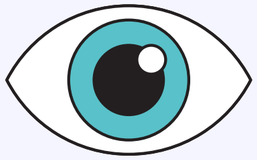
Dry Eye Disease

Team Members:
Javier Merino
Meyliani Sanjaya
Angeli De los Reyes
Nay Zaw Lin



Today's Agenda

1 Data Description

2 Objectives

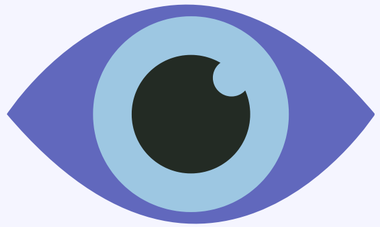
3 Analysis

4 Results

5 Conclusions

6 Next Steps

1 Data Description - Dataset



What?

Data Set describe the diagnostic of Dry Eye Disease(“Yes” or “No”) and key attributes of the subjects related to health and life choices.

Who?

Individuals aged 18 to 45 from India of both genders, whose sleep patterns, lifestyle choices, and medical conditions have been recorded for studying Dry Eye Disease.

Where?

Sourced from Kaggle in CSV format https://www.kaggle.com/datasets/dakshnagra/dry-eye-disease/data

1 Data Description - Features

The dataset includes 20,000 observations organized into 26 columns

PREPARATION

1. Sample of 1%: 200 Observations
2. Scale of the numerical features

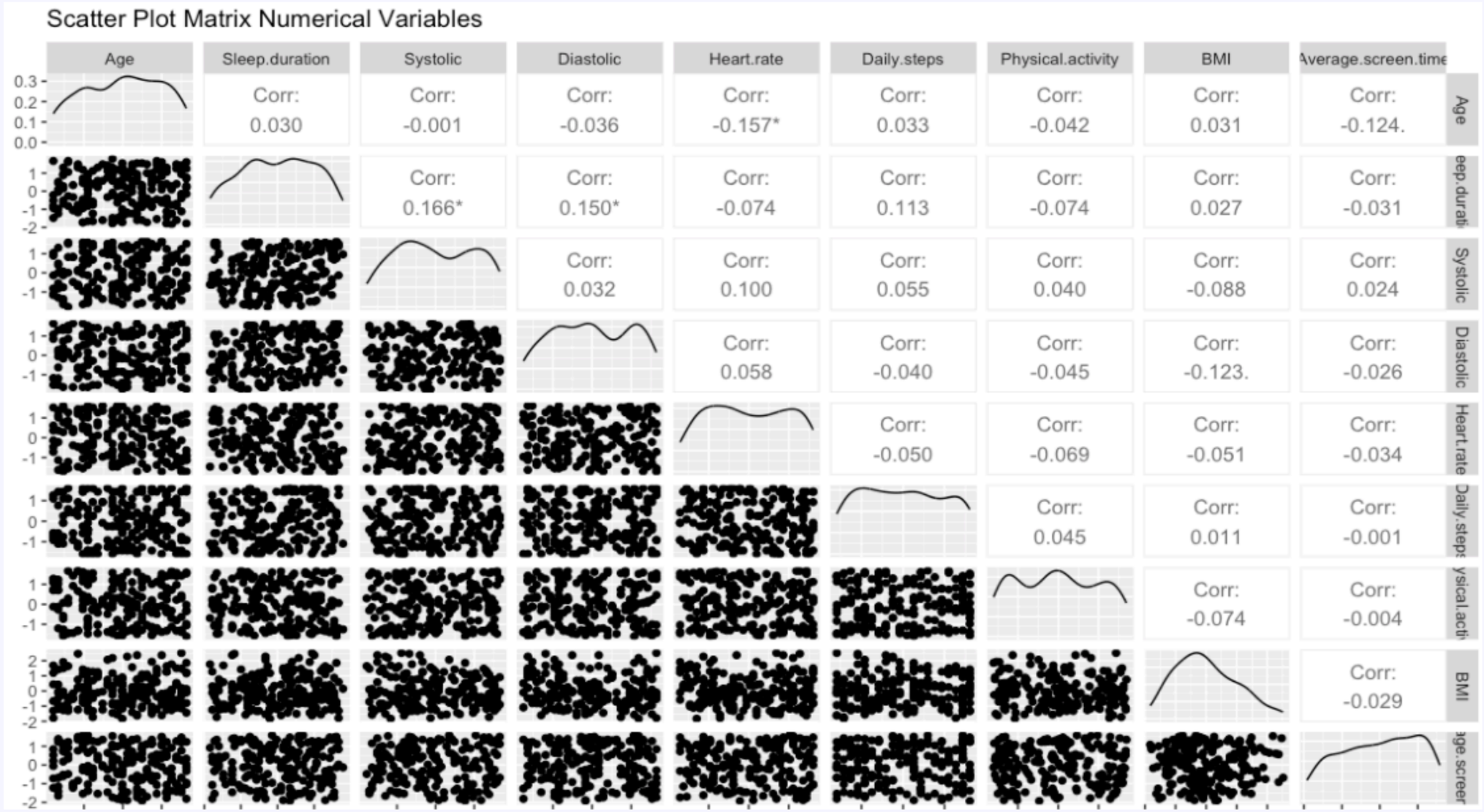
Numerical

1. Age (years)
2. Sleep duration (hours)
3. Diastolic pressure (mmHg)
4. Systolic pressure (mmHg)
5. Heart rate (beats/min)
6. Daily steps (steps/day)
7. Physical activity (min/day)
8. BMI (kg/m2)
9. Average screen time (hours/day)

Categorical

1. Gender (M/F)
2. Sleep quality (1-5)
3. Stress level (1-5)
4. Sleep disorder (Y/N)
5. Wake up during night (Y/N)
6. Feel sleepy during day (Y/N)
7. Caffeine consumption (Y/N)
8. Alcohol consumption (Y/N)
9. Smoking (Y/N)
10. Medical issue (Y/N)
11. Ongoing medication (Y/N)
12. Smart device before bed (Y/N)
13. Blue-light filter (Y/N)
14. Discomfort Eye-strain (Y/N)
15. Redness in eye (Y/N)
16. Itchiness/Irritation in eye (Y/N)
17. Dry Eye Disease (Y/N)

