

Introduction

The following project documentation was written as work assignment for the module "Multivariate Analysis" of the Master in Statistics for Data Science at the Universidad Carlos III de Madrid. It contains the Multivariate Analysis of a Kaggel dataset on Sleep Health and Lifestyle (<https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>). The work is split into two parts, where in a first part a exploratory data analysis is performed, where required, data preprocessing steps are performed and a Prinicipal Component Analysis (PCA) is performed. In the second part, based on the learnings of part one, a (XXXXXXXXXX) is performed to (XXXXXXXXXX).

The dataset at hand is composed out of the collumns shown in the table below. It has been modified compared to the kaggle source data by turning the "Sleep Disorder" Variable into a binary variable (yes/no) and by seperating the blood pressure variables into the two variables blood pressure systolic and blood pressure diastolic.

Variable	Description
Person ID	An identifier for each individual.
Gender	The gender of the person (Male/Female).
Age	The age of the person in years.
Occupation	The occupation or profession of the person.
Sleep Duration (hours)	The number of hours the person sleeps per day.
Quality of Sleep (scale: 1-10)	A subjective rating of the quality of sleep, ranging from 1 to 10.
Physical Activity Level (minutes/day)	The number of minutes the person engages in physical activity daily.
Stress Level (scale: 1-10)	A subjective rating of the stress level experienced by the person, ranging from 1 to 10.
BMI Category	The BMI category of the person (e.g., Underweight, Normal, Overweight).
Blood Pressure (systolic)	The blood pressure measurement of the person (systolic pressure)
Blood Pressure (diastolic)	The blood pressure measurement of the person (diastolic pressure)
Heart Rate (bpm)	The resting heart rate of the person in beats per minute.
Daily Steps	The number of steps the person takes per day.

Variable	Description
Sleep Disorder	The presence or absence of a sleep disorder in the person (Binary)

Furthermore some variable renaming and data type modifications are performed. All these preprocessing steps are performed in the following code chunk.

Based on the value counts and the variable description, the numeric and categorical variables are identified and two lists are set up containing the variables for later use.

```
Out[124...  person_id          374
          gender           2
          age            31
          occupation      11
          sleep_duration   27
          quality_of_sleep  6
          physical_activity_level 16
          stress_level      6
          bmi_category      4
          blood_pressure   25
          heart_rate       19
          daily_steps      20
          sleep_disorder    2
          blood_pressure_systolic 18
          blood_pressure_diastolic 17
          dtype: int64
```

Reviewing the unique values of the identified categorical values, a duplicate in bmi_category for "Normal" and "Normal Weight" can be identified.

```
Out[125...  gender          [Male, Female]
          occupation  [Software Engineer, Doctor, Sales Representati...
          quality_of_sleep  [6, 4, 7, 5, 8, 9]
          stress_level   [6, 8, 7, 4, 3, 5]
          bmi_category   [Overweight, Normal, Obese, Normal Weight]
          sleep_disorder  [0, 1]
          dtype: object
```

This is solved by replacing all instances of "Normal Weight" with "Normal"

```
Out[126...  bmi_category
0    Overweight
1    Normal
2    Obese
```

Doing so concludes the required preprocessing steps and allows to begin with the first part of this project work.

Part 1 - Exploratory Analysis and Dimension Reduction via PCA

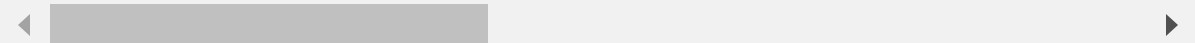
The first part of this work contains the initial exploratory analysis of the dataset as well as a Principal Component Analysis (PCA) of the dataset.

Explortary Data Analysis

The Exploratory Data Analysis contains a general overview of the datasets structure and correlations. To begin, the first 5 rows of the data Set are shown to give a first insight into the structure of the dataset.

Out[127...

	person_id	gender	age	occupation	sleep_duration	quality_of_sleep	physical_activity
0	1	Male	27	Software Engineer	6.1	6	
1	2	Male	28	Doctor	6.2	6	
2	3	Male	28	Doctor	6.2	6	
3	4	Male	28	Sales Representative	5.9	4	
4	5	Male	28	Sales Representative	5.9	4	



The categorical Variables can be sperated into the following categories with:

One Binary variable:

- gender
- sleep disorder

Three ordinal variables:

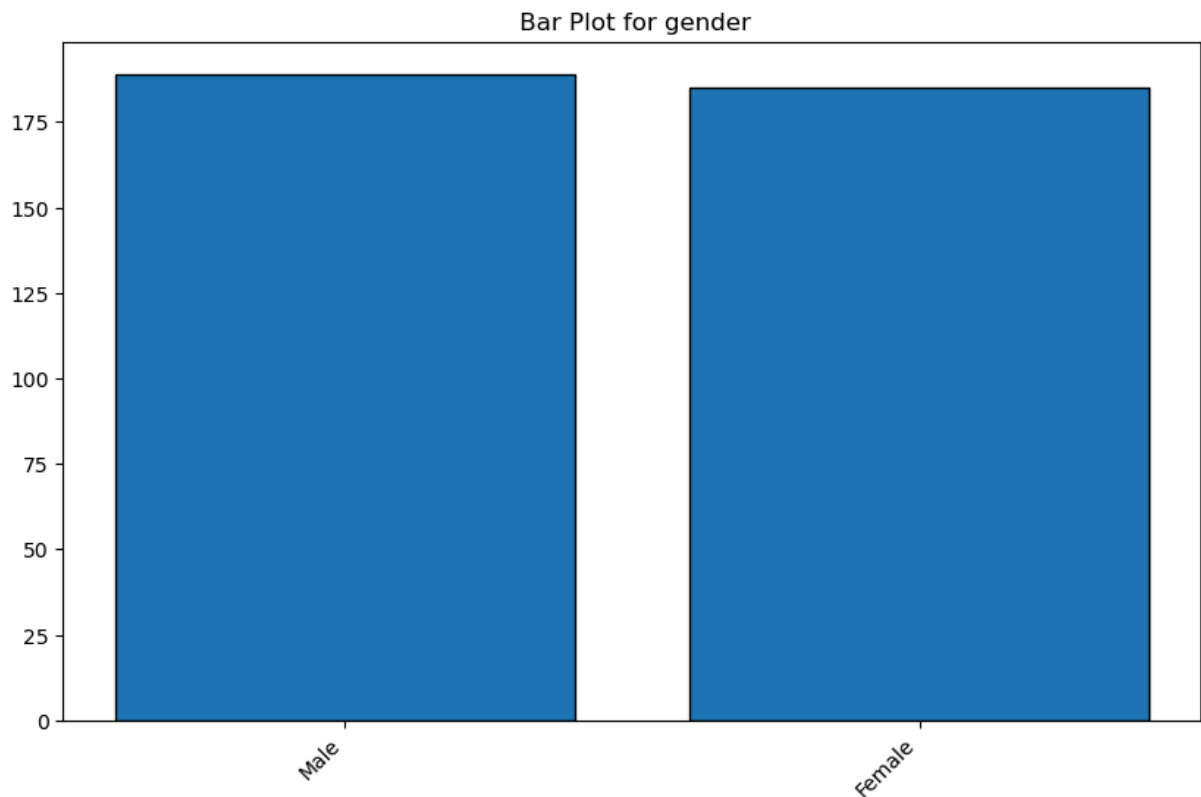
- quality of sleep
- stress level
- bmi_category

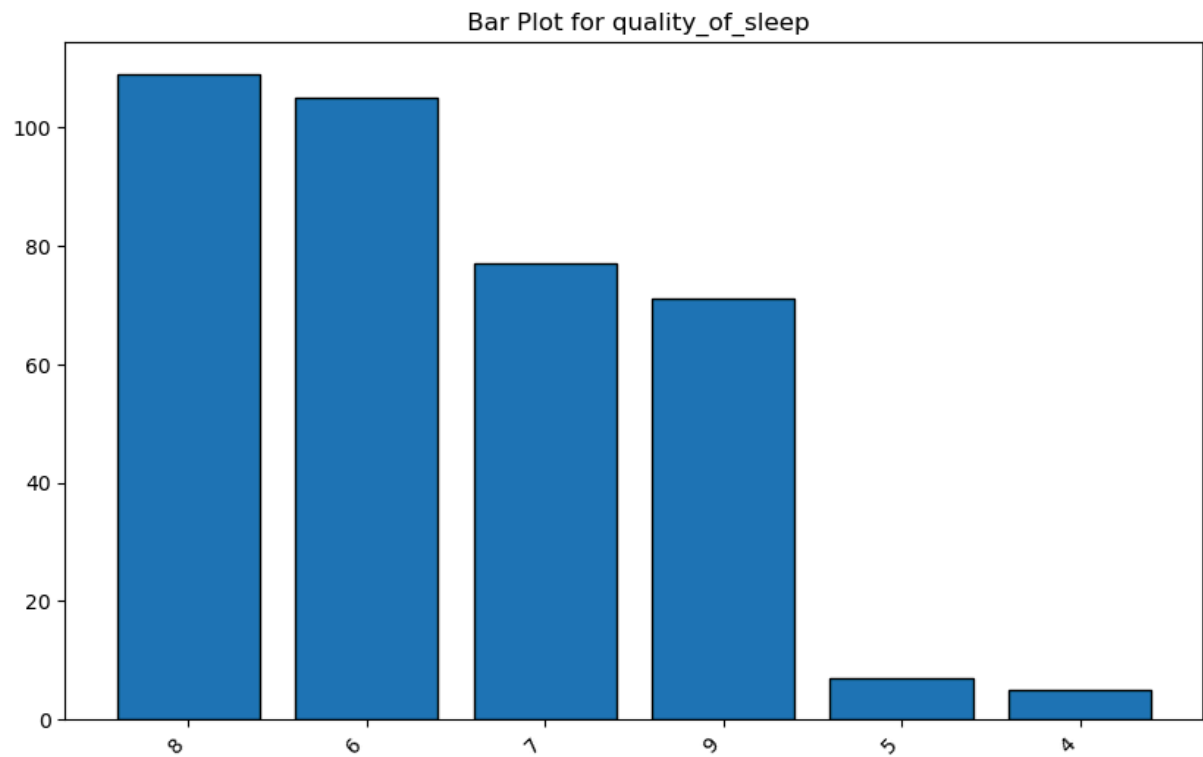
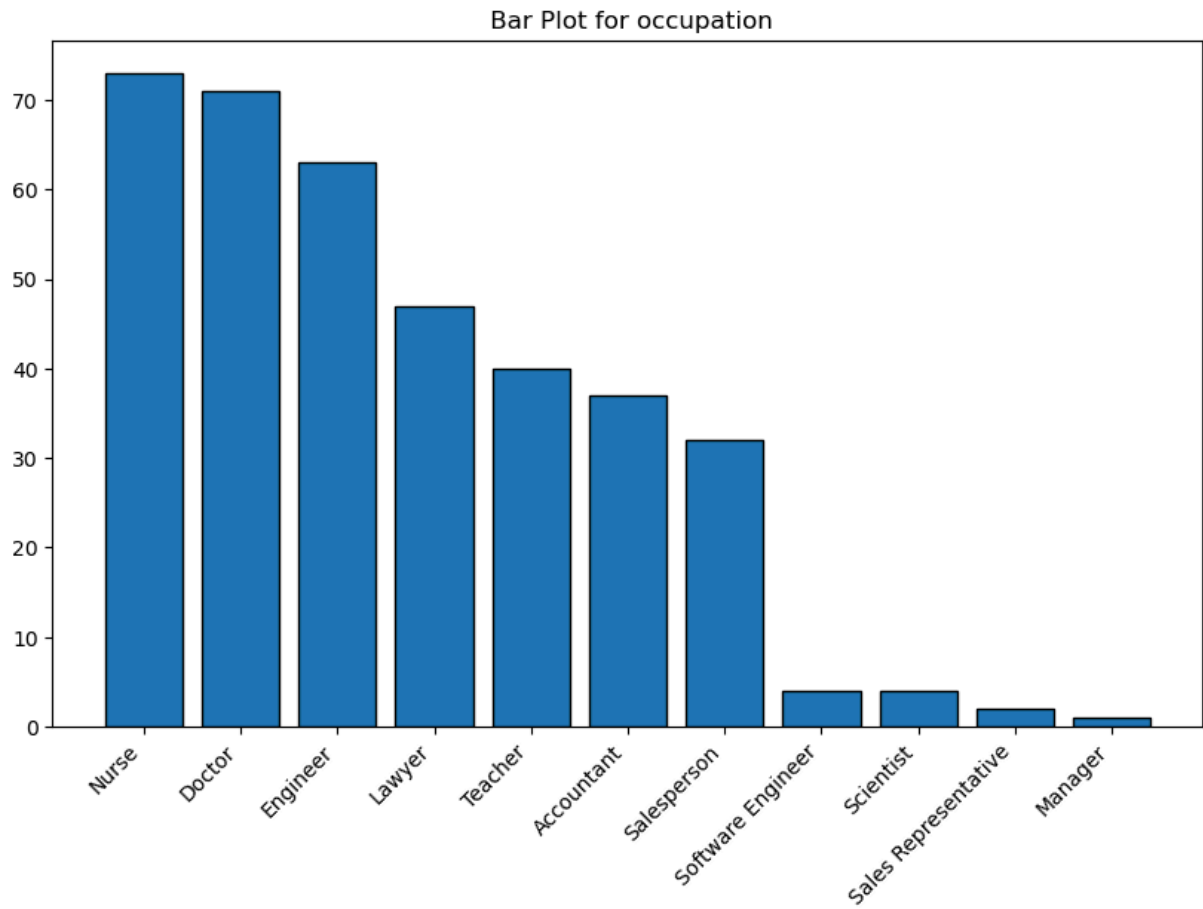
And one nominal variables:

