Big Data in Social Sciences

Week 4. Measurement

Nicola Barban

How to participate?

1 Go to wooclap.com

2 Enter the event code in the top banner

Event code
BCEXXG

Copy participation link

“The most important aspect of a statistical analysis is not what you do with the data, it’s what data you use”

Andrew Gelman

This week

Types and sources of data bias

Missing values

Measurement errors

Correlation

Principa Component Analysis

Concepts & measurement

- **Social science is about understanding causal relationships:**
  - Does minimum wage change levels of employment?
- **Relationships are between concepts:**
  - Minimum wage, unemployment, outgroup contact, views on immigration.
- **Important to consider how we measure these concepts.**
  - Some more straightforward: what is your age?
  - Others more complicated: what does it mean to “be liberal”?
  - Operational definition: mapping of concept to numbers in our data.

Measurement and bias

- **Measurement error: chance variation in our measurements.**
- **individual measurement = exact value + chance error**
- chance errors tend to cancel out when we take averages.
- **Bias: systematic errors for all units in the same direction.**
- **individual measurement = exact value + bias + chance error.**
- “What did you eat yesterday?” ⇝ underreporting

Simple Random Sampling (SRS)

- **Random Sampling means every element of the population has an equal probability of being chosen.**
- When a sample is selected at random from a population, it is said to be an **unbiased sample.**
- However, if the sample is selected incorrectly it may result in a biased sample.

Bias

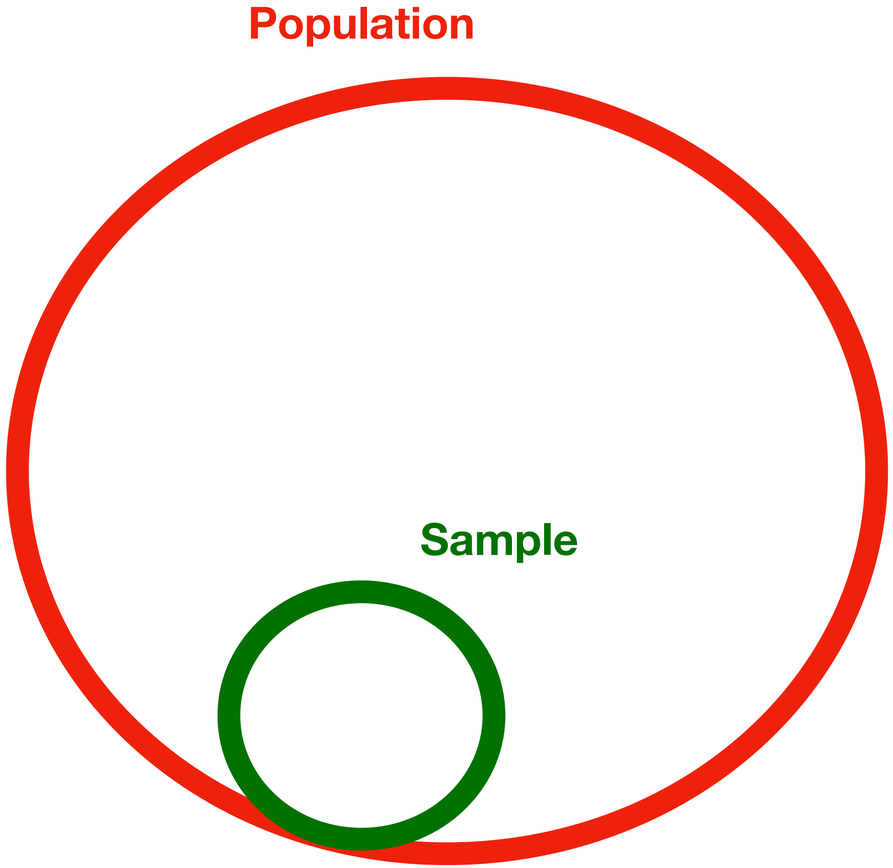
- If the sample is not representative of the population, then the results may be inaccurate.
- If a survey is ambiguous, subjective, or biased, then the results may be inaccurate.
- Bias is any factor that favours certain outcomes or responses, or influences an individual's responses.
- Bias may be unintentional (accidental), or intentional (to achieve certian results)

Sampling bias

- **The sample does not represent adequately the population of interest**

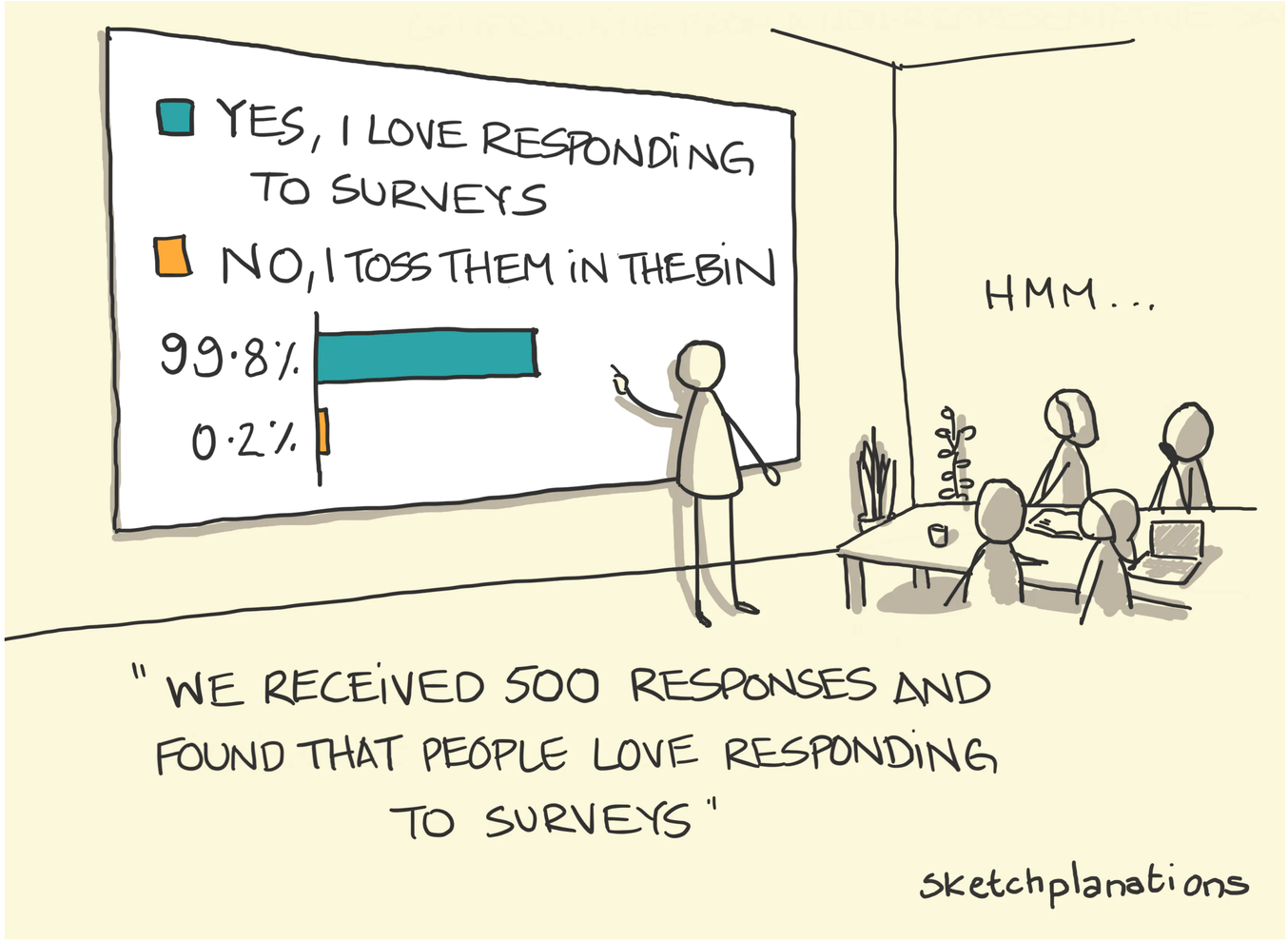
  Example:

  Attendees at a Star Trek convention may report that their favourite genre is a science fiction. While this may be representative of the population at the convention, it might not reflect the general population's opinion.



Non-response bias

- Certain groups are under-represented because they elect not to participate.



Response Bias

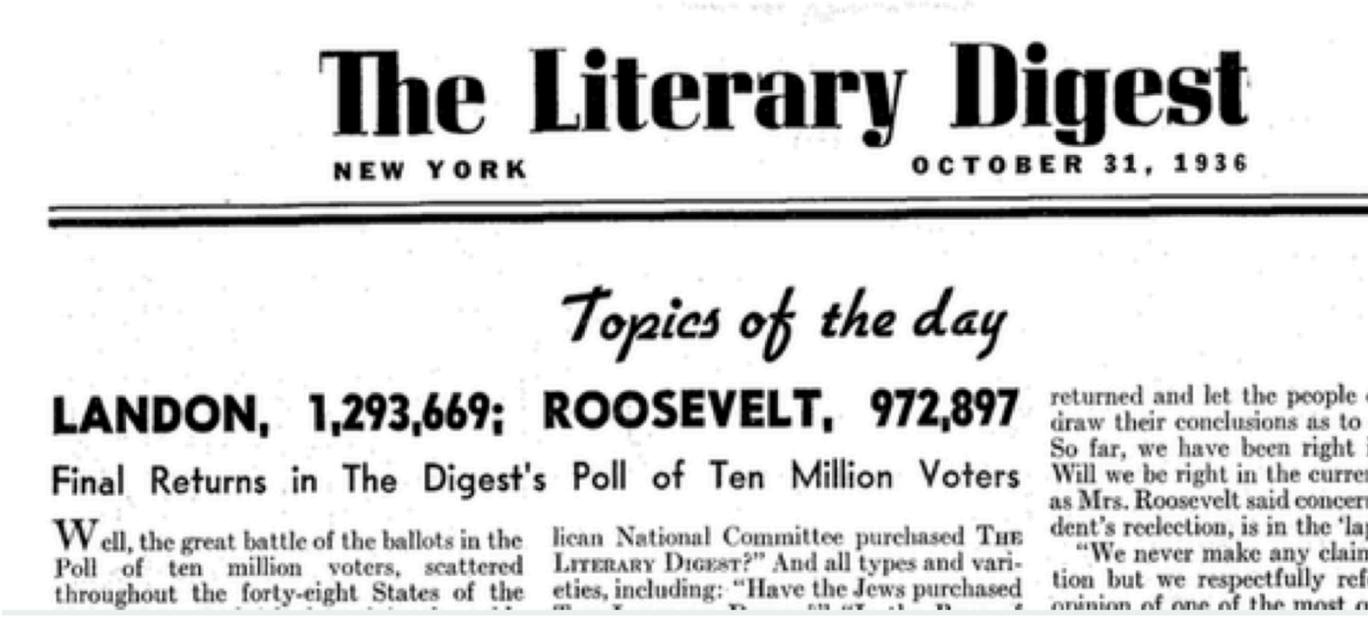
- Factors in the sampling method influence the data obtained.

  Example1: A respondent may answer questions in the way she thinks the questioner wants her to answer or a man may respond differently to a woman when asked questions about domestic violence.

  Example 2: A psychologist is conducting a research study concerning sexual activities. The survey is administered over the phone and many of the questions are personal. Some participants feel uncomfortable and do not answer honestly. Their responses are biased toward what they perceive as being socially acceptable.

The 1936 Polling Disaster

- In 1936 The Literary Digest made a prediction that Republican Alf Landon would beat the incumbent Democratic president Franklin Delano Roosevelt in a landslide.
- Rather than a landslide for Landon instead Roosevelt earned a rather comfortable win.
- This rates as arguably the biggest disaster of opinion polling.



The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people
draw their conclusions as to
So far, we have been right
Will we be right in the curre
as Mrs. Roosevelt said concer
dent's reelection, is in the 'la
"We never make any clain
tion but we respectfully ref

Survey methodology

- Literary Digest predicted elections using mail-in polls.
- Source of addresses: automobile registrations, phone books, etc.
- In 1936, sentout 10 million ballots, over 2.3 million returned.
- George Gallup used only 50,000 respondents.

	FDR's vote share
Literary Digest	43
Gallup	56
Presidential results	63

Pool Fail

- **Selection bias:** ballots skewed toward the wealthy (with cars, phones)
  - Only 1 in 4 households had a phone in 1936.
- **Nonresponse bias:** respondents differ from nonrespondents.
  - When selection procedure is biased, adding more respondents won’t help

Why random sampling works?

- **Law Large Numbers (LLN):** when sample size tends to infinity, the sample mean equals to population mean.

  as N $\rightarrow \infty$, $E[\bar{X} = \mu]$

- **Central Limit Theorem (CLT):** when sample size tends to infinity, the sample mean will be normally distributed.

