Part 2

Laura Silvana Alvarez, Florencia Luque and Simon Schmetz

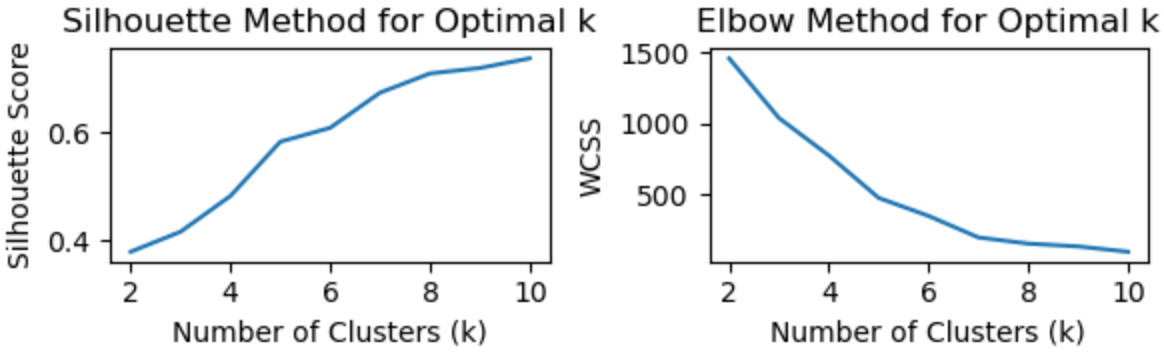
The following project documentation was written as work assignment for the module "Multivariate Analysis" of the Master in Statistics for Data Science at the Universidad Carlos III de Madrid. It contains the Multivariate Analysis of a Kaggel dataset on Sleep Health and Lifestyle (https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset). The work is split into two parts, where in a first part a exploratory data analysis is performed, some data preprocessing steps are taken and a Prinicipal Component Analysis (PCA) is performed. In the second part, distance based metrics in the form of Kmeans will then be applied to identify clusters emerging from the PCA. Additionally, as an alternative to PCA, Multi Dimensional Scaling is applied as second dimension reduction technique and compared to PCA in its effectiveness to reduce dimensions.

Kmeans

In the first step, we will compare different types of clustering using the K-Means algorithm. The first approach focuses solely on numeric variables, comparing two scenarios: one using the raw data (excluding variables with high correlations) and another using the PCA components derived in the first part of the analysis.

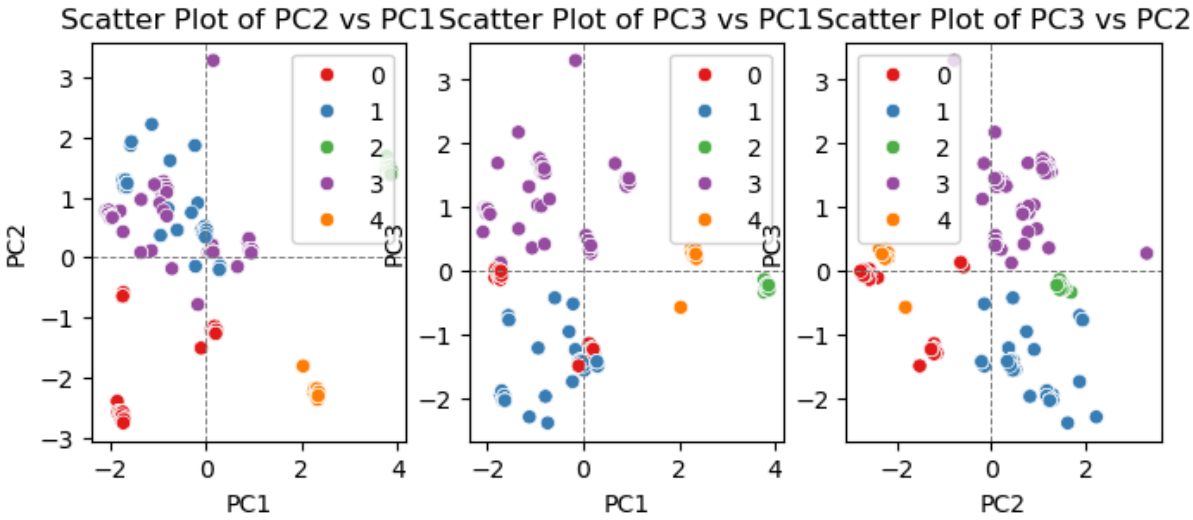
For the second approach, K-Means will be applied to mixed data, combining both numeric and categorical variables.

Kmeans with PCA data



If we compare the graphs, there is stabilization between 5 and 7 clusters in the Silhouette method. Similarly, in the Elbow method, the slope becomes less steep around 5 and 7 clusters.

Between the cluster 3 and 1 is some elapsing data within the PC1 vs PC2. This also ocurre within the 1 and the 0 in the PC1 vs PC3. The other clusters all well separarted in the first factorial plane.




```
Inertia (WCSS): 469.44182341248046
Silhouette Score: 0.5704171764668877
```

If we use 5 cluster there's seem to be more clarity between the clusters. The silhouette is also almost over 0.5 and it visible in the graphs.

If we compare the information we can said that the best number of cluster for the PCA data is 7, but we decide to keep 5, because it does not make sense to have that amount of clusters when the sample is not that large.

The principal characteristics of the clusters are given in the next table. Where in the categorical variables we have the mode of the cluster, and in the end we have the proportion of the cluster that has sleep disorders.

- In cluster 2 and 4 it is shown that there are more nurse females with overweight. Despite the fact that the cluster 0 has a good perception of their quality of sleep, it is shown that both clusters have the biggest % sleep disorder (over 90%)
- Cluster 3 have the lowest proportions of sleep disorder, most of the members of this cluster are men, lawyers with a normal bmi.

In general it can be concluded that the perception of the sleep quality is not directly related with having or not a sleep disorder.

