Health Care Exploratory Data Analysis

—

By Jalal Haider

Structure

1. Theoretical Information about the Data Set
2. Data Cleaning and Processing
3. Descriptive Statistics
4. Data Visualization (Histograms)
5. Correlation between Drowsiness & PPG Wavelengths
6. Time period Analysis
7. Key findings & Takeaways
8. Certain Limitations

Theoretical Information about the Data Set

- The data set is showing different heart rate levels and their corresponding drowsiness levels, the drowsiness levels are ranging from 0, 1 or 2)
- The ppg measurements (red, green and infrared(IR) are different wavelengths that describe the intestinal wavelength absorption of the tissue when these devices are worn
- The data set is extremely large and the number of entries taken are ranging to approximately 4.9million samples
- There are different types of devices and they can be worn in different ways, i.e: wrist, forehead, ear etc

Data Cleaning and Processing

- Data Extracted from Kaggle named file as ‘drowsiness_dataset.csv’
- Looked at the dataset and derived the data types of each column, found in Fig 3.1
- Data was cleaned by checking for any null values (none were found as seen on Fig 3.2)

	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.3.1

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

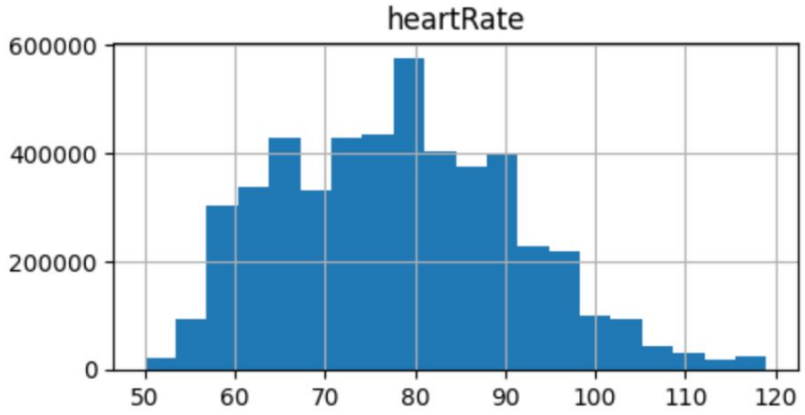
Fig 3.2

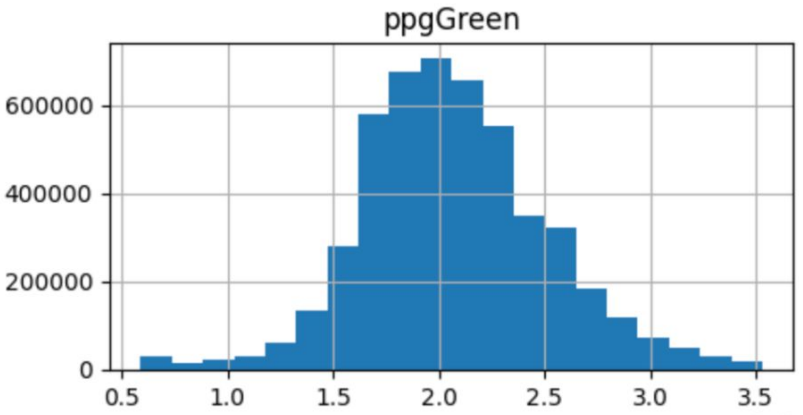
Descriptive Statistics

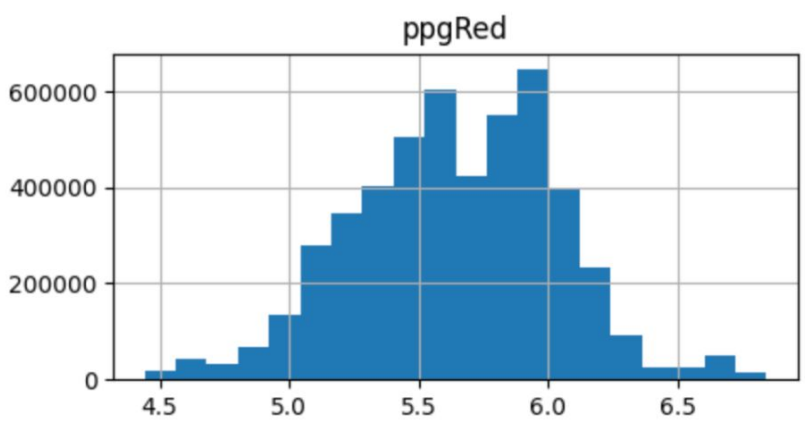
	heartRate	ppgGreen	ppgRed	ppgIR	drowsiness
count	4.890260e+06	4.890260e+06	4.890260e+06	4.890260e+06	4.890260e+06
mean	7.814245e+01	2.073589e+06	5.643653e+06	5.728191e+06	8.593592e-01
std	1.296635e+01	4.418773e+05	3.909626e+05	4.313052e+05	8.370285e-01
min	5.000000e+01	5.897580e+05	4.441989e+06	4.409976e+06	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