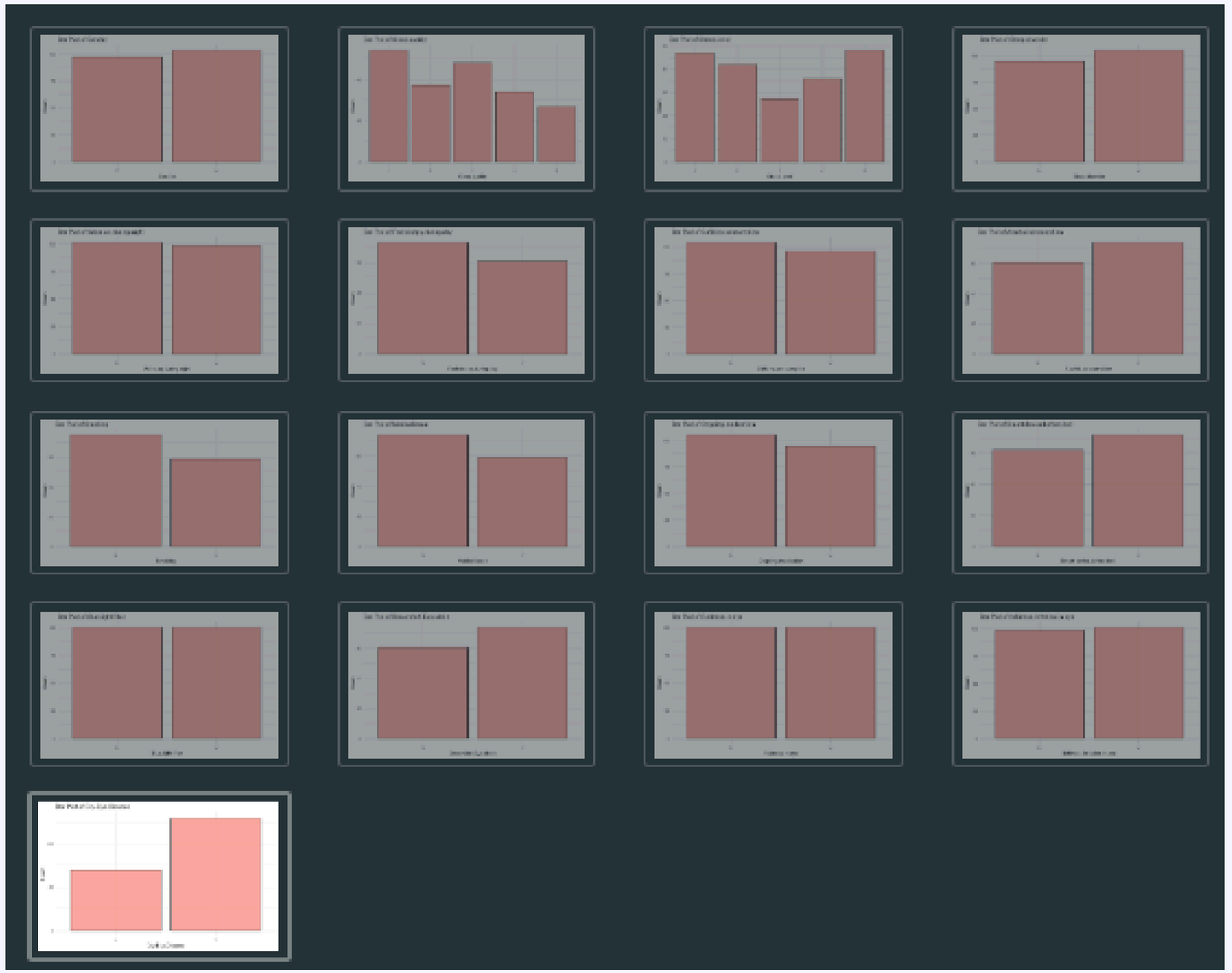
1 Data Description - Possible Bias



Scatter plots appeared to have a uniform / random distribution

No Correlation

1 Data Description - Possible Bias



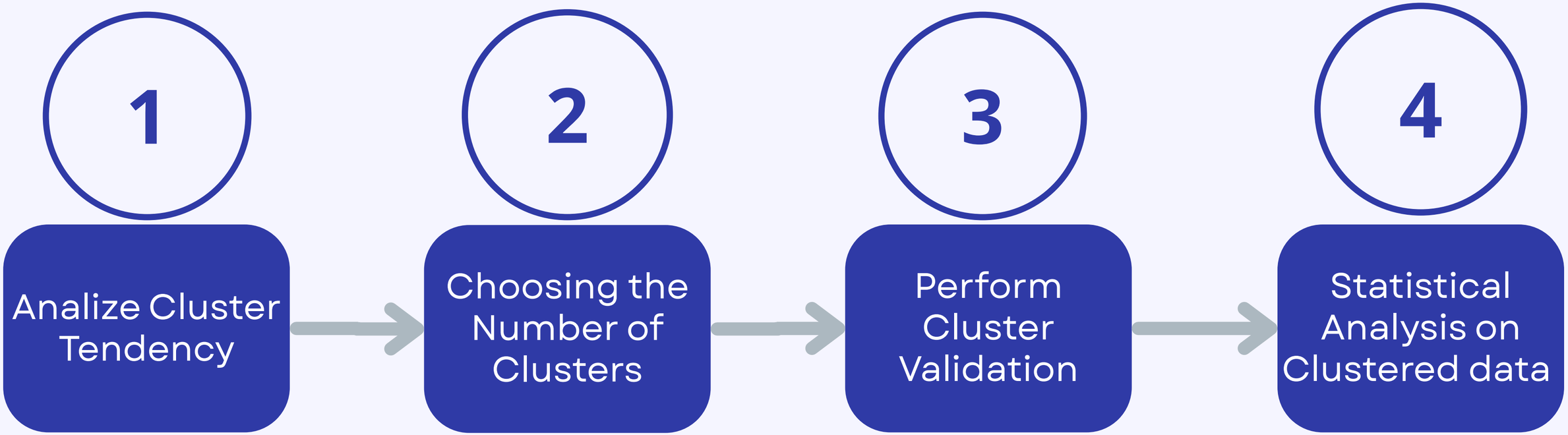
Categorical Variables had similar counts for each class size

2. Objectives of the Study

1. What distinct clusters can be identified based on lifestyle choices, and how do these clusters correlate with Dry Eye Disease outcomes?