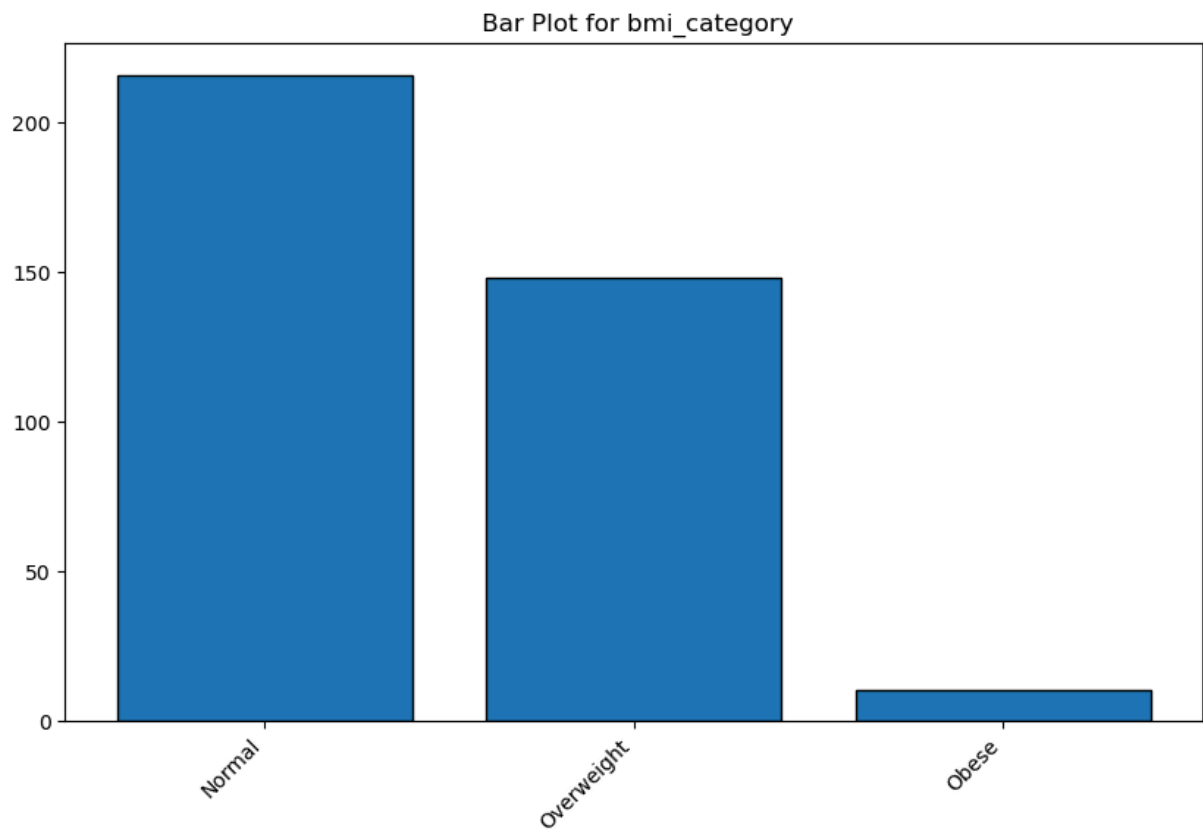
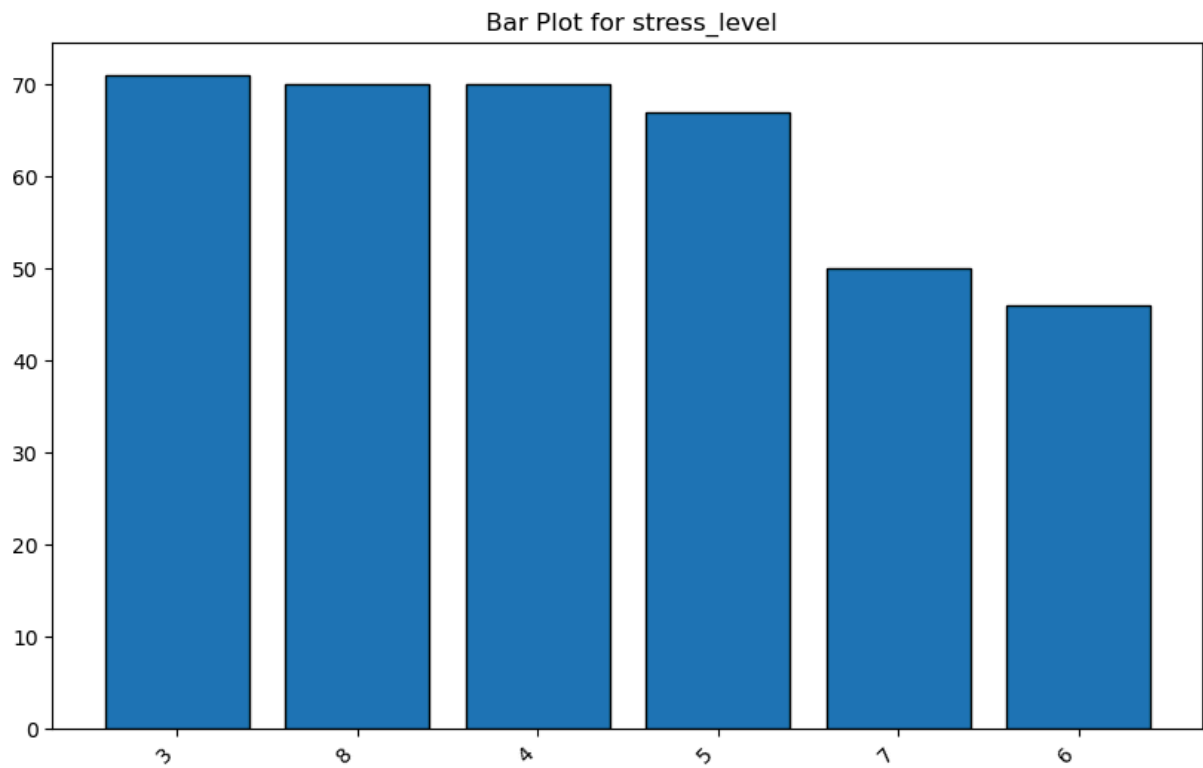
- occupation

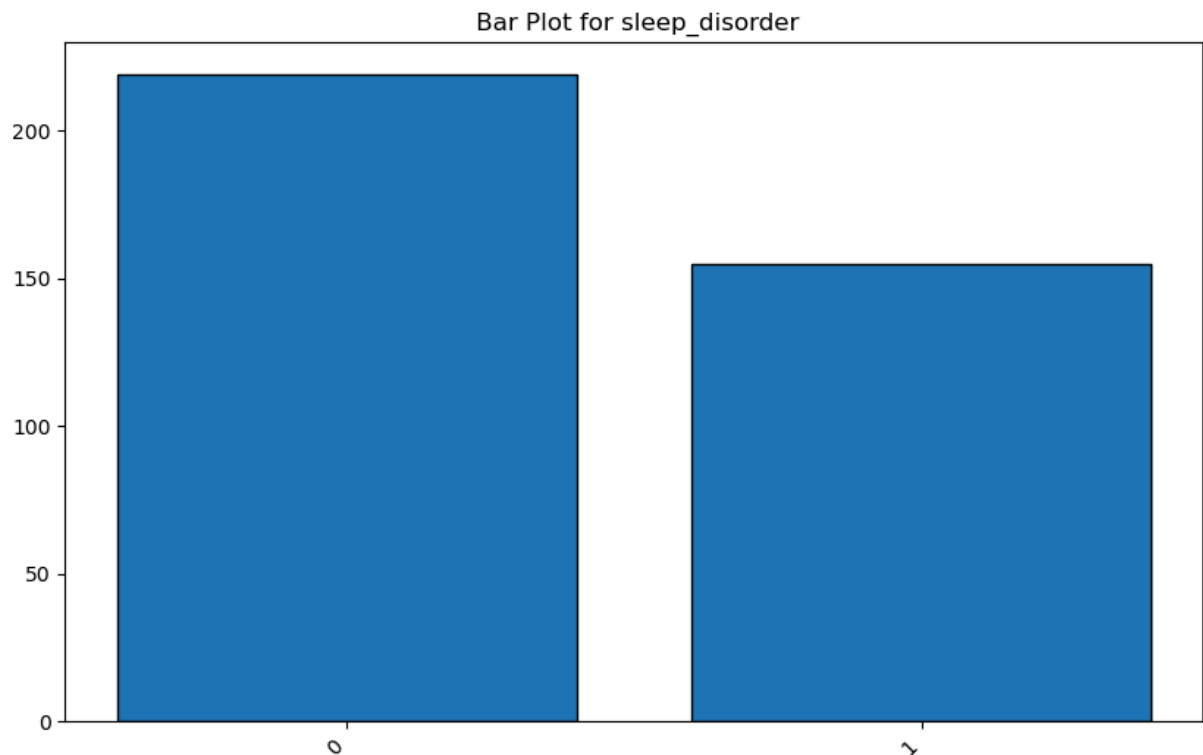
Plotting barplots for the categorical variables, beginning with the binary variables of gender and sleep disorder shows an even distribution between male and female and a fairly even distribution between observations with and without sleep disorder. In the meantime, stress level shows a decrease in observations towards higher stress level, and equally quality of sleep and body mass index (bmi) show a continuous decrease in observations for quality of sleep from 8 to 4 and for a bmi categories from normal to obese. Lastly, the variable occupation shows a somewhat uneven distribution of observations between the different occupations, with the two occupations with the individually biggest contribution to the overall dataset are Nurses and Doctors. The bias a potential overrepresentation of the medical field with irregular working ours in shifts can not be futhere anaylzed in this work, but has to be taken into account in later conclusions.