Out[49]:

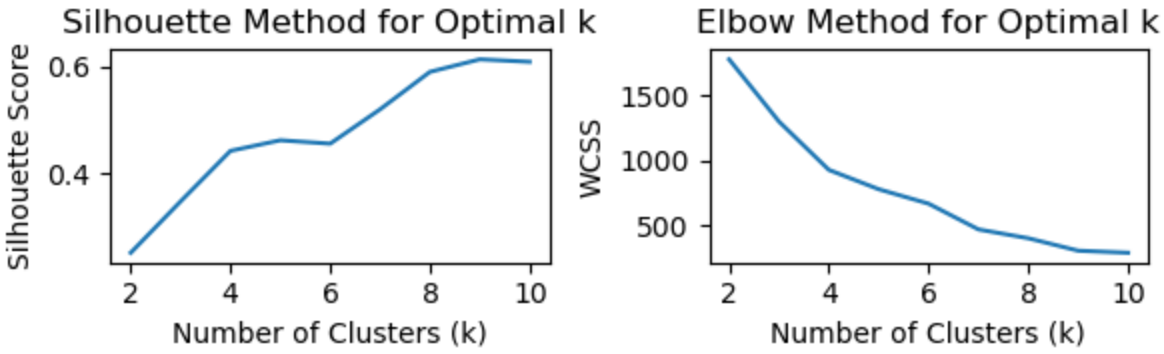
	gender	occupation	quality_of_sleep	stress_level	bmi_category	sleep_disorder	sleep_d
0	Female	Engineer	9	3	Normal	0	
1	Male	Salesperson	6	7	Overweight	0	
2	Female	Nurse	6	8	Overweight	1	
3	Male	Lawyer	8	5	Normal	0	
4	Female	Nurse	9	3	Overweight	1	

Out[50]:

	age	sleep_duration	physical_activity_level	heart_rate	daily_steps	blood_pressu
0	47.646154	7.558462	38.538462	65.000000	5415.384615	1
1	38.731707	6.320732	39.560976	72.146341	5734.146341	1
2	49.750000	6.065625	90.000000	75.000000	10000.000000	1
3	37.070423	7.483803	71.035211	69.112676	7754.929577	1
4	58.030303	8.093939	75.000000	68.242424	6878.787879	1

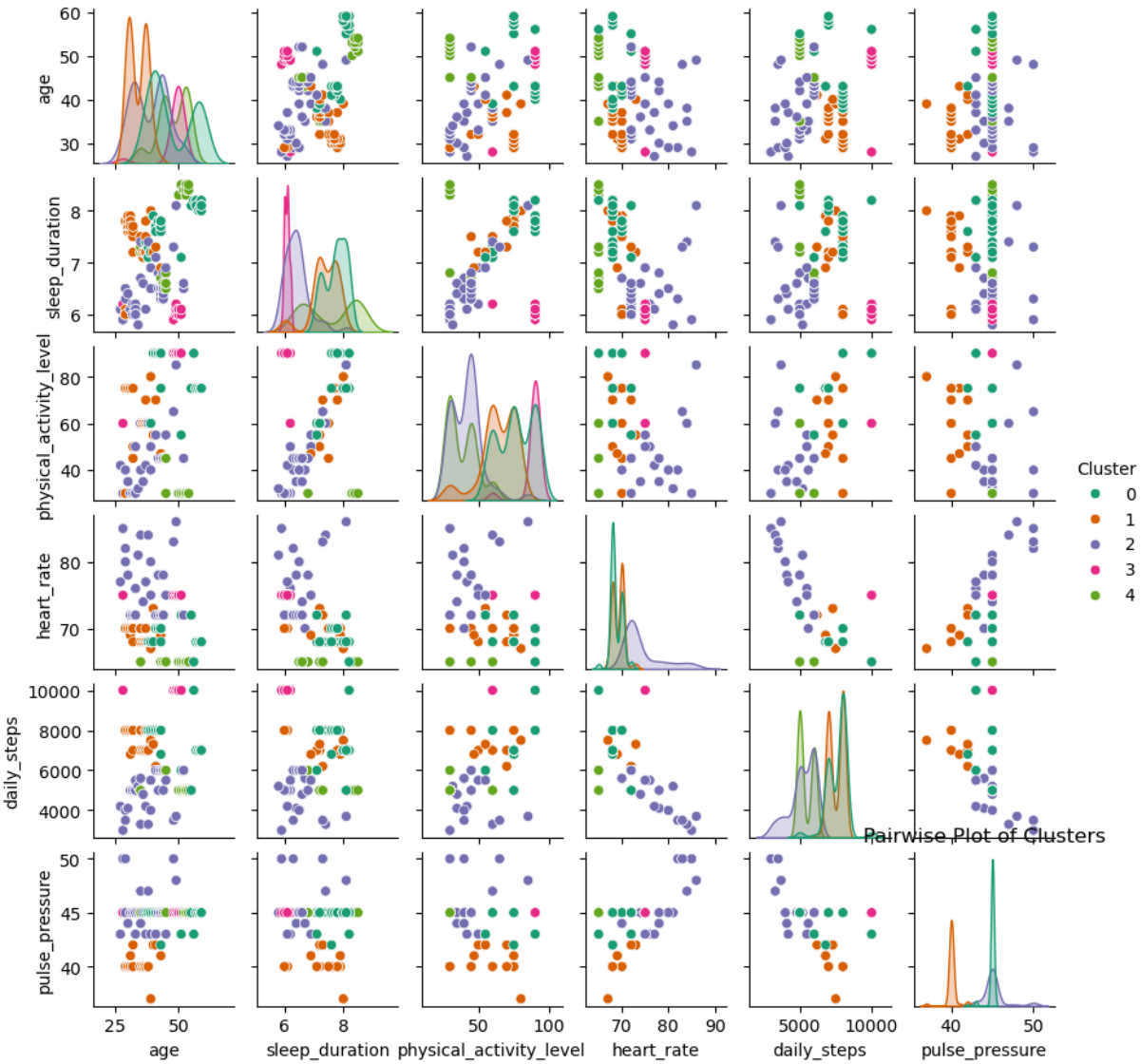
Kmeans with raw data

For this, we need data without strong correlations. The variables related to blood pressure have a high correlation of almost 0.97. To avoid losing any information, we decided to create a new variable called pulse pressure. Pulse pressure is calculated as the difference between systolic and diastolic blood pressure.



If you compare the two graphs, there is stabilization around 4 clusters in the Silhouette Score graph and a noticeable decrease in the slope (tangent) in the Elbow Method graph. However, using 5 clusters appears slightly better, as it balances both metrics.

```
Inertia (WCSS): 694.6414160666794
Silhouette Score: 0.57221538910709
```



Inertia (WCSS): 694.6414160666794
Silhouette Score: 0.57221538910709

There is a larger decrease in WCSS and a slight improvement in the Silhouette Score when choosing 5 clusters. Additionally, using PCA to visualize the data reveals better separation of the clusters compared to 4 clusters. Choosing 5 clusters reduces overlaps between the groups, providing a more distinct and interpretable clustering structure.

The data description for the cluster is as follows.