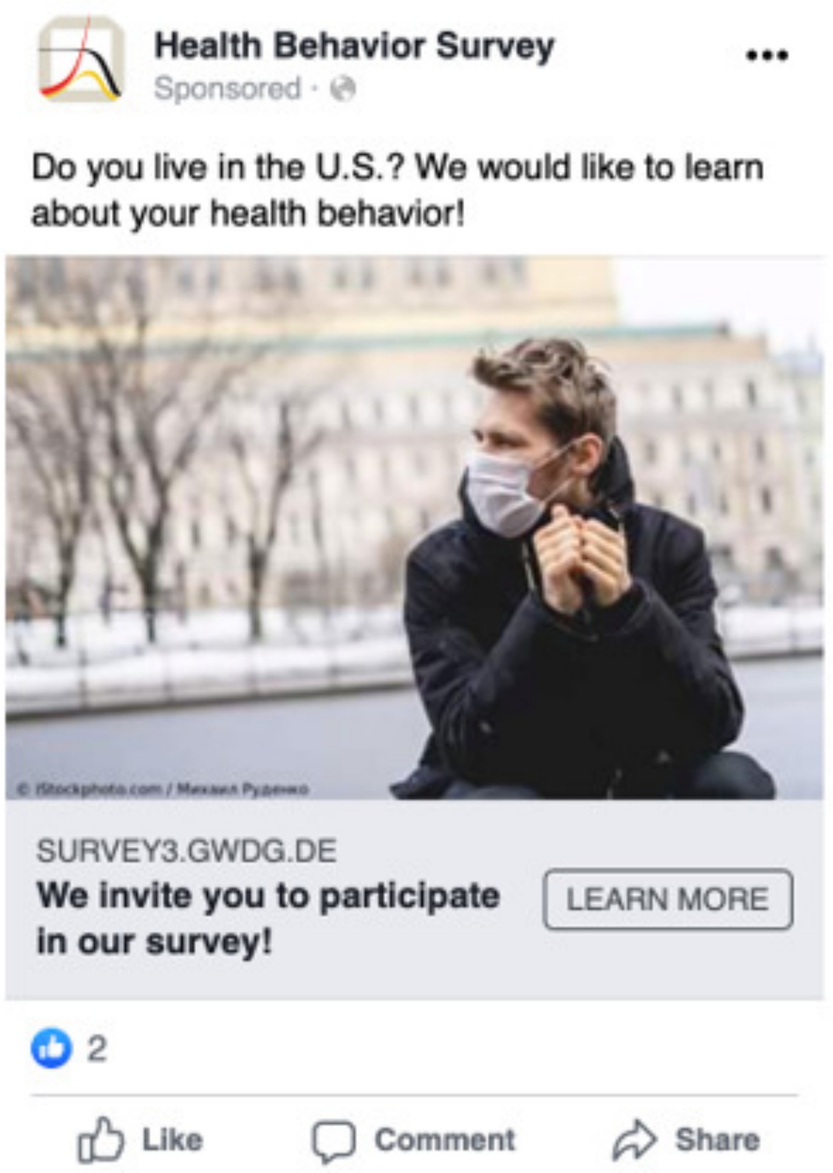
- as N $\rightarrow \infty$, $\bar{X} \sim N(\mu, \sigma/\sqrt{N})$

Sampling glossary

- **Target population:** set of people we want to learn about.
  - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample.
  - Frame bias: list of registered voters (frame) might include nonvoters!
- **Sample:** set of people contacted.
- **Respondents:** subset of sample that actually responds to the survey.
  - Unit non-response: sample ≠ respondents.
  - Not everyone picks up their phone.
- **Completed items:** subset of questions that respondents answer.
  - Item non-response: refusing to disclose their vote preference.

Difficulties of sampling

- **Problems of telephone survey**
  - Cell phones (double counting for the wealthy)
  - Caller ID screening (unit non-response)
  - Response rates down to 9%!
- **An alternative: Internet surveys**
  - Opt-in panels, respondent-driven sampling ⇝ **non-probability sampling**
  - Cheaper, but non-representative
  - Digital divide: rich vs. poor, young vs. old
  - Correct for potential sampling bias via statistical methods.

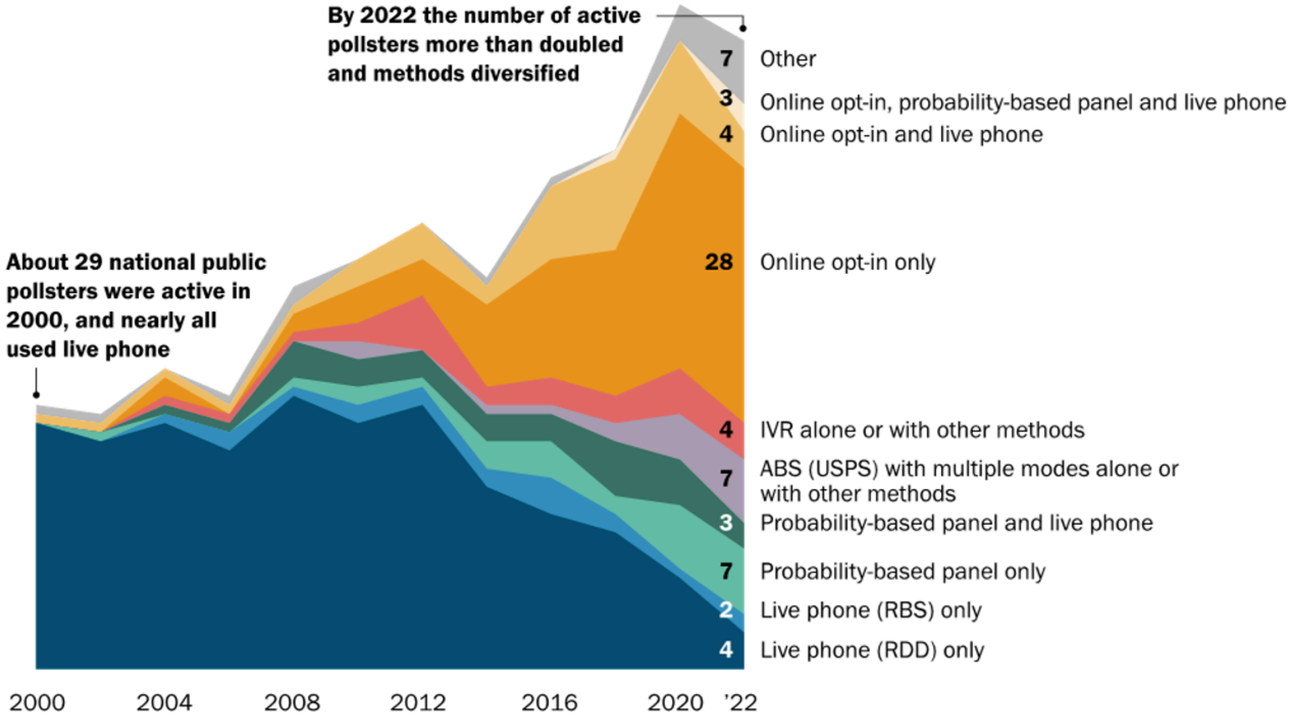


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7744148/

https://www.pewresearch.org/methods/2023/04/19/how-public-polling-has-changed-in-the-21st-century/

Polling has entered a period of unprecedented diversity in methods

Number of national public pollsters in the U.S. using method(s)



Note: Figures represent the number of active national public pollsters in each year and the method(s) that they used. IVR refers to interactive voice response, also known as robo-polling. ABS refers to address-based sampling. RBS refers to voter registration-based sampling. RDD refers to random-digit-dial sampling.
Source: Pew Research Center analysis of external data. See Methodology for details.
"How Public Polling Has Changed in the 21st Century"

PEW RESEARCH CENTER

https://www.pewresearch.org/methods/2023/04/19/how-public-polling-has-changed-in-the-21st-century/

How to make survey questions?

- The way you ask questions influences the outcome
  - Language in a multi-country study
  - open/closed questions
  - Interviewer effect
  - Scale used (1-10; strongly agree-strongly disagree)
  - Even small wording differences can substantially affect the answers people provide.
  - Question order matters

Fewer people mention the economy in open-ended version

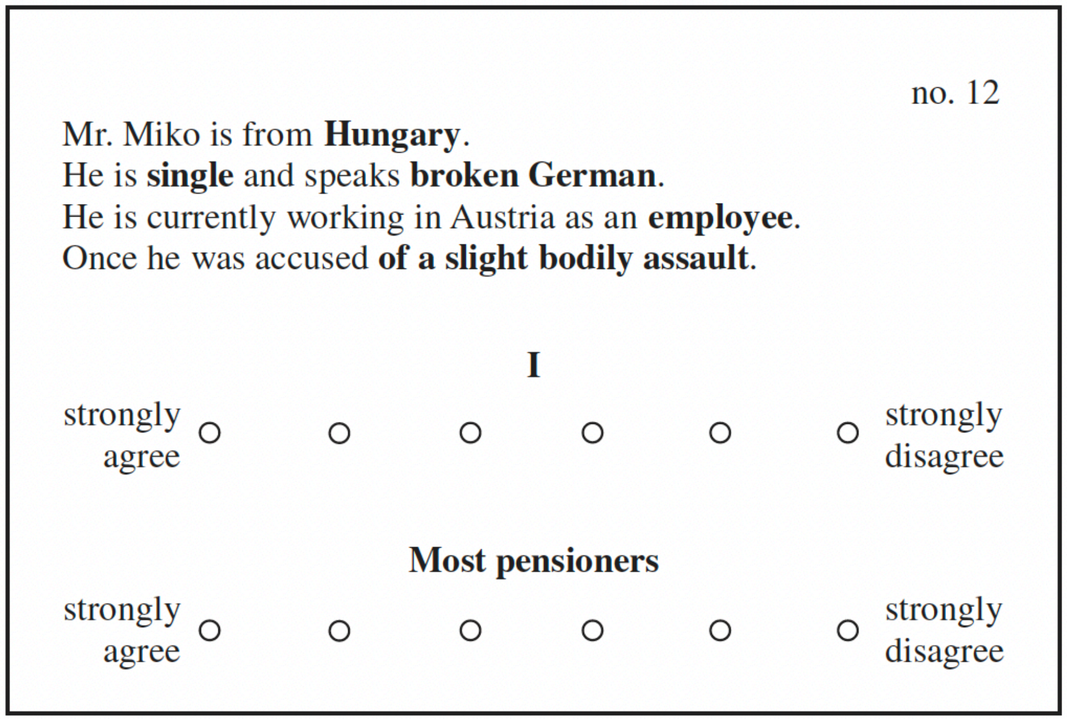
% answering that the issue matter most in deciding their vote for president in 2008

	Open-ended	Closed-ended
The economy	**35**	**58**
The war in Iraq	5	10
Health care	4	8
Terrorism	6	8
Energy policy	*	6
Other	43	8
Candidate mentions	9	–
Moral values/social issues	7	–
Taxes/distribution of income	7	–
Other issues	5	–
Other political mentions	3	–
Change	3	–
Other	9	–
Don't know	7	2
	100	100

Note: Open-ended figures reflect respondents' unprompted first response. Close-ended figures reflect respondents' first choice from five options read by the interviewer.
Source: Survey conducted November 2008.

PEW RESEARCH CENTER

Vignette study



no. 12

Mr. Miko is from **Hungary**.
He is **single** and speaks **broken German**.
He is currently working in Austria as an **employee**.
Once he was accused **of a slight bodily assault**.

I

strongly agree ○ ○ ○ ○ ○ ○ strongly disagree

Most pensioners

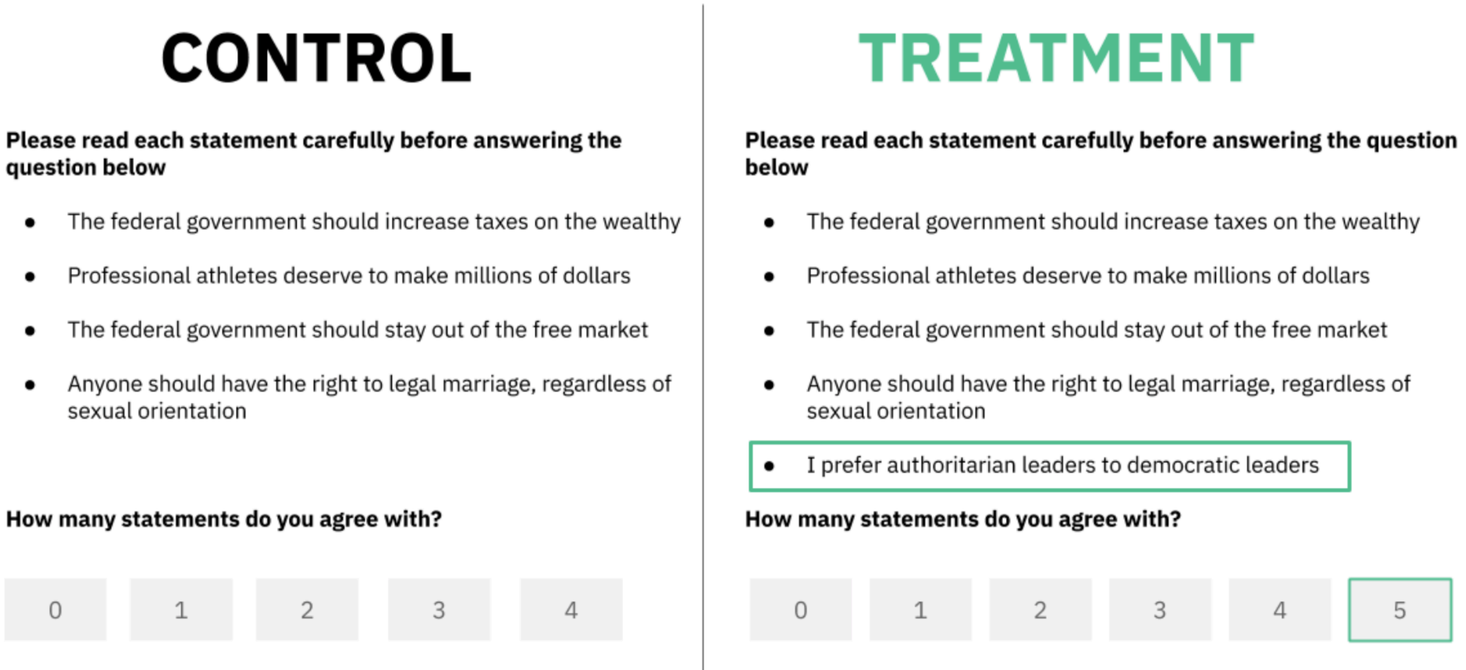
strongly agree ○ ○ ○ ○ ○ ○ strongly disagree

Figure 1. Vignette of an applicant for the Austrian citizenship. The question to be answered by a student was: "Should this applicant become an Austrian citizen?"

List experiment

A list experiment requires that you randomly divide the sample into two groups: the Direct Response Group and the Veiled Response Group.

A typical list experiment question looks like this:



Using randomization to elicit sensible information

The prevalence of an opinion is devised through the randomization process, and is simply the difference in means between the number of endorsed statements by the treatment and control group.

When to use list experiments?

