Understanding Political Risk Data

COMPSS 224B: Quantitative Political Risk

Iris Malone, PhD and Mark Rosenberg, PhD

April 18, 2025

Announcements

Short response feedback
PS1 due next week
Final assignment released soon

Recap

Where We've Been:

Recap

Where We've Been:

- Quant political risk analysis leans on prediction over inference

Recap

Where We've Been:

- Quant political risk analysis leans on prediction over inference
- 2 classes ML applications: supervised and unsupervised

Recap

Where We've Been:

- Quant political risk analysis leans on prediction over inference
- 2 classes ML applications: supervised and unsupervised
- "Good" model performance aims to minimize cost function $\sum_{i=1}^{n}(y_i - \hat{y}_i)$

Recap

Where We've Been:

- Quant political risk analysis leans on prediction over inference
- 2 classes ML applications: supervised and unsupervised
- "Good" model performance aims to minimize cost function $\sum_{i=1}^{n}(y_i - \hat{y}_i)$

New Terminology:

- Prediction Problem
- Loss function (single); cost function (ensemble)
- MLOps

Agenda

1. Common Data Challenges

2. Data Wrangling and Preparation

3. EDA and Data Mining

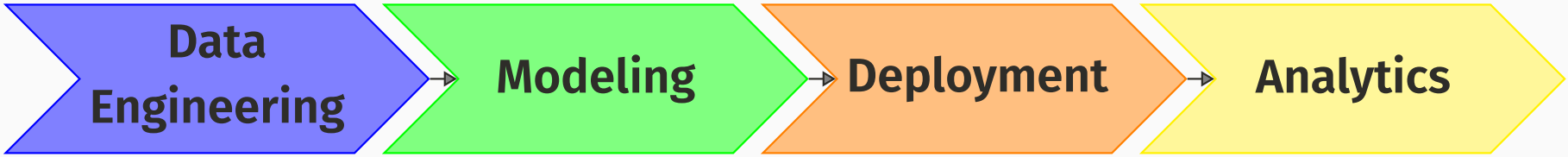
    Principal Component Analysis

    Clustering

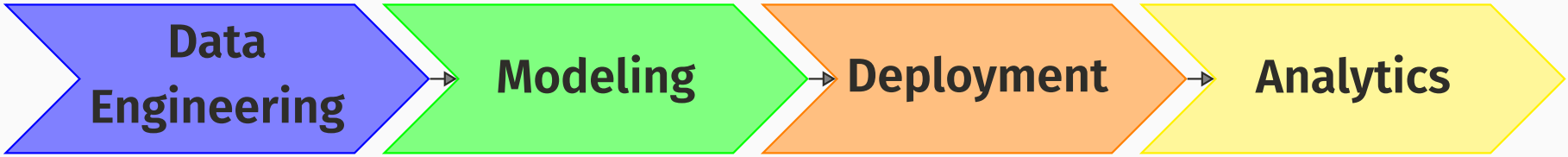
4. Special Topic: Missing Data

Common Data Challenges

Data engineering means understanding your data

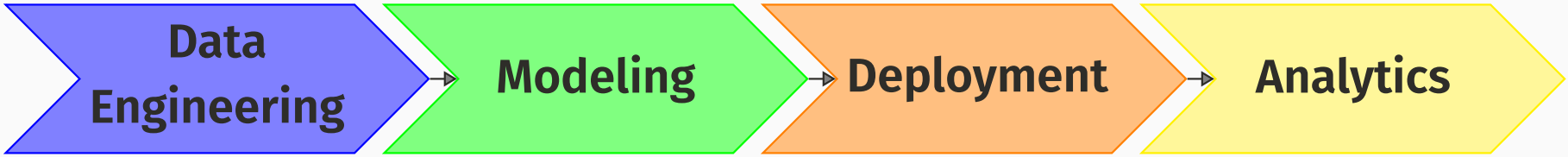


Data engineering means understanding your data



Data engineering is often the most time-intensive part of the pipeline because you have to become an expert on the data.

Data engineering means understanding your data

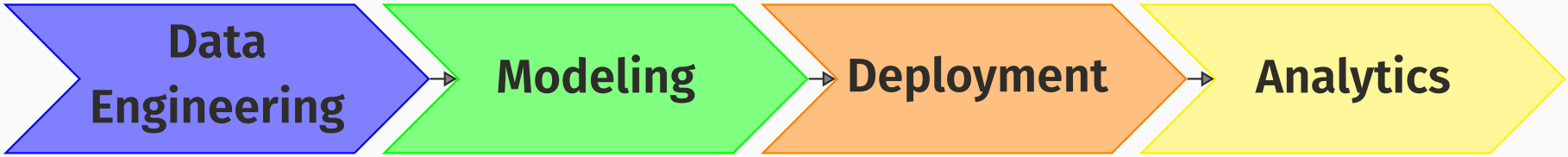


Data engineering is often the most time-intensive part of the pipeline because you have to become an expert on the data.

- **Data Wrangling:** What's in the data?
- **Data Preparation:** Is the data any "good"?
- **Exploratory Data Analysis:** What does the data describe?
- **Data Mining:** What does the data signal?

Figure 1: Be an expert on your data.
Source: skillup

Data engineering means understanding your data



Data engineering is often the most time-intensive part of the pipeline because you have to become an expert on the data.

- **Data Wrangling:** What's in the data?
- **Data Preparation:** Is the data any "good"?

Figure 2: Be an expert on your data.
Source: skillup

Common Data Challenges

Quantitative political risk requires large-n data, but...

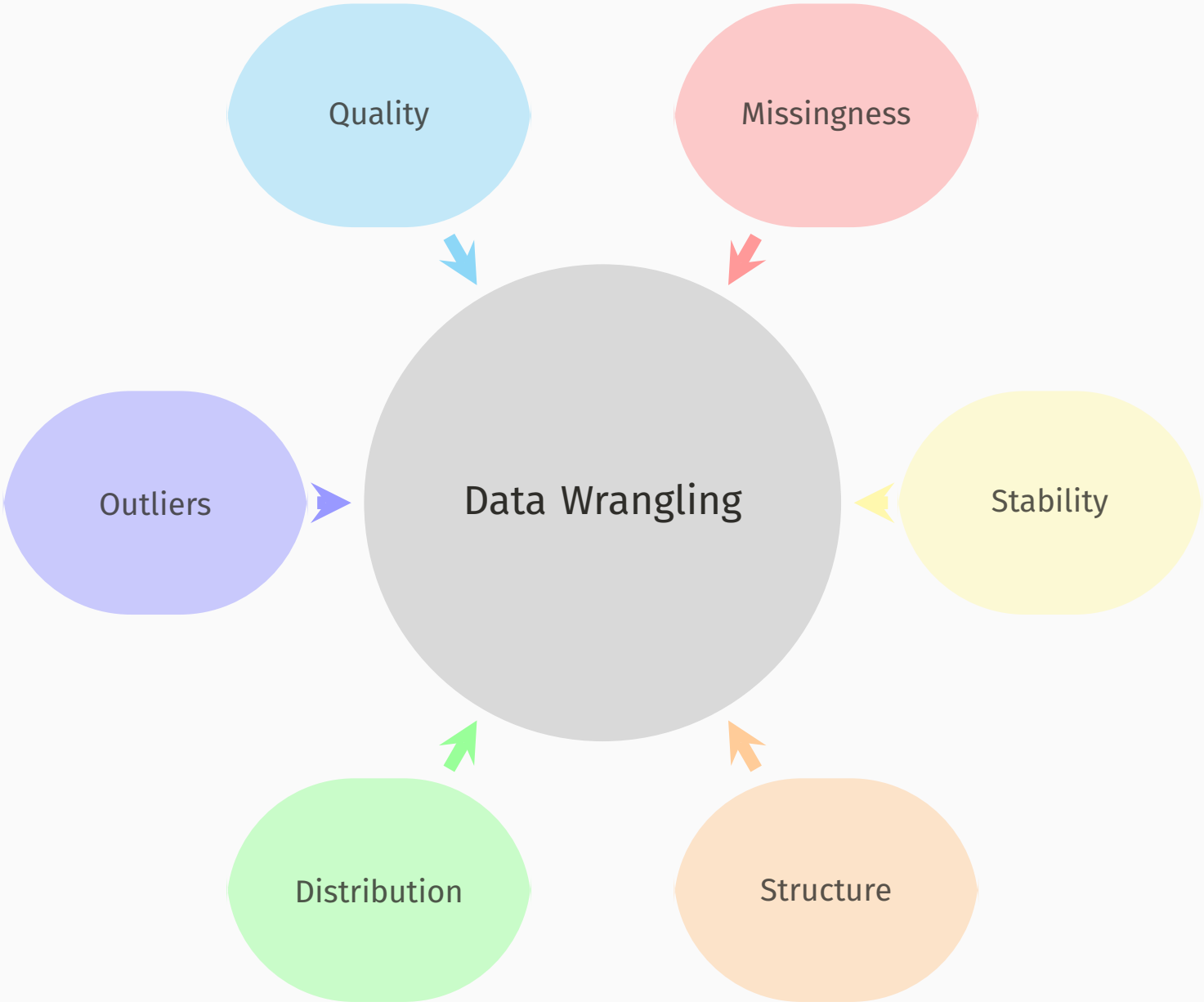
- **Selection bias:** Hard to measure many political concepts (e.g. culture, human rights, corruption)
- **Different DGPs:** observational, survey, experimental
- **Backward-Looking:** Historical and/or rarely updated
- **Missingness:** Non-standardized coverage
- **Quality concerns:** $n = 1$ dataset

Example Quantitative Political Datasets

Example	World Bank Governance Indicators	Social Conflict Analysis Database	Ethnic Power Relations	Eurobarometer
Measure	Governance	Social Conflict	Ethno-religious inequality	Public opinion (varied)
Coverage	Global (204 economies)	Africa and Latin America (minimum 1mn pop.)	Global (minimum 250k pop.)	Europe
Time Series	1996-2023	1990-2011	1946-2021	1974-2023
Updates	Annual (September)	No	Every 5 years (hopefully)	Annual
DGP	Survey	Observational	Observational	Survey
Forecasts	No	No	No	No

Data Wrangling and Preparation

Data Wrangling and Cleaning



What We're Looking For

- **Quality:** Is the data trustworthy? Representative?

What We're Looking For

- **Quality:** Is the data trustworthy? Representative?
- **Missingness:** How much of the data is missing?

What We're Looking For

- **Quality:** Is the data trustworthy? Representative?
- **Missingness:** How much of the data is missing?
- **Outliers:** Are there any outliers in the data?

What We're Looking For

- **Quality:** Is the data trustworthy? Representative?
- **Missingness:** How much of the data is missing?
- **Outliers:** Are there any outliers in the data?
- **Distributions:** Is a variable normal, uniform, exponential, etc. distributed?

What We're Looking For

- **Quality:** Is the data trustworthy? Representative?
- **Missingness:** How much of the data is missing?
- **Outliers:** Are there any outliers in the data?
- **Distributions:** Is a variable normal, uniform, exponential, etc. distributed?
- **Stability:** Is the data stable over time or does it change? Do the measurements look consistent across units?

What We're Looking For

- **Quality:** Is the data trustworthy? Representative?
- **Missingness:** How much of the data is missing?
- **Outliers:** Are there any outliers in the data?
- **Distributions:** Is a variable normal, uniform, exponential, etc. distributed?
- **Stability:** Is the data stable over time or does it change? Do the measurements look consistent across units?
- **Structure:** How is the data organized? How do different indicators relate to each other?

Data Cleaning Solutions

Diagnostic	Solution

Data Cleaning Solutions

	Diagnostic	Solution
Quality	Case Study	Social Science Expertise

Data Cleaning Solutions

	Diagnostic	Solution
Quality	Case Study	Social Science Expertise
Missingness	Matrix plots Frequency tables	Delete Ignore Impute Improve

Data Cleaning Solutions

	Diagnostic	Solution
Quality	Case Study	Social Science Expertise
Missingness	Matrix plots Frequency tables	Delete Ignore Impute Improve
Outliers	Box/violin plot IQR Z-Score Anomaly detection	Investigate Retain Scope

Data Cleaning Solutions

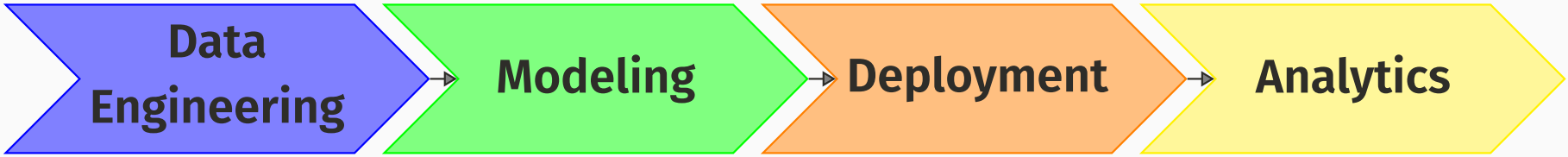
	Diagnostic	Solution
Quality	Case Study	Social Science Expertise
Missingness	Matrix plots Frequency tables	Delete Ignore Impute Improve
Outliers	Box/violin plot IQR Z-Score Anomaly detection	Investigate Retain Scope
Distribution	Histogram Shapiro-Wilk Kolmogorov-Smirnov	Transform Modeling Choices: Non-Linear Non-Parametric

Data Cleaning Solutions

	Diagnostic	Solution
Quality	Case Study	Social Science Expertise
Missingness	Matrix plots Frequency tables	Delete Ignore Impute Improve
Outliers	Box/violin plot IQR Z-Score Anomaly detection	Investigate Retain Scope
Distribution	Histogram Shapiro-Wilk Kolmogorov-Smirnov	Transform Modeling Choices: Non-Linear Non-Parametric
Stability	Correlation Heatmap Line Plots ACF Stationarity tests	Detrend Difference Modeling Choices: AR, ETS, MA

EDA and Data Mining

The fun stuff: EDA and Data Mining



- **Data Wrangling:** What's in the data?
- **Data Preparation:** Is the data any "good"?

