Part1- Cluster Analysis

We have been provided a dataset of people's responses from the European Working Conditions Survey 2016, where 7813 individuals were asked to fill surveys about their mental health, work-life balance, and job satisfaction. Our dataset consists of mixed data type variables with 11 dimensions such as numerical variable Age and Categorical variable Gender. Our task is to explore the dataset and provide a description of the data by unsupervised learning methods such as Principal Component Analysis, Clustering.

I will be using Clustering Unsupervised Learning Algorithm for exploring the dataset. There are different types of Clustering analysis such as K-means clustering, Hierarchical clustering, etc. K-means clustering algorithm is an elegant approach for partitioning the dataset into K distinct clusters where an analyst first has to define the hyper-parameter K which is the required number of centroids and then K clusters are created by allocating each data point to the nearest centroid based on Euclidean distance. Since this distance is valid only for continuous variables and our dataset is of mixed data type variables, we won't be using Euclidean distance methodology instead we will be measuring similarity across individuals using Gower Distance. Gower distance fits well with the PAM algorithm i.e. Partitioning Around Medoids which is very similar to K-means but it is more robust to noise & outliers and produces clusters of very similar individuals which are useful during interpretation.

The optimal number of clusters

To find the optimal number of clusters, we are going to use the Silhouette coefficient. "Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified" (Wikipedia, 2019) [1].

Finally, we have visualized our clusters in low-dimensionality by using the Rtsne function of R which stands for "t-Distributed Stochastic Neighbor Embedding. tsne is a technique for

dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets" (Filaire 2018) [2].

DataPreprocessing-

1) There were some unusual values in the dataset such as -999 which have been replaced with the mode value of that particular column. 2) For better readability, column names have been changed intuitively such as Q2a to 'Gender' and Q87a to 'Cheerful'. 3) Since the structure of discreet answers of people for every categorical question has been interpreted as an integer in R, they have been converted into factors i.e. categories. 4) For better interpretability, I have assigned given Labels to every discreet number such as for 1 it is "All of the time", for 2 it is "Most of the time" and so on. 5) Gower distance has been calculated between every individual, male & female in our case, using the daisy function of R. Through the matrix, based on Gower distance, we can find that the distance between same variables is intuitively zero. To view the most similar and dissimilar individuals we can use min and max function on this matrix of the dataset and here are the results-

Most Similar case:

Gender	Age	Cheerful	Calm	Active	Fresh	InterestingLife
1: Male	62	More than half of the time	More than half of the time	More than half of the time	More than half of the time	More than half of the time
2: Male	63	More than half of the time	More than half of the time	More than half of the time	More than half of the time	More than half of the time

Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
Most of the time	Most of the time	Most of the time	Most of the time
Most of the time	Most of the time	Most of the time	Most of the time

Most Dissimilar case:

Gender	Age	Cheerful	Calm	Active	Fresh	InterestingLife
1: Male	87	Most of the time	Most of the time	Most of the time	Most of the time	Most of the time
2: Female	18	All of the time	All of the time	All of the time	All of the time	All of the time

Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
Always	Most of the time	Sometimes	Most of the time
Most of the time	Never	Most of the time	Always

In the most similar case, the distance between two male individuals, aged 62 and 63, found to be the lowest among all and their responses to every question of the survey found similar to each other. 6) Next, using the Silhouette coefficient we found the optimal number of clusters which is 3 in our case. 7) We fit the model using the PAM function at K=3 and below is the visualization and patterns discovered in each cluster.

Cluster 1

Gender	Age	Cheerful	Calm	Active	Fresh
Male : 933	Min. :18.00	All of the time : 31	All of the time : 29	All of the time : 46	All of the time : 19
Female:1359	1st Qu.:37.00	Most of the time : 337	Most of the time : 263	Most of the time : 315	Most of the time : 214
	Median :47.00	More than half of the time:1297	More than half of the time:1223	More than half of the time:1257	More than half of the time:1151
	Mean :46.28	Less than half of the time: 372	Less than half of the time: 467	Less than half of the time: 411	Less than half of the time: 485
	3rd Qu.:55.00	Some of the time : 229	Some of the time : 257	Some of the time : 222	Some of the time : 332
	Max. :87.00	At no time : 26	At no time : 53	At no time : 41	At no time : 91

InterestingLife	Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
All of the time : 79	Always : 82	Always : 164	Always : 226	Always : 671
Most of the time : 369	Most of the time: 780	Most of the time: 683	Most of the time: 704	Most of the time:1367
More than half of the time:1142	Sometimes :1221	Sometimes :1044	Sometimes :1045	Sometimes : 202
Less than half of the time: 395	Rarely : 168	Rarely : 321	Rarely : 257	Rarely : 40
Some of the time : 272	Never : 41	Never : 80	Never : 60	Never : 12
At no time : 35				