Categorical Values: ['gender', 'occupation', 'quality_of_sleep', 'stress_level', 'bmi_category', 'sleep_disorder']









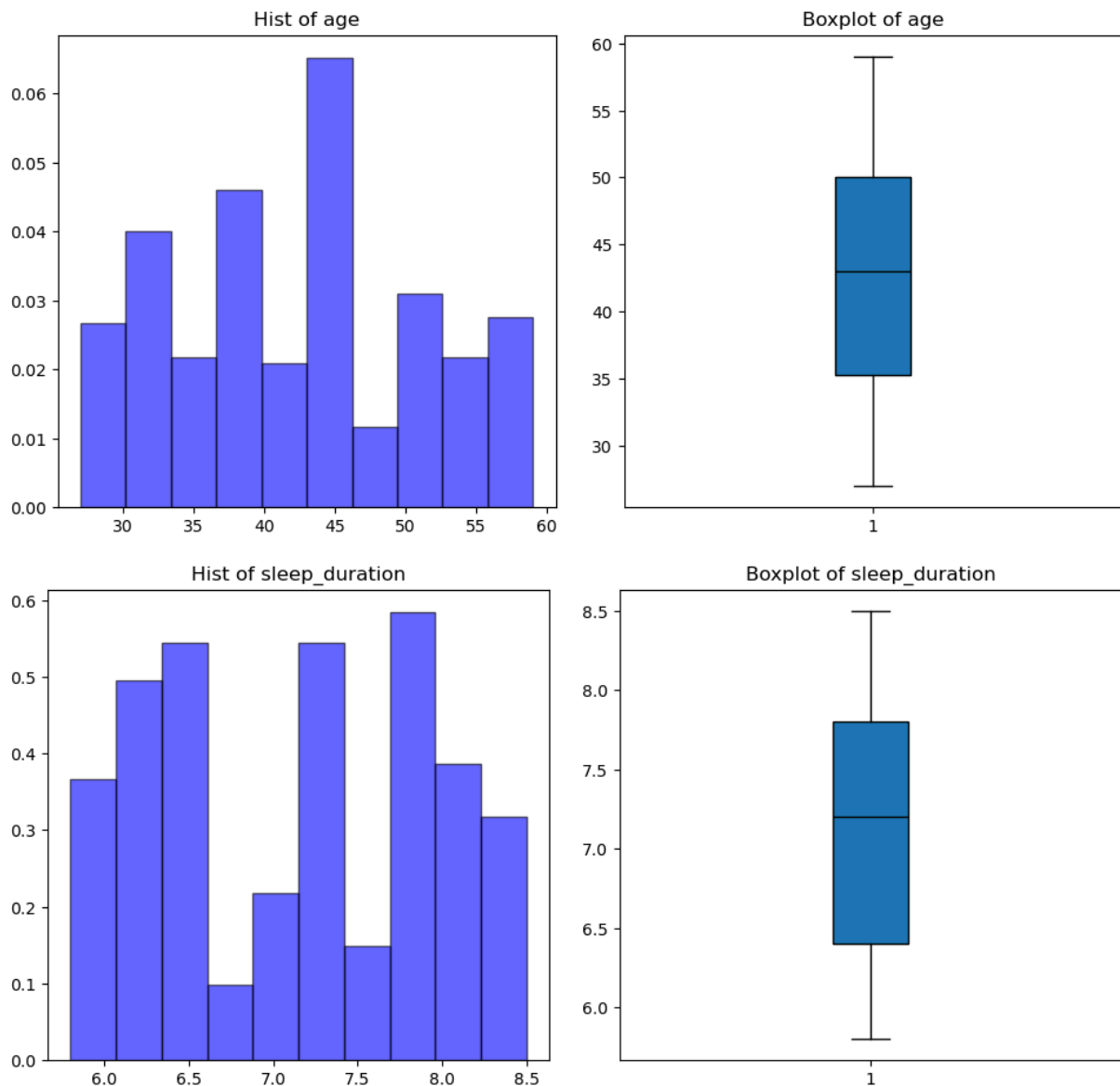
Of the seven numerical variables, all are continuous. Plotting histograms and boxplots side by side, two categories emerge:

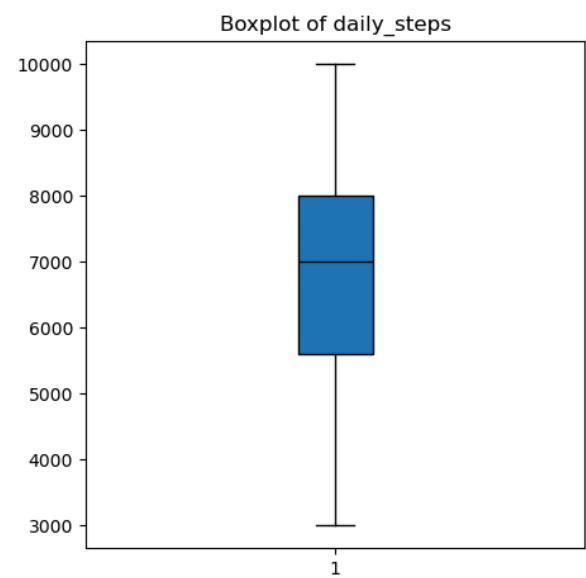
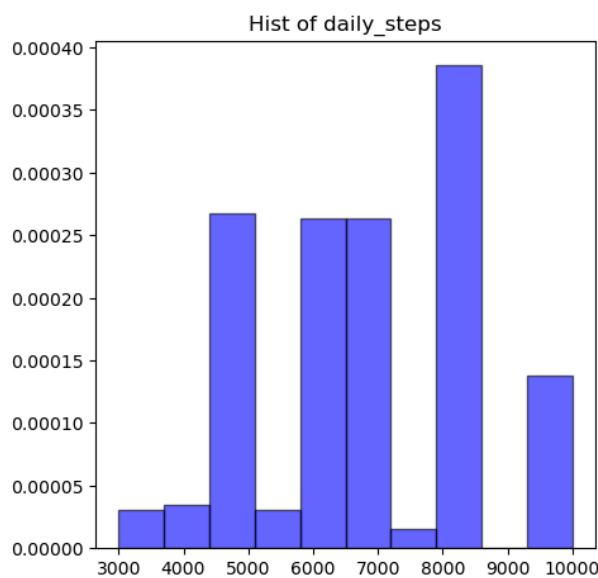
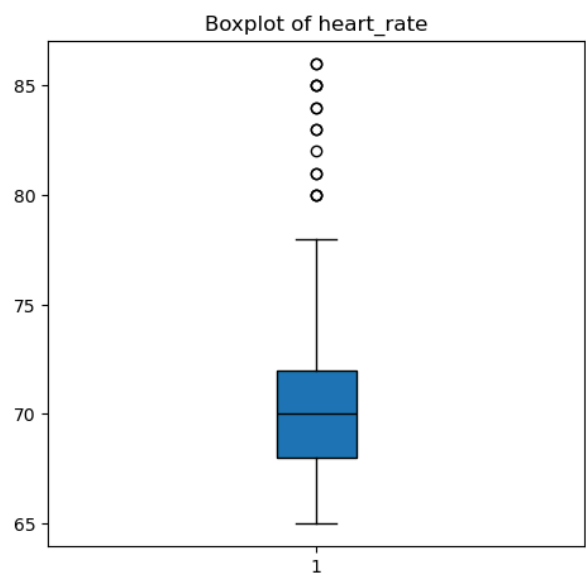
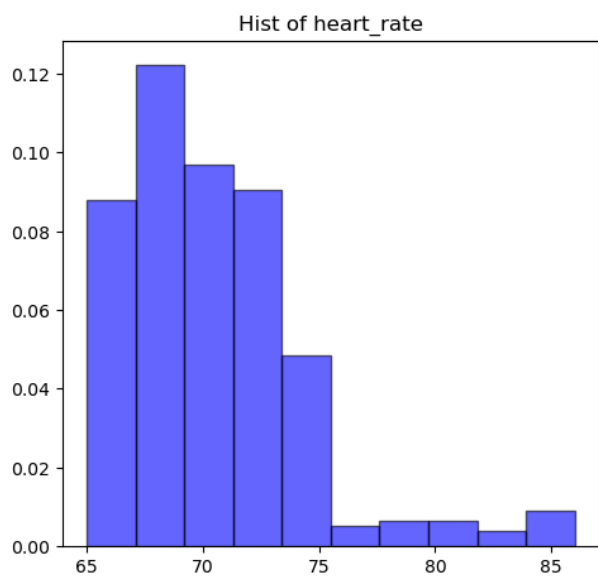
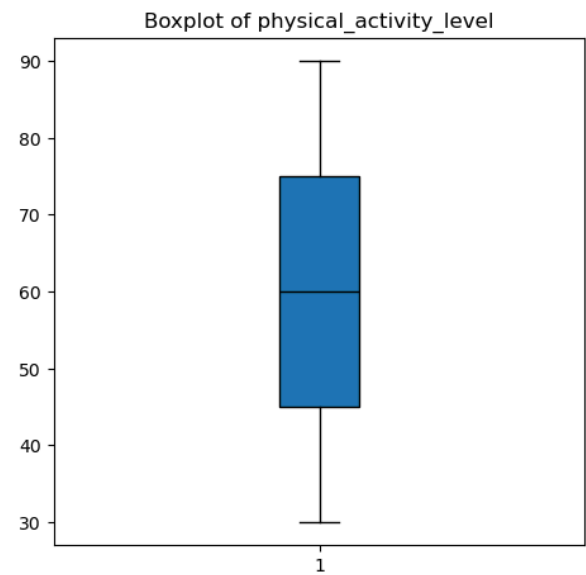
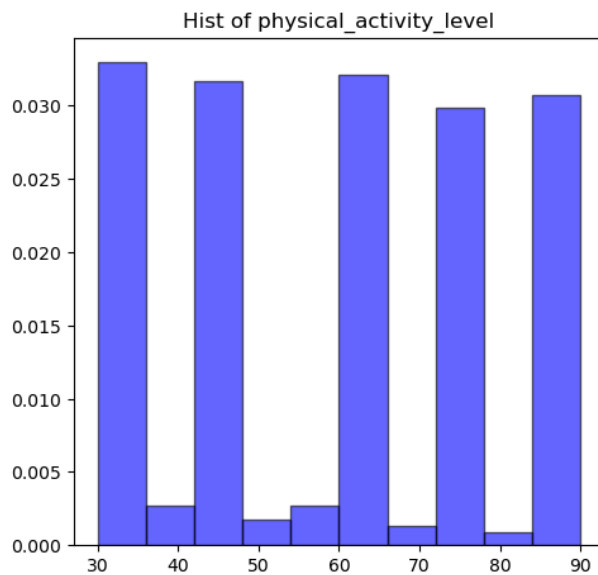
- measured variables
- estimated/rounded variables (variables that probably were estimated as part of a questionnaire by its participants)

