INSEAD

The Business School
for the World®

Prof. Anton Ovchinnikov

Prof. Spyros Zoumpoulis

Data Science for Business

Sessions 9-10, February 11, 2020

Dimensionality Reduction; Clustering and Segmentation

Structure of the course

- SESSIONS 1-2 (AO): Data analytics process; from Excel to R
  - Tutorial 1: Getting comfortable with R
- SESSIONS 3-4 (AO): Time Series Models
- SESSIONS 5-6 (AO): Introduction to classification
  - Tutorial 2: Midterm R help / classification
- SESSIONS 7-8 (SZ): Advanced Classification; Overfitting and Regularization; From .R to Notebooks
  - Tutorial 3: Setup with GitHub and knitting notebooks
- **SESSIONS 9-10 (SZ): Dimensionality Reduction; Clustering and Segmentation**
- SESSIONS 11-12 (SZ): AI in Business; The Data Science Process; Guest speaker
  - Hands-on help with projects
- SESSIONS 13-14 (AO+SZ): Project presentations

Plan for the day
Learning objectives

- Derived attributes and dimensionality reduction
  - Generate (a small number of) new manageable/ interpretable attributes that capture most of the information in the data
- Clustering and segmentation
  - Group observations in a few segments so that data within any segment are similar while data across segments are different
- Work on business solution template for market segmentation (Assignment 3) for the Boats (A) case

Derived Attributes and Dimensionality Reduction

- What is dimensionality reduction?
  - Generate (a small number of) new attributes that are (linear) combinations of the original ones, and capture most of the information in the original data
  - Often used as the first step in data analytics

- Why do dimensionality reduction?
  - Computational and statistical reasons: with thousands of features, very expensive and hard to estimate a good model
  - Managerial reason: the new attributes are interpretable and actionable

- The key idea of dimensionality reduction
  - Transform the original variables into a smaller set of **factors**
  - Understand and interpret the factors
  - Use the factors for subsequent analysis

Dimensionality Reduction: Key Questions

INSEAD
The Business School for the World®

1. How many factors do we need?
2. How would you name the factors? What do they mean?
3. How interpretable and actionable are the factors we found?

Applying Dimensionality Reduction: Evaluation of MBA Applications

Variables available:

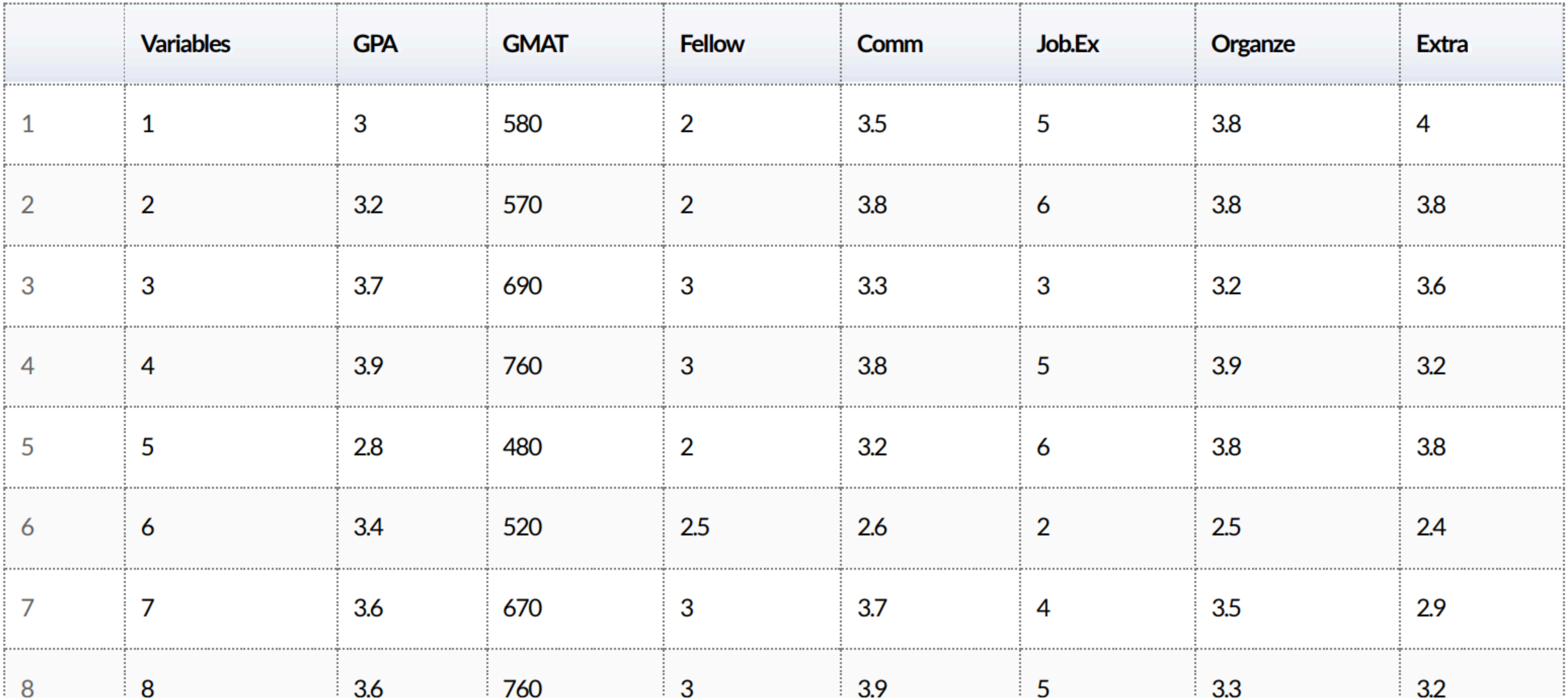
1. GPA
2. GMAT score
3. Scholarships, fellowships won
4. Evidence of communications skills
5. Prior job experience
6. Organizational experience
7. Other extra curricular achievements

Which variables are correlated? What do these groups of variables capture?

(A) Process for Dimensionality Reduction

1. Confirm the data is metric
2. Scale the data
3. Check correlations
4. Choose number of factors
5. Interpret the factors
6. Save factor scores

Step 1: Confirm data is metric

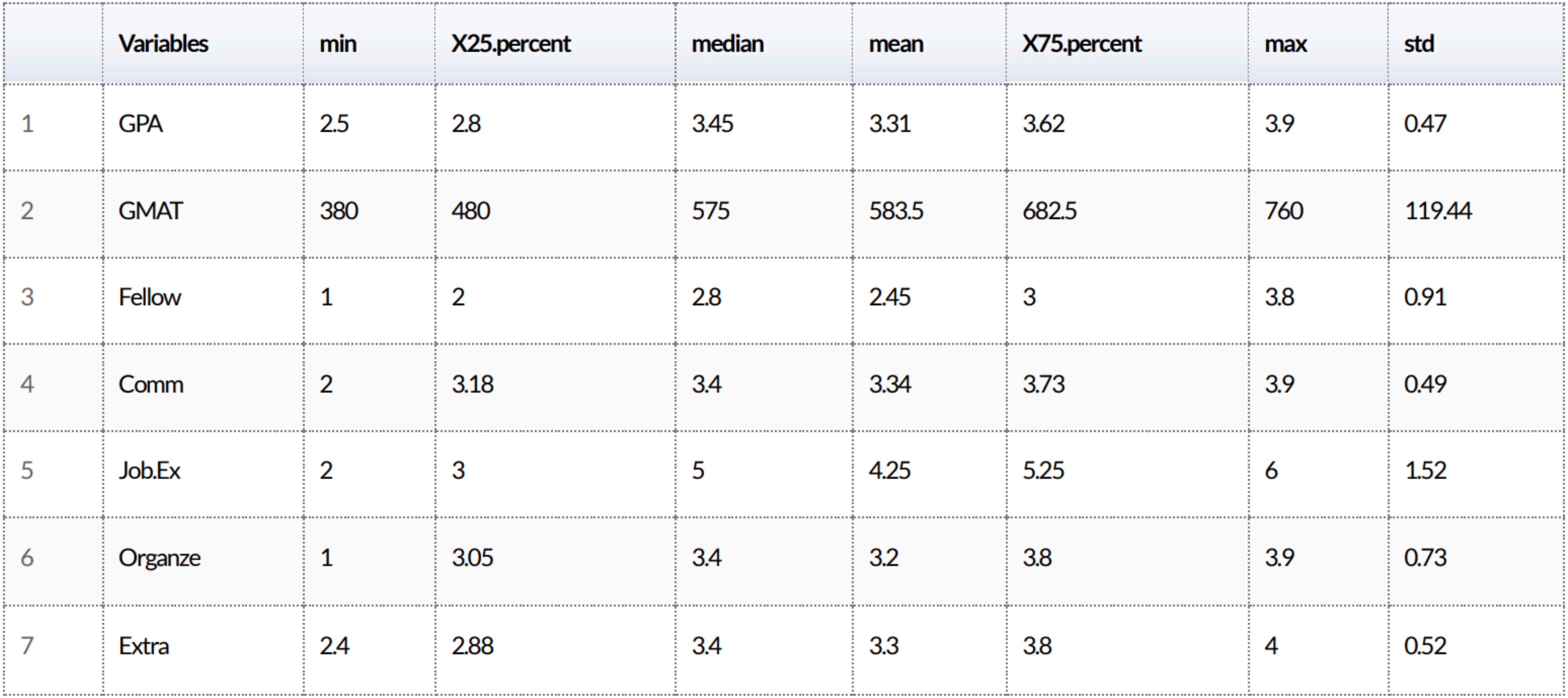


	Variables	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
1	1	3	580	2	3.5	5	3.8	4
2	2	3.2	570	2	3.8	6	3.8	3.8
3	3	3.7	690	3	3.3	3	3.2	3.6
4	4	3.9	760	3	3.8	5	3.9	3.2
5	5	2.8	480	2	3.2	6	3.8	3.8
6	6	3.4	520	2.5	2.6	2	2.5	2.4
7	7	3.6	670	3	3.7	4	3.5	2.9
8	8	3.6	760	3	3.9	5	3.3	3.2

Step 2: Scale the data

INSEAD
The Business School for the World®

Before standardization



	Variables	min	X25.percent	median	mean	X75.percent	max	std
1	GPA	2.5	2.8	3.45	3.31	3.62	3.9	0.47
2	GMAT	380	480	575	583.5	682.5	760	119.44
3	Fellow	1	2	2.8	2.45	3	3.8	0.91
4	Comm	2	3.18	3.4	3.34	3.73	3.9	0.49
5	Job.Ex	2	3	5	4.25	5.25	6	1.52
6	Organze	1	3.05	3.4	3.2	3.8	3.9	0.73
7	Extra	2.4	2.88	3.4	3.3	3.8	4	0.52

