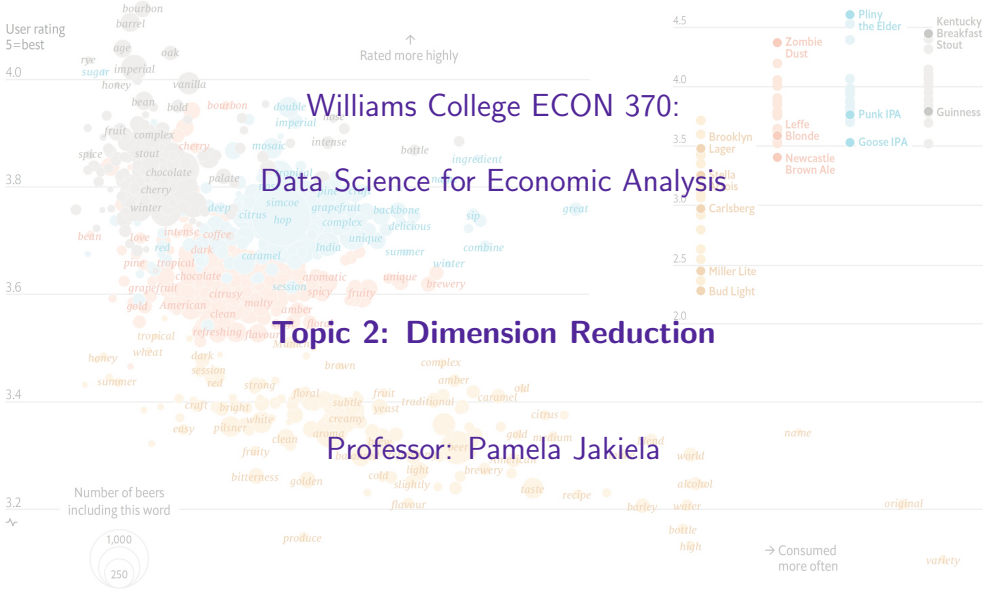


Williams College ECON 370:

Data Science for Economic Analysis

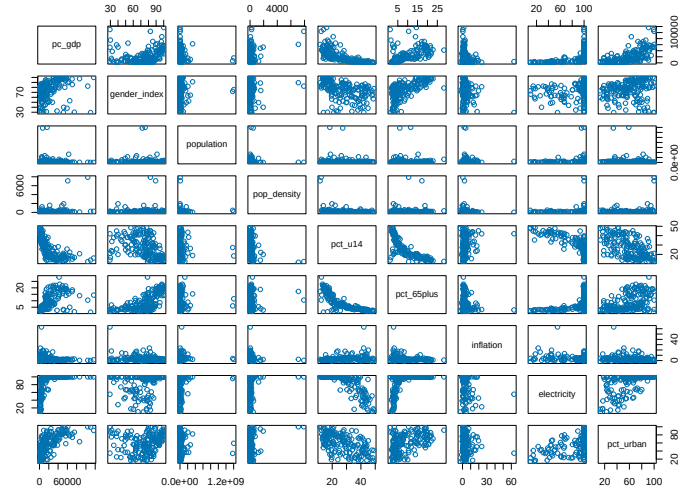
Topic 2: Dimension Reduction

Professor: Pamela Jakiela

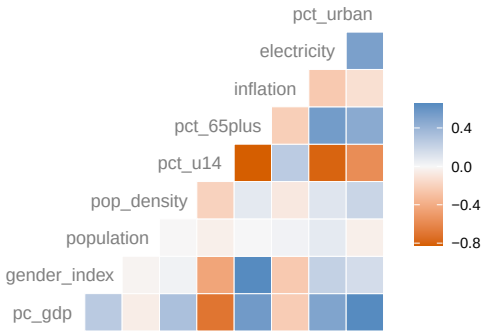
Outline

- Principal components analysis
- k-means clustering

Exploring Relationships Between Many Variables Simultaneously



Exploring Relationships Between Many Variables Simultaneously

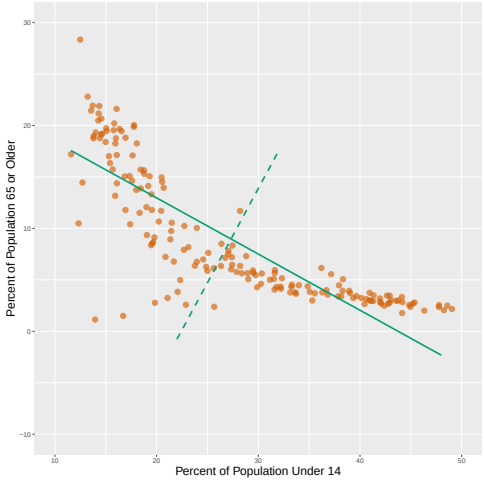


Principal Components Analysis (PCA)