- **Exploratory Data Analysis:** What does the data describe?
- **Data Mining:** What does the data signal?

Figure 3: Be an expert on your data.
Source: skillup

Unsupervised Learning for Data Mining

Unsupervised learning is common approach for exploratory analysis, data mining, and index creation.

- Principal Component Analysis (PCA)
  - New risk indices
  - Alpha generation input for predictive modeling
  - Anomaly detection tools (product and analytics)
- Clustering
  - Early warning system (e.g., Political Instability Task Force)
  - Market and country segmentation (e.g., fragility dashboard)
- Text Analysis
  - New risk indices
  - New data series
  - Early warning models

EDA Toolkit

	Univariate	Multivariate
Qualitative	Case Study	Case Study
Descriptive	Five number summary Frequency table	Correlations Dimensionality Reduction Factor Analysis Clustering
Visual	Histogram Violin or box plot Matrix plot Aggregation plot	Scatter Plot Heatmaps Spider chart

Unsupervised Learning Methods

- **Principal Component Analysis (PCA)**
- Clustering
- Text Analysis

Why PCA? Explosion of Big Data

- $p >> n$ is pretty common due to experimental advances, cheaper computers, explosion unstructured data
- But high-p creates curse of dimensionality problems (high noise, risk overfitting, false positives)
- Need method to summarize high-dimensional data

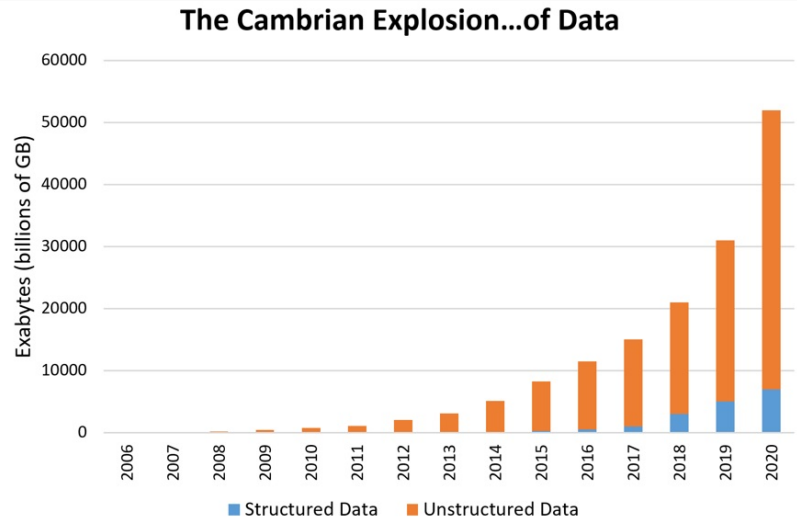


Figure 4: Source: Patrick Chessman

Principal Component Analysis

- This is the most popular unsupervised procedure ever
- First theorized by Karl Pearson (1901)
- Developed by Harold Hotelling (1933)
- Provides a way to visualize and summarize information about high-dimensional data

Principal Component Analysis

Main Idea: Define a small set of M dimensions which summarize the information in all p predictors.

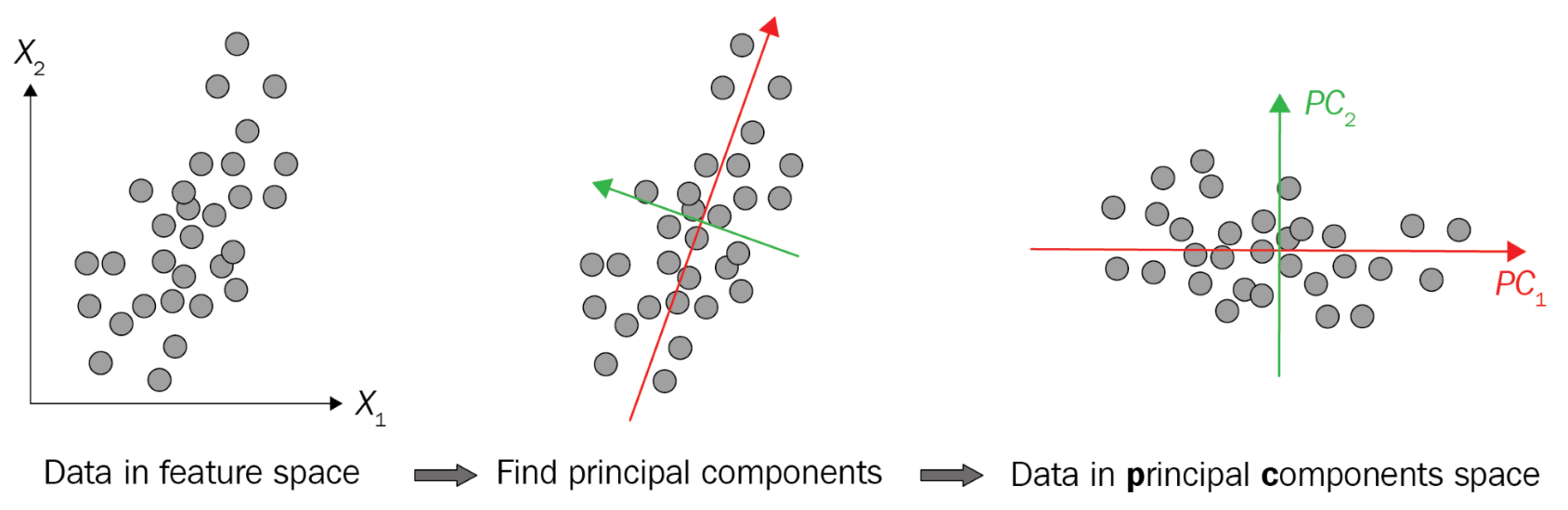


Figure 5: Example Dimensionality Reduction

Principal Components

- Each of the n observations live in p-dimensional space, but not all dimensions equally interesting.
- PCA seeks to find the most interesting dimensions, meaning the dimensions with the largest amount of variation among observations

Figure 6: Can't. Look. Away. (Source: Pixar Post)

Principal Components Procedure

Procedure:

Pre-process the data Identify similarities between groups of predictors $X_1, X_2, \ldots, X_p$Transform groups of predictors into M linear combination known as **principal component** Z

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

Pre-Processing the Data: Centering and Scaling

- Centering is key to ensure dimensions look at variance and not mean of predictors

Pre-Processing the Data: Centering and Scaling

- Centering is key to ensure dimensions look at variance and not mean of predictors
- Scale the variance to have mean zero and look for the linear combination with the largest sample variance

Pre-Processing the Data: Centering and Scaling

- Centering is key to ensure dimensions look at variance and not mean of predictors
- Scale the variance to have mean zero and look for the linear combination with the largest sample variance
- Why?
  - Interpretability: Scaling is key to good interpretation
  - Skewness: Unscaled data means the PCA loading vector will have a very large loading for the variable with the highest variance

Transform Groups Predictors

Procedure:

- Pre-process the data
- **Identify similarities between groups of predictors** $X_1, X_2, \ldots, X_p$
- Transform groups of predictors into M linear combination known as **principal component** Z

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

Transform Groups Predictors

Procedure:

- Pre-process the data
- Identify similarities between groups of predictors $X_1, X_2, \ldots, X_p$
- **Transform groups of predictors into M linear combination known as principal component Z**

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

Principal Component Characteristics

- Solution to an optimization problem where the first two principal components span a plane which is closest to the data
- First and second principal components must be orthogonal (i.e., they don't explain the same types of variation)

PCA Cost Function

Let X be the data matrix with n samples and p variables. From each variable, we subtract the mean of the column (i.e. we center the variables).

PCA Cost Function

Let X be the data matrix with n samples and p variables. From each variable, we subtract the mean of the column (i.e. we center the variables).

To find the first principal component we minimize variance of the n samples projected onto ϕ_1:

$$max_{\phi_{11}\ldots\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2 \qquad \text{subject to} \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

PCA Cost Function

Let X be the data matrix with n samples and p variables. From each variable, we subtract the mean of the column (i.e. we center the variables).

To find the first principal component we minimize variance of the n samples projected onto ϕ_1:

$$max_{\phi_{11}\ldots\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2 \qquad \text{subject to} \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

Projection of the i^{th} sample onto ϕ_1 is the score Z_{i1}

Finding the second principal component

Let X be the data matrix with n samples and p variables. From each variable, we subtract the mean of the column (i.e. we center the variables).

To find the second principal component we solve:

$$max_{\phi_{11}...\phi_{p2}} \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} \phi_{j2} x_{ij})^2 \qquad \text{subject to } \sum_{j=1}^{p} \phi_{j2}^2 = 1 \& \sum_{j=1}^{p} \phi_{j1} \phi_{j2} = 1$$

Loadings

- **Loadings**: Aspects of the vector which passes the closest to a cloud of observations in terms of squared Euclidean distance

Loadings

- **Loadings**: Aspects of the vector which passes the closest to a cloud of observations in terms of squared Euclidean distance
- The loading make up an element of the principal component loading vector ($\phi_1 = (\phi_{11}, \phi_{21}, \dots)$)

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

Loadings

- **Loadings**: Aspects of the vector which passes the closest to a cloud of observations in terms of squared Euclidean distance
- The loading make up an element of the principal component loading vector ($\phi_1 = (\phi_{11}, \phi_{21}, \dots)$)

$$Z_m = \sum_{i=1}^{p} \phi_{jm} X_j$$

- Loadings:
  - Size describes how much a variable contributes to a particular principal component
  - Sign explains correlation between elements

What is the first principal component?

The first principal component (the most interesting dimension) is the vector which passes the closest to a cloud of samples in terms of squared **Euclidean distance**.

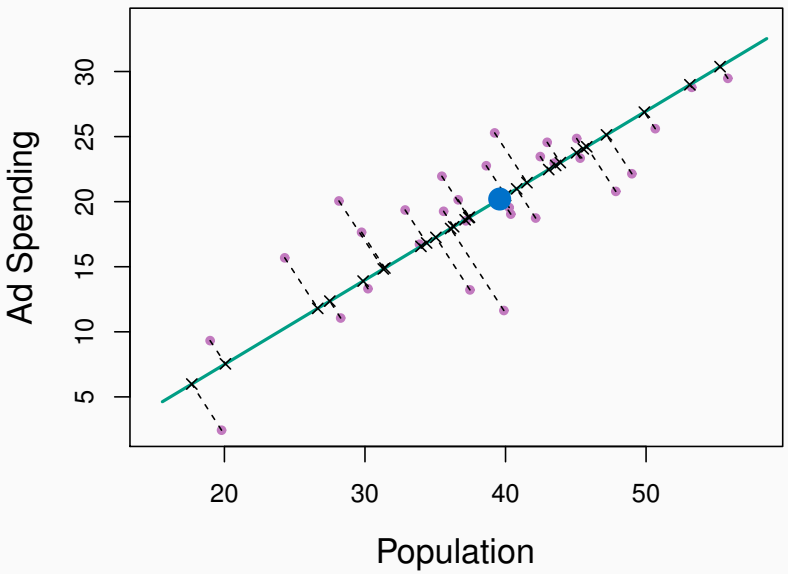
$$d\left(x_{i'j}, x_{ij}\right) = \sqrt{\sum_{i=1}^{n}\left(x_{ij} - x_{i'j}\right)^2}$$

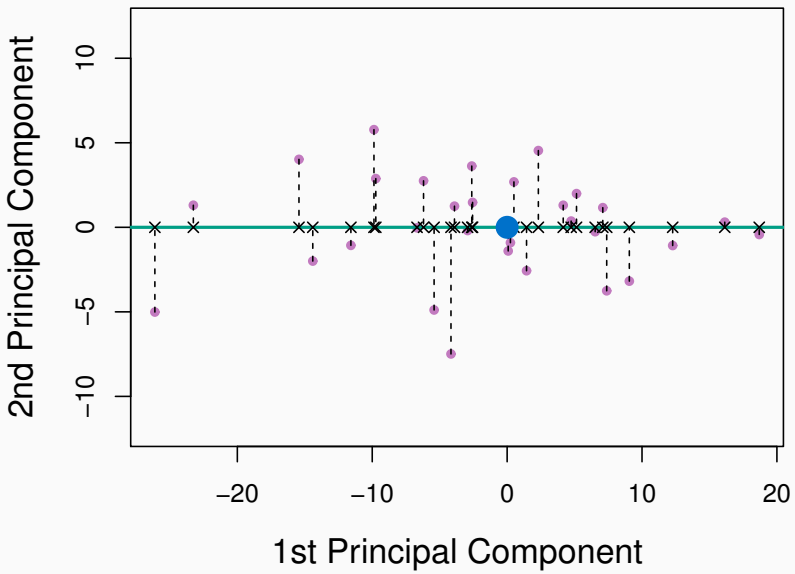
Interpretation:

- We expect some observations to be 'closer' (more similar) to each other
- When distance is small, observation pairs are more similar
- When distance is larger, observation pairs are more dissimilar

Example Euclidean Distance

Vector which passes the closest to a cloud of samples in terms of squared **Euclidean distance**, i.e. the green line minimizing the average squared length of the dotted lines





What does this look like with 3 variables?

