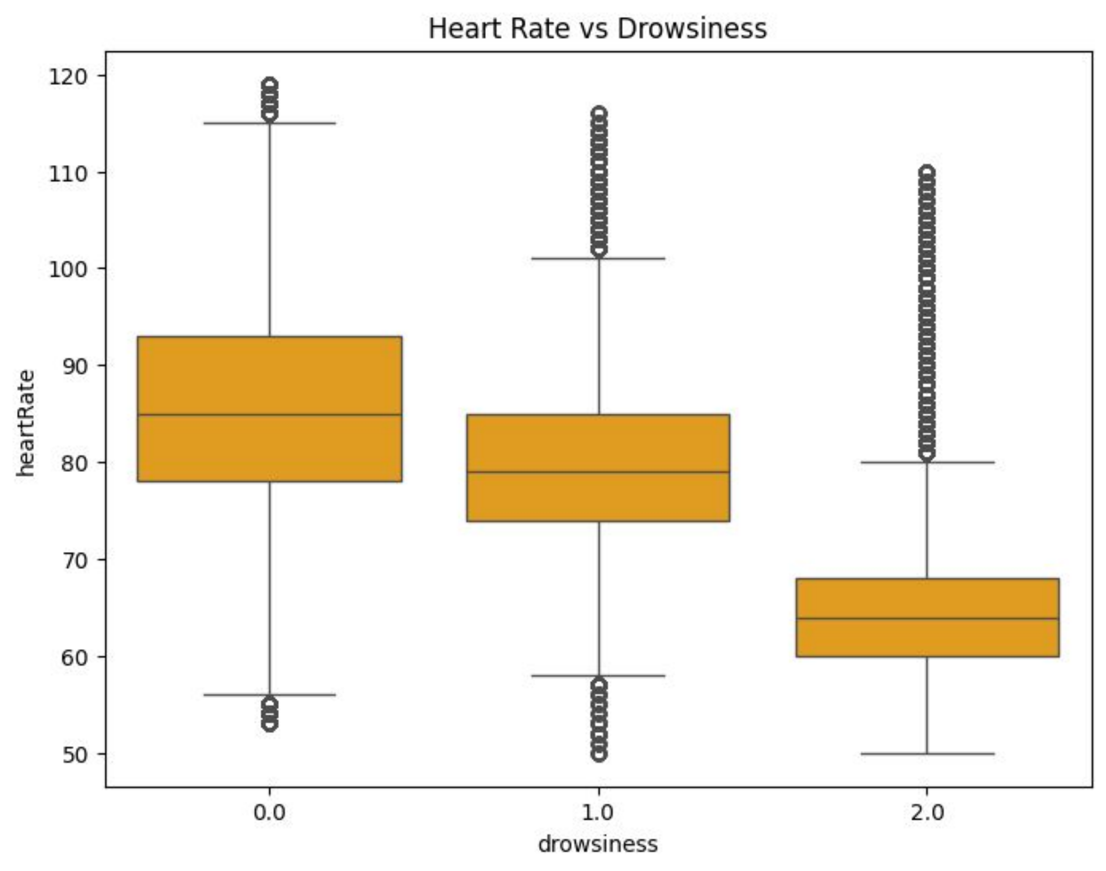
Exploring Potential Correlations

- **The previous visuals provided general information about the data present in drowsiness_dataset.csv.**