Out[56]:

	gender	occupation	quality_of_sleep	stress_level	bmi_category	sleep_disorder	sleep_d
0	Male	Lawyer	8	5	Normal	0	
1	Male	Doctor	8	4	Normal	0	
2	Male	Salesperson	6	7	Overweight	1	
3	Female	Nurse	6	8	Overweight	1	
4	Female	Engineer	9	3	Normal	0	

Out[57]:

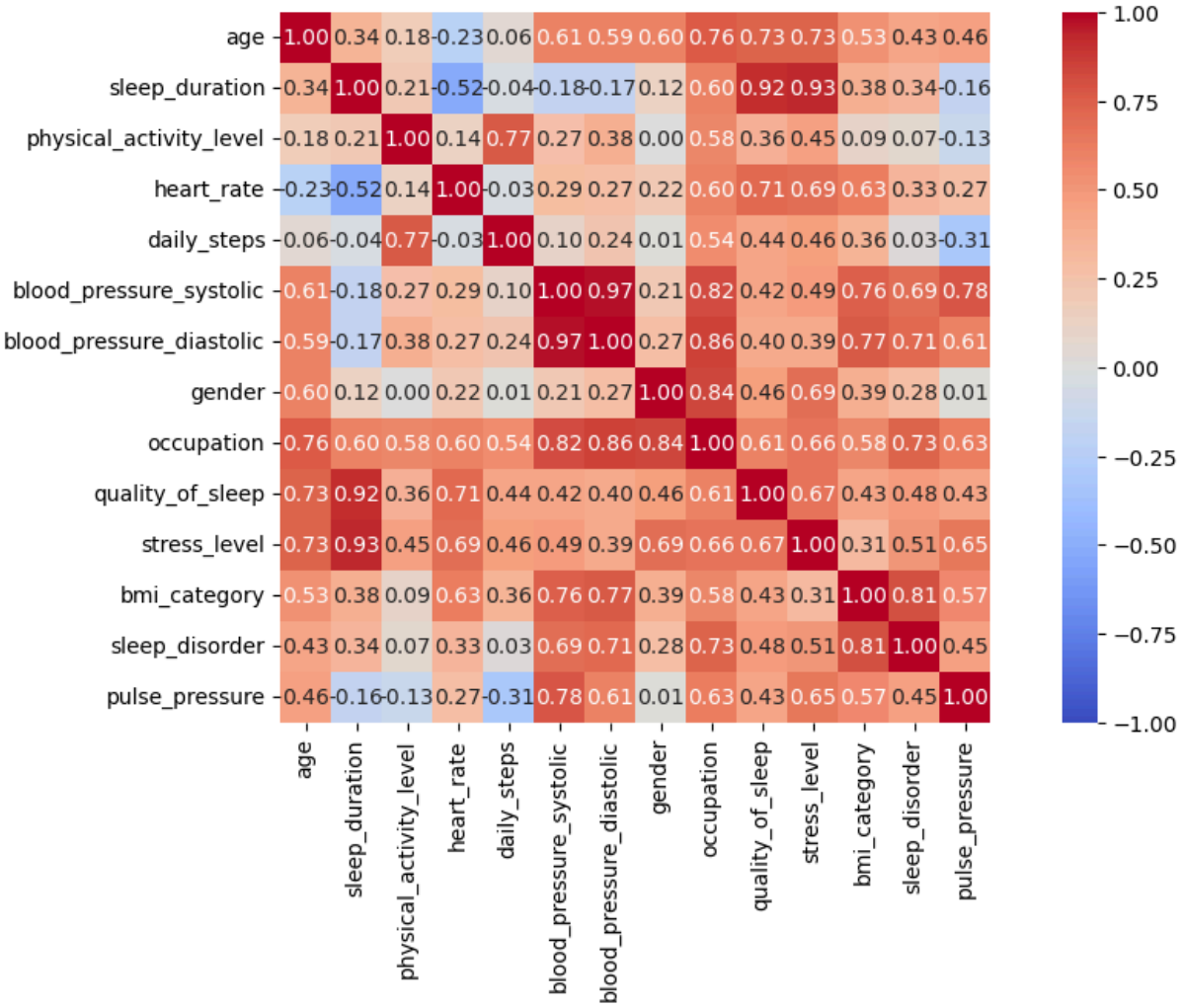
	age	sleep_duration	physical_activity_level	heart_rate	daily_steps	blood_pressu
0	47.031579	7.735789	76.368421	68.715789	7618.947368	1
1	33.882353	7.362353	63.788235	69.176471	7510.588235	1
2	38.778947	6.409474	41.557895	74.305263	5213.684211	1
3	48.470588	6.073529	88.235294	75.000000	10000.000000	1
4	47.646154	7.558462	38.538462	65.000000	5415.384615	1

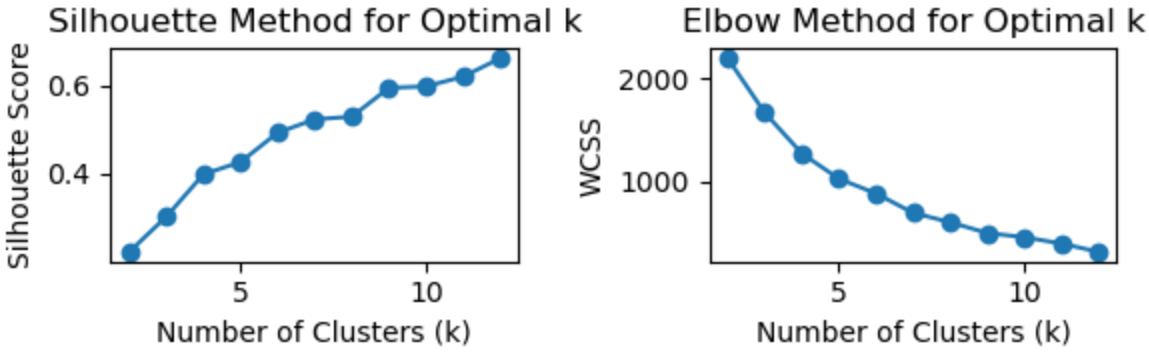
Kmeans with mixed data

For this, we will also use the pulse pressure variable and exclude the sleep disorder variable. This is because the sleep disorder variable is the target variable that this dataset aims to explain.

