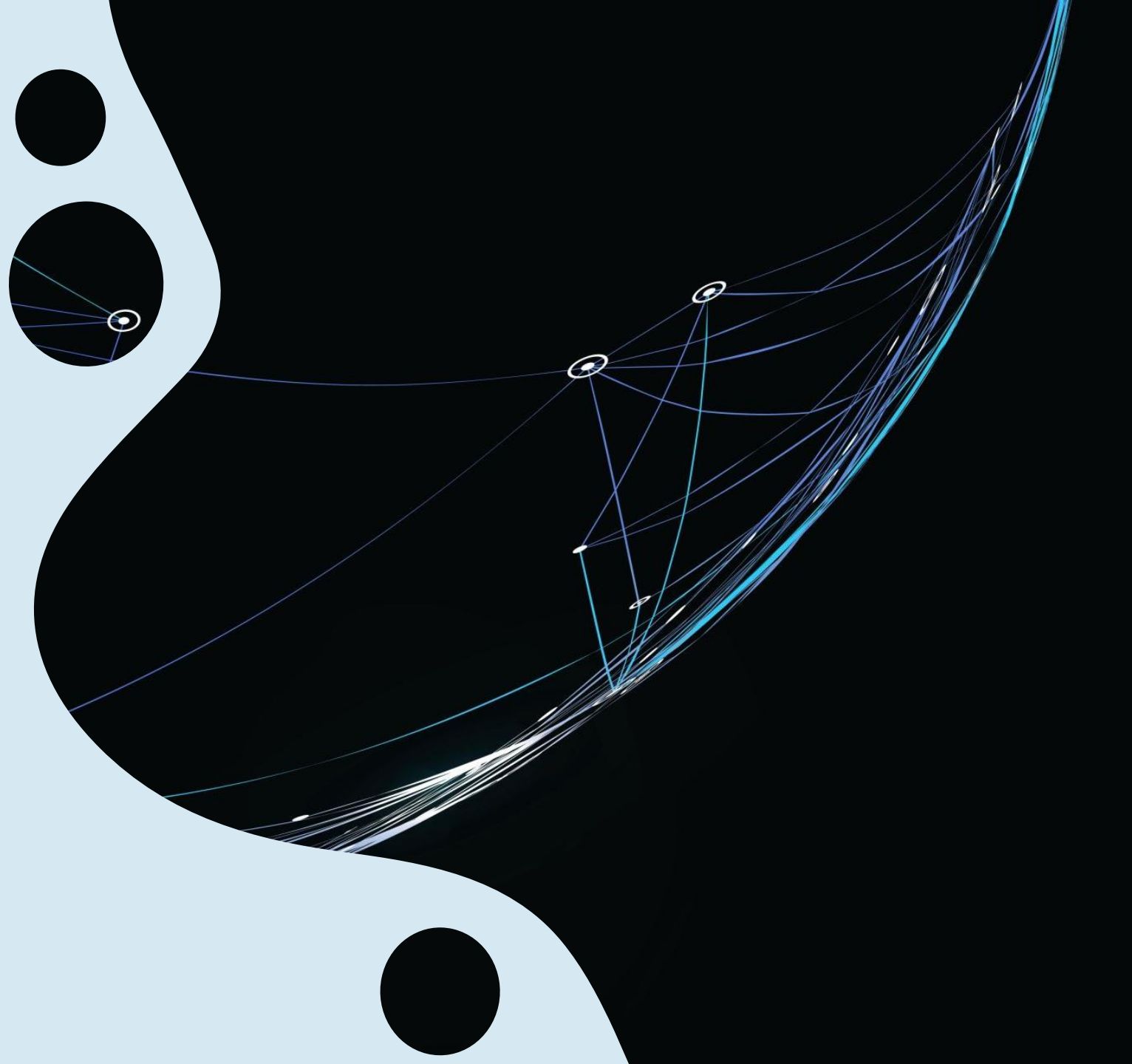


Exploratory Data Analysis (EDA)

Balkrishna M Mottannavar



About the Dataset



The Dataset was created using a Galaxy watch 4 smart watch vital signs sensors.



Data were collected from Heart Rate and PPG Sensors. PPG Includes 3 Variations

1. Green
2. Red
3. Infrared (IR)



The Drowsiness column refers to the label assigned by the user based on an adaption of the Karolinska sleepiness scale (KSS). Labels range from 0.0 to 2.0 where,

- 0.0 represents level 1 on the KSS scale (Alert)
- 1.0 corresponds to level 6 (some signs of sleepiness), and
- 2.0 indicates level 8 (sleepy, but some effort to stay awake)