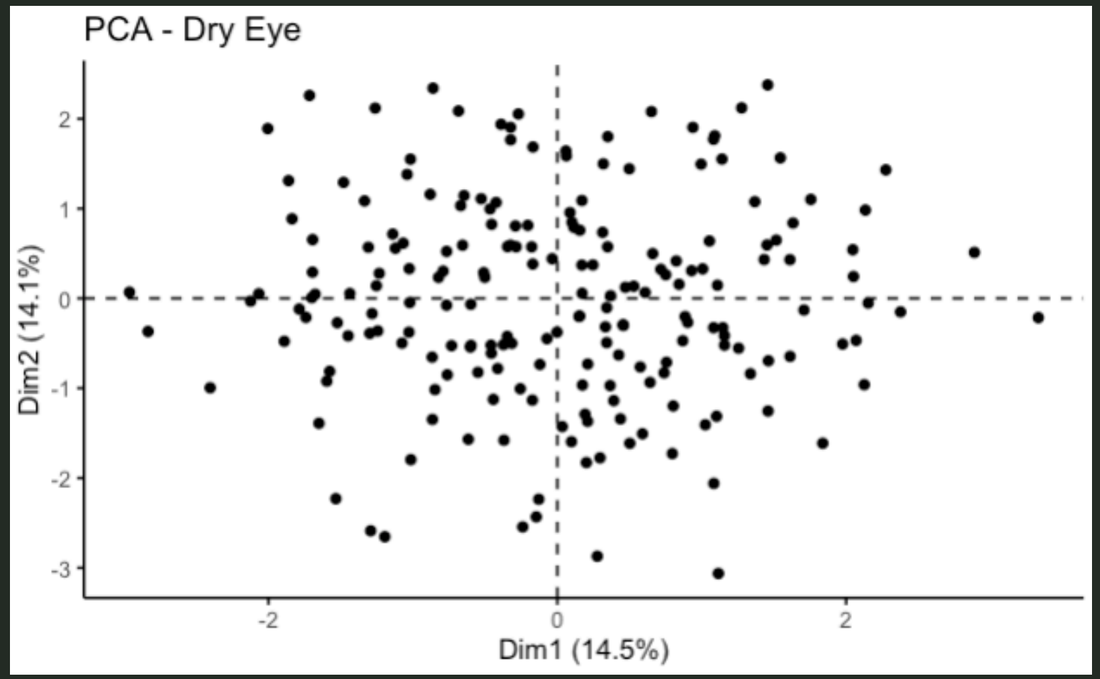
2. Does BMI and physical health contribute to the development of Dry Eye Disease, and can clustering analysis reveal subgroups at higher risk?