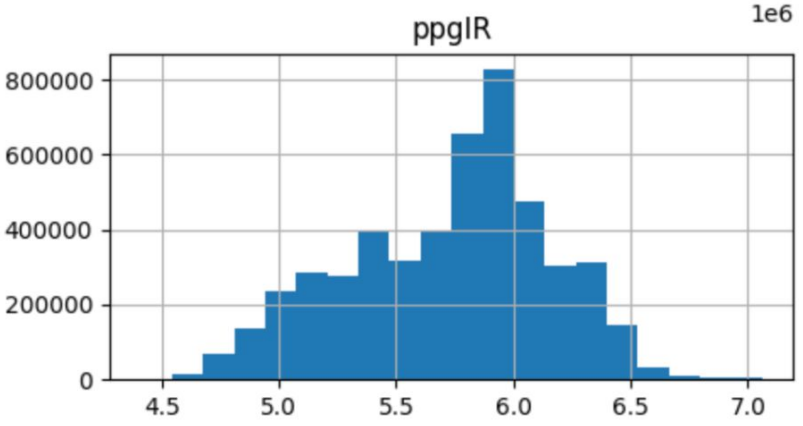
- The dataset comprises a large number of observations (nearly 4.9 million) for each variable, providing a comprehensive overview of heart rate, PPG signals in green, red, and infrared, and drowsiness levels.
- The heart rate data shows moderate variability with a mean around 78 bpm. The PPG signals have considerable variability, especially the green and infrared signals, with their means in the millions of units.
- The drowsiness level is generally low, with a mean below 1 and a maximum of 2.x

Data Visualization





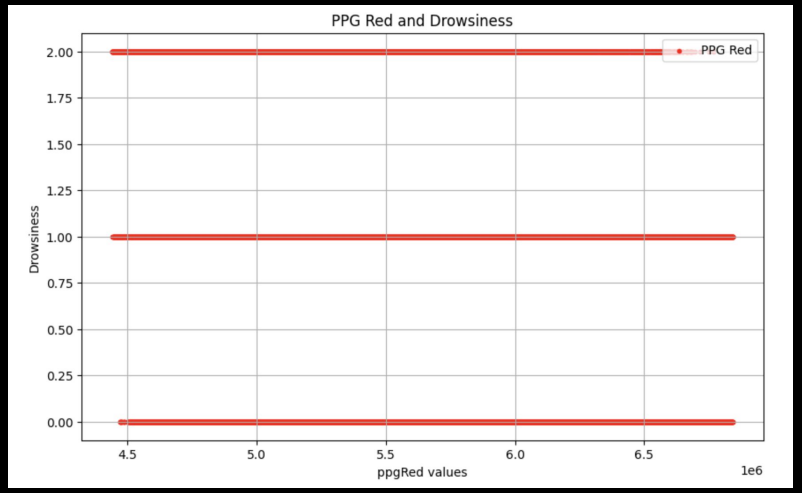


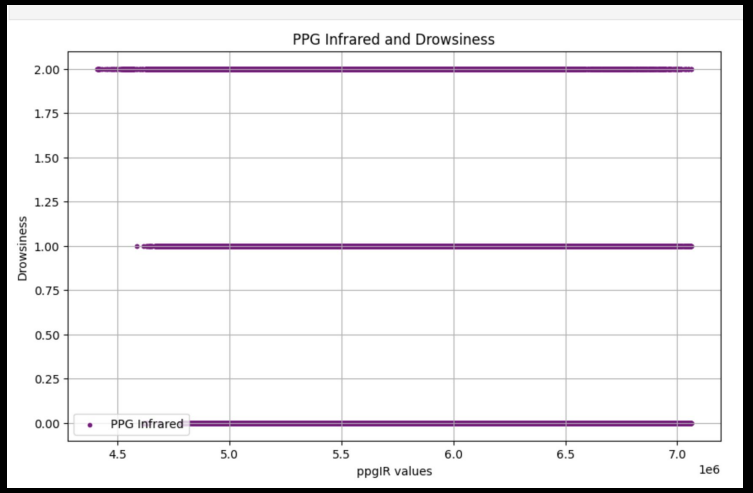


Histogram Explanations

- **heartRate**: The distribution is approximately normal, peaking around 80 bpm, with most values between 60 and 100 bpm, and a high concentration between 75 and 85 bpm
- **ppgGreen**: The distribution is right-skewed, peaking around 2 million units, with most values between 1.5 million and 2.5 million units
- **ppgRed**: The distribution is approximately normal, peaking around 5.6 million units, with most values between 5 million and 6.5 million units, and a high concentration between 5.5 million and 6 million units
- **ppgIR**: The distribution is approximately normal, peaking around 6 million units, with most values between 5 million and 6.5 million units, and a notable peak around 6 million units