Now we have the assosiations between numerical and categorical variables finding some interesting relationships:

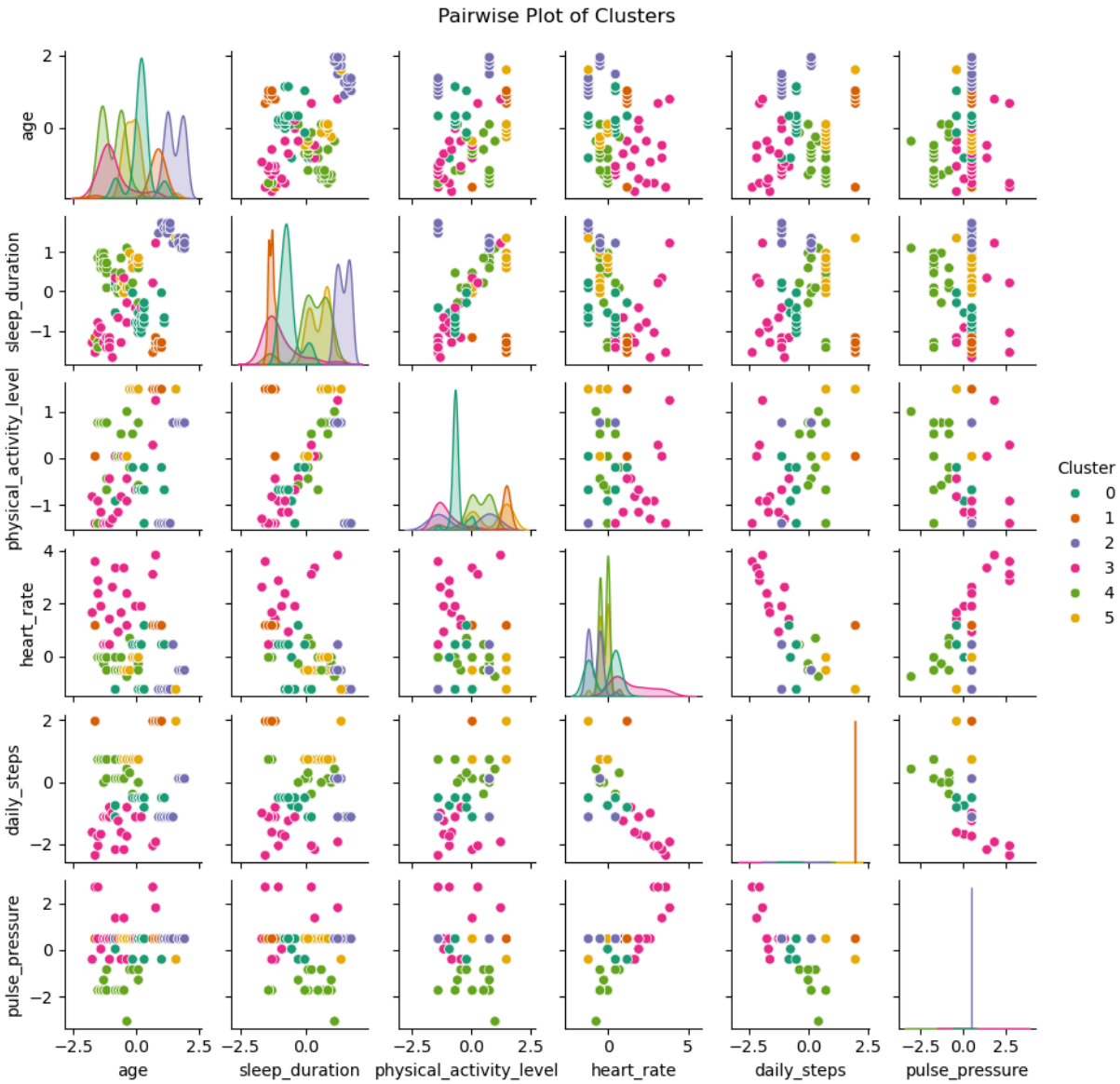
- Age is highly associated with occupation, quality of sleep and stress level.
- The hear rate is associated with quality of sleep, stress level and BMI category.
- Other important relations are gender with occupation and sleep disorder with bmi category.





In this case the elbow method does not show a clear point where the trend of the line change, but the silhouette method suggest to take 6 or 8 clusters.

```
Inertia (WCSS): 654.0051973653807
Silhouette Score: 0.5304769553918965
```




```
Inertia (WCSS): 848.3962463374016
Silhouette Score: 0.45668915344071426
```

If we compared the silhouette score and the cost. The best quantity of cluster for the mixed kmeans is 8 clusters. But taking into account the size of the data base, we consider that 8 clusters are too much, so we decide to keep 6.

What we see before about the positive association of the bmi category and occupation with having or not sleep disorder is reflected in the clusters 0 and 1 where orverweight and being a Nurse/Salesperson increase the ratio of sleep disorder.

From cluster number 4 and the association matrix we can infere that the bmi category is the principal driver for having or not a sleep disorder.

Out[68]:

	gender	occupation	quality_of_sleep	stress_level	bmi_category	sleep_disorder	sleep_d
0	Male	Salesperson	7	7	Overweight	1	
1	Female	Nurse	6	8	Overweight	1	
2	Female	Nurse	9	3	Overweight	0	
3	Male	Doctor	6	8	Normal	0	
4	Male	Doctor	8	4	Normal	0	
5	Male	Lawyer	8	5	Normal	0	

Out[69]:

	age	sleep_duration	physical_activity_level	heart_rate	daily_steps	blood_pressu
0	43.649351	6.592208	46.168831	68.987013	5905.194805	1
1	48.470588	6.073529	88.235294	75.000000	10000.000000	1
2	55.446154	8.256923	52.846154	66.646154	5953.846154	1
3	34.269231	6.348077	38.326923	76.230769	4588.461538	1
4	33.988372	7.365116	63.918605	69.162791	7502.325581	1
5	40.983333	7.551667	77.500000	68.933333	8066.666667	1

Multi Dimensional Scaling

As alternative to Principal Component Analysis, Multi Dimensional Scaling (MDS) offers a distance based dimension reduction method. Input to MDS are a distance (or dissimilarity) Matrix generated by a chosen distance. This distance has to fulfill the following properties:

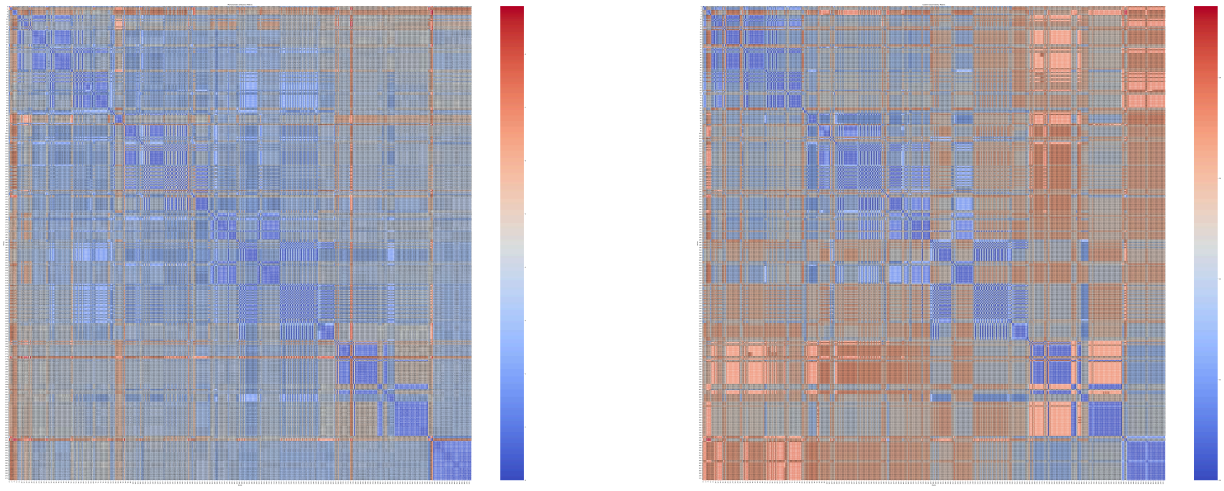
- Symmetry (of distance between two points, A->B = B->A)
- Non-Negativity (distance between two points must be non-negative)
- Identity of Indiscernibles (distance A->B is only zero if A=B)
- Triangular Inequality (A->C <= A->B->C)