Principal components analysis (PCA) is a **dimension reduction** technique that summarizes the variability in a data set of p variables into $m \leq p$ linear combinations of those variables

- Consider a data set with p variables, $X_1, X_2, \ldots, X_p$
  - PCA is an example of **unsupervised learning** because we analyze X variables with no Y
- Each **principal component** is characterized by a set of **loadings** $\phi_1, \ldots \phi_p$ on the variables
- The **score** is the linear combination that uses the loadings as the weights, $\sum_{j=1}^{p} \phi_j x_{ij}$
- Loadings for the first principal component to maximize the variance of the score
  - Loadings for the second principal component to maximize the variance of the score, but restricting attention to loadings that are orthogonal to the first principal component
- PCA has an analytical solution based on eigendecomposition, but we won't use it

Example: PCA with Two Variables



First principal component loadings:

- $\phi_{11} = -0.877$
- $\phi_{21} = 0.479$

First principal component direction: ϕ_{21}/ϕ_{11}

- Passes through mean X_1, X_2
- Minimizes sum of squared distances
- PC direction and OLS not identical

2nd PC direction is perpendicular to 1st

- $\phi_{12} = -0.479$
- $\phi_{22} = -0.877$

Why Constrain the Loadings?

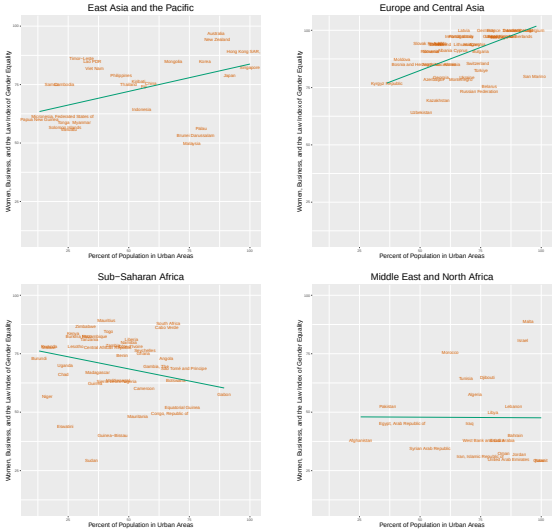
The first principal component is the set of loadings that solves:

$$\max_{\phi_1,\ldots\phi_p} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{p} \phi_j x_{ij} \right)^2 \right] \text{ subject to } \sum_{j=1}^{p} \phi_j^2 = 1$$

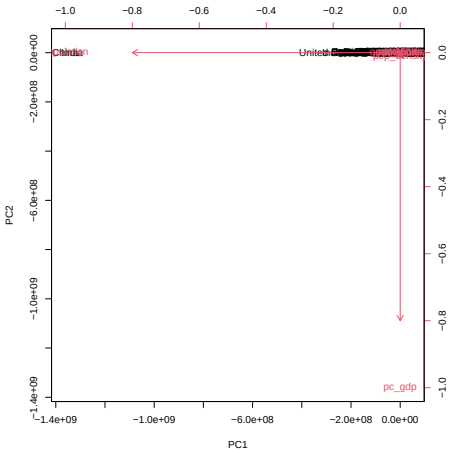
Why do we impose the constraint $\sum_{j=1}^{p} \phi_j^2 = 1$?

