



**Università
di Catania**

Data Analysis Report - IoT Telemetry Data
Academic year 2022-2023

Giovanni Spadaro

Contents

1 Data description	5
1.1 Overall dataset exploration	6
2 Univariate analysis	8
2.1 Time stamp	8
2.2 Device	8
2.3 Carbon monoxide	8
2.3.1 Carbon monoxide fitting	8
2.4 Humidity	8
2.4.1 Humidity fitting	9
2.5 LPG	9
2.5.1 LPG fitting	10
2.6 Light	10
2.7 Motion	11
2.8 Smoke	12
2.8.1 Smoke fitting	13
2.9 Temperature	13
2.9.1 Temperature fitting	14
3 Principal component analysis	17
3.1 Multivariate exploratory data analysis	17
3.2 Applying PCA	17
3.3 Choosing the number of principal components	18
3.4 Data visualization in the principal components space	19
4 Cluster analysis	21
4.1 Cluster validation	21
4.1.1 Assessing clustering tendency	21
4.1.2 Determining the optimal number of clusters	22
4.1.3 Cluster validation statistics	22
4.2 Hierarchical clustering	25
4.3 Partitioning clustering	26
4.3.1 Kmeans	26
4.3.2 Kmedoids	30
4.4 Model based clustering	30
Bibliography	36

List of Figures

1.1	Detail image of sensors [1].	6
2.1	Device barplot.	8
2.2	Left: Boxplot of carbon monoxide. Right: Histogram of carbon monoxide.	9
2.3	Carbon monoxide fitting.	9
2.4	Carbon monoxide fitting with 2 mixtures.	10
2.5	Carbon monoxide fitting with 3 mixtures.	10
2.6	Left: Boxplot of humidity. Right: Histogram of humidity.	10
2.7	Humidity fitting.	11
2.8	Left: Boxplot of LPG. Right: Histogram of LPG.	11
2.9	LPG fitting.	12
2.10	LPG fitting with 2 mixtures.	12
2.11	LPG fitting with 3 mixtures.	13
2.12	Barplot of light.	13
2.13	Barplot of motion.	13
2.14	Left: Boxplot of smoke. Right: Histogram of smoke.	14
2.15	Smoke fitting.	14
2.16	Smoke fitting with 2 mixtures.	15
2.17	Smoke fitting with 3 mixtures.	15
2.18	Left: Boxplot of temperature. Right: Histogram of temperature.	15
2.19	Temperature fitting.	16
2.20	Temperature fitting with 2 mixtures.	16
3.1	Pairplot of numeric dataset.	17
3.2	Correlation matrix.	18
3.3	Pairplot of numeric dataset with hue.	18
3.4	Scree plot.	19
3.5	PCA plot.	20
3.6	Biplot.	20
4.1	Pairplot on clusterization dataset.	21
4.2	Pairplot on benchmark dataset.	22
4.3	Pcaplot on benchmark dataset.	22
4.4	Pcaplot on clusterization dataset.	23
4.5	Distance matrix for clusterization dataset.	23
4.6	Distance matrix for benchmark dataset.	24
4.7	Elbow method.	24
4.8	Silhouette method.	24
4.9	GAP method.	25
4.10	Kmeans visualization.	25
4.11	Silhouette plot for kmeans clusterization.	25
4.12	Clara visualization.	26
4.13	Silhouette plot for clara clusterization.	26

4.14	Dissimilarity matrix.	27
4.15	Dendrogram of single linkage method.	27
4.16	Dendrogram of complete linkage method.	28
4.17	Dendrogram of average linkage method.	28
4.18	Dendrogram of ward linkage method.	28
4.19	Pairplot with average partition hue.	29
4.20	Pairplot with ward partition hue.	29
4.21	Pcaplot with average partition hue.	30
4.22	Pcaplot with ward partition hue.	30
4.23	Pairplot with kmeans partition hue.	31
4.24	Pairplot with kmeans partition hue and centroids.	31
4.25	Pcaplot with kmeans partition hue.	32
4.26	Pairplot with clara partition hue.	32
4.27	BIC plot.	33
4.28	Pairplot with model based partition hue.	34
4.29	Pcaplot with model based partition hue.	35
4.30	Pcaplot with classification uncertainty.	35
4.31	Bivariate marginalization.	35

List of Tables

1.1	Devices and environments description.	5
1.2	Columns description.	6
1.3	Dataset head.	7
1.4	Sampled dataset head.	7
1.5	Dataset summary.	7
2.1	Carbon monoxide fitting results.	9
2.2	Humidity fitting results.	11
2.3	LPG fitting results.	12
2.4	Smoke fitting results.	14
2.5	Temperature fitting results.	16
3.1	Principal component rotation.	19
3.2	Principal component center.	19
4.1	Confusion matrix comparing kmeans clusterization with external information.	26
4.2	Confusion matrix comparing clara clusterization with external information.	27
4.3	Clustering division according to the average linkage method.	29
4.4	Clustering division according to the ward linkage method.	29
4.5	Confusion matrix comparing average and ward clusterization.	30
4.6	Kmeans centroids.	31

Chapter 1

Data description

[1]

The data was generated from a series of three identical, custom-built, breadboard-based sensor arrays. Each array was connected to a Raspberry Pi devices. Each of the three IoT devices was placed in a physical location with varied environmental conditions.

Each breadboard-based sensor array is connected to a Raspberry Pi single-board computer (SBC), the popular, low cost, credit-card sized Linux computer. The IoT devices were purposely placed in physical locations that vary in temperature, humidity, and other environmental conditions.

device	environmental conditions
00:0f:00:70:91:0a	stable conditions, cooler and more humid
1c:bf:ce:15:ec:4d	highly variable temperature and humidity
b8:27:eb:bf:9d:51	stable conditions, warmer and dryer

Table 1.1: Devices and environments description.

Each device includes the following sensors:

- MQ135 Air Quality Sensor Hazardous Gas Detection Sensor: CO, LPG, Smoke
- DHT22/AM2302 Digital Temperature and Humidity Sensor
- Onyehn IR Pyroelectric Infrared PIR Motion Sensor
- Anmbest Light Intensity Detection Photosensitive Sensor

Each IoT device collected a total of seven different readings from the four sensors on a regular interval. Sensor readings include temperature, humidity, carbon monoxide (CO), liquid petroleum gas (LPG), smoke, light, and motion. The data spans the period from 07/12/2020 00:00:00 UTC – 07/19/2020 23:59:59 UTC. There is a total of 405,184 rows of data.

The sensor readings, along with a unique device ID and timestamp, were published as a single message, using the ISO standard Message Queuing Telemetry Transport (MQTT) network protocol. Below is an example of an MQTT message payload.

The script securely publishes the sensor readings, along with a device ID and timestamp, as a single message, to AWS using the ISO standard Message Queuing Telemetry Transport (MQTT) network protocol. Below is an example of an MQTT message payload, published by the collector script.

```
1 {  
2   "data": {  
3     "co": 0.006104480269226063,  
4     "humidity": 55.099998474121094,  
5     "light": true,
```

```

6   "lpg": 0.008895956948783413,
7   "motion": false,
8   "smoke": 0.023978358312270912,
9   "temp": 31.799999237060547
10  },
11 "device_id": "6e:81:c9:d4:9e:58",
12 "ts": 1594419195.292461
13 }

```

There are nine columns in the dataset.

column	description	units
ts	timestamp of event	epoch
device	unique device name	string
co	carbon monoxide	ppm (%)
humidity	humidity	percentage
light	light detected?	boolean
lpg	liquid petroleum gas	ppm (%)
motion	motion detected?	boolean
smoke	smoke	ppm (%)
temp	temperature	Fahrenheit

Table 1.2: Columns description.

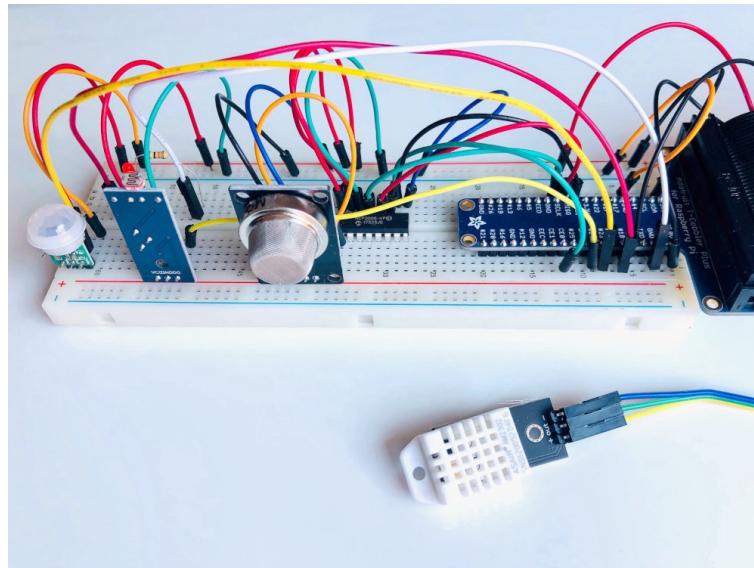


Figure 1.1: Detail image of sensors [1].