Distance Matrices

in the following, we utilize the Mahalanobis distance and the Gower dissimilarity to set up two distance matrices, one of which will then used to apply MDS.

Plotting the corresponding distance matrix gives a subjective idea in the differences in the distance matrices, while it does not deliver any profunder conclusions on a data point level.

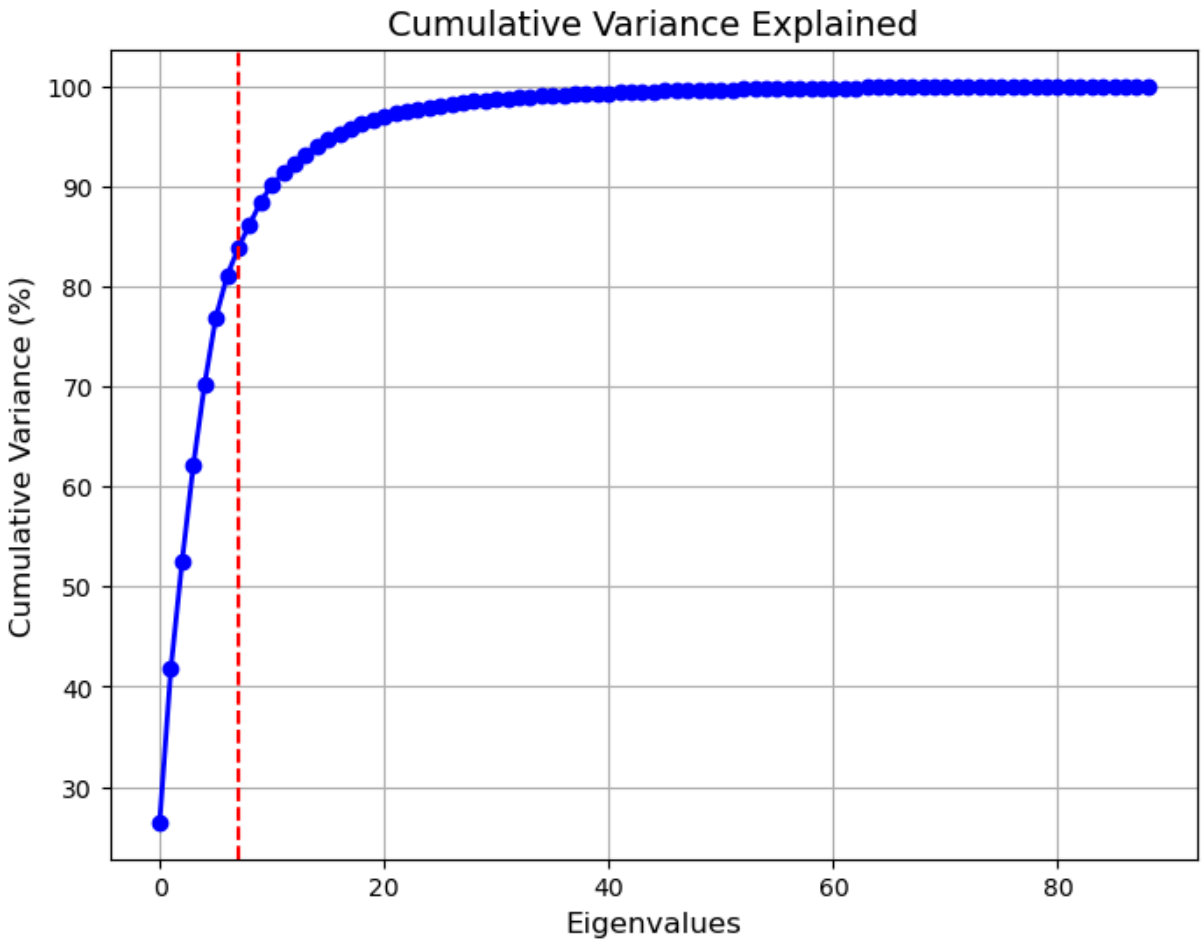
Out[73]: Text(8734.358585858585, 0.5, 'Points')



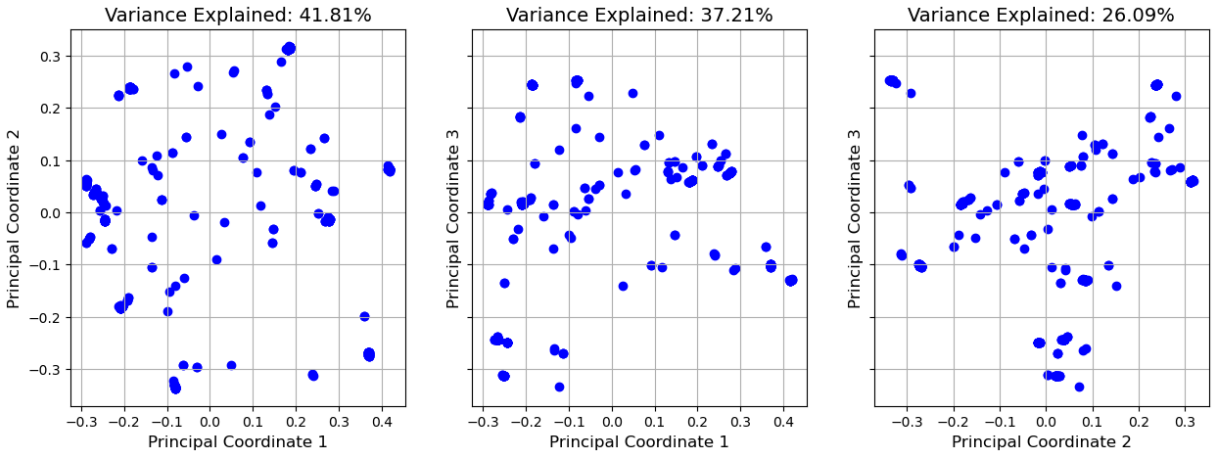
Applying Multi Dimension Scaling

With the distance/dissimilarity Matrices set up, the MDS is performed using Gowers dissimilarity matrix as it also contains the categorical/binary that can be found in the dataset data, whereas Mahalanobis distance was only created with the numerical variables.