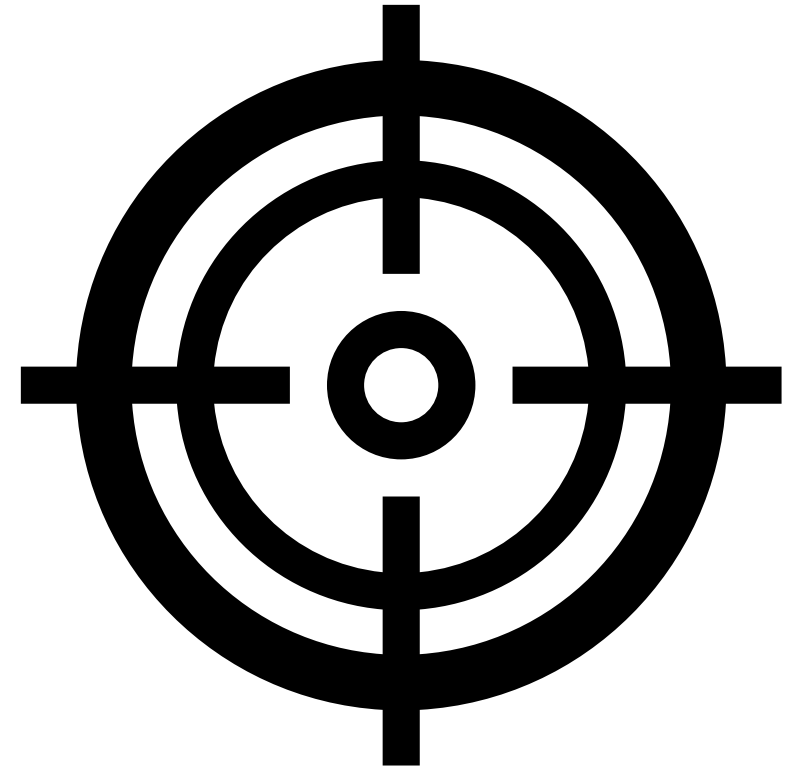
Source: <https://www.kaggle.com/datasets/vitoraugustx/drowsiness-dataset>

Objective

Perform an Exploratory Data Analysis (EDA) on a Dataset collected from these smart watches.

The Dataset includes various physiological parameters along with a 'drowsiness' label, which indicates the level of sleepiness based on an adapted Karolinska Sleepiness Scale (KSS)

For more about KSS see: [Karolinska Sleepiness Scale \(KSS\)](#)



Overview of the Dataset

The Data has:

Rows: 4890260

Columns: 5



	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

Fig. First 5 rows of the dataset

	heartRate	ppgGreen	ppgRed	ppgIR	drowsiness
4890255	63.0	2286384.0	5783226.0	6356797.0	2.0
4890256	63.0	2289887.0	5783786.0	6357004.0	2.0
4890257	63.0	2291928.0	5784221.0	6358348.0	2.0
4890258	63.0	2295386.0	5785012.0	6358565.0	2.0
4890259	63.0	2296992.0	5783386.0	6357466.0	2.0

Fig. Bottom 5 rows of the dataset

Finding and Handling the Missing Values

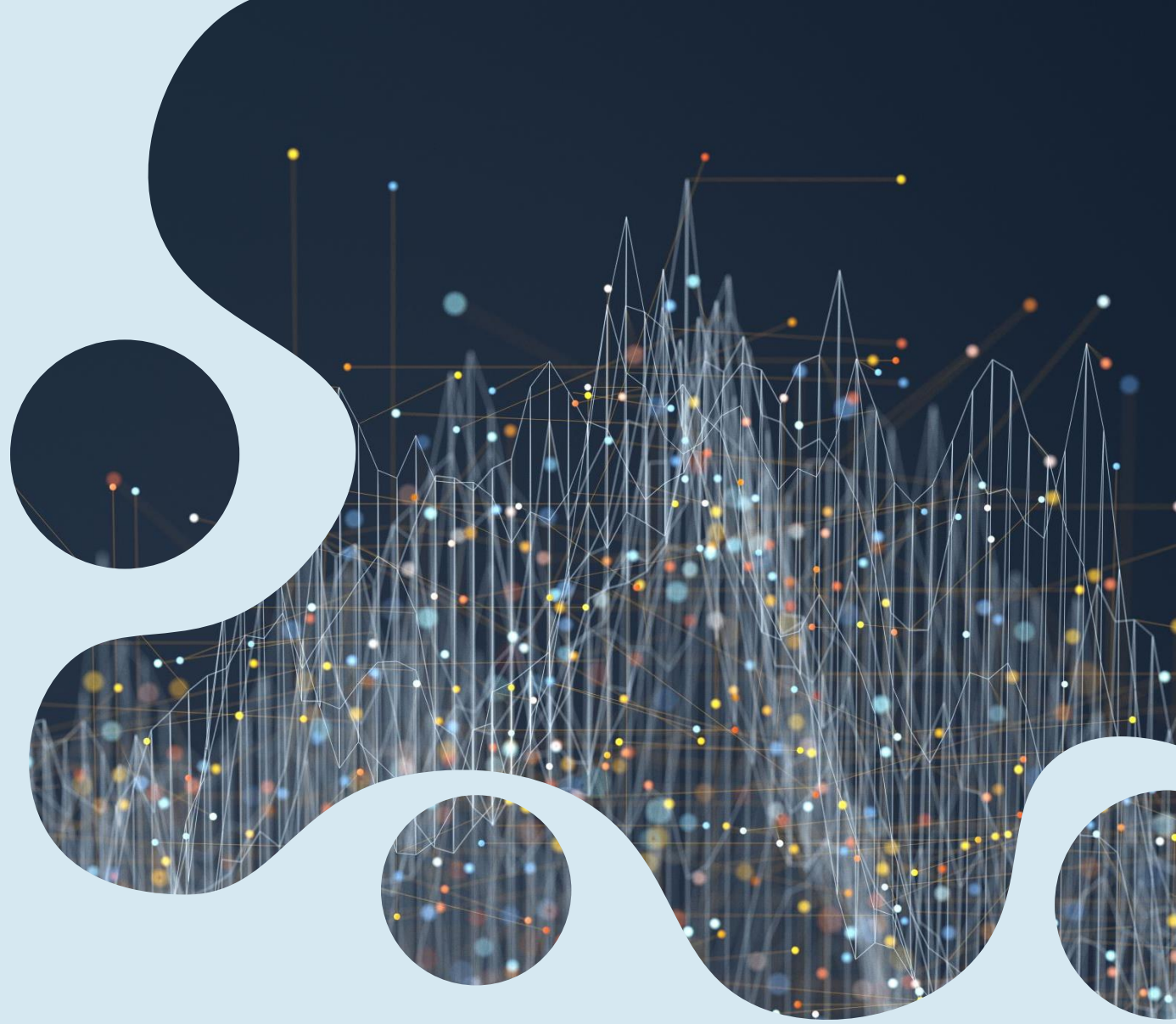
```
missing_val = data.isnull().sum()  
missing_val
```

