



TIPOLOGIA I CICLE DE VIDA DE LES DADES

PAC1: Què son les dades i quin el seu cicle de vida?

Marc Cervera Rosell

Semestre: febrer 2025 - juny 2025

Màster en ciència de dades

Índex

Exercici 1	1
Pregunta 1	1
Pregunta 2	1
Pregunta 3	1
Pregunta 4	4
Pregunta 5	4
Pregunta 6	4
Bibliografia	5

Índex de figures

1	Exemple de la tasca "Panorama general"	2
2	Exemple de la tasca "Zoom"	2
3	Exemple de la tasca "Filtratge"	3
4	Exemple de la tasca "Relcions"	3

Exercici 1

Pregunta 1

Per aplicar la tècnica de normalització caldria, en primer lloc, seleccionar aquelles variables numèriques a les quals volem aplicar la tècnica i després ajustar-les a una escala més reduïda, per tal de, fer-les més comprensibles. Típicament, les escales són entre -1.0 i 1.0 o entre 0.0 i 1.0.

Per aplicar la tècnica de discretització cal, com en el cas anterior, seleccionar aquelles variables numèriques a les quals volem aplicar la tècnica. Seguidament, cal substituir els valors numèrics per una sèrie d'etiquetes les quals poden ser conceptuals o intervals. Aquesta tècnica permet consolidar un possible criteri d'agrupació de dades. Per exemple, un *dataset* amb els resultats dels balanços d'empreses. Es podria substituir els resultats dels balanços per 'positiu', 'negatiu' i 'neutre' i, posteriorment, agrupar pel resultat del balanç.

Aquestes tècniques són molt importants per als models de predicció, ja que permeten que les dades siguin més senzilles d'interpretar i això comporta que el rendiment del model sigui millor i que aquest tingui la capacitat de trobar patrons més fàcilment, és a dir, el model es fa més senzill.

Pregunta 2

Mentre un enginyer de dades treballa en el desenvolupament, construcció, prova i manteniment d'arquitectures, un científic de dades és una persona que es dedica a obtenir informació rellevant sobre les dades, a partir de preguntes que s'ha plantejat, i a transmetre aquesta informació de manera senzilla. A més, tot i que ambdós perfils tenen coneixements de programació, l'enginyer de dades no aprofundeix tant en coneixements d'estadística i matemàtiques com el científic de dades.

[Oferta de *data scientist*](#)

[Oferta de *data engineer*](#)

Pregunta 3

Les set tasques que permeten un major nivell d'abstracció en la visualització de dades són:

- Panorama general → Permet una visió general de les dades.
- Acostament → Permet l'acostament a un punt concret.
- Filtratge → Permet filtrar els elements que no són d'interès.
- Detalls a petició → Permet obtenir detalls sobre un element o grup d'elements.

- Relaciones → Permet la visualització de tendències entre diferents elements.
- Historial → Permet revisar/repetir passos anteriors.
- Extracció → Permet l'extracció de subconjunts i els seus detalls.

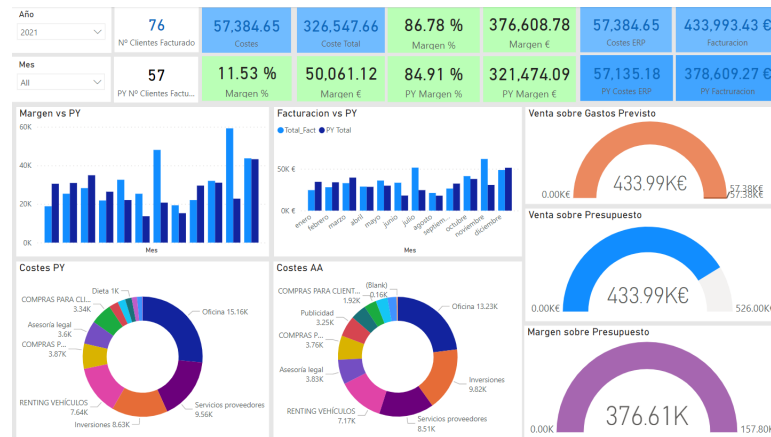


Figura 1: Exemple de la tasca "Panorama general"