Having performed the MDS, as a first step the cumulative explained variance is plotted to show how many Principal Coordinates are required to represent the full Variance of the Data Set. The plot below shows, how 80 % of the variance can be covered with 7 Principal Components. Notable is how from the 7th principal coordinate onwards, the added variance decreases significantly as the cumulative variance goes towards 100 %



Looking at the first three principal coordinates, does not give significant insight into how these principal coordinates can be interpreted.

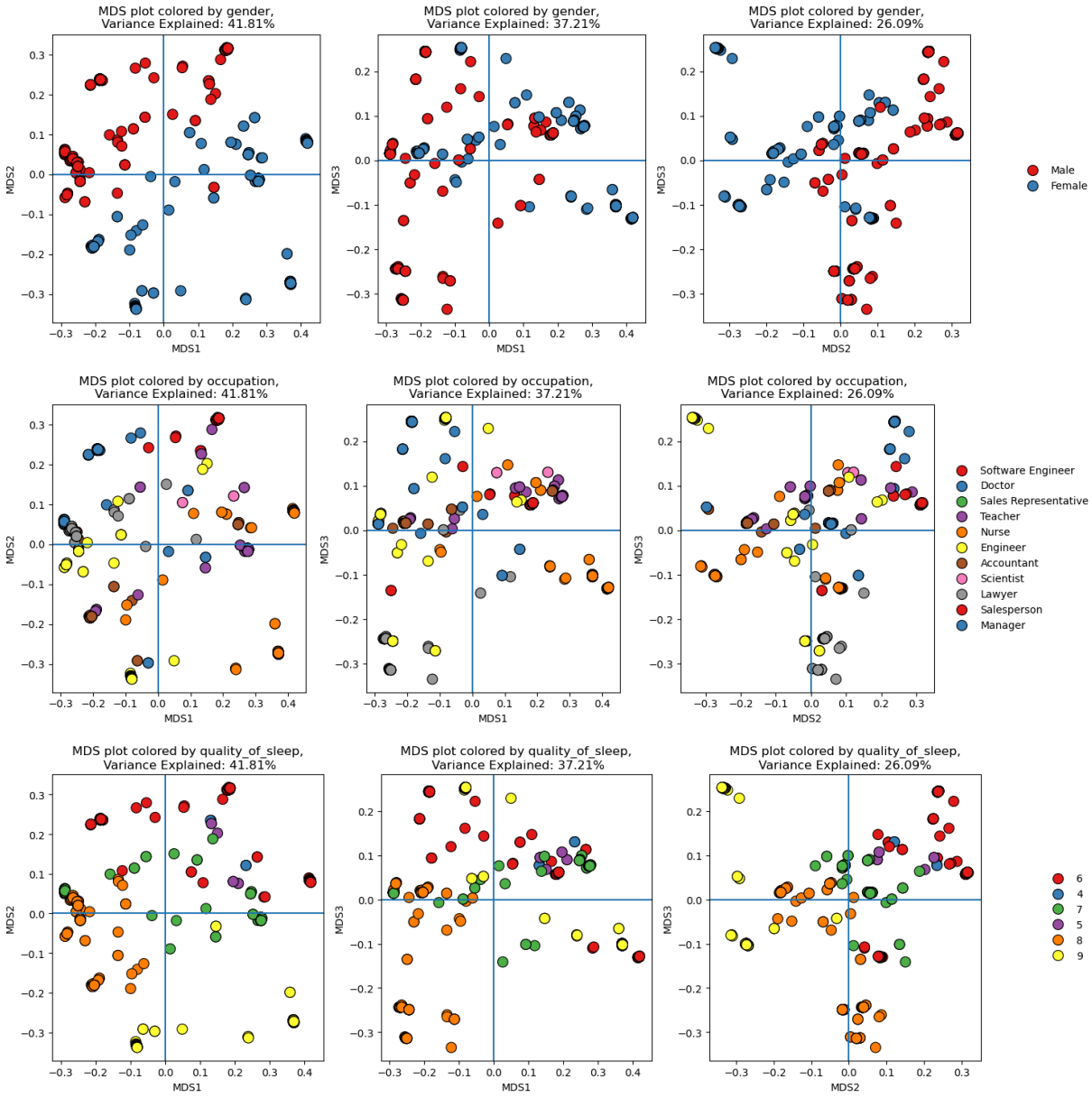


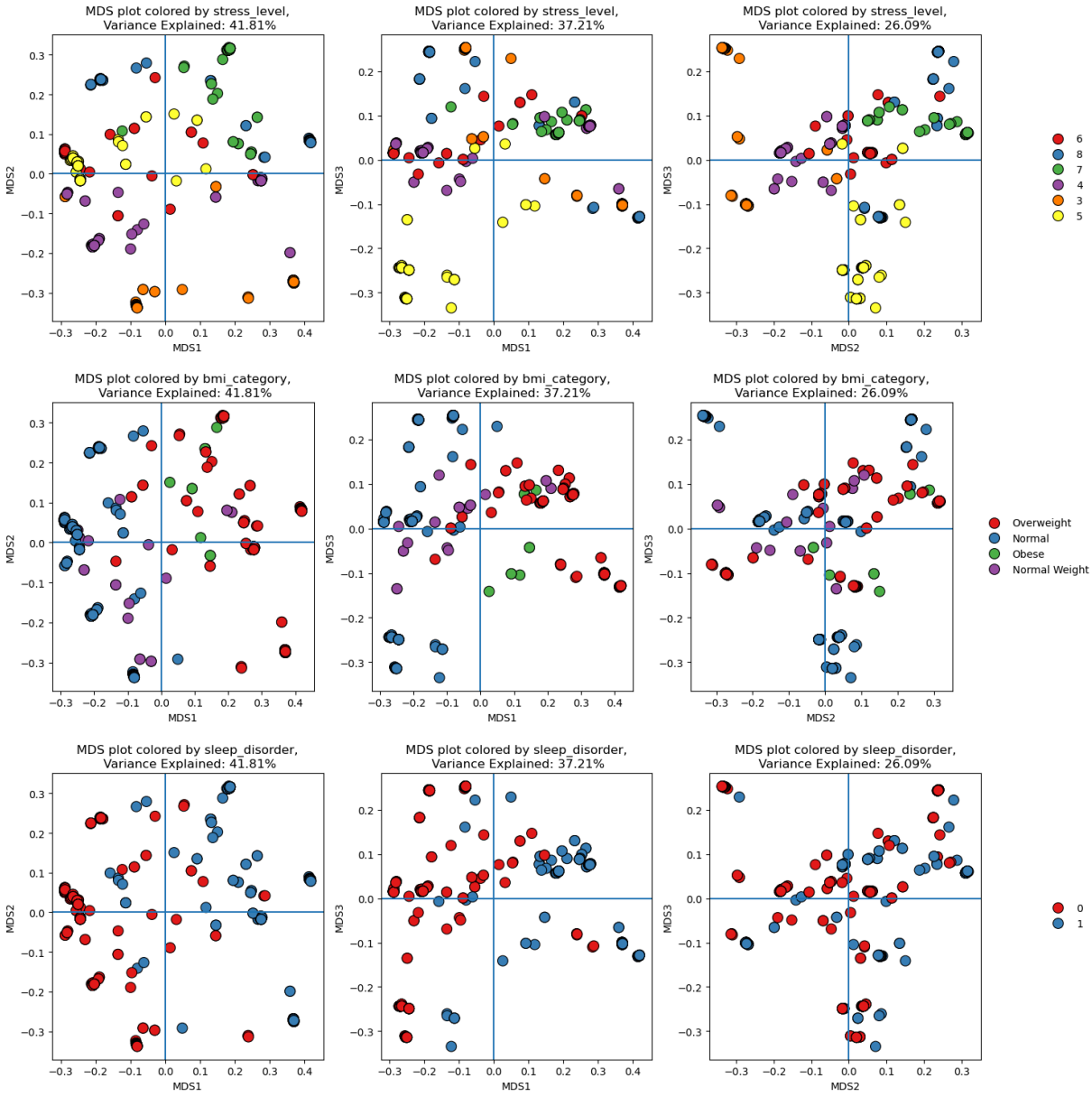
To gain a deeper understanding on how to interprete the principal coordinates, the same plot is done with a color coding of the categorical and the binary Variables.