- Multiplying loadings by a positive constant would increase the variance of the scores
- Loadings also depend on the variances of the X_j variables

Example: Urbanization and Gender Equality Across Regions



Example: PCA Loadings Are Not Unit Invariant



Variable	PC1	PC2	PC3
pc_gdp	0	-1	-0.012
gender_index	0	0	-0.001
population	-1	0	0
pop_density	0	-0.012	1
pct_u14	0	0	0
pct_65plus	0	0	-0.001
inflation	0	0	0
electricity	0	-0.001	-0.002
pct_urban	0	-0.001	0

When to Scale X Variables

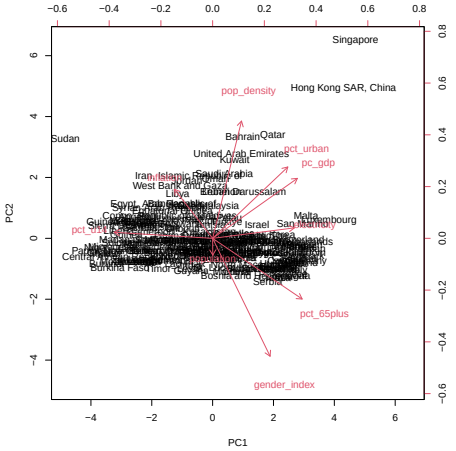
PCA over-weights high-variance variables, and variance is not unit invariant

- Common practice is to scale all variables to be mean 0, SD 1

When all variables are measured in the same units, normalization is not appropriate

- Examples: dummy variables, test items, responses in choice experiments

Example: PCA with Nine Variables



SD	Prop.
1.961	0.245
1.086	0.135
1.019	0.127
0.964	0.120
0.913	0.114
0.723	0.090
0.585	0.073
0.477	0.060
0.286	0.036

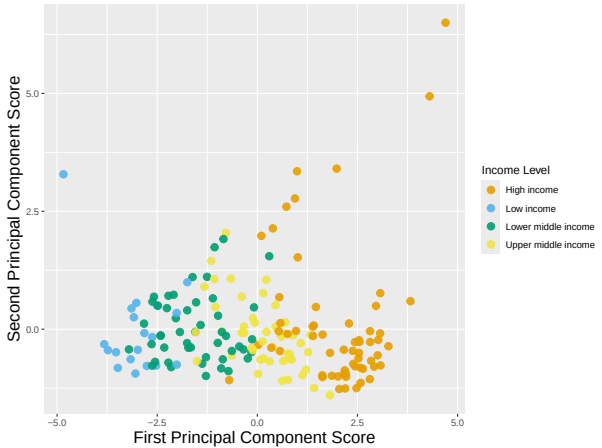
Example: PCA with Nine Variables

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
pc_gdp	0.409	0.290	-0.074	0.018	-0.001	-0.363	0.645	0.401	-0.185
gender_index	0.278	-0.569	-0.168	-0.128	0.470	-0.164	-0.29	0.463	0.088
population	0.001	-0.081	0.907	-0.342	0.067	-0.221	0.020	0.003	-0.012
pop_density	0.140	0.566	-0.133	-0.565	0.474	0.251	-0.162	-0.079	-0.039
pct_u14	-0.476	0.029	-0.105	-0.088	0.011	-0.277	-0.205	0.169	-0.779
pct_65plus	0.433	-0.293	0	0.112	0.239	0.035	0.138	-0.643	-0.472
inflation	-0.185	0.237	0.257	0.670	0.605	0.117	0.015	0.118	0
electricity	0.395	0.051	0.221	0.133	-0.320	0.585	-0.258	0.378	-0.347
pct_urban	0.363	0.345	0.016	0.243	-0.154	-0.543	-0.588	-0.140	0.077