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

There are NO Missing Values



Data Visualization





Distribution of the Data

```
# Distribution plots for each feature
plt.figure(figsize=(16, 12), facecolor='lightgrey')

# Heart Rate
plt.subplot(2, 3, 1)
sns.histplot(data['heartRate'], kde=True, bins=30, color = '#300000')
plt.title('Heart Rate Distribution')

# PPG Green
plt.subplot(2, 3, 2)
sns.histplot(data['ppgGreen'], kde=True, bins=30, color = 'green')
plt.title('PPG Green Distribution')

# PPG Red
plt.subplot(2, 3, 3)
sns.histplot(data['ppgRed'], kde=True, bins=30, color = 'red')
plt.title('PPG Red Distribution')

# PPG IR
plt.subplot(2, 3, 4)
sns.histplot(data['ppgIR'], kde=True, bins=30, color = 'grey')
plt.title('PPG IR Distribution')

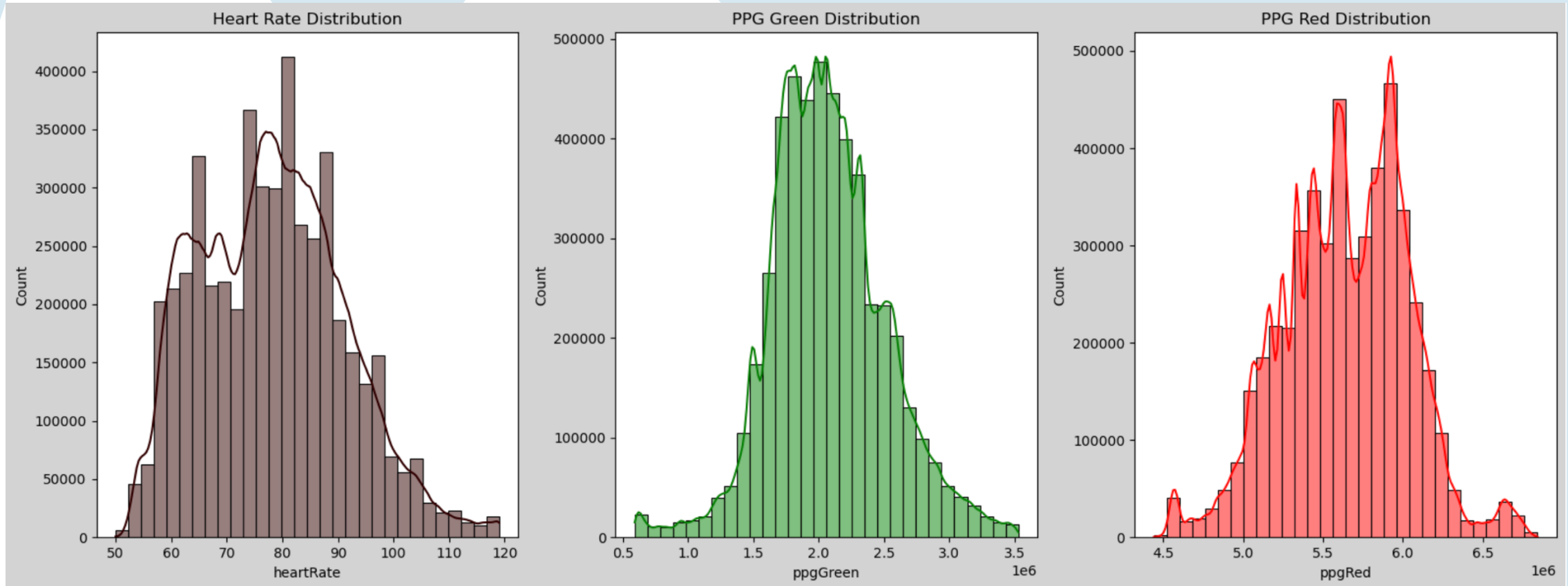
# Drowsiness
plt.subplot(2, 3, 5)
sns.histplot(data['drowsiness'], kde=True, bins=3, discrete=True, color = 'orange')
plt.title('Drowsiness Distribution')

plt.tight_layout()
plt.show()
```