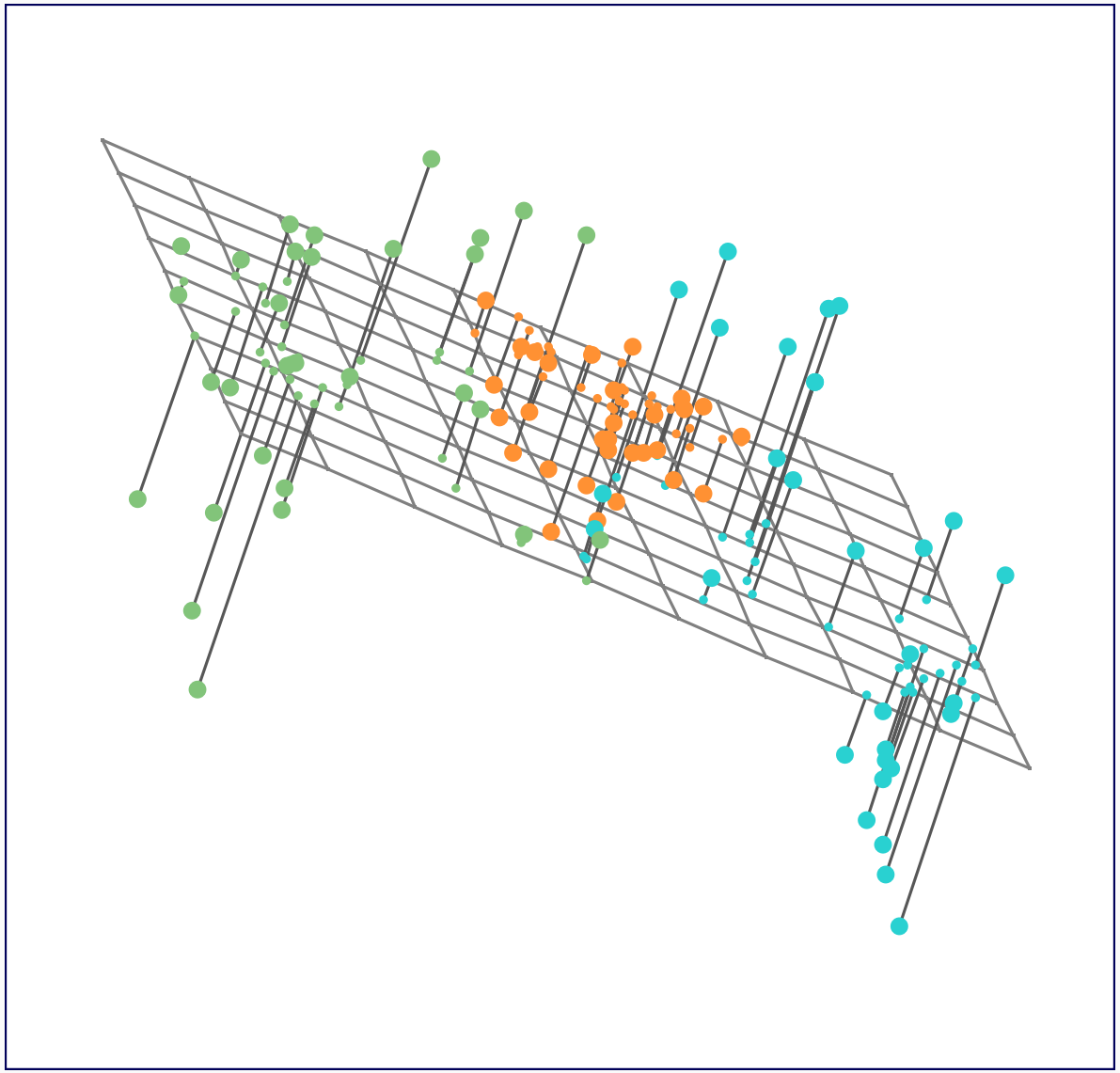


Figure 7: PC spans the plane that best fits the data (like SVM hyperplane)

What is PCA Good For?

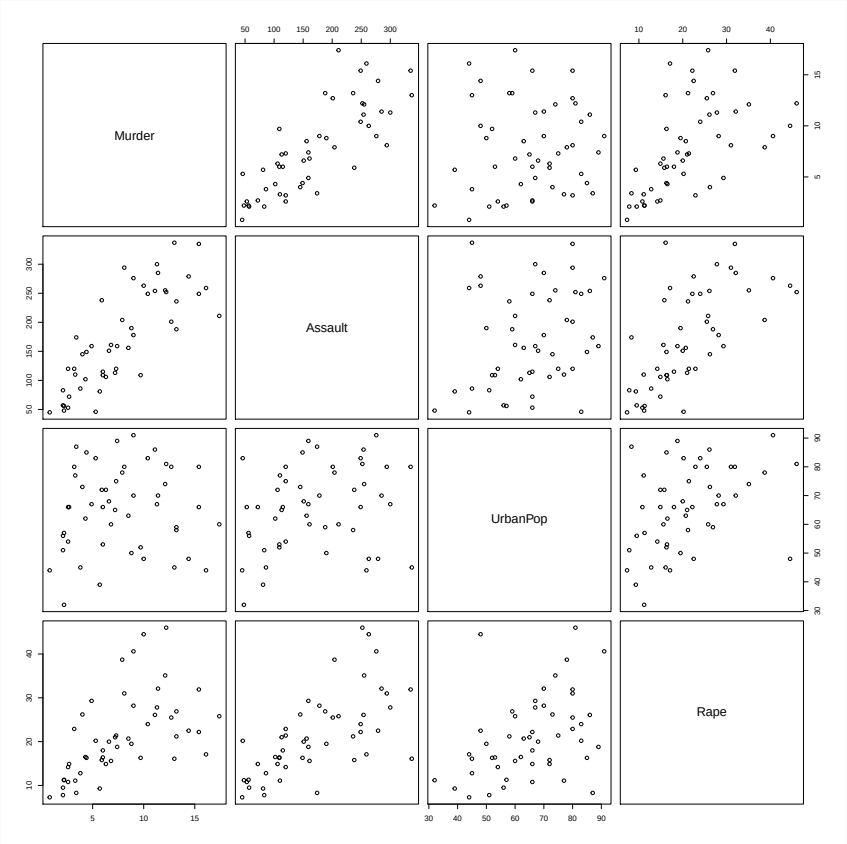
Example: US Arrests data contains info on 3 crime statistics (assault, murder, rape) and population ($p = 4$) for 50 states ($n = 50$).

Potential Research Questions:

- Do crimes correlate with each other?
- Do states with larger urban populations see more crime?

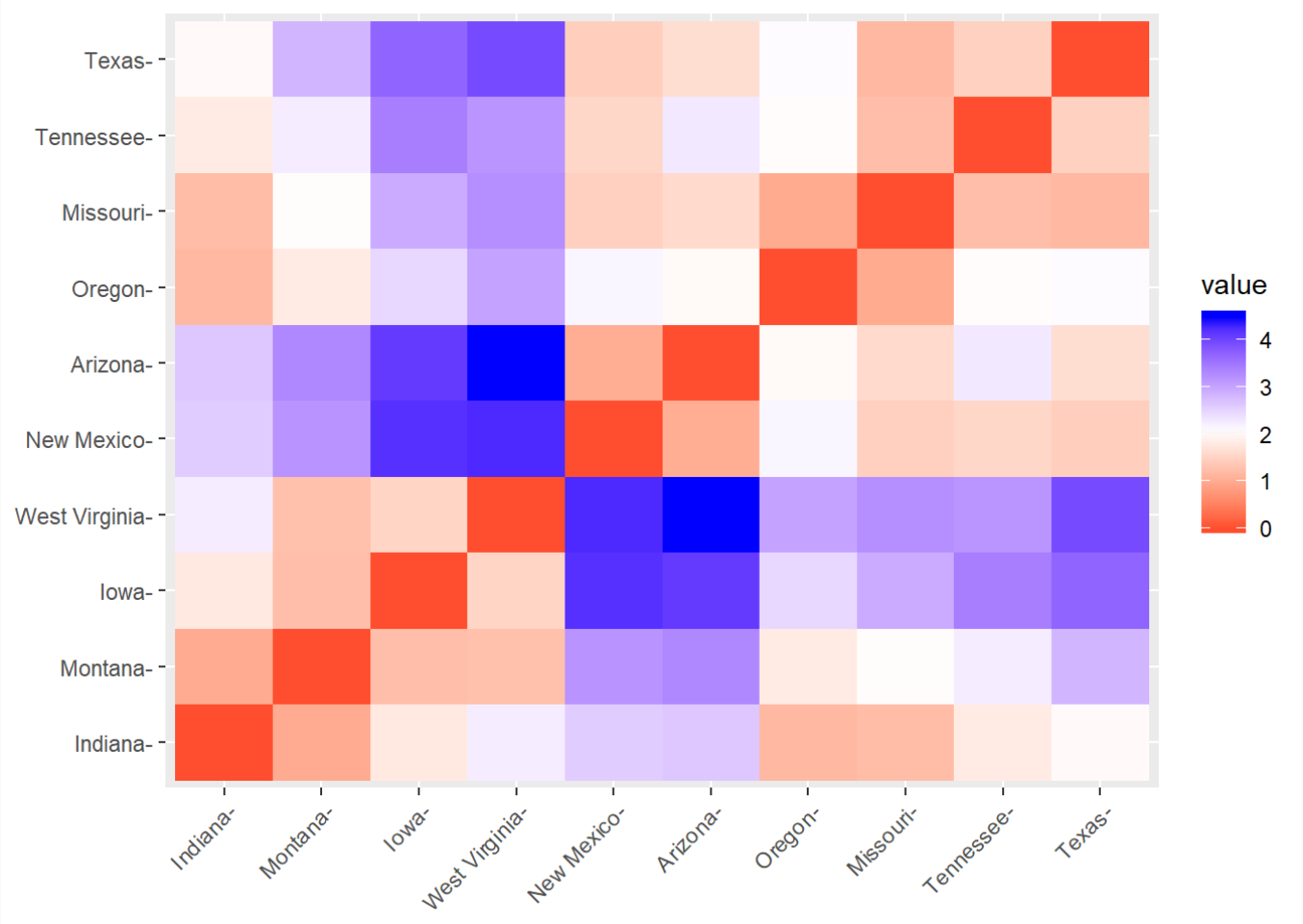
What is PCA Good For?

US Arrests data contains info on 3 crime statistics (assault, murder, rape) and population ($p = 3$) for 50 states ($n = 50$).



Example Euclidean Distance

- Compare how similar states are based on crime statistics
- When distance is small, observation pairs are more similar (red)
- When distance is larger, observation pairs are more dissimilar (blue)



