Resit exam questions (2021-07-15)

Resit exam (July 2021)

- This was an oral resit exam
- Students had to answer ~20 questions in ~20-25 minutes
- Questions were chosen at random by throwing a dice, advancing the number shown on the dice, and asking that question, then throwing the dice again, and so on ...

Exam protocol

- Choose language es/ca/en
- We are recording now, the recording will stay in the platform with access only to me, me unless the university authorities request it for some reason
- Please place your mobile in airplane mode (unless you’re using it for communicating with me)
- Please briefly show me the room where you are giving your exam
- Please briefly share with me (“present”) your entire computer screen
- We will start with a topic you think you’ve studied more, then we will go back to slide #3 and roll the dice to determine each question; if we land on a question you’ve already answered or a non-question slide, I ask you the next one; if we get to the end we restart
- I’ll ask you questions for 20 minutes starting now – pick the initial topic please

TT02 Data, methods, scenarios

What is dependency-oriented data?

TT02 Data, methods, scenarios

If you use one-hot encoding for “Investments” and remove the original column, how many columns will the resulting table have?

Id	Name	Email	Investments
231	Albert Master	albert.master@gmail.com	Bonds
210	Alfred Alan	aalan@gmail.com	Stocks
256	Alison Smart	asmart@biztalk.com	Residential Property
211	Ally Emery	allye@easymail.com	Stocks
248	Andrew Phips	andyp@mycorp.com	Stocks
234	Andy Mitchel	andym@hotmail.com	Stocks
226	Angus Robins	arobins@robins.com	Bonds
241	Ann Melan	ann_melan@iinet.com	Residential Property
225	Ben Bessel	benb@hotmail.com	Stocks
235	Bensen Romanolf	benr@albert.net	Bonds

TT03. Data preparation: data types

How do you convert a **numerical** variable into a **categorical** value?

TT03. Data preparation: data types

Suppose data are:
0, 4, 12, 16, 16, 18, 24, 26

Divide into three equi-width bins.

Divide into three equi-depth bins.

TT04. Data prep.: integration & cleaning

What does it mean to do **object matching**?

TT04. Data prep.: integration & cleaning

Name two reasons why we perform data cleaning

TT04. Data prep.: integration & cleaning

What is a **range constraint**?

TT04. Data prep.: integration & cleaning

Give an example of **cross-field validation**

TT04. Data prep.: integration & cleaning

Suppose in a database for sales of a company we have the zip code of a sale but not the province. What shall we do?

TT04. Data prep.: integration & cleaning

What is **min-max scaling**?

TT04. Data prep.: integration & cleaning

We have a variable taking values {1, 2, 3, 4, 5}
$\mu=3.0$, $\sigma=1.41$

Normalize by using standardization

TT04. Data prep.: integration & cleaning

Describe clearly how to perform seasonal standardization

TT04. Data prep.: integration & cleaning

We have a variable taking values {-4, 0, 5}

Normalize by using min-max scaling

TT05. Reduction and transformation

Suppose a population contains 10 women and 15 men. Describe how to sample 4 elements using **stratified sampling by gender**

TT06. Similarity on numerical data

Suppose data points are represented by a single feature, which is categorical/nominal

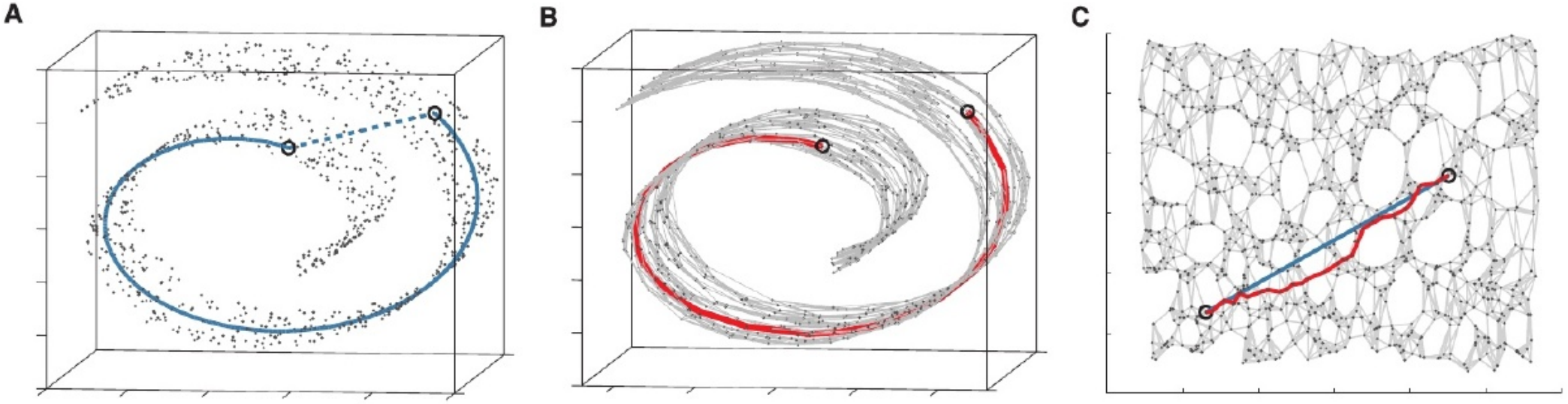
Describe for this dataset (1) a similarity function, and (2) a distance function