- When you want to learn about the prevalence of sensitive attitudes, perceptions, or behaviors
- When privacy is imperative for honest self-reporting
- When you want to know if there's a gap between public and private opinions
- When you want to know *who* is more likely to withhold their public opinion

Missing values

- Missing data can arise from various places in data:
  - A survey was conducted and values were just randomly missed when being entered in the computer.
  - A respondent chooses not to respond to a question like `Have you ever recreationally used opioids?'.
- The source of missing values in data can lead to the major types of missingness:

Type of missingness

There are 3 major types of missingness to be concerned about:

1. **Missing Completely at Random (MCAR)** - the probability of missingness in a variable is the same for all units. Like randomly poking holes in a data set.

2. **Missing at Random (MAR)** - the probability of missingness in a variable depends only on available information (in other predictors).

3. **Missing Not at Random (MNAR)** - the probability of missingness depends on information that has not been recorded and this information also predicts the missing values.

What are examples of each these 3 types?

How to deal with missing data

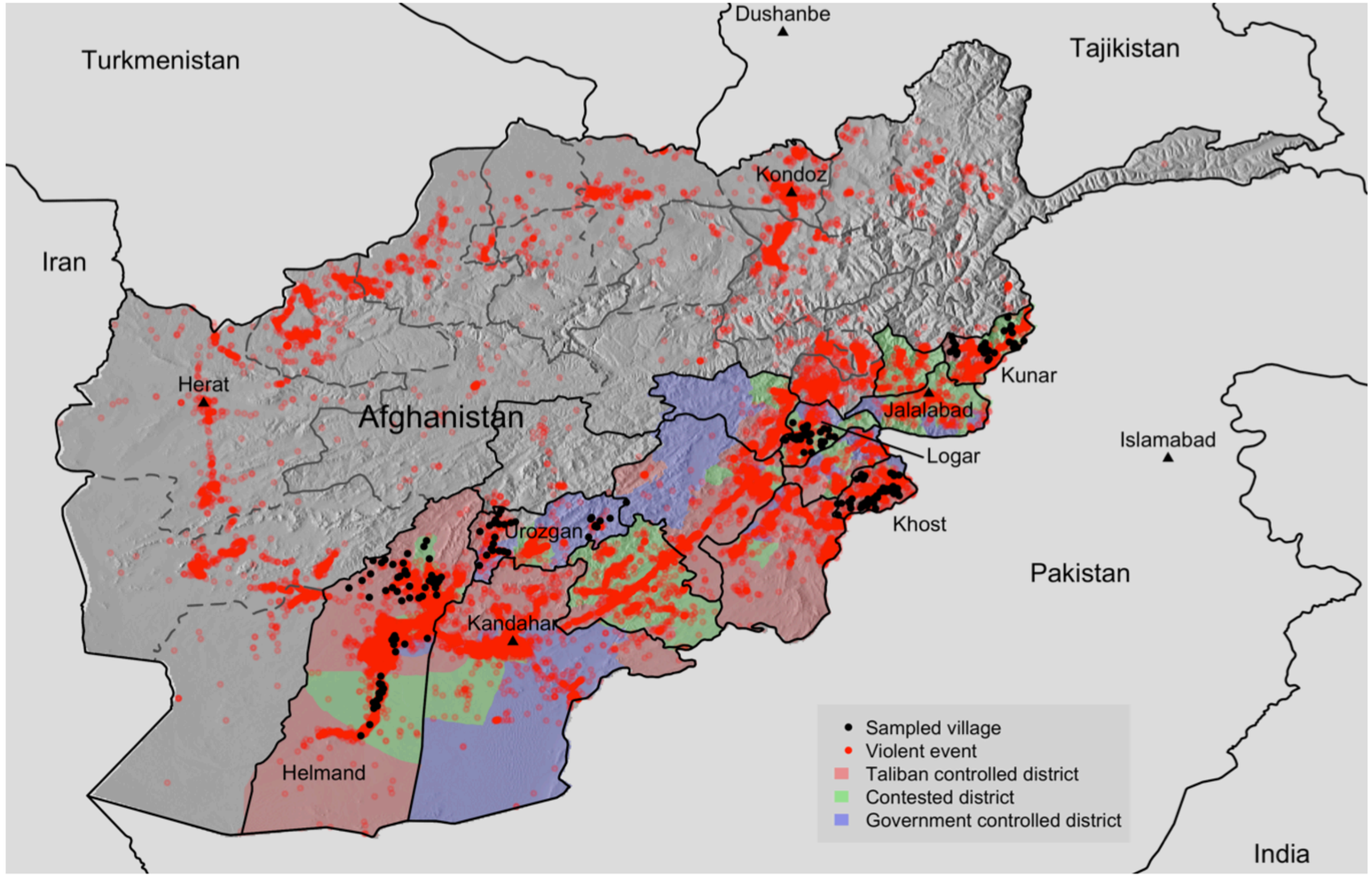
There are three main approaches to handle missing data:

1. Omission: Ignoring (dropping) samples that has missing value from the analysis

2. Imputation: Filling missing values

3. Analysis: Using statistical methods unaffected by missing data.

Afghanistan study: sample civilians on their exposure to violence and support for Taliban, coalition forces



Script for the control group:

I’m going to read you a list with the names of different groups and individuals on it. After I read the entire list, I’d like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don’t tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

Karzai Government; National Solidarity Program; Local Farmers

•

Script for the TREATMENT group:

I’m going to read you a list with the names of different groups and individuals on it. After I read the entire list, I’d like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don’t tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

Karzai Government; National Solidarity Program; Local Farmers; **ISAF (Taliban)**

•

Surveying Afghan Population

- One problem with randomization: **need a list to sample from.**
  - Random digit dialing: all phone numbers.
  - Other polls are using voter files.
- **No comprehensive list of citizens in Afghanistan to use**
- **Alternative: multi-stage cluster sampling**
  - Randomly choose villages from a list of all villages
  - Go to each village and randomly choose households.

Measuring difficult “things”

Some examples:

- Fertility intentions
- Intelligence
- Political Ideology
- Social Class
- Retrospective questions (what did you eat 2 weeks ago; How tall were you at age 14?)

Fertility preferences/intentions

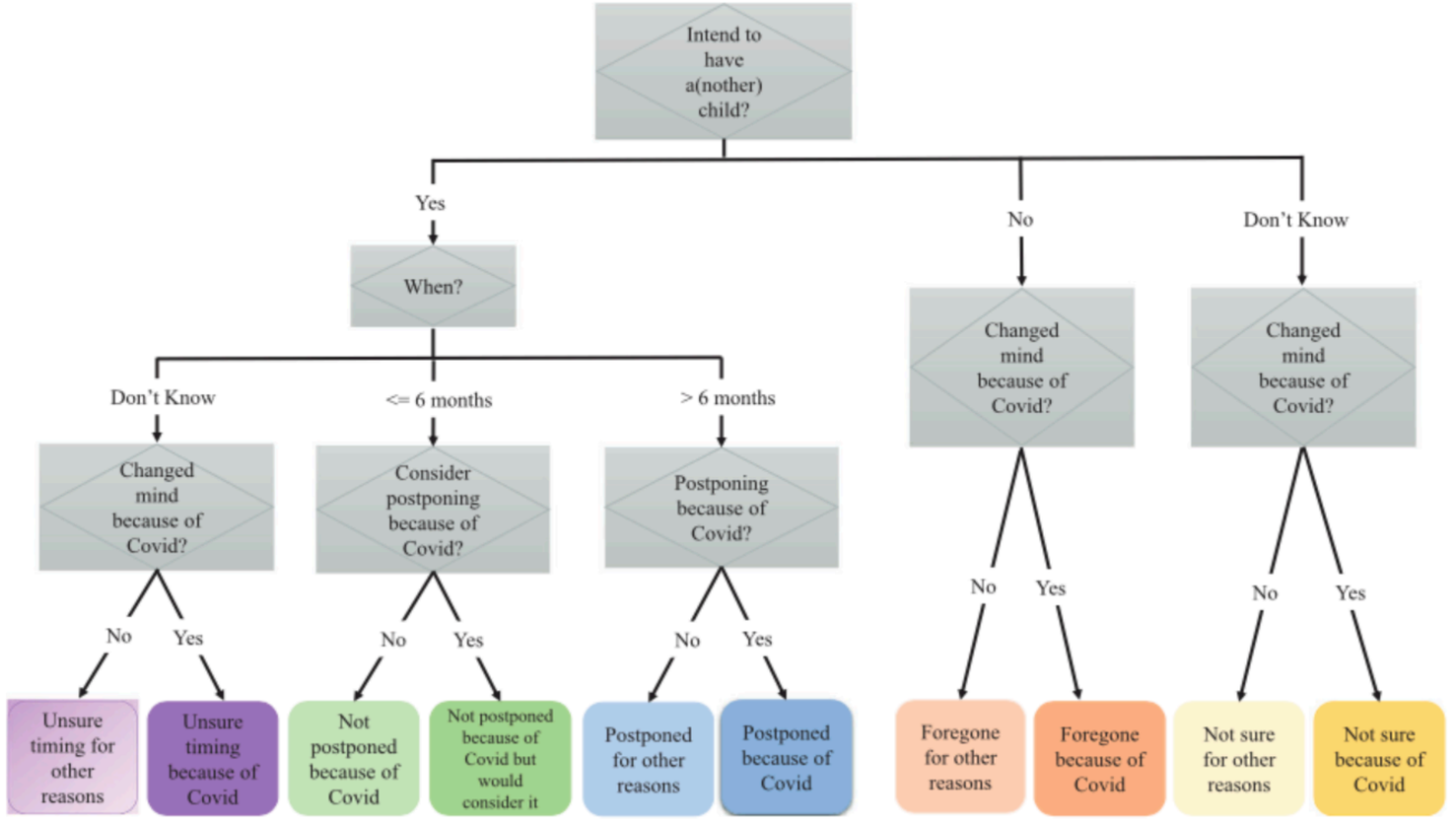
How many children do you prefer to have?

Fertility preferences/intentions

How many children do you prefer to have?

Do you intend to become a parent in the next 3 years?

FIGURE 1 Flow chart for construction of fertility intentions typology using DeCodE, Wave 1



NOTE: Target sample for questions on fertility intentions is women, ages 18-34, who were not sterilized or declared infecund (N= 3,753).
Timing regarding when to have another child refers to the duration within which the respondent intends to get pregnant. Among currently pregnant respondents, timing refers to how soon after giving birth the respondent would like to pregnant again.

What Is Principal Component Analysis?

Principal component analysis (PCA) is a **dimensionality reduction** and **machine learning method** used to **simplify a large data set into a smaller set while still maintaining significant patterns and trends.**

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity.

So, to sum up, the idea of PCA is simple: **reduce the number of variables of a data set, while preserving as much information as possible.**

How does it work?

- PCA project the data to a lower dimensionality (from $\mathbb{R}^d$ to $\mathbb{R}^q$, $q < d$)

- Works with numerical variables.
- Produces a **low-dimensional representation of a dataset**:
- Finds **linear combinations** of the original features that have maximal variance
- The derived variables are mutually uncorrelated

Step 1: Standardization

The aim of this step is to **standardize the range of the continuous initial variables** so that each one of them contributes equally to the analysis.

- More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables.

$$Z = \frac{X - \bar{X}}{\sigma}$$ with $\bar{X}$ = mean of variable and σ its standard deviation