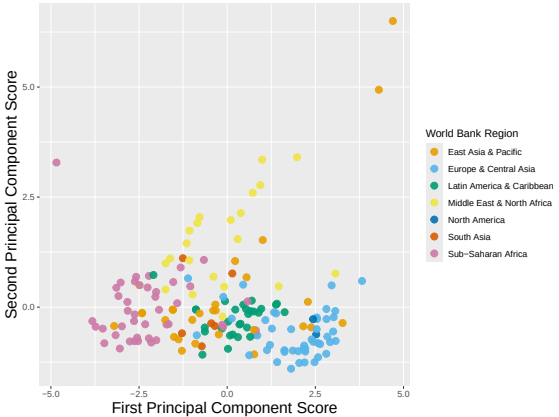
Example: PCA with Nine Variables

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
pc_gdp	0.409	0.290	-0.074	0.018	-0.001	-0.363	0.645	0.401	-0.185
gender_index	0.278	-0.569	-0.168	-0.128	0.470	-0.164	-0.29	0.463	0.088
population	0.001	-0.081	0.907	-0.342	0.067	-0.221	0.020	0.003	-0.012
pop_density	0.140	0.566	-0.133	-0.565	0.474	0.251	-0.162	-0.079	-0.039
pct_u14	-0.476	0.029	-0.105	-0.088	0.011	-0.277	-0.205	0.169	-0.779
pct_65plus	0.433	-0.293	0	0.112	0.239	0.035	0.138	-0.643	-0.472
inflation	-0.185	0.237	0.257	0.670	0.605	0.117	0.015	0.118	0
electricity	0.395	0.051	0.221	0.133	-0.320	0.585	-0.258	0.378	-0.347
pct_urban	0.363	0.345	0.016	0.243	-0.154	-0.543	-0.588	-0.140	0.077

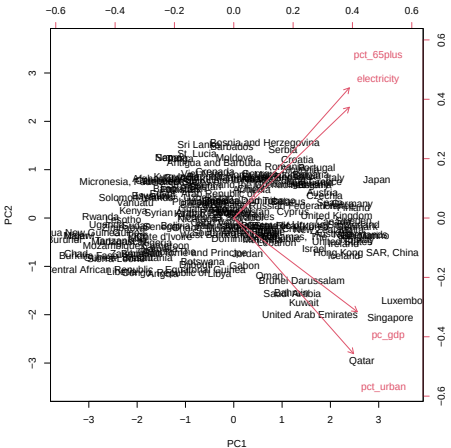
Example: PCA with Nine Variables



Example: PCA with Nine Variables

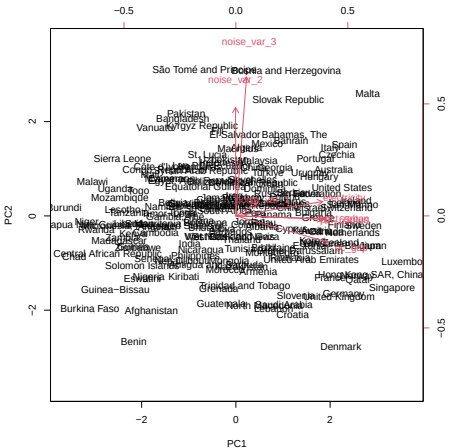


PCA with Sparse Data: Example



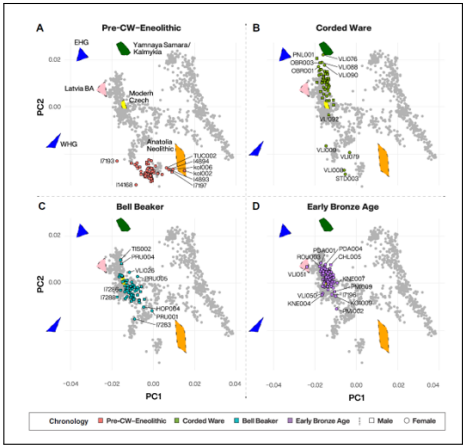
Variable	PC1	PC2	PC3
pc_gdp	0.521	-0.395	0.322
electricity	0.487	0.466	-0.702
pct_65plus	0.487	0.548	0.593
pct_urban	0.505	-0.571	-0.227

PCA with Sparse Data: Adding Noise



Variable	PC1	PC2	PC3
pc_gdp	0.515	-0.147	0.097
electricity	0.484	0.08	-0.031
pct_65plus	0.485	-0.018	-0.036
pct_urban	0.497	-0.015	0.192
noise_var_1	0.12	0.072	-0.787
noise_var_2	-0.002	0.605	0.498
noise_var_3	0.06	0.775	-0.291

PCA in the Wild



Source: Papac et al. (2021)

PCA: Practical Considerations

Preprocessing:

- Which variables should be included?
- Should variables be re-scaled to have a common mean and variance?