TT06. Similarity on numerical data

Explain what is **the curse of dimensionality**

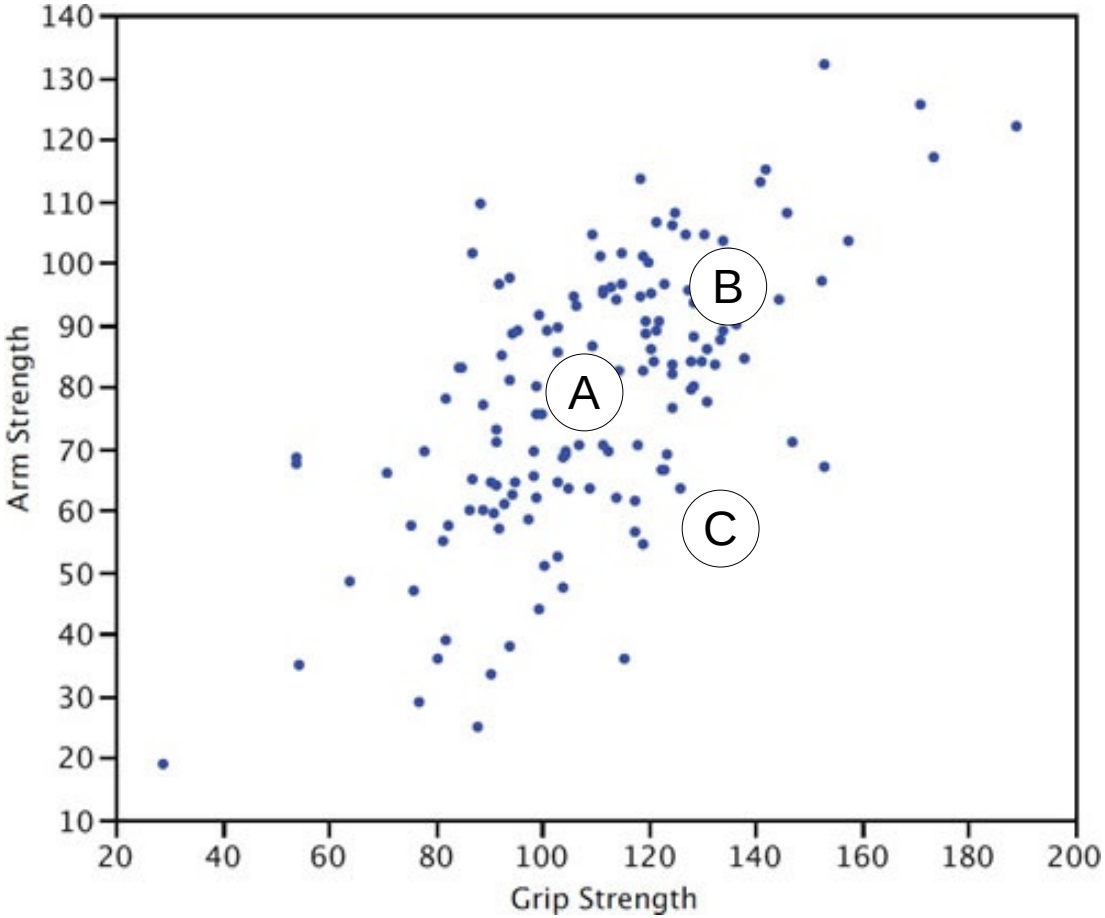
TT06. Similarity on numerical data

Explain how ISOMAP works



TT06. Similarity on numerical data

The Mahalanobis distance would consider that d(A,B) < d(A,C) or d(A,B) > d(A,C) ?