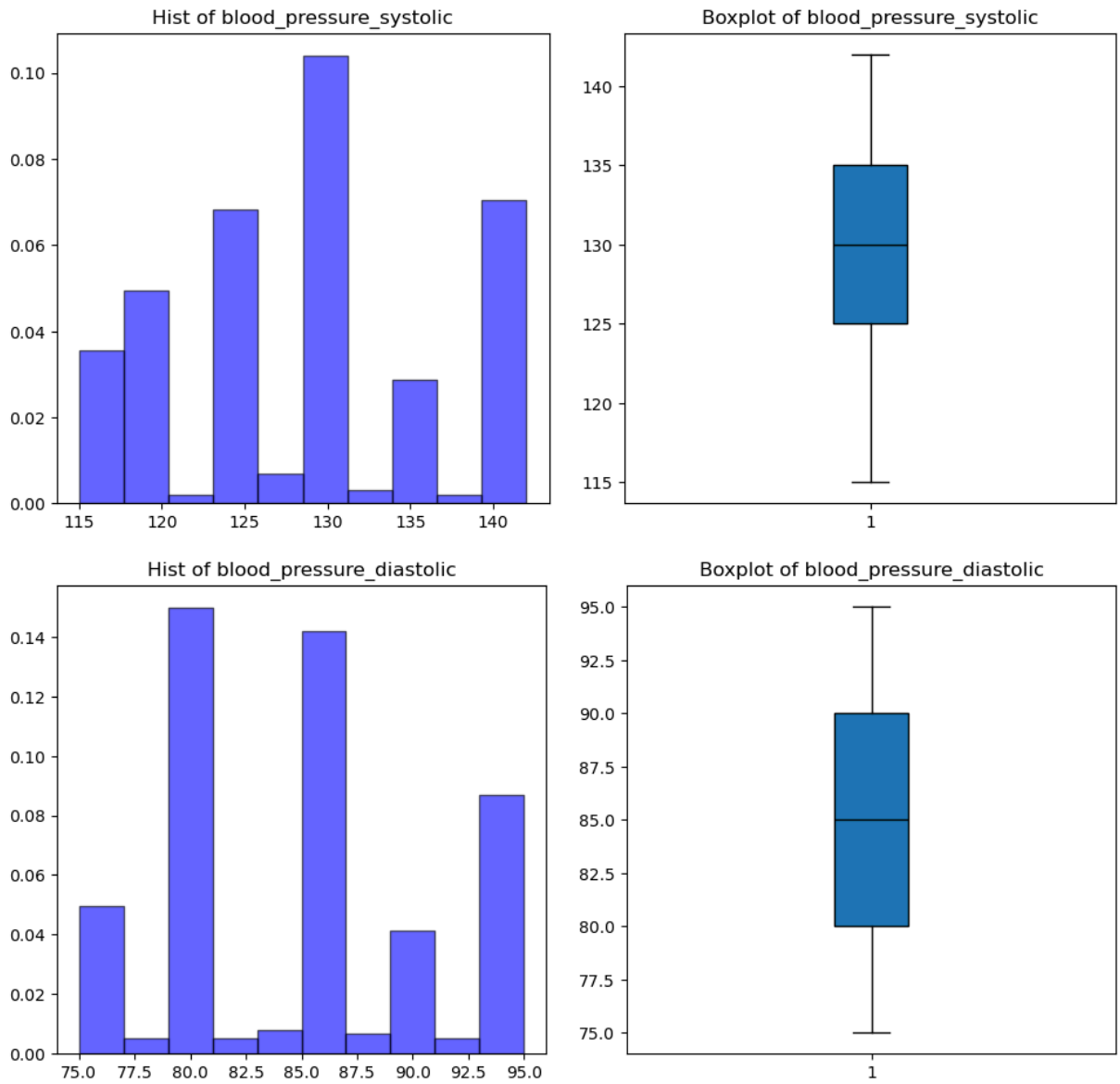
For the measured variables age, heart rate and sleep duration, a more or less continuous distribution can be identified for age, with while showing variance still is somewhat uniformly distributed in between the limits of around 30 to 60. Furthermore, appears to show three groups with one group having very little sleep (up to 6,5 hours) and one group sleeping for longer times (more than 8 hours), while in between these two a dip in sleeping time can be observed. As last variable of this group, the heart rate is shown to have the majority of its observations in between 65 and 75 with visual hints to a normal distribution, but shows both in histogram and boxplot significant outliers for higher heart rates which will have to be addressed in the preprocessing step.

The estimated/rounded variables (physical activity level, daily steps, blood_pressure_systolic/diastolic) shows a specific characteristic with a back and forth between highs and lows which can be attributed to the way people estimate numeric values in increments, like evaluating the physical activity level (mins/day) mostly in increments of 15 min (half an hour, 45 min, one hour, etc.) The lows in between then show observations with more specific answer like 42 min, leading to the somewhat irregular appearance of the histograms. Taking this into account, the physical activity level shows a fairly uniform distribution of observation, while daily steps tends more into the direction of a right skewed uniform distribution, leading to an overall potentially above average fit sample of people.

The high percentage of medical workers with a great walking distances as part of their profession might further contribute to this. Meanwhile, the blood preassure systolic/diastolic appears to both be more or less evenly distributed.







As a first attempt to evaluate the correlations found in this dataset, the following set of Metrics is applied and plotted.

$$q_1 = \left(1 - \frac{\min \lambda_j}{\max \lambda_j}\right)^{p+2},$$

$$q_2 = 1 - \frac{p}{\sum_{j=1}^p (1/\lambda_j)},$$

$$q_3 = 1 - \sqrt{|R|},$$

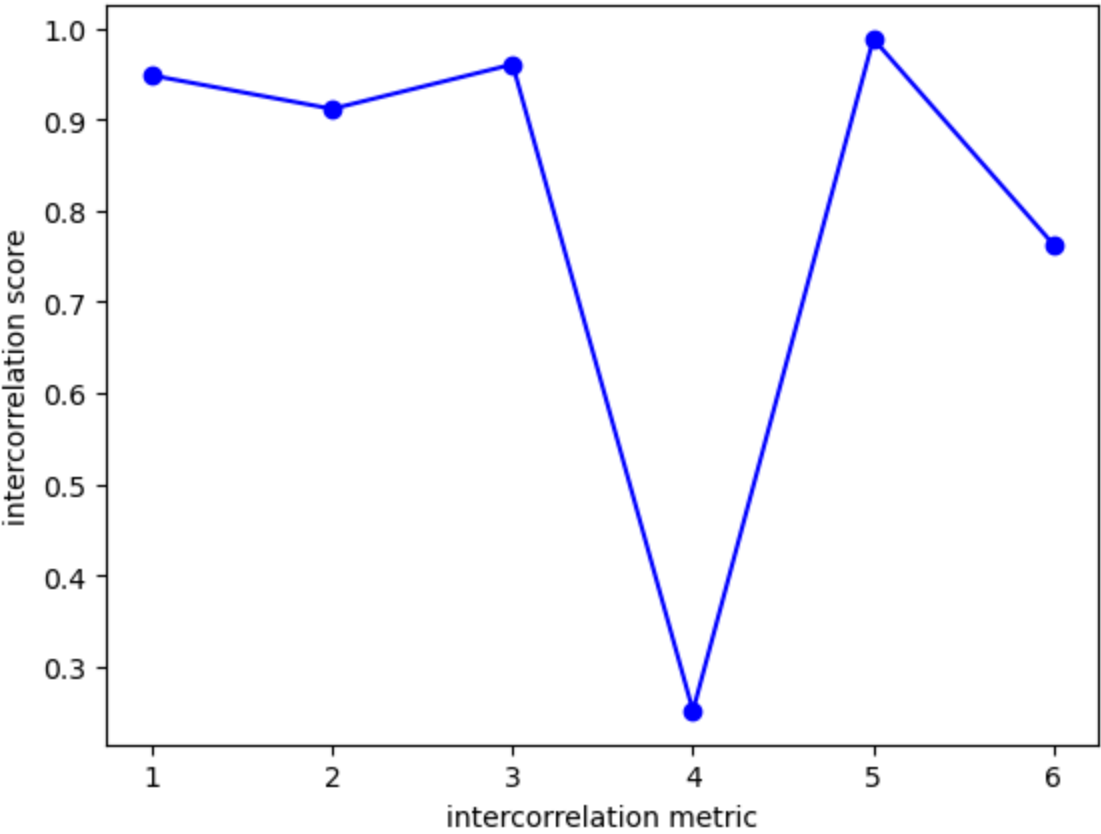
$$q_4 = \left(\frac{\max \lambda_j}{p}\right)^{3/2},$$

$$q_5 = \left(1 - \frac{\min \lambda_j}{p}\right)^5,$$

$$q_6 = \sum_{j=1}^p \frac{1 - 1/r_{ij}}{p}$$

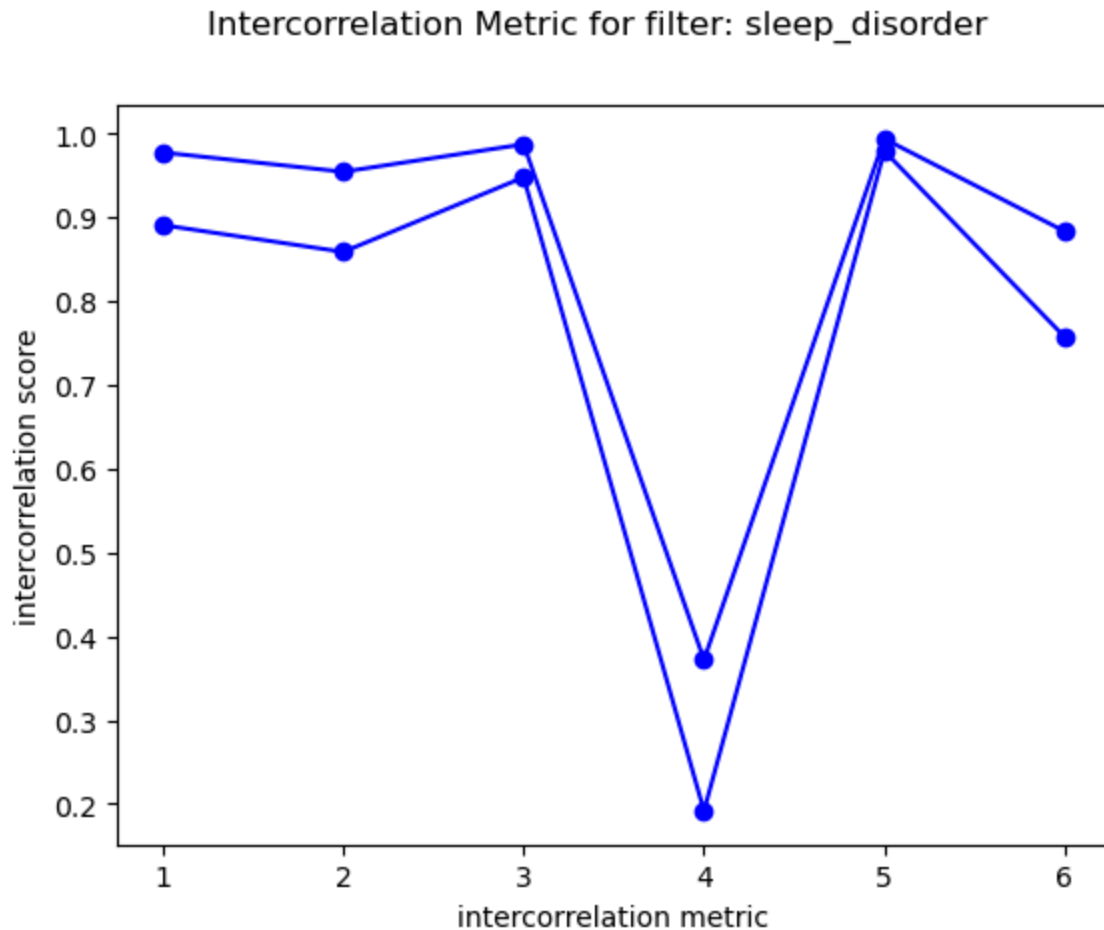
The resulting plot shows high correlation metric scores for all metric except for metric 4.