Interpretation:

- How many principal components to retain in subsequent analysis
- What do the different principal components represent?

k-Means Clustering

In clustering, we partition the sample into K distinct subsets or **clusters**

- Objective is to identify unobserved groups of observations
- Clustering is an **unsupervised learning** technique: no outcome variable to predict
- Groups are chosen by finding observations that look similar in terms of their X variables
- In practice, we must define "similar" in terms of some distance metric (e.g. Euclidean)

Examples of clustering in economics:

- Identifying markets and/or competitors, fields of study using text descriptions
- Identifying types of individuals, e.g. patterns of behavior in lab experiments

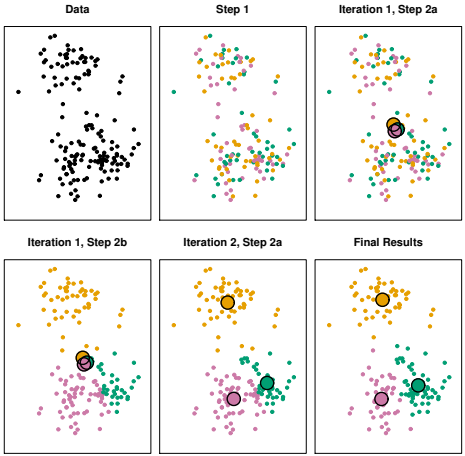
k-Means Clustering Minimizes (?) the Distance Between Observations

- Distance between $x_a = (x_{a1}, \ldots x_{ap})$ and $x_b = (x_{b1}, \ldots x_{bp})$: $d_{ab} = \sqrt{\sum_{j=1}^{p}(x_{aj} - x_{bj})^2}$
- Variation within cluster C_k if C_k contains N_{C_k} observations:

$$\sum_{a,b \in C} \sum_{j=1}^{p} (x_{aj} - x_{bj})^2 / N_{C_k} = \sum_{a,b \in C} d_{ab}^2$$

In (other) words, k-means clustering minimizes the sum of squared Euclidean distances across all pairs of points within a cluster, so that variance is between (not within) clusters

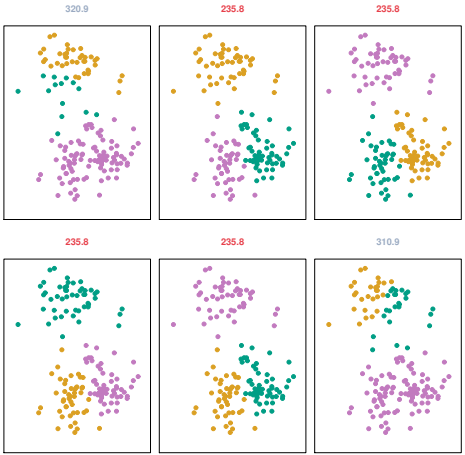
A k-Means Clustering Algorithm



Source: James et al. (2021)

1. Randomly assign initial clusters
2. Compute **centroid** of each cluster
3. Reassign clusters, mapping each observation to the nearest centroid
4. Recalculate centroids
5. Repeat Steps 3 and 4 until process converges, and no observations are reassigned at the next iteration

k-Means Clustering Identifies a Local (Not a Global) Optimum



Source: James et al. (2021)

Example: Identifying a Global Optimum (?)

Iteration 1: seed = 314159

Cluster 1 ($N = 11$)	Cluster 2 ($N = 162$)	Cluster 3 ($N = 2$)
Bangladesh; Brazil; Egypt, Arab Republic of; Indonesia; Japan; Mexico; Nigeria; Pakistan; Philippines; Russian Federation; United States	Afghanistan; Albania; Algeria; Angola; Antigua and Barbuda; Armenia; Australia; Austria; Azerbaijan; Bahamas, The; Bahrain; Barbados; Belarus; Belgium; Belize; Benin...	China; India

Iteration 2: seed = 8675309

Cluster 1 ($N = 162$)	Cluster 2 ($N = 11$)	Cluster 3 ($N = 2$)
Afghanistan; Albania; Algeria; Angola; Antigua and Barbuda; Armenia; Australia; Austria; Azerbaijan; Bahamas, The; Bahrain; Barbados; Belarus; Belgium; Belize; Benin...	Bangladesh; Brazil; Egypt, Arab Republic of; Indonesia; Japan; Mexico; Nigeria; Pakistan; Philippines; Russian Federation; United States	China; India

Example: Characterizing Clusters

Variable	Cluster 1	Cluster 2	Cluster 3
pc_gdp	22,463	19,786	11,507
gender_index	75.21	69.72	73.75
population	15,322,408	182,900,876	1,388,709,532
pop_density	246.92	261.28	305.88
pct_u14	27.37	27.34	22.83
pct_65plus	8.95	9.26	8.77
inflation	3.78	5.5	3.01
electricity	84.14	93.52	97.85
pct_urban	59.1	62.15	46.59