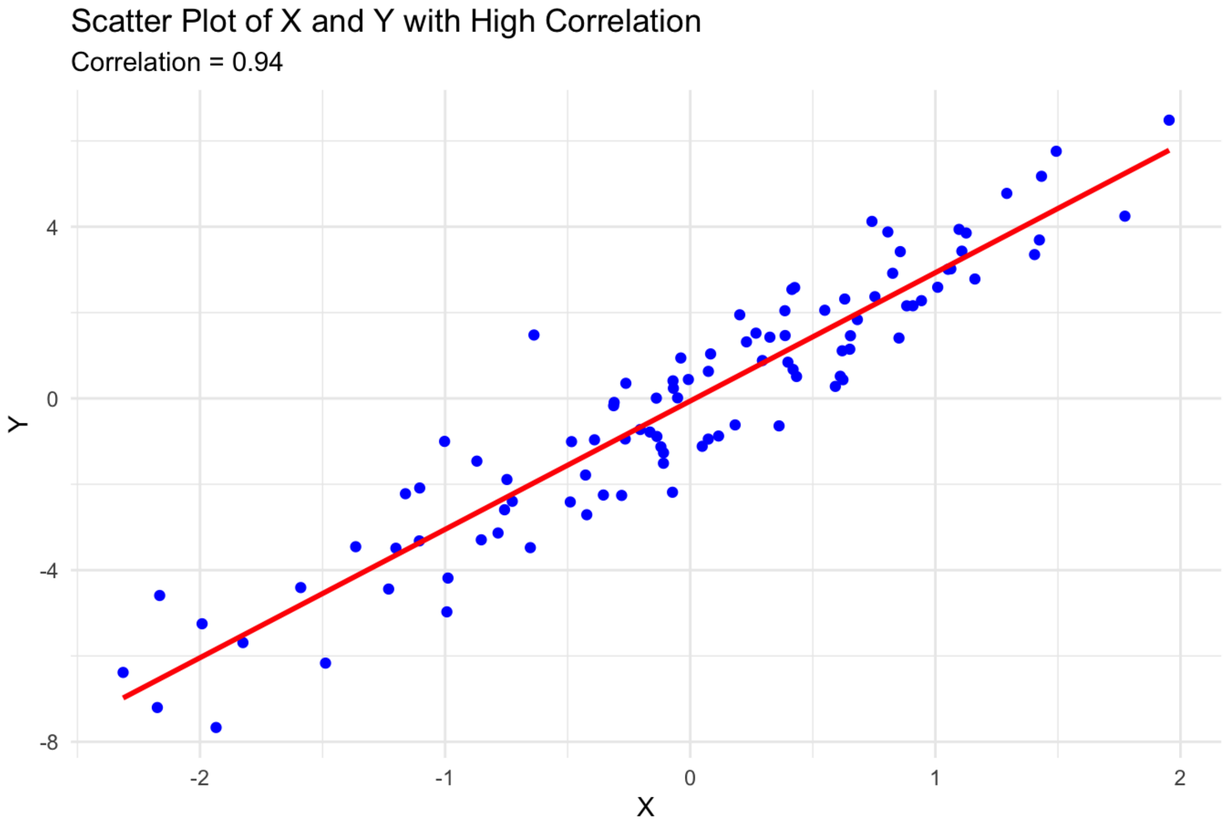
Step 2. Covariance Matrix Computation

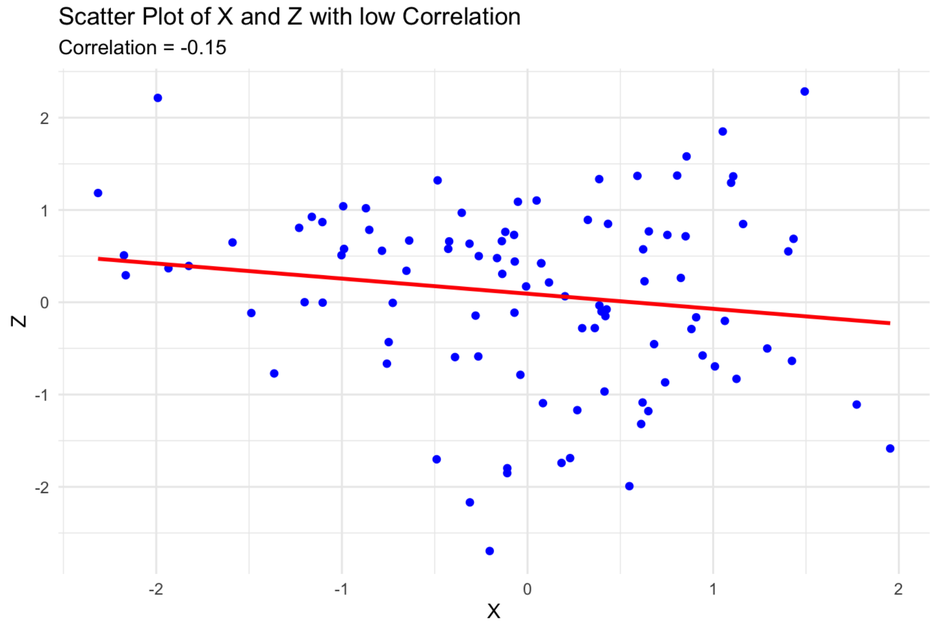
- The aim of this step is to understand if there is any relationship between the variables.
- Variables that are highly correlated contain redundant information.
- The covariance matrix is a p × p symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables x, y, and z, the covariance matrix is a 3×3 data matrix of this form:

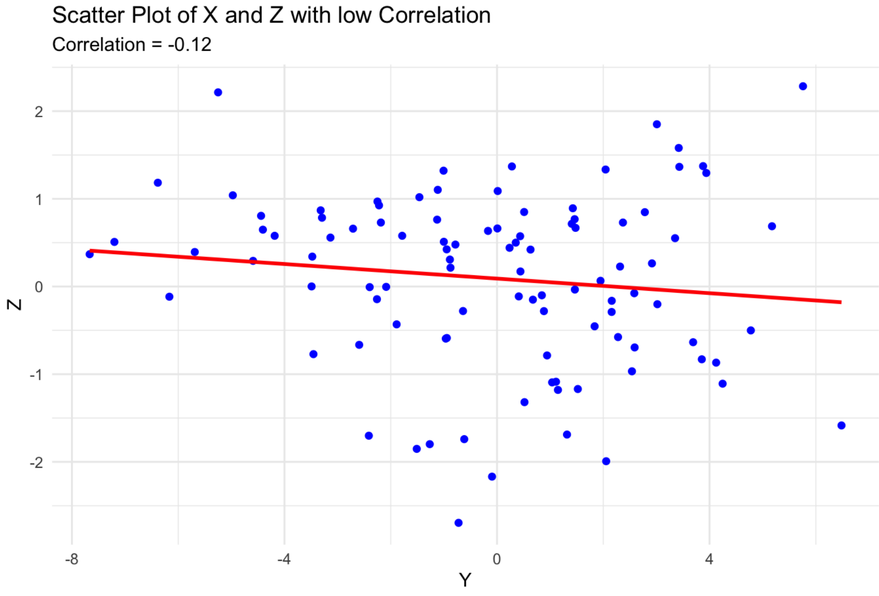
$$\begin{bmatrix} \text{Cov}(x,x) & \text{Cov}(x,y) & \text{Cov}(x,z) \\ \text{Cov}(y,x) & \text{Cov}(y,y) & \text{Cov}(y,z) \\ \text{Cov}(z,x) & \text{Cov}(z,y) & \text{Cov}(z,z) \end{bmatrix}$$

To demonstrate high covariance between two variables X and Y we can generate a dataset where Y is strongly linearly related to X (e.g., Y=3X+ small noise). This will result in a high correlation, close to 1.

	scale.X.	scale.Y.	scale.Z.
scale.X.	1.0000000	0.9392921	-0.1539882
scale.Y.	0.9392921	1.0000000	-0.1245823
scale.Z.	-0.1539882	-0.1245823	1.0000000


Scatter Plot of X and Y with High Correlation
Correlation = 0.94
Y
X
4
0
-4
-8
-2
-1
0
1
2


Scatter Plot of X and Z with low Correlation
Correlation = -0.15
Z
X
2
1
0
-1
-2
-2
-1
0
1
2


Scatter Plot of X and Z with low Correlation
Correlation = -0.12
Z
Y
2
1
0
-1
-2
-8
-4
0
4

Properties of covariance

- Covariance

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

- Covariance of X,X is variance X

$$\text{Cov}(X, X) = \text{Var}(X)$$

- Covariance is sensitive to scaling.

$$\text{Cov}(aX, bY) = a \cdot b \cdot \text{Cov}(X, Y)$$

- Covariance is not affected by shifts (adding a constant).

$$\text{Cov}(X + c, Y + d) = \text{Cov}(X, Y)$$

Properties of correlation

- Correlation

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Correlation of X,X is 1

$$\text{Cor}(X, X) = 1$$

- Correlation is not affected byscaling.

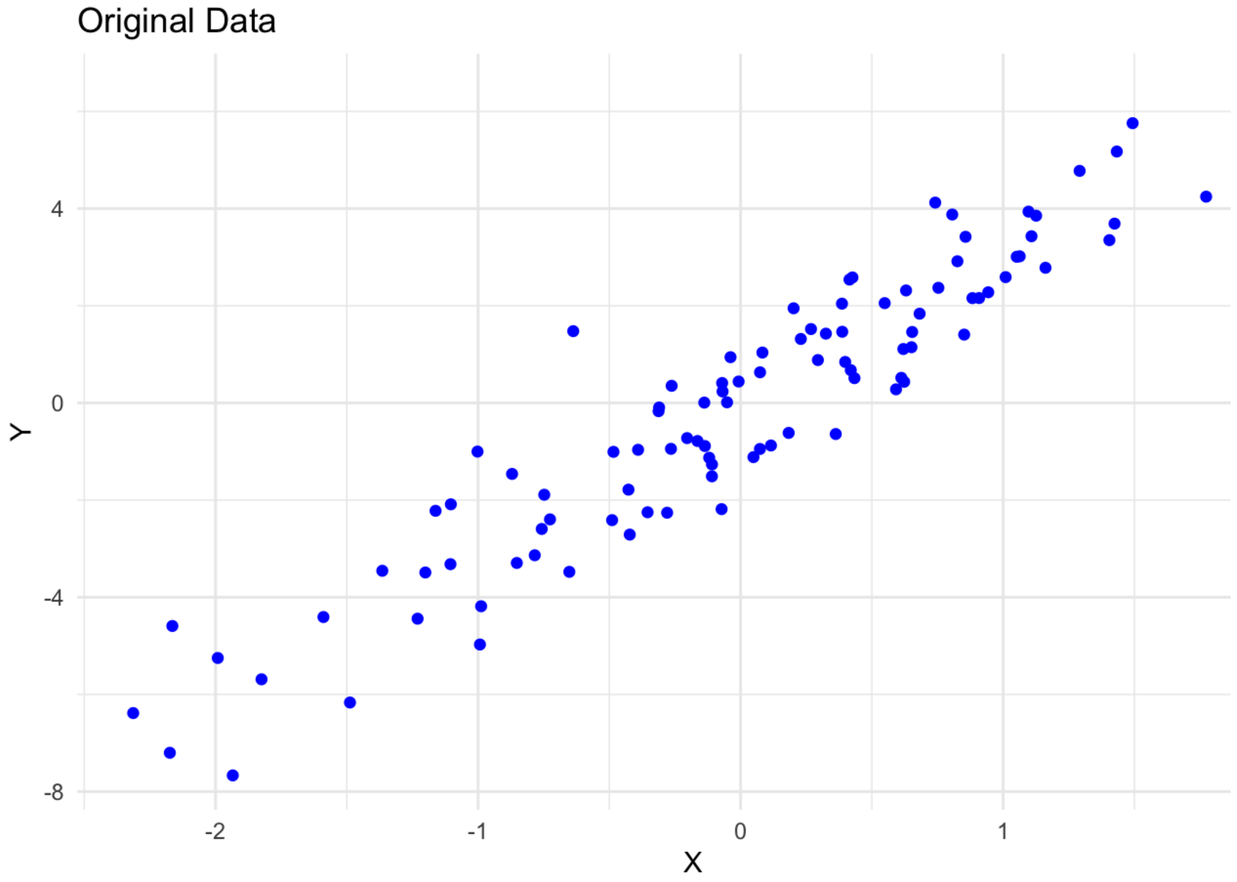
$$\text{Cor}(aX, bY) = \text{Cor}(X, Y)$$

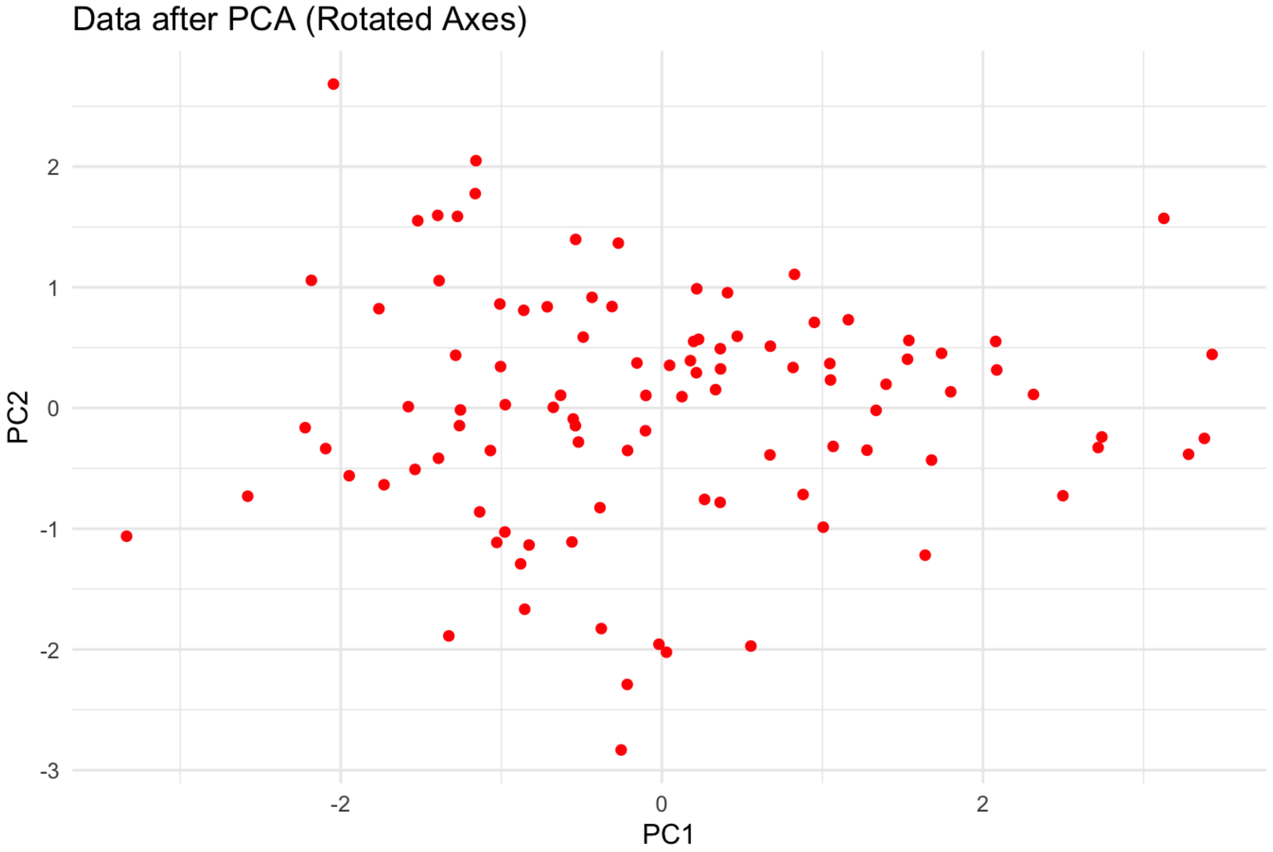
- Correlation is not affected by shifts (adding a constant).