Code Snippet

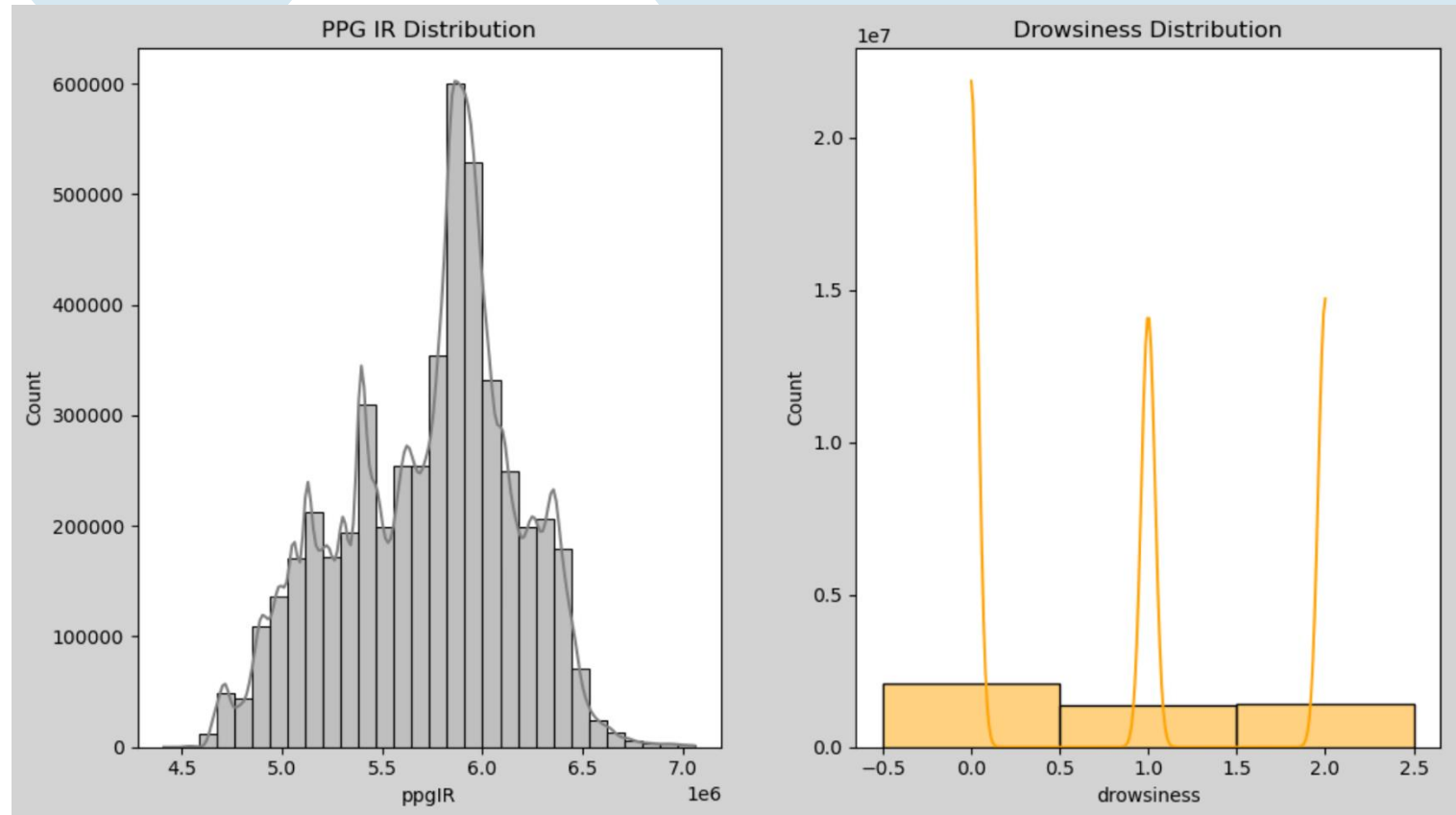


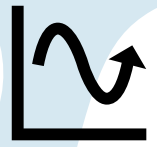
Distribution of the Data





Distribution of the Data





Heart rate and PPG Varying across Drowsiness

```
# Set up the plotting area
plt.figure(figsize=(16, 12))

# Heart Rate vs Drowsiness with custom palette
plt.subplot(2, 2, 1)
sns.boxplot(x='drowsiness', y='heartRate', data=data, palette={0.0: '#d50000', 1.0: '#ff5252', 2.0: '#ef9a9a'})
plt.title('Heart Rate vs Drowsiness')
plt.xlabel('Drowsiness')
plt.ylabel('Heart Rate')

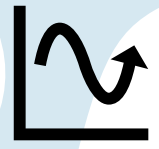
# PPG Green vs Drowsiness with custom palette
plt.subplot(2, 2, 2)
sns.boxplot(x='drowsiness', y='ppgGreen', data=data, palette={0.0: '#0eff00', 1.0: '#063b00', 2.0: '#089000'})