[0.948209 0.9115284 0.96033816 0.25164604 0.98831354 0.76247781]
Intercorrelation Metric for filter: whole Dataset

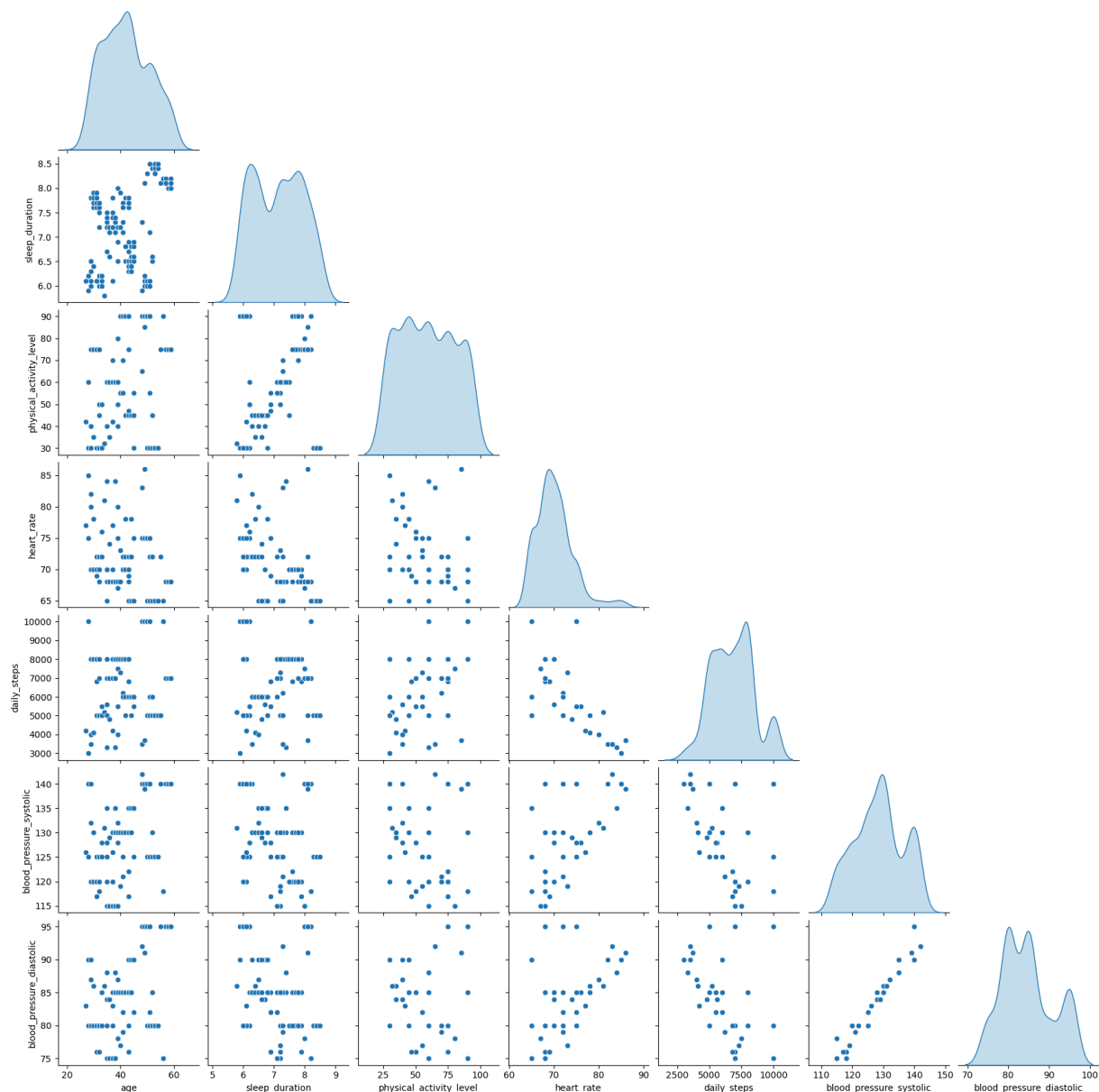


Applying the same method to a dataset filtered on the binary variable sleep disorder shows an overall higher than before correlation in the subset for (?????) while the subset for (?????) shows lower correlation metrics than the joined dataset.

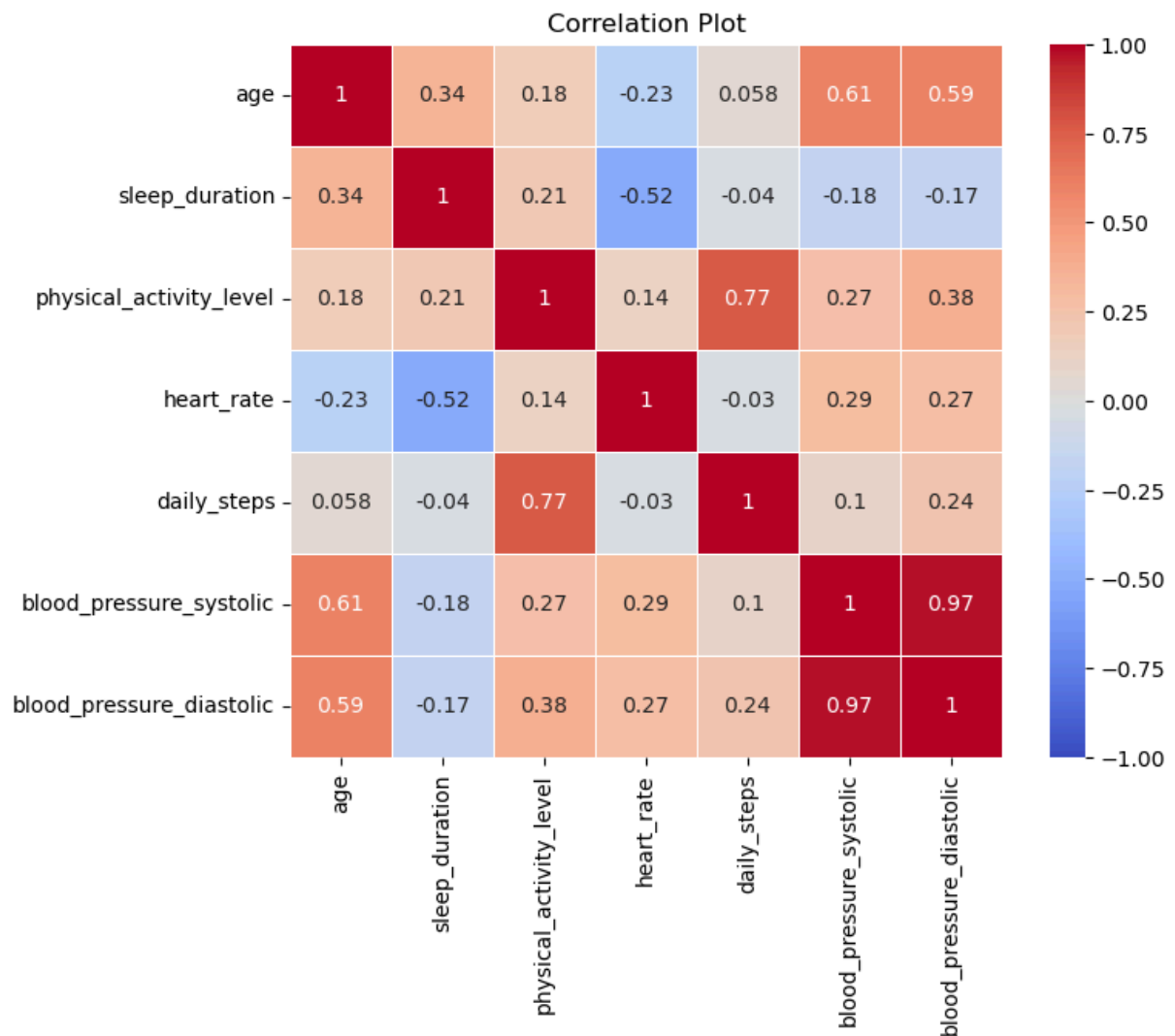
[0.89043129 0.85848947 0.9475162 0.19168981 0.97888426 0.75689052]
[0.97710909 0.95408716 0.98689285 0.37193839 0.99337259 0.88299737]



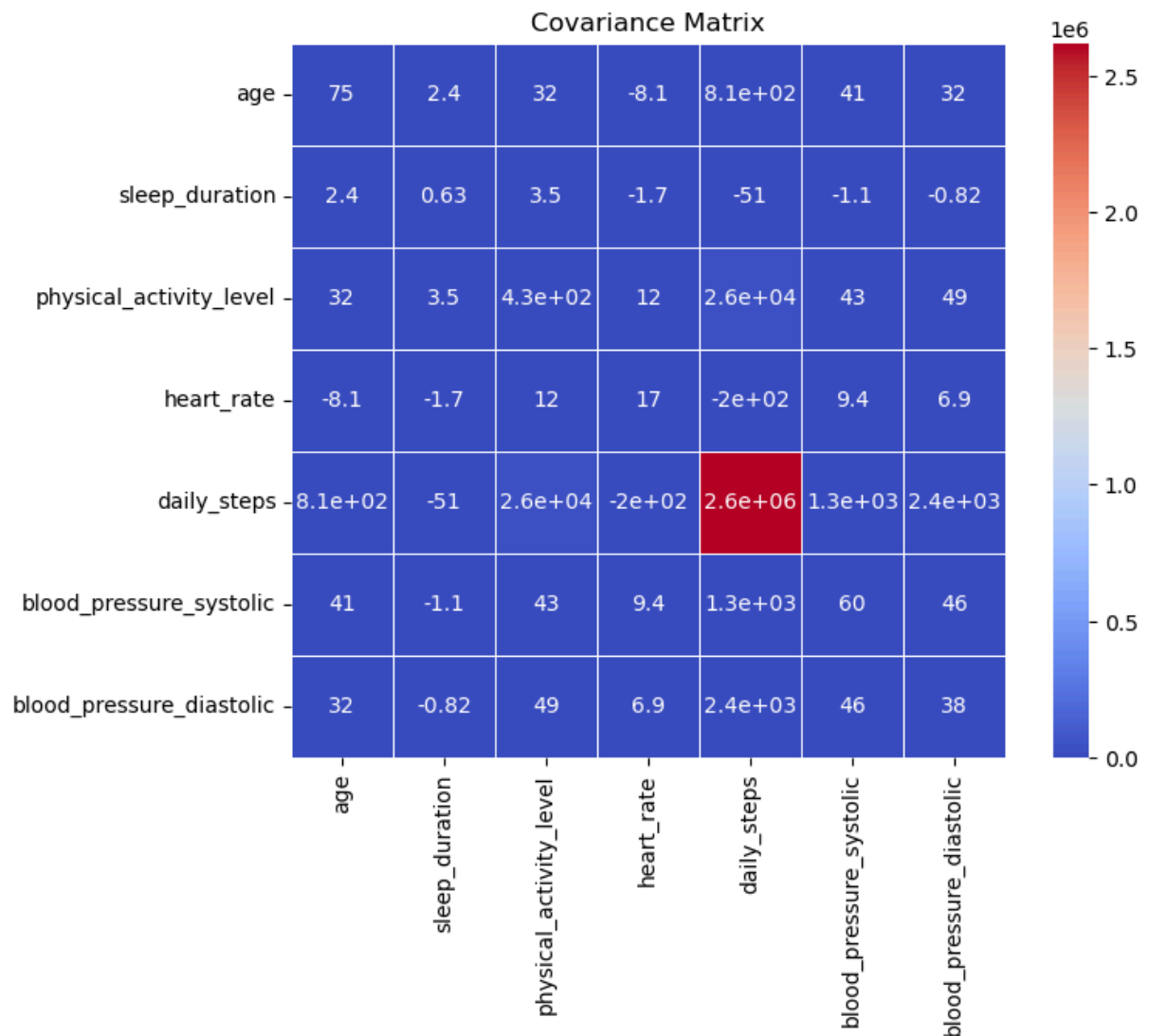
Expanding the coreelation analysis with a paiplot shows a veriaty of potential correlations between variables, the most notable being between the linear correlation between blood pressure systolic and diastolic as well as between daily steps and heart rate/ blood pressure. Further correlation can bee seen inbetween physical activity level and sleep duration while the plot age vs sleep duration appears to show certain clusters that may be further analyzed in the second part of this work.



