Lecture 3:
K-Means Clustering

(because hierarchical clustering is slow)

Dr Matt Wheeler

K-Means

Hierarchical clustering assumes that we know the distance between all observations

- For 1000 observations this is ok.
- For 100,000,000 observations this is time consuming and memory intensive!
  - There are $0.5*100{,}000{,}000^2$ distances we need to compute!
  - That equates to approximately 40,000 terabytes of storage!!!
- For K-means we say we want K clusters and try to find the K clusters that minimizes this distance.
  - K*100,000,000 distances << $0.5*100{,}000{,}000^2$ distances
  - K*11 MB << 40,000 TB
- hclust() will not run when there are more than 65k observations.

Algorithm:

CLUST <- randomly pick K cluster centroids.

Until Convergence

    For Each observation i

        Find closest centroid k in CLUST for observation i

        Label i with cluster k

    Loop

    CLUST <- recompute cluster centroids

Loop

Idea: We want to save time so we assume there are K clusters. We start off by throwing out K random cluster centers (i.e. the average d(x,y) for cluster k) then we find the data closest to each cluster, recompute the centroids, and repeat until the centroid center's don't change that much. EASY PEASY.

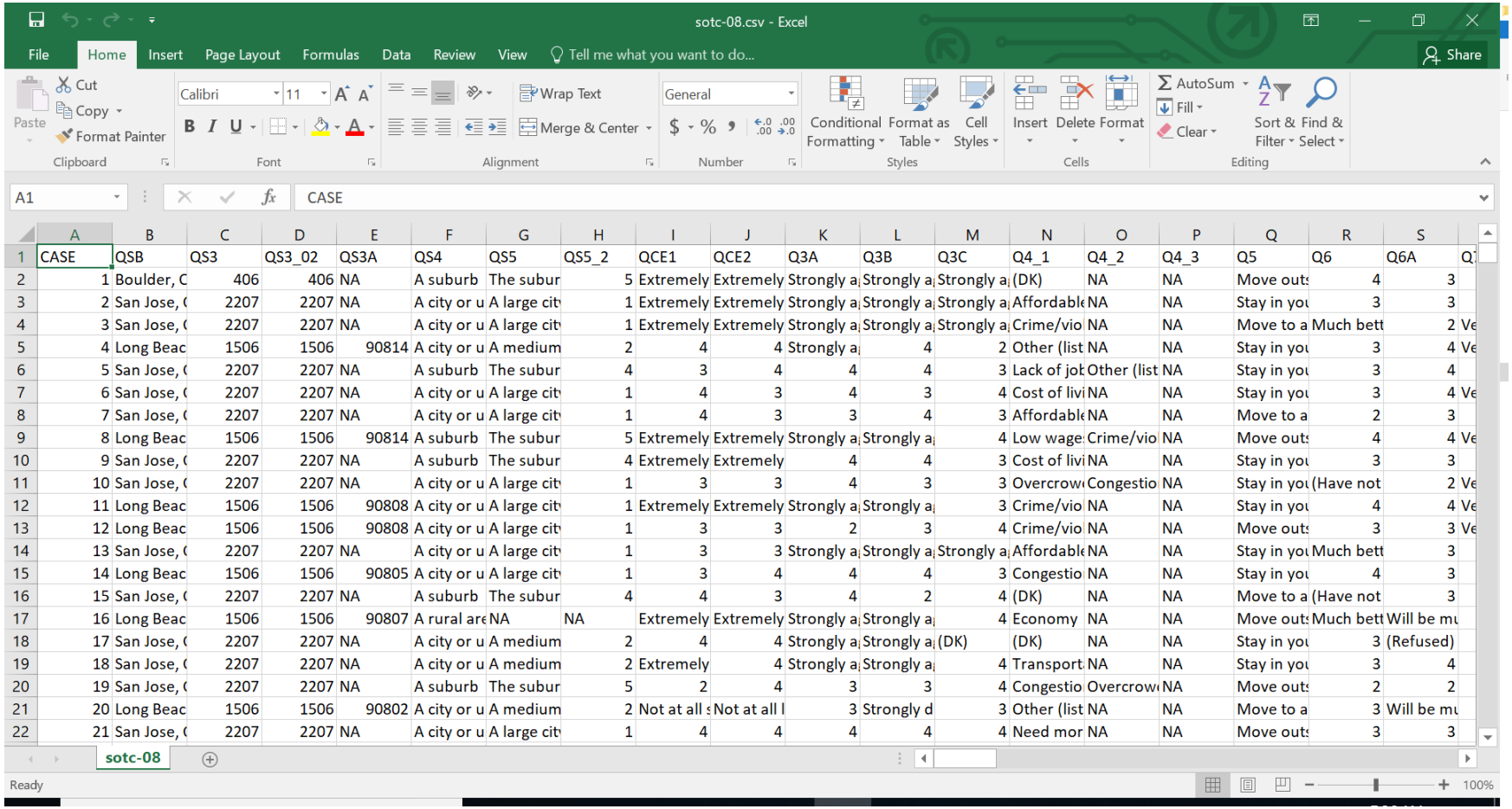
Our Data

- This time we are going to look at real data. This comes from the "**Knight Foundation Soul of the Community survey.**"

- This survey looks at a variety of metropolitan areas and attempts looking at community involvement.

- Let's start looking at this assuming a Euclidean distance metric.

- The Data
  - Based upon survey responses.
    - 1 = Low
    - 7 = High
  - Also have demographic characteristics (Race/Education/Income/Sex etc.)
  - Messy:
    - It is not coded the same for every respondent.
    - Also is a weighted study. We are going to ignore that for the time being.