Example: Variables the Distinguish Between Clusters

Rank	Cluster 1	Cluster 2	Cluster 3
1	pc_gdp	pc_gdp	population
2	pct_urban	pct_urban	pop_density
3	gender_index	electricity	electricity
4	pct_u14	pct_u14	gender_index
5	pct_65plus	inflation	pct_65plus
6	inflation	pct_65plus	inflation
7	electricity	gender_index	pct_u14
8	pop_density	pop_density	pct_urban
9	population	population	pc_gdp

Variables are ranked based on the difference in means between cluster j and the average of cluster-level means across all other clusters.

Example: k-Means Clustering in Rescaled Data

Cluster 1 ($N = 41$)	Cluster 2 ($N = 82$)	Cluster 3 ($N = 52$)
Afghanistan; Angola; Benin; Burkina Faso; Burundi; Cameroon; Central African Republic; Chad; Congo, Republic of; Côte d'Ivoire; Eswatini; Gambia, The; Guinea; Guinea-Bissau; Haiti; Kenya; Lesotho; Liberia; Madagascar; Malawi; Mali; Mauritania; Micronesia, Federated States of; Mozambique; Myanmar; Namibia; Niger; Nigeria; Papua New Guinea; Rwanda; São Tomé and Principe; Senegal; Sierra Leone; Solomon Islands; Sudan; Tanzania; Togo; Uganda; Vanuatu; Zambia; Zimbabwe	Algeria; Antigua and Barbuda; Armenia; Azerbaijan; Bahrain; Bangladesh; Barbados; Belize; Bhutan; Bolivia, Plurinational State of; Botswana; Brazil; Brunei Darussalam; Cabo Verde; Cambodia; China; Colombia; Costa Rica; Djibouti; Dominica; Dominican Republic; Ecuador; Egypt, Arab Republic of; El Salvador; Equatorial Guinea; Fiji; Gabon; Georgia; Ghana; Grenada; Guatemala; Guyana; Honduras; India; Indonesia; Iran, Islamic Republic of; Iraq; Jamaica; Jordan; Kazakhstan; Kiribati; Kuwait; Kyrgyz Republic; Lao PDR; Lebanon; Libya; Malaysia; Maldives...	Albania; Australia; Austria; Bahamas, The; Belarus; Belgium; Bosnia and Herzegovina; Bulgaria; Canada; Chile; Croatia; Cyprus; Czechia; Denmark; Estonia; Finland; France; Germany; Greece; Hong Kong SAR, China; Hungary; Iceland; Ireland; Israel; Italy; Japan; Korea; Latvia; Lithuania; Luxembourg; Malta; Montenegro; Netherlands; New Zealand; North Macedonia; Norway; Poland; Portugal; Romania; Russian Federation; San Marino; Serbia; Singapore; Slovak Republic; Slovenia; Spain; Sweden; Switzerland; Ukraine; United Kingdom; United States...

Example: k-Means Clustering in Rescaled Data

Rank	Cluster 1	Cluster 2	Cluster 3
1	pct_u14	electricity	pct_65plus
2	inflation	population	pc_gdp
3	population	pct_urban	gender_index
4	pop_density	pop_density	pct_urban
5	gender_index	inflation	electricity
6	pc_gdp	pct_u14	pop_density
7	pct_urban	pc_gdp	population
8	pct_65plus	gender_index	inflation
9	electricity	pct_65plus	pct_u14

Variables are ranked based on the difference in means between cluster j and the average of cluster-level means across all other clusters.

Example: k-Means Clustering in Rescaled Data

Rank	Cluster 1	Cluster 2	Cluster 3
pc_gdp	-0.849	-0.187	0.964
gender_index	-0.461	-0.354	0.921
population	-0.129	0.128	-0.100
pop_density	-0.177	-0.059	0.233
pct_u14	1.365	-0.019	-1.046
pct_65plus	-0.922	-0.397	1.352
inflation	0.470	-0.060	-0.276
electricity	-1.601	0.416	0.606
pct_urban	-0.936	-0.023	0.774

How Many Clusters?