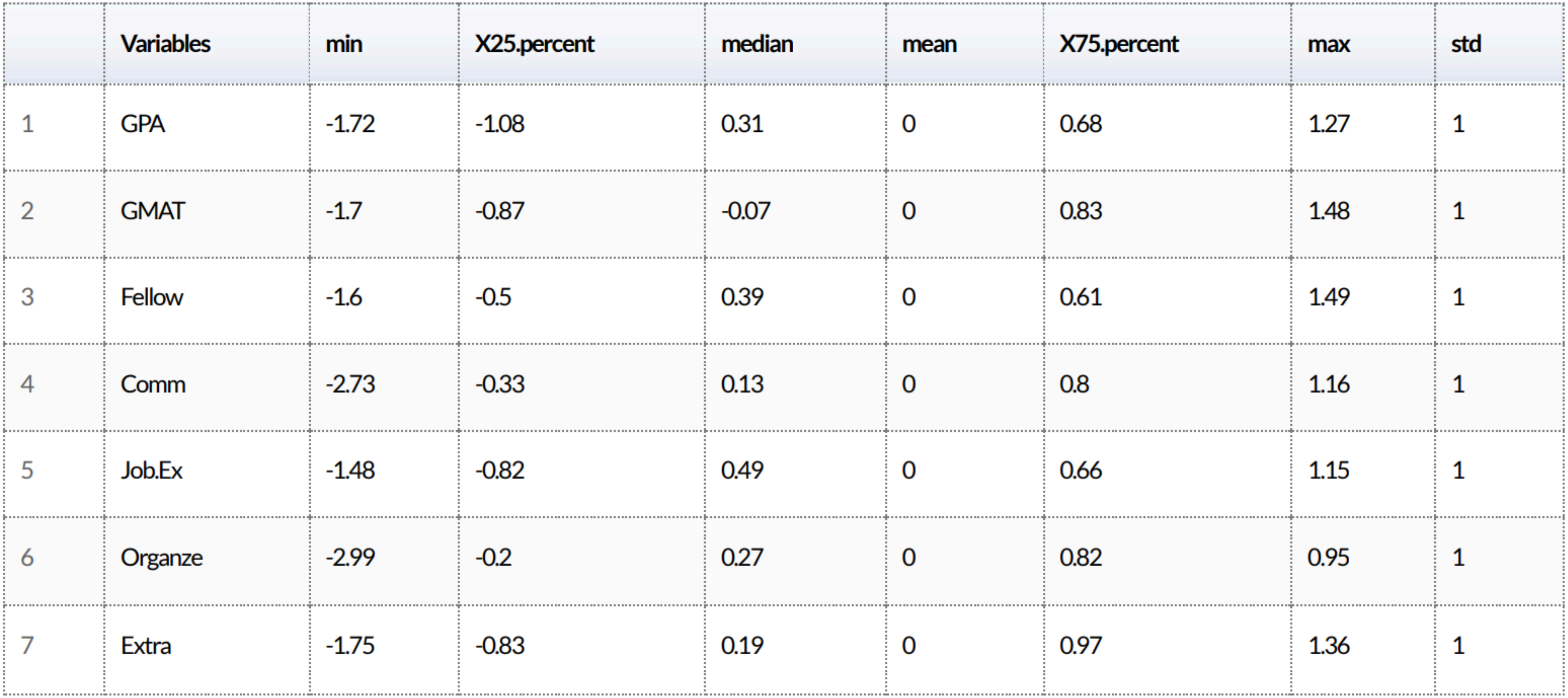
Step 2: Scale the data

Standardization….

```
ProjectDatafactor_scaled=apply(ProjectDataFactor,2, function(r) { #"2" applies the function over columns
  if (sd(r)!=0) {
    res=(r-mean(r))/sd(r)
  } else {
    res=0*r
  }
  res
})
```

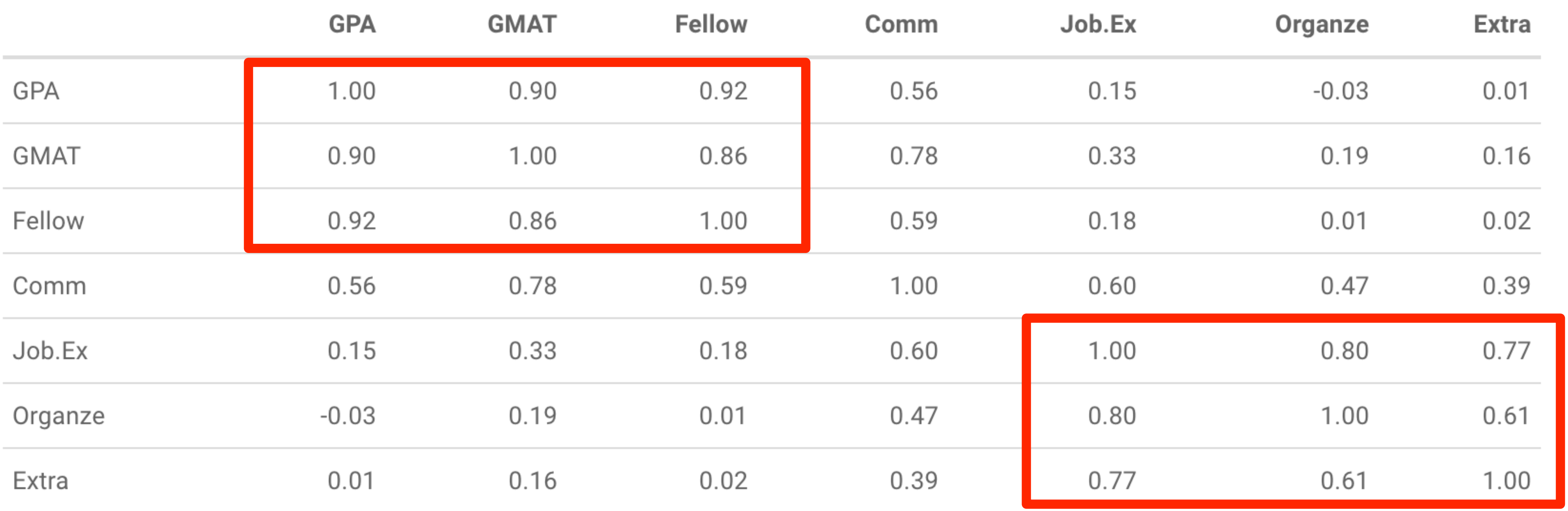
Step 2: Scale the data

After standardization



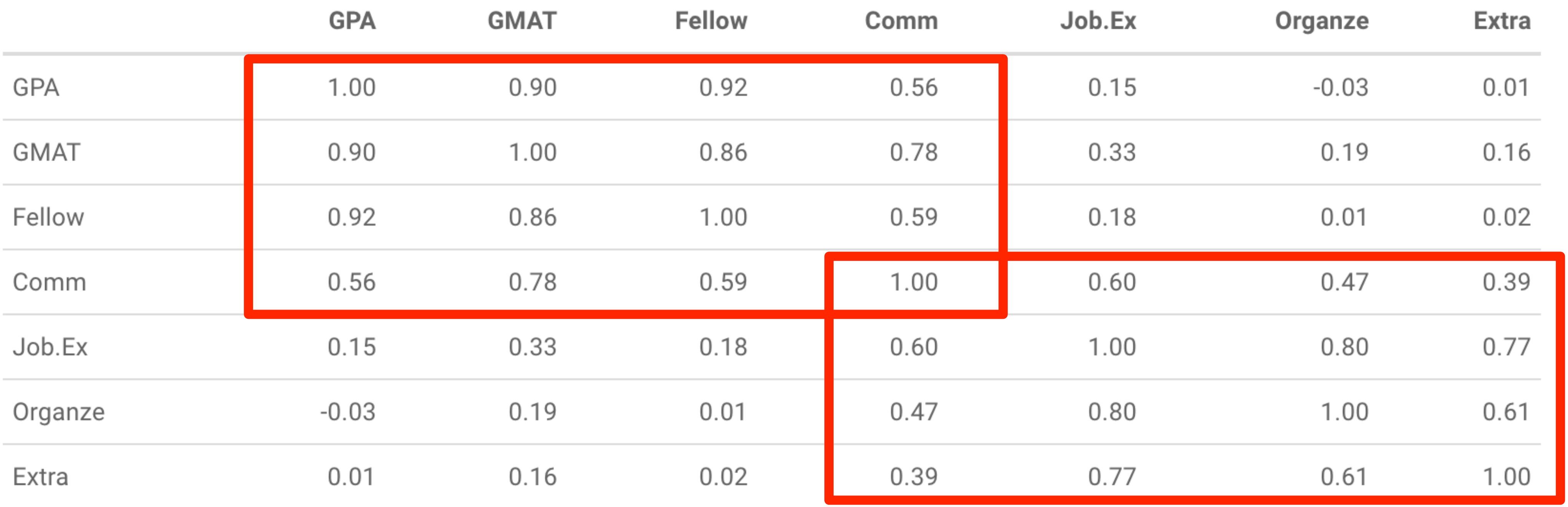
	Variables	min	X25.percent	median	mean	X75.percent	max	std
1	GPA	-1.72	-1.08	0.31	0	0.68	1.27	1
2	GMAT	-1.7	-0.87	-0.07	0	0.83	1.48	1
3	Fellow	-1.6	-0.5	0.39	0	0.61	1.49	1
4	Comm	-2.73	-0.33	0.13	0	0.8	1.16	1
5	Job.Ex	-1.48	-0.82	0.49	0	0.66	1.15	1
6	Organze	-2.99	-0.2	0.27	0	0.82	0.95	1
7	Extra	-1.75	-0.83	0.19	0	0.97	1.36	1

Step 3: Check correlations



	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
GPA	1.00	0.90	0.92	0.56	0.15	-0.03	0.01
GMAT	0.90	1.00	0.86	0.78	0.33	0.19	0.16
Fellow	0.92	0.86	1.00	0.59	0.18	0.01	0.02
Comm	0.56	0.78	0.59	1.00	0.60	0.47	0.39
Job.Ex	0.15	0.33	0.18	0.60	1.00	0.80	0.77
Organze	-0.03	0.19	0.01	0.47	0.80	1.00	0.61
Extra	0.01	0.16	0.02	0.39	0.77	0.61	1.00

Step 3: Check correlations



	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
GPA	1.00	0.90	0.92	0.56	0.15	-0.03	0.01
GMAT	0.90	1.00	0.86	0.78	0.33	0.19	0.16
Fellow	0.92	0.86	1.00	0.59	0.18	0.01	0.02
Comm	0.56	0.78	0.59	1.00	0.60	0.47	0.39
Job.Ex	0.15	0.33	0.18	0.60	1.00	0.80	0.77
Organze	-0.03	0.19	0.01	0.47	0.80	1.00	0.61
Extra	0.01	0.16	0.02	0.39	0.77	0.61	1.00