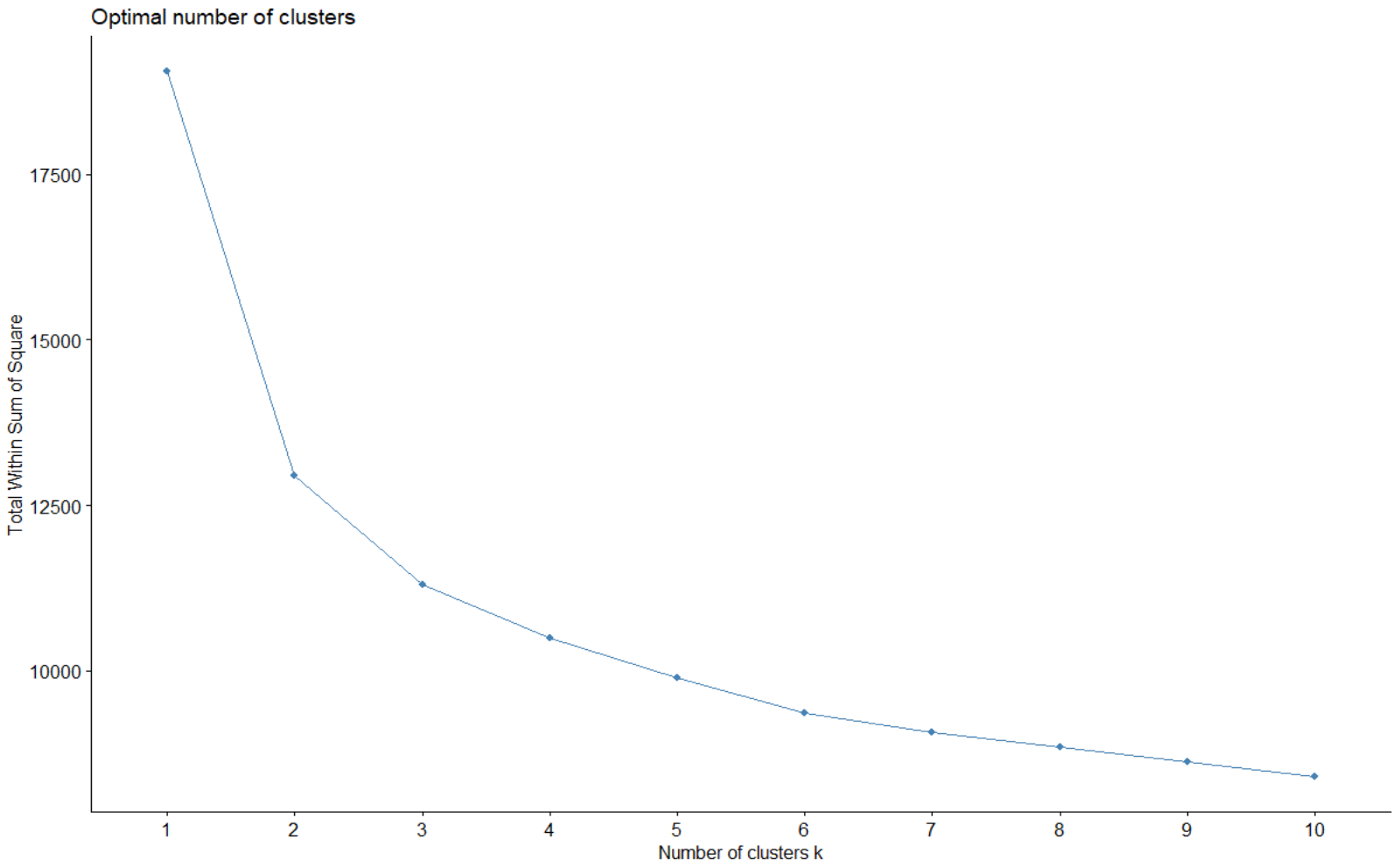
sotc-08.csv - Excel

CASE	QSB	QS3	QS3_02	QS3A	QS4	QS5	QS5_2	QCE1	QCE2	Q3A	Q3B	Q3C	Q4_1	Q4_2	Q4_3	Q5	Q6	Q6A	Q
1	Boulder, C	406	406	NA	A suburb	The subur	5	Extremely	Extremely	Strongly a	Strongly a	Strongly a	(DK)	NA	NA	Move outs	4	3	
2	San Jose, (	2207	2207	NA	A city or u	A large cit	1	Extremely	Extremely	Strongly a	Strongly a	Strongly a	Affordable	NA	NA	Stay in you	3	3	
3	San Jose, (	2207	2207	NA	A city or u	A large cit	1	Extremely	Extremely	Strongly a	Strongly a	Strongly a	Crime/vio	NA	NA	Move to a	Much bett	2	Ve
4	Long Beac	1506	1506	90814	A city or u	A medium	2	4	4	Strongly a	4	2	Other (list	NA	NA	Stay in you	3	4	Ve
5	San Jose, (	2207	2207	NA	A suburb	The subur	4	3	4	4	4	3	Lack of jot	Other (list	NA	Stay in you	3	4	
6	San Jose, (	2207	2207	NA	A city or u	A large cit	1	4	3	4	3	4	Cost of livi	NA	NA	Stay in you	3	4	Ve
7	San Jose, (	2207	2207	NA	A city or u	A large cit	1	4	3	3	4	3	Affordable	NA	NA	Move to a	2	3	
8	Long Beac	1506	1506	90814	A suburb	The subur	5	Extremely	Extremely	Strongly a	Strongly a	4	Low wage	Crime/vio	NA	Move outs	4	4	Ve
9	San Jose, (	2207	2207	NA	A suburb	The subur	4	Extremely	Extremely	4	4	3	Cost of livi	NA	NA	Stay in you	3	3	
10	San Jose, (	2207	2207	NA	A city or u	A large cit	1	3	3	4	3	3	Overcrow	Congestio	NA	Stay in you	(Have not	2	Ve