Diagnostics for assessing the quality of a partition into clusters:

- Stability of the global maximum: does changing the seed lead to different clusters?
  - How different? We often observe clusters that are mostly but not completely stable, suggesting that there is a global maximum that is relatively flat near the peak
- How much of the total sum of squared distances between observations is between clusters?
  - Between sum of squares increases with more clusters, but by how much?
- Does the partition make sense?
  - Unsupervised learning does not usually include explicit validation, but we often have some prior knowledge of the relations between observations that we can use as a sanity check

Cluster Diagnostics

Clusters	Stability	Between SS	Proportion	Marginal
3	4	646	0.413	
4	4	808	0.516	0.103
5	3	963	0.615	0.099
6	3	1059	0.676	0.061
7	2	1149	0.734	0.057
8	2	1195	0.763	0.029
9	4	1233	0.787	0.024
10	1			
12	2	1301	0.831	0.017
16	2	1359	0.868	0.009

Variables are ranked based on the difference in means between cluster j and the average of cluster-level means across all other clusters.

Does the Partition Make Sense?

Cluster 1 ($N = 1$)	Cluster 2 ($N = 82$)	Cluster 3 ($N = 52$)
Sudan	China; India	Hong Kong; Singapore

Cluster 4 ($N = 7$)	Cluster 5 ($N = 22$)	Cluster 6 ($N = 26$)
Bahrain; Brunei Darussalam; Kuwait; Oman; Qatar; Saudi Arabia; United Arab Emirates	Afghanistan; Algeria; Bangladesh; Botswana; Djibouti; Egypt, Arab Republic of; Equatorial Guinea; Eswatini; Gabon; Ghana; Indonesia; Iran, Islamic Republic of; Iraq; Jordan; Lebanon; Libya; Malaysia; Pakistan; São Tomé and Principe; Syrian Arab Republic; Uzbekistan; West Bank and Gaza	Antigua and Barbuda; Barbados; Belize; Bhutan; Cambodia; Fiji; Grenada; Guatemala; Guyana; Honduras; Kiribati; Kyrgyz Republic; Lao PDR; Maldives; Micronesia, Federated States of; Myanmar; Nepal; Nicaragua; Philippines; Samoa; Sri Lanka; St. Kitts and Nevis; St. Lucia; Timor-Leste; Tonga; Viet Nam

Does the Partition Make Sense?

Cluster 7 ($N = 35$)	Cluster 8 ($N = 39$)	Cluster 9 ($N = 41$)
Angola; Benin; Burkina Faso; Burundi; Cameroon; Central African Republic; Chad; Congo, Republic of; Côte d'Ivoire; Gambia, The; Guinea; Guinea-Bissau; Haiti; Kenya; Lesotho; Liberia; Madagascar; Malawi; Mali; Mauritania; Mozambique; Namibia; Niger; Nigeria; Papua New Guinea; Rwanda; Senegal; Sierra Leone; Solomon Islands; Tanzania; Togo; Uganda; Vanuatu; Zambia; Zimbabwe	Australia; Austria; Belgium; Bulgaria; Canada; Croatia; Cyprus; Czechia; Denmark; Estonia; Finland; France; Germany; Greece; Hungary; Iceland; Ireland; Italy; Japan; Korea; Latvia; Lithuania; Luxembourg; Malta; Netherlands; New Zealand; Norway; Poland; Portugal; Romania; San Marino; Serbia; Slovak Republic; Slovenia; Spain; Sweden; Switzerland; United Kingdom; United States	Albania; Armenia; Azerbaijan; Bahamas, The; Belarus; Bolivia, Plurinational State of; Bosnia and Herzegovina; Brazil; Cabo Verde; Chile; Colombia; Costa Rica; Dominica; Dominican Republic; Ecuador; El Salvador; Georgia; Israel; Jamaica; Kazakhstan; Mauritius; Mexico; Moldova; Mongolia; Montenegro; Morocco; North Macedonia; Palau; Panama; Paraguay; Peru; Russian Federation; Seychelles; South Africa; St. Vincent and the Grenadines; Thailand; Trinidad and Tobago; Tunisia; Türkiye; Ukraine; Uruguay

Does the Partition Make Sense?