Plotting the correlation corresponding matrix reflects some of the observations made in the pairplot, with the various near 100% correlation between blood pressure systolic and diastolic very apparent. Two previously less apparent correlations are the ones between age and blood pressure as well as the correlation between physical activity and daily steps.



Plotting the covariance matrix however shows an issue with the current format of the data, where daily steps outweighs all other variances due to its scale (3000 - 10000). This issue will be adressed in the next section of part one of this work.



Having a general overview of structure and correlation in the data, the next step is to scaling and outlier issues in the next subsection.

Preprocessing

The following two issues in the current data set:

- Outliers in variable "heart rate"
- Scaling issue to (among others) variable "daily steps"

This section corrects outliers, validates skewness and standardizes the numeric variables.

Outliers and Skewness

The aim of this part of the preprocessing, is to obtain symmetric variables without outliers in order to apply in a correct form the PCA.

It is observed that only one variable has outliers and positive skewness problems (heart rate). Therefore, the first step is to cut the outliers (4% of the dataframe), and then, check if the skewness problem is also corrected.

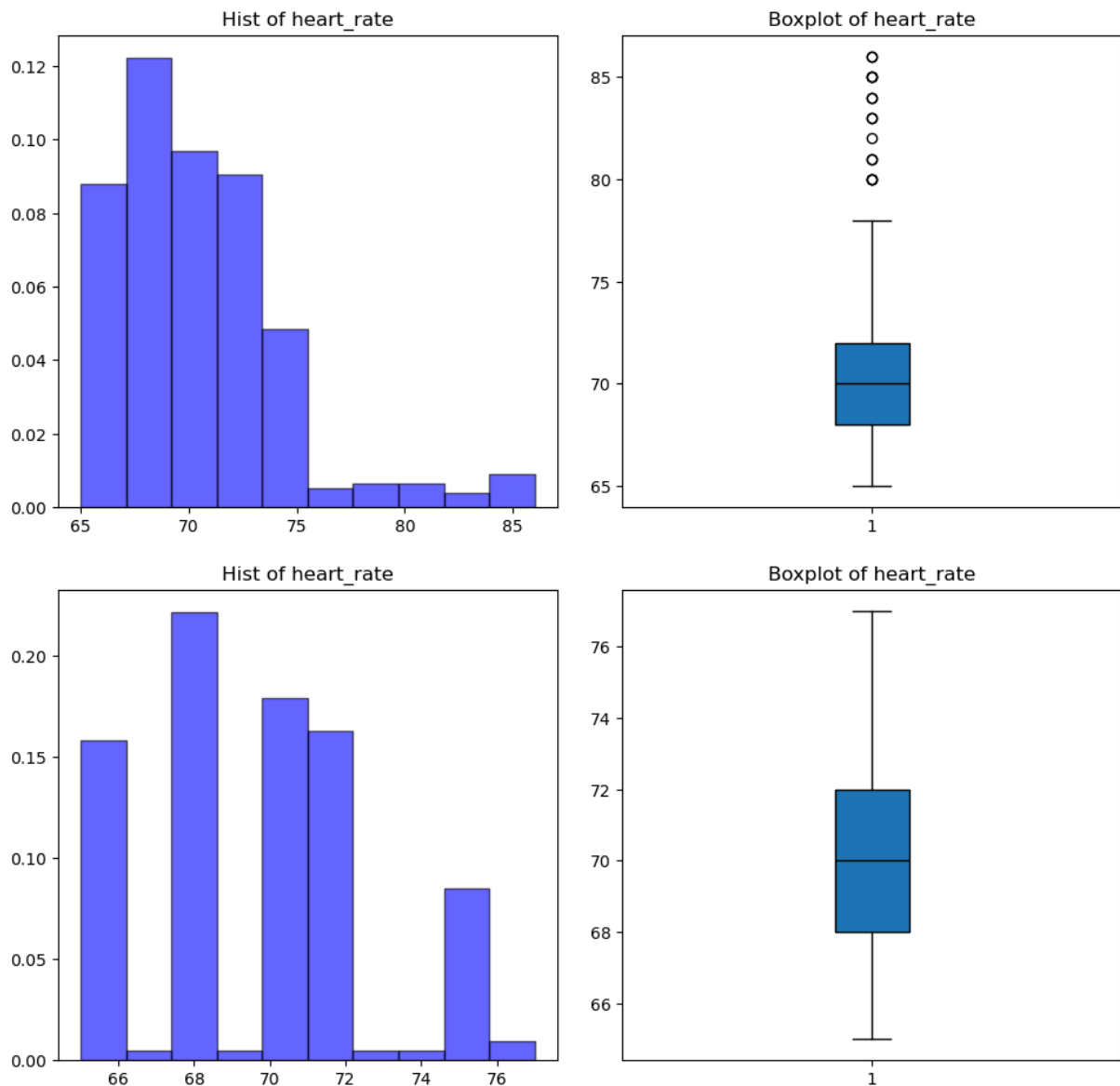
Skewness of age : 0.2561893511793312
 Skewness of sleep_duration : 0.037403602518975176
 Skewness of physical_activity_level : 0.07418782500797434
 Skewness of heart_rate : 1.2199056700731632
 Skewness of daily_steps : 0.17756151681455
 Skewness of blood_pressure_systolic : -0.03552565092220491
 Skewness of blood_pressure_diastolic : 0.37705009626387237

	Variable	Count	Min	Mean	Percentile 90%	\
0	age	374	27.0	42.184492	54.0	
1	sleep_duration	374	5.8	7.132086	8.2	
2	physical_activity_level	374	30.0	59.171123	90.0	
3	heart_rate	374	65.0	70.165775	75.0	
4	daily_steps	374	3000.0	6816.844920	8000.0	
5	blood_pressure_systolic	374	115.0	128.553476	140.0	
6	blood_pressure_diastolic	374	75.0	84.649733	95.0	
	Percentile 95%	Percentile 99%	Max	Variance		
0	58.0	59.0	59.0	7.522324e+01		
1	8.4	8.5	8.5	6.330696e-01		
2	90.0	90.0	90.0	4.339224e+02		
3	78.0	85.0	86.0	1.710381e+01		
4	10000.0	10000.0	10000.0	2.617651e+06		
5	140.0	140.0	142.0	6.003333e+01		
6	95.0	95.0	95.0	3.796546e+01		

The threshold for age upper outliers is 72.125
 then there are 0 outliers in this variable, representing the 0.0 % of the dataset
 The threshold for sleep_duration upper outliers is 9.899999999999999
 then there are 0 outliers in this variable, representing the 0.0 % of the dataset
 The threshold for physical_activity_level upper outliers is 120.0
 then there are 0 outliers in this variable, representing the 0.0 % of the dataset
 The threshold for heart_rate upper outliers is 78.0
 then there are 15 outliers in this variable, representing the 4.01 % of the dataset
 The threshold for daily_steps upper outliers is 11600.0
 then there are 0 outliers in this variable, representing the 0.0 % of the dataset
 The threshold for blood_pressure_systolic upper outliers is 150.0
 then there are 0 outliers in this variable, representing the 0.0 % of the dataset
 The threshold for blood_pressure_diastolic upper outliers is 105.0
 then there are 0 outliers in this variable, representing the 0.0 % of the dataset

We can see that the skewness was also corrected by cutting the outliers observations. For that reason, there is not needed another type of transformation.

Skewness of heart_rate : 0.207482395234077



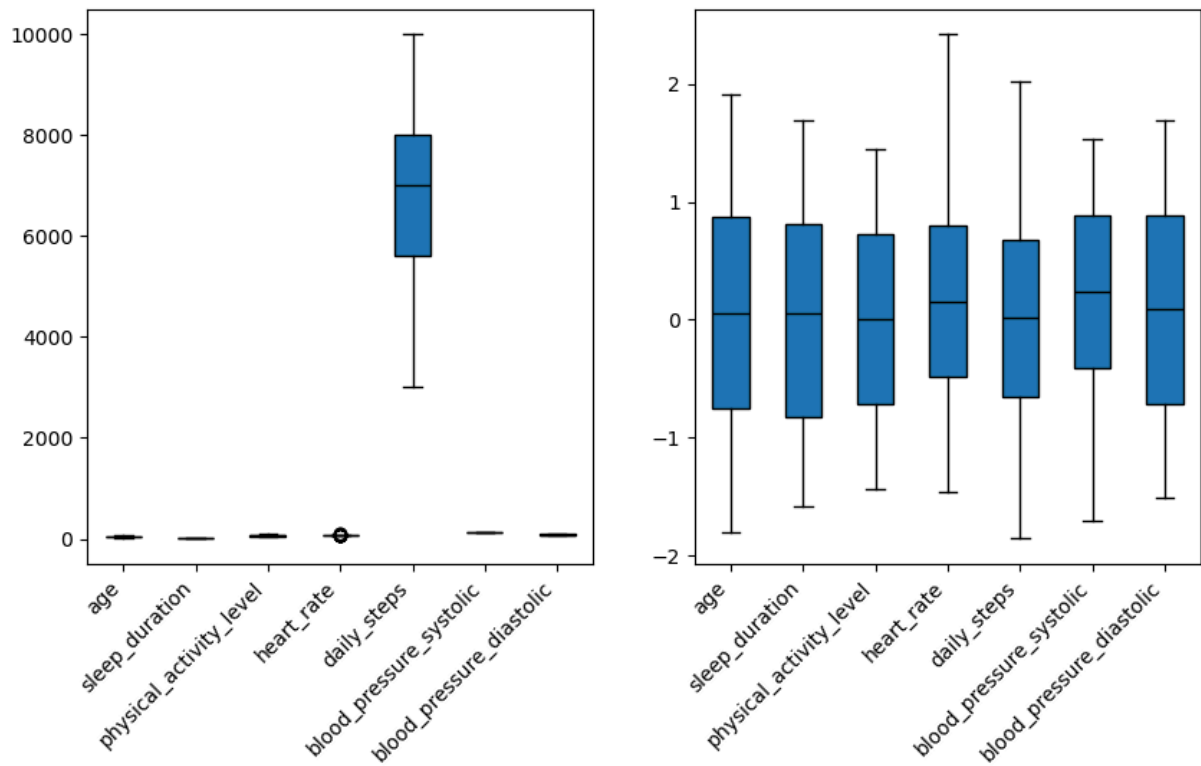
Standardize numeric variables

Having seen in the exploratory data analysis that there exists a strong imbalance in scale between numerical variables in the dataset, the dataset is standardized in this step to mean 0 and scaled on its standarddeviation.