plt.title('PPG Green vs Drowsiness')
plt.xlabel('Drowsiness')
plt.ylabel('PPG Green')

# PPG Red vs Drowsiness with custom palette
plt.subplot(2, 2, 3)
sns.boxplot(x='drowsiness', y='ppgRed', data=data, palette={0.0: '#300000', 1.0: '#b10000', 2.0: '#f44336'})
plt.title('PPG Red vs Drowsiness')
plt.xlabel('Drowsiness')
plt.ylabel('PPG Red')

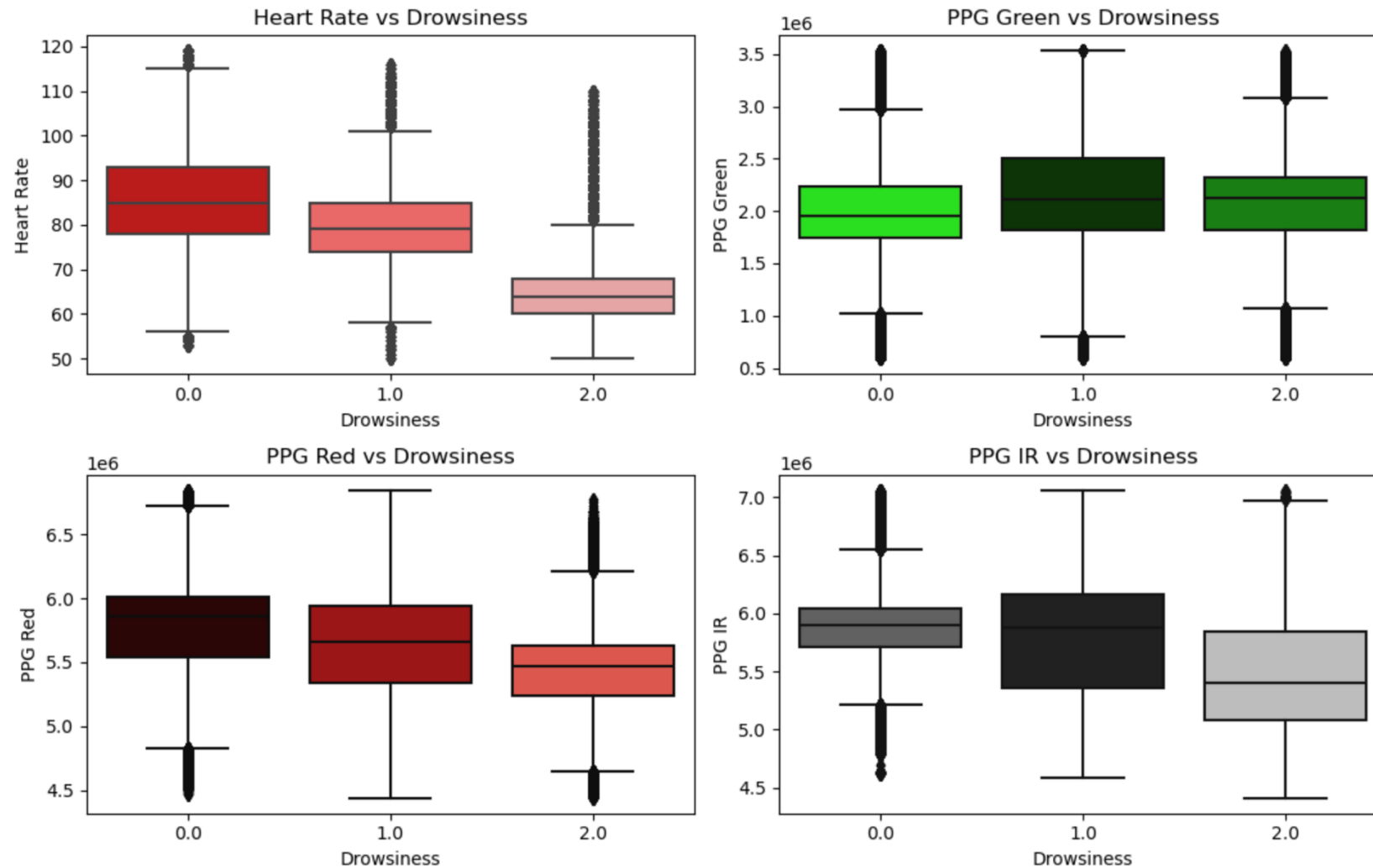
# PPG IR vs Drowsiness with custom palette
plt.subplot(2, 2, 4)
sns.boxplot(x='drowsiness', y='ppgIR', data=data, palette={0.0: '#616161', 1.0: '#212121', 2.0: '#bdbdbd'})
plt.title('PPG IR vs Drowsiness')
plt.xlabel('Drowsiness')
plt.ylabel('PPG IR')

plt.tight_layout()
plt.show()
```