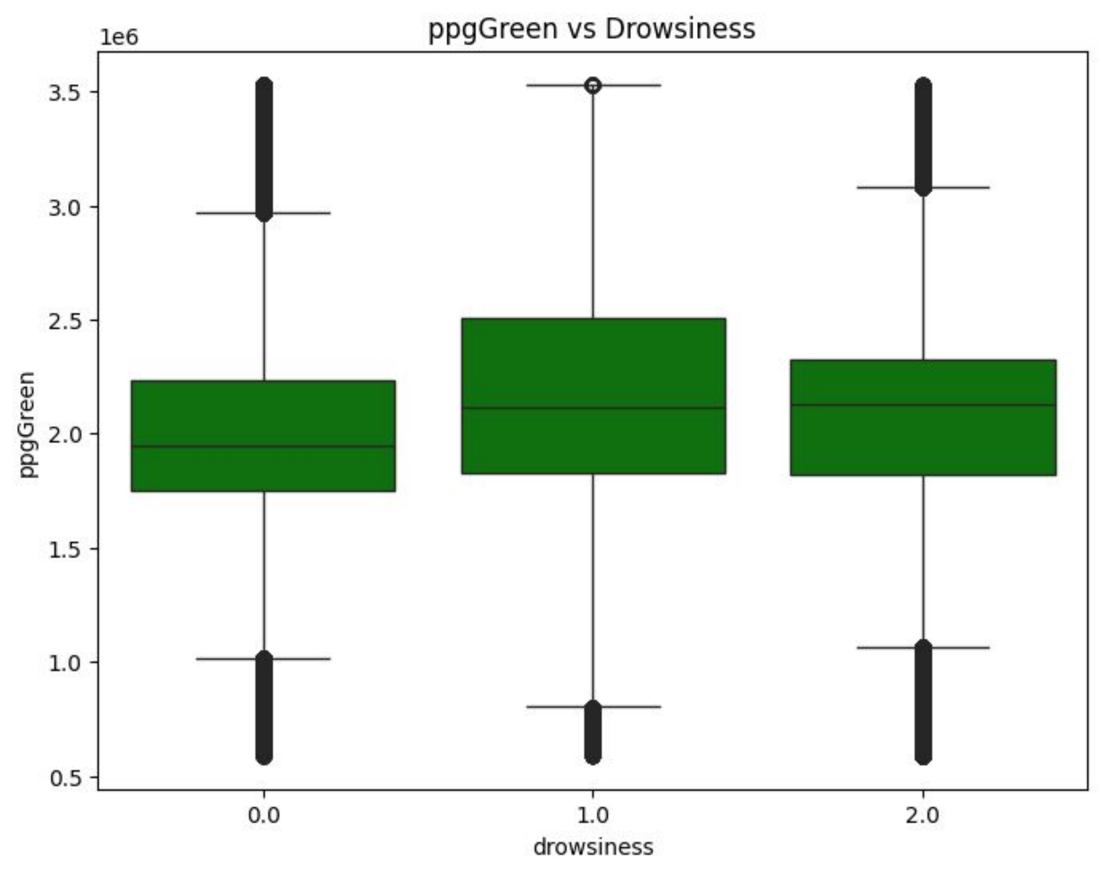
- **Exploring potential correlations between the variables can further deepens one’s understanding of the data.**

- **The following slides represent box plots of each variable in relation to drowsiness, as well as a correlation matrix.**