1.1 Overall dataset exploration

All code is available at [2]

```

head(df)
head(df) sampled taking 5000 observations
summary(df)

```

	ts	device	co	humidity	light	lpg	motion	smoke	temp
1	1594512094.38597	b8:27:eb:bf:9d:51	0.00496	51.00000	false	0.00765	false	0.02041	22.70000
2	1594512094.73557	00:0f:00:70:91:0a	0.00284	76.00000	false	0.00511	false	0.01327	19.70000
3	1594512098.07357	b8:27:eb:bf:9d:51	0.00498	50.90000	false	0.00767	false	0.02048	22.60000
4	1594512099.58915	1c:bf:ce:15:ec:4d	0.00440	76.80000	true	0.00702	false	0.01863	27.00000
5	1594512101.76123	b8:27:eb:bf:9d:51	0.00497	50.90000	false	0.00766	false	0.02045	22.60000
6	1594512104.46841	1c:bf:ce:15:ec:4d	0.00439	77.90000	true	0.00701	false	0.01859	27.00000

Table 1.3: Dataset head.

	ts	device	co	humidity	light	lpg	motion	smoke	temp
129000	1594733136.86349	Device 1	0.00491	55.70000	false	0.00760	false	0.02027	22.50000
50910	1594599159.13134	Device 3	0.00452	58.20000	true	0.00716	false	0.01902	27.00000
231919	1594907138.65742	Device 1	0.00625	50.70000	false	0.00905	false	0.02441	22.50000
274855	1594981012.56385	Device 1	0.00589	53.00000	false	0.00867	false	0.02333	22.20000
65686	1594624376.44225	Device 3	0.00467	59.70000	true	0.00733	false	0.01949	24.40000
258769	1594953228.90825	Device 1	0.00628	49.50000	false	0.00908	false	0.02451	21.70000

Table 1.4: Sampled dataset head.

ts	device	co	humidity	light
Min. : 2020-07-12 02:04:06.25	Length: 5000	Min. : 0.001181	Min. : 9.30	Length: 5000
1st Qu.: 2020-07-14 03:39:39.08	Class : character	1st Qu.: 0.003919	1st Qu.: 51.00	Class : character
Median : 2020-07-16 02:08:25.09	Mode : character	Median : 0.004816	Median : 54.70	Mode : character
Mean : 2020-07-16 02:31:04.67		Mean : 0.004652	Mean : 60.57	
3rd Qu.: 2020-07-18 01:37:13.19		3rd Qu.: 0.005439	3rd Qu.: 74.30	
Max. : 2020-07-20 02:03:09.08		Max. : 0.012863	Max. : 91.70	
lpg	motion	smoke	temp	
Min. : 0.002711	Length: 5000	Min. : 0.006738	Min. : 6.30	
1st Qu.: 0.006456	Class : character	1st Qu.: 0.017024	1st Qu.: 19.90	
Median : 0.007494	Mode : character	Median : 0.019964	Median : 22.30	
Mean : 0.007251		Mean : 0.019303	Mean : 22.46	
3rd Qu.: 0.008183		3rd Qu.: 0.021931	3rd Qu.: 23.50	
Max. : 0.015253		Max. : 0.042653	Max. : 30.60	

Table 1.5: Dataset summary.

Chapter 2

Univariate analysis

2.1 Time stamp

This variable represents the time of data acquisition.
(not sure if maintaining this variable)

2.2 Device

This is a categorical variable which indicates the device who took the measurements.
useful for cluster analysis given that each device is located in a different place

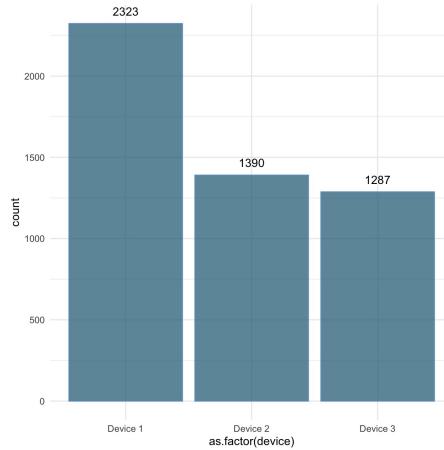


Figure 2.1: Device barplot.

2.3 Carbon monoxide

This variable is defined in R^+ and takes values between 0.001181 and 0.013397.
variance of co = 1.634931e-06 skewness of co = 0.3461658 kurtosis of co = 6.047793

2.3.1 Carbon monoxide fitting

2.4 Humidity

This variable is defined in the bounded interval [0, 100] and takes values between 5.80 and 99.90.

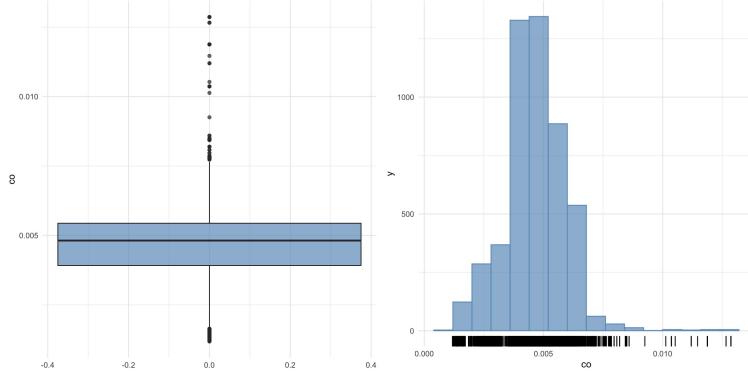


Figure 2.2: Left: Boxplot of carbon monoxide. Right: Histogram of carbon monoxide.

	df	logLik	AIC	BIC
NO	2	-5208.56191	10421.12383	10434.15661
LO	2	-4615.49080	9234.98161	9248.01440
GU	2	-5999.76216	12003.52433	12016.55711
RG	2	-7289.77785	14583.55571	14596.58849
exGAUS	3	-5208.83677	10423.67354	10443.22272
TF	3	-4392.86555	8791.73109	8811.28027

Table 2.1: Carbon monoxide fitting results.

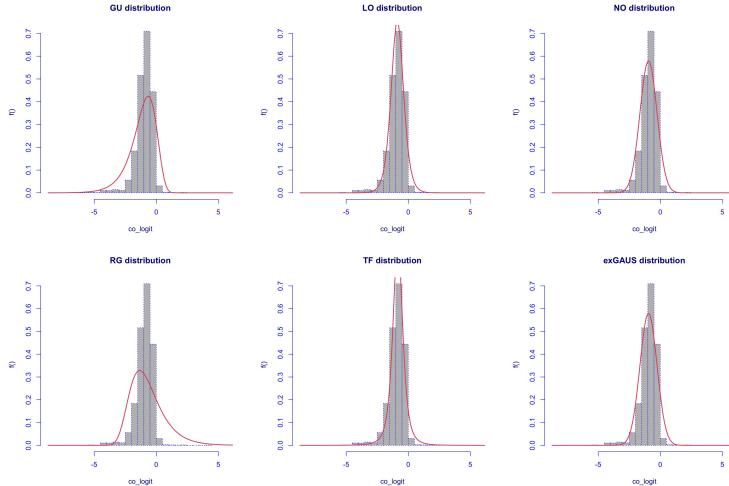


Figure 2.3: Carbon monoxide fitting.

var of humidity = 132.2306 skewness of humidity = 0.4784462 kurtosis of humidity = 1.819171
Dalla distribuzione delle variabili è evidente la presenza di cluster (variabili bimodali)

2.4.1 Humidity fitting

2.5 LPG

This variable is defined in $[0, 1000000]$ and takes values between 0.002711 and 0.015709

(liquefied petroleum gas)

variance of lpg = 2.164596e-06 skewness of lpg = -0.1105976 kurtosis of lpg = 4.828982

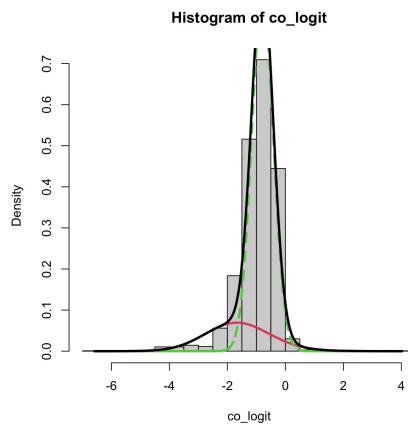


Figure 2.4: Carbon monoxide fitting with 2 mixtures.

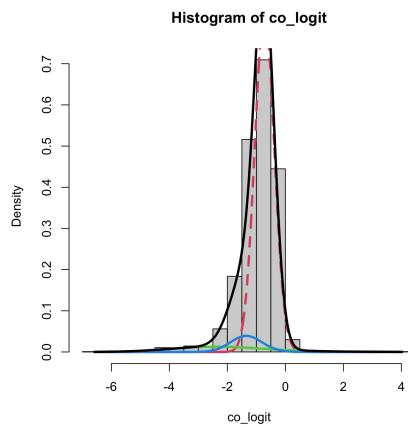


Figure 2.5: Carbon monoxide fitting with 3 mixtures.

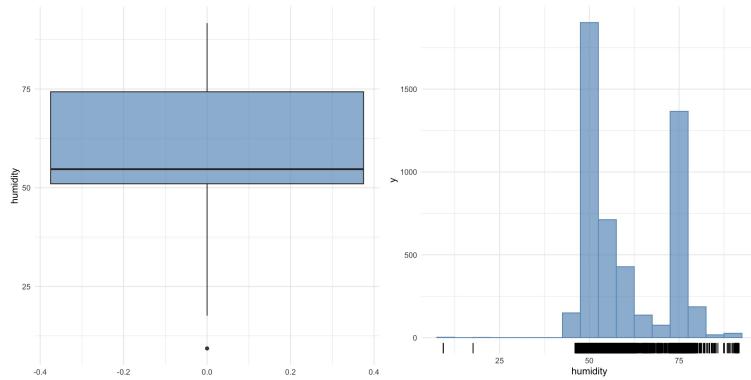


Figure 2.6: Left: Boxplot of humidity. Right: Histogram of humidity.