- Gender is divided in the third quadrant with the males and higher values in the first and second coordinate with females.
- Occupation does not seems to have an specific order in any of the planes.

- Quality of sleep is slightly organize over the second coordinate, greater values of that axis represent a better quality of sleep.
- Strees level seems to have the opposite organization as quality of sleep (Higher values of the second axis represents lower values of stress level).
- BMI and Sleep Disorder are organized over the first axis.

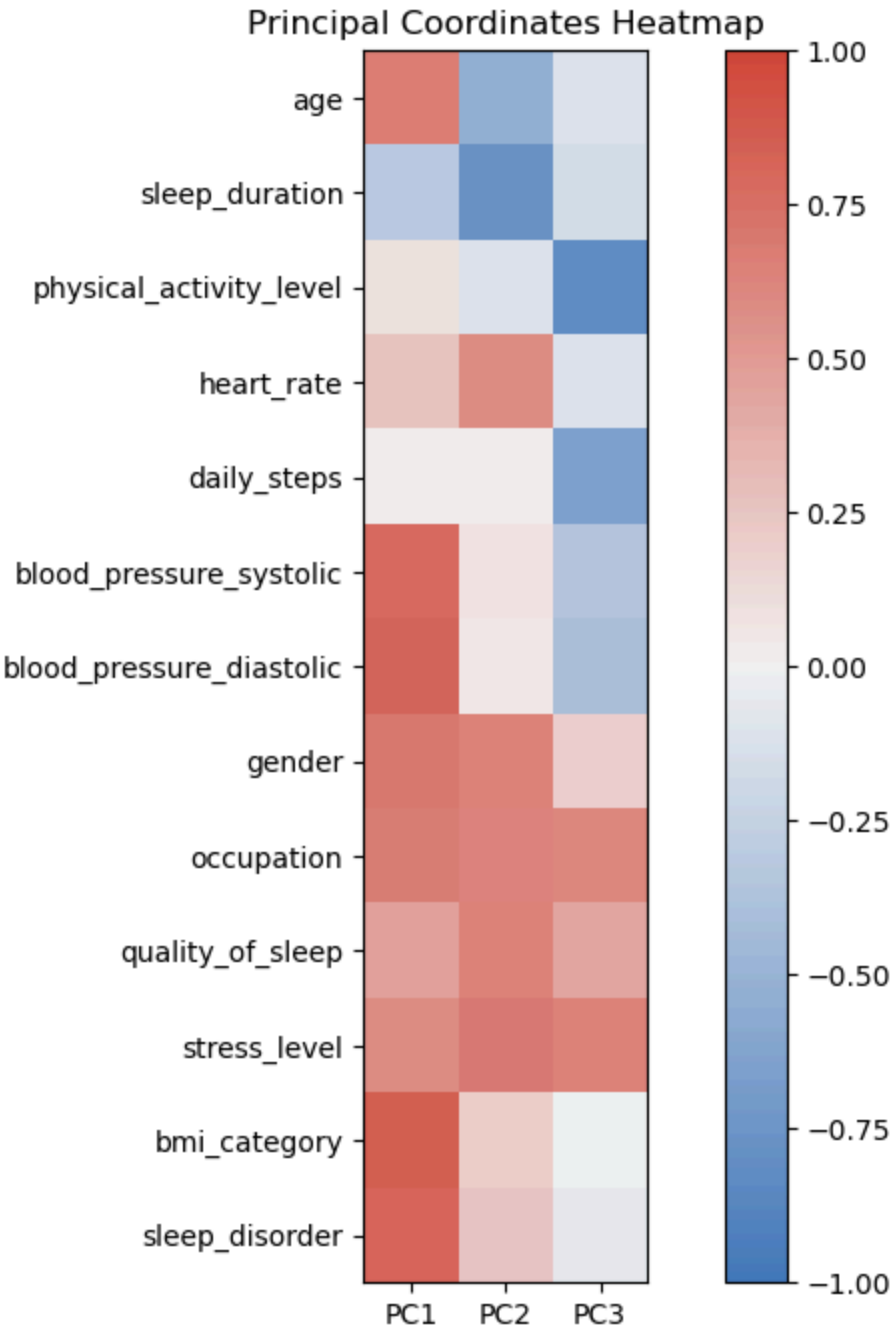
This interpretations are an initial approach of how we can interpret the MDS, however we are going to check the influence of the Categorical variables.