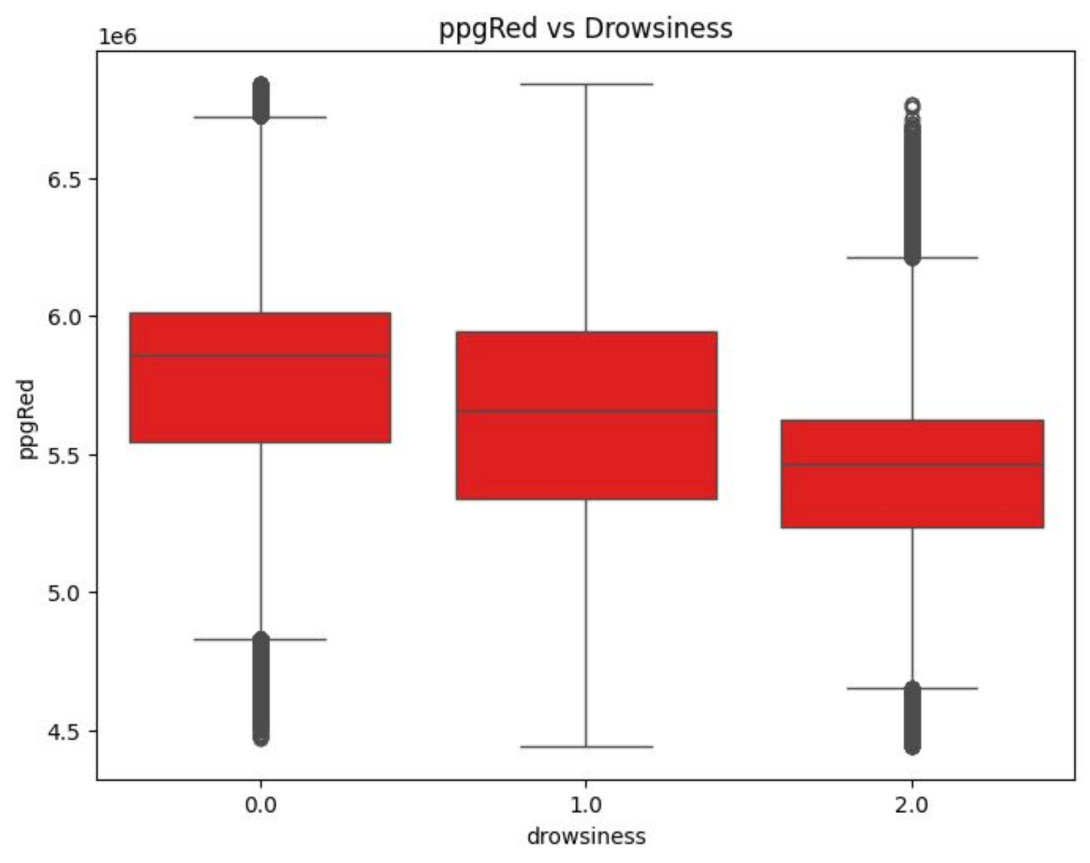
Heart Rate vs Drowsiness



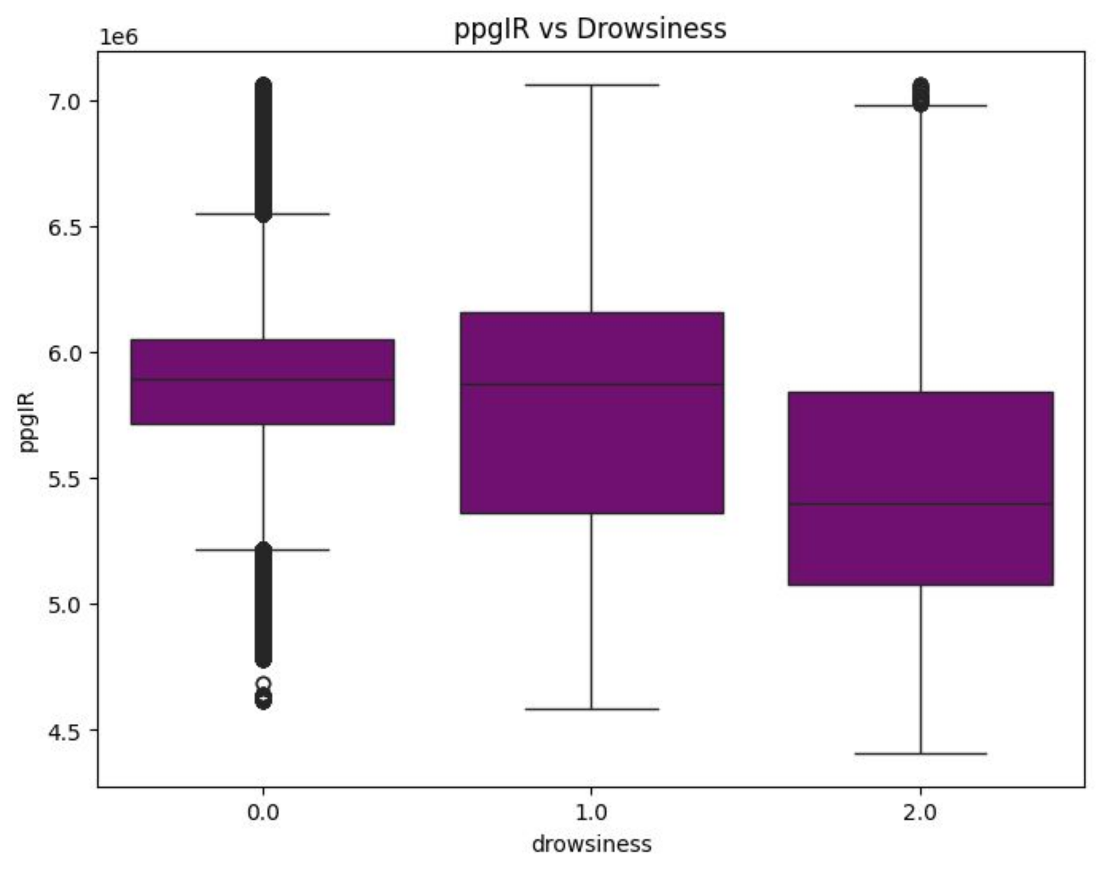
ppgGreen vs Drowsiness



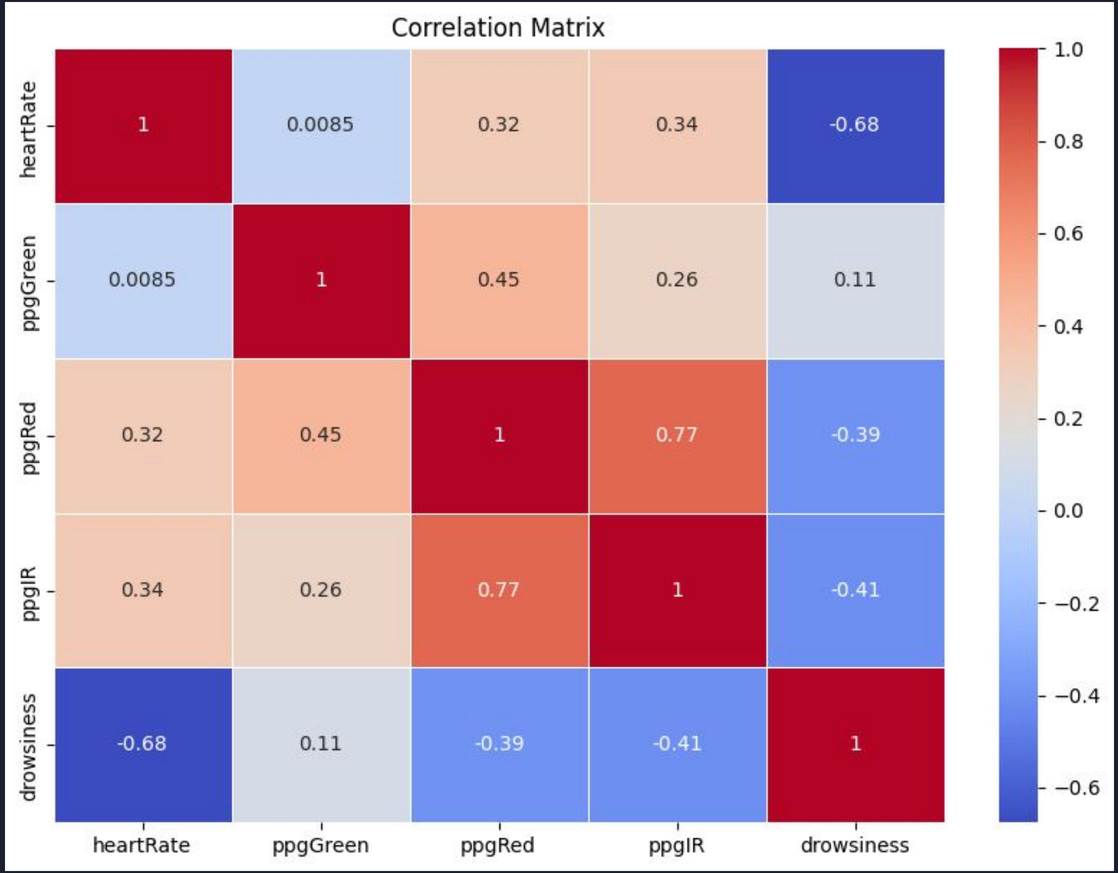
ppgRed vs Drowsiness



ppgIR vs Drowsiness



Correlation Matrix of Data



	heartRate	ppgGreen	ppgRed	ppgIR	drowsiness
heartRate	1	0.0085	0.32	0.34	-0.68
ppgGreen	0.0085	1	0.45	0.26	0.11
ppgRed	0.32	0.45	1	0.77	-0.39
ppgIR	0.34	0.26	0.77	1	-0.41
drowsiness	-0.68	0.11	-0.39	-0.41	1