Step 4: Choose the number of factors

We use Principal Component Analysis

Package: psych

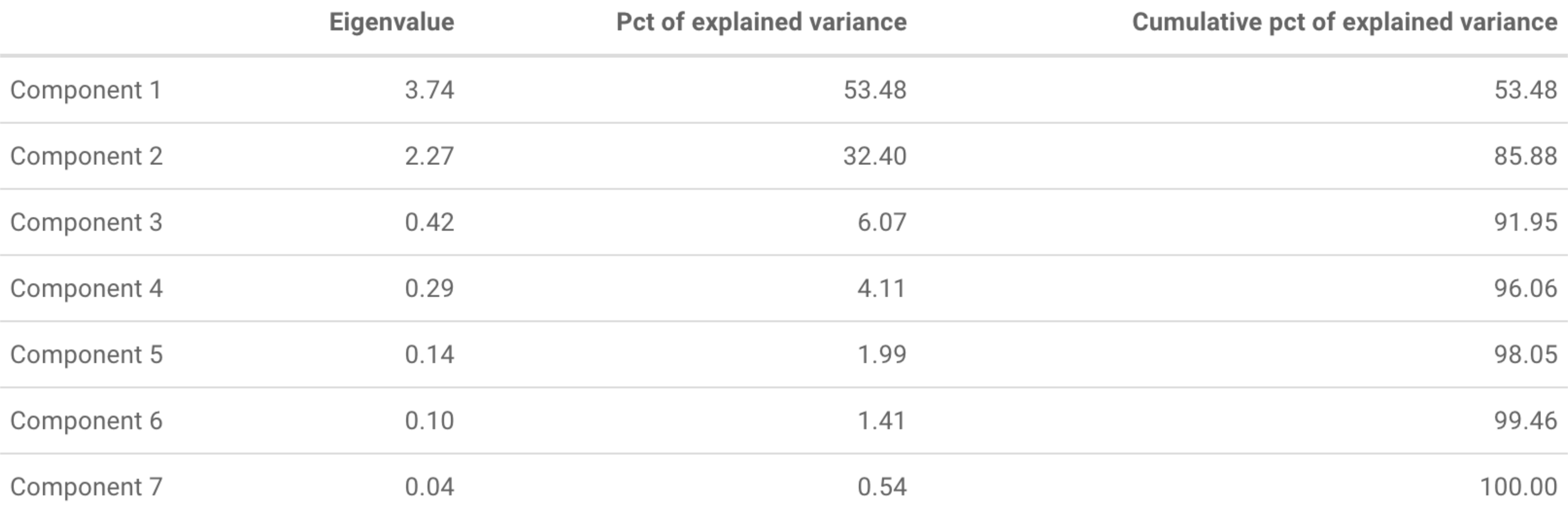
```
UnRotated_Results<-principal(ProjectDataFactor, nfactors=ncol(ProjectDataFactor),
                                                    rotate="none”, score=TRUE)
```

- Factors are linear combinations of the original raw attributes…
- …so that they capture as much of the variability in the data as possible
- Factors are uncorrelated, and as many as the variables
- Each factor has an associated “eigenvalue” – which corresponds to the amount of variance captured by that factor
- First factor has the highest eigenvalue and explains most of the variance, then the second, …, and so on

Step 4: Choose the number of factors

Package: FactoMineR

```
Variance_Explained_Table_results<-PCA(ProjectDataFactor, graph=FALSE)
Variance_Explained_Table<-Variance_Explained_Table_results$eig
```



	Eigenvalue	Pct of explained variance	Cumulative pct of explained variance
Component 1	3.74	53.48	53.48
Component 2	2.27	32.40	85.88
Component 3	0.42	6.07	91.95
Component 4	0.29	4.11	96.06
Component 5	0.14	1.99	98.05
Component 6	0.10	1.41	99.46
Component 7	0.04	0.54	100.00

```
> Variance_Explained_Table[1,1]/sum(Variance_Explained_Table[,1])    ??
[1] 0.5347987
```

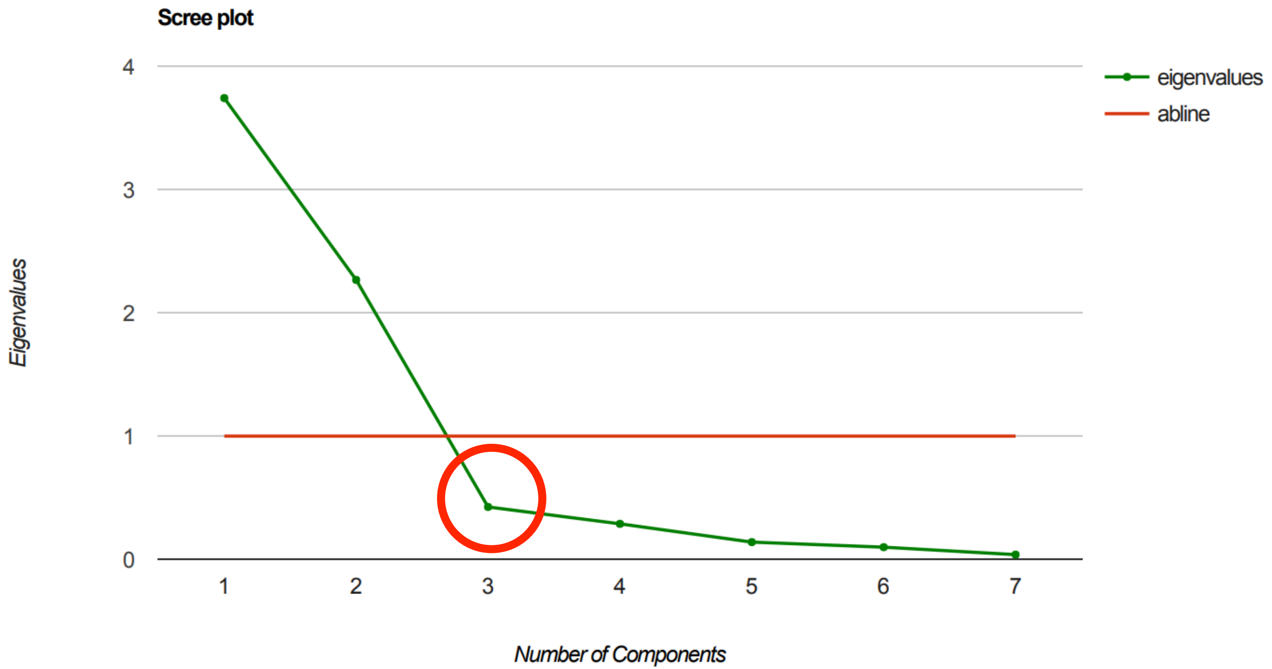
Step 4: Choose the number of factors

INSEAD

The Business School for the World®

We want to capture as much of the variance as possible, with as few factors as possible. How to choose the factors? Three criteria to use:

- Select all factors with eigenvalue > 1
- Select factors with highest eigenvalues up to exceeding a threshold (e.g. 65%) in cumulative % of explained variance
- Select factors up to the "elbow" of the scree plot



Step 5: Interpret the factors

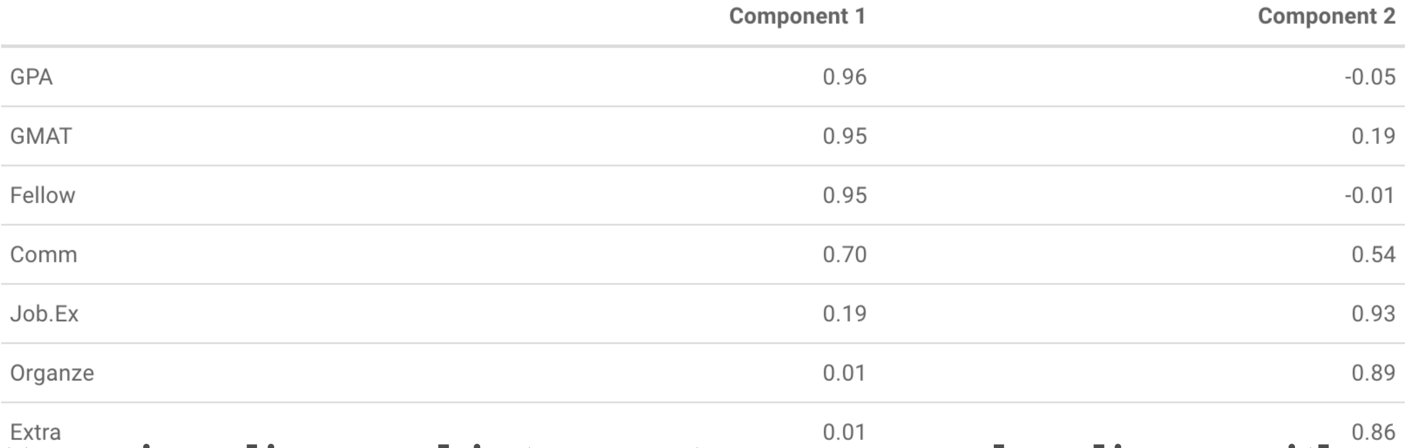
To interpret the factors, we want them to use only a few, non-overlapping original attributes

- Factor “rotations” transform the estimated factors into new ones that satisfy that, while capturing the same information

Step 5: Interpret the factors

Package: psych

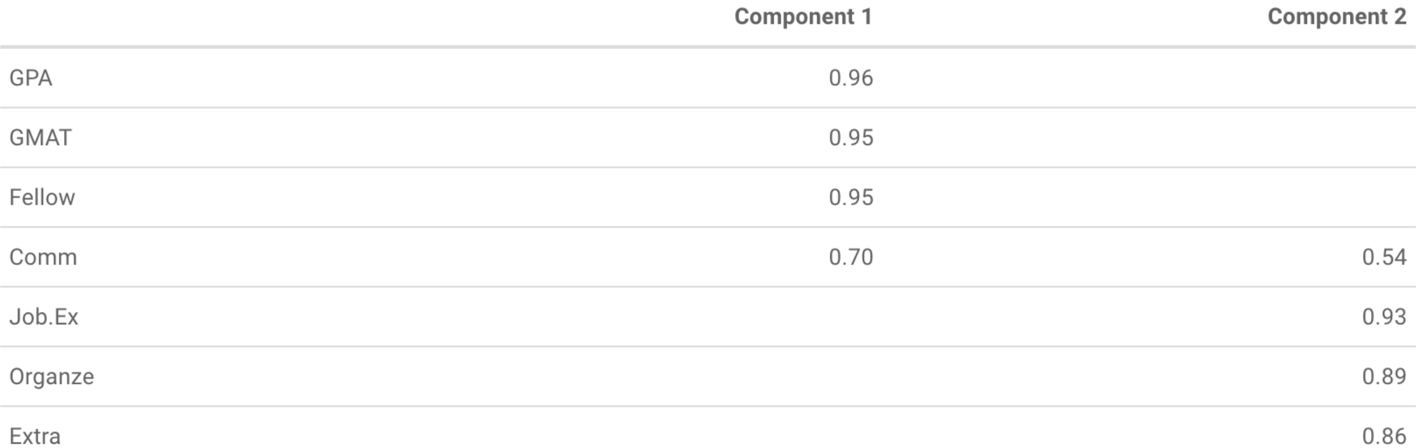
```
Rotated_Results<-principal(ProjectDataFactor, nfactors=max(factors_selected),
                            rotate="varimax", score=TRUE)
Rotated_Factors<-round(Rotated_Results$loadings,2)
```



