Snooze Clustering

Billie Kim, Casey Ng, Ethan Panal

Problem

Stakeholders: Medical Providers & Sleep Coaches

Problem: Providers may only be looking at sleep quality from physical symptom perspective

Solution: Patient segmentation enables for personalized treatment based individual's life circumstances

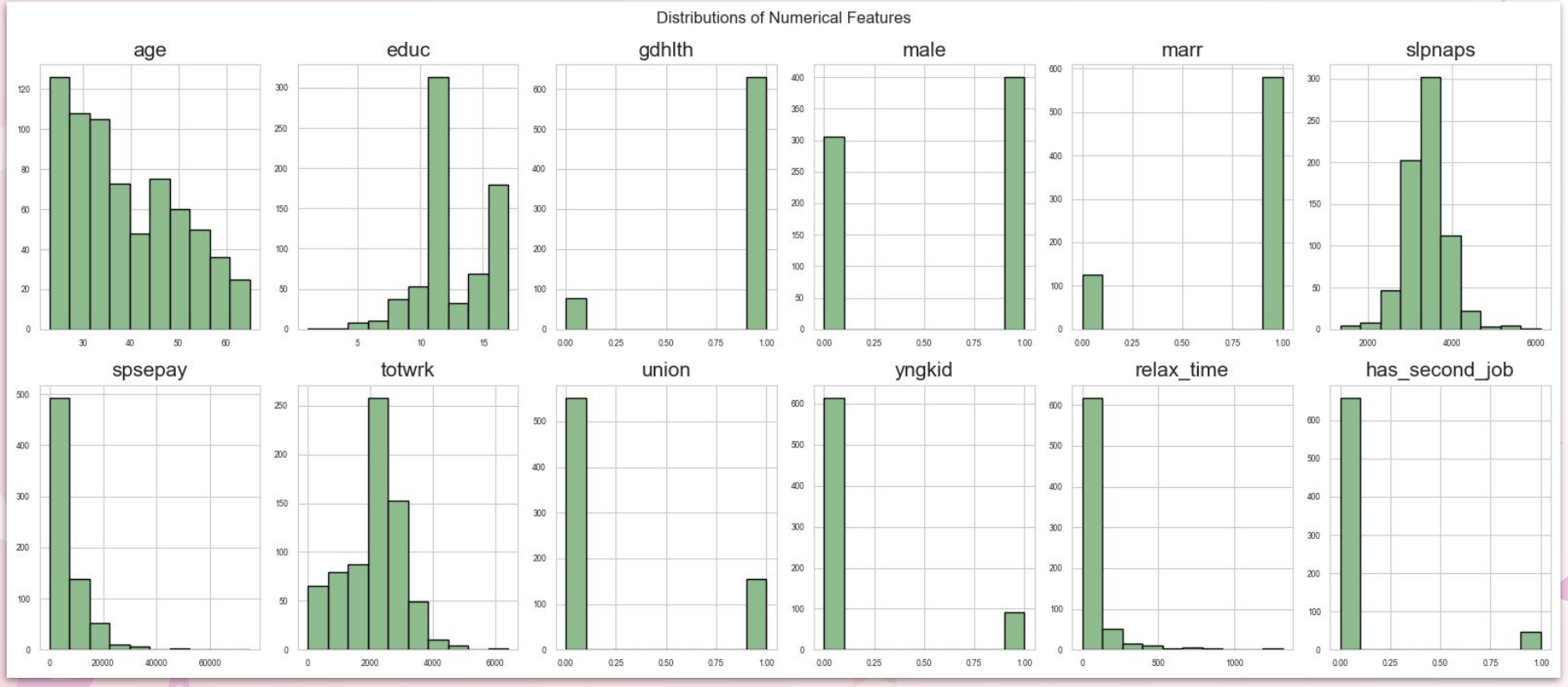
Dataset

	age	educ	gdhlth	male	marr	slpnaps	spsepay	totwrk	union	yngkid	relax_time	has_second_job
0	32	12	0	1	1	3163	0	3438	0	0	0	0
1	31	14	1	1	0	2920	0	5020	0	0	0	0
2	44	17	1	1	1	2760	20000	2815	0	0	278	0
3	30	12	1	0	1	3083	5000	3786	0	0	0	0
4	64	14	1	1	1	3493	2400	2580	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
701	45	12	1	0	1	3385	16000	2026	0	0	25	0
702	34	10	0	1	1	3535	0	675	1	0	0	1
703	37	12	1	0	1	3510	12000	1851	0	0	135	0
704	54	17	1	0	1	3000	35000	1961	1	0	88	1
705	30	16	1	0	1	3415	0	2363	0	1	0	0