11	Long Beac	1506	1506	90808	A city or u	A large cit	1	Extremely	Extremely	Strongly a	Strongly a	3	Crime/vio	NA	NA	Stay in you	4	4	Ve
12	Long Beac	1506	1506	90808	A city or u	A large cit	1	3	3	2	3	4	Crime/vio	NA	NA	Move outs	3	3	Ve
13	San Jose, (	2207	2207	NA	A city or u	A large cit	1	3	3	Strongly a	Strongly a	Strongly a	Affordable	NA	NA	Stay in you	Much bett	3	
14	Long Beac	1506	1506	90805	A city or u	A large cit	1	3	4	4	4	3	Congestio	NA	NA	Stay in you	4	3	
15	San Jose, (	2207	2207	NA	A suburb	The subur	4	4	3	4	2	4	(DK)	NA	NA	Move to a	(Have not	3	
16	Long Beac	1506	1506	90807	A rural are	NA	NA	Extremely	Extremely	Strongly a	Strongly a	4	Economy	NA	NA	Move outs	Much bett	Will be mu	
17	San Jose, (	2207	2207	NA	A city or u	A medium	2	4	4	Strongly a	Strongly a	(DK)	(DK)	NA	NA	Stay in you	3	(Refused)	
18	San Jose, (	2207	2207	NA	A city or u	A medium	2	Extremely	4	Strongly a	Strongly a	4	Transport	NA	NA	Stay in you	3	4	
19	San Jose, (	2207	2207	NA	A suburb	The subur	5	2	4	3	3	4	Congestio	Overcrow	NA	Move outs	2	2	
20	Long Beac	1506	1506	90802	A city or u	A medium	2	Not at all s	Not at all l	3	Strongly d	3	Other (list	NA	NA	Move to a	3	Will be mu	
21	San Jose, (	2207	2207	NA	A city or u	A large cit	1	4	4	4	4	4	Need mor	NA	NA	Move outs	3	3	