Code Snippet



Heart rate and PPG Varying across Drowsiness





Heart rate and PPG Varying across Drowsiness

Heart Rate vs Drowsiness:

1. Heart Rate is inversely proportional to drowsiness. i.e., the median heart rate decreases as drowsiness levels increase, this indicates that higher drowsiness levels might be associated with the lower heart rates.
2. We can find a positive increase in the outliers for heart rate despite an increase in drowsiness.. The Potential reasons would be:
 - i. Engagement in Physical activity shortly before the measurement.
 - ii. Stress or Anxiety.
 - iii. Caffeine intake and Sleep Disorders.
 - iv. Other Medical Conditions like Hyperthyroidism or Cardiovascular issues.



Heart rate and PPG Varying across Drowsiness

PPG Green vs Drowsiness:

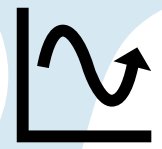
1. PPG (photoplethysmography) is a technique in which green light is emitted in the skin, where the skin absorbs, reflects, or emits the light, and the photodetector measures the amount of light reflected or transmitted in the tissue.
2. The drowsiness level 0 has outliers at both ends, indicating that some of the individuals feel active, irrespective of whether the photodetector measured the highest or lowest reflected green light.
3. The drowsiness level 2 also has outliers at both ends, indicating that some of the individuals feel drowsier, irrespective of whether the photodetector measured the highest or lowest reflected green light.
4. PPG green values shows high variability and the presence of outliers across all the drowsiness levels and median values remain relatively stable.
5. This make it less reliable as a standalone indicator of drowsiness.



Heart rate and PPG Varying across Drowsiness

PPG Red vs Drowsiness:

1. The median values of PPG Red across the levels of drowsiness shows a trend that drowsiness increases as the PPG Red values decrease.
2. At drowsiness level 0, the highest outliers are seen at lower rates of PPG Red, indicating that some individuals feel active irrespective of the PPG Red values.
3. At drowsiness level 1, there is an equal distribution, and no outliers are present indicating less variability.
4. At drowsiness level 2, Variability in PPG Red values persists even at high drowsiness level.
5. High variability and weak correlation of PPG Red with drowsiness levels suggests that PPG red is not a strong indicator of drowsiness.



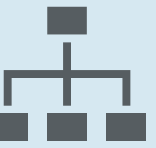
Heart rate and PPG Varying across Drowsiness

PPG IR vs Drowsiness:

1. The median of PPG IR slightly decreases as drowsiness increase.
2. At drowsiness level 1, there is an equal distribution, and no outliers are present indicating less variability.
3. As there are few outliers at level 2 of the drowsiness, we can somewhat loosely conclude that drowsiness increases as PPG IR value decrease.
4. But considering the highest outliers at the drowsiness level 0 leaves us with weak correlation of PPG IR across different levels of drowsiness.
5. PPG IR is not a standalone strong indicator of drowsiness.