	Component 1	Component 2
GPA	0.96	-0.05
GMAT	0.95	0.19
Fellow	0.95	-0.01
Comm	0.70	0.54
Job.Ex	0.19	0.93
Organze	0.01	0.89
Extra	0.01	0.86

To better visualize and interpret: suppress loadings with small values

```
Rotated_Factors_thres <- Rotated_Factors
Rotated_Factors_thres[abs(Rotated_Factors_thres) < 0.5]<- NA
```

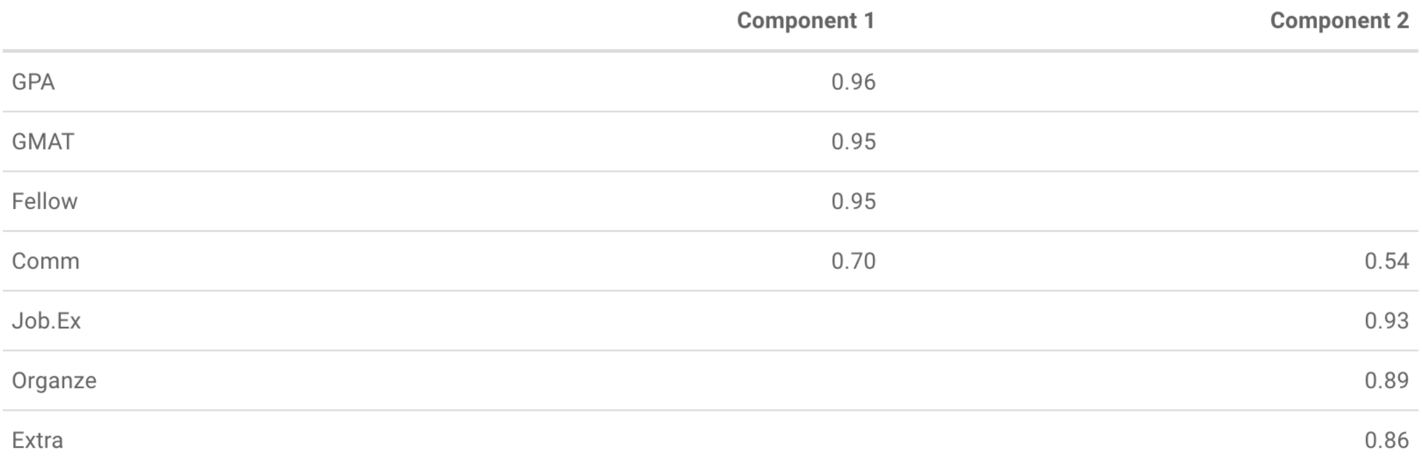


	Component 1	Component 2
GPA	0.96	
GMAT	0.95	
Fellow	0.95	
Comm	0.70	0.54
Job.Ex		0.93
Organze		0.89
Extra		0.86

Step 5: Interpret the factors

What factor loads “look good"? Three technical quality criteria:

1. For each factor (column) only a few loadings are large (in absolute value)
2. For each raw attribute (row) only a few loadings are large (in absolute value)
3. Any pair of factors (columns) should have different "patterns" of loading



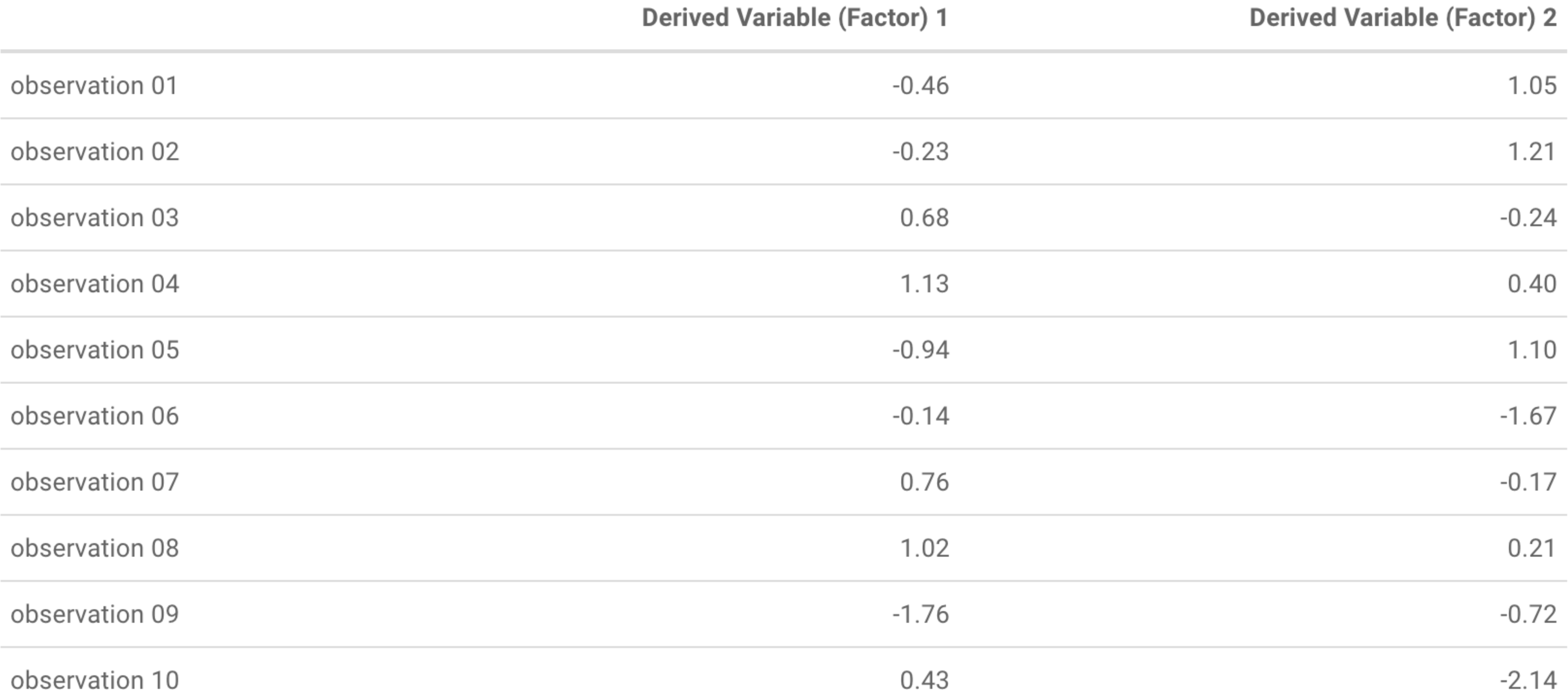
	Component 1	Component 2
GPA	0.96	
GMAT	0.95	
Fellow	0.95	
Comm	0.70	0.54
Job.Ex		0.93
Organze		0.89
Extra		0.86

Step 6: Save factor scores

Replace the original data with a new dataset where each observation (row) is described using the selected derived factors

- For each row, estimate the **factor scores**: how the observation "scores" for each of the selected factors

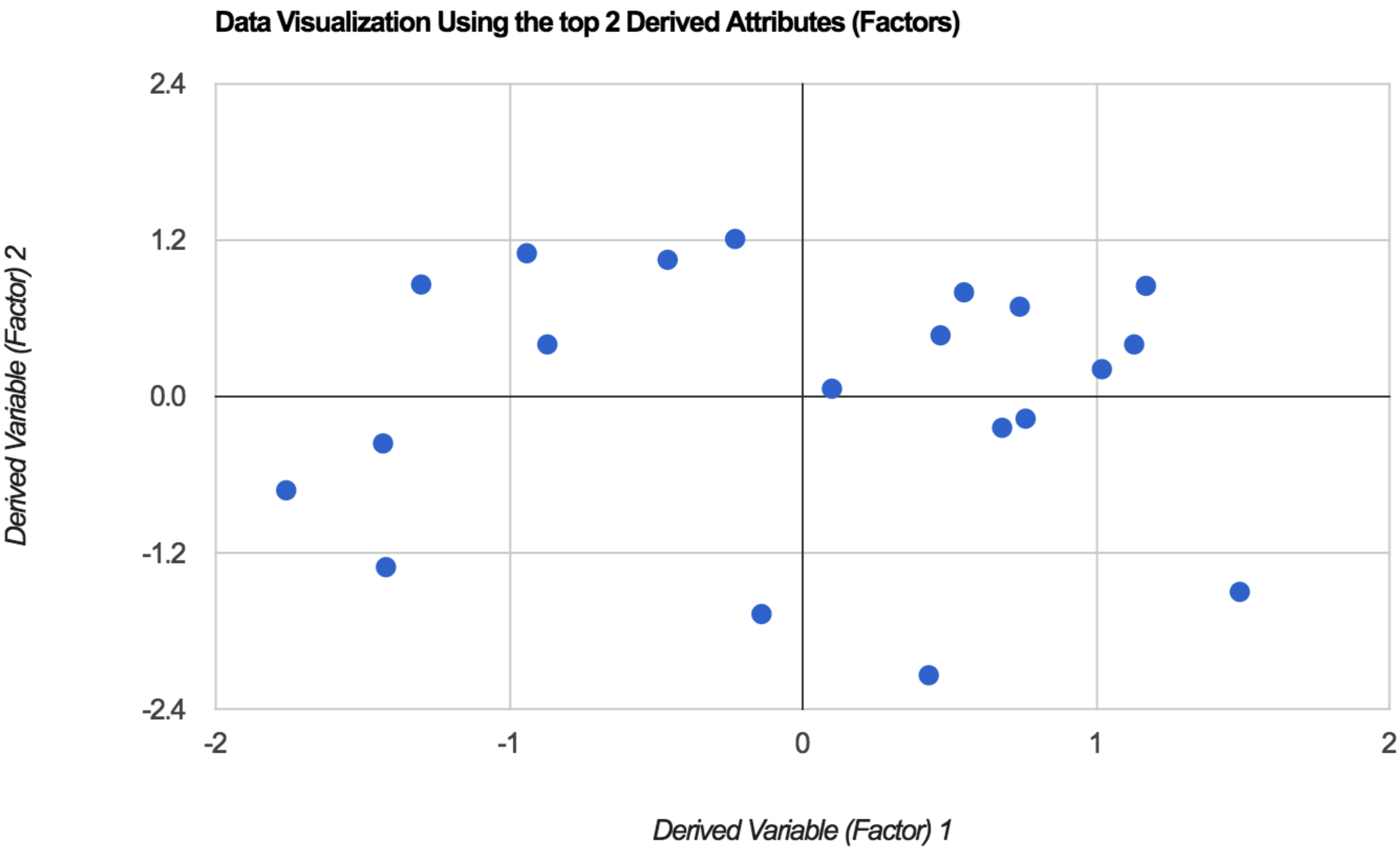
Package: psych
NEW_ProjectData <- round(Rotated_Results$scores[,1:factors_selected],2)