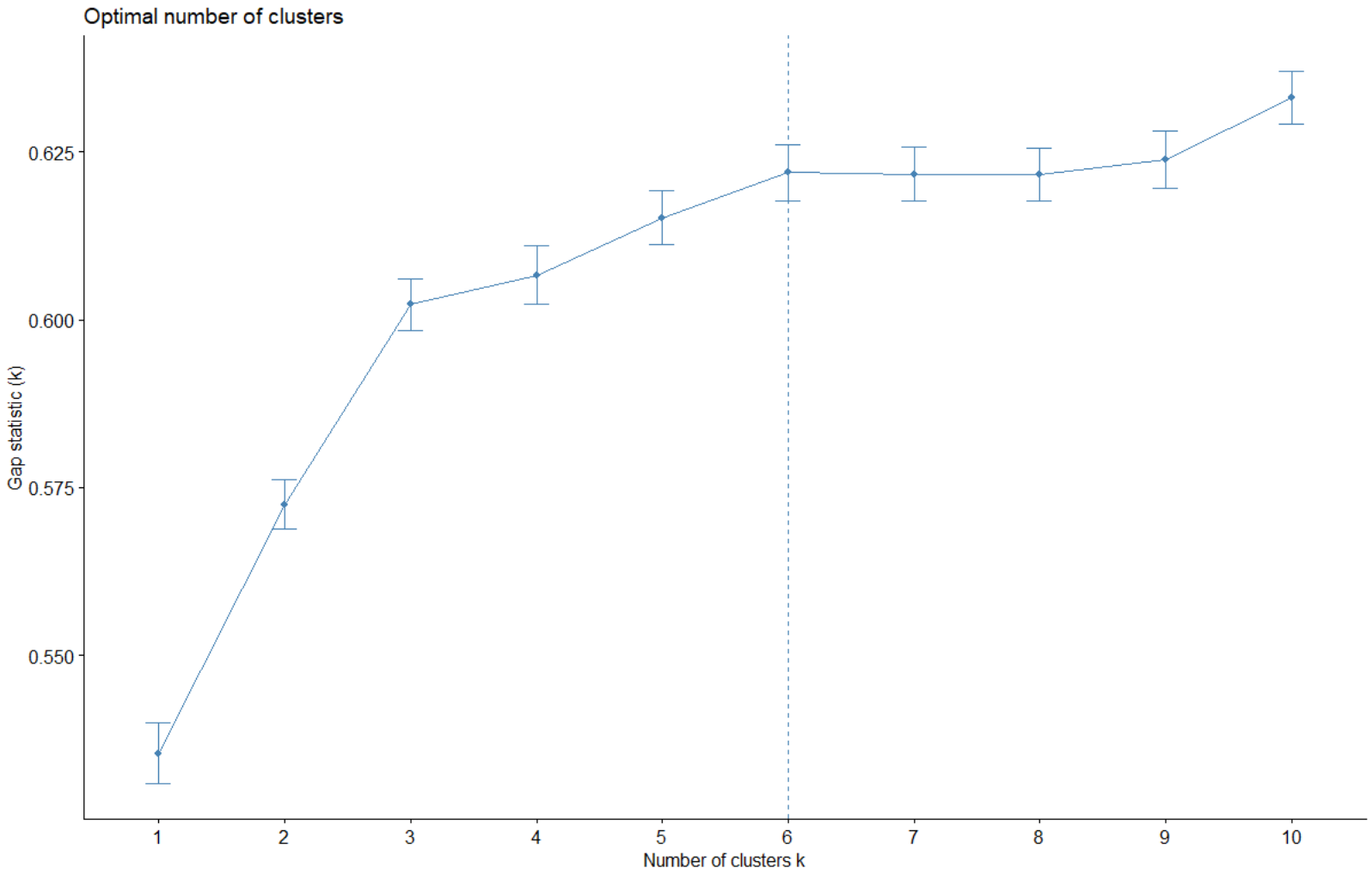
First Find the “optimal” cluster size

```
df <- scale(df[,-15]) #do this to make everthing mean = 0, sd = 1

reduced.df <- sample_frac(as.data.frame(df),size = 0.10) # randomly sample 10% of the data
                                                         # full dataset is too large

fviz_nbclust(reduced.df, kmeans, method = "wss")
fviz_nbclust(reduced.df, kmeans, method = "gap")
fviz_nbclust(reduced.df, kmeans, method = "silhouette")
```