Interp: PC pairs states with similar crime rates

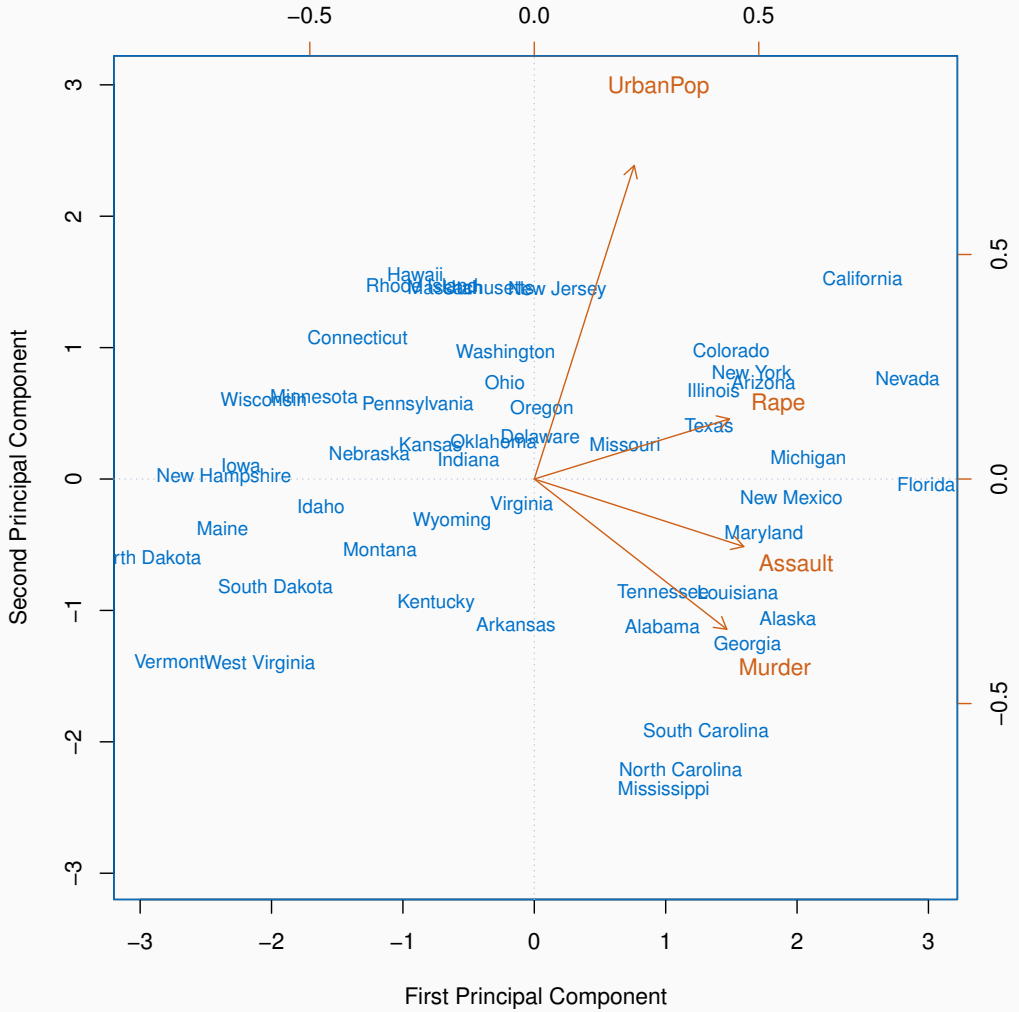
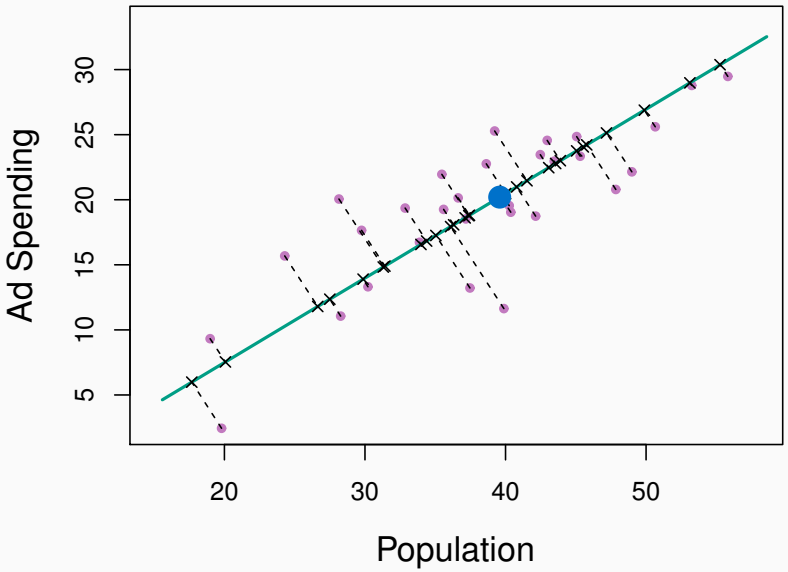
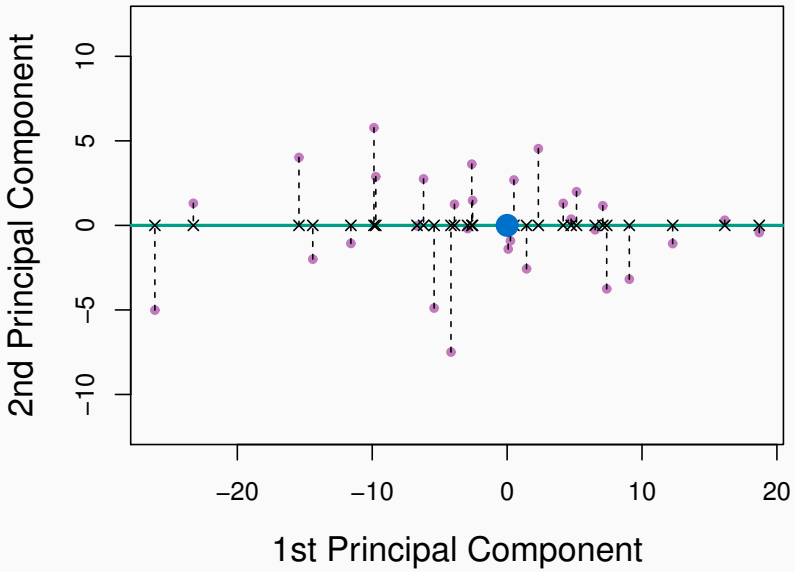


Figure 8: A **biplot** of the first two principal component scores and loading vectors.

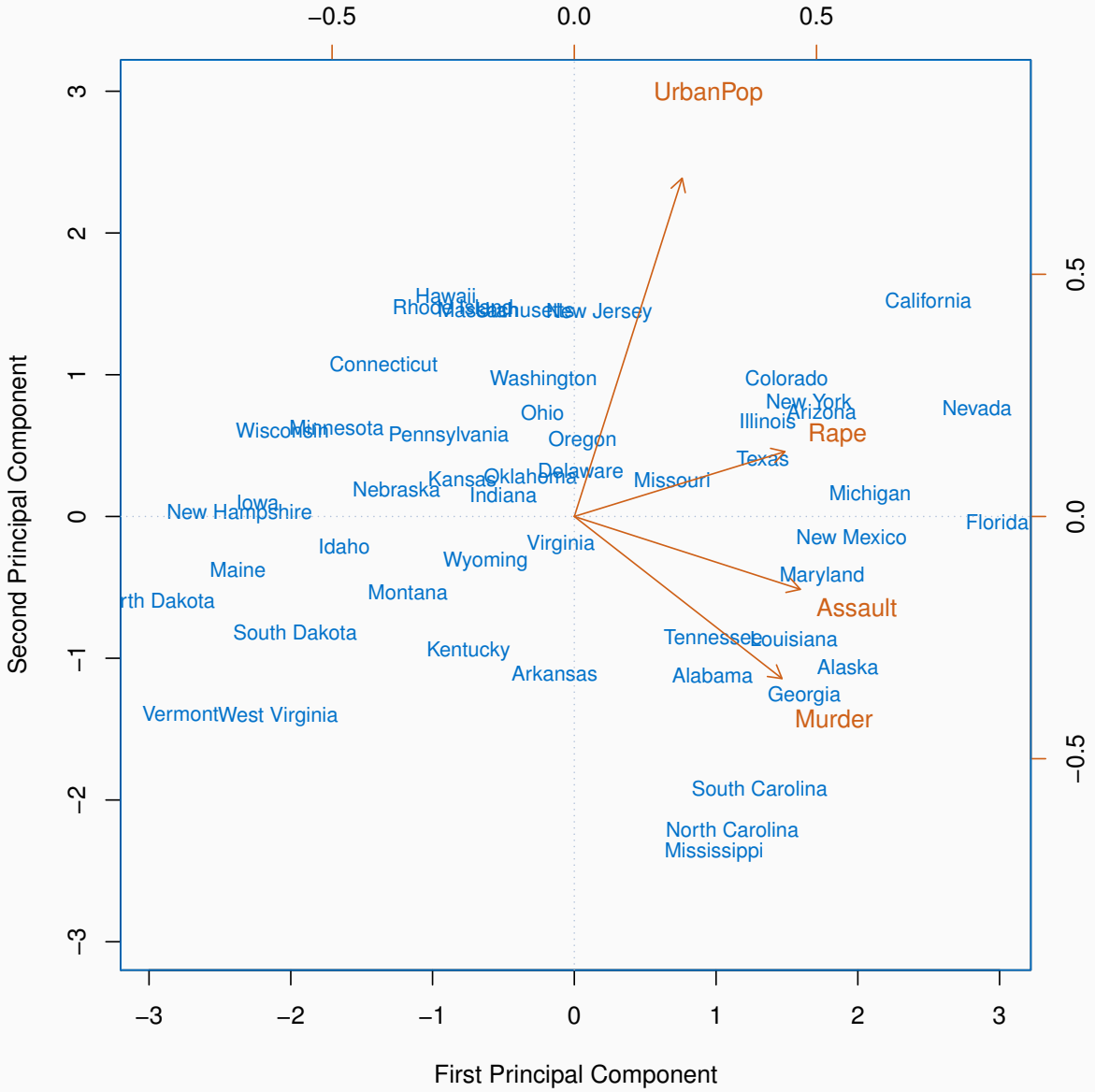
A second interpretation

Another way to explain the first principal component is that it is the dimension with the highest variance between variables



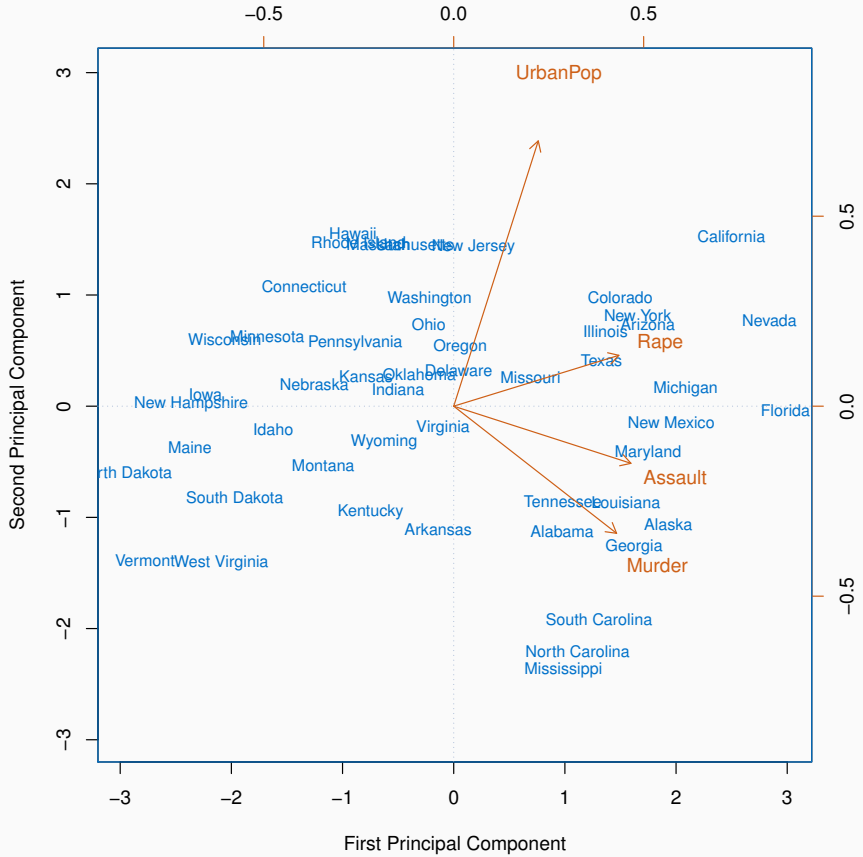


Interp: PC explains variation in crime levels



Example: USArrests Interpretation

- States with high levels of rape also have high levels of assault and murder
- Urban population is orthogonal (unrelated) to crime rates
- Different types of states have different crime rates



How many principal components are enough?

Rule of Thumb: 2 Principal Components capture most of the relevant information.

How many principal components are enough?

Rule of Thumb: 2 Principal Components capture most of the relevant information.

More Precise Answer:

- **Proportion of Variance Explained** (PVE): tells us sum of the variance explained by the m-th principal component over the total variance
- Can assess how much variation in data: low PVE = noisy data; high PVE = highly separable data

Proportion of Variance Explained

- First principal component explains the direction in space in which the data vary the most

Proportion of Variance Explained

- First principal component explains the direction in space in which the data vary the most
- Second principal component explains the direction in space in which the data vary the second most, etc.

Proportion of Variance Explained

- First principal component explains the direction in space in which the data vary the most
- Second principal component explains the direction in space in which the data vary the second most, etc.
- Total variance of the score vectors is the same as the total variance of the original variables:

$$\sum_{i=1}^{p} \frac{1}{n} \sum_{j=1}^{n} z_{ji}^2 = \sum_{k=1}^{p} Var(x_k)$$

Proportion of Variance Explained

The variance of the m^{th} score variable is:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{jm}x_{ij})^2$$

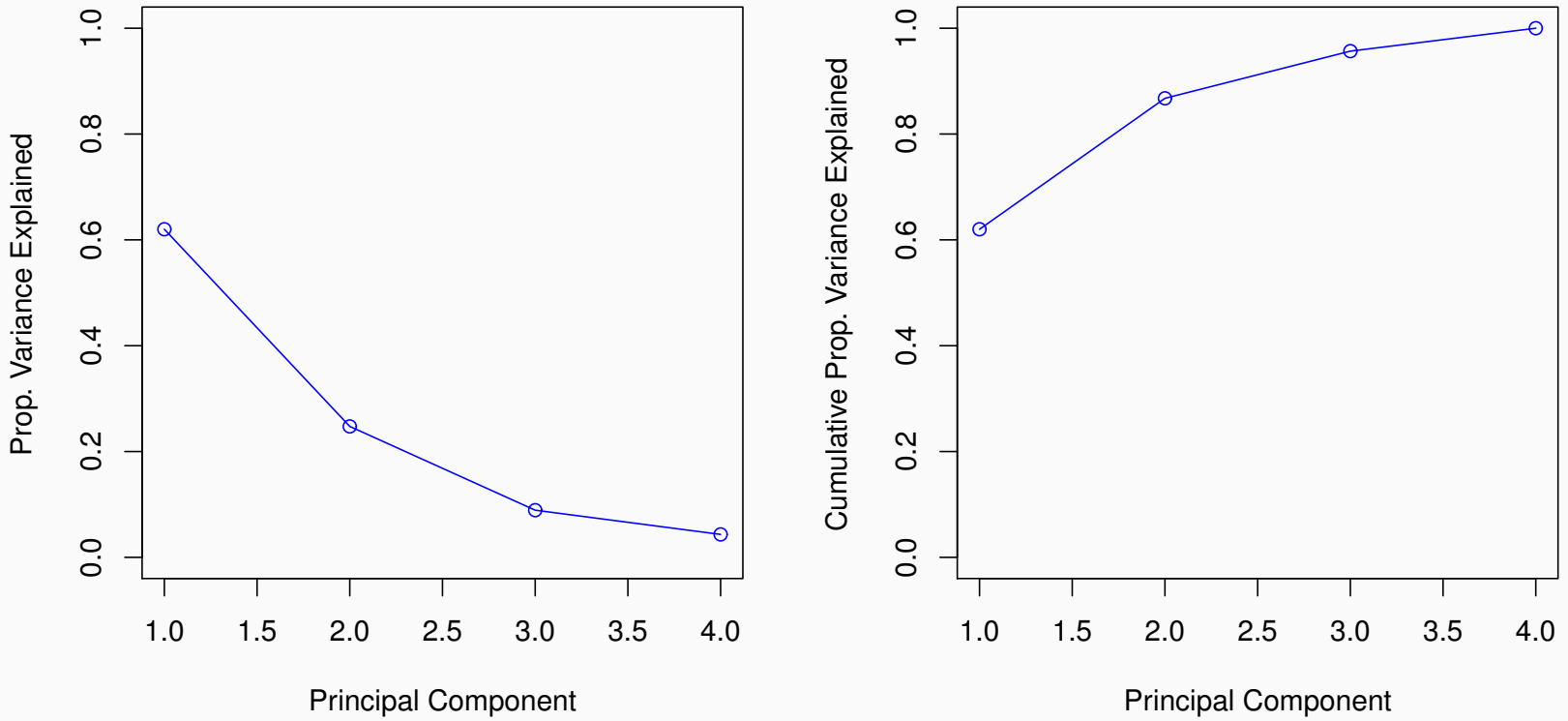


Figure 10: Scree plot showing proportion variance explained by each principal component