	Derived Variable (Factor) 1	Derived Variable (Factor) 2
observation 01	-0.46	1.05
observation 02	-0.23	1.21
observation 03	0.68	-0.24
observation 04	1.13	0.40
observation 05	-0.94	1.10
observation 06	-0.14	-1.67
observation 07	0.76	-0.17
observation 08	1.02	0.21
observation 09	-1.76	-0.72
observation 10	0.43	-2.14

Step 6: Save factor scores

Then continue the analysis (e.g., make decision, or do clustering, etc.) with the new attributes



Clustering and Segmentation

- What is clustering and segmentation?
  - Processes and tools to organize data in a few segments, with data being as similar as possible within each segment, and as different as possible across segments

- Applications
  - Market segmentation
  - Co-moving asset classes
  - Geo-demographic segmentation
  - Recommender systems
  - Text mining

(A) Process for Clustering

1. Confirm the data is metric
2. Scale the data
3. Select segmentation variables
4. Define similarity measure
5. Visualize pair-wise distances
6. Method and number of segments
7. Profile and interpret the segments
8. Robustness analysis

Step 3. Select segmentation variables

Critically important decision for the solution

- Requires lots of contextual knowledge and creativity

Segmentation attributes vs. **profiling attributes**

For market research:

- Use attitudinal data for segmentation, so as to segment customers based on attitudes/needs
  - If ran dimensionality reduction before: segmentation attributes can be the original attributes with the highest absolute factor loading for each factor
- Use demographic and behavioral data for profiling the clusters found

Step 4. Define similarity measure

Important: need to understand what makes two observations "similar" or "different"

There are infinitely many rigorous mathematical definitions of distance between two observations

Euclidean distance:

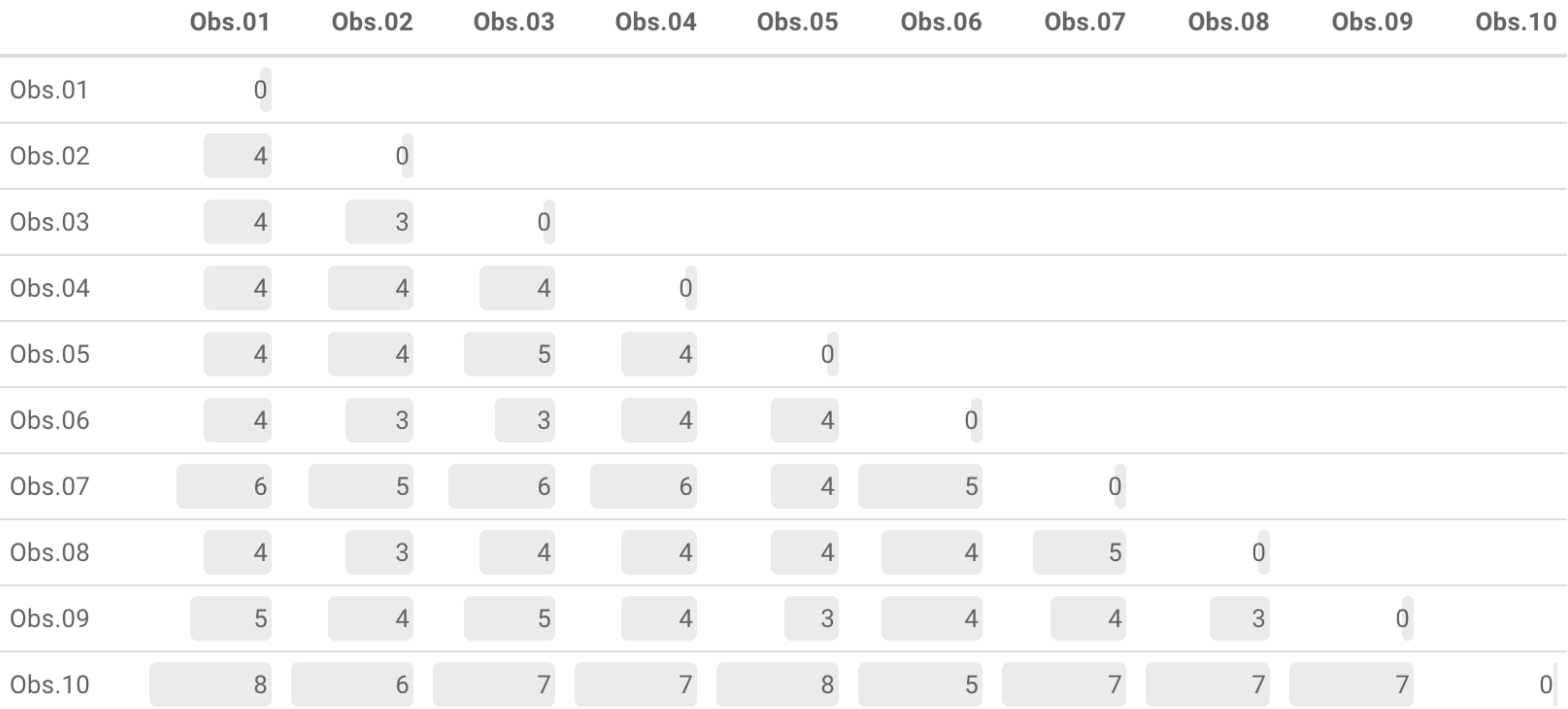
$$\|x - z\|_2 = \sqrt{(x_1 - z_1)^2 + \ldots (x_p - z_p)^2}$$

Manhattan distance:

$$\|x - z\|_1 = |x_1 - z_1| + \ldots + |x_p - z_p|$$

Step 4. Define similarity measure

Using Euclidean distance:



	Obs.01	Obs.02	Obs.03	Obs.04	Obs.05	Obs.06	Obs.07	Obs.08	Obs.09	Obs.10
Obs.01	0									
Obs.02	4	0								
Obs.03	4	3	0							
Obs.04	4	4	4	0						
Obs.05	4	4	5	4	0					
Obs.06	4	3	3	4	4	0				
Obs.07	6	5	6	6	4	5	0			
Obs.08	4	3	4	4	4	4	5	0		
Obs.09	5	4	5	4	3	4	4	3	0	
Obs.10	8	6	7	7	8	5	7	7	7	0

Step 4. Define similarity measure

Can also define distance manually

- Let’s say that the management team believes that two customers are similar for an attitude if they do not differ in their ratings for that attitude by more than 2 points
- We can manually assign a distance of 1 for every question for which two customers gave an answer that differs by more than 2 points, and 0 otherwise

```
My_Distance_function<-function(x,y){ # x, y are vectors (answers of customers)
                sum(abs(x - y) > 2)
}
```