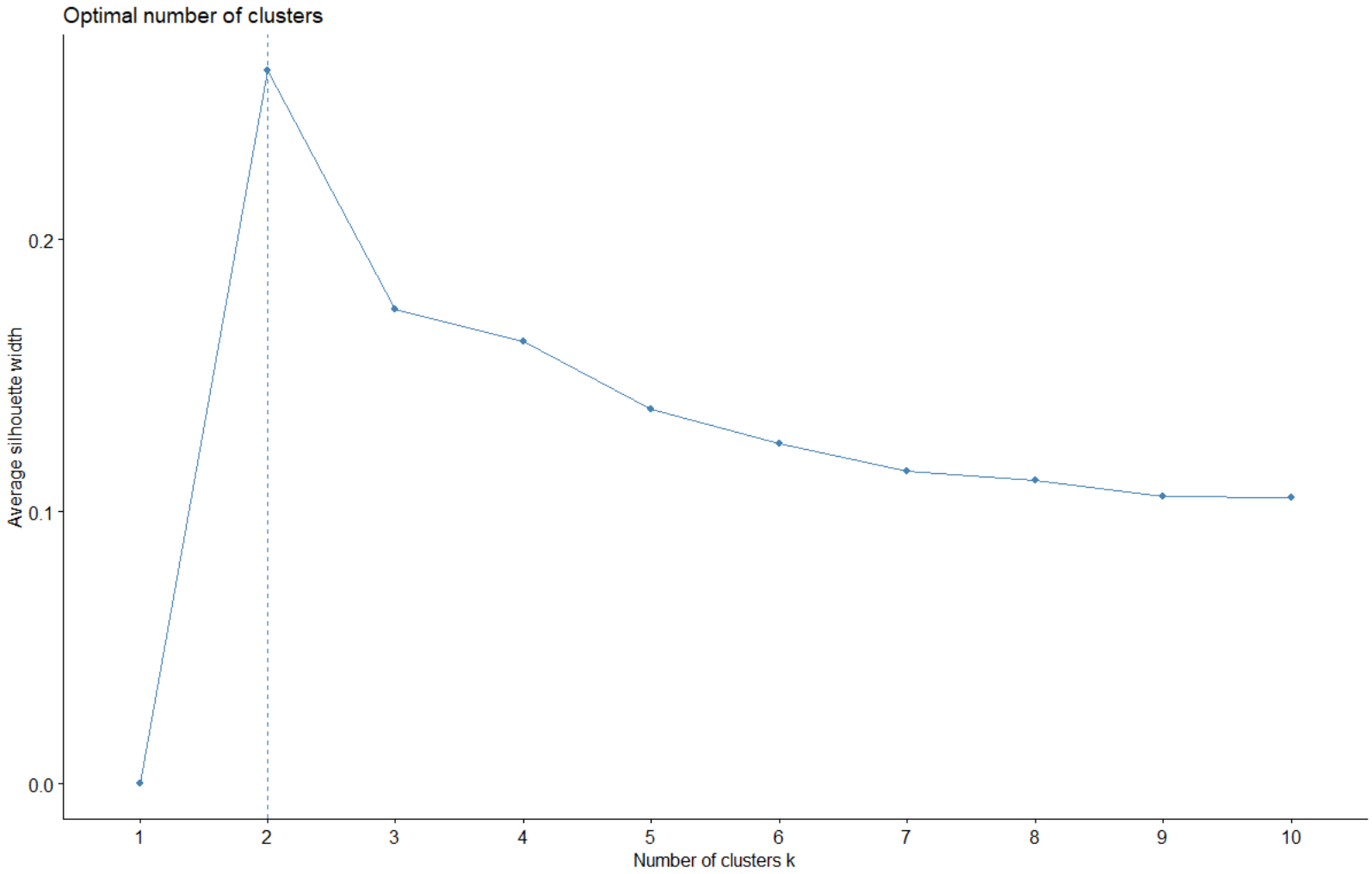
WSS


Optimal number of clusters
Total Within Sum of Square
17500
15000
12500
10000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

GAP


Optimal number of clusters
Gap statistic (k)
0.625
0.600
0.575
0.550
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Silhouette


Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
1
2
3
4
5
6
7
8
9
10
Number of clusters k

K-Means in R

```
kmeans(x, centers, iter.max = 10,
       nstart = 1,
       algorithm =
       c("Hartigan-Wong", "Lloyd",
         "Forgy", "MacQueen"),
       trace=FALSE)
```

Number of random centers

Data

I have clusters, what do I do with them.

Chances are your boss isn't going to ask you to "find" some clusters and be done. Segmentation is about trying to understand the differences, so you can tell your boss something useful.

THIS IS AN ART!!!

What would you say about these centroids?

LOYALTY	PASSION	CCE	LEADERSH	EDUCATIO	SAFETY	AESTHETI
-1.10	-1.13	-1.18	-0.60	-0.76	-0.41	-0.83
0.20	0.27	0.25	-0.20	-0.03	-0.02	0.04
0.88	0.79	0.88	1.00	0.90	0.48	0.85
INVOLVEM	OPENNESS	SOCIAL_C	DOMAINS	ECONOMY	SOCIAL_O	COMMUNITY
-0.20	-0.77	-0.34	-0.85	-0.57	-0.80	-0.98
0.03	-0.12	0.04	-0.06	-0.14	-0.09	-0.11
0.17	1.05	0.30	1.03	0.86	1.03	1.26

“I love this place”

The “MEH” group

“I gotta get outa here”

What can we say about these groups?