2.5.1 LPG fitting

2.6 Light

This variable is binary.

	df	logLik	AIC	BIC
NO	2	-5320.93070	10645.86140	10658.89458
LO	2	-5328.19709	10660.39418	10673.42737
GU	2	-7142.55456	14289.10913	14302.14232
RG	2	-4677.80659	9359.61318	9372.64636
exGAUS	3	-3943.92306	7893.84612	7913.39590
TF	3	-5233.70913	10473.41825	10492.96803

Table 2.2: Humidity fitting results.

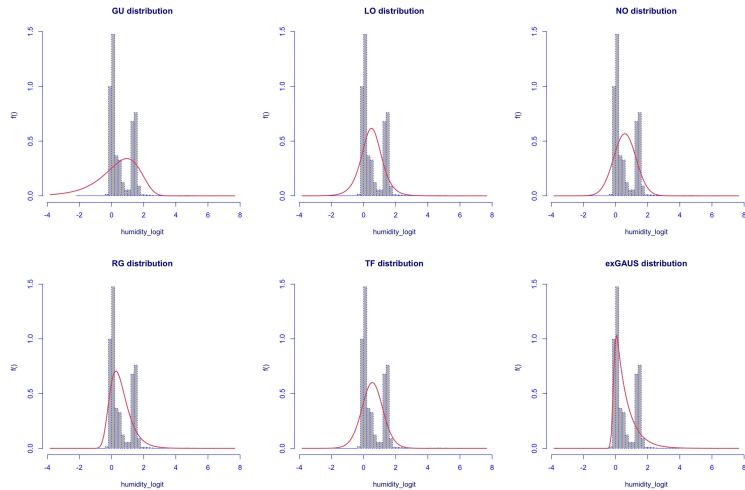


Figure 2.7: Humidity fitting.

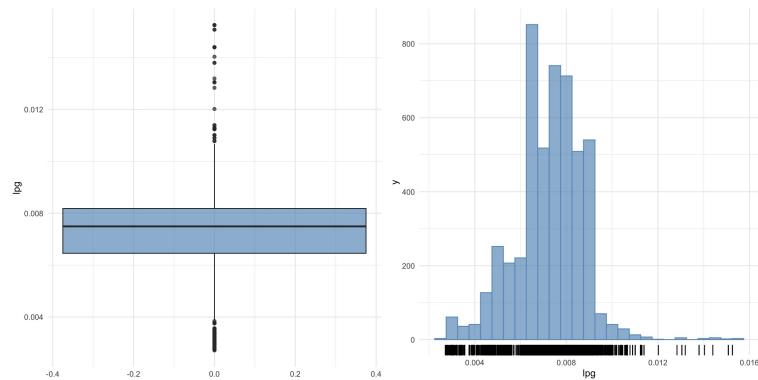


Figure 2.8: Left: Boxplot of LPG. Right: Histogram of LPG.

can be modeled as a bernoulli variable (need to check) ??? need to convert to numeric variable, unclass function doesn't work

2.7 Motion

This variable is binary.

	df	logLik	AIC	BIC
NO	2	-5005.78199	10015.56398	10028.59677
LO	2	-4391.64937	8787.29873	8800.33152
GU	2	-5858.55361	11721.10721	11734.14000
RG	2	-7128.95916	14261.91833	14274.95112
exGAUS	3	-5006.05828	10018.11656	10037.66574
TF	3	-4159.25674	8324.51348	8344.06266

Table 2.3: LPG fitting results.

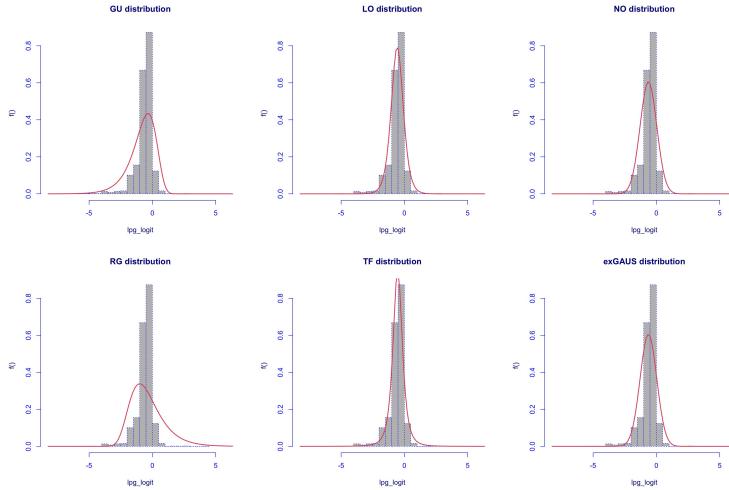


Figure 2.9: LPG fitting.

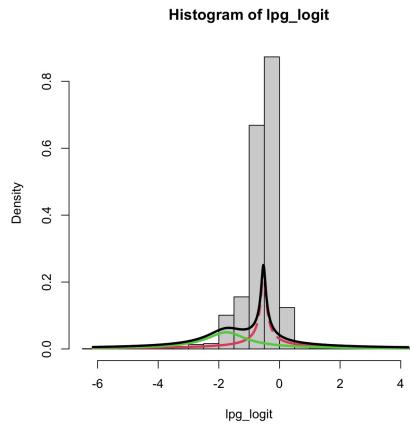


Figure 2.10: LPG fitting with 2 mixtures.

2.8 Smoke

This variable is defined in R^+ and takes values between 0.006738 and 0.044015

variance of smoke = 1.73533e-05 skewness of smoke = -0.03321988 kurtosis of smoke = 4.969701

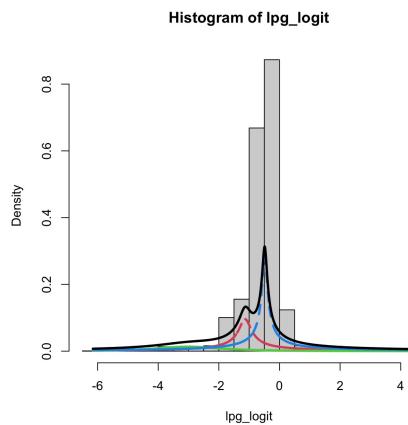


Figure 2.11: LPG fitting with 3 mixtures.

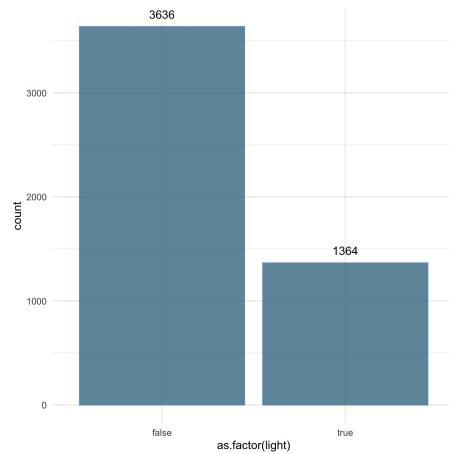


Figure 2.12: Barplot of light.

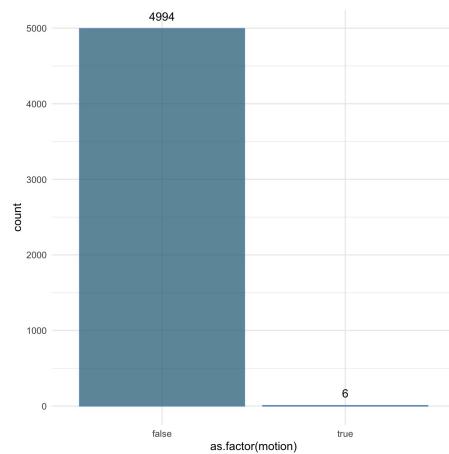


Figure 2.13: Barplot of motion.

2.8.1 Smoke fitting

2.9 Temperature

This variable is defined in *R* and takes values between 0.00 and 30.60. Temperature reported as Fahrenheit but it looks like it's wrongly reported as so, cause analyzing the ranges it seems like it's recorded in Celsius.

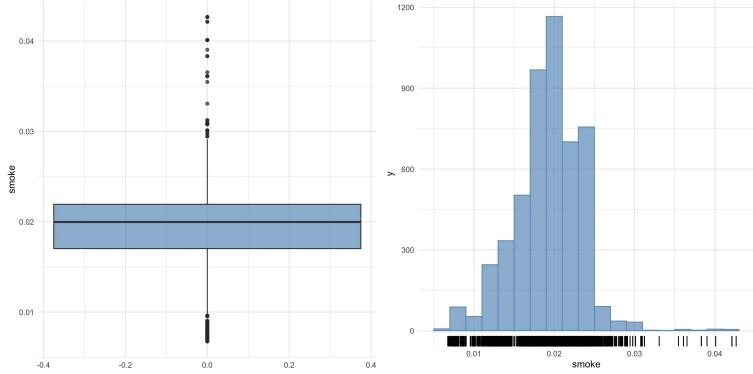


Figure 2.14: Left: Boxplot of smoke. Right: Histogram of smoke.

	df	logLik	AIC	BIC
NO	2	-5038.07210	10080.14420	10093.17699
LO	2	-4427.34973	8858.69946	8871.73225
GU	2	-5880.91568	11765.83137	11778.86415
RG	2	-7154.46931	14312.93862	14325.97140
exGAUS	3	-5038.34817	10082.69635	10102.24553
TF	3	-4196.53773	8399.07547	8418.62465

Table 2.4: Smoke fitting results.