706 rows × 12 columns

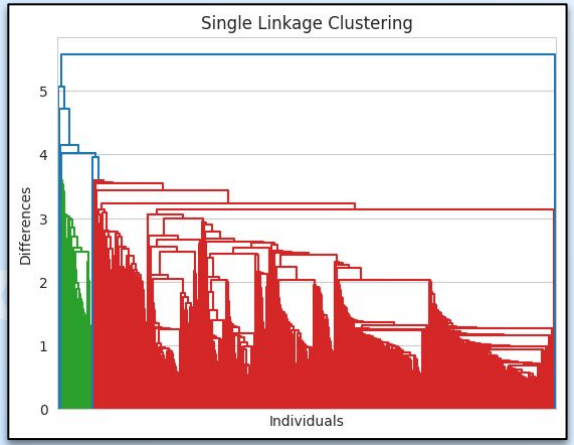
- Combination of demographic data, economic circumstances, and sleep metrics
- Feature selection - Avoid highly correlated variables

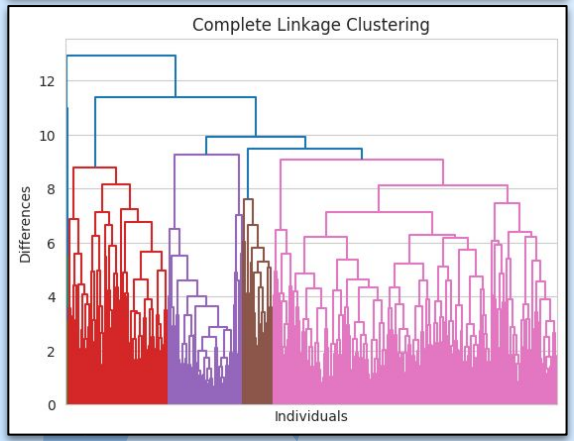
Exploratory Data Analysis



Cluster Models

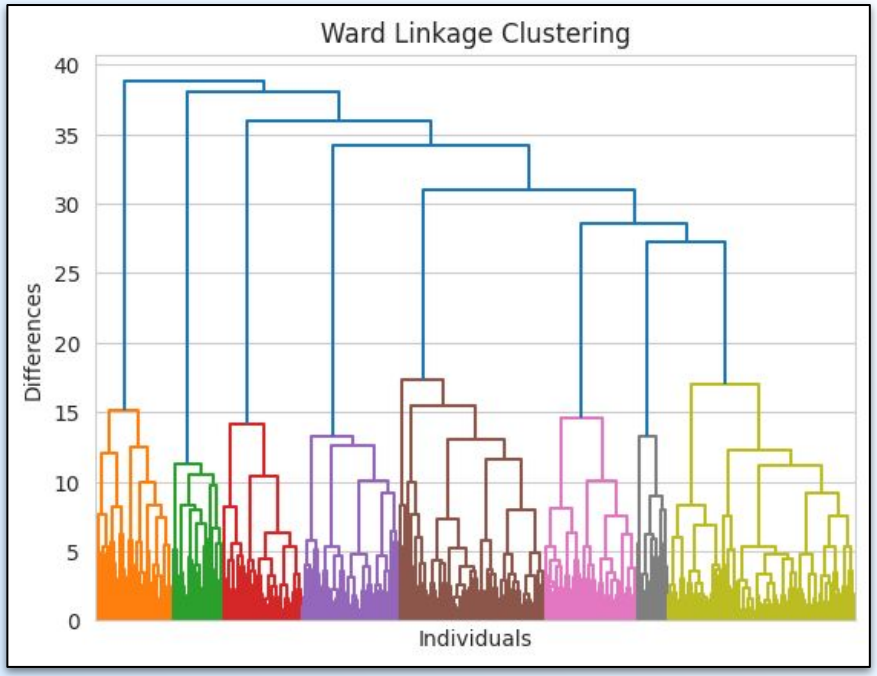
Using Hierarchical:





BEST LINKAGE METHOD: WARD

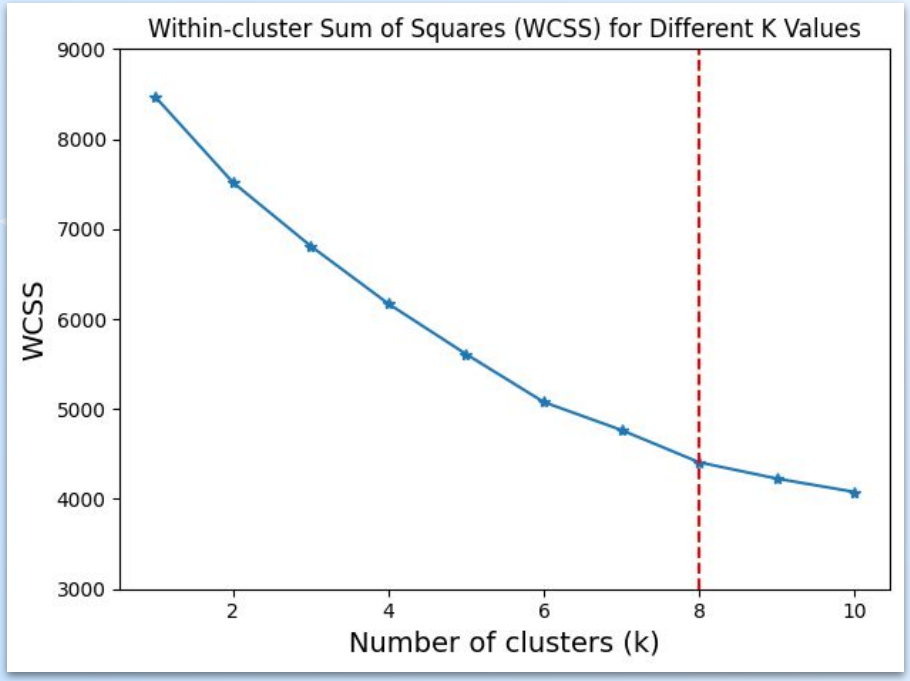
- 8 distinct balanced clusters



Cluster Models Cont.

Using K-Means:

- Sum-of-Squares Error/Elbow-Method
- n_clusters = 8

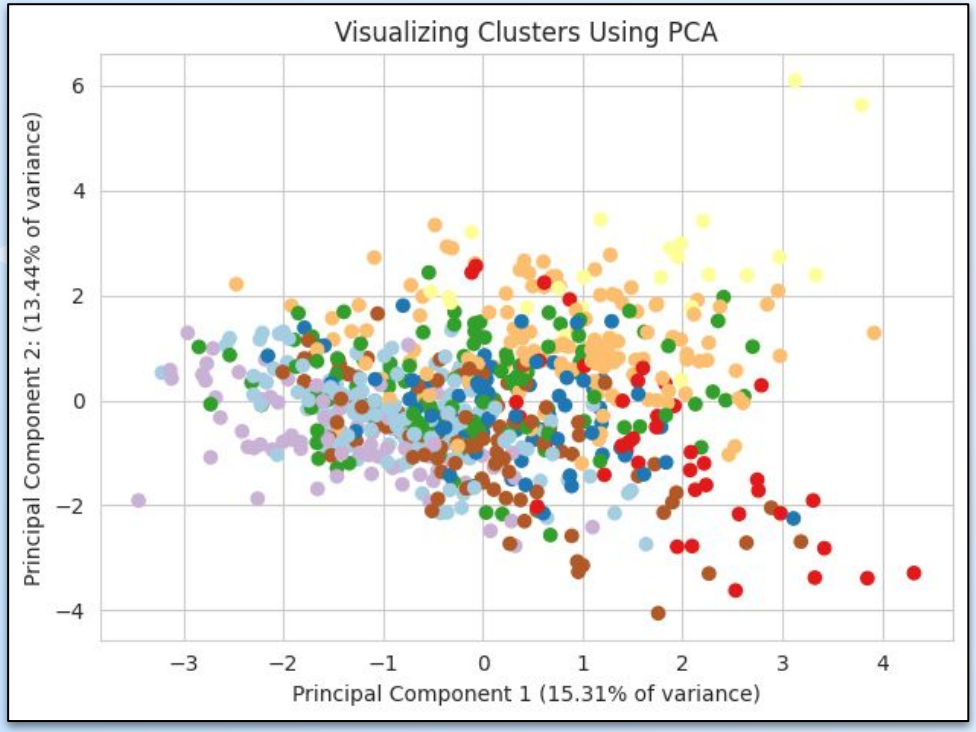



```
In [26]: # initializing a kmeans clustering model with 8 clusters
         kmModel = KMeans(n_clusters=8, random_state=143)
         # fitting the model
         kmModel.fit(X)
         # getting cluster labels for the data
         clusters = kmModel.fit_predict(X)
         # counts for each cluster
         pd.Series(clusters).value_counts()

Out[26]: 5    177
         1    130
         6     92
         3     88
         7     78
         0     70
         4     47
         2     24
         dtype: int64
```