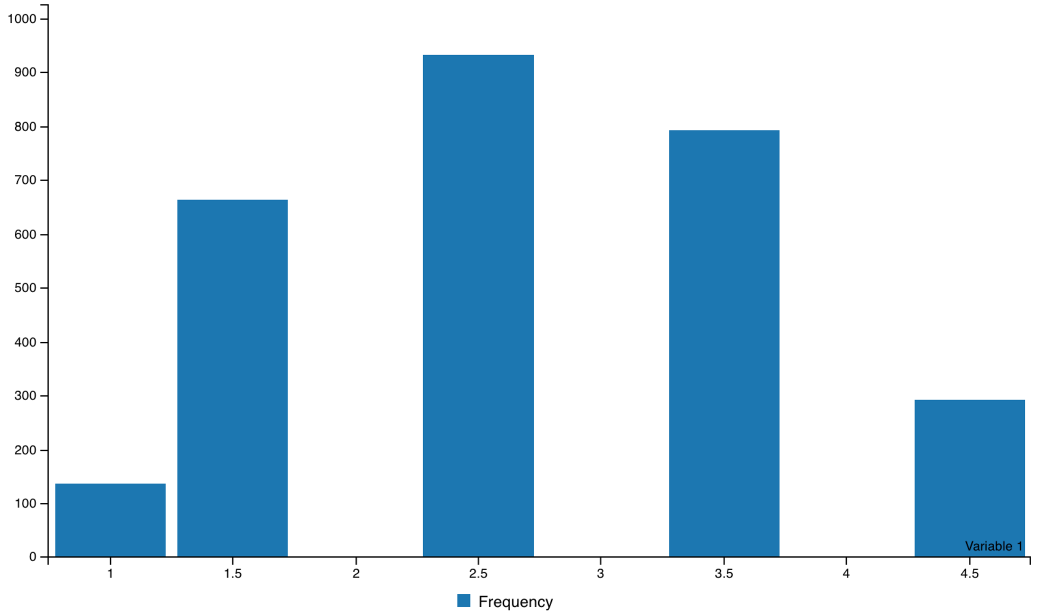
Step 5. Visualize pairwise distances

INSEAD

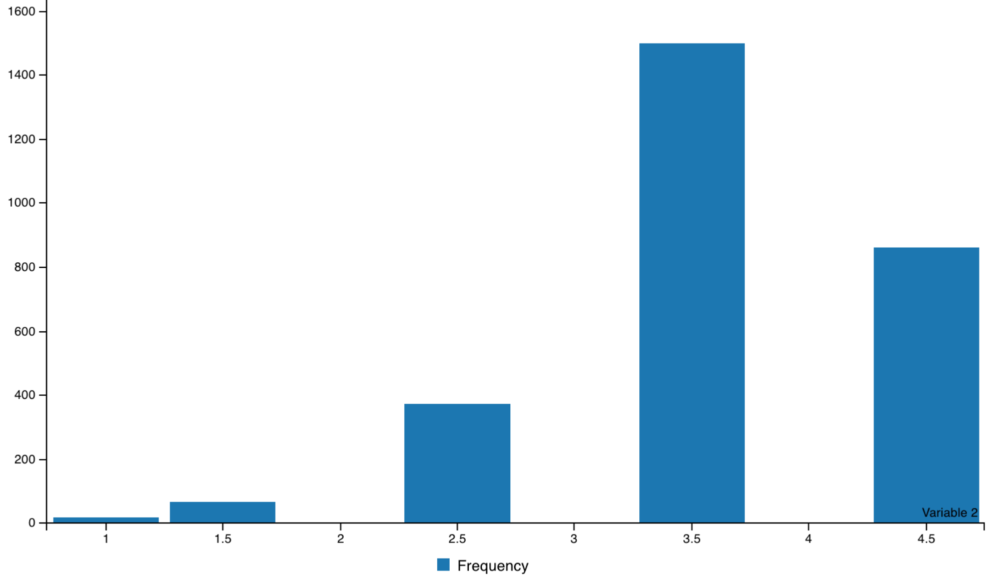
The Business School for the World®

Visualize individual attributes…

Q1.27: Boating is the number one thing I do in my spare time

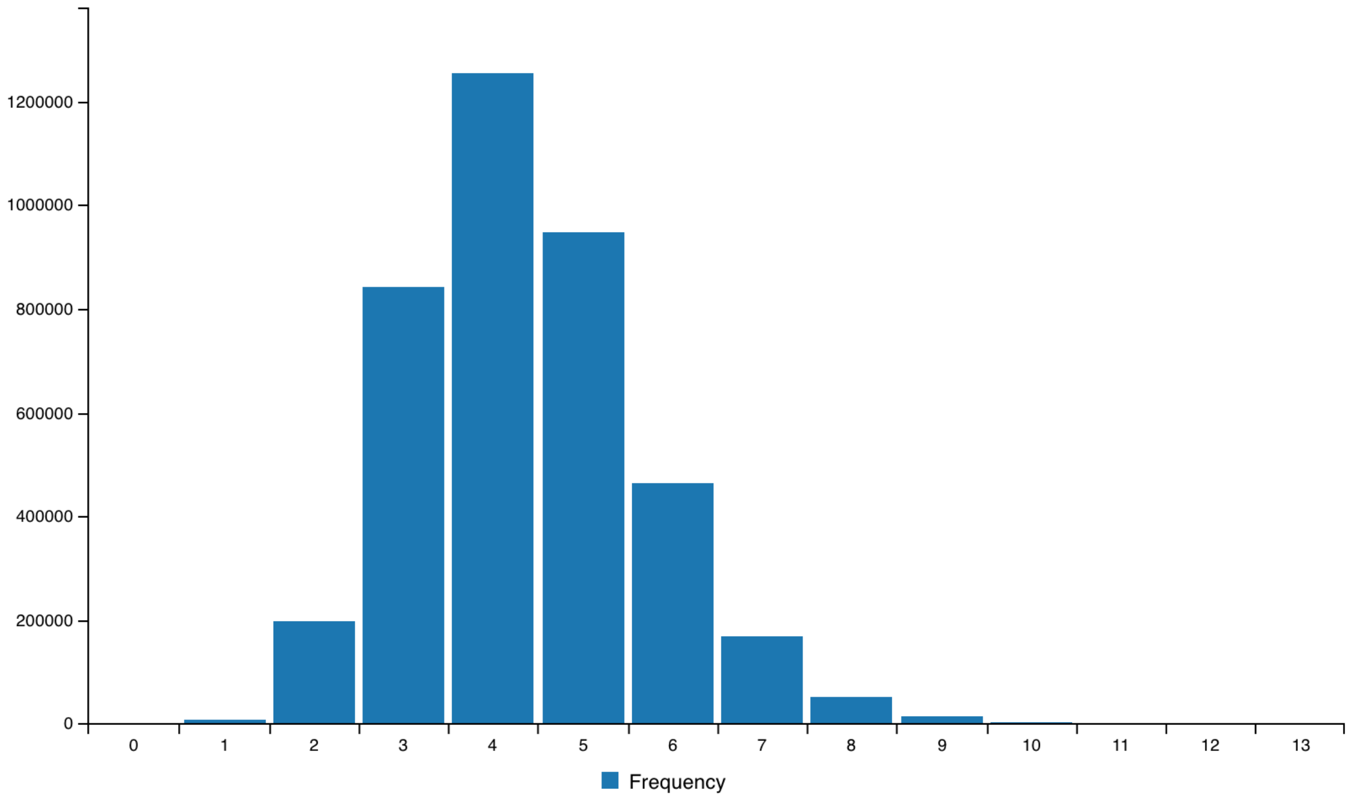


Q1.24: Boating gives me an outlet to socialize with family and/or friends



Step 5. Visualize pairwise distances

… and pairwise distances



Step 6. Method and number of segments

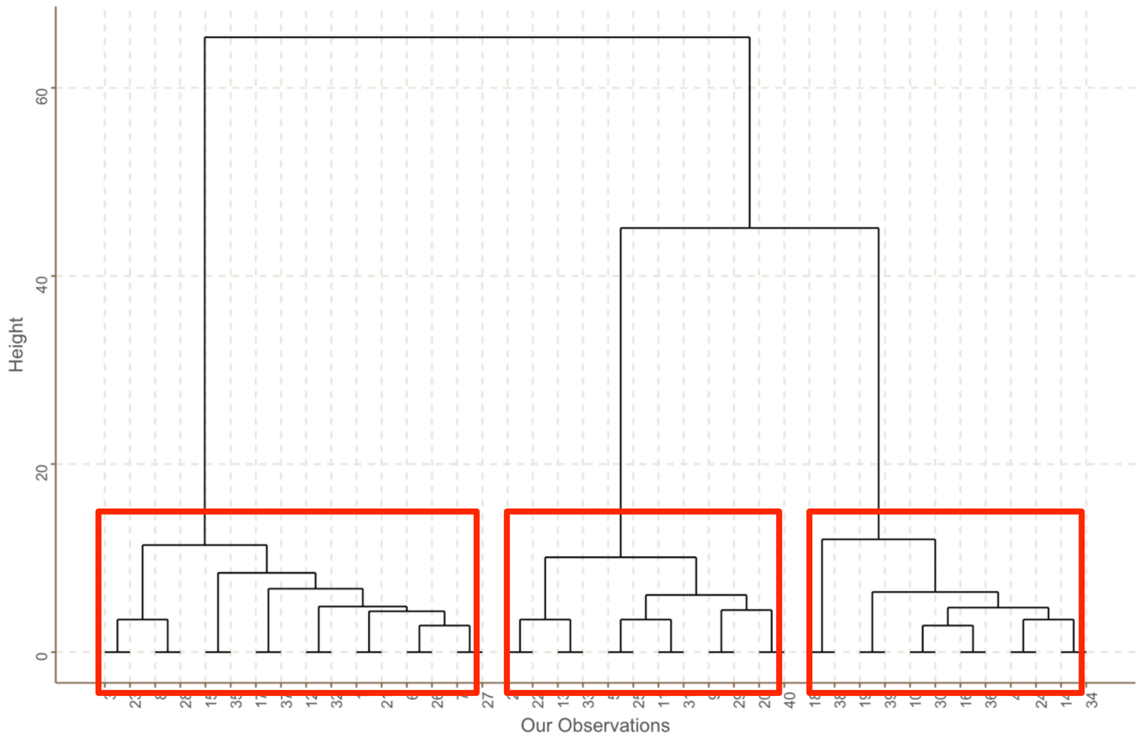
Many clustering methods. In practice, we want to use various approaches and select the solution that is robust, interpretable, actionable.

- Hierarchical clustering
- K-means

We can plug-and-play this “black box” in our analysis – with care

Step 6. Method and number of segments

Hierarchical Clustering



"Dendrogram"

- Observations that are the closest to each other are grouped together
- Start with pairs
- Merge smaller groups into larger ones
- Eventually all our data are merged into one segment
- Heights of the branches of the tree indicate how different are the clusters merged at that level of the tree
- Then cut the tree so as to create the desired number of clusters

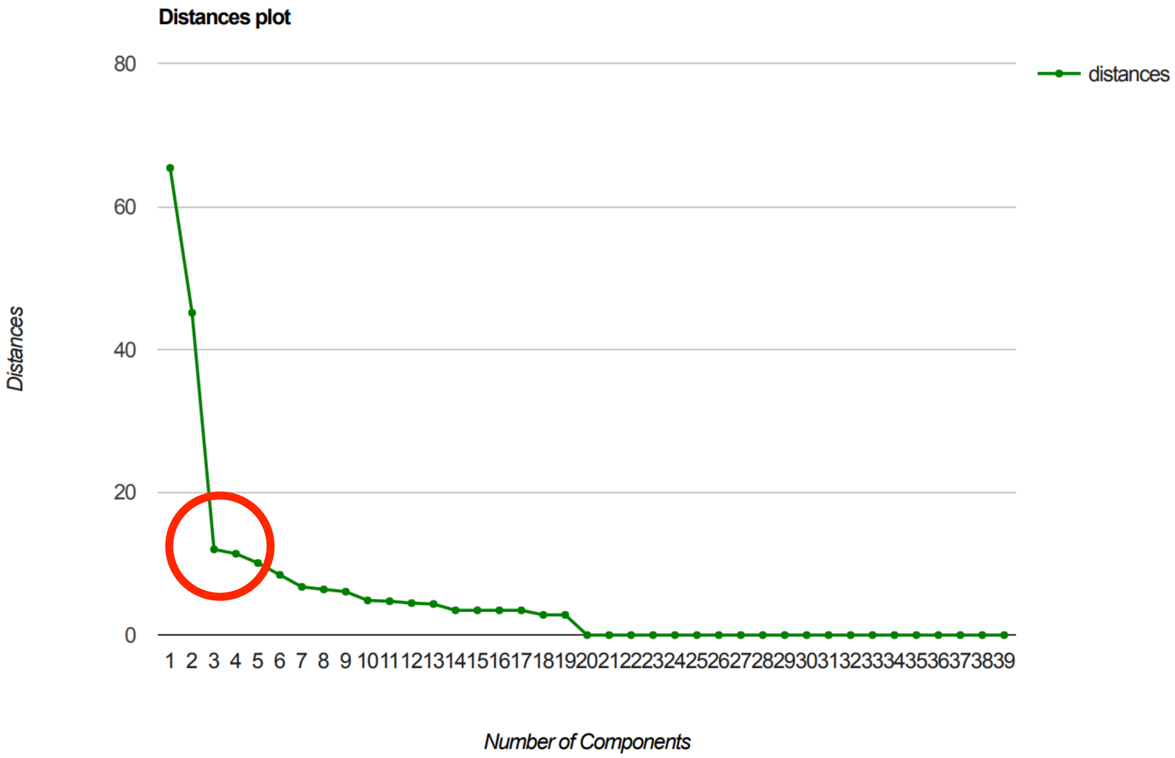
Step 6. Method and number of segments

Hierarchical Clustering

```
ProjectData_segment <- ProjectData[,segmentation_attributes_used]
Hierarchical_Cluster_distances <- dist(ProjectData_segment, method="euclidean")
Hierarchical_Cluster <- hclust(Hierarchical_Cluster_distances, method="ward.D")
# Display dendrogram
iplot.dendrogram(Hierarchical_Cluster)
```