Correlation Matrix





Heat Map and Scatter Plots

```
import seaborn as sns

# Calculating the correlation matrix
correlation_matrix = data.corr()

# Plotting the correlation matrix
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix')
plt.show()

# Scatter plots to explore relationships between PPG signals and drowsiness
fig, axes = plt.subplots(2, 2, figsize=(15, 12))

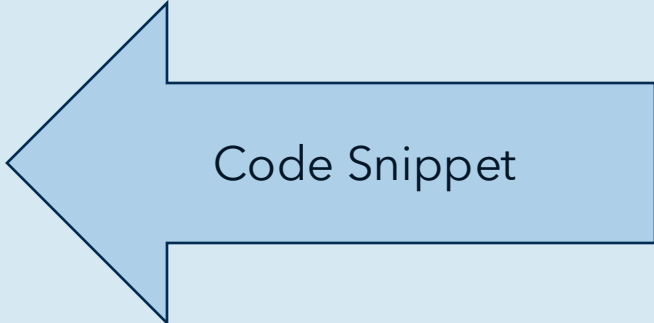
# PPG Green vs Drowsiness
sns.scatterplot(x='ppgGreen', y='drowsiness', data=data.sample(10000), ax=axes[0, 0], alpha=0.3, color='green')
axes[0, 0].set_title('PPG Green vs Drowsiness')

# PPG Red vs Drowsiness
sns.scatterplot(x='ppgRed', y='drowsiness', data=data.sample(10000), ax=axes[0, 1], alpha=0.3, color='red')
axes[0, 1].set_title('PPG Red vs Drowsiness')

# PPG IR vs Drowsiness
sns.scatterplot(x='ppgIR', y='drowsiness', data=data.sample(10000), ax=axes[1, 0], alpha=0.3, color='grey')
axes[1, 0].set_title('PPG IR vs Drowsiness')

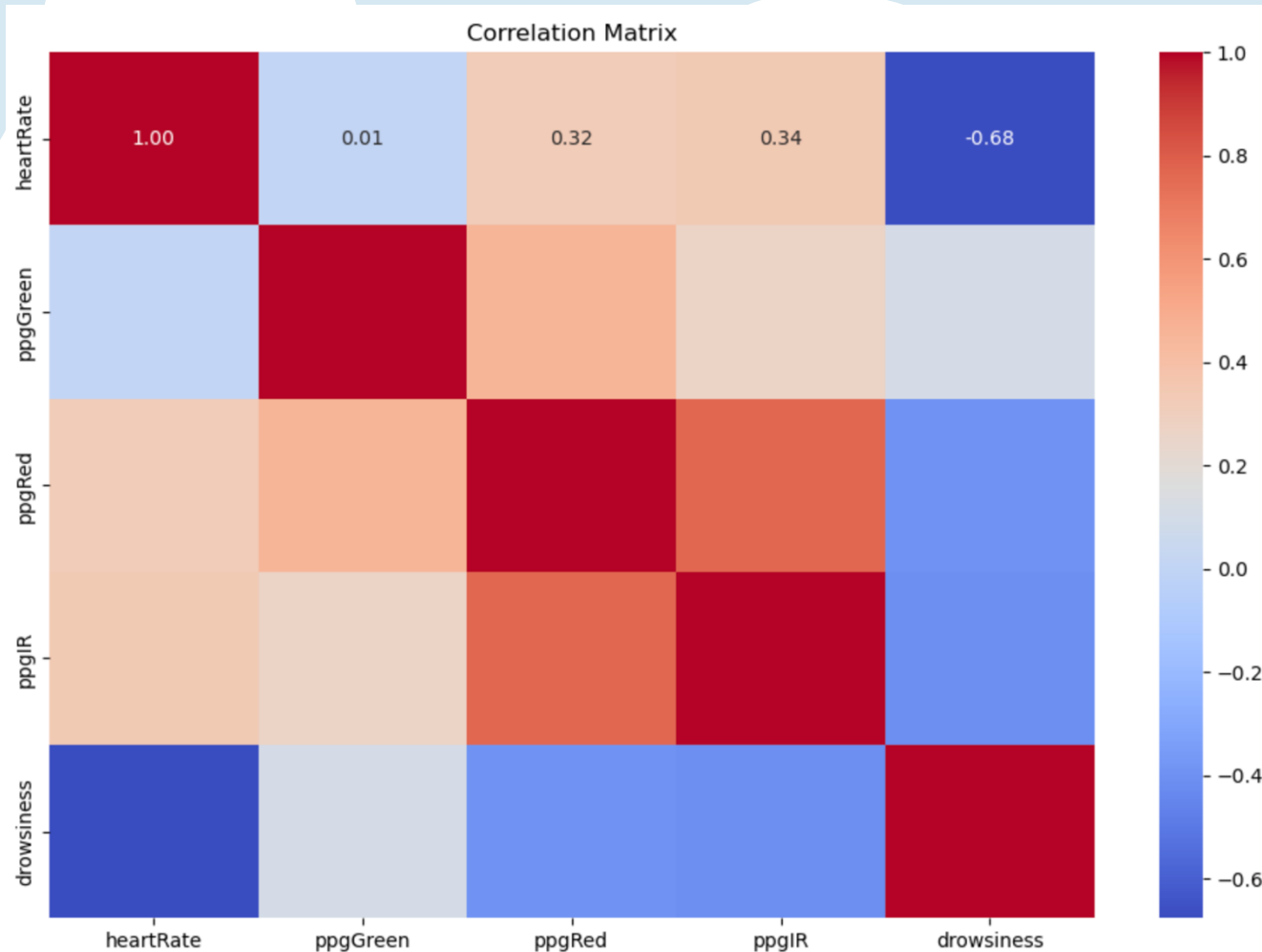
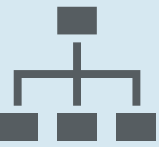
# Heart Rate vs Drowsiness
sns.scatterplot(x='heartRate', y='drowsiness', data=data.sample(10000), ax=axes[1, 1], alpha=0.3, color='#6b0001')
axes[1, 1].set_title('Heart Rate vs Drowsiness')

plt.tight_layout()
plt.show()
```



Code Snippet

Heat Map and Scatter Plots



Heart rate:

- Shows a significant negative correlation with drowsiness, making it a more reliable indicator of drowsiness levels.