TT07. Similarity: beyond numerical

Write and explain the formula for the Goodall measure

TT07. Similarity: beyond numerical

Compute the **Jaccard similarity** between these two sets:

{carrot, apple, banana}

{tomato, banana}

TT07. Similarity: beyond numerical

What is a **PAM matrix**?

TT08. Near duplicates

Explain the principle behind min hashing, and give an example

TT08. Near duplicates

How many **different** 3-word-gram shingles are contained in the phrase “time is a flat circle”?

TT08. Near duplicates

Permutation π **Rows=Shingles, Columns=Documents**

	D1	D2	D3	D4
5	0	0	0	1
3	1	0	1	1
6	1	1	0	1
1	0	1	0	1
4	0	0	0	1
2	1	0	0	0

Compute the signature vector under π

D1	D2	D3	D4

TT08. Near duplicates

What is the similarity between each pair of documents, in this signature matrix?

	D1	D2	D3	D4
π_1	1	1	2	2
π_2	2	3	3	2
π_3	5	5	4	5

TT09. LSH

Suppose you use a hash function to do LSH with this signature matrix. For LSH to be useful, what is the maximum size of the generated hash table?

	D1	D2	D3	D4
π_1	1	1	2	2
π_2	2	3	3	2
π_3	5	5	4	5

TT11. Itemsets

Give 3 examples of transactions
from 3 different application domains

TT11. Itemsets

What is the minimum possible support of an itemset that exists in a database?

TT11. Itemsets

Indicate the support of the itemset
“Paper, Scissors”

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT11. Itemsets

Explain the support monotonicity property using as example the itemset {“Pencil”}

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT11. Itemsets

What is a closed itemset?

TT11. Itemsets

What is a maximal itemset?

TT11. Itemsets

What is a closed itemset in this database?

What is a non closed itemset in this database?

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT11. Itemsets

Draw an itemset lattice for the following transactions:

TID	Items
1	Tomato, Pear
2	Strawberry, Pear, Apple
3	Apple, Pear
4	Apple, Strawberry, Tomato, Pear

TT12. Association rules

Explain the formula of the confidence of a rule

Is it a problem that the denominator could be in theory zero?

TT12. Association rules

Indicate the confidence of the rule
{Eraser} => {Pencil}

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT12. Association rules

What does it mean if the **lift** of a rule is strictly larger than 1.0?

TT13. Association rule mining

Explain the apriori algorithm on this dataset, with minsup=0.5
Tip: first write a table with itemsets of size 1 (itemset, support)

tid	Set of items
1	x1 x2 x3
2	x2 x3 x4
3	x4 x5
4	x1 x2 x4
5	x1 x2 x3 x5
6	x1 x2 x3 x4

TT13. Association rule mining

Explain the confidence monotonicity property

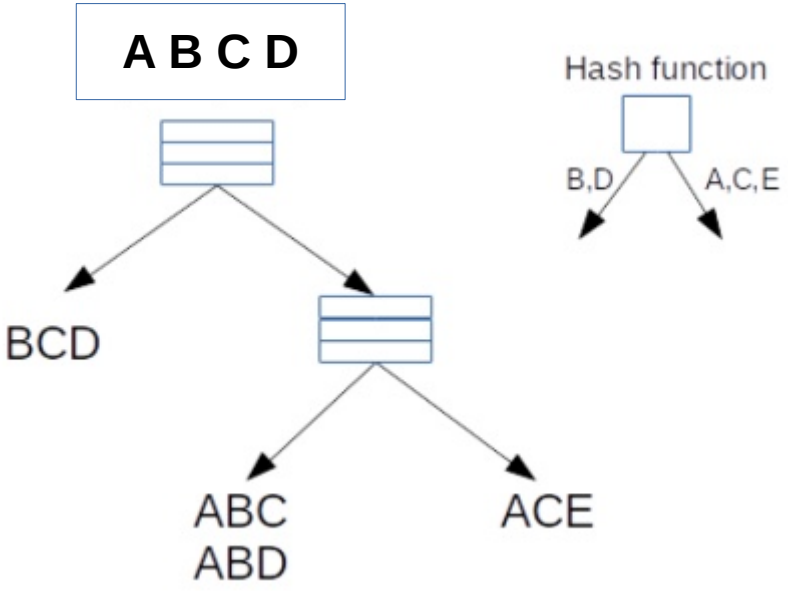
Let X_S, X_L, I be itemsets; assume $X_S \subset X_L \subset I$