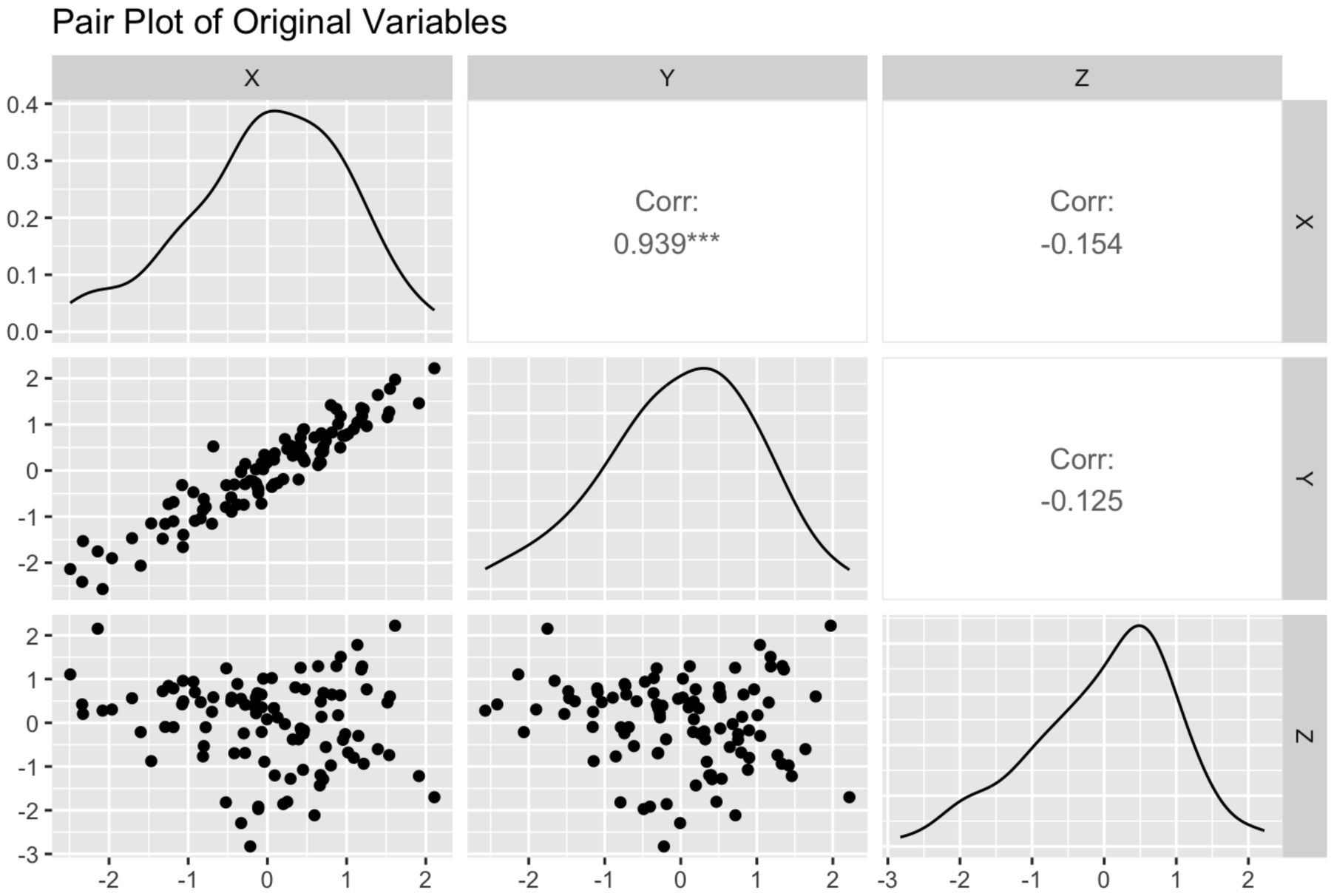
$$\text{Cor}(X + c, Y + d) = \text{Cor}(X, Y)$$

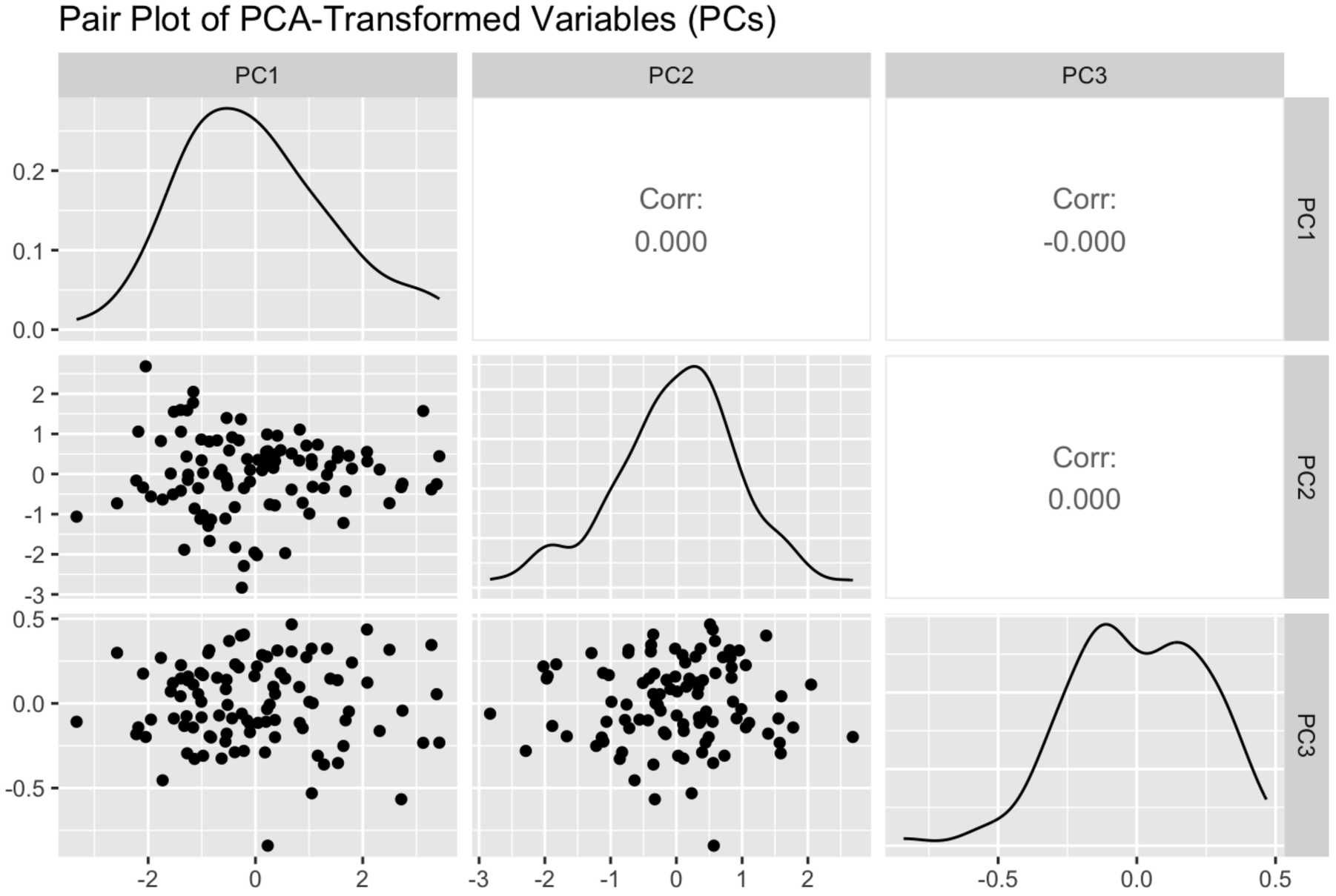
Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components

- Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the ***principal components*** of the data.
- **The eigenvectors** of the Covariance matrix are actually **the directions of the axes where there is the most variance (most information)** and that we c**all Principal Components.**
- **The eigenvalues** are simply the coefficients attached to eigenvectors, which **give the amount of variance carried in each Principal Component.**
- By ranking your eigenvectors in order of their eigenvalues, highest to lowest, you get the principal components in order of significance.


Original Data
4
0
-4
-8
Y
-2
-1
0
1
X


Data after PCA (Rotated Axes)
2
1
0
-1
-2
-3
PC2
-2
0
2
PC1


Pair Plot of Original Variables
X
Y
Z
Corr:
0.939***
Corr:
-0.154
Corr:
-0.125
X
Y
Z
0.4
0.3
0.2
0.1
0.0
2
1
0
-1
-2
-3


Pair Plot of PCA-Transformed Variables (PCs)
PC1
PC2
PC3
Corr:
0.000
Corr:
-0.000
Corr:
0.000
0.2
0.1
0.0
2
1
0
-1
-2
-3
0.5
0.0
-0.5
-2
0
2
-3
-2
-1
0
1
2
-0.5
0.0
0.5

Step 4: Create a Feature Vector

- Computing the eigenvectors and ordering them by their eigenvalues in descending order, allow us to **find the principal components in order of significance.**

- In this step, what we do is, to **choose whether to keep all these components or discard those of lesser significance** (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call **Feature vector.**

- The feature vector is simply a matrix that has **as columns the eigenvectors of the components that we decide to keep.** This makes it the first step towards dimensionality reduction, because if we choose to keep only q eigenvectors (components) out of d, the final data set will have only q dimensions.

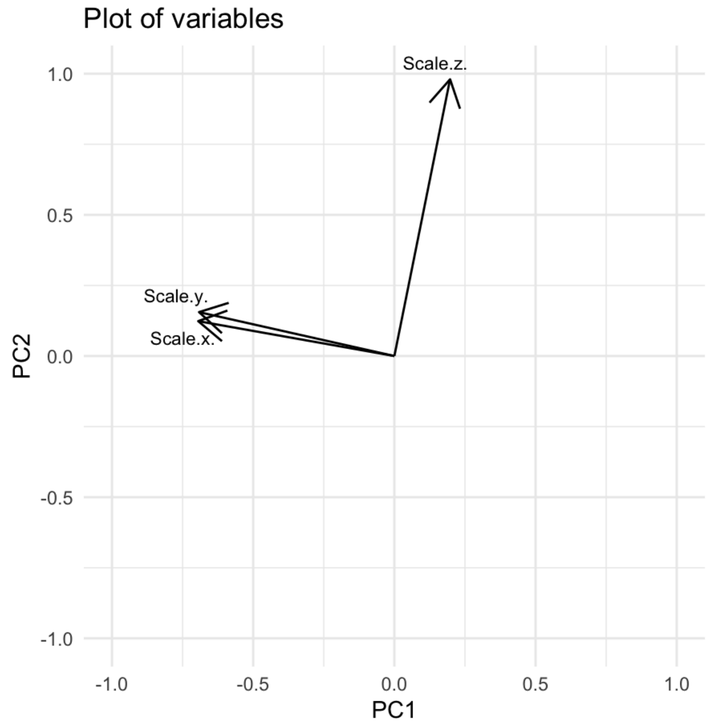
```
> eigenvalues
[1] 1.97893687 0.96082566 0.06023747
```

In this case we might choose to keep only the first **2 eigenvectors** since most of the variance is explained by these dimensions

Step 5: Recast the Data Along the Principal Components Axes

The aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components (hence the name Principal Components Analysis).

```
[1] "Feature Vector (Top k Eigenvectors):"
> print(feature_vector)
         PC1       PC2
X -0.6947170 0.1235059
Y -0.6916924 0.1555242
Z  0.1973068 0.9800809
```



Example: Social capital

https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/principalcomponentanalysisofsocialcapitalindicators

- Social Capital refers to the connections between people and collective attitudes that result in a well-functioning and close-knit society. Connections have been noted between increased social capital and positive well-being, economic growth and sustainability.
- measure of community involvement and cohesion
- source of insight for those wishing to facilitate community well-being and social cohesion.
- Despite this, and growing policy interest in the topic, social capital has remained a difficult concept to measure.

Measuring social capital

three forms; bonding, bridging and linking capital:

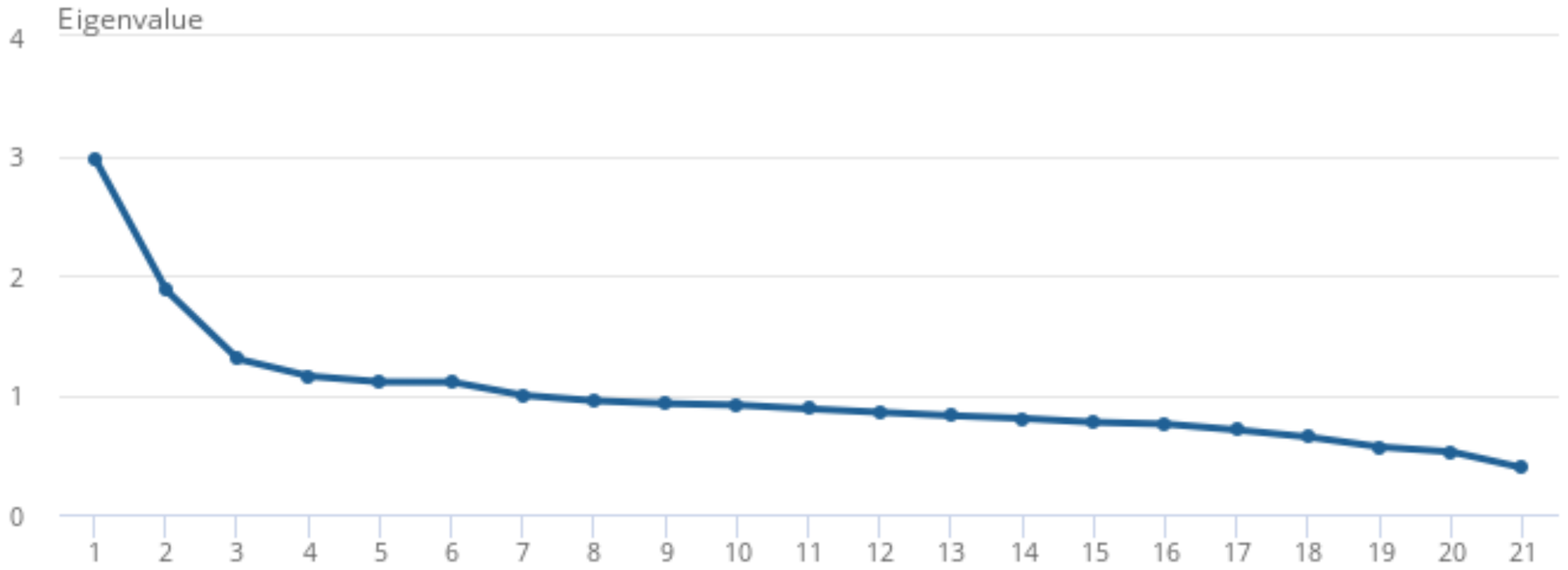
- bonding capital refers to horizontal ties within a group; this can mean the relationships between friends and family, or relationships between people of the same sex, ethnicity or religious group
- bridging capital refers to ties between individuals that exist between social groups, such as those between colleagues or neighbours
- linking capital refers to the ties between an individual and others with greater resources or power, such as a boss or a teacher
- **To capture these different facets, social capital is currently measured by the Office for National Statistics (ONS) through a set of 25 headline**

The goal of principal component analysis (PCA) is to transform a set of possibly correlated variables into a smaller set of uncorrelated variables called principal components.