Advantages and Disadvantages to PCA

Advantages

Disadvantages

Advantages and Disadvantages to PCA

Advantages

- Fast and easy
- Makes no assumptions about the underlying structure → no human bias
- Makes no assumption about latent unobserved variables (alternative to factor analysis)

Disadvantages

Advantages and Disadvantages to PCA

Advantages

- Fast and easy
- Makes no assumptions about the underlying structure → no human bias
- Makes no assumption about latent unobserved variables (alternative to factor analysis)

Disadvantages

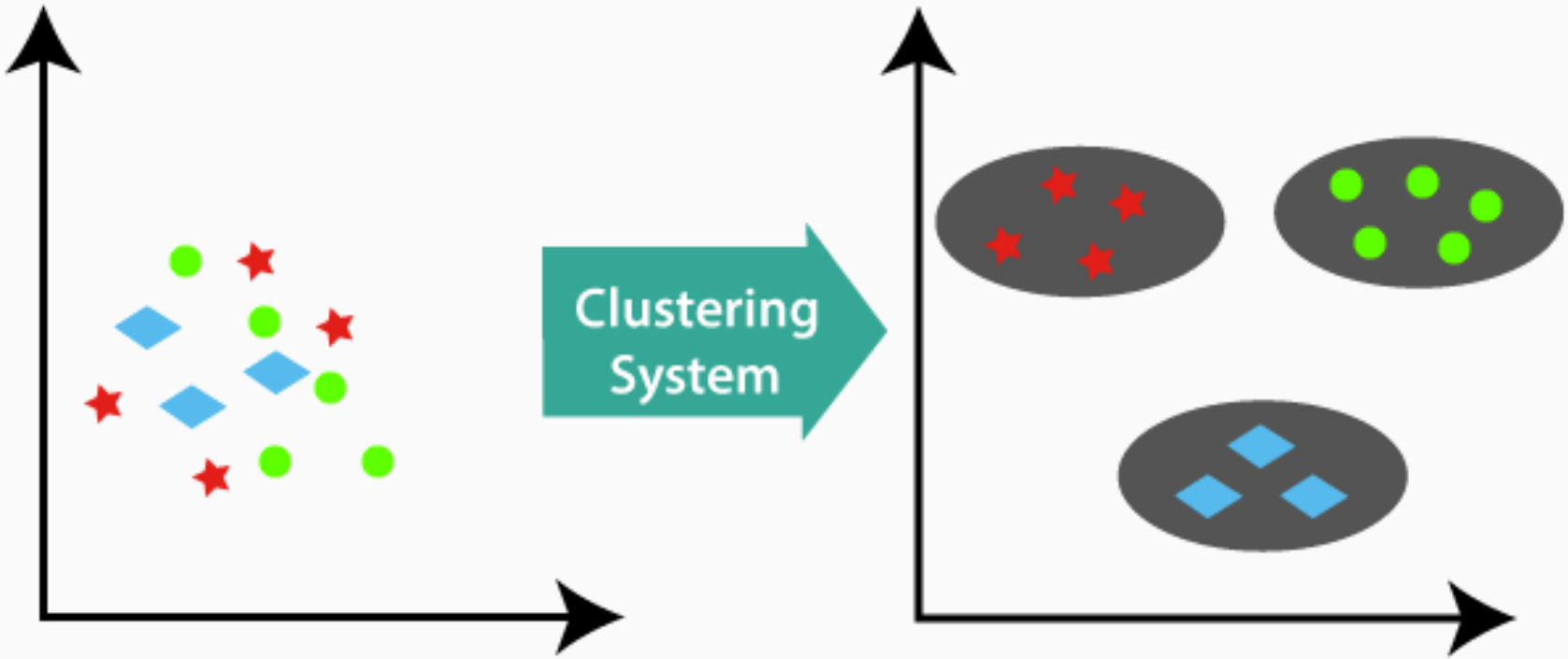
- Simplification → loss of relevant information
- Hard to interpret
- Different (human-chosen) inputs → different PC

Unsupervised Learning Methods

- Principal Component Analysis (PCA)
- **Clustering**
- Text Analysis

Clustering

Main Idea: Partition the data into distinct groups based on intra-group similarities (or inter-group differences)



PCA vs Clustering

- **PCA:** Simplify multiple predictors into small number of principal components to explain variance
- **Clustering:** Find subgroups among observations based on individual or combination of predictors

Types of Clustering

1. **K-Means Clustering:** Partition observations into pre-set number of clusters

Types of Clustering

1. **K-Means Clustering:** Partition observations into pre-set number of clusters
2. **Hierarchical Clustering:** Partition observations, but with no pre-set number of clusters

K-Means Clustering

Main Idea: Partition observations into pre-set number of clusters

K-Means Clustering

Main Idea: Partition observations into pre-set number of clusters

- Goal:
  - Maximize the similarity of samples within each cluster
  - Minimize within-cluster variation
- Cost Function:

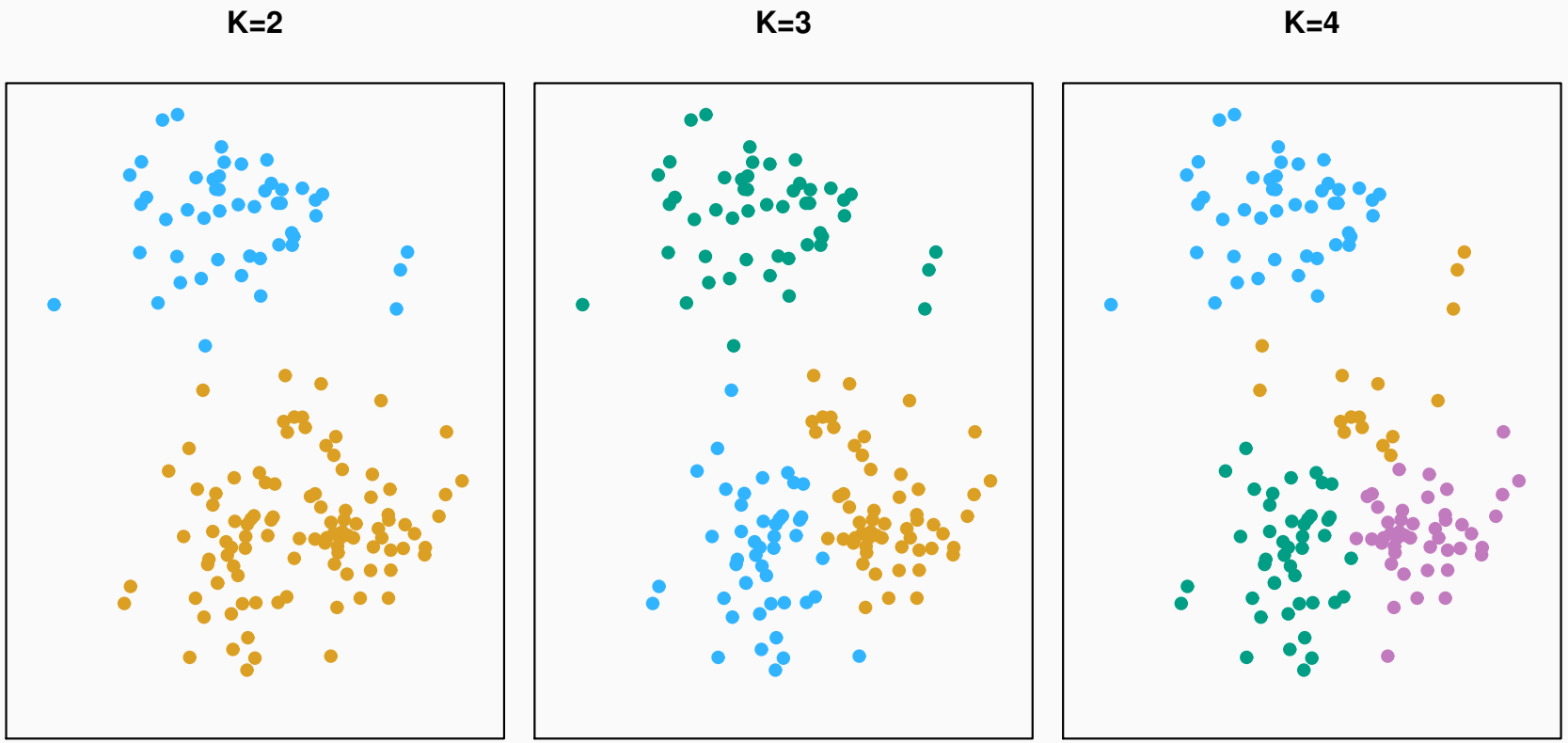
$$min_{C_1,\ldots,C_k} \sum_{l=1}^{K} W(C_l)$$

- $W(C_l)$ is measure of similarity between pairs of observations

$$W(C_l) = \frac{1}{|C_l|} \sum_{i,j \in C_l} \text{Distance}^2(x_i, x_j)$$

- K is the number of clusters and must be fixed in advance
- $\text{Distance}^2(x_i, x_j)$ is Euclidean distance between observations

K-Means Clustering



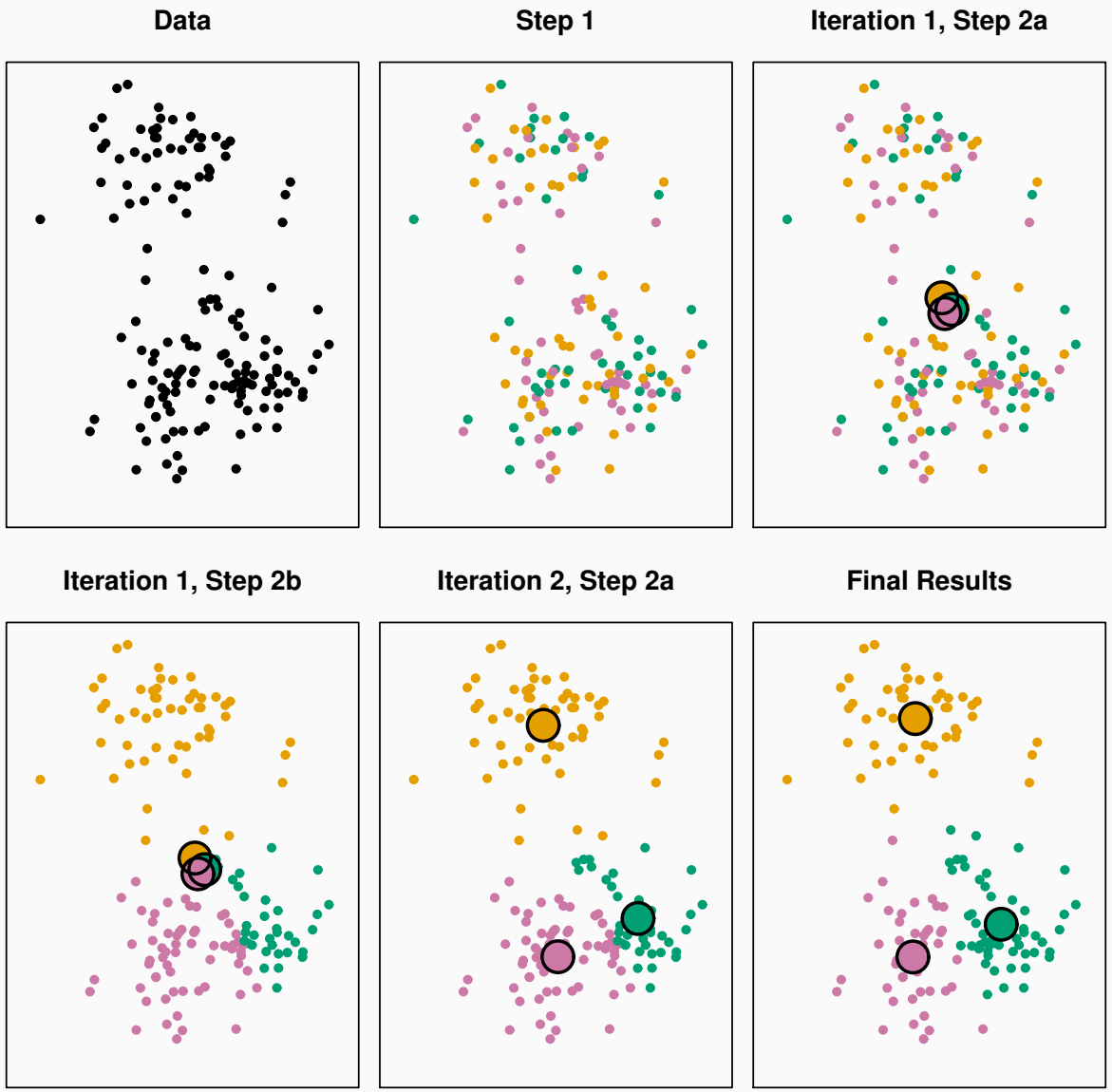
K-Means Procedure

- Assign each sample to a cluster from 1 to K arbitrarily, e.g. at random
- Assign obs. to closest centroid
  - Find the centroid of each cluster , i.e. the average $\bar{x}$ of all the samples in the cluster

$$x_{l,j} = \frac{1}{|C_l|} \sum_{i,j \in C_l} x_{i,j} \text{for} j = 1, \ldots, p$$