Cluster2

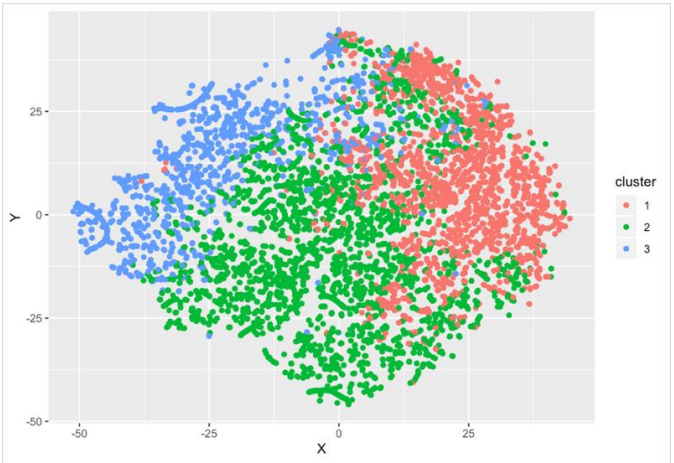
Gender	Age	Cheerful	Calm	Active	Fresh
Male :2305	Min. :15.00	All of the time : 254	All of the time : 237	All of the time : 305	All of the time : 182
Female:1480	1st Qu.:33.00	Most of the time :2607	Most of the time :2343	Most of the time :2650	Most of the time :2275
	Median :42.00	More than half of the time: 598	More than half of the time: 706	More than half of the time: 511	More than half of the time: 753
	Mean :42.69	Less than half of the time: 173	Less than half of the time: 262	Less than half of the time: 158	Less than half of the time: 280
	3rd Qu.:51.00	Some of the time : 131	Some of the time : 183	Some of the time : 131	Some of the time : 224
	Max. :87.00	At no time : 22	At no time : 54	At no time : 30	At no time : 71

InterestingLife	Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
All of the time : 474	Always : 506	Always : 791	Always : 792	Always :2200
Most of the time :2370	Most of the time:2695	Most of the time:2068	Most of the time:1988	Most of the time:1403
More than half of the time: 562	Sometimes : 465	Sometimes : 641	Sometimes : 765	Sometimes : 149
Less than half of the time: 187	Rarely : 87	Rarely : 206	Rarely : 170	Rarely : 19
Some of the time : 155	Never : 32	Never : 79	Never : 70	Never : 14
At no time : 37				

Cluster 3

Gender	Age	Cheerful	Calm	Active	Fresh
Male :739	Min. :17.00	All of the time :1174	All of the time :1061	All of the time :1219	All of the time :975
Female:997	1st Qu.:31.00	Most of the time : 287	Most of the time : 264	Most of the time : 247	Most of the time :318
	Median :40.00	More than half of the time: 103	More than half of the time: 121	More than half of the time: 88	More than half of the time:126
	Mean :40.57	Less than half of the time: 62	Less than half of the time: 93	Less than half of the time: 54	Less than half of the time: 92
	3rd Qu.:49.25	Some of the time : 69	Some of the time : 120	Some of the time : 80	Some of the time :136
	Max. :82.00	At no time : 41	At no time : 77	At no time : 48	At no time : 89

InterestingLife	Energetic	WorkEnthusiastic	Workaholic	JobSatisfaction
All of the time :1232	Always :1146	Always :1175	Always :1098	Always :1459
Most of the time : 240	Most of the time: 372	Most of the time: 289	Most of the time: 294	Most of the time: 231
More than half of the time: 102	Sometimes : 137	Sometimes : 145	Sometimes : 250	Sometimes : 30
Less than half of the time: 51	Rarely : 47	Rarely : 63	Rarely : 54	Rarely : 5
Some of the time : 72	Never : 34	Never : 64	Never : 40	Never : 11
At no time : 39				



Cluster1: Female Cluster

The majority of individuals in cluster 1 are females with a median age of 47 years and more than half of the time they have been feeling Cheerful, Calm, Active, Fresh over the last two weeks including the things that interest their daily life. It seems individuals in this cluster are pretty satisfied with their job though they don't seem very workaholic, enthusiastic and energetic. We can call this cluster of individuals, satisfied people.

Cluster2: Male Cluster

People in this cluster are mainly males and this Male dominating cluster is happier and diligent than the Female dominating cluster. Most of the time individuals in this cluster feel Cheerful, Calm, Fresh in their routine life and they seem passionate about their work with a high level of job satisfaction. An interesting pattern to note is, this cluster has more number of individuals than any other one. We can call this cluster of individuals, happy and hard-working people.

Cluster3- Amalgam of Males and Females

This is the smallest cluster, with 1736 individuals in total out of which 739 are males and 997 are females. This cluster has a better male to female ratio than the other two clusters. With a median age of 40, the majority of the participants in this cluster are in good mental health and enjoy their work. A whopping 1459 (84%) people are satisfied with their job all the time and over 1230 people (71%) consider their life interesting all the time whereas over 1150 (68%) people feel cheerful all the time. In each survey question, there are less than 90 people with a negative response to either mental health or work satisfaction related questions. These impressive numbers tell us that people in this cluster are generally happy with both their life and work with a higher degree of satisfaction.