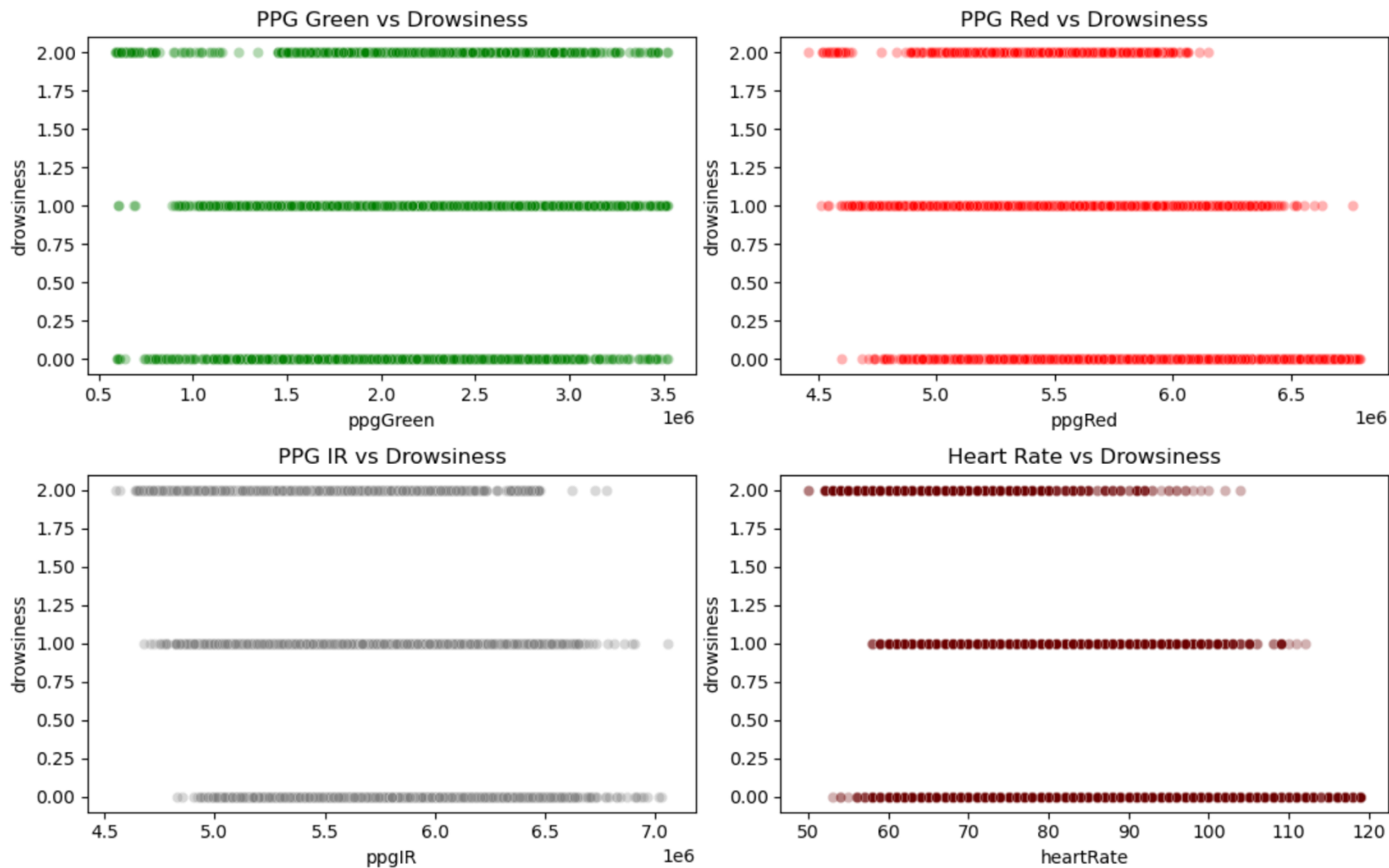
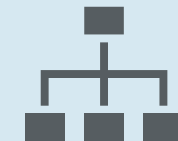
PPG Signals (Red, Green, IR):

- Show strong inter correlations among themselves and moderate positive correlation with heart rate but weak correlation with drowsiness.

Drowsiness:

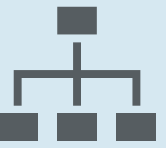
- Best inferred from heart rate rather than PPG signals alone, due to weak correlation of PPG signals with drowsiness.

Heat Map and Scatter Plots

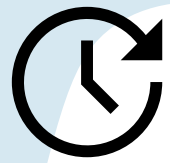




Heat Map and Scatter Plots



PPG signals (Green, Red, IR) are not as good at predicting drowsiness as heart rate is, according to scatter plots. Even though PPG signals are helpful in assessing physiological conditions, they don't significantly correlate with sleepiness on their own, integrating heart rate information with additional physiological and contextual data may yield a more thorough evaluation of sleepiness levels.



Future Improvements and Way forward

Research and Development: Improving the accuracy of drowsiness detection methods by exploring new technologies and sensors that can enhance the quality and reliability of physiological measurement.

Data Quality: Noise reduction in PPG signals by enhancing data collection and integrity and considering data from multiple sensors like temperature sensor, accelerometer, gyroscope which gives the insights behind the outliers.

Machine Learning and AI: Training ML models using labeled data to recognize patterns associated with drowsiness.

Real-time continuous monitoring: with more data, the analysis would be more accurate.

Collaboration with healthcare providers: Integrating technology and healthcare.

↪ Source

Click on 'Jupyter Notebook' below to redirect to the google drive link of the the Jupyter Notebook of the above analysis.

[Jupyter Notebook](#)