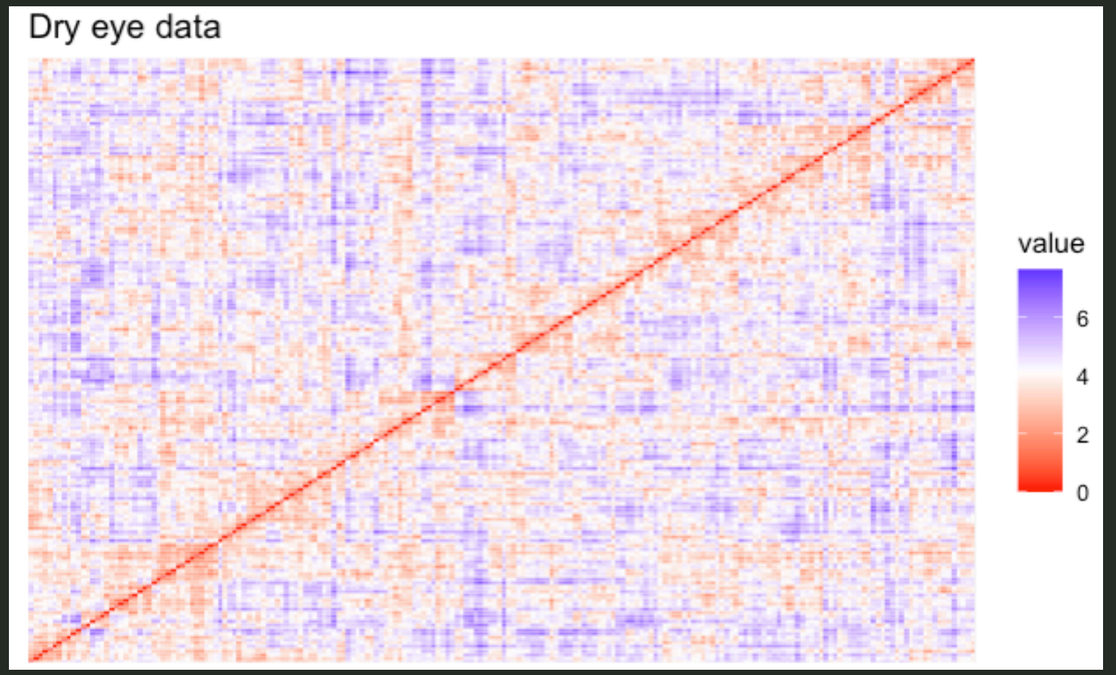
3. Analysis: Project flow



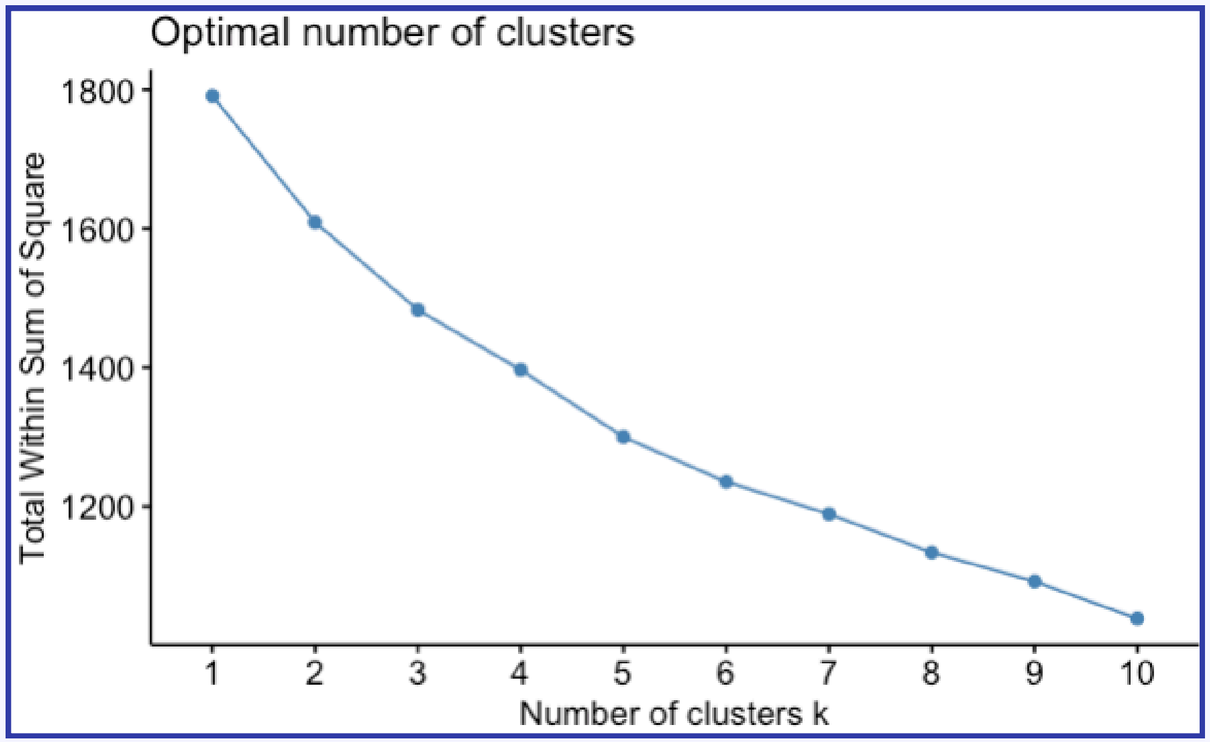
3.1 Cluster Tendency

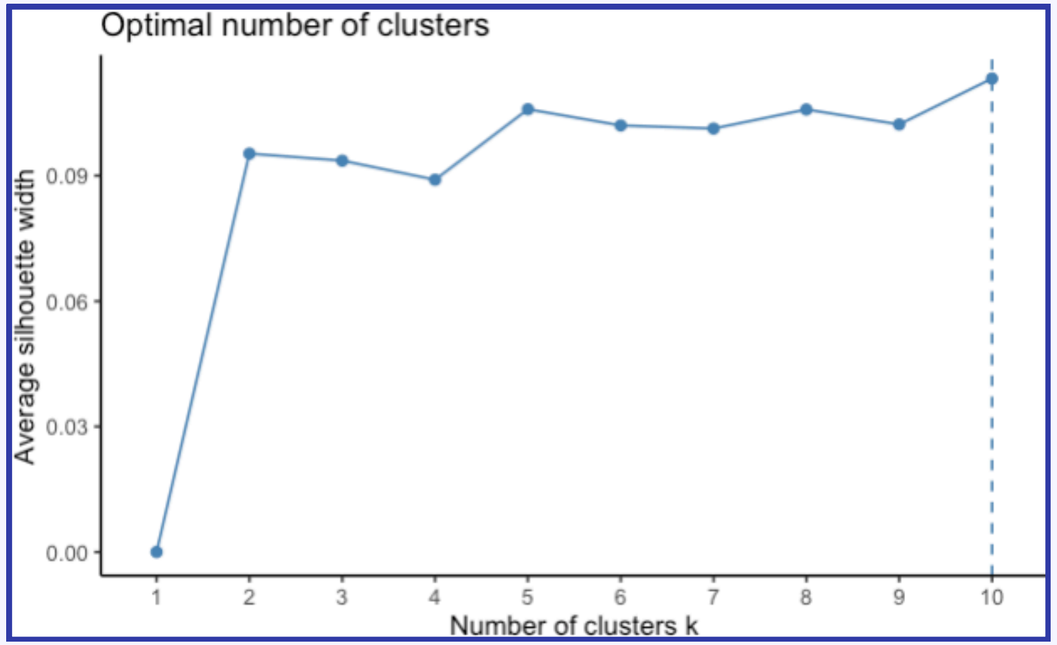
Hopkins Statistic:
0.4982412

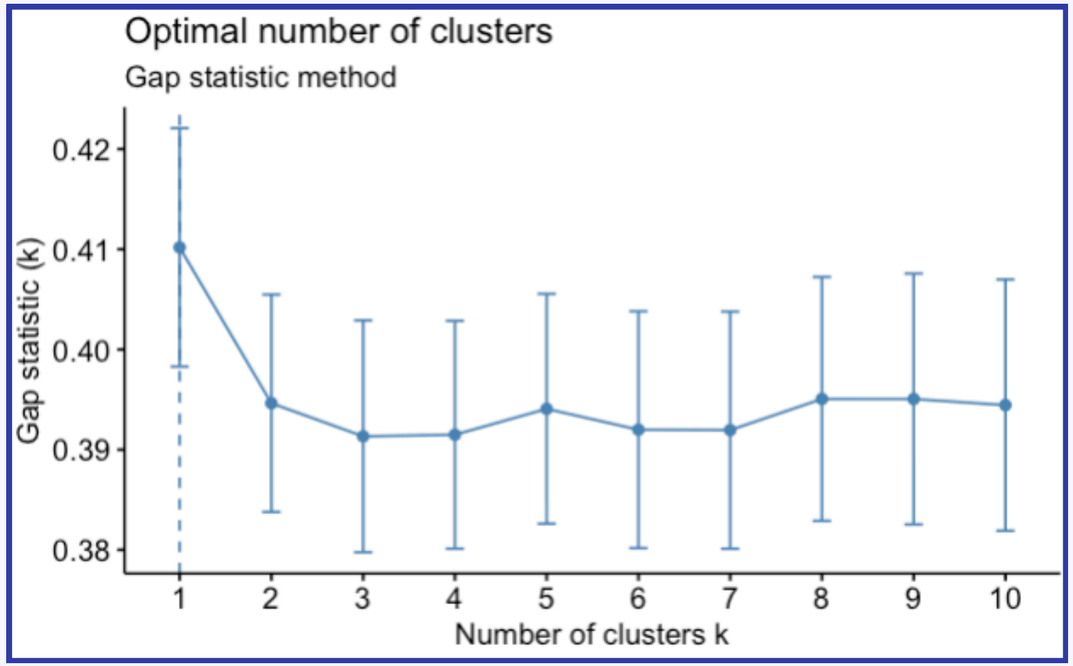




3.2 Choosing Clusters







3.2 Choosing Clusters

K number:
2 and 4 had the best scores.