Feel that you belong to your neighbourhood	Interested in politics
People around where you live are willing to help their neighbours	Feel they can affect decisions in the local area
Regularly stop and talk with people in your neighbourhood	Feel most people can be trusted
Feel most people in their neighbourhood can be trusted	Receive help from a child over 16 not living with you
Borrow things and exchange favours with neighbours	Give help to a sick, disabled or elderly person living or not living with them
Participate in the activities of an organisation or group	Have at least one close friend
Member of an organisation or group	Go out socially or visit friends when you feel like it
Volunteered more than once in the last 12 months	Belong to any social network
Participate in the activities of a political party or trade union	Have a spouse or partner, friend or family member to rely on if you have a serious problem
Feel that voting is a civic duty	Are friends with people of a different age, ethnicity, level of education or income
	Feel safe walking alone after dark

Figure 1: Scree plot showing eigenvalues from principal component analysis, UK

between 2009 to 2011 and 2016 to 2018



The five-component model explained 42% of the variance in the data.

Source: Understanding Society, the UK Longitudinal Household Survey

	Component 1	Component 2	Component 3	Component 4
	Neighbourhood relationships'	Political engagement'	Organised social and civic engagement'	Friendship and safety'
Feel that you belong to your neighbourhood	0.74*	0,09	0,03	-0.05
Regularly stop and talk with people in your neighbourhood	0.73*	-0.06	0,06	-0.1
People around where you live are willing to help their neighbours	0.7*	0,16	-0.01	0,08
Borrow things and exchange favours with neighbours	0.63*	-0.03	0,1	0,11
Feel most people in their neighbourhood can be trusted	0.62*	0,29	-0.01	0,04
Interested in politics	-0.05	0.64*	0,15	-0.11
Feel that voting is a civic duty	0,05	0.64*	0,12	-0.24
Feel they can affect decisions in the local area	-0.01	0.55*	0,04	0,15
Feel most people can be trusted	0,15	0.49*	0,12	0,05
Participate in the activities of an organisation or group	0,09	0,09	0.8*	0,09
Member of an organisation or group	0,06	0,19	0.76*	0,09
Volunteered more than once in the last 12 months	0,04	0,15	0.51*	0,02
Participate in the activities of a political party or trade union	-0.02	-0.02	0.42*	-0.03
Belong to any social network	-0.14	-0.13	0,03	0.63*
Have at least one close friend	0,11	-0.04	0,07	0.44*
Feel safe walking alone after dark	0,16	0,29	0,03	0.38*
Go out socially or visit friends when you feel like it	0,17	0,18	0,04	0,28
Have a spouse or partner, friend or family member to rely on if you have a serious problem	0,14	0,27	-0.01	0,1
Give help to a sick, disabled or elderly person living or not living with them	0,03	-0.19	0,21	-0.43
Are friends with people of a different age, ethnicity, level of education or income	-0.14	-0.14	0,09	0,24

	Component 1	Component 2	Component 3	Component 4
	Neighbourhood relationships'	Political engagement'	Organised social and civic engagement'	Friendship and safety'
Feel that you belong to your neighbourhood	0.74*	0,09	0,03	-0.05
Regularly stop and talk with people in your neighbourhood	0.73*	-0.06	0,06	-0.1
People around where you live are willing to help their neighbours	0.7*	0,16	-0.01	0,08
Borrow things and exchange favours with neighbours	0.63*	-0.03	0,1	0,11
Feel most people in their neighbourhood can be trusted	0.62*	0,29	-0.01	0,04
Interested in politics	-0.05	0.64*	0,15	-0.11
Feel that voting is a civic duty	0,05	0.64*	0,12	-0.24
Feel they can affect decisions in the local area	-0.01	0.55*	0,04	0,15
Feel most people can be trusted	0,15	0.49*	0,12	0,05
Participate in the activities of an organisation or group	0,09	0,09	0.8*	0,09
Member of an organisation or group	0,06	0,19	0.76*	0,09
Volunteered more than once in the last 12 months	0,04	0,15	0.51*	0,02
Participate in the activities of a political party or trade union	-0.02	-0.02	0.42*	-0.03
Belong to any social network	-0.14	-0.13	0,03	0.63*
Have at least one close friend	0,11	-0.04	0,07	0.44*
Feel safe walking alone after dark	0,16	0,29	0,03	0.38*
Go out socially or visit friends when you feel like it	0,17	0,18	0,04	0,28
Have a spouse or partner, friend or family member to rely on if you have a serious problem	0,14	0,27	-0.01	0,1
Give help to a sick, disabled or elderly person living or not living with them	0,03	-0.19	0,21	-0.43
Are friends with people of a different age, ethnicity, level of education or income	-0.14	-0.14	0,09	0,24

	Component 1	Component 2	Component 3	Component 4
	Neighbourhood relationships'	Political engagement'	Organised social and civic engagement'	Friendship and safety'
Feel that you belong to your neighbourhood	0.74*	0,09	0,03	-0.05
Regularly stop and talk with people in your neighbourhood	0.73*	-0.06	0,06	-0.1
People around where you live are willing to help their neighbours	0.7*	0,16	-0.01	0,08
Borrow things and exchange favours with neighbours	0.63*	-0.03	0,1	0,11
Feel most people in their neighbourhood can be trusted	0.62*	0,29	-0.01	0,04
Interested in politics	-0.05	0.64*	0,15	-0.11
Feel that voting is a civic duty	0,05	0.64*	0,12	-0.24
Feel they can affect decisions in the local area	-0.01	0.55*	0,04	0,15
Feel most people can be trusted	0,15	0.49*	0,12	0,05
Participate in the activities of an organisation or group	0,09	0,09	0.8*	0,09
Member of an organisation or group	0,06	0,19	0.76*	0,09
Volunteered more than once in the last 12 months	0,04	0,15	0.51*	0,02
Participate in the activities of a political party or trade union	-0.02	-0.02	0.42*	-0.03
Belong to any social network	-0.14	-0.13	0,03	0.63*
Have at least one close friend	0,11	-0.04	0,07	0.44*
Feel safe walking alone after dark	0,16	0,29	0,03	0.38*
Go out socially or visit friends when you feel like it	0,17	0,18	0,04	0,28
Have a spouse or partner, friend or family member to rely on if you have a serious problem	0,14	0,27	-0.01	0,1
Give help to a sick, disabled or elderly person living or not living with them	0,03	-0.19	0,21	-0.43
Are friends with people of a different age, ethnicity, level of education or income	-0.14	-0.14	0,09	0,24

	Component 1	Component 2	Component 3	Component 4
	Neighbourhood relationships'	Political engagement'	Organised social and civic engagement'	Friendship and safety'
Feel that you belong to your neighbourhood	0.74*	0,09	0,03	-0.05
Regularly stop and talk with people in your neighbourhood	0.73*	-0.06	0,06	-0.1
People around where you live are willing to help their neighbours	0.7*	0,16	-0.01	0,08
Borrow things and exchange favours with neighbours	0.63*	-0.03	0,1	0,11
Feel most people in their neighbourhood can be trusted	0.62*	0,29	-0.01	0,04
Interested in politics	-0.05	0.64*	0,15	-0.11
Feel that voting is a civic duty	0,05	0.64*	0,12	-0.24
Feel they can affect decisions in the local area	-0.01	0.55*	0,04	0,15
Feel most people can be trusted	0,15	0.49*	0,12	0,05
Participate in the activities of an organisation or group	0,09	0,09	0.8*	0,09
Member of an organisation or group	0,06	0,19	0.76*	0,09
Volunteered more than once in the last 12 months	0,04	0,15	0.51*	0,02
Participate in the activities of a political party or trade union	-0.02	-0.02	0.42*	-0.03
Belong to any social network	-0.14	-0.13	0,03	0.63*
Have at least one close friend	0,11	-0.04	0,07	0.44*
Feel safe walking alone after dark	0,16	0,29	0,03	0.38*
Go out socially or visit friends when you feel like it	0,17	0,18	0,04	0,28
Have a spouse or partner, friend or family member to rely on if you have a serious problem	0,14	0,27	-0.01	0,1
Give help to a sick, disabled or elderly person living or not living with them	0,03	-0.19	0,21	-0.43
Are friends with people of a different age, ethnicity, level of education or income	-0.14	-0.14	0,09	0,24

Limitation of PCA

Interpretability of Components: Principal components are linear combinations of original features, which can make them difficult to interpret in a meaningful way, especially in applications where understanding specific features is important.

Training and Validation Datasets: PCA is typically performed on a specific dataset, meaning the principal components are specific to the patterns in that data. The same components may not capture the essential structure in a different dataset, especially if it has variations in distributions, features, or noise levels. This limits the generalizability of the PCA results to new datasets, impacting external validity.

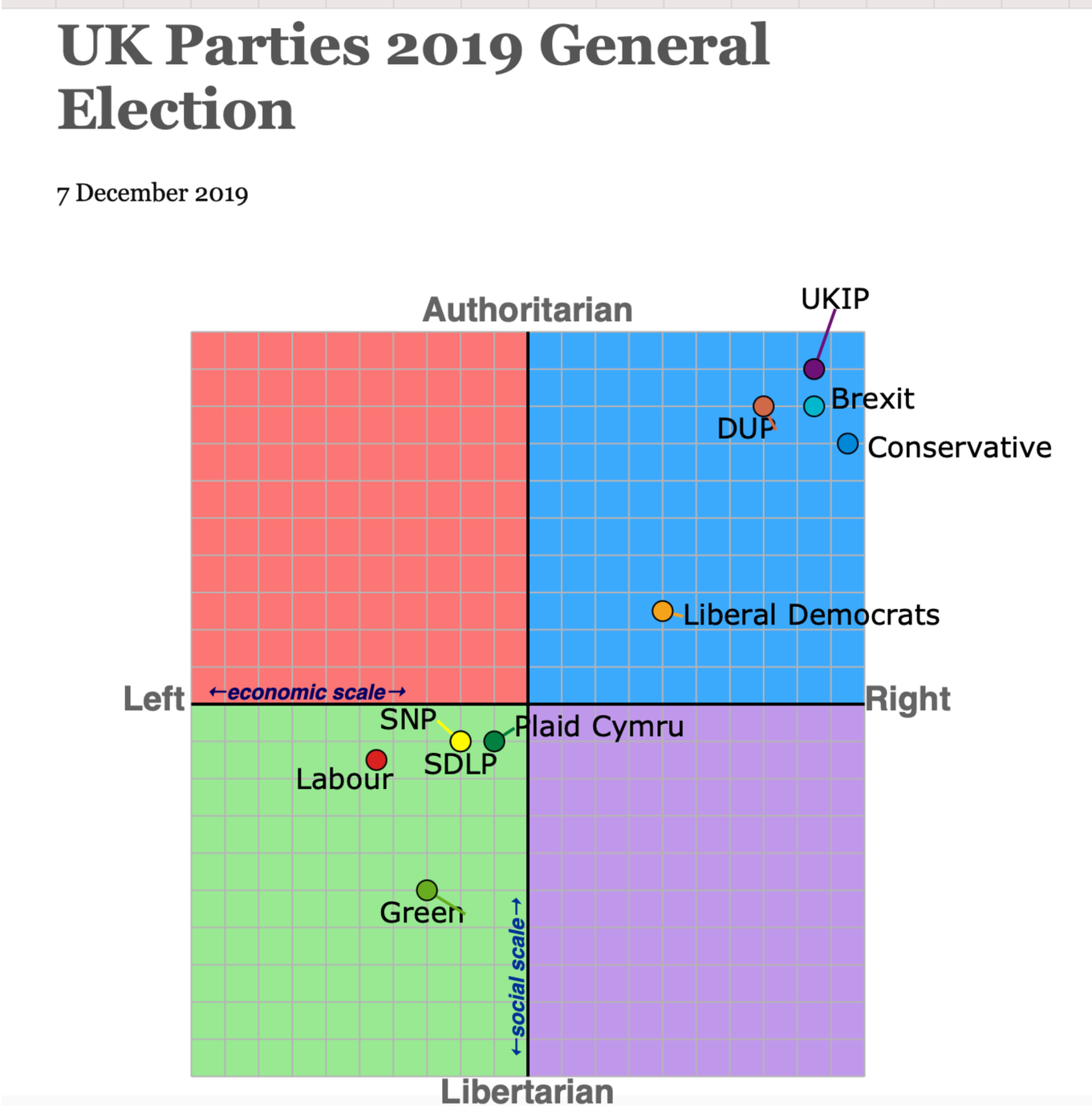
Limitation of PCA

Feature Scaling: PCA is sensitive to the scaling of features. If data is preprocessed differently (e.g., standardized vs. unstandardized), the resulting components may not generalize to datasets where different scaling has been applied.

Nonlinearity and Complexity: PCA assumes linearity in the data, which may not hold in complex real-world applications. Many real-world patterns are nonlinear, so components derived in controlled, simplified datasets may not accurately reflect the true relationships in more intricate or interactive real-world data.

Political ideology

- Political ideology is a good example of complex variable
- It ranges from economic to social issues, environment, solidarity, migrations etc.
- **Need to use variable reduction techniques**
- **Often use Principal Component Analysis**



Principal Component analysis

- Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed.

- The idea of PCA is simple — **reduce the number of variables of a data set, while preserving as much information as possible.**

Evolution of songs' features over the years.