  - Reassign each sample to the nearest centroid
- Reposition centroids to new center
- Iterate until centroid position becomes static and local optimum is reached

K-Means Clustering



Advantages and Disadvantages to K-Means Clustering

Advantages

Disadvantages

Advantages and Disadvantages to K-Means Clustering

Advantages

- Guaranteed convergence
- Good scalability for large-p multi-dimensional data
- If large p, then faster than hierarchical clustering

Disadvantages

Advantages and Disadvantages to K-Means Clustering

Advantages

- Guaranteed convergence
- Good scalability for large-p multi-dimensional data
- If large p, then faster than hierarchical clustering

Disadvantages

- Need to make assumption about number of clusters
- What if a sample can belong to more than one cluster?
- Algorithm focuses on minimizing local differences rather than global ones (greedy!)
- Different initializations $\rightarrow$ different results

Correlation Distance

Rule of Thumb: Cluster using Euclidean Distance to measure similarities and dissimilarities

Alternative Approach: Correlation Distance

- Considers 2 observations similar if their features are highly correlated

Correlation Distance

Example: Suppose that we want to cluster countries in the UN for (investing) market segmentation.

- Samples are countries
- Each variable corresponds to a specific country development indicator and measures the amount of economic growth over time

Approaches:

- Euclidean distance would cluster all countries who have the highest growth (quantity)
- But we might want to cluster countries by the type of economic growth (e.g. capital vs labor)to distinguish Developed vs Emerging Markets (category)
- Here, correlation distance may be more appropriate dissimilarity between samples

Correlation Distance

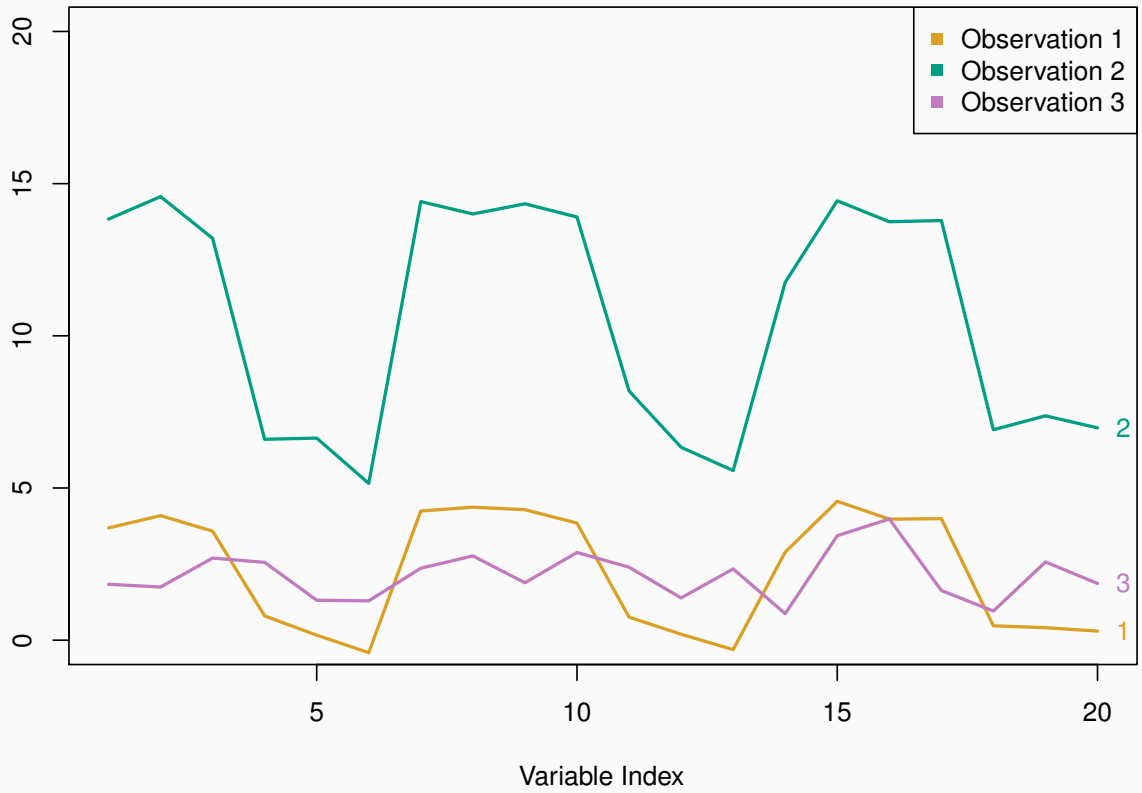


Figure 11: Observations 1 and 3 have similar values so there is small Euclidean distance, but large correlation; observations 1 and have different values so large Euclidean distance, but highly correlated

Special Topic: Missing Data

Problem of Missing Data

Motivation: Country risk data is largely observational. Observational data often has missing data. When there is a large number of predictors, higher likelihood you might be missing at least one predictor.

- What is your household income in the year 2025?
- Have you committed a crime before?
- What was the CO_2 amount of every country in 1990?

Problem of Missing Data

Motivation: Country risk data is largely observational. Observational data often has missing data. When there is a large number of predictors, higher likelihood you might be missing at least one predictor.

- What is your household income in the year 2025?
  Extremely rich people may refuse to answer
- Have you committed a crime before?

- What was the CO_2 amount of every country in 1990?

Problem of Missing Data

Motivation: Country risk data is largely observational. Observational data often has missing data. When there is a large number of predictors, higher likelihood you might be missing at least one predictor.

- What is your household income in the year 2025?
  Extremely rich people may refuse to answer
- Have you committed a crime before?
  There will be consequence if yes.
- What was the CO_2 amount of every country in 1990?

Problem of Missing Data

Motivation: Country risk data is largely observational. Observational data often has missing data. When there is a large number of predictors, higher likelihood you might be missing at least one predictor.

- What is your household income in the year 2025?
  Extremely rich people may refuse to answer
- Have you committed a crime before?
  There will be consequence if yes.
- What was the CO2 amount of every country in 1990?
  Governments did not report the data (documentation, choice, error).

Why is missing data a problem

- What is your household income in the year 2025?
- Have you committed a crime before?
- What was the CO2 amount emitted by every country in 1990?

Why is missing data a problem

- What is your household income in the year 2025?
  We lose the richest group in our analysis.
- Have you committed a crime before?

- What was the CO2 amount emitted by every country in 1990?

Why is missing data a problem

- What is your household income in the year 2025?
  We lose the richest group in our analysis.
- Have you committed a crime before?
  We can't catch criminals!
- What was the CO2 amount emitted by every country in 1990?

Why is missing data a problem

- What is your household income in the year 2025?
  We lose the richest group in our analysis.
- Have you committed a crime before?
  We can't catch criminals!
- What was the CO2 amount emitted by every country in 1990?
  We want to be able to study all types of countries.

Types of Missing Data

1. **Missing Completely at Random (MCAR):** No systematic pattern in which observations are missing

Types of Missing Data

1. **Missing Completely at Random (MCAR):** No systematic pattern in which observations are missing
2. **Missing at Random (MAR):** Observations are missing for some values, but not due to a specific attribute (censored, truncation, survival bias, etc.)

Types of Missing Data

1. **Missing Completely at Random (MCAR):** No systematic pattern in which observations are missing
2. **Missing at Random (MAR):** Observations are missing for some values, but not due to a specific attribute (censored, truncation, survival bias, etc.)
3. **Missing Not at Random (MNAR):** Observations are missing for some values as a function of a particular attribute/mechanism

Example of Missing Data

- What is your household income in the year 2025?
- Have you committed a crime before?
- What was the CO2 amount emitted by every country in 1990?

Example of Missing Data

- What is your household income in the year 2025?
  Missing at random (conditional on being in rich group)
- Have you committed a crime before?

- What was the CO_2 amount emitted by every country in 1990?

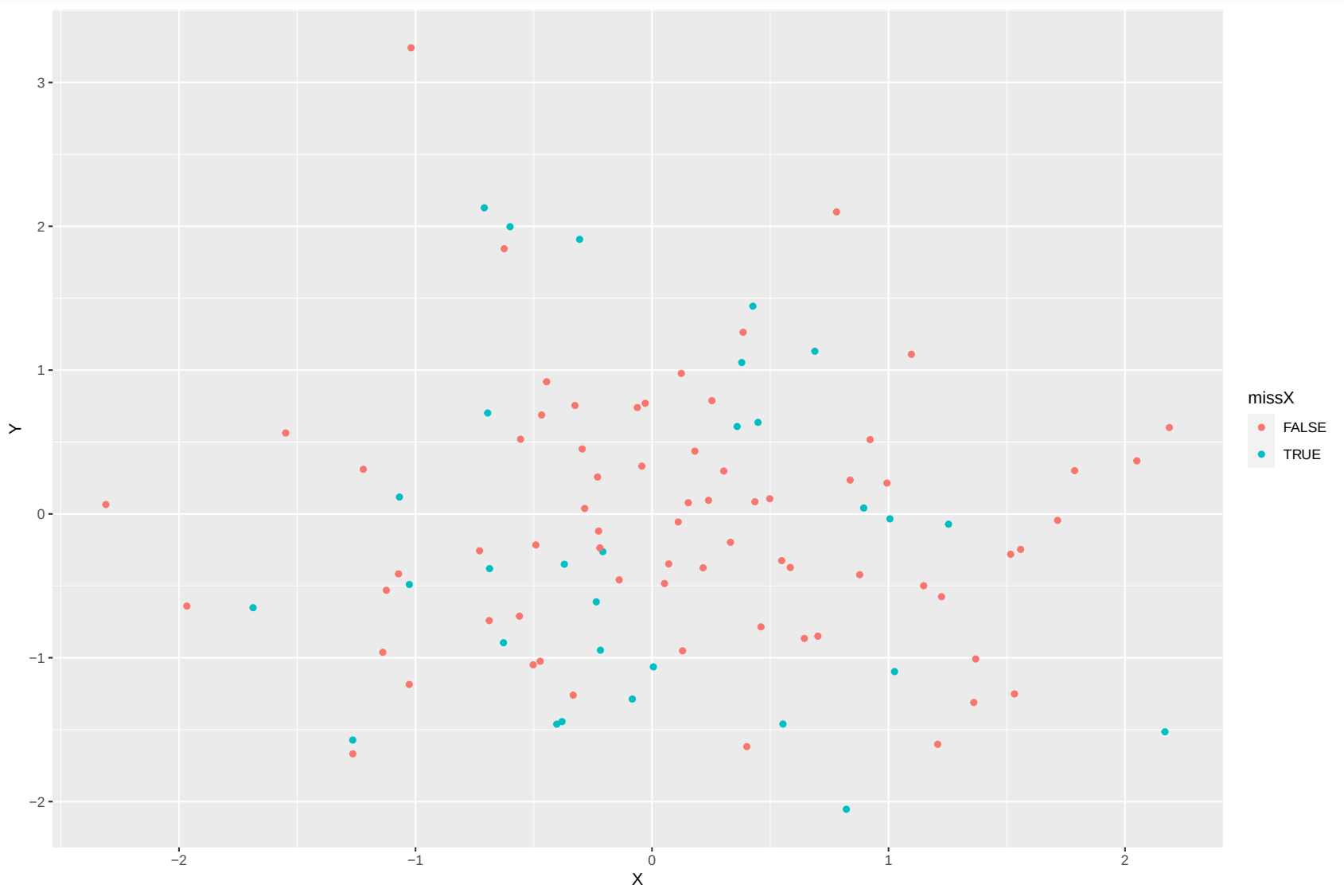
Example of Missing Data

- What is your household income in the year 2025?
  Missing at random (conditional on being in rich group)
- Have you committed a crime before?
  Missing not at random (consequence for doing crime)
- What was the CO_2 amount emitted by every country in 1990?