Then:

$$\mathrm{conf}(X_L \Rightarrow I - X_L) \geq \mathrm{conf}(X_S \Rightarrow I - X_S)$$

TT14. Improved assoc. rule mining

Indicate in this hash tree which candidates are visited if we are looking for itemsets contained in { B, C, D}



TT16. Recommender systems

What is a utility matrix?

What kind of values we can found in its cells?

TT16. Recommender systems

What is a content-based recommender system?

TT16. Recommender systems

Indicate what is the concept in “?” and explain the in-class exercise we did with the electric scooters

- Database of ~100 electric scooters, of which 12 have been rated on a scale 1-5
- We have done [redacted] on:

  price [$], battery capacity [Wh], range [km]
- Which would be your top-3 recommended scooter among the remaining ones?

TT17. Recommender systems

What is a neighborhood-based method for recommender systems?

Can you explain it in terms of an item-based recommender?

TT17. Recommender systems

Compute the similarity between users u and v in this dataset

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_{u,v}} (u_i - \hat{u}) \cdot (v_i - \hat{v})}{\sqrt{\sum_{i \in I_{u,v}} (u_i - \hat{u})^2 \cdot \sum_{i \in I_{u,v}} (v_i - \hat{v})^2}}$$

	2			4	5	
	5		4			1
			5		2	
u		1		5		4
			4			2
v	4	5		1		

TT17. Recommender systems

Suppose you have computed all similarities of users to u.

Explain how do you recommend movies to user u using the formula below

	2			4	5	
	5		4			1
			5		2	
u		1		5		4
			4			2
	4	5		1		

$$\text{score}(u,i) = \hat{u} + \frac{\sum_{v:v_i \neq \text{NULL}} \text{sim}(v,u) \cdot (v_i - \hat{v})}{\sum_{v:I_{u,v} \neq \emptyset} |\text{sim}(v,u)|}$$

TT18. Factorization-based recsys

How does a factorization model for recommender systems works?

TT18. Factorization-based recsys

Why is it advantageous to use **non-negative factorization** in recommender systems?

TT19. Outlier detection

Name 3 reasons why a dataset may have outliers

TT19. Outlier detection

What is the difference between an internal (unsupervised) and external (supervised) criteria for outlier detection