Dataset of Top 100 Billboard. To this data set is associated a characterization of the songs according to several features (danceability, mode, tempo…), provided by the Spotify API.

```
# Load data
# Billbord ranking
library(tidyverse)
billboard <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mast
# Songs features based on Spotify API
features <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste

head(features)
```

https://bjnnowak.netlify.app/2021/09/15/r-pca-with-tidyverse/

`billboard.csv`

variable	class	description
url	character	Billboard Chart URL
week_id	character	Week ID
week_position	double	Week position 1: 100
song	character	Song name
performer	character	Performer name
song_id	character	Song ID, combo of song/singer
instance	double	Instance (this is used to separate breaks on the chart for a given song. Example, an instance of 6 tells you that this is the sixth time this song has appeared on the chart)
previous_week_position	double	Previous week position
peak_position	double	Peak position as of that week
weeks_on_chart	double	Weeks on chart as of that week

Data from Spotify

variable	class	description
song_id	character	Song ID
performer	character	Performer name
song	character	Song
spotify_genre	character	Genre
spotify_track_id	character	Track ID
spotify_track_preview_url	character	Spotify URL
spotify_track_duration_ms	double	Duration in ms
spotify_track_explicit	logical	Is explicit
spotify_track_album	character	Album name

danceability	double	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
energy	double	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
key	double	The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
loudness	double	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
mode	double	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
speechiness	double	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
acousticness	double	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
instrumentalness	double	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
liveness	double	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
valence	double	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
tempo	double	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
time_signature	double	Time signature
spotify_track_popularity	double	Popularity

Cleaning the data

```
library(tidyverse)

bill_prep<-billboard |>
  # Keep only 1st appearance on Billboard
  filter(
    (weeks_on_chart==1)&(instance==1)
  ) |>
  # Add Year column
  mutate(year=format(
    as.Date(week_id,"%m/%d/%Y"),format="%Y")
  ) |>
  # Set year as numeric
  mutate(year=as.numeric(year))

# Add year to songs' features data
features_prep<-features |>
  left_join(bill_prep,by="song_id")
```

Selecting variables

```
PCA_data<-features_prep |>
  select(
    # Variables of interest for PCA
    c(danceability,energy,instrumentalness,
      key,acousticness,mode,valence,tempo,
      time_signature,speechiness,loudness,liveness,
    # Add year as supplementary variable
      year
    )
  ) |>
  # Remove rows with NA
  drop_na()
```

Correlation

- On average, how do two variables move together?
- Positive (negative) correlation: When x is larger than its mean, y is likely (unlikely) to be larger than its mean
- Positive (negative) correlation: data cloud slopes up (down) High correlation: data cluster tightly around a line
- Mathematical definition of correlation coefficient:

$$\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x})}{SD_x}\times\frac{y_i-\bar{y})}{SD_y}\right)$$

Properties of Correlation Coefficient

- Correlation is between –1 and 1
- Order does not matter: cor(x, y) = cor(y, x)
- Not affected by changes of scale:
  - cor(x,y) = cor(ax+b,cy+d) for any numbers a, b, c, and d
  - Celsius vs. Fahrenheit; cm vs. inch; yen vs. dollar etc.

Correlation matrix

```
#install.packages("corrplot")
library(corrplot)
```

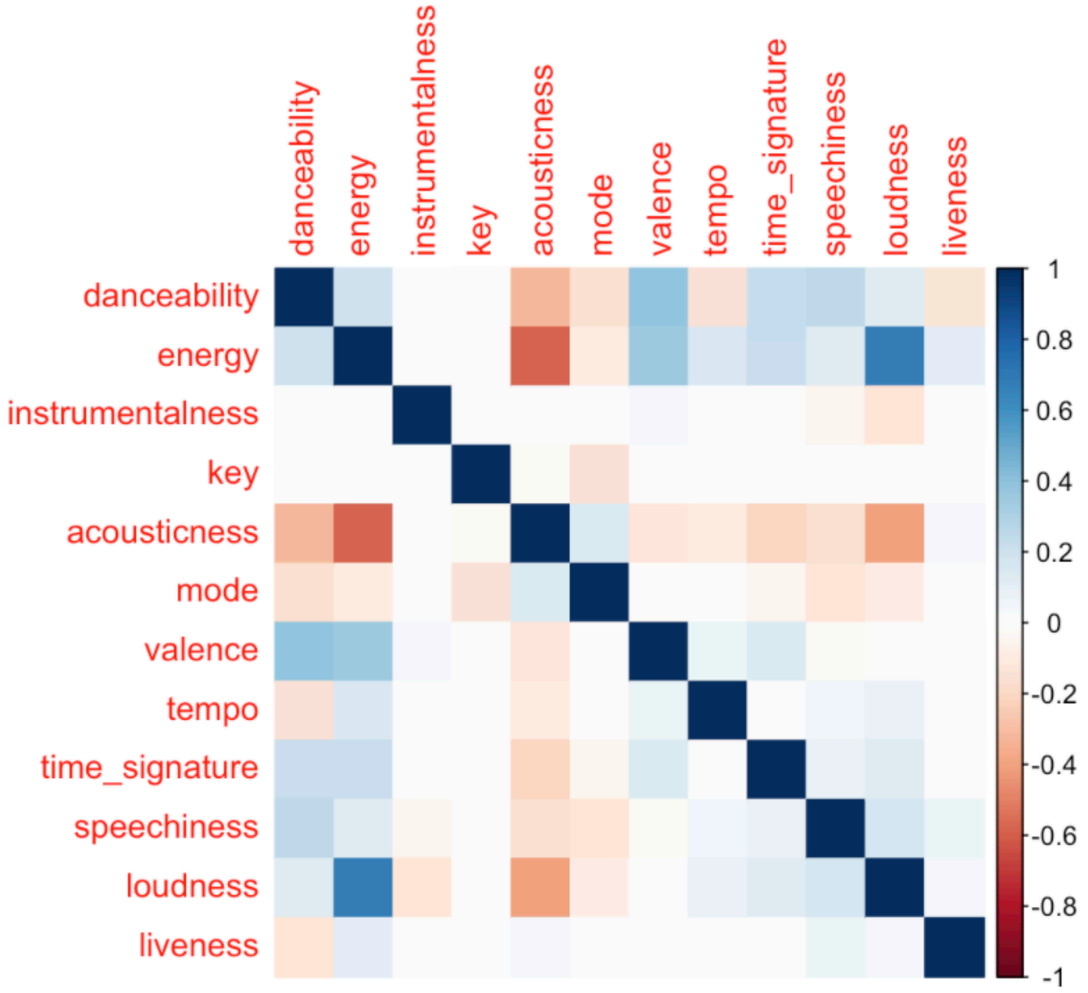
```
corrplot 0.92 loaded
```

```
corrMatrix<-PCA_data |>
  select(-year) |>
  cor()
corrMatrix
```

```
                danceability      energy instrumentalness          key
danceability      1.00000000  0.20184246     -0.001288590  0.014098210
energy            0.20184246  1.00000000     -0.001072350  0.021919303
instrumentalness -0.00128859 -0.00107235      1.000000000  0.003452150
key               0.01409821  0.02191930      0.003452150  1.000000000
acousticness     -0.31458912 -0.58583638      0.028317785 -0.021313380
mode             -0.16047340 -0.10185943     -0.010026540 -0.142663131
valence           0.38517701  0.35427457      0.049672657  0.012421507
tempo            -0.14500045  0.15984193      0.002528736 -0.014622950
time_signature    0.22298832  0.22761631      0.008582511  0.009891870
speechiness       0.25504753  0.13313155     -0.057595385  0.022231053
loudness          0.13752454  0.68440146     -0.134586882  0.008191408
liveness         -0.13044205  0.11196988     -0.010956020 -0.001773604
                acousticness        mode     valence        tempo
danceability     -0.31458912 -0.16047340  0.38517701 -0.145000451
```

Correlation plot

```
corrplot(corrMatrix, method="color")
```



principal components

- Principal components are “new variables” that explain correlation structure of original data
- They are ordered, i.e. the first explain the most
- In this case, PC1 explain 21% of the variability in the dataset, PC2 11%
- If we chose to use the first two PCs we explain 33% of the original variability

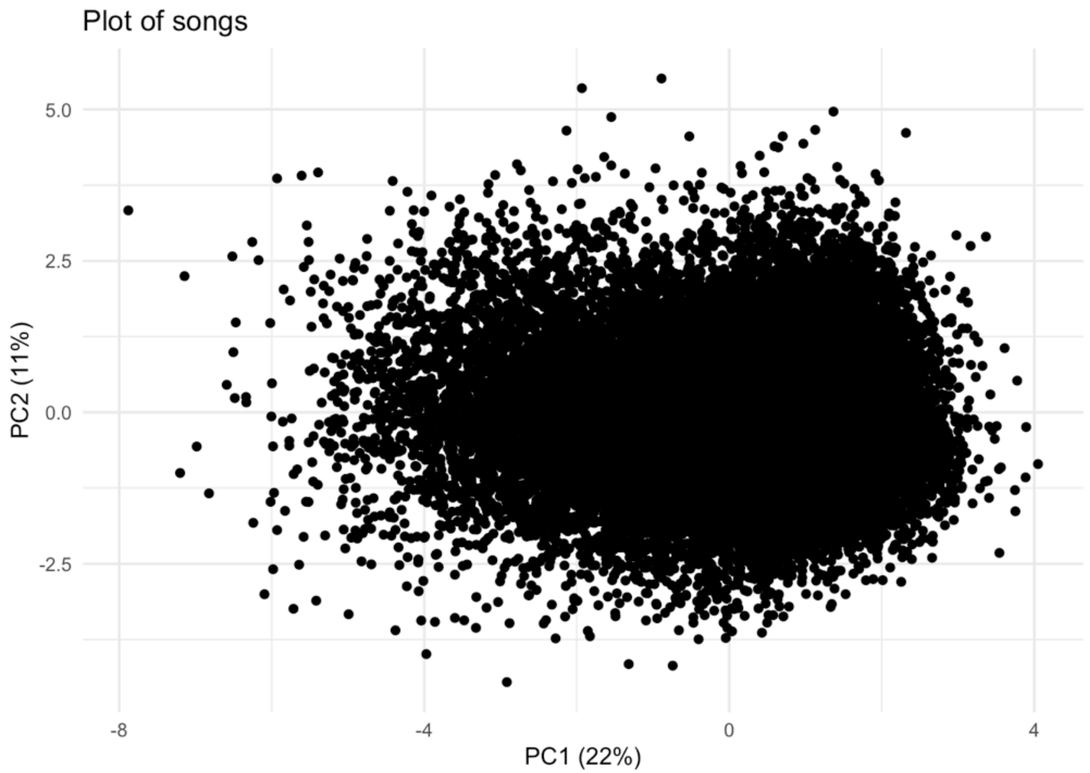
```r
library(broom)

PCA |>
  tidy(matrix = "eigenvalues")
```

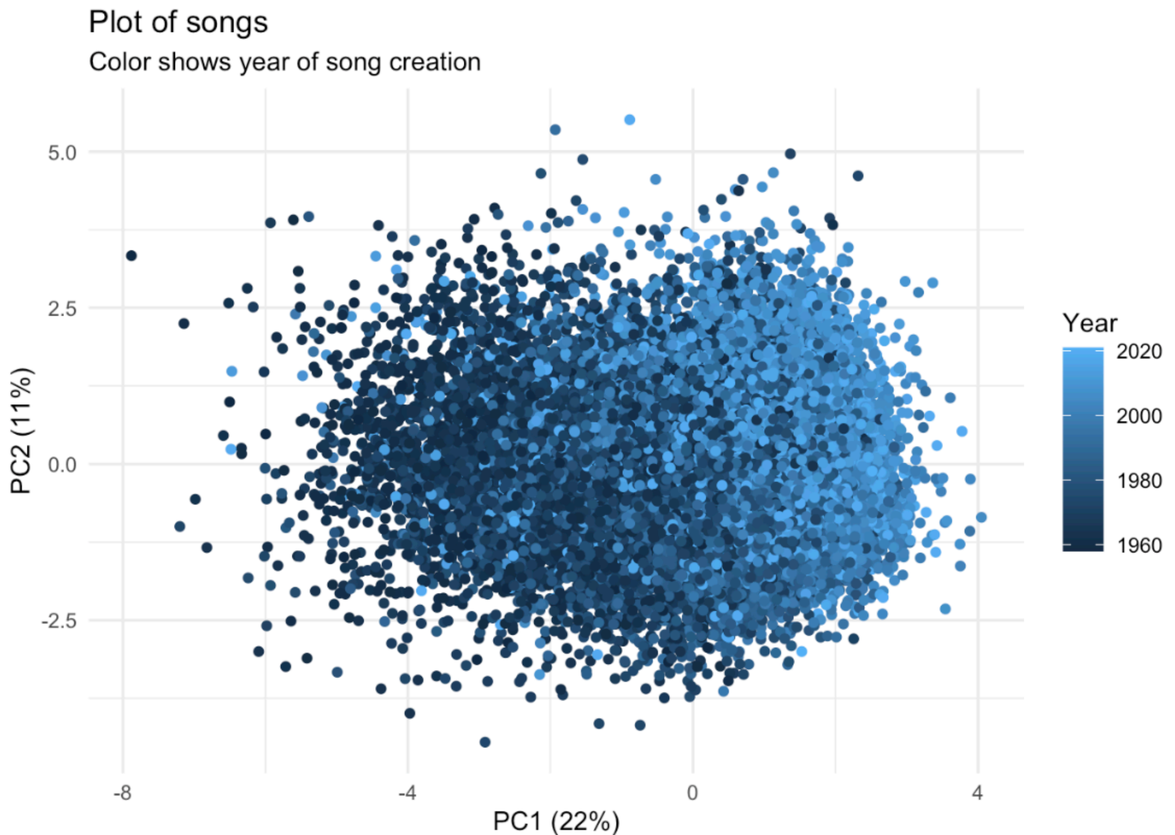
```
# A tibble: 12 × 4
      PC std.dev percent cumulative
   <dbl>   <dbl>   <dbl>      <dbl>
 1     1   1.61   0.217       0.217
 2     2   1.17   0.114       0.331
 3     3   1.09   0.0993      0.431
 4     4   1.04   0.0893      0.520
 5     5   1.01   0.0850      0.605
 6     6   0.980  0.0801      0.685
 7     7   0.965  0.0776      0.762
 8     8   0.928  0.0717      0.834
 9     9   0.894  0.0666      0.901
10    10   0.760  0.0482      0.949
11    11   0.660  0.0363      0.985
12    12   0.420  0.0147      1
```

```
PCA_indiv<-PCA%>%
  broom::augment(PCA_data)

# Plot of individuals
ggplot(
  data=PCA_indiv,
  aes(.fittedPC1, .fittedPC2))+
  geom_point()+
  labs(
    title = 'Plot of songs',
    x='PC1 (22%)',
    y='PC2 (11%)',
    color='Year'
  )+
  theme_minimal()
```



```r
# Plot of individuals
ggplot(
  data=PCA_indiv,
  aes(.fittedPC1, .fittedPC2,color=year))+
  geom_point()+
  labs(
    title = 'Plot of songs',
    subtitle = 'Color shows year of song creation',
    x='PC1 (22%)',
    y='PC2 (11%)',
    color='Year'
  )+
  theme_minimal()
```



```r
PCA_var<-PCA |>
  # Extract variable coordinates
  tidy(matrix = "rotation") %>%
  # Format table form long to wide
  pivot_wider(names_from = "PC", names_prefix = "PC", values_from = "value") |>
  # Rename column with variable names
  rename(Variable=column) |>
  # 'Clean' variable names
  # Upper case on first letter
  mutate(Variable=stringr::str_to_title(Variable)) |>
  # Change '_' for space
  mutate(Variable=stringr::str_replace_all(Variable,"_"," "))

head(PCA_var)
```