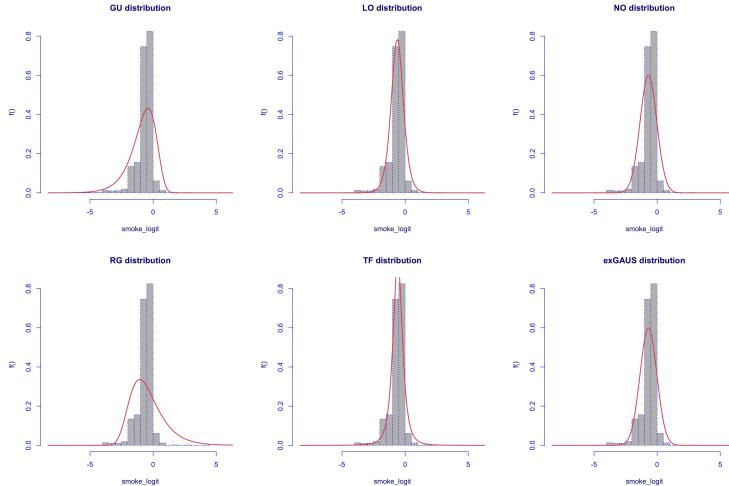


Figure 2.15: Smoke fitting.

variance of temp = 7.349724 skewness of temp = 0.6207737 kurtosis of temp = 3.812327
Dalla distribuzione delle variabili è evidente la presenza di cluster (variabili bimodali)

2.9.1 Temperature fitting

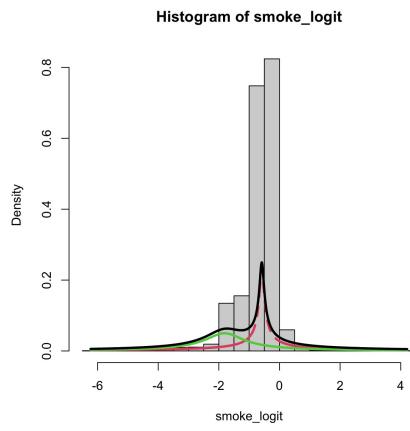


Figure 2.16: Smoke fitting with 2 mixtures.

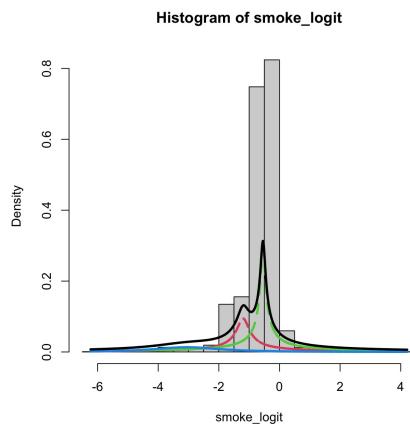


Figure 2.17: Smoke fitting with 3 mixtures.

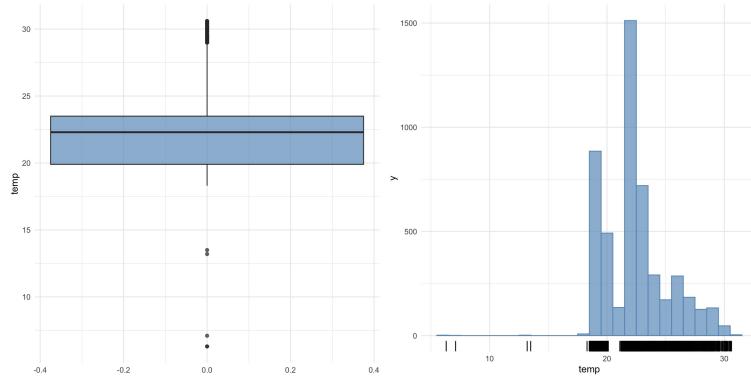


Figure 2.18: Left: Boxplot of temperature. Right: Histogram of temperature.

	df	logLik	AIC	BIC
NO	2.00000	-12080.84949	24165.69897	24178.73336
LO	2.00000	-12037.19654	24078.39309	24091.42747
GU	2.00000	-12908.87197	25821.74395	25834.77833
RG	2.00000	-12504.09408	25012.18815	25025.22254
exGAUS	3.00000	-11875.06112	23756.12224	23775.67382
TF	3.00000	-12043.93827	24093.87655	24113.42813

Table 2.5: Temperature fitting results.

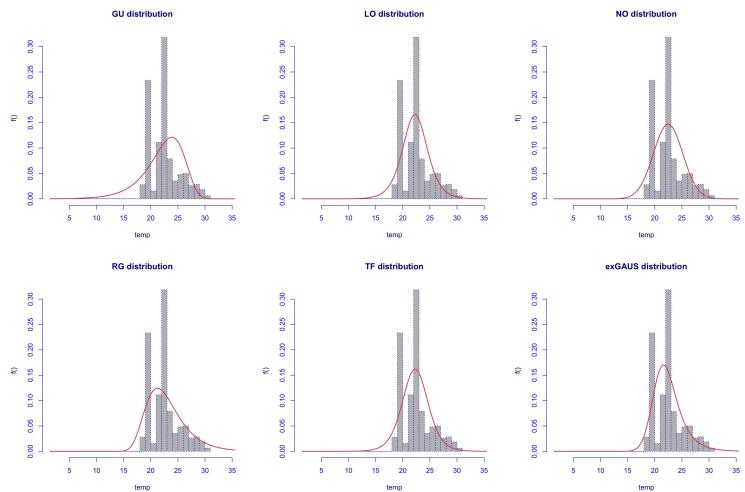


Figure 2.19: Temperature fitting.

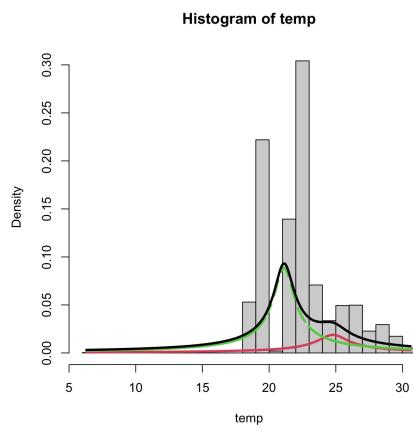


Figure 2.20: Temperature fitting with 2 mixtures.

Chapter 3

Principal component analysis

3.1 Multivariate exploratory data analysis

co, lpg and smoke are very positively correlated between each other

Evaluating the variance of Humidity and Temperature print(var[2]/sum(var)*100): humidity -> 94.7344

print(var[5]/sum(var)*100): temp -> 5.265588

df numeric scaled cov = cov(df numeric scaled) df numeric scaled eigen = eigen(df numeric scaled cov) str(df numeric scaled eigen)

values : num [1:5] 3.56 1.09 3.52e-01 2.68e-03 2.35e-07 vectors: num [1:5, 1:5] 0.516 -0.419 0.519 0.519 0.14 ... - attr(*, "class")= chr "eigen"

Need to have all variables with the same measurement unit ex. humidity is measured as a percentage, temperature is measured in °C Need also to have all variables with the same magnitude

```
1 df_numeric = df %>% select_if(is.numeric)
2 df_numeric_scaled = as.data.frame(scale(df_numeric))
```

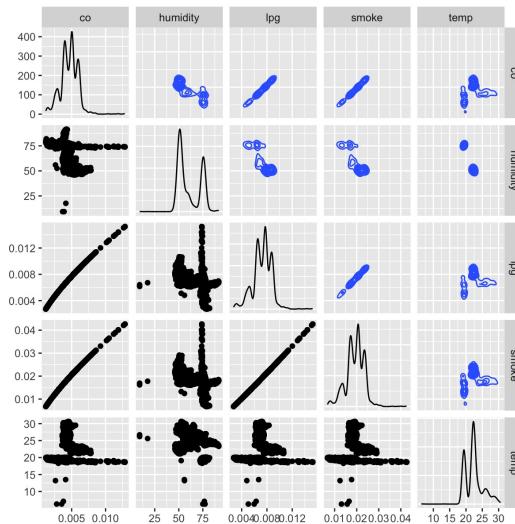


Figure 3.1: Pairplot of numeric dataset.

3.2 Applying PCA

```
print(pr.out -> sdev) 1.8862180887 1.0427915424 0.5933709464 0.0517475185 0.0004850959
```

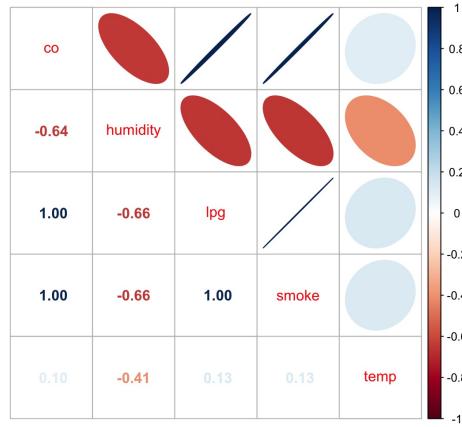


Figure 3.2: Correlation matrix.