Cluster 1 ($N = 1$)	Cluster 2 ($N = 82$)	Cluster 3 ($N = 52$)
Sudan	China; India	Hong Kong; Singapore
inflation $+$ gender equality $-$ population under 14 $+$	population $+$	population density $+$ GDP per capita $+$ percent urban $+$

Does the Partition Make Sense?

Cluster 4 ($N = 7$)	Cluster 5 ($N = 22$)	Cluster 6 ($N = 26$)
Bahrain; Brunei Darussalam; Kuwait; Oman; Qatar; Saudi Arabia; United Arab Emirates	Afghanistan; Algeria; Bangladesh; Botswana; Djibouti; Egypt, Arab Republic of; Equatorial Guinea; Eswatini; Gabon; Ghana; Indonesia; Iran, Islamic Republic of; Iraq; Jordan; Lebanon; Libya; Malaysia; Pakistan; São Tomé and Principe; Syrian Arab Republic; Uzbekistan; West Bank and Gaza	Antigua and Barbuda; Barbados; Belize; Bhutan; Cambodia; Fiji; Grenada; Guatemala; Guyana; Honduras; Kiribati; Kyrgyz Republic; Lao PDR; Maldives; Micronesia, Federated States of; Myanmar; Nepal; Nicaragua; Philippines; Samoa; Sri Lanka; St. Kitts and Nevis; St. Lucia; Timor-Leste; Tonga; Viet Nam
GDP per capita $+$ gender equality $-$ percent urban $+$		

Does the Partition Make Sense?

Cluster 7 ($N = 35$)	Cluster 8 ($N = 39$)	Cluster 9 ($N = 41$)
Angola; Benin; Burkina Faso; Burundi; Cameroon; Central African Republic; Chad; Congo, Republic of; Côte d'Ivoire; Gambia, The; Guinea; Guinea-Bissau; Haiti; Kenya; Lesotho; Liberia; Madagascar; Malawi; Mali; Mauritania; Mozambique; Namibia; Niger; Nigeria; Papua New Guinea; Rwanda; Senegal; Sierra Leone; Solomon Islands; Tanzania; Togo; Uganda; Vanuatu; Zambia; Zimbabwe	Australia; Austria; Belgium; Bulgaria; Canada; Croatia; Cyprus; Czechia; Denmark; Estonia; Finland; France; Germany; Greece; Hungary; Iceland; Ireland; Italy; Japan; Korea; Latvia; Lithuania; Luxembourg; Malta; Netherlands; New Zealand; Norway; Poland; Portugal; Romania; San Marino; Serbia; Slovak Republic; Slovenia; Spain; Sweden; Switzerland; United Kingdom; United States	Albania; Armenia; Azerbaijan; Bahamas, The; Belarus; Bolivia, Plurinational State of; Bosnia and Herzegovina; Brazil; Cabo Verde; Chile; Colombia; Costa Rica; Dominica; Dominican Republic; Ecuador; El Salvador; Georgia; Israel; Jamaica; Kazakhstan; Mauritius; Mexico; Moldova; Mongolia; Montenegro; Morocco; North Macedonia; Palau; Panama; Paraguay; Peru; Russian Federation; Seychelles; South Africa...
population under 14 + electricity −	population 65 and over + gender equality +	

k-Means Clustering: Practical Considerations

- Should X variables be normalized or not?
- Which X variables should be included?
- How many clusters? Should they be small or large?
  - Do the clusters identify meaningful groups?
- How many random starts are needed to avoid local maxima? How robust are the clusters?
  - Check for (unexpected) clusters that are very small, exclude outliers or adapt preprocessing
  - Clustering can be very sensitive to small changes in sample construction (so beware)

Lab #2

- pc_gdp
- gender_index
- population
- pop_density
- pct_u14
- pct_65plus
- inflation
- electricity
- pct_urban
- infant_mort
- life_exp
- energy_use
- pc_co2
- trade
- pct_tropical
- temperature
- precipitation

Lab #2

1. Load the data directly from github
2. Create a numeric data frame, name/index rows, normalize (i.e. re-scale) the variables
3. Implement PCA
4. Adapt code for k-means clustering
   - 4.1 Increase the number of clusters as much as possible while maintaining stable clusters
   - 4.2 Analyze the resulting clusters using one of the methods we've discussed
     - Look at the cluster centers
     - Look at the members of each cluster
     - Look at the important variables distinguishing between clusters