Variation between Drowsiness and PPG Types

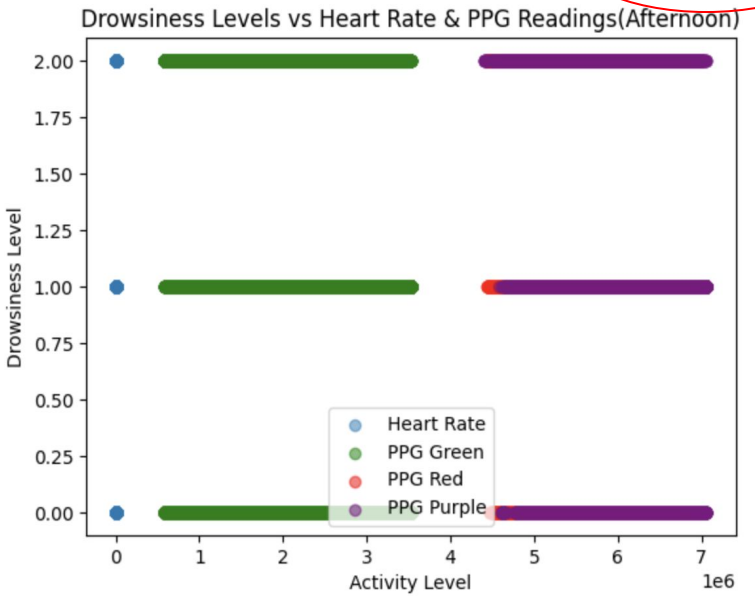




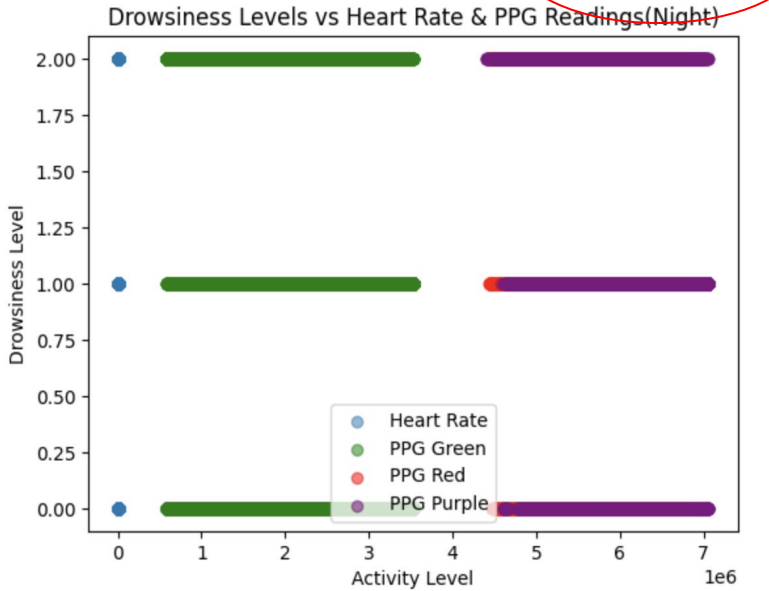
- We can see that PPG red has less variation with drowsiness levels as the line seems to slightly increase from 0 to 1
- On the other hand, PPG IR has a larger variability as the drowsiness level increases from 0 to 2, the difference between the levels has dramatically increased

Variations to different times of the Day(Afternoon & Night)

Correlation between drowsiness and heart rate(Afternoon): -0.7295576553744954
Correlation between drowsiness and PPG green(Afternoon): 0.24140212200218605
Correlation between drowsiness and PPG red(Afternoon): 0.036138940883998674
Correlation between drowsiness and PPG Infrared(Afternoon): -0.2655153229308298