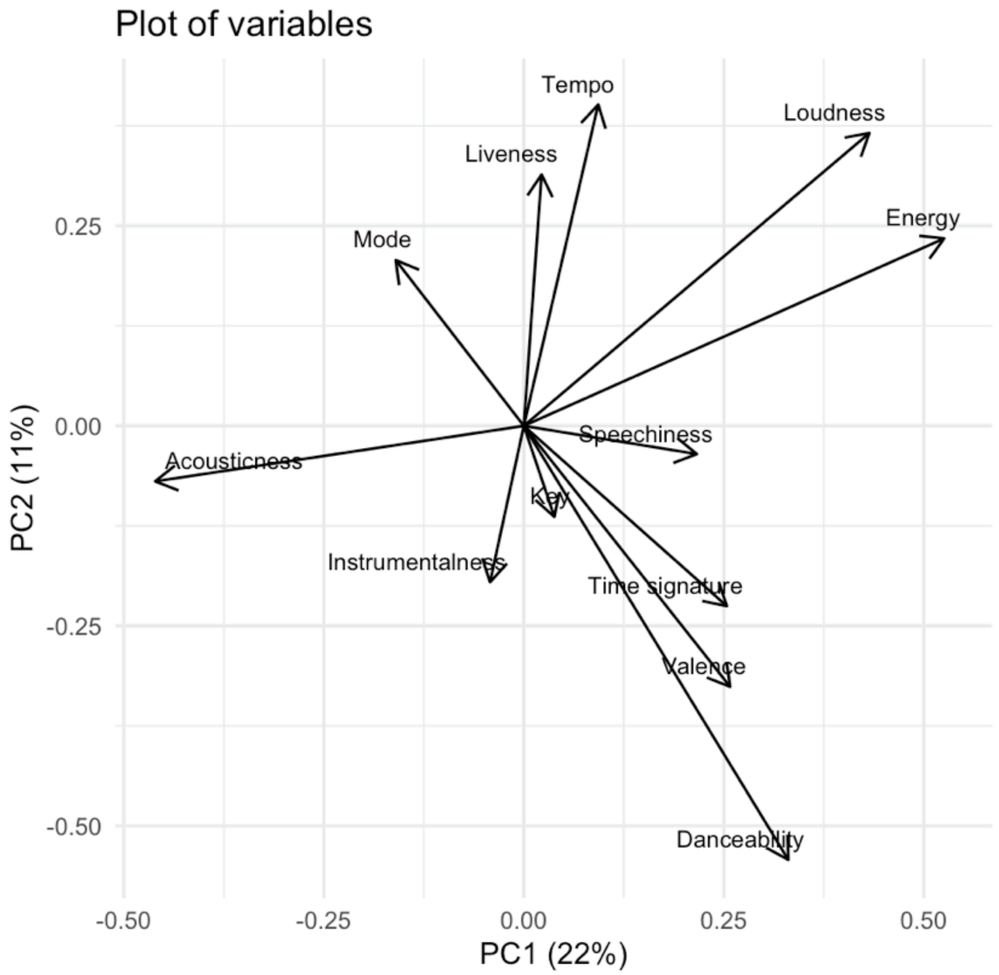
```
# A tibble: 6 × 13
  Variable          PC1     PC2      PC3     PC4     PC5     PC6     PC7     PC8
  <chr>           <dbl>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 Danceability   0.331  -0.542   0.00516 -0.164  -0.133  -0.0879  0.140   0.155
2 Energy         0.525   0.234  -0.161    0.0851  0.0569  0.145  -0.0808  0.114
3 Instrumental… -0.0422 -0.195  -0.304    0.521   0.0581 -0.327  -0.635   0.106
4 Key            0.0382 -0.114   0.382    0.559   0.233   0.395   0.248  -0.245
5 Acousticness  -0.461  -0.0693 -0.0153   0.0601 -0.212   0.0326  0.149  -0.0402
6 Mode          -0.160   0.207  -0.497   -0.329  -0.0900  0.107   0.0612 -0.220
# ℹ 4 more variables: PC9 <dbl>, PC10 <dbl>, PC11 <dbl>, PC12 <dbl>
```

Plot the variables

```r
# Load ggrepel to avoid variable names to overlap
library(ggrepel)

var<-ggplot(data=PCA_var,aes(PC1, PC2)) +
  # Add variables arrows
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow(
      length = unit(0.03, "npc"),
      ends = "first"
    )
  )+
  # Add variables names
  geom_text_repel(
    aes(label = Variable),
    hjust = 1,size=3,
    min.segment.length = Inf,
    nudge_x=0.01,nudge_y=0.01
  ) +
  coord_fixed()+
  labs(
    title = 'Plot of variables',
    x='PC1 (22%)',
    y='PC2 (11%)',
    color='Year'
  )+
  theme_minimal()

var
```



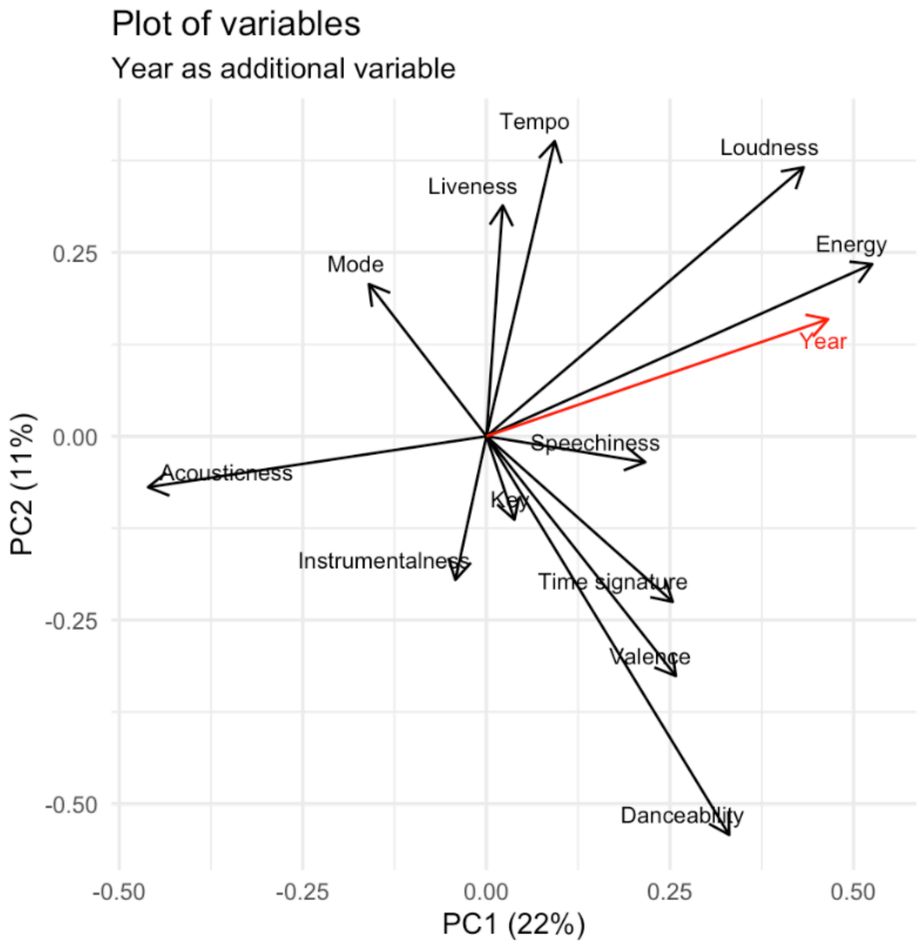
Adding Year

```
year_coord<-as.data.frame(
  # Calculate correlation of year with PCA axis
  cor(PCA_data$year,PCA$x)
  )%>%
  # Add name of the variable
  mutate(Variable="Year")

year_coord
```

```
        PC1       PC2       PC3        PC4       PC5        PC6        PC7
1 0.4652744 0.1592881 0.3657629 -0.2336913 0.1282045 -0.1212724 -0.1761371
          PC8        PC9       PC10      PC11       PC12 Variable
1 -0.02687627 0.09279391 -0.1115629 -0.221641 0.04197899     Year
```

```
var+
  geom_segment(
    data=year_coord,
    color="red",
    xend = 0, yend = 0,
    arrow = arrow(
      length = unit(0.03, "npc"),
      ends = "first"
    )
  )+
  geom_text_repel(
    data=as.data.frame(year_coord),
    aes(label = Variable),
    color="red",hjust = 1,size=3,
    min.segment.length = Inf,
    nudge_x=0.02,nudge_y=-0.02
  )+
  labs(
    subtitle="Year as additional variable"
  )
```



We can see that the “Energy” variable is the most strongly correlated with the “Year” variable: Billboard hits tend to become more and more energetic over the years.

Other examples of PCA

Intelligence

- In1904,Charles Spearman noted that children'sperformance across unrelated school subjects, like Classics, Math, and Music, were positively correlated.
- He hypothesized that all cognitive ability could be traced to a single "general intelligence" factor, which he called the *g* factor.
- Later, IQ tests were designed to try to measure this *g* factor. It attempts to quantify intelligence along a single dimension.
- Psycological research uses statistical methods such as PCA to identify "cognitive factors"

.

Other examples of PCA

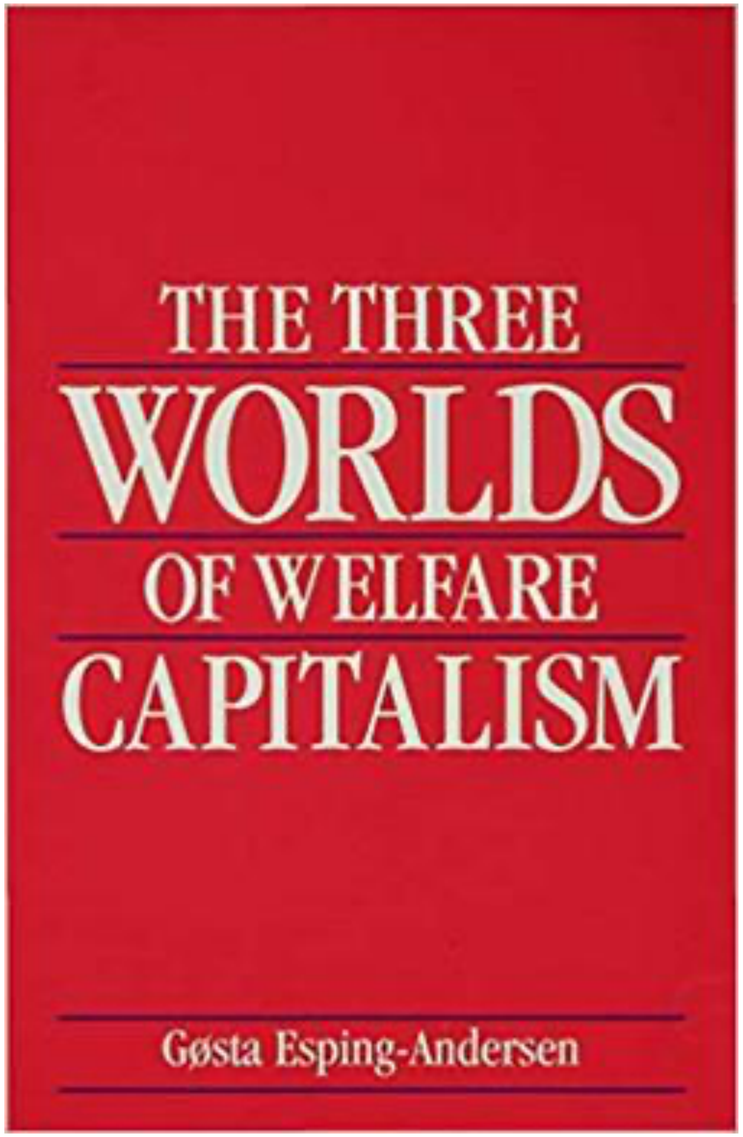
Classification of Welfare regimes

- EA (1990) analysed several policy indicators to classify countries into three welfare regimes

1. Liberal regimes

2. Conservative regimes

3. Social-Democratic regimes



Other examples of PCA

Genetics

- DNA is composed by millions of “variables”
- DNA variations have a correlation structure due to history of migration of ancestral populations
- PCA is used to simplify and identify ancestral populations