CLUSTER LABELS

Insights into Clusters



High

High

Low

```
# summary statistics for cluster 1
clust_dfs[1].mean()
```

```
age                38.092308
educ               13.123077
gdhlth              1.000000
male                0.023077
marr                1.000000
slpnaps          3454.346154
spsepay         13239.615385
totwrk           1564.738462
union               0.000000
yngkid              0.000000
relax_time         44.500000
has_second_job      0.000000
clusters            1.000000
dtype: float64
```

```
# summary statistics for cluster 2
clust_dfs[2].mean()
```

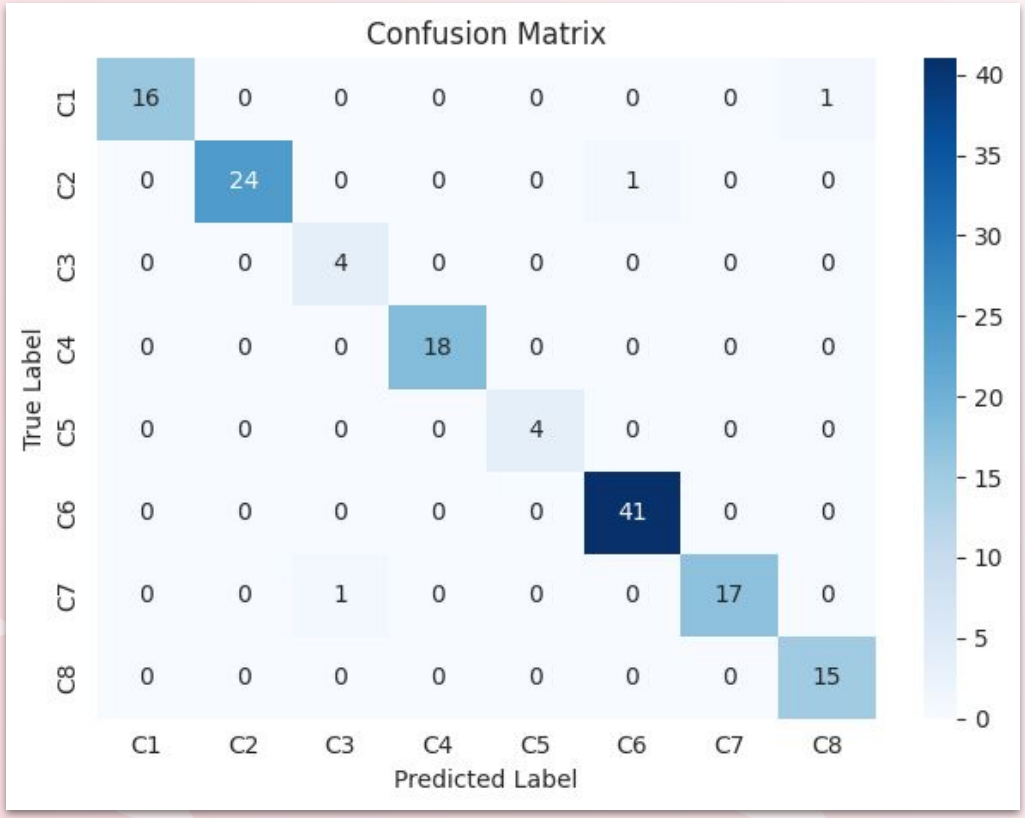
```
age                40.541667
educ               12.625000
gdhlth              0.958333
male                0.583333
marr                0.875000
slpnaps          3519.708333
spsepay          5377.625000
totwrk           1761.833333
union               0.125000
yngkid              0.041667
relax_time        602.458333
has_second_job      0.000000
clusters            2.000000
dtype: float64
```

```
# summary statistics for cluster 7
clust_dfs[7].mean()
```

```
age                29.192308
educ               13.730769
gdhlth              0.961538
male                0.666667
marr                0.948718
slpnaps          3338.602564
spsepay          3640.384615
totwrk           2192.012821
union               0.141026
yngkid              1.000000
relax_time         28.961538
has_second_job      0.000000
clusters            7.000000
dtype: float64
```

KNN Classifier Results

```
Accuracy Score: 0.9788732394366197
```



Next Steps...

- Hyper parameter tuning for K
- More, more, & more... data
- Further analysis into clusters