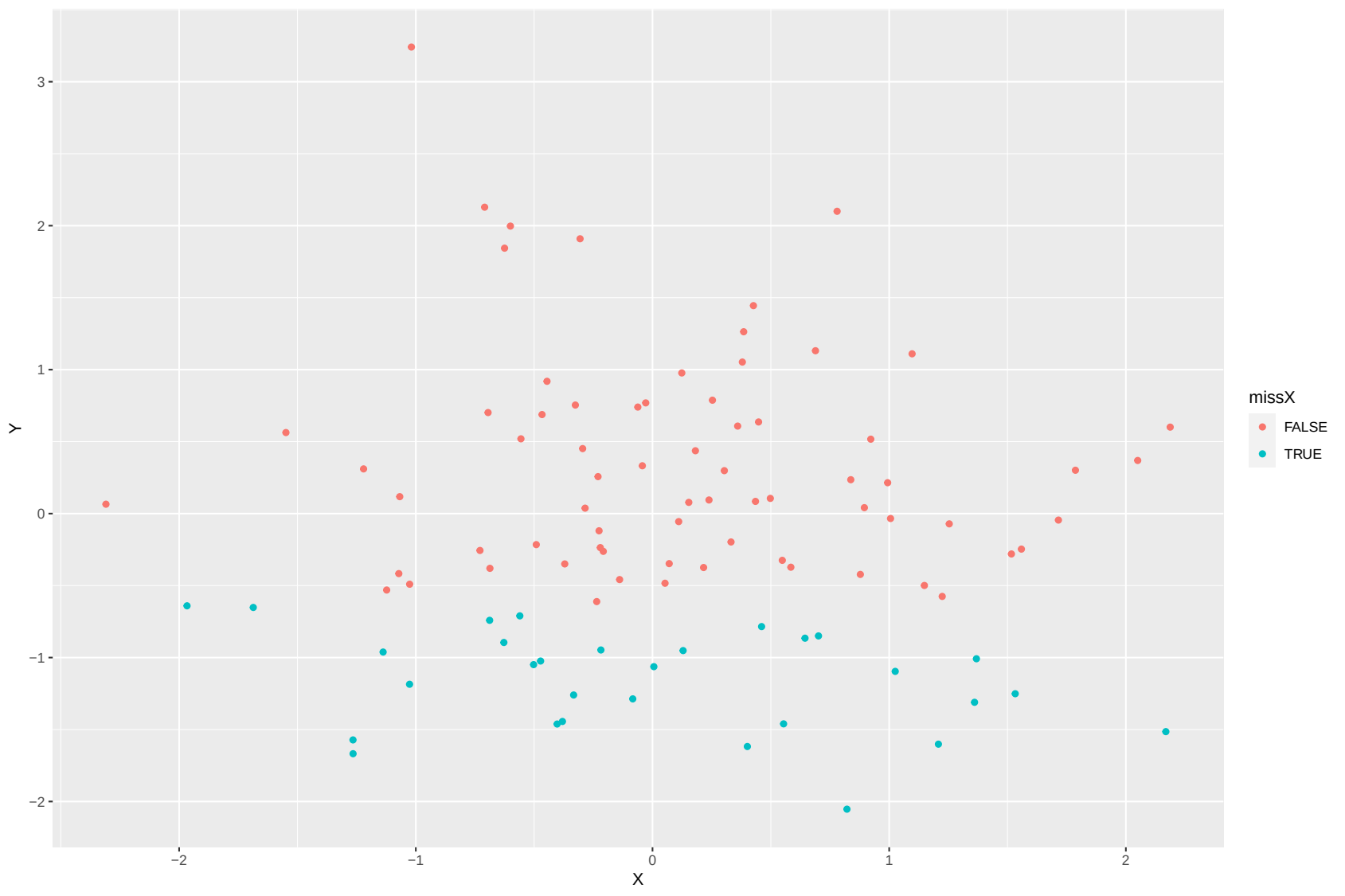
Example of Missing Data

- What is your household income in the year 2025?
  Missing at random (conditional on being in rich group)
- Have you committed a crime before?
  Missing not at random (consequence for doing crime)
- What was the CO_2 amount emitted by every country in 1990?
  Optimist: Missing completely at random
  Pessimist: Missing not at random (hide bad information)

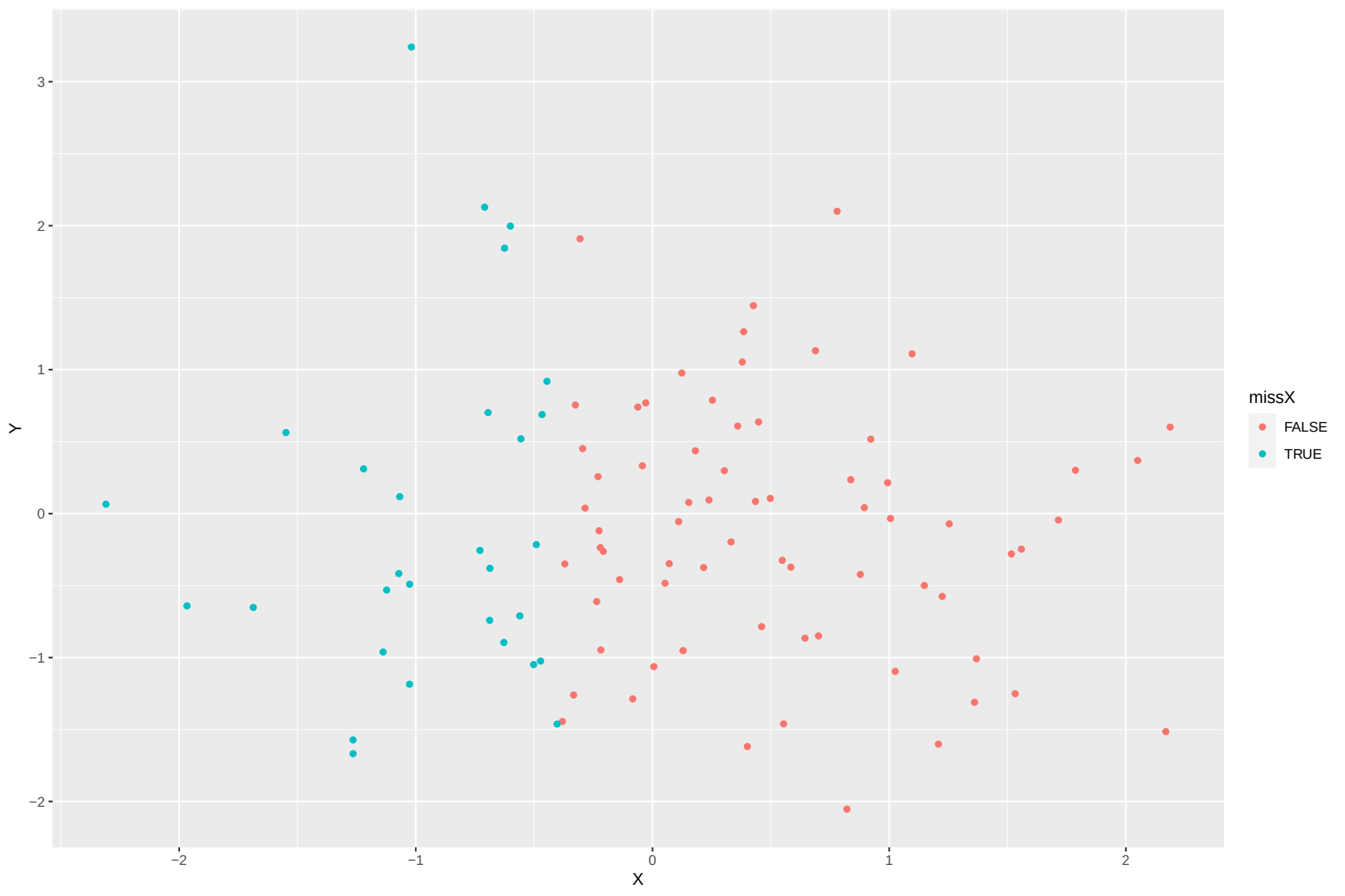
Missing Completely at Random



Missing (Conditionally) at Random



Missing Not at Random



Consequences of Missing Data

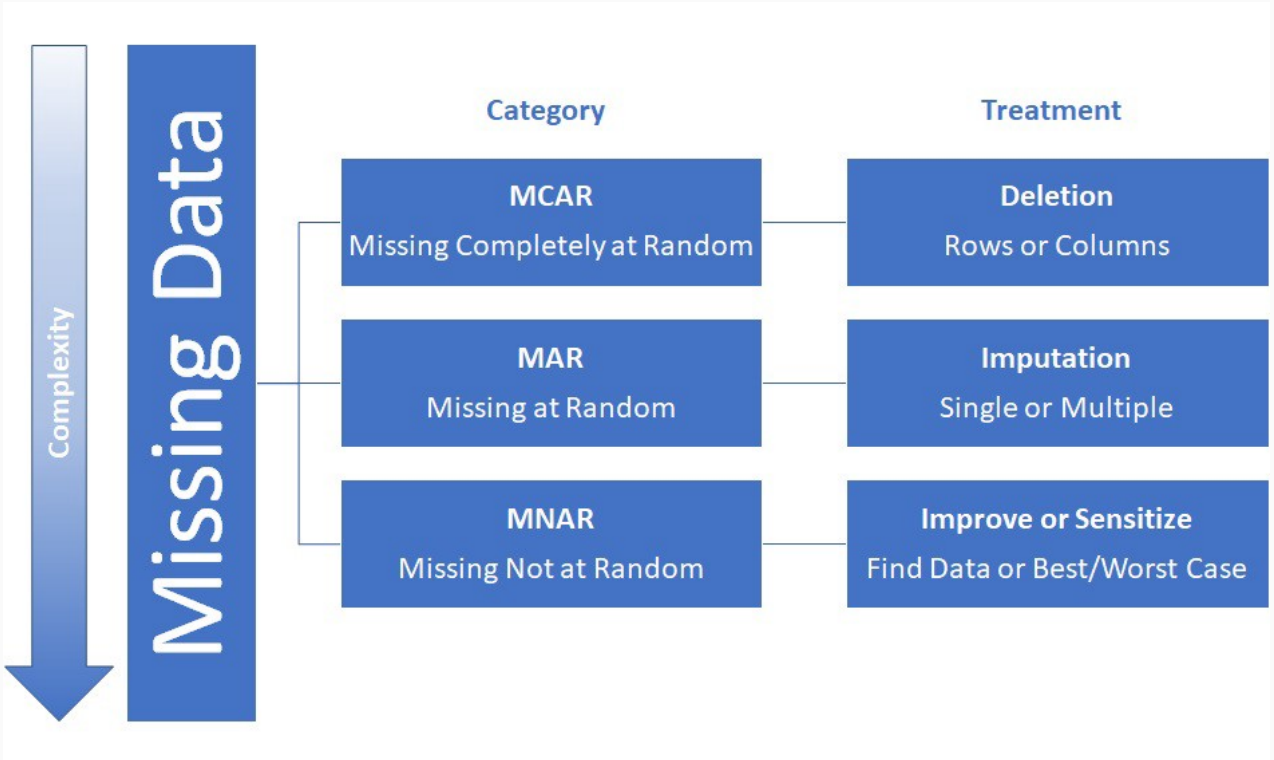
- Less learning power (fewer n)
- Selection bias (less representative)
- Omitted variable bias (biased estimates)

Main Takeaway: Never omit missing missing observations without understanding what type of missing data you have.

Solutions to Missing Data

1. Delete Missing Observations
2. Ignore Missing Observations
3. Impute Missing Observations
4. Improve Missing Observations

Missing Data Type Governs Solution

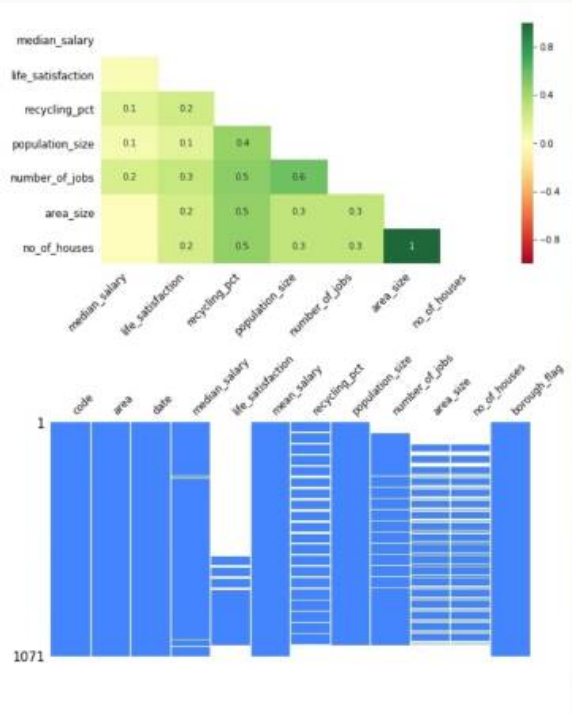


Visualizing Missing Data

Step 1: Inspect the data!

Step 2: Classify the type of missingness

Step 3: Solve based on type



Python: missingno (msno)

Solutions to Missing Data

1. Delete Missing Observations
2. Ignore Missing Observations
3. Impute Missing Observations
4. Improve Missing Observations

Deletion Methods

If data is MCAR ...

- **Listwise Deletion:** Delete all rows where one or more values are missing.

Deletion Methods

If data is MCAR ...