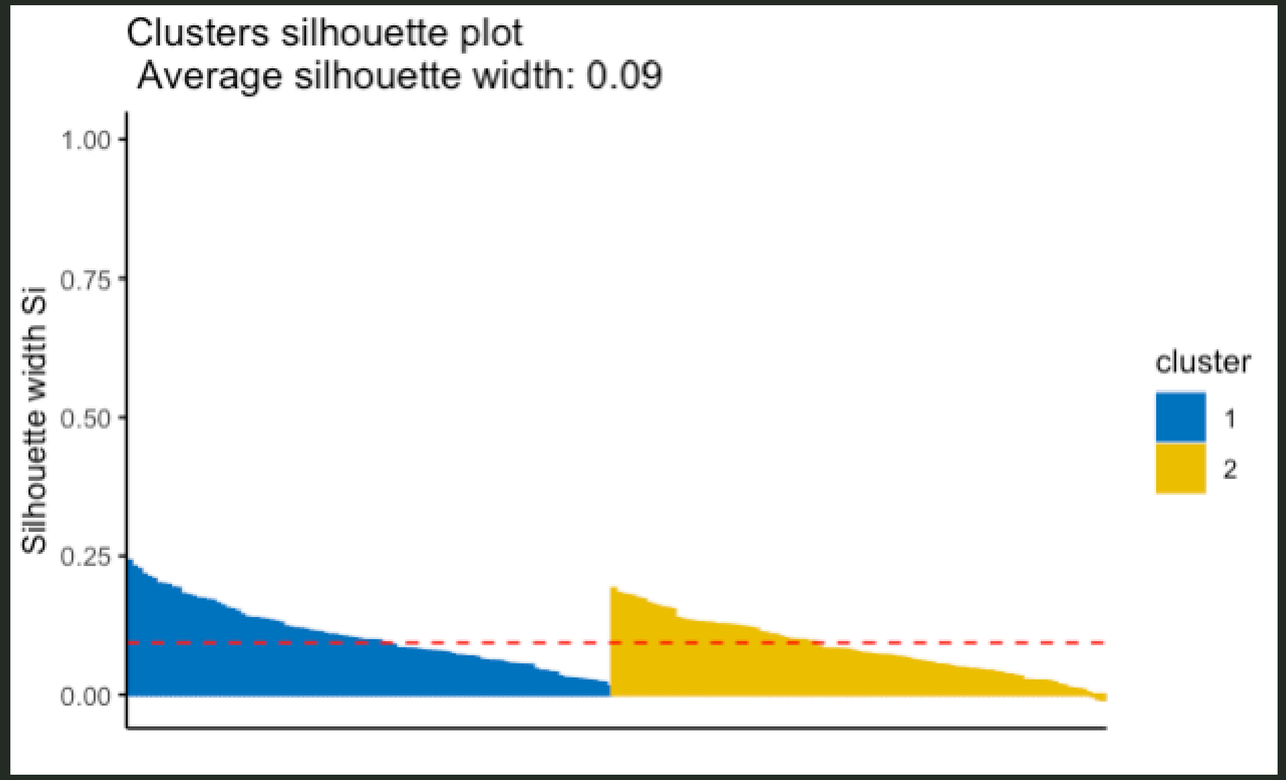
Best Algorithm:
Kmeans, PAM and Hierarchical - Complete Linkage excel in at least one score.

Validation Method		Clustering Technique	2	3	4
Internal	Connectivity	Hierarchical-Comp	141.6738	147.0984	180.2702
		Kmeans	**108.9905**	149.9813	178.9163
		PAM	147.4833	184.8659	223.8028
	Dunn	Hierarchical-Comp	0.2445	0.2511	**0.2568**
		Kmeans	0.2115	0.2235	0.2451
		PAM	0.1769	0.1769	0.1970
	Silhouette	Hierarchical-Comp	0.0422	0.0311	0.0326
		Kmeans	0.0939	0.0884	**0.1023**
		PAM	0.0655	0.0667	0.0644
Stability	APN	Hierarchical-Comp	0.4000	0.5995	0.6452
		Kmeans	0.4270	0.5511	0.5237
		PAM	**0.3235**	0.5192	0.5404
	AD	Hierarchical-Comp	4.1315	4.1167	4.0777
		Kmeans	4.0929	4.0387	**3.9138**
		PAM	4.0695	4.0504	3.9724
	ADM	Hierarchical-Comp	0.8868	1.1681	1.3771
		Kmeans	1.1371	1.4104	1.4000
		PAM	**0.7583**	1.1892	1.2582
	FOM	Hierarchical-Comp	1.0001	1.0008	**0.9987**
		Kmeans	1.0013	1.0004	1.0011
		PAM	1.0004	1.0032	1.0010

