Part I: Selecció del conjunt de dades

1. Justificació de la selecció

https://www.kaggle.com/datasets/yogape/logistics-operations-database

He seleccionat aquest conjunt de dades relacionat amb operacions logístiques d'una companyia de transport de mercaderies. Tot i que són simulades, han estat generades per una persona amb experiència real en la indústria logística, de manera que reflecteixen patrons i condicions reals de l'operació. Treballar amb aquest dataset ofereix l'oportunitat de treballar amb dades organitzades en diverses taules relacionades, com en el món real (ex. bases de dades relacionals), i permet posar-te en el lloc d'un científic de dades, explorar les variables disponibles i escollir les visualitzacions més adients per transmetre de manera clara les idees i la informació més rellevant.

Així, la meva finalitat és crear un dashboard empresarial que permeti analitzar l'eficiència de la flota, els fluxos de rutes o els patrons operatius, entre d'altres, incloent en la mesura del possible visualitzacions més avançades com *ribbon plots, bubble maps, heatmaps, sankey charts* o altres elements visuals que permetin resumir grans volums de dades de forma intuïtiva i impactant.

En definitiva, treballar amb aquest dataset permet combinar anàlisi exploratori, visualització de dades i generació de coneixement, replicant processos reals en la indústria logística i el treball de científics de dades professionals.

2. Rellevància del conjunt de dades

- **Són dades actuals?**

Les dades són actuals en termes d'estructura i valors típics de la indústria, i cobreixen un període de 3 anys amb informació detallada de més de 85.000 viatges, 120 camions, 150 conductors o 50 magatzems de distribució. Tot i que són dades simulades, la simulació basada en l'experiència professional garanteix que els patrons observats siguin coherents i aplicables a situacions reals, tot i no representar dades confidencials de cap empresa concreta.

- **Tracten un tema important per algun col·lectiu concret?**

L'estudi d'aquestes dades són rellevants en la indústria logística, ja que permeten identificar rutes eficients, planificar manteniments preventius, optimitzar el consum de combustible, veure el rendiment dels conductors, entre d'altres. Així, són útils per a gestors de flota, analistes logístics i responsables de planificació de rutes, ja que poden obtenir coneixement rellevant a partir de les visualitzacions creades.

- **S'ha tingut en compte la perspectiva de gènere?**

Tot i que les dades originals no incloïen informació de gènere, s'ha creat manualment una columna addicional a la taula de conductors, on s'ha inferit el gènere de la persona basant-

se en el seu nom. Aquesta acció permet incorporar la perspectiva de gènere en algunes anàlisis i visualitzacions, com comparacions de rendiment o incidències.

3. Complexitat

El conjunt de dades està format per 13 taules enllaçades (1 a N, N a 1...), el que permet un volum suficient per anàlisis realistes i complexes. Les taules i nombre de registres de cada una són:

1. trips (85410 files, 12 variables)
2. delivery_events (170820 files, 9 variables)
3. drivers (150 files, 13 variables)
4. driver_monthly_metrics (4464 files, 9 variables)
5. trucks (120 files, 11 variables)
6. truck_utilization_metrics (3312 files, 11 variables)
7. loads (85410 files, 12 variables)
8. maintenance_records (2920 files, 12 variables)
9. fuel_purchases (196442 files, 11 variables)
10. routes (58 files, 9 variables)
11. facilities (50 files, 11 variables)
12. customers (200 files, 8 variables)
13. safety_incidents (170 files, 15 variables)

Així, comptem amb dades tant quantitatives (ex. actual_distance_miles, actual_duration_hours, fuel_gallons_used, average_mpg, downtime_hours, total_cost, labor_hours, parts_cost, total_revenue, trips_completed, weight_lbs, gallons, entre d'altres), categòriques (ex. trip_status, truck_model, fuel_type, event_type, facility_type, load_status, load_type, customer_type, route_id, incident_type, employment_status, cdl_class, entre d'altres), geogràfiques (ex. latitude, longitude) o temporals (ex. dispatch_date, maintenance_date, purchase_date, load_date, incident_date, month). A més a més, també es poden derivar altres mètriques com "cost per milla" (total_cost / actual_distance_miles), segons la necessitat i intenció final de la visualització.

4. Originalitat

Aquest dataset permet combinar i enriquir les dades de diverses maneres per aportar un enfocament innovador. A més a més, les taules relacionades formen un dataset que reflecteix la complexitat del món real, similar a bases de dades relacionals professionals. El dataset és recent de Kaggle i no s'han trobat visualitzacions prèvies del conjunt de dades. El conjunt de les 13 taules no necessita ser enriquit amb dades externes a aquest ja que és

molt complet i permet molts enfocs diferents segons el que es vulgui estudiar (eficiència dels conductors, antiguitat dels camions...).

Les noves mètriques que es puguin necessitan es crearan a posteriori, durant la creació del dashboard final.

5. Qüestions i diccionari de variables

A partir