Significant Findings

- On average, people with lower heart rates are more likely to experience higher levels of drowsiness.
- However, the variable that has the least correlation to drowsiness is also heart rate, as seen by the matrix.
- ppgRed and ppgIR have the highest correlation values (0.32, 0.34) in the matrix (excluding self correlations).
- There are no outliers in the data for ppgRed and ppgIR for people with level 1.0 drowsiness.

Significant Findings (Continued)

- **For all variables except ppgIR, the boxplots for level 2.0 drowsiness have the smallest quartile range.**

- **ppgGreen is the only variable with nearly no correlation to drowsiness (given its correlation value 0.0085).**

- **Putting everything together, the data suggests ppgRed and ppgIR contributes most to drowsiness, while ppgGreen has no significant effect. Additionally, heart rate is the least significant/reliable variable to drowsiness.**

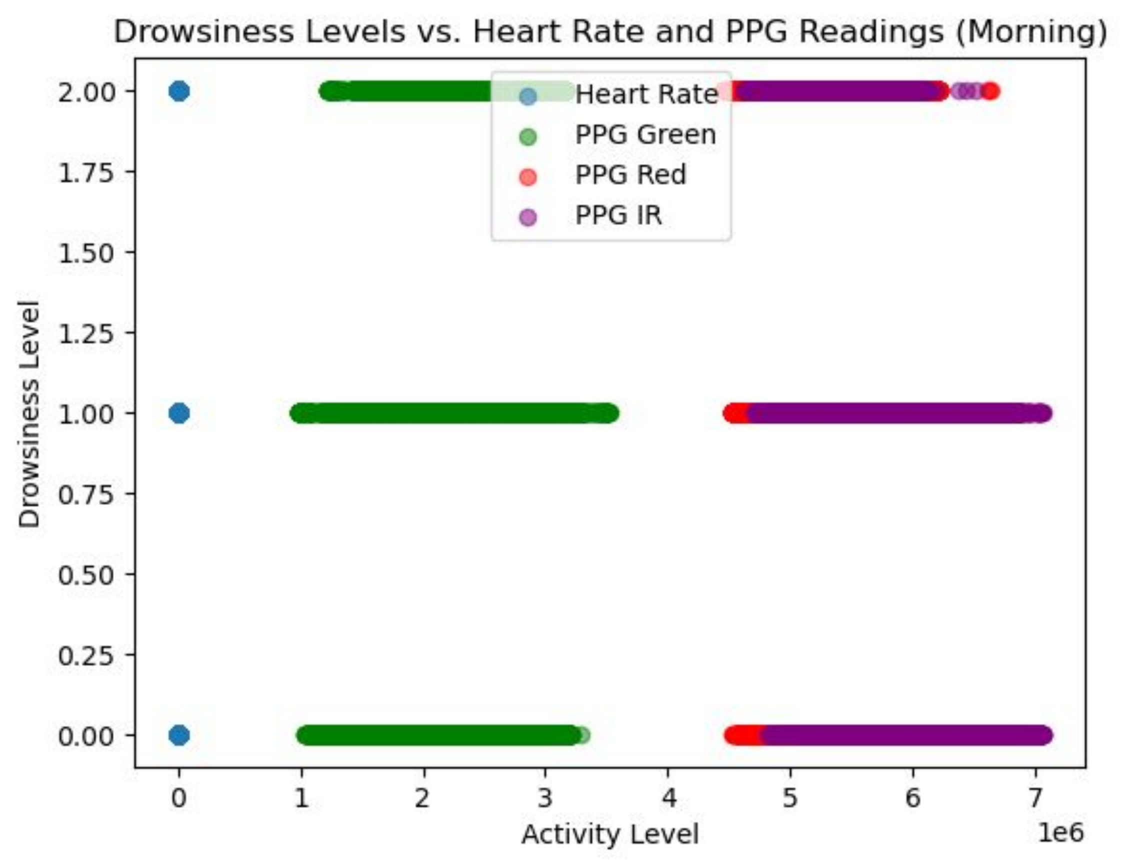
Splitting the Data

- Recall that the data comes from a technology company that produces smartwatches.

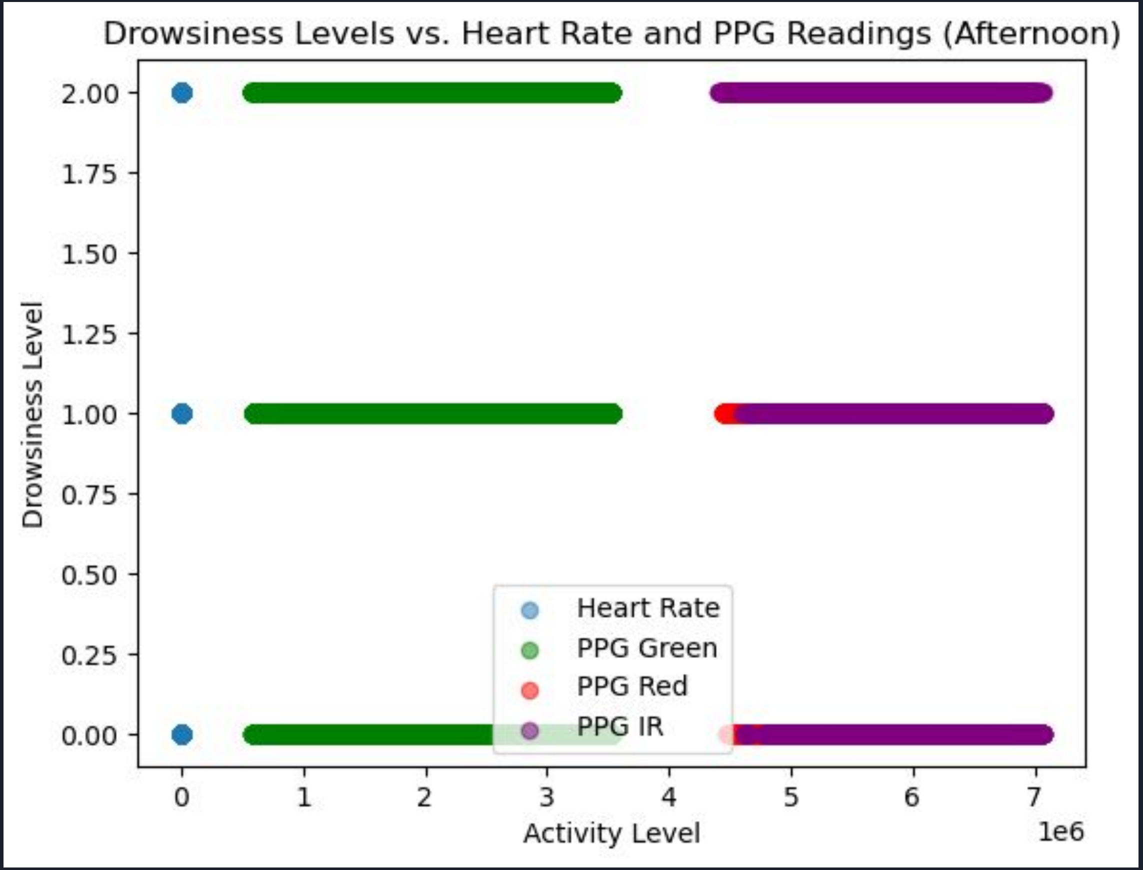
- Given this, each data entry comes from a specific point in time during the day from morning to night.