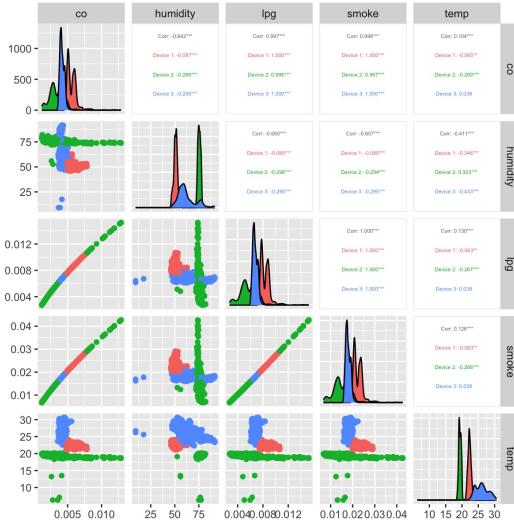


Figure 3.3: Pairplot of numeric dataset with hue.

3.3 Choosing the number of principal components

Cumulative proportion of explained variance

Kaiser rule

there's the -1 cause i'm working on the standardized data (1 it's the average of the eigen values)

```

1 kaiserVector = df_numeric_scaled$eigen$values - 1    # Positive values will be chosen
2
3 kaiserNumOfPC = 0
4 for(x in kaiserVector){
5   if(x > 0)
6     kaiserNumOfPC = kaiserNumOfPC + 1
7 }
8
9 print(paste("Number of PCs according to Kaiser Rule: ", kaiserNumOfPC))

```

Scree plot

	PC1	PC2	PC3	PC4	PC5
co	-0.51572	-0.18863	0.19467	0.80123	0.13636
humidity	0.41918	-0.33941	0.84195	-0.01459	-0.00040
lpg	-0.51916	-0.16140	0.18462	-0.52407	0.62904
smoke	-0.51879	-0.16626	0.18591	-0.28799	-0.76532
temp	-0.14021	0.89192	0.42962	0.01538	-0.00014

Table 3.1: Principal component rotation.

	V1
-1.90301108204949e-14	1
-3.1106561770855e-15	1
5.6656901392671e-15	1
7.45379313826788e-15	1
8.43272118800087e-15	1

Table 3.2: Principal component center.

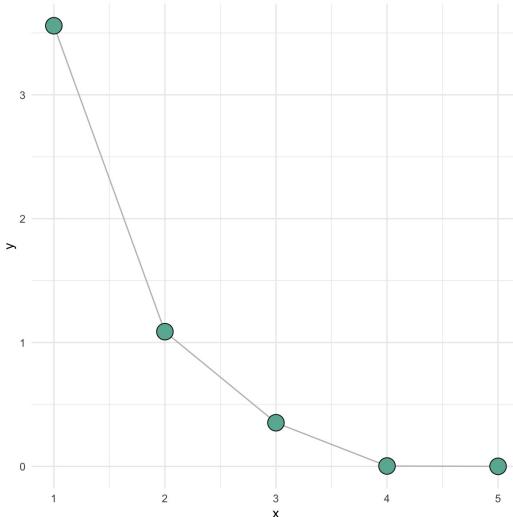


Figure 3.4: Scree plot.

According to the first 2 methods an optimal choice would be 2 PCs. The third method doesn't give a clear outcome. Personal consideration: the optimal choice is to pick 2 PCs cause the first one is correlated to co, lpg and smoke variables, the second one is correlated to temperature.

3.4 Data visualization in the principal components space

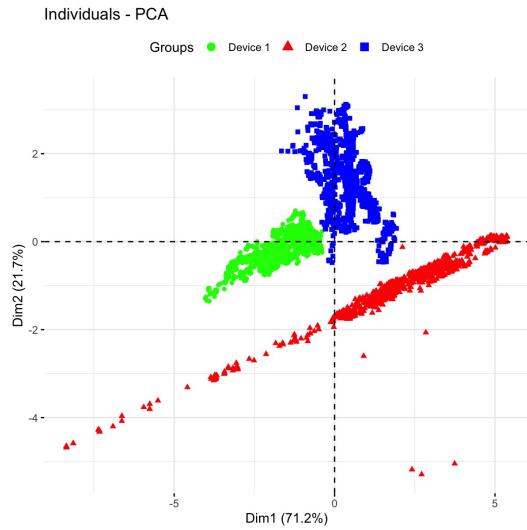


Figure 3.5: PCA plot.

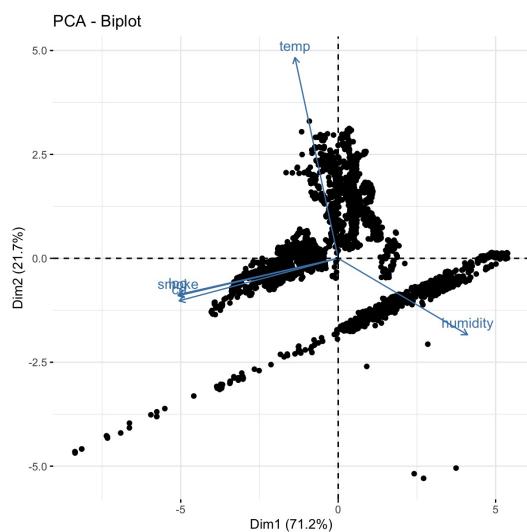


Figure 3.6: Biplot.

Chapter 4

Cluster analysis

dataset sampled to 500 observation for this analysis

4.1 Cluster validation

4.1.1 Assessing clustering tendency

```
1 # Generating the benchmark dataset from a uniform distribution
2 benchmark_df = apply(df_cl_numeric, 2, function(x){runif(length(x), min(x), max(x))})
3 benchmark_df = as.data.frame(benchmark_df)
4 benchmark_df_scaled = as.data.frame(scale(benchmark_df))
```

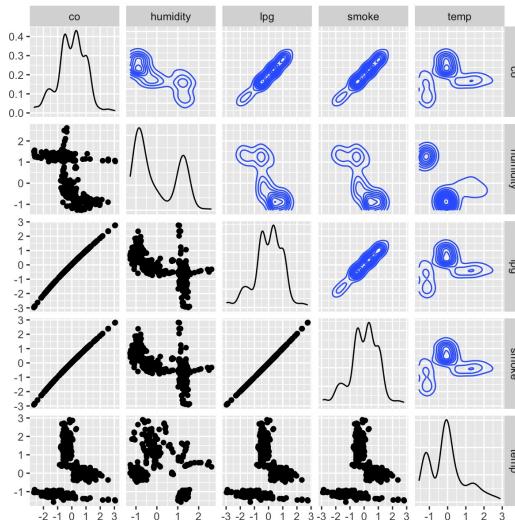


Figure 4.1: Pairplot on clusterization dataset.

We can see that the first PC influences the most the clusterization, while considering the projection on the second PC the red cluster and the blue one are mostly overlapped

Hopkins

hopkins(df cl numeric) = 0.9999879 hopkins(benchmark df) = 0.2625113

VAT

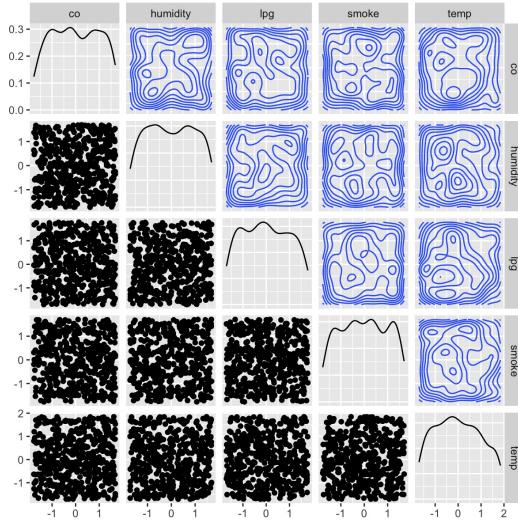


Figure 4.2: Pairplot on benchmark dataset.

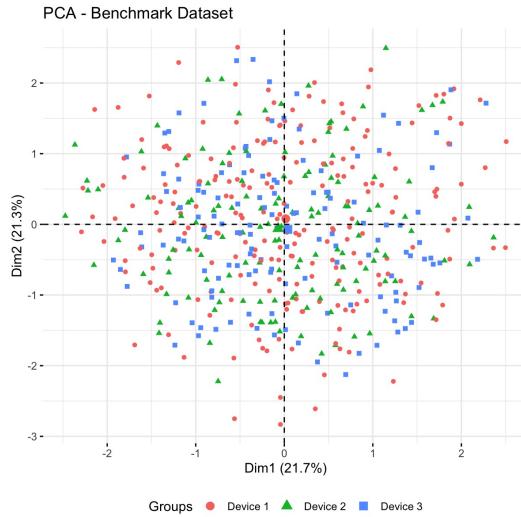


Figure 4.3: Pcaplot on benchmark dataset.