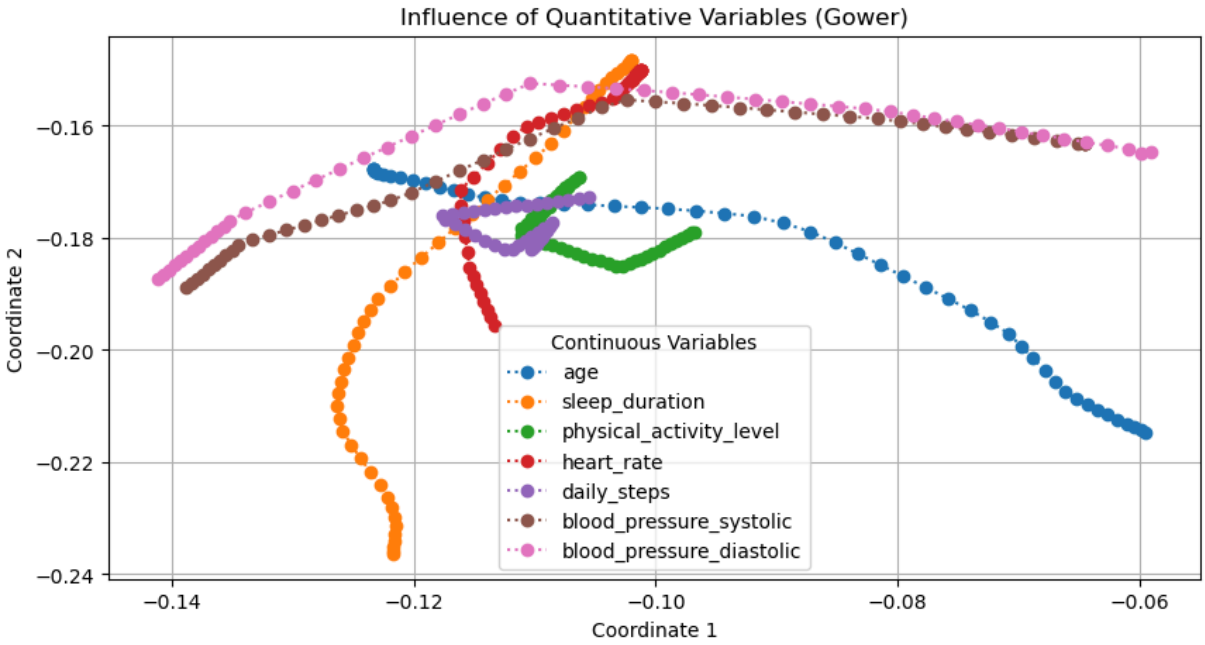
Correlations between original Variables and first three Principal Coordinates

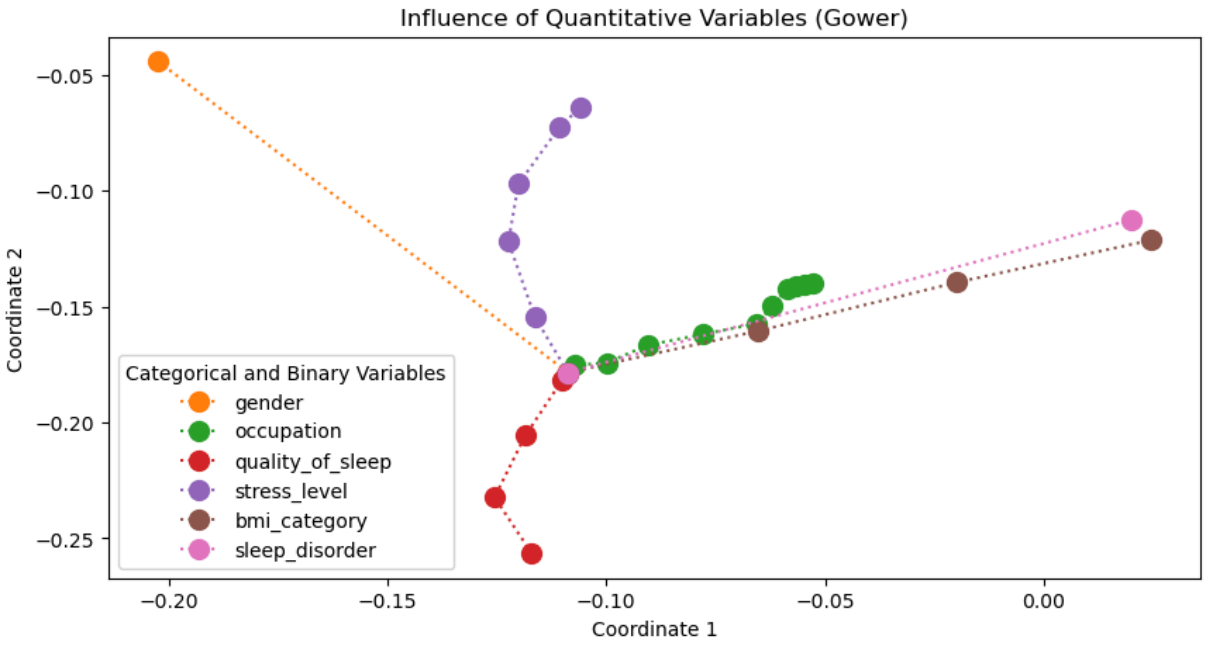
- **First axis:** Blood pressure, BMI category and sleep disporder have a strong positive correlation.
- **Second axis:** Sleep duration has the largest positive correlation and hear reat the largest negative one.
- **Third axis:** Physical activity is the most important variable.



From the Influence of Quantitative and Categorical Variables in the MDS configuration computed from Gower's distance we confirm the results obtained with the correlation matrix for the first two axes, and obtain more information:

- **Quantitative:** On the first axis the most influential variables are the blood pressure (distolic and systolic) and the age. While in the second axis sleep duration and heart rate are the most important ones.
- **Qualitative:** The second axis is clearly separated by quality of sleep and stress level. Lowe values of the first axis are influentiated by the gender and higher by the BMI category and sleep disorder.





Conclusion

Looking at the findings of the Multi Dimensiona Scaling with Gowers distance, it is apparent that the MDS compared to the PCA, is not able to contain as much of the information (variance) in a small number of variables. Where the PCA was able to represent 90 % of the variability with 3 principal components, the MDS requires 6 principal coordinates to represent 80% of the variability with significant more to reach the respective 90%. With a total of 13 Variables, when the objective is the maximum possible dimension reduction, MDS thus appear to be the significantly better approach. None the less, the interpretability of the resulting new dimensions might however favour MDS, where the Influence of the categorical variables shows to be nearly orthogonal for the PC1 vs PC2 dimensions.

The MDS could be further improved by modifying its configuration with new distance types as done in ReIMS, using different distances from the traditional Gower's distance to archieve greater robustness & stability. This however would not change the fact, that PCA appears to be better fit for dimension reduction in this particular data set.