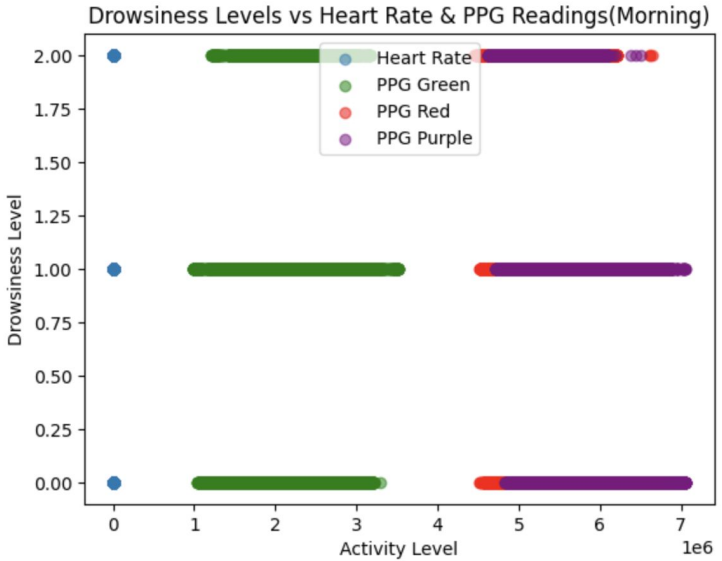
Correlation between drowsiness and heart rate(Night): -0.7295576553744954
Correlation between drowsiness and PPG green(Night): 0.24140212200218605
Correlation between drowsiness and PPG red(Night): 0.036138940883998674
Correlation between drowsiness and PPG Infrared(Night): -0.2655153229308298



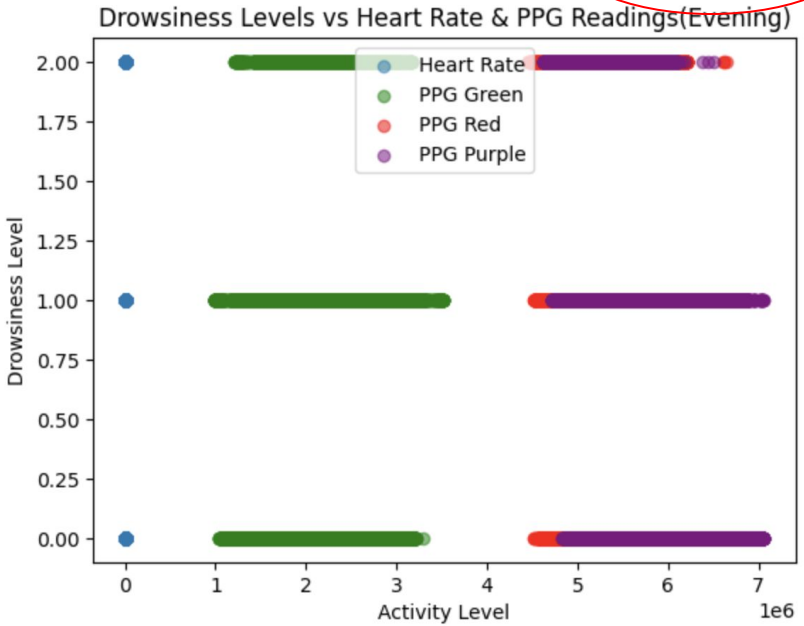
Variations to different times of the Day(Morning & Evening)

```
Correlation between drowsiness and heart rate(Morning): -0.6319114762754944
Correlation between drowsiness and PPG green(Morning): -0.08278873014767459
Correlation between drowsiness and PPG red(Morning): -0.6565096863047643
Correlation between drowsiness and PPG Infrared(Morning): -0.5784903897390827
```

```
/Library/Frameworks/Python.framework/Versions/3.10/lib/python3.10/site-packages
with loc="best" can be slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```




```
Correlation between drowsiness and heart rate(Evening): -0.6319114762754944
Correlation between drowsiness and PPG green(Evening): -0.08278873014767459
Correlation between drowsiness and PPG red(Evening): -0.6565096863047643
Correlation between drowsiness and PPG Infrared(Evening): -0.5784903897390827
```



Key Findings & Takeaways

- When the data set was divided into 4 different time periods, the variation in drowsiness and the PPG types was the same as in the Day & Evening
- The same case mentioned above was the same in the Afternoon and Night as well
- This idea concludes that the activity levels and the drowsiness levels in those given time frames are very much similar
- The Photoplethysmography types (red, green and infra-red all have varying ranges and therefore, capture different variations to the drowsiness levels

Certain Limitations

- The dataset seems to be biased, the range of the drowsiness values only ranges from 0 to 2, whereas the Karolinska Sleepiness Scale (KSS) has a longer range (0 to 9)
- More evidence is needed to back up the fact that the different time frames are in fact, related to the times of the day, the sample size is only confined to the heart rate ranging from 54.0 to 63.0
- The data set also does not consider people with heart diseases/disabilities as its normal distribution shows that there are not any outliers

Bibliography

https://web.stanford.edu/class/ee368/Project_Winter_1819/Reports/prabhu_barnhart_seetharaman.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6426305/

https://www.optalert.com/what-is-drowsiness-and-how-can-it-be-measured/#:~:text=The%20most%20common%20measurement%20of,to%209%20(see%20table).

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