4.1.2 Determining the optimal number of clusters

4.1.3 Cluster validation statistics

Internal validation

Clustering Methods: hierarchical kmeans clara

Cluster sizes: 3 4

Validation Measures: 3 4

hierarchical APN 0.0008 0.0019 AD 2.5245 2.5117 ADM 0.0070 0.0110 FOM 0.9301 0.9260 Connectivity 8.2071
10.8861 Dunn 0.2053 0.1870 Silhouette 0.4299 0.3691 kmeans APN 0.0661 0.0837 AD 1.8031 1.3599 ADM 0.2753
0.2460 FOM 0.7184 0.5956 Connectivity 17.7159 19.7036 Dunn 0.0553 0.0924 Silhouette 0.5297 0.5474 clara APN
0.1101 0.0691 AD 1.5314 1.2689 ADM 0.3448 0.1997 FOM 0.6987 0.6151 Connectivity 32.9131 37.7024 Dunn 0.0043
0.0051 Silhouette 0.5164 0.4469

Optimal Scores:

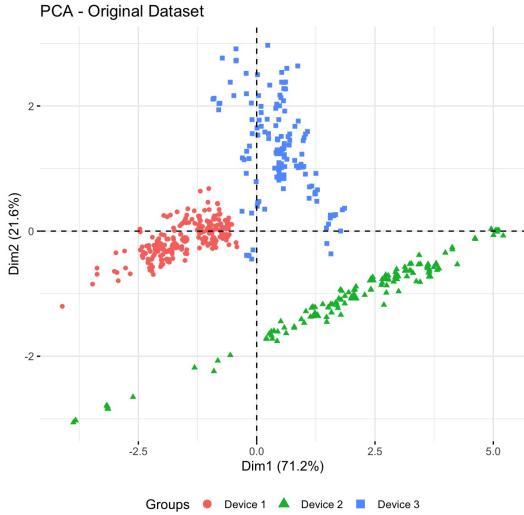


Figure 4.4: Pcaplot on clusterization dataset.

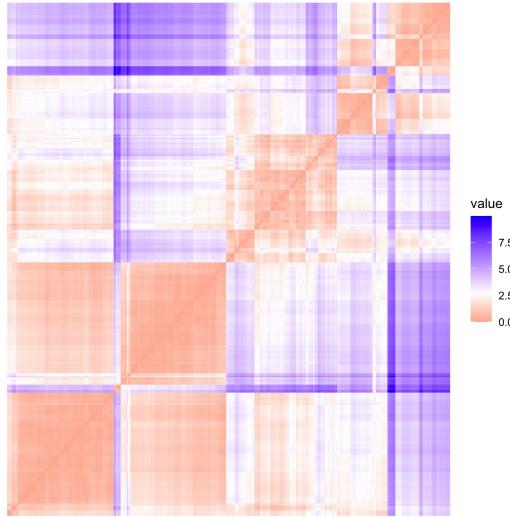


Figure 4.5: Distance matrix for clusterization dataset.

Score Method Clusters APN 0.0008 hierarchical 3 AD 1.2689 clara 4 ADM 0.0070 hierarchical 3 FOM 0.5956 kmeans 4 Connectivity 8.2071 hierarchical 3 Dunn 0.2053 hierarchical 3 Silhouette 0.5474 kmeans 4

External validation

KMEANS

```
1 km_res = factoextra::eclust(df_cl_numeric_scaled, FUNcluster="kmeans", k=3, nstart=10,
  graph=FALSE)
```

The results are decently good for each of the 3 clusters

Missclassified points: $15+14+6=35$ that is the 3.8% of the data
silhouette plot on kmeans

NEED TO REVISE In the first cluster there some units with a silhouette value lower than 0, this is a bad result for kmeans algorithm

CLARA

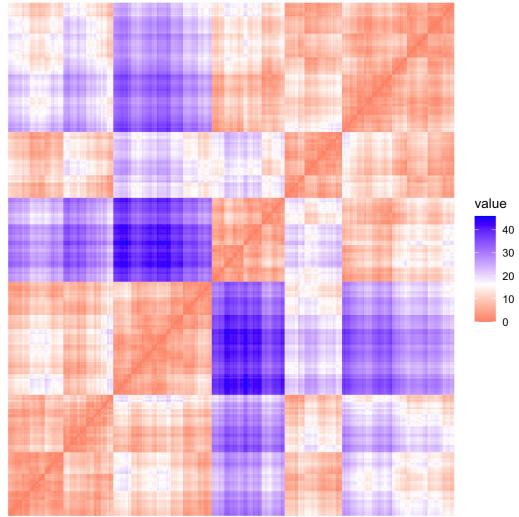


Figure 4.6: Distance matrix for benchmark dataset.

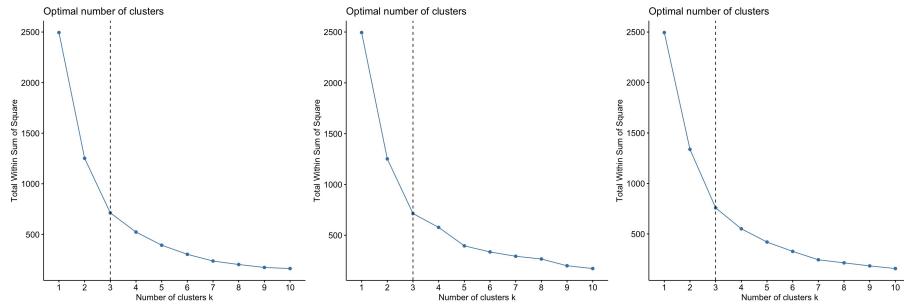


Figure 4.7: Elbow method.

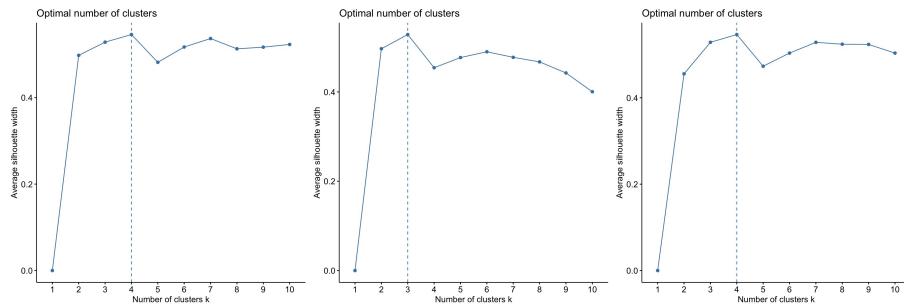


Figure 4.8: Silhouette method.

```
1 clara_res = factoextra::eclust(df_cl_numeric_scaled, k=3, FUNcluster="clara", graph=FALSE)
```

The results are decently good for each of the 3 clusters and they are practically the same of the kmeans partition
 Missclassified point: $17+8+10=35$ that is the 3.8% of the data
 silhouette plot on clara
 The results are good

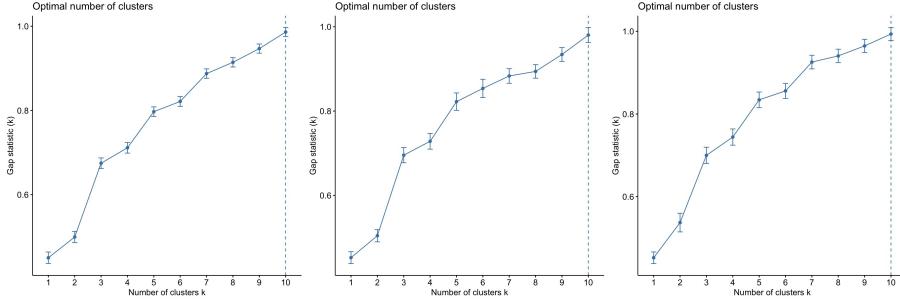


Figure 4.9: GAP method.

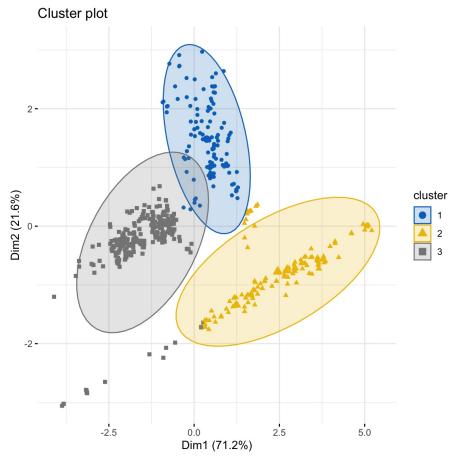


Figure 4.10: Kmeans visualization.

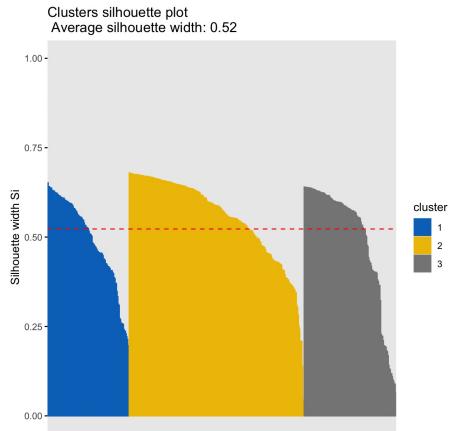


Figure 4.11: Silhouette plot for kmeans clusterization.

4.2 Hierarchical clustering

Dissimilarity matrix

Dendograms

cophenetic distances

Correlation between original distance vector and cophenetic of single method = 0.7732038

Correlation between original distance vector and cophenetic of complete method = 0.7110569

	Device 1	Device 2	Device 3
1	0	0	117
2	0	121	14
3	227	15	6

Table 4.1: Confusion matrix comparing kmeans clusterization with external information.