- In addition to all previous analysis, it would be ideal to divide the data into separate times of the day to see how drowsiness may also be correlated to a specific time of day (morning, afternoon, evening, night).

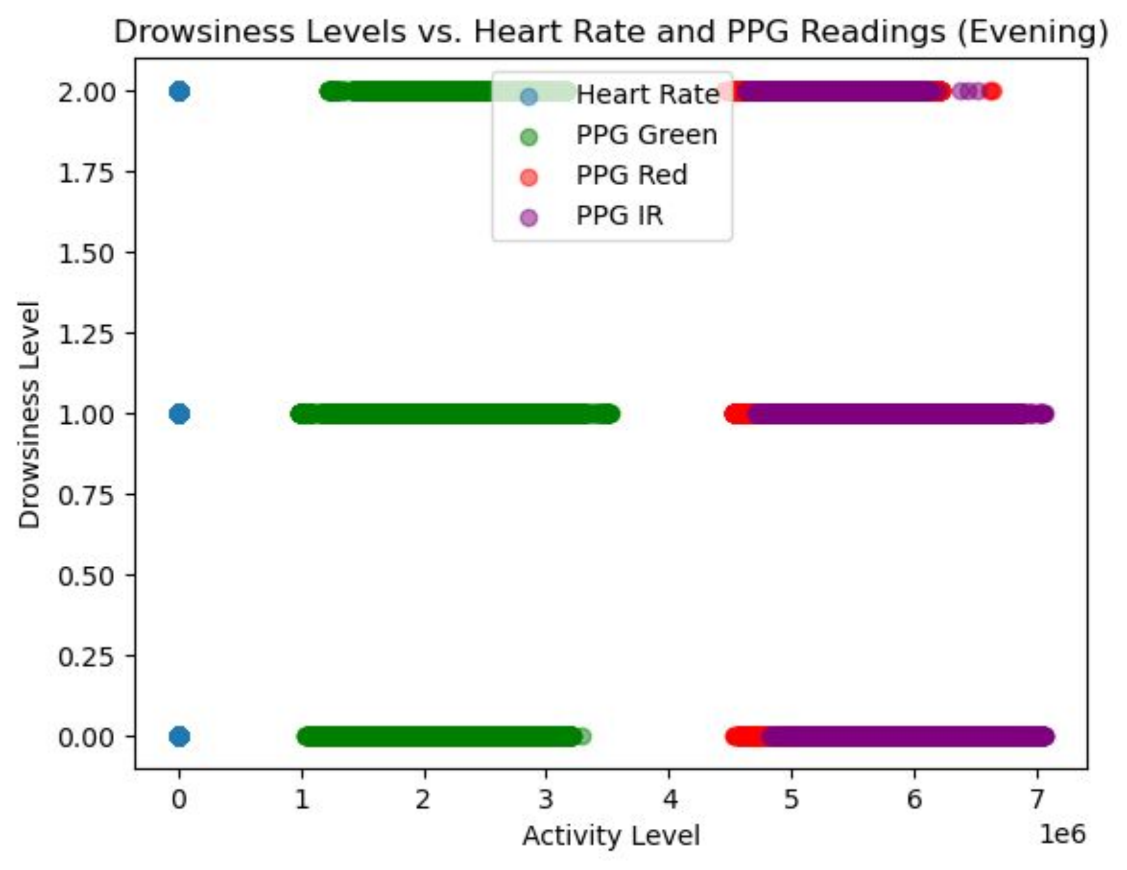
Morning Data



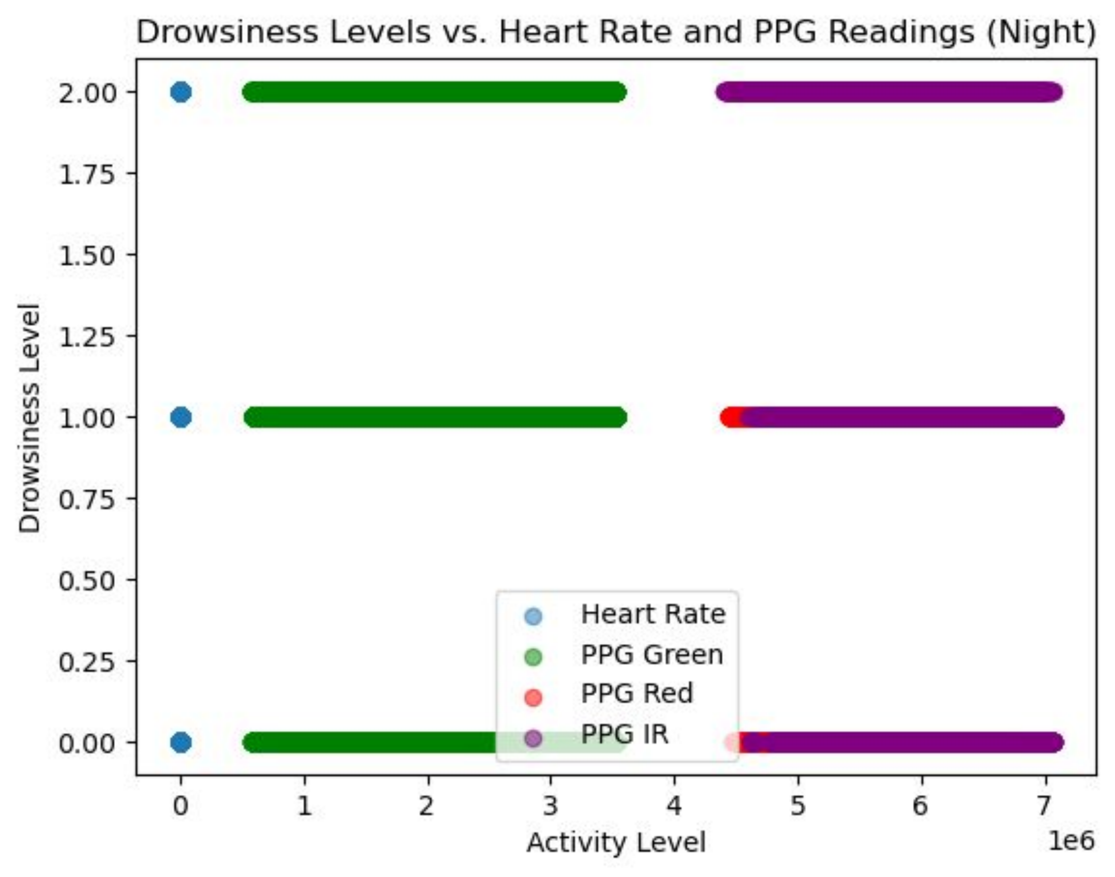
Afternoon Data



Evening Data



Night Data


Drowsiness Levels vs. Heart Rate and PPG Readings (Night)
Drowsiness Level
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Heart Rate
PPG Green
PPG Red
PPG IR
0
1
2
3
4
5
6
7
1e6
Activity Level

Significant Findings

- Once more, it can be observed how heart rate has no significant correlation with drowsiness.

- There is noticeable overlap between ppgRed and ppgIR levels

- On average, ppgIR levels are smaller in the morning and evening compared to afternoon and night.

Conclusion

Based on all analysis conducted, a person with level 2.0 drowsiness should contain some of the following traits:

- **Average heart rate of 65 BPM.**
- **Average ppgGreen levels of 2.2 x 10^6.**
- **Average ppgRed levels of 5.5 x 10^6.**
- **Average ppgIR levels of 5.4 X 10^6.**
- **Experience 2.0 drowsiness during the afternoon/night.**

Note: Recall ppgRed and ppgIR has the highest correlation to drowsiness.

Conclusion (Continued)

On the other hand, a person with level 0.0 drowsiness should contain some of the following traits:

- **Average heart rate of 84 BPM.**
- **Average ppgGreen levels of 1.9 x 10^6.**
- **Average ppgRed levels of 5.8 x 10^6.**
- **Average ppgIR levels of 5.8 X 10^6.**
- **Experience 0.0 drowsiness during the morning/evening.**

Note: Recall ppgRed and ppgIR has the highest correlation to drowsiness.