Step 6. Method and number of segments

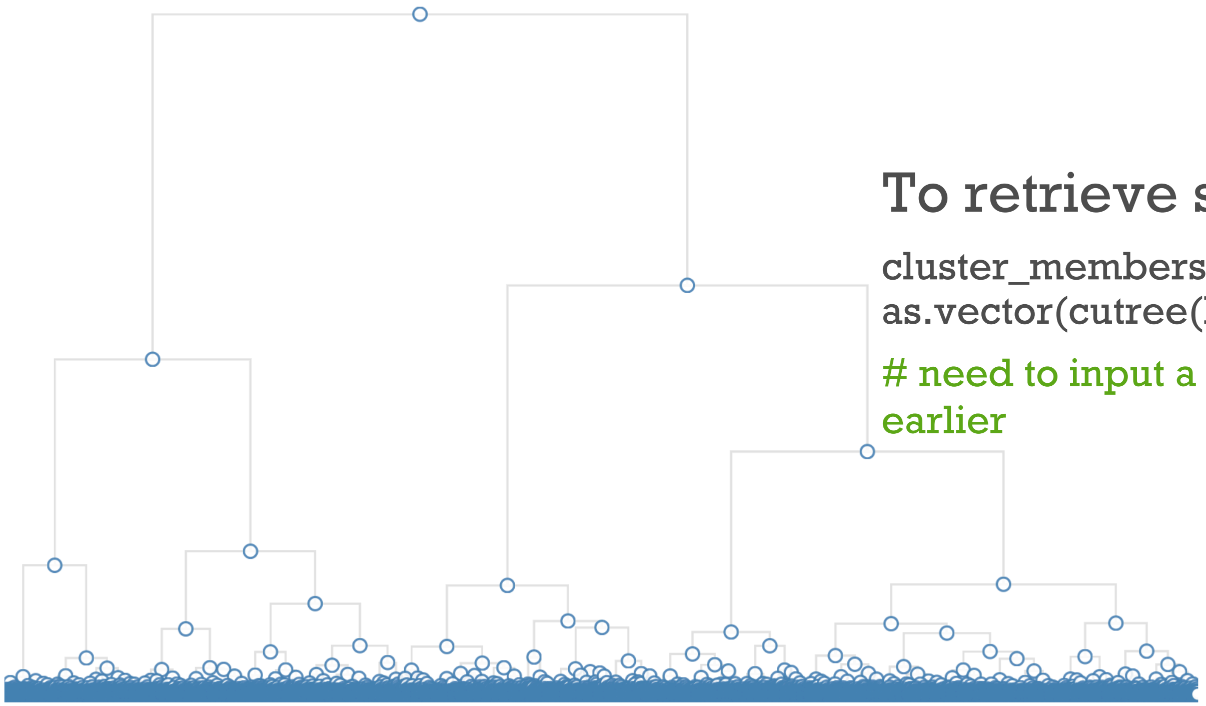
Hierarchical clustering: Choosing the number of clusters



- Rule of thumb: set number of clusters as the "elbow" of the plot
- In practice: start with above rule, then explore different numbers of clusters
- Select final solution using also interpretability

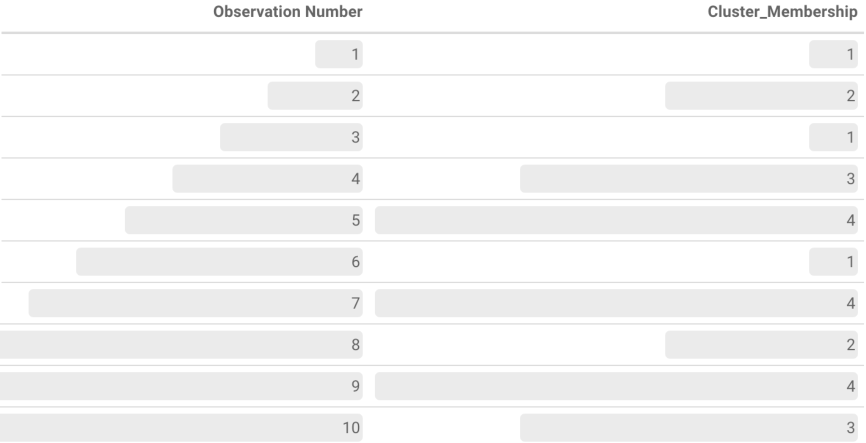
Step 6. Method and number of segments

Hierarchical clustering on Boats data



To retrieve segment membership:

```
cluster_memberships_hclust <-
as.vector(cutree(HierarchicalCluster, k=numb_clusters_used))
# need to input a number of clusters for cutting the tree, not
earlier
```



Observation Number	Cluster_Membership
1	1
2	2
3	1
4	3
5	4
6	1
7	4
8	2
9	4
10	3

Step 6. Method and number of segments

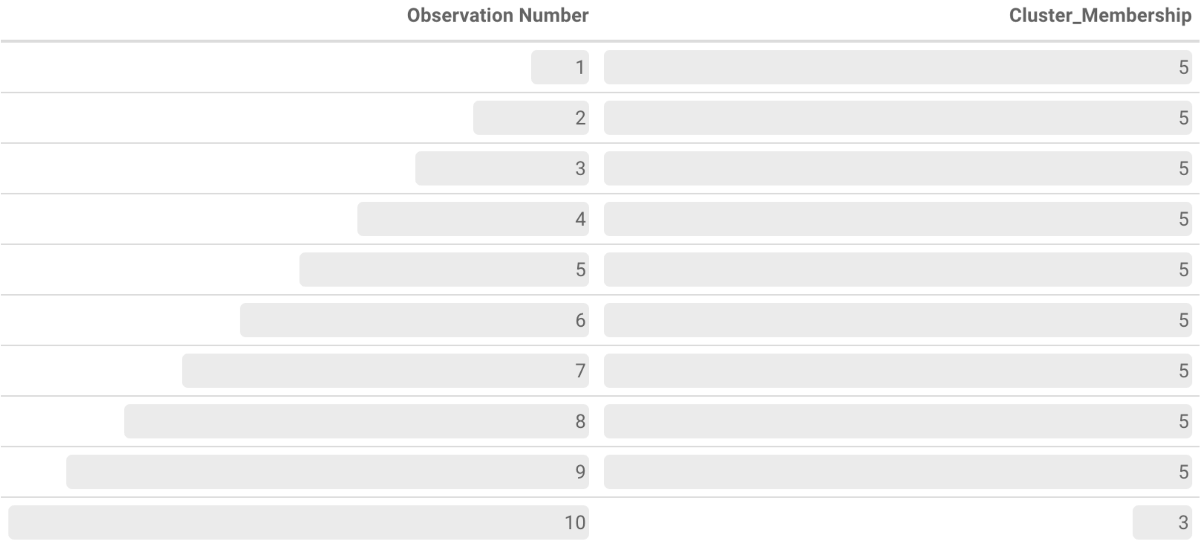
K-means clustering aims to partition the observations into k sets so as to minimize the sum of within-cluster variances

- In each iteration, every observation is assigned to the nearest mean. Then means are recalculated.
- K-means does not necessarily lead to the same solution every time you run it

```
kmeans_clusters <- kmeans(ProjectData_segment,centers = numb_clusters_used, iter.max =
2000, algorithm="Lloyd")
# need to input number of clusters as
```

To retrieve segment membership:

```
kmeans_clusters$cluster
```



Observation Number	Cluster_Membership
1	5
2	5
3	5
4	5
5	5
6	5
7	5
8	5
9	5
10	3

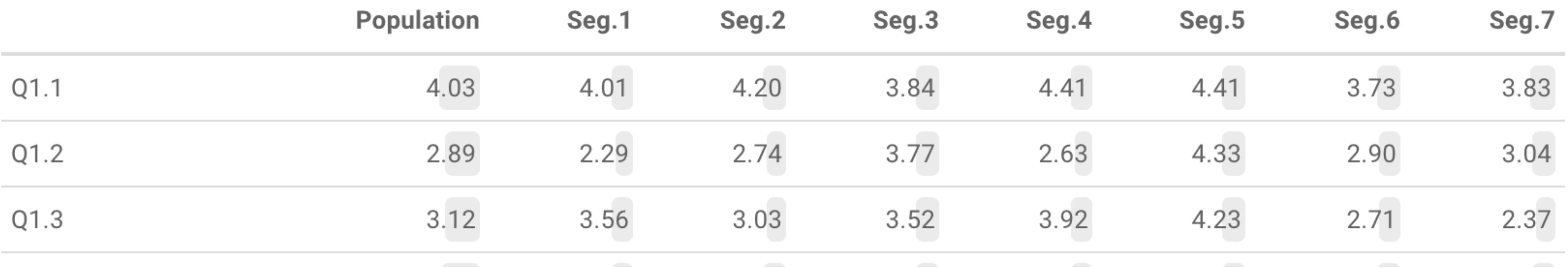
Different methods may put observations in different clusters

Step 7. Profile and interpret the segments

What are the resulting segments? We need to be able to understand and interpret the clustering solution

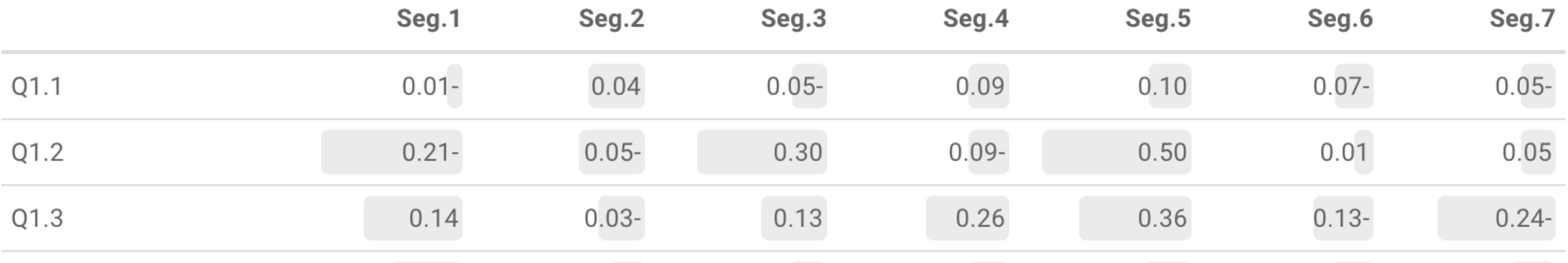
- Profile the segments using the profiling attributes