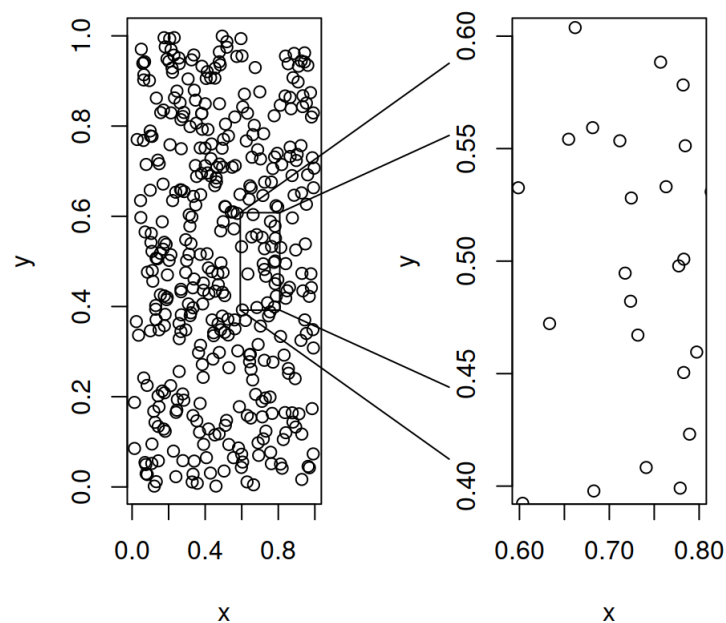


Figura 2: Exemple de la tasca "Zoom"

	A	B	C	D	E	F
1	Región	Vendedor	Orden	Fecha	Total	
A Z	Ordenar de A a Z		00	01/01/2014	\$404	
Z A	Ordenar de Z a A		01	02/01/2014	\$789	
	Ordenar por color		02	03/01/2014	\$955	
	Borrar filtro de "Vendedor"		03	04/01/2014	\$556	
	Filtrar por color		04	05/01/2014	\$806	
	Filtros de texto		05	06/01/2014	\$174	
	Buscar		06	07/01/2014	\$149	
	<input checked="" type="checkbox"/> (Seleccionar todo)		07	08/01/2014	\$639	
	<input type="checkbox"/> Alejandra		08	09/01/2014	\$218	
	<input type="checkbox"/> Brenda		09	10/01/2014	\$134	
	<input type="checkbox"/> Carolina		10	11/01/2014	\$899	
	<input type="checkbox"/> Diana		11	12/01/2014	\$924	
	<input checked="" type="checkbox"/> Hugo		12	13/01/2014	\$436	
	<input type="checkbox"/> Juan		13	14/01/2014	\$844	
	<input type="checkbox"/> Luis		14	15/01/2014	\$511	
	<input type="checkbox"/> Paco		15	16/01/2014	\$142	
	Aceptar	Cancelar	16	17/01/2014	\$898	
			17	18/01/2014	\$939	
			18	19/01/2014	\$663	
21	Sur	Luis	119	20/01/2014	\$598	
22	Este	Luis	120	21/01/2014	\$686	

Figura 3: Exemple de la tasca "Filtratge"

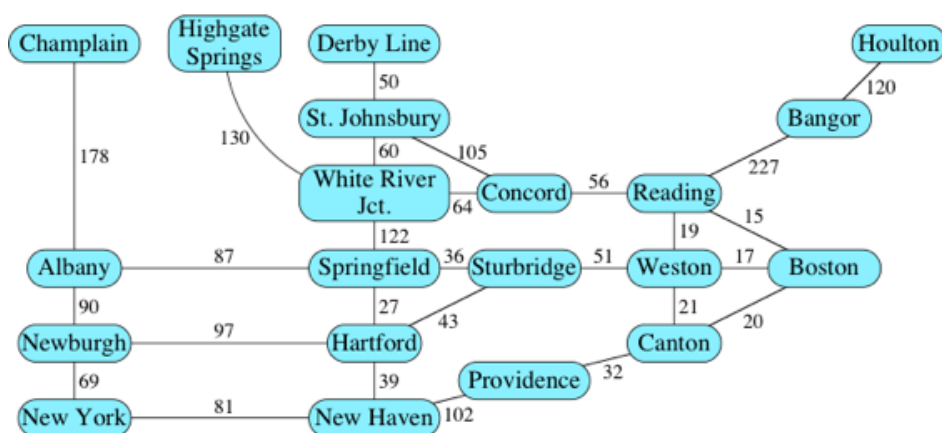


Figura 4: Exemple de la tasca "Relcions"