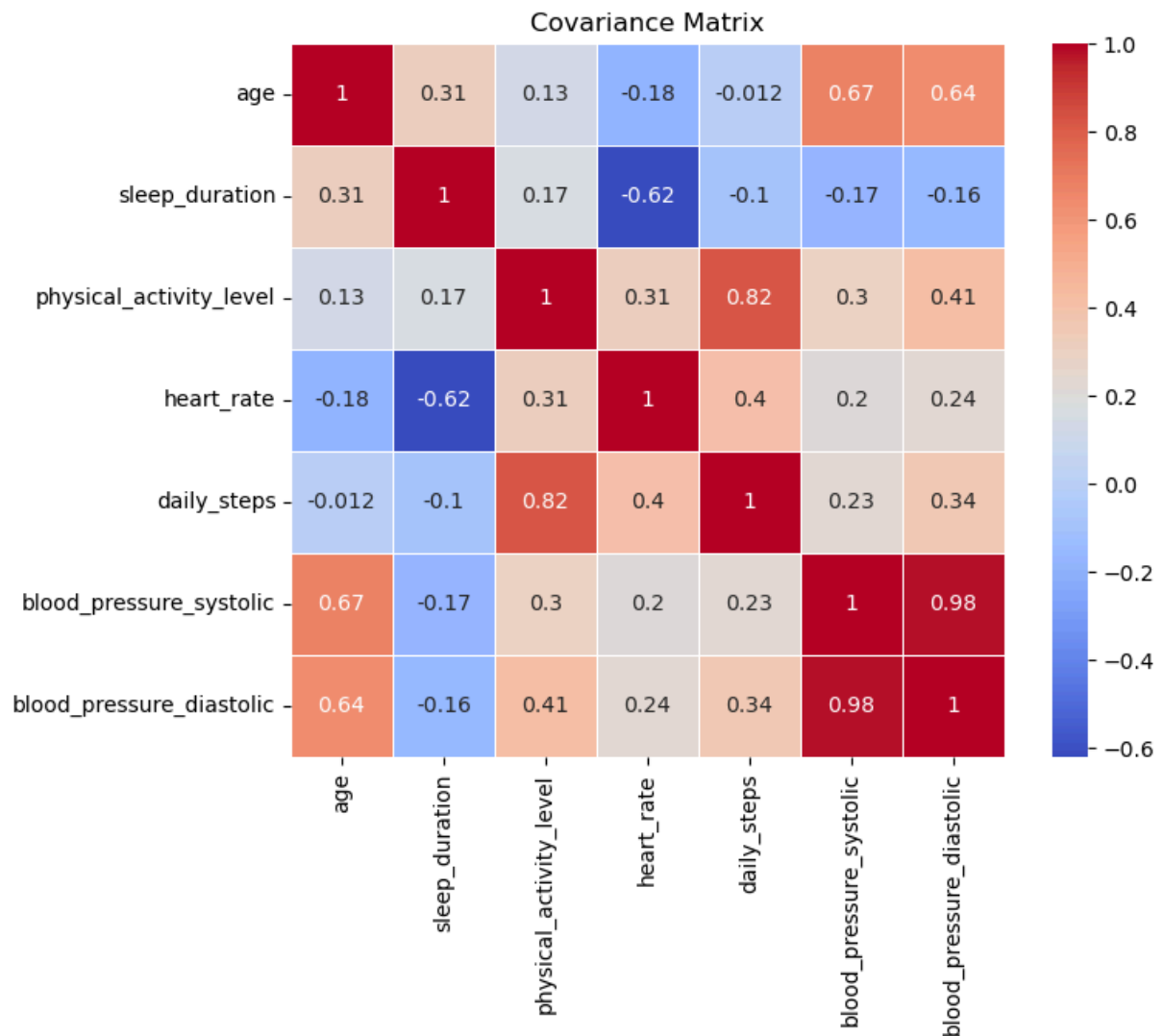
Comparing the boxplots pre-standardized and post-standardized shows the major impact the rescaling has, where daily steps previously dominated and now an even distribution for all numeric variables can be seen.

```
Out[141...] Text(0.5, 0.98, 'Boxplots not-standardized vs standardized')
```

Boxplots not-standardized vs standardized



As a result from scaling, now the covariance matrix can be constructed, showing similar results compared to the previously analyzed correlation matrix.



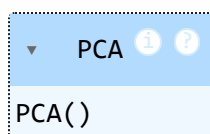
PCA

Having analyzed the data and its characteristic and highly correlated variables identified, as well as having eliminated outliers as well as having standardized the numeric variables, principal component analysis can now be applied in an attempt to reduce dimensionality.

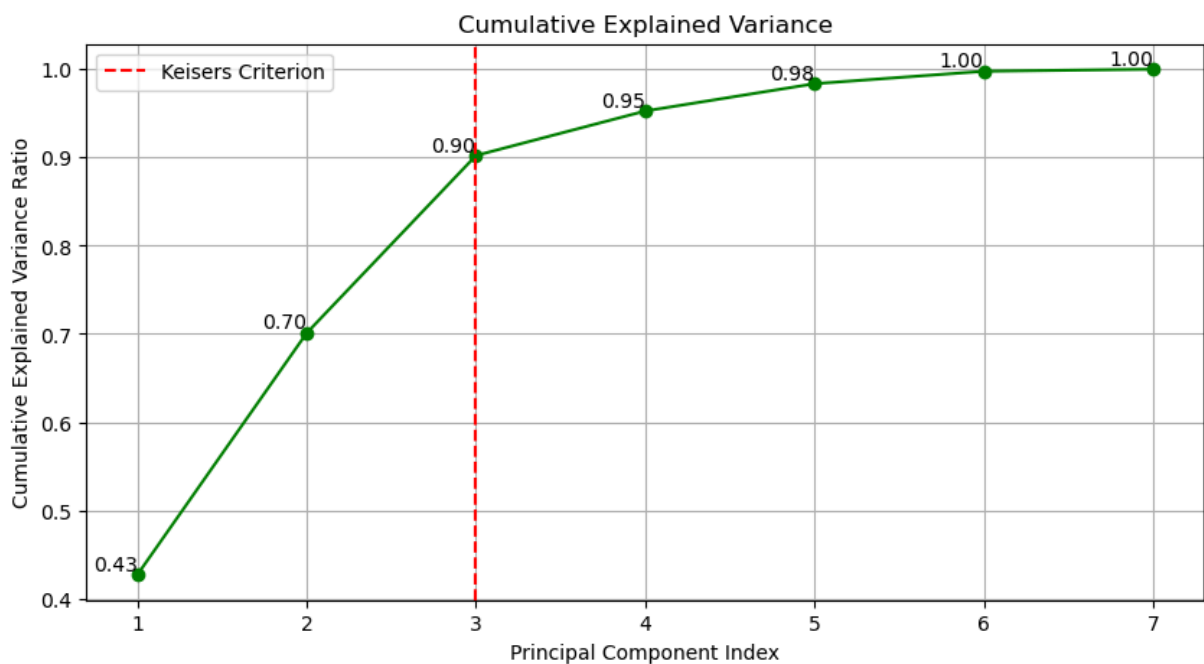
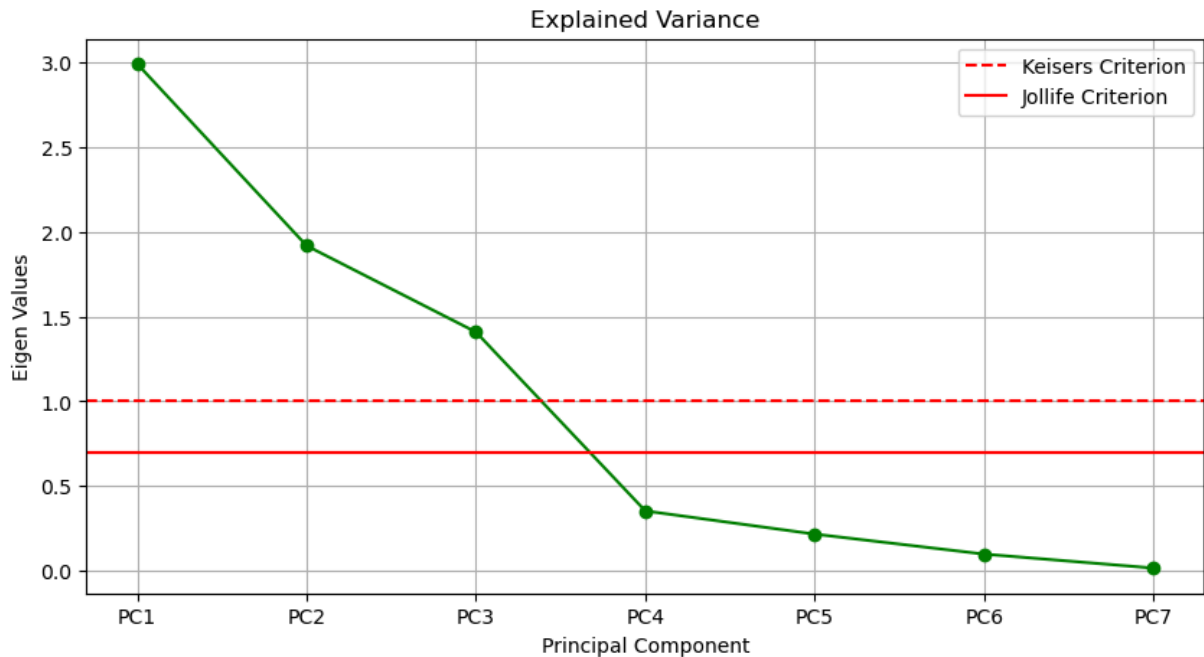
Number of Principal Components

By analysing the Explained Variance (eigenvalues) trend, and the Joliffe's and Kaiser's criterion, for this project there are selected 3 Principal components that explain the 90% of the variability.