Just naming these groups is one thing. Can we use these 'labels' to know anything about our groups.

That is can we look at information not used in the cluster analysis (demographic etc.)

Here: Cluster 1 = BAD, 2 = Meh, 3 = "Yeah"

Probability of Falling in each cluster

clust	n	prop
1	3928	0.2920880
2	5966	0.4436347
3	3554	0.2642772

With regard to wages

wage	clust	n	prop
$100,000 or over	1	708	0.2401628
$100,000 or over	2	1425	0.4833786
$100,000 or over	3	815	0.2764586

wage	clust	n	prop
Under $15,000	1	248	0.3568345
Under $15,000	2	251	0.3611511
Under $15,000	3	196	0.2820144

What about Race?

race	clust	n	prop
Asian (If necessary, read:) includes Chinese, Filipino...	1	82	0.2827586
Asian (If necessary, read:) includes Chinese, Filipino...	2	149	0.5137931
Asian (If necessary, read:) includes Chinese, Filipino...	3	59	0.2034483
Black or African-American	1	510	0.3635068
Black or African-American	2	545	0.3884533
Black or African-American	3	348	0.2480399
White	1	2945	0.2764999
White	2	4817	0.4522580
White	3	2889	0.2712421

Education?

edu	clust	n	prop
Post-graduate work or degree	1	769	0.2556516
Post-graduate work or degree	2	1434	0.4767287
Post-graduate work or degree	3	805	0.2676197
Some high school	1	164	0.3147793
Some high school	2	184	0.3531670
Some high school	3	173	0.3320537

???

So how can we name these “groups” now that we know more about them?

Are there other questions you can ask about the data?