3.2 Choosing Clusters

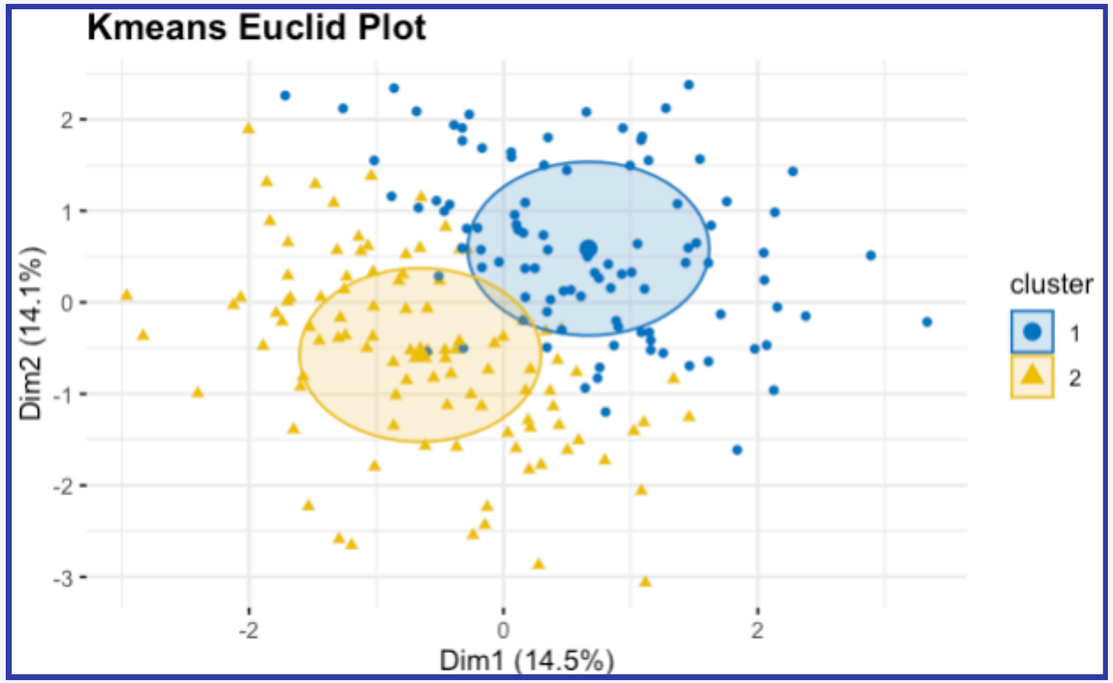
Chosen Algorithm:
K-means

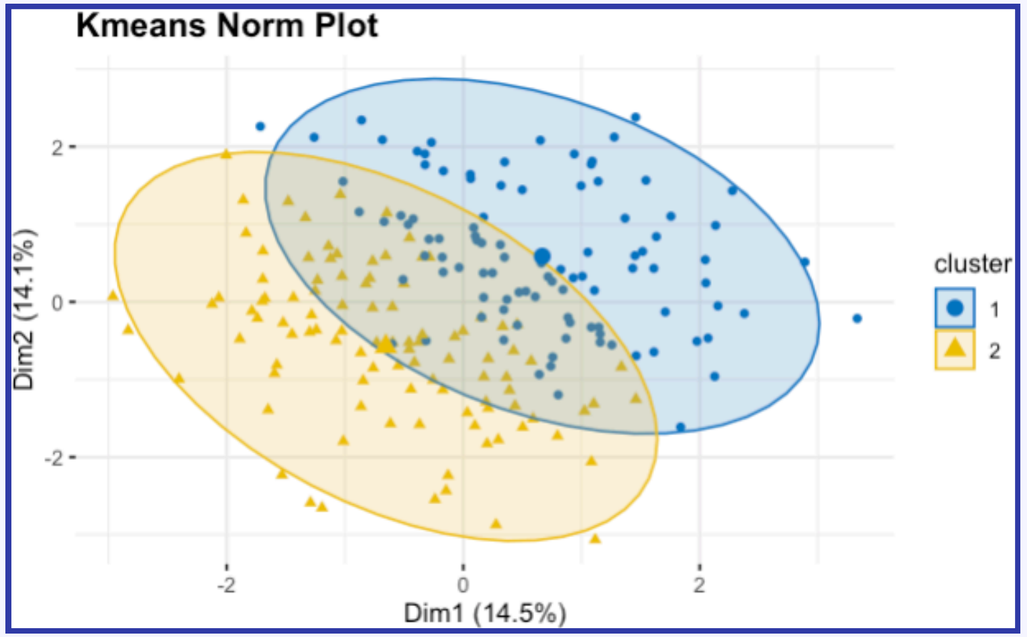
Chosen K Clusters:
2

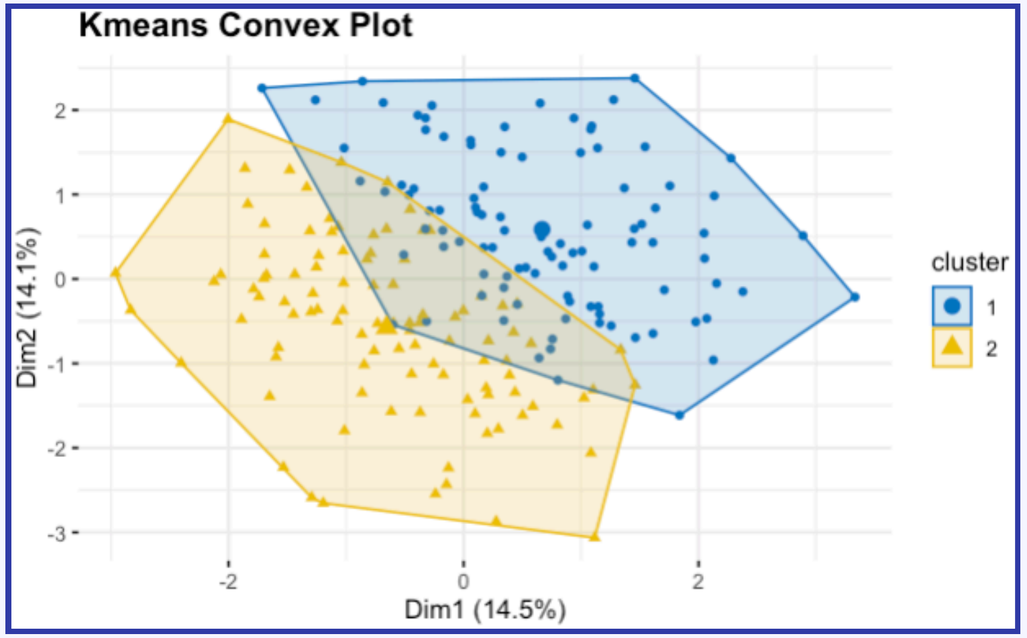


Clustering Technique	Cluster 1 Size	Cluster 2 Size	Cluster 3 Size	Cluster 4 Size	Average Sil Width	Percentage Negative
Kmeans	99	101			**0.09**	**1.50%**
PAM	98	102			0.07	9.50%
Hierarchical-Comp	100	100			0.04	22%
Kmeans	38	48	55	59	0.1	2.50%
PAM	35	56	45	64	0.07	14%
Hierarchical-Comp	61	39	95	5	0.03	35%

3.2 Choosing Clusters





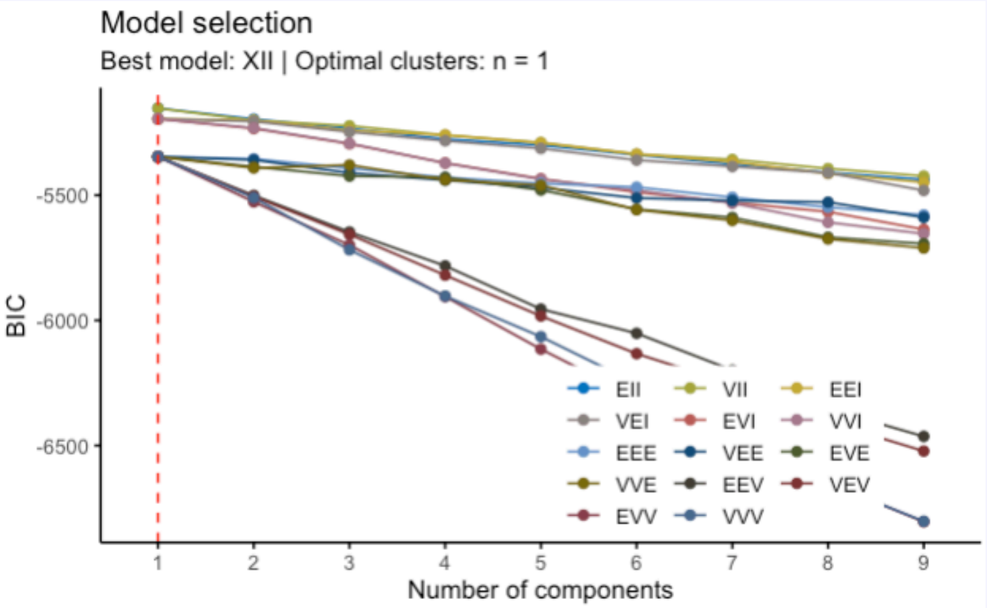


3.3 Cluster Validation

True Label vs Cluster Label analysis showed that the clustering was not better than a random assignment

	No Dry Eye	Yes Dry Eye	No Caffeine	Yes Caffeine	No Alcohol	Yes Alcohol	No Smoking	Yes Smoking	No SD before Bed	Yes SD before Bed	No BL Filter	Yes BL Filter	Female	Male
Cluster 1 (99 obs)	34	65	54	45	43	56	54	45	44	55	48	51	48	51
Cluster 2 (101 Obs)	36	65	50	51	47	54	58	43	49	52	52	49	49	52
Corrected Rand	-0.0045		-0.00253		-0.0041		-0.004074		-0.003418		-0.004145		-0.00505	
Vi	1.340358		1.382899		1.380265		1.378185		1.382127		1.385344		1.385794	