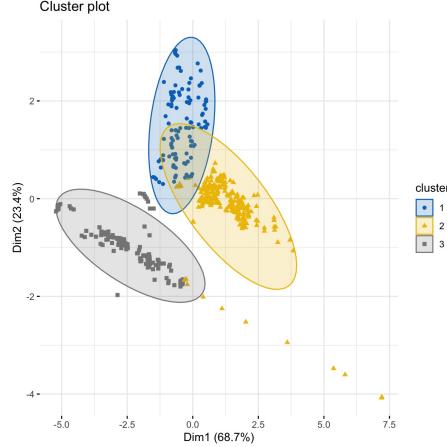


Figure 4.12: Clara visualization.

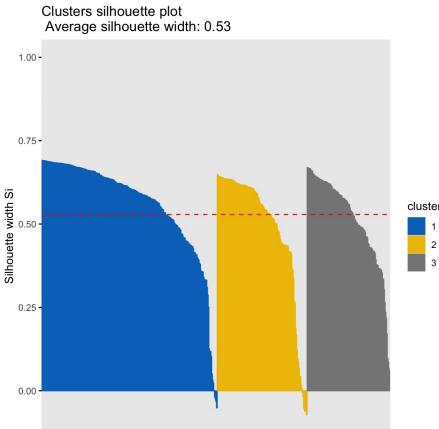


Figure 4.13: Silhouette plot for clara clusterization.

Correlation between original distance vector and cophenetic of average method = 0.8204009

Correlation between original distance vector and cophenetic of ward method = 0.6995259

NEED TO REVISE According to the two previous results I think there are outliers cause some clusters are containing only 1 or 2 points From the scatterplot it's clearer their presence

PAIRPLOTS

Tell which are the best variable for clustering

PCAPLOTS

4.3 Partitioning clustering

4.3.1 Kmeans

	Device 1	Device 2	Device 3
1	227	17	8
2	0	119	10
3	0	0	119

Table 4.2: Confusion matrix comparing clara clusterization with external information.

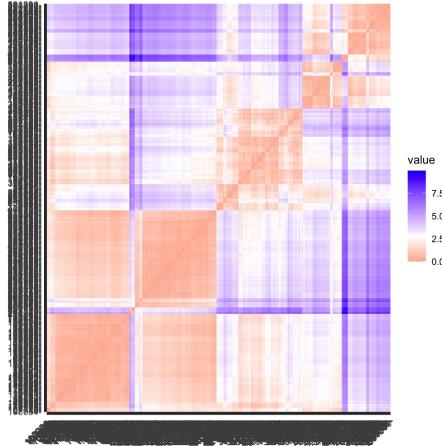


Figure 4.14: Dissimilarity matrix.

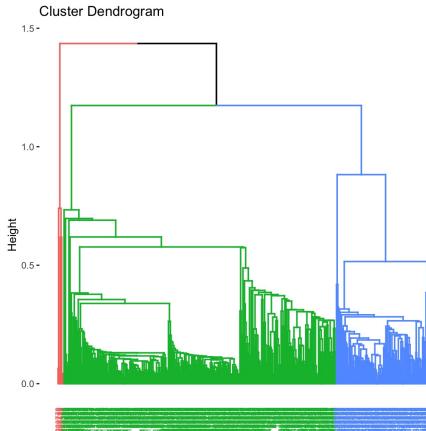


Figure 4.15: Dendrogram of single linkage method.

```

1 km = kmeans(df_cl_numeric_scaled, centers=3, nstart=25)
km$size = 135 117 248 Important to evaluate the presence of outliers
km$centers
Data is scaled so the mean for each variable is 0
Units in cluster 1 are characterized by an above average humidity and the other values are below average
Units in cluster 2 are characterized by an above average temperature, slightly below average values co, lpg and
smoke and more or less average humidity
Units in cluster 3 are characterized by an above average co, lpg and smoke, an average temperature and a below
average humidity
This process is useful in order to assign new units to the clusters
Variability explained by the clustering
1 clusterVariability = scales::label_percent()(km$betweenss / km$totss)

```

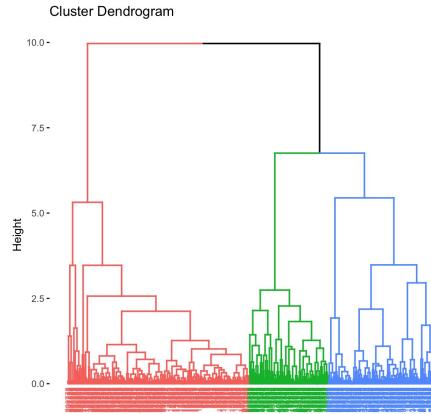


Figure 4.16: Dendrogram of complete linkage method.

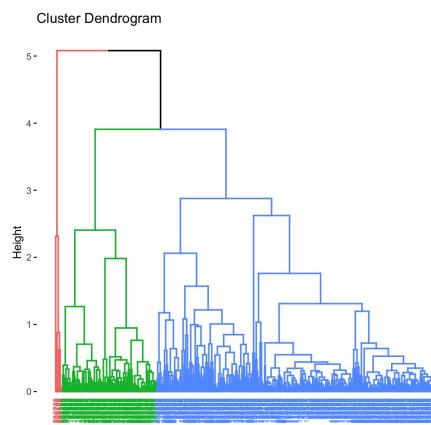


Figure 4.17: Dendrogram of average linkage method.

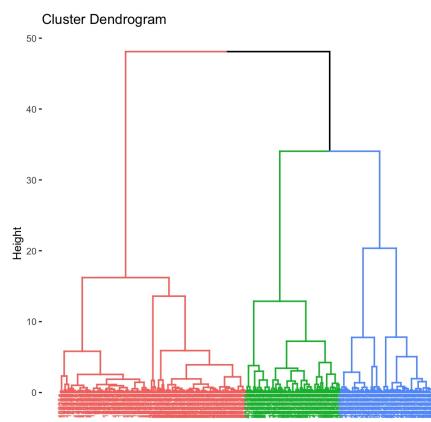


Figure 4.18: Dendrogram of ward linkage method.

```
2 print(clusterVariability)
```

So this clustering method explains the 71% of the original variability

	groupsAverage
1	367
2	124
3	9

Table 4.3: Clustering division according to the average linkage method.

	groupsWard
1	247
2	128
3	125

Table 4.4: Clustering division according to the ward linkage method.

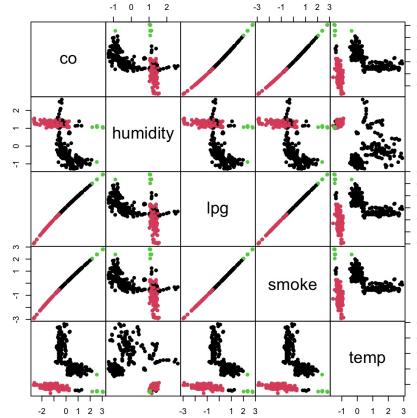


Figure 4.19: Pairplot with average partition hue.

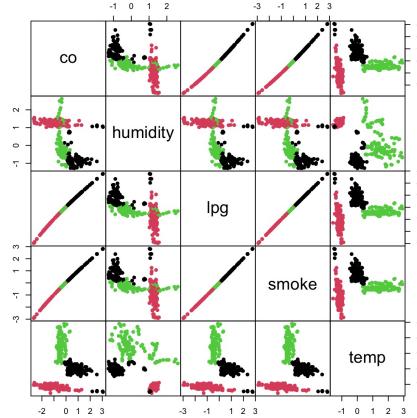


Figure 4.20: Pairplot with ward partition hue.

Useful to select the best clustering method given the number of clusters K (the higher the better)

```

1 RDatasetCentr_scaled = rbind(df_cl_numeric_scaled, km$centers)
2 clusterMembership = km$cluster
3 clusterMembership_new = c(km$cluster, rep(4,3)) # Adding 3 units with cluster membership
  4

```

	1	2	3
1	238	4	125
2	0	124	0
3	9	0	0

Table 4.5: Confusion matrix comparing average and ward clusterization.

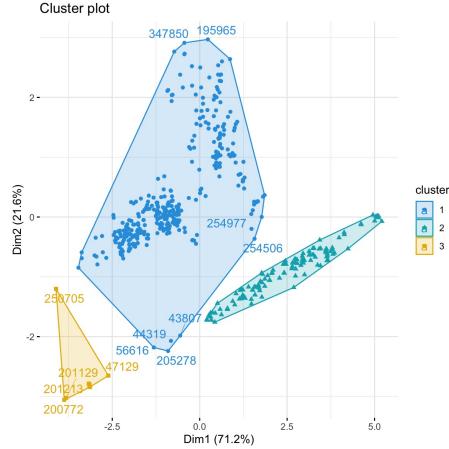


Figure 4.21: Pcaplot with average partition hue.

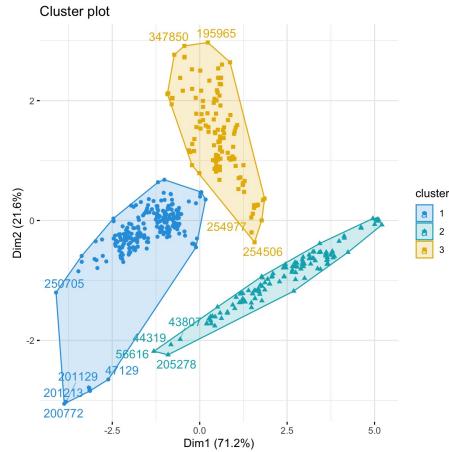


Figure 4.22: Pcaplot with ward partition hue.