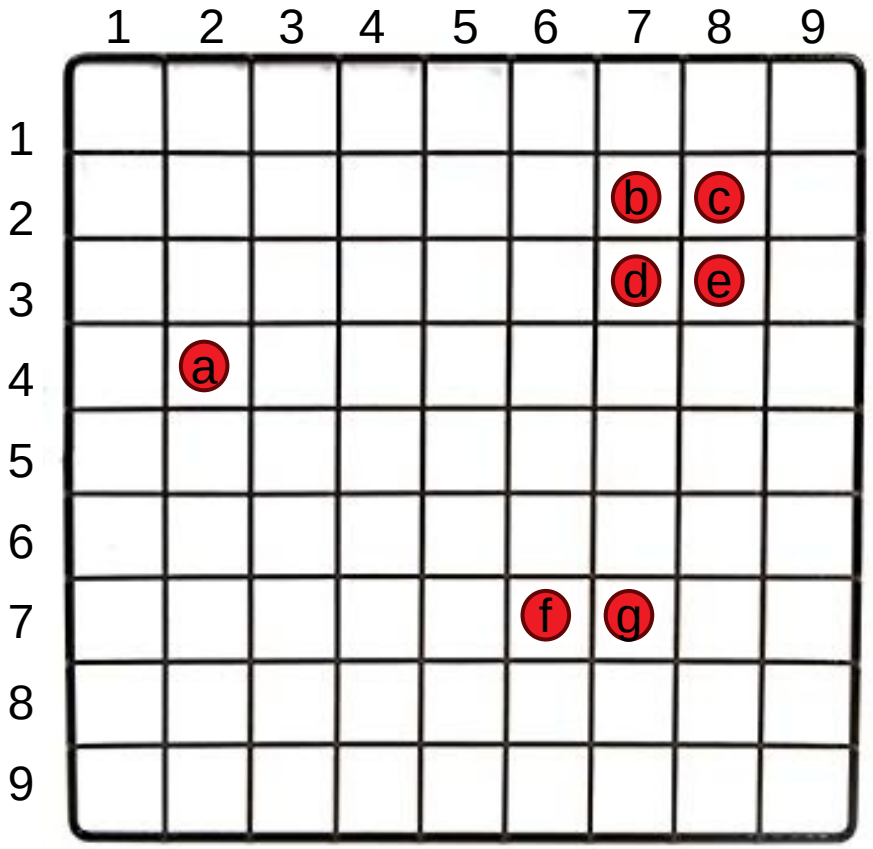
TT19. Outlier detection

Describe one situation in which extreme value analysis is appropriate for finding outliers

TT21. Outlier detection

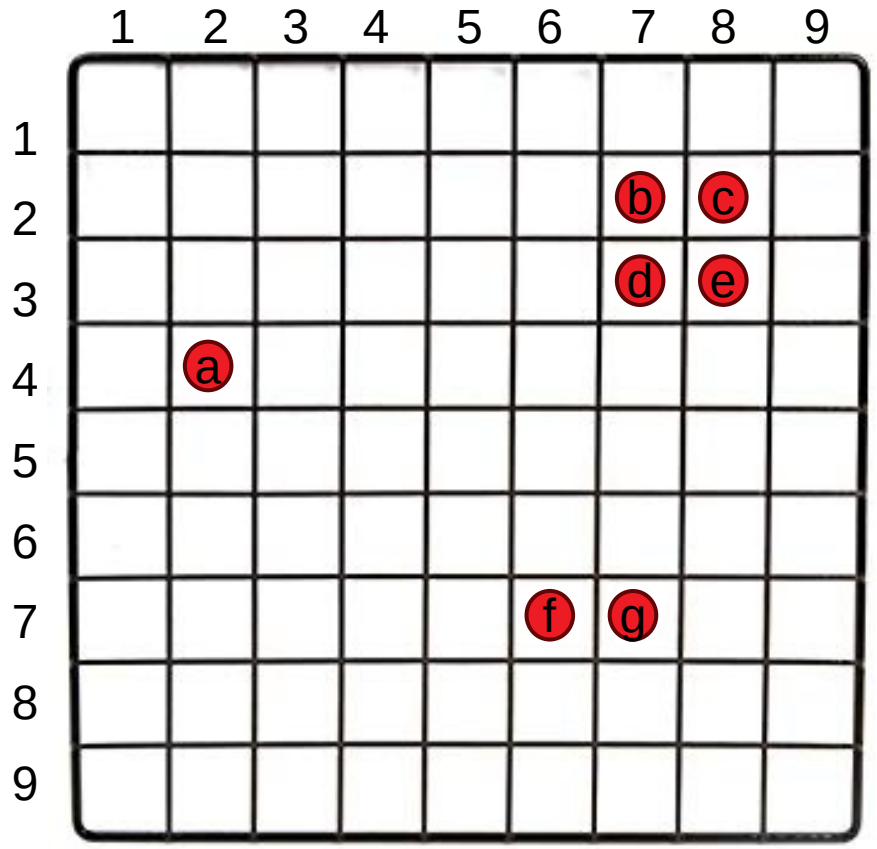
Indicate how do you create an isolation forest over the graph on the right

Explain what the outlier score for a point depends on (no need to give a formula)



TT21. Outlier detection

Indicate how a grid-based method would work to find outliers in this dataset



TT22. Streams

What does it mean that a data stream is transient?

TT22. Streams

Suppose we have a stream of the type *(a, b)* indicating that user *a* purchased book *b*

Indicate how to sample 1% of the users and the books they have bought from this stream

TT22. Streams

Suppose we have a stream of photos from a photo sharing site

Indicate how to sample 100 photos from this stream **uniformly at random**

TT22. Streams

Explain what is a standing query in a stream-processing architecture

TT22. Streams

What is load shedding?

TT23. Streams

Explain how reservoir sampling works

TT24. Bloom filters

In a bloom filter, why would we sometimes want to increase the number of bits of the filter?