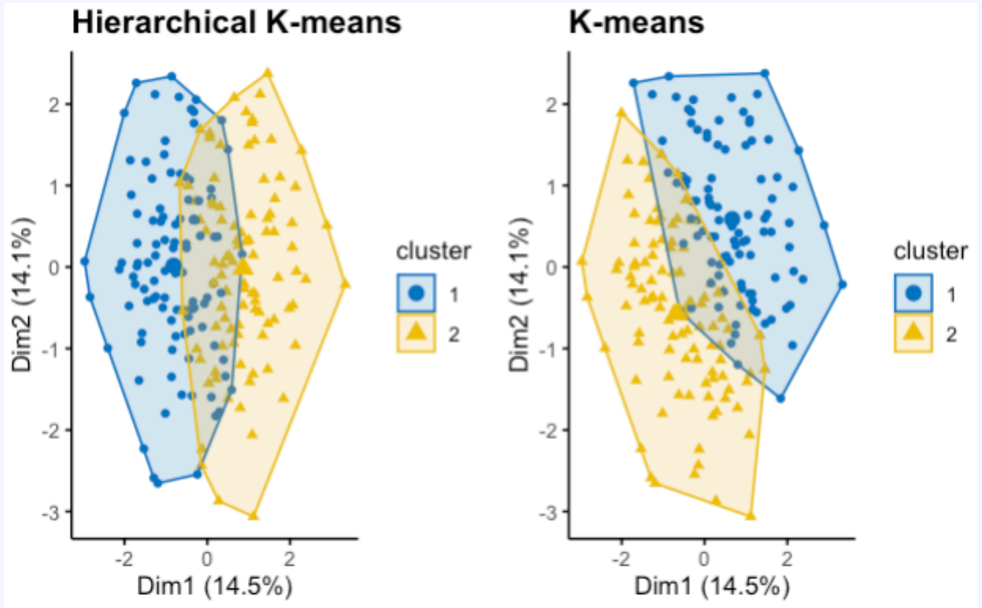
Advanced Clustering

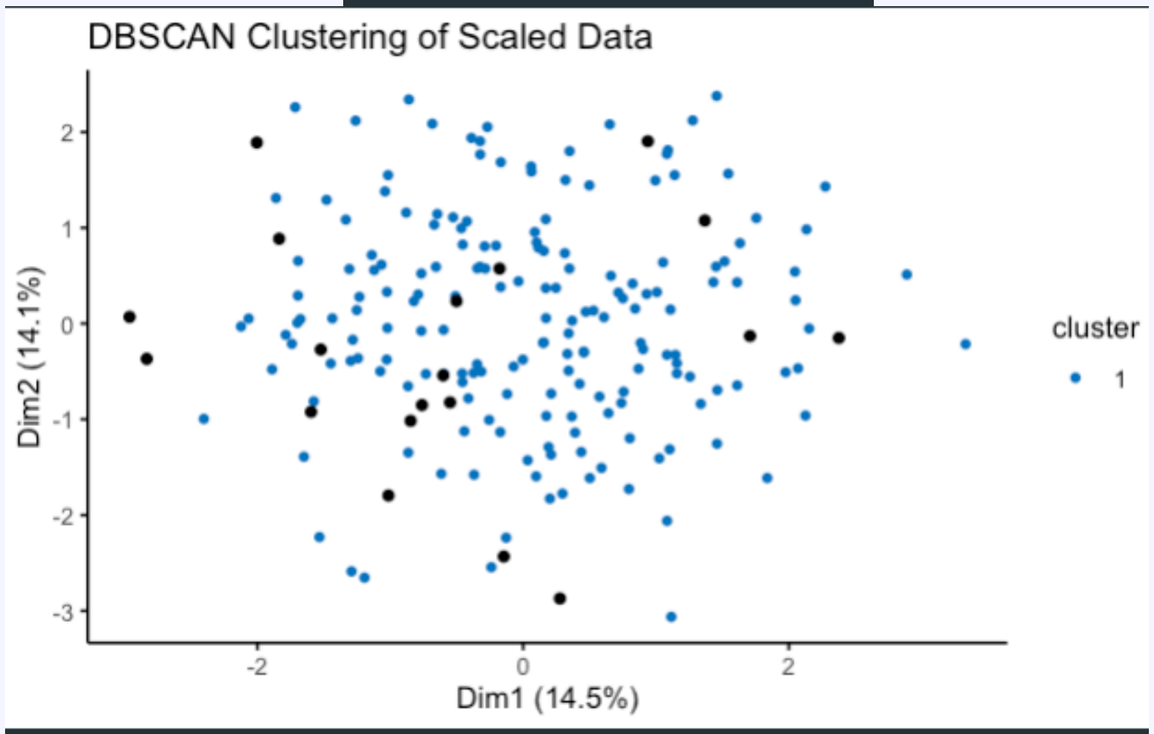


Fuzzy Clustering

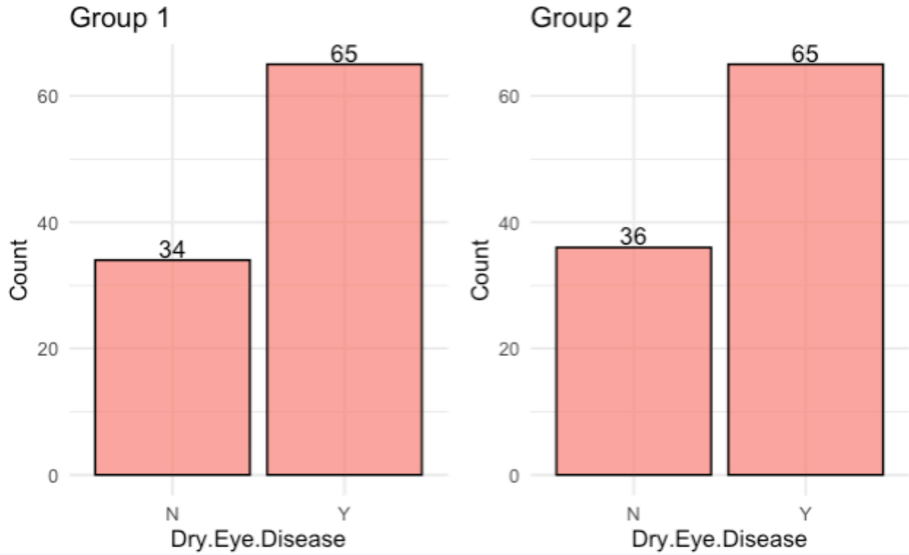
```{r}
fanny.res <- fanny(sample.scale, 2)

head(fanny.res$membership, 10)
```

```
Warning: the memberships are all ver
[,1] [,2]
 [1,]  0.5  0.5
 [2,]  0.5  0.5
 [3,]  0.5  0.5
 [4,]  0.5  0.5
 [5,]  0.5  0.5
 [6,]  0.5  0.5
 [7,]  0.5  0.5
 [8,]  0.5  0.5
 [9,]  0.5  0.5
```