Analizing the clusters means of the original dataset

4.3.2 Kmedoids

I expect a better result cause I think in the dataset there are some outliers

```
1 clara_cl = cluster::clara(df_cl_numeric_scaled, 3, metric="euclidean", stand=FALSE)
```

Confusion matrix to compare the clara result with external information

Confusion matrix to compare the clara result with kmeans result

4.4 Model based clustering

	co	humidity	lpg	smoke	temp
1	-1.12815	1.33131	-1.14747	-1.14445	-0.98356
2	-0.37662	0.00366	-0.33577	-0.34336	1.45026
3	0.79179	-0.72643	0.78304	0.78498	-0.14879

Table 4.6: Kmeans centroids.

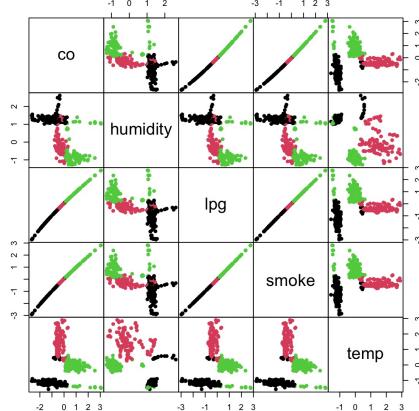


Figure 4.23: Pairplot with kmeans partition hue.

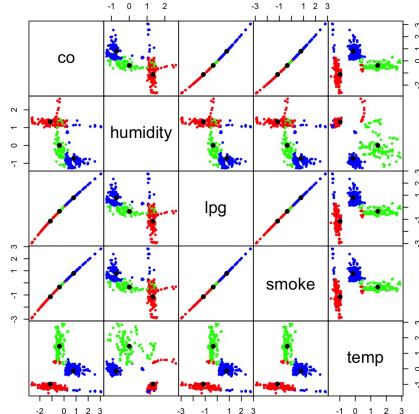


Figure 4.24: Pairplot with kmeans partition hue and centroids.

```

1 mbc = mclust::Mclust(df_cl_numeric_scaled, G=1:10) # Considering K from 1 to 10
Best model: print(mbc$modelName) = VVV print(mbc$G) = 10
Head of matrix of soft classification
Head of matrix of hard classification
External cluster validation
Adjusted rand index
0.4908627
0 = random partition, 1 = perfect agreement

```

	cluster	co	humidity	lpg	smoke	temp
1	1	0.00324	76.37185	0.00558	0.01460	19.86000
2	2	0.00418	61.07521	0.00676	0.01789	26.33675
3	3	0.00564	52.66331	0.00839	0.02252	22.08145

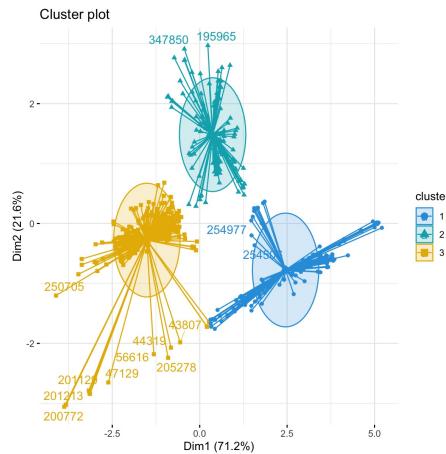


Figure 4.25: Pcaplot with kmeans partition hue.

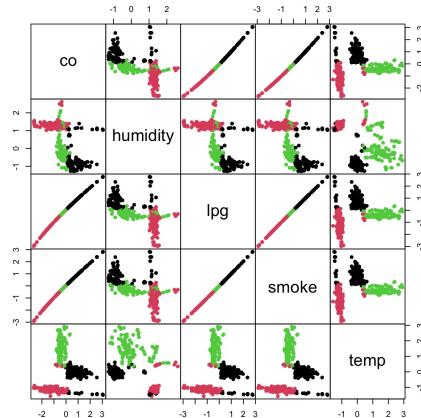


Figure 4.26: Pairplot with clara partition hue.

	Device 1	Device 2	Device 3
1	227	17	8
2	0	119	10
3	0	0	119

	1	2	3
1	2	0	115
2	2	129	4
3	248	0	0

	VVV,10	VVV,9	VVV,8
BIC	15292.21	15275.51957	15041.6859
BIC diff	0.00	-16.68917	-250.5228

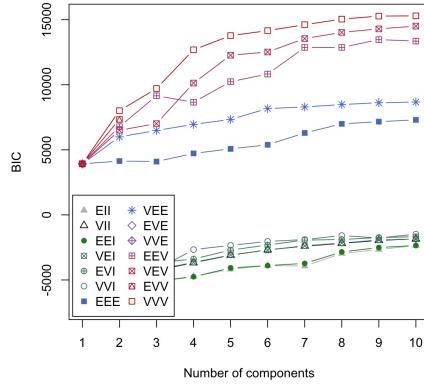


Figure 4.27: BIC plot.

	1	2	3	4	5	6	7	8	9	10
76925	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
210667	0.99859	0.00000	0.00141	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
124790	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
125528	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
216303	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
4670	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
24890	0.00000	0.00000	0.00019	0.00000	0.00000	0.00000	0.00000	0.00000	0.99981	0.00000
191005	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
177054	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
300335	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
116782	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
181467	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
147682	0.99967	0.00000	0.00033	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
383749	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
38641	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
192610	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
2425	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
27056	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
34967	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
41062	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

	1	2	3	4	5	6	7	8	9	10
76925	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
210667	0.99859	0.00000	0.00141	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
124790	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
125528	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
216303	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
4670	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
24890	0.00000	0.00000	0.00019	0.00000	0.00000	0.00000	0.00000	0.00000	0.99981	0.00000
191005	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
177054	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
300335	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
116782	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
181467	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
147682	0.99967	0.00000	0.00033	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
383749	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
38641	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
192610	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
2425	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
27056	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
34967	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
41062	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

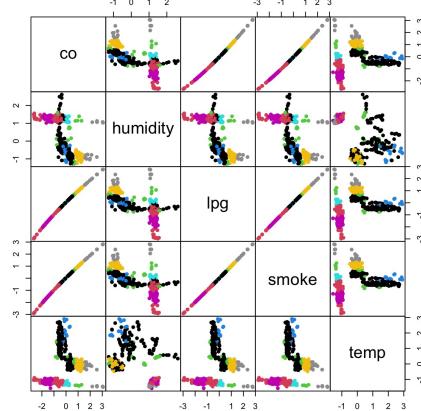


Figure 4.28: Pairplot with model based partition hue.

	1	2	3	4	5	6	7	8	9	10
Device 1	110	0	7	0	0	0	99	11	0	0
Device 2	0	39	4	0	15	57	0	8	0	13
Device 3	0	0	17	16	0	0	0	0	104	0

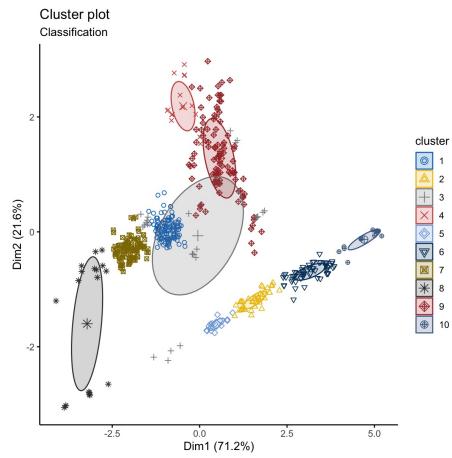


Figure 4.29: Pcaplot with model based partition hue.

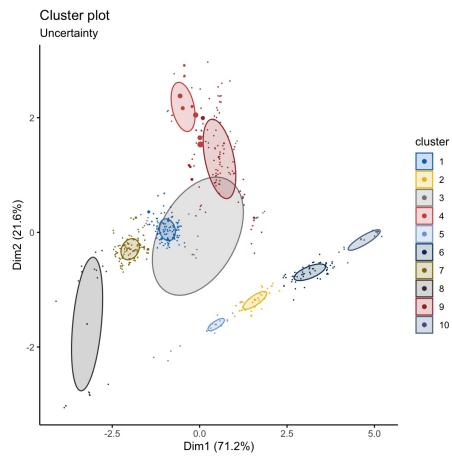


Figure 4.30: Pcaplot with classification uncertainty.

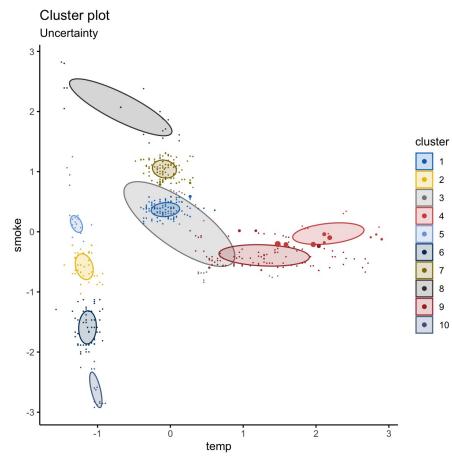


Figure 4.31: Bivariate marginalization.

Bibliography

- [1] *Environmental sensor telemetry data*, www.kaggle.com. [Online]. Available: <https://www.kaggle.com/datasets/garystafford/environmental-sensor-data-132k>.
- [2] G. Spadaro, *Iot-telemetry-analysis*, GitHub, 2023. [Online]. Available: <https://github.com/Giovo17/iot-telemetry-analysis>.