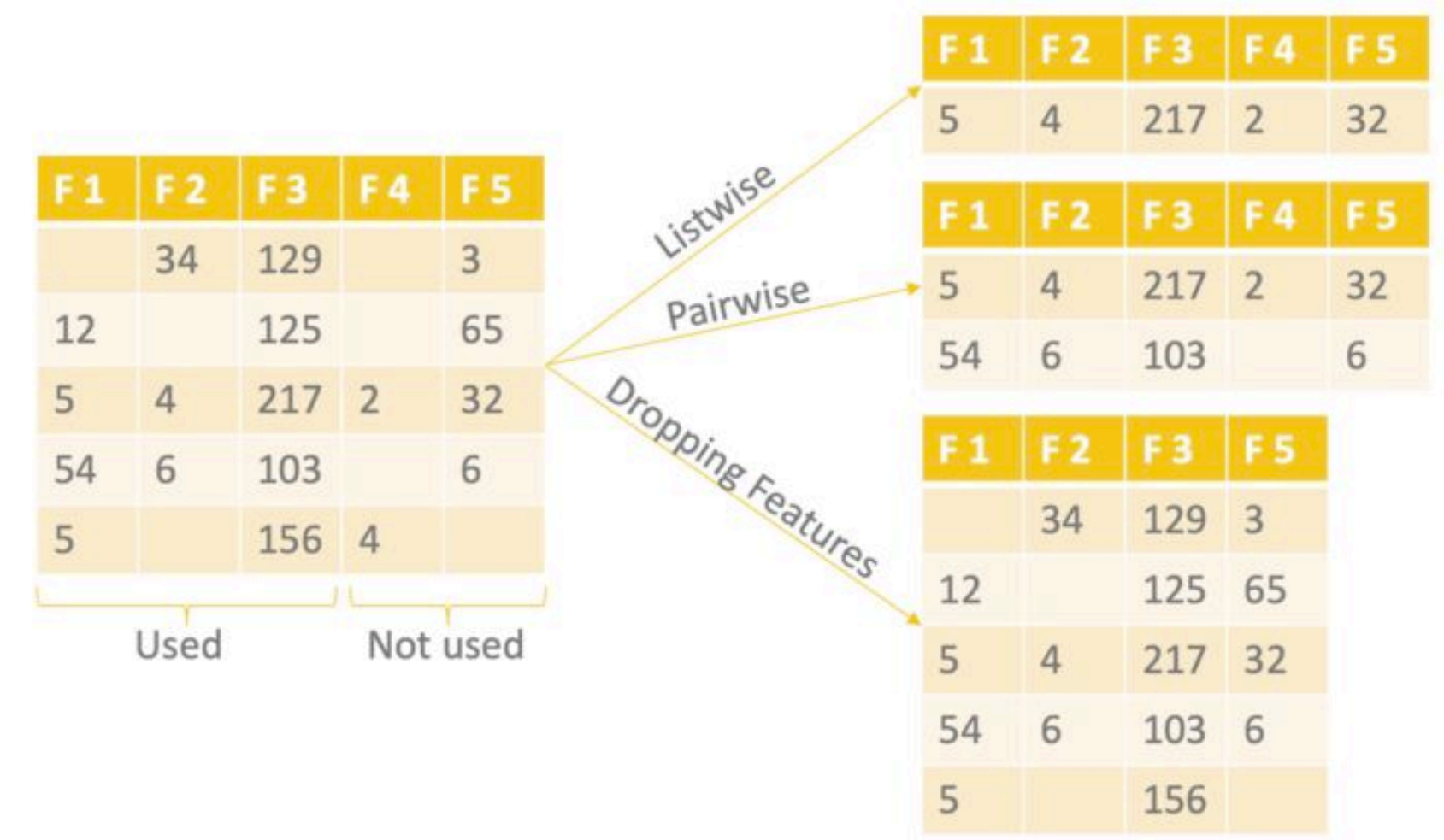
- **Listwise Deletion:** Delete all rows where one or more values are missing.
- **Pairwise Deletion:** Delete only the rows that have missing values in the columns used for the analysis.

Deletion Methods

If data is MCAR ...

- **Listwise Deletion:** Delete all rows where one or more values are missing.
- **Pairwise Deletion:** Delete only the rows that have missing values in the columns used for the analysis.
- **Dropping Features:** Drop entire columns with more missing values than a given threshold, e.g. 60

Deletion Methods



Solutions to Missing Data

1. Delete Missing Observations
2. **Ignore Missing Observations**
3. Impute Missing Observations
4. Improve Missing Observations

Ignore Methods

If data is MCAR for only some observations, then you may alternatively ignore by passing through the data (dropna())

Solutions to Missing Data

1. Delete Missing Observations
2. Ignore Missing Observations
3. **Impute Missing Observations**
4. Improve Missing Observations

Imputation Methods

If data is MAR (and it often is), then you may impute using:

- Zero/Constant Values
- Mean or Median Values
- KNN Values
- Multivariate Imputed Chained Equations (MCMC)
- Deep Learning Imputation

Zero/Constant Imputation

Method: replaces the missing values with either zero or any constant value you specify (often mode)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

Problem: Can skew results depending on input value.

Mean or Median Imputation

Method: Calculate the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Advantages and Disadvantages to Mean Imputation

Advantages:

Disadvantages:

Advantages and Disadvantages to Mean Imputation

Advantages:

- Easy and fast.
- Works well with small numerical datasets.

Disadvantages:

Advantages and Disadvantages to Mean Imputation

Advantages:

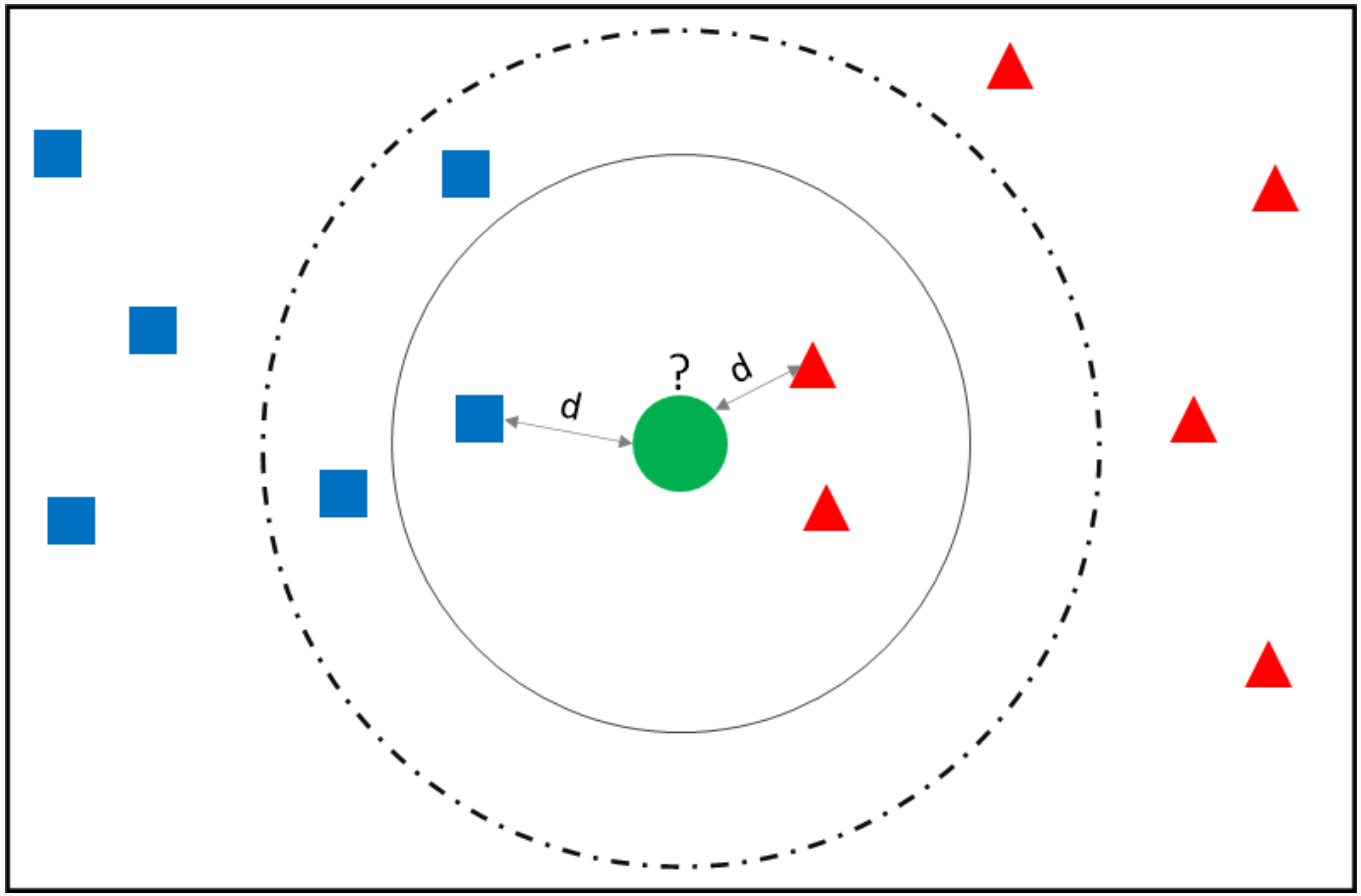
- Easy and fast.
- Works well with small numerical datasets.

Disadvantages:

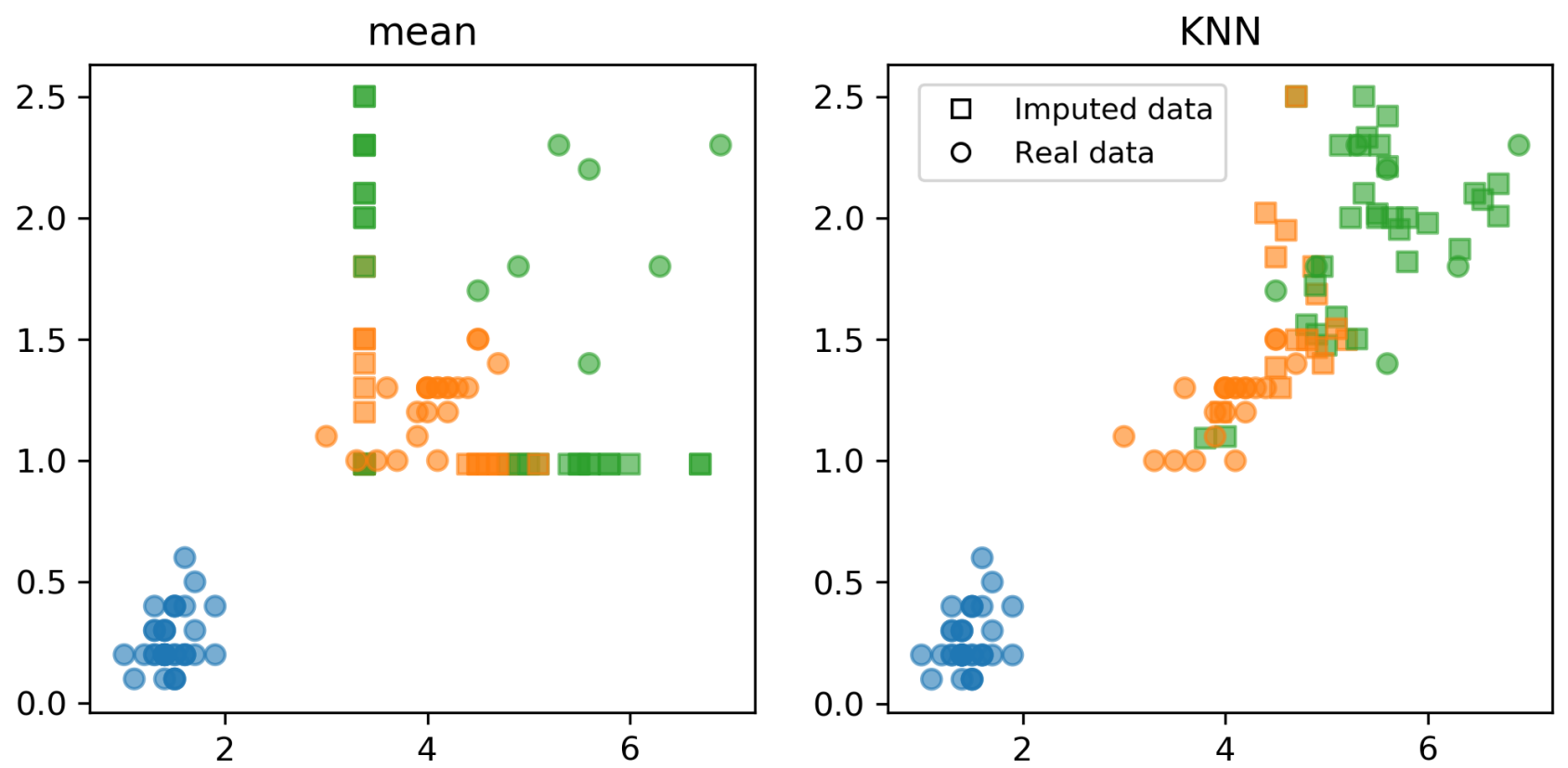
- Doesn’t factor the correlations between features. It only works on the column level.
- Will give poor results on encoded categorical features
- Not very accurate.
- Doesn’t account for uncertainty in the imputations

KNN Imputation

Method: Finding the k's closest neighbours to the observation with missing data and then imputing those values based on the non-missing values in the neighborhood.



Imputation Example



Advantages and Disadvantages to KNN Imputation

Advantages:

Disadvantages:

Advantages and Disadvantages to KNN Imputation

Advantages:

- Can be much more accurate than the mean, median or most frequent imputation methods
- Requires very few assumptions

Disadvantages:

Advantages and Disadvantages to KNN Imputation

Advantages:

- Can be much more accurate than the mean, median or most frequent imputation methods
- Requires very few assumptions

Disadvantages:

- Computationally expensive. Requires storing the whole training dataset in memory.
- Sensitive to outliers in the data

Solutions to Missing Data

1. Delete Missing Observations
2. Ignore Missing Observations
3. Impute Missing Observations
4. **Improve Missing Observations**

Solutions to Missing Data

1. Delete Missing Observations
2. Ignore Missing Observations
3. Impute Missing Observations
4. **Improve Missing Observations** → Collect more data!

Conclusion

- Data engineering requires understanding your data, not deleting “bad” observations
- PCA good at exploring trends across multi-dimensional data
- Clustering good at identifying similarities between groups of observations
- Missing data common but resolvable challenge