Pregunta 4

La primera etapa és la captura, que té com a objectiu la recollida de les dades que es generen en un procés. Un exemple d'aquesta etapa podria ser, les dades d'un sensor IoT que recollisca un broker de Kafka.

La segona etapa és l'emmagatzematge, que té com a objectiu guardar les dades capturades en un format tal que sigui adequat per a la posterior explotació. L'exemple en aquest cas, seria l'emmagatzematge de les lectures del sensor IoT en una base de dades.

La tercera etapa és el preprocessament, que té com a objectiu l'aplicació de diferents tècniques per a preparar les dades per la seva posterior anàlisi. L'exemple podria ser la neteja de les dades.

La quarta etapa és l'anàlisi, que té com a objectiu crear models per explicar les dades. L'exemple seria construir un model de predicció de falles del sensor IoT.

La penúltima etapa és la visualització que té com a objectiu presentar les dades. L'exemple seria un panell de PowerBI.

Finalment, la publicació que documenta els resultats. L'exemple seria la redacció d'un *report* amb els resultats obtinguts.

Pregunta 5

La millor opció per a un sistema de monitoratge ambiental basat en sensors seria una base de dades no relacional a causa del fet que els sensors poden generar dades molt diverses. És a dir els sensors poden generar dades de tipus text, numèriques, imatges, so, etc. Per tant, considerant aquesta gran varietat que deriva en diferents esquemes per cada sensor es considera una millor opció una base de dades no relacional gràcies a la seva flexibilitat amb els esquemes de les dades i els seus tipus. A més, cal destacar que aquest tipus de bases de dades tenen un elevat rendiment de lectura i escriptura. Per tant, en un sistema com aquest, que captura volums de dades molt grans, és important poder tant consultar com inserir noves dades de la manera més ràpida possible. Finalment, un altre motiu pel qual una base de dades NoSQL és una millor opció és la capacitat d'escalar horitzontalment. És a dir, la capacitat d'afegir més recursos per a poder treballar amb volums de dades majors que en les bases de dades relacionals.

Pregunta 6

La primera llibreria és [Pandas](#). Les seves funcionalitats clau són la lectura i escriptura de dades, la manipulació de dades, el tractament de *nulls*, l'agrupació, fusió i combinació i el tractament i anàlisi de *time series*. Aquesta llibreria conté una funció per eliminar duplicats ([drop_duplicates\(\)](#)), una funció que es pot usar per a corregir errates tipogràfiques, tot i que no està dissenyada específicament per això ([replace\(\)](#)) i una funció per a la conversió

de formats ([*astype\(\)*](#)). La segona eina és [PyJanitor](#). És una llibreria basada en Pandas, per tant, les seves funcionalitats se centren simplificar algunes operacions de Pandas. Per eliminar duplicats es pot utilitzar la funció *remove_na()*. Per corregir errors tipogràfics i transformar formats, aquesta llibreria, no implementa cap funcionalitat específica. En conseqüència, aquestes dues operacions s'haurien de realitzar amb Pandas o amb tercers llibreries.

Bibliografia

- [Imatge de l'exemple de "Panorama general"](#)
- [Imatge de l'exemple de "Zoom"](#)
- [Imatge de l'exemple de "Filtratge"](#)
- [Imatge de l'exemple de "Relacions"](#)
- [Informació Bases de dades NoSQL](#)
- [Pandas: La Herramienta Esencial para Data Science en Python](#)
- [Pandas: Cambiar los tipos de datos en los DataFrames](#)
- [*Most Helpful Data Cleaning Tools in Python for 2025*](#)