Average values within each segment and in total population



	Population	Seg.1	Seg.2	Seg.3	Seg.4	Seg.5	Seg.6	Seg.7
Q1.1	4.03	4.01	4.20	3.84	4.41	4.41	3.73	3.83
Q1.2	2.89	2.29	2.74	3.77	2.63	4.33	2.90	3.04
Q1.3	3.12	3.56	3.03	3.52	3.92	4.23	2.71	2.37

avg(segment)/avg(population) - 1

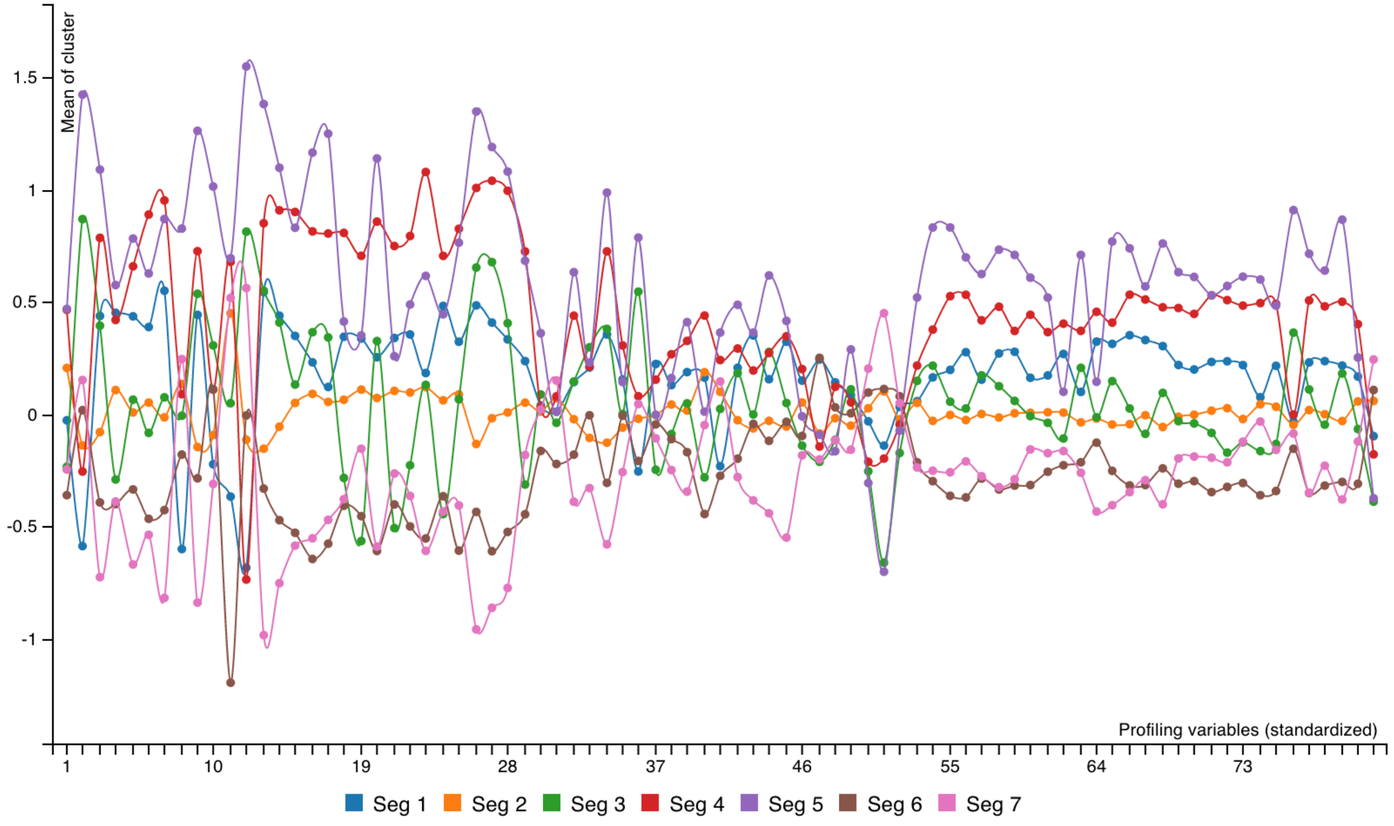


	Seg.1	Seg.2	Seg.3	Seg.4	Seg.5	Seg.6	Seg.7
Q1.1	0.01-	0.04	0.05-	0.09	0.10	0.07-	0.05-
Q1.2	0.21-	0.05-	0.30	0.09-	0.50	0.01	0.05
Q1.3	0.14	0.03-	0.13	0.26	0.36	0.13-	0.24-

Step 7. Profile and interpret the segments

INSEAD
The Business School for the World®

Snake plots for each cluster: means of (standardized) profiling variables



Step 8. Robustness analysis

The segments found should be relatively robust to changes in the clustering methodology

- Large changes indicate that segmentation is not valid

Two basic tests for statistical robustness and stability of interpretation:

1. How much overlap is there between the clusters found using Hierarchical vs. Kmeans?
2. How similar are the profiles of the segments found?

Also try different

- subsets of the original data
- variations of the original segmentation attributes
- different distance metrics
- different numbers of clusters

Data Science is an iterative process…

Assignment 3 & Break-out Rooms

INSEAD

The Business School for the World®

- Assignment: Parts 1 and 2 of MarketSegmentationProcessInClass
- Answer the questions (in Parts 1 and 2 only) in the .Rmd notebook
- BORs: 320-326, 327A, 327B
- I will go around and help with the concepts.
- Varun is available remotely. Email him and he can Skype with you.

Summary of Sessions 9-10

- Derived attributes and dimensionality reduction
  - Principal Component Analysis, how to choose number of factors
  - Then continue analysis on the new attributes

- Clustering and segmentation
  - Create groups of similar observations
  - Hierarchical clustering, K-means clustering

- Template for market segmentation (Assignment 3) for the Boats (A) case

Next…

- Assignment 3 (due Feb 14):
  - Complete the market segmentation process for the Boats (A) case
    - Answer the questions in Parts 1 and 2 of MarketSegmenationProcessInClass.Rmd

- Proposal for Final Project (due Feb 14)
  - A short notebook with description of the business problem, your business solution process, sample of the data, and data dictionary

- Sessions 11-12 [Fri Feb 14]
  - Guest speaker: advanced analytics leader in BCG's Financial Institutions and Insurance practices
  - AI in Business
    - Detailed discussion of specific cases
    - Open Q&A

Final Project (due the day of last class)

INSEAD

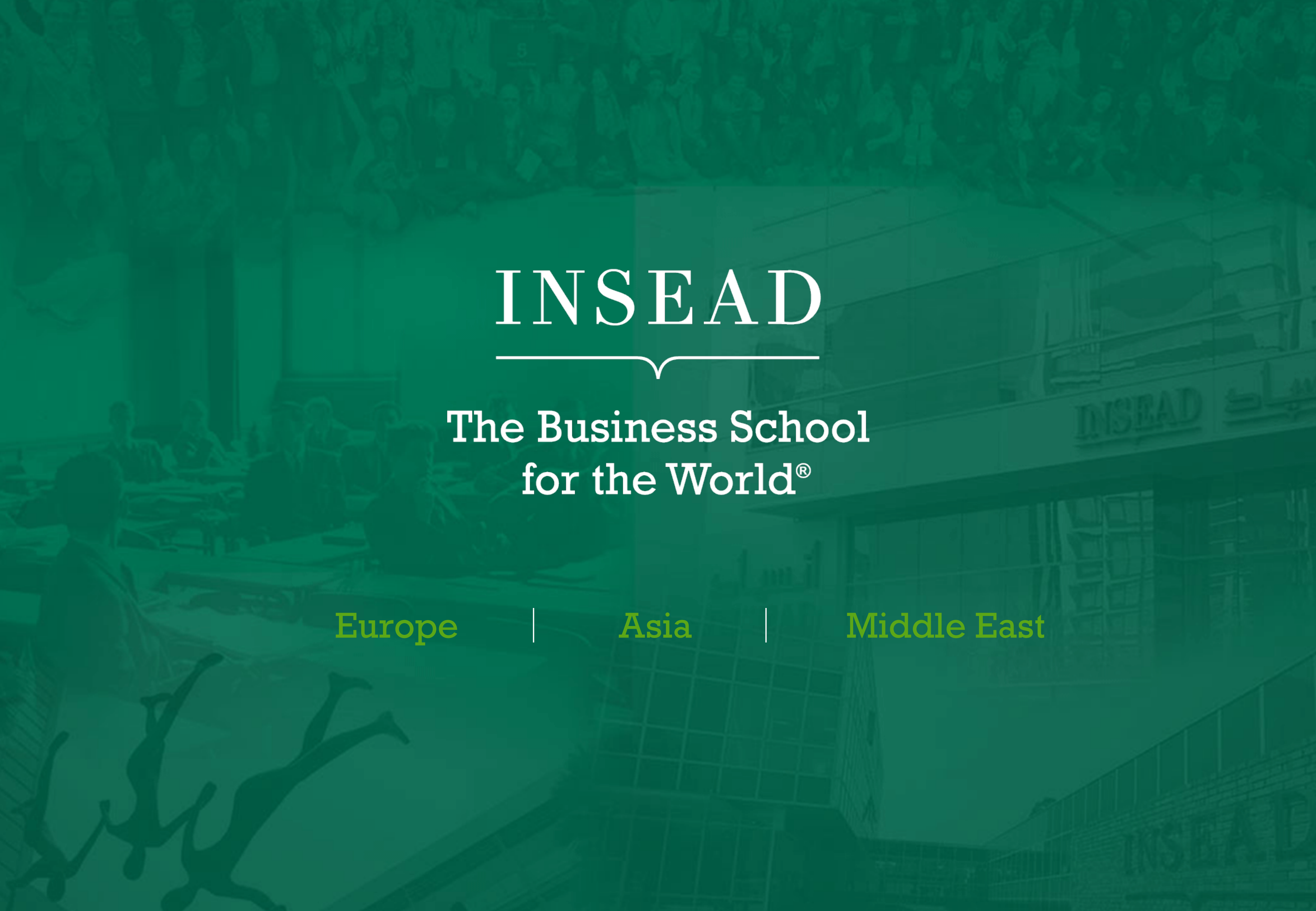
The Business School for the World®

- Develop a data analytics solution to a business problem
  - Relevant business problem, ideally from your past or future workplace
  - Develop a process for how to solve the problem with steps codified in a notebook
  - Show application on a dataset
  - Draw relevant and actionable business insights

- You are expected to share the data you use

- Examples of past projects on GitHub course website

- You will present in class



INSEAD

The Business School for the World®

Europe | Asia | Middle East