TT24. Bloom filters

When a bloom filter says an object
is <u>not</u> a member of a set, it is:
(a) always right (b) sometimes right

When a bloom filter says an object
is a member of a set, it is:
(a) always right (b) sometimes right

TT27. Time series

Interpolate the following time series using **linear interpolation** to obtain the values on Monday at midnight and Tuesday at noon

Monday 12:00 – 36ºC

Monday 23:59 – ???

Tuesday 06:00 – 30ºC

Tuesday 12:00 – ???

Tuesday 18:00 – 35ºC

TT27. Time series

Compute a moving average with k=2 in the following series:

t	1	2	3	4	5	6	7	8	9	10
y_t	10	12	16	20	30	100	500	1000	1050	1070
y_t^{MA2}										

TT27. Time series

Explain how dynamic time warping works and indicate what is it can be used for

TT29. Time series forecasting

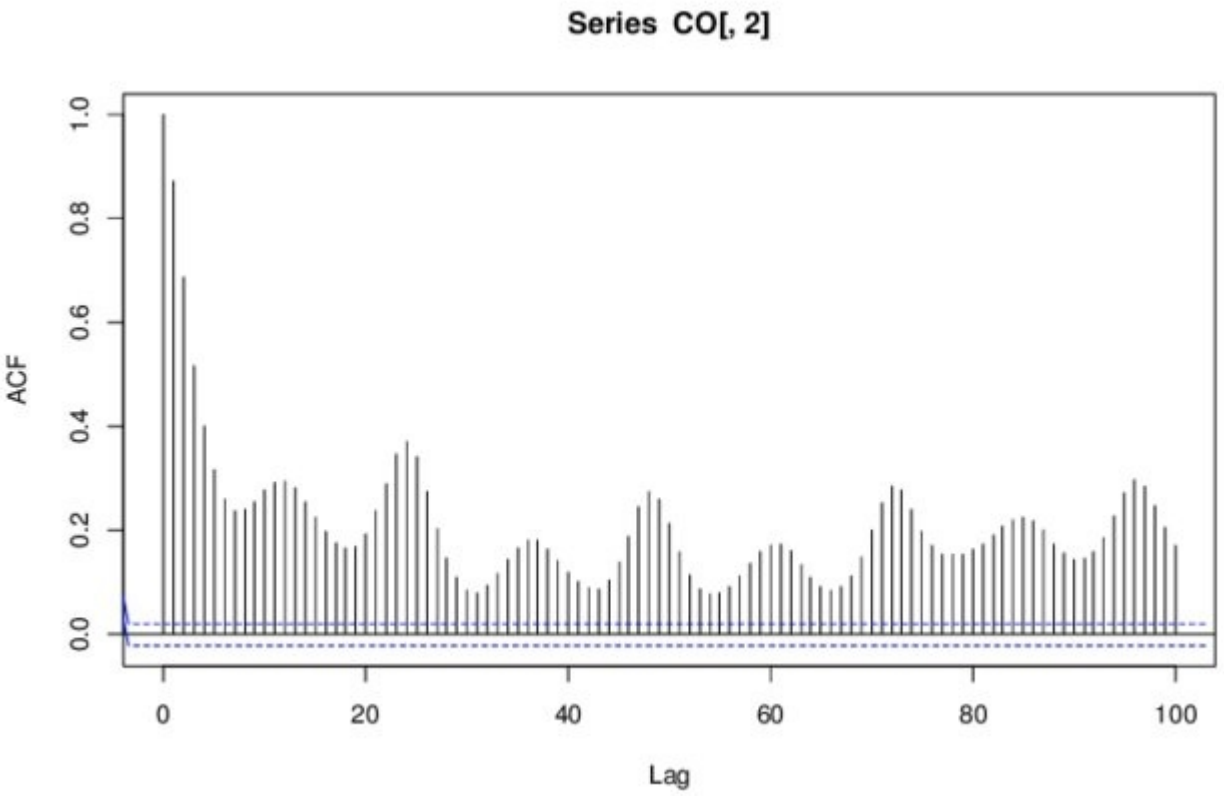
What is the difference between a stationary and a non-stationary process?

TT29. Time series forecasting

What is first-order differencing in a time series?

TT29. Time series forecasting

Explain this auto-correlation plot of carbon monoxide (CO) concentration in an urban monitoring station; lags are in hours



TT29. Time series forecasting

What is the difference between an autoregressive moving average (ARMA) model and an autoregressive (AR) model