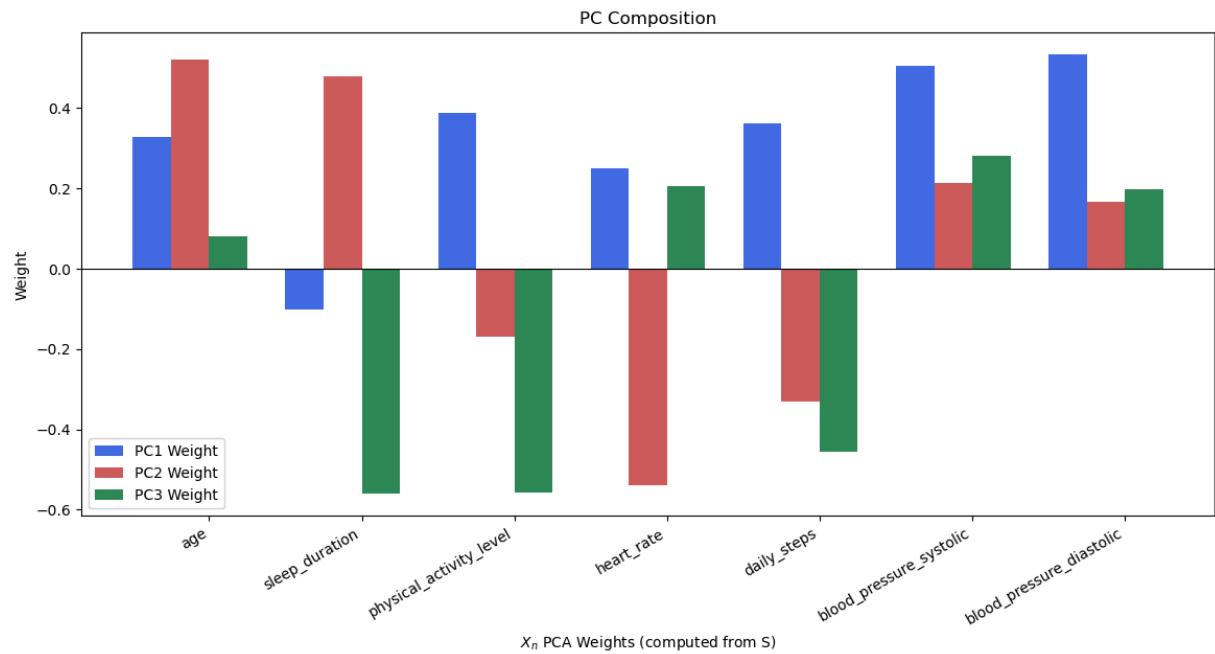
Out[143...



Explained Variance: [2.98968211 1.91745272 1.4083181 0.35319134 0.21650964 0.09865361 0.0161925]
Singular Values: [32.4862707 26.01654875 22.29655327 11.16586514 8.74230527 5.90124756 2.39080551]
trace: 7.000000000000005



Plotting the Composition of the Principal Components as a Barplot shows a general overlap in the contribution of all variables to the TODO: Interpret Barplot



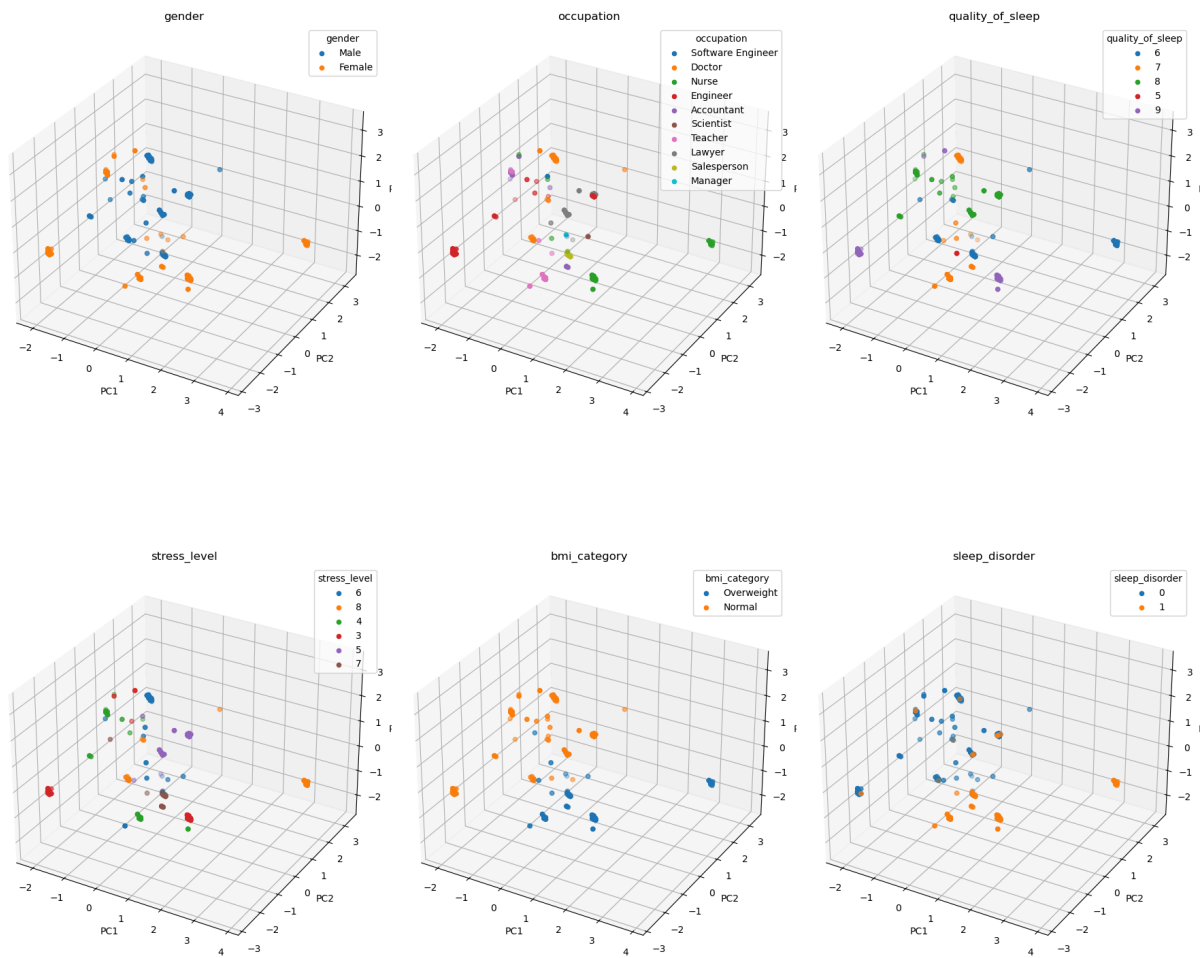
Out[148...

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	index	ger
0	-1.124016	2.215918	-2.298034	1.154400	-1.533155	0.032912	0.166127	0	M
1	0.156448	3.280637	0.273857	-0.164556	0.474852	-0.964458	-0.353457	1	M
2	0.156448	3.280637	0.273857	-0.164556	0.474852	-0.964458	-0.353457	2	M
3	-0.941439	1.204074	1.702637	-0.073055	-0.624588	-0.332685	0.124738	7	M
4	-0.941439	1.204074	1.702637	-0.073055	-0.624588	-0.332685	0.124738	8	M
...
349	2.344172	-2.306957	0.262360	0.259204	-0.226505	-0.012552	0.137651	369	Fer
350	2.356797	-2.246671	0.191758	0.215299	-0.183942	0.045334	0.138202	370	Fer
351	2.344172	-2.306957	0.262360	0.259204	-0.226505	-0.012552	0.137651	371	Fer
352	2.344172	-2.306957	0.262360	0.259204	-0.226505	-0.012552	0.137651	372	Fer
353	2.344172	-2.306957	0.262360	0.259204	-0.226505	-0.012552	0.137651	373	Fer

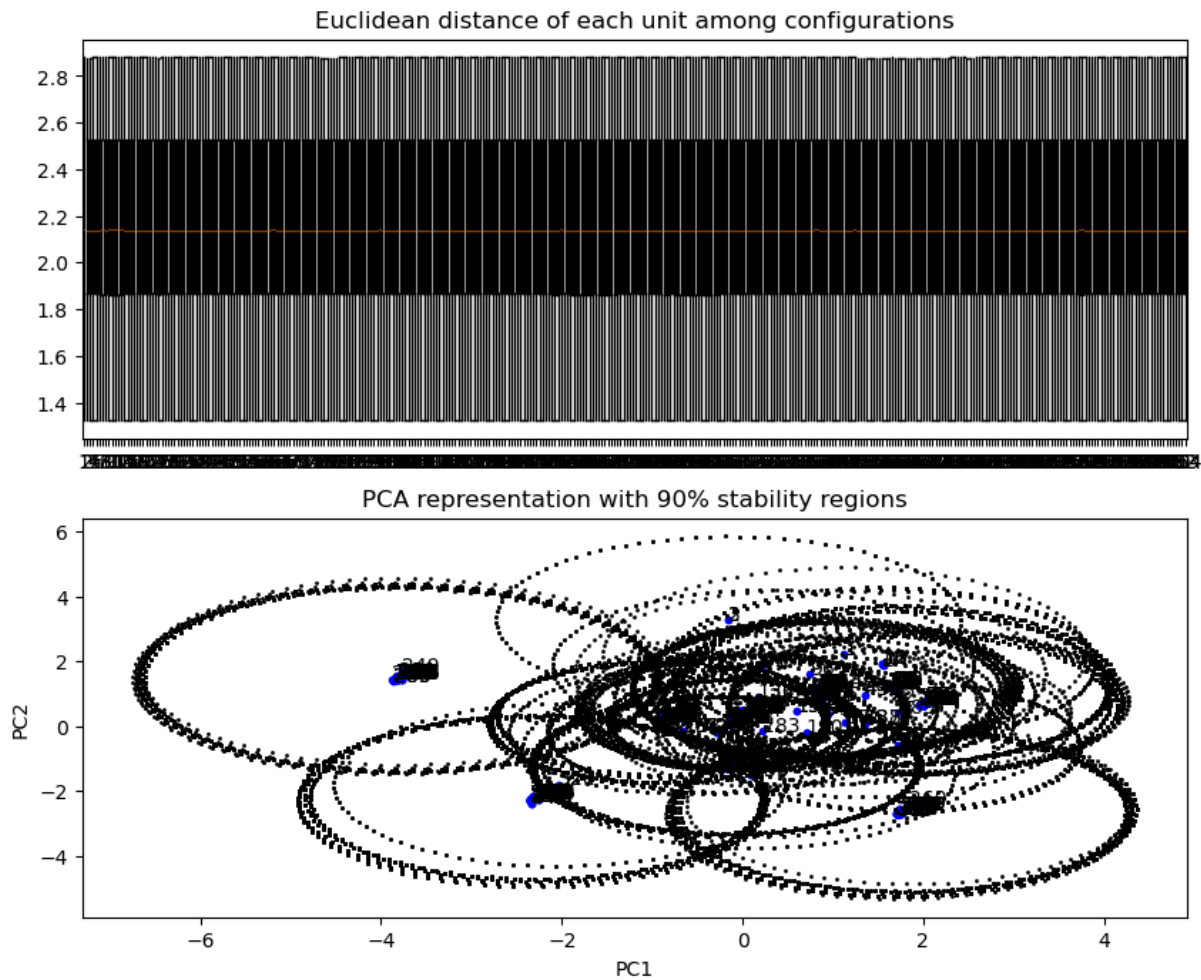
354 rows × 14 columns



TODO: Interpret Scatterplots



Estability of the Principal components



Out[152...

	Mean	Standard Deviation	Median	Mean Absolute Deviation (MAD)
PCA 1	2.169042	0.396796	2.139104	0.306660
PCA 2	2.170711	0.396445	2.140058	0.305647
PCA 3	2.169899	0.396737	2.139598	0.306719
PCA 4	2.169952	0.396728	2.139642	0.306678
PCA 5	2.170067	0.396691	2.139747	0.306680
PCA 6	2.170119	0.396667	2.139757	0.306644

To perform the stability analysis using the leave-one-out method, we obtained 354 different outputs, which makes it challenging to interpret the results visually.

The values in the table are quite consistent across the different PCAs, indicating that they all exhibit similar average Euclidean distances and very low standard deviations. This suggests that the principal components are stable, remaining unaffected by changes in the data. Consequently, this implies that the PCA captures the underlying structure of the data reliably, providing confidence in the robustness of the results. In general, this suggests that the PCA model is likely to generalize well to new data.

use of chatgpt to redact some of the text