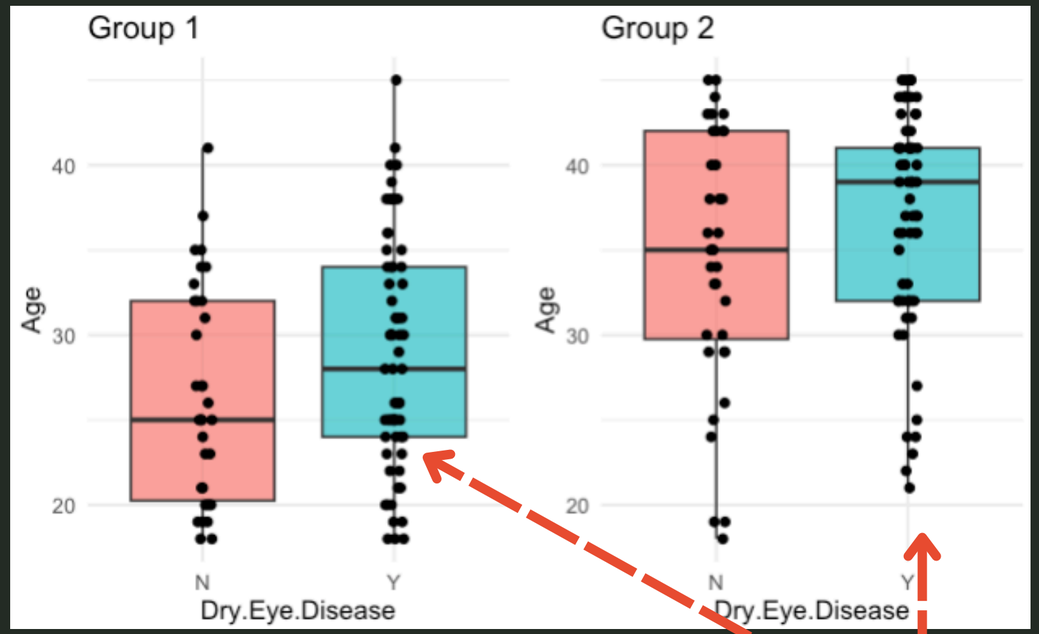




3.4 Statistical Analysis



Similar cluster sizes. There is a statistical difference between the 2 groups.



Variable	Group1 Dry Eye Yes Shapiro Test	Group2 Dry Eye Yes Shapiro Test	Wilcoxon test
Age	Normal	Not Normal	9.18E-10
Diastolic	Not Normal	Not Normal	5.17E-04
Systolic	Not Normal	Not Normal	1.11E-01
Heart Rate	Not Normal	Not Normal	2.27E-07
Physical Activity	Not Normal	Not Normal	4.38E-02
Daily Steps	Not Normal	Not Normal	3.50E-02
Average Screen Time	Not Normal	Not Normal	1.35E-05
Sleep Duration	Normal	Not Normal	1.20E-01

Statistical test reviewed in “A Study of Clustered Data and Approaches to Its Analysis” by Sally Galbraith, James A. Daniel, and Bryce Vissel.

Statistical test reviewed in “Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis by Minlei Liao, Yunfeng Li, Farid Kianifard, Engels Obi and Stephen Arcona

4. Results of the Study

1 What distinct clusters can be identified based on lifestyle choices, and how do these clusters correlate with Dry Eye Disease outcomes?

Averages and Modes by Dry Eye Disease and Cluster

cluster	Dry.Eye.Disease	Sleep.duration - Average	Daily.steps - Average	Physical.activity - Average	Average.screen.time - Average	Caffeine.consumption - Mode	Alcohol.consumption	Smoking - Mode	Smart.device.before.bed - M	Blue.light.filter - Mode
1	N	7.13	9500.00	79.68	7.01	Y	N	N	Y	N
2	N	6.89	10250.00	99.22	5.22	N	Y	N	Y	N
1	Y	7.30	9615.38	76.94	6.83	N	Y	N	Y	Y
2	Y	6.83	11938.46	96.08	4.87	Y	Y	N	N	Y
Influence?		No	No	No	No	No	Might	No	No	Might

People who drink alcohol and use Blue light filter might be at higher risk of suffering a positive diagnostic of Dry Eye Disease.

4. Results of the Study

2 Does BMI and physical health contribute to the development of Dry Eye Disease, and can clustering analysis reveal subgroups at higher risk?

cluster	Dry.Eye.Disease	Gender - Mode	Age - Average	Sleep.duration - Average	Systolic - Average	Diastolic - Average	Heart.rate - Average	BMI - Average	Sleep.quality - Mode	Stress.level - Mode	Sleep.disorder - Mode	FeeL.sleepy.during.day - M	MedicaLissue - Mode	Ongoing.medication - M	Discomfort.Eye.strain - M	Redness.in.eye - Mode	Itchiness.Irritation.in.eye - M
1	N	M	26.41	7.13	118.82	78.79	88.15	23.29	1	5	N	N	Y	N	N	Y	N
2	N	M	34.61	6.89	118.42	72.72	73.83	27.82	1	2	Y	Y	N	N	N	N	N
1	Y	F	28.80	7.30	116.55	78.12	86.18	23.11	1	5	Y	N	N	N	Y	N	Y
2	Y	F	37.00	6.83	112.34	72.65	74.94	25.82	1	1	Y	Y	Y	N	Y	Y	Y
Influence?		Might	No	No	No	No	No	No	No	No	Might	No	No	No	Might	No	Might

Females who have been diagnosed with sleep disorder, eye strain discomfort and eye irritation might be at higher risk of having a positive diagnose of Dry Eye Disease.

5. Conclusions

1. The Hopkins statistic revealed low clustering tendency, and the Rand and VI indexes also revealed values that suggests a random clustering assignment for this data set.

2. Neither of the advanced clustering algorithms learned in class provided a better result than the ones explored in this study: kmeans, pam, and hierarchical clustering.

3. The wilcoxon test revealed that there was a statistical difference between the clusters numerical features, nevertheless the classes for the categorical features appeared to have similar distributions for both clusters.

6. Next Steps

1. Run supervised ML algorithms in the data grouped in each cluster and compare the results obtained.

2. Run the analysis used in this report in a bigger sample of the data set.

3. Find new sources of data related to dry eye disease and compare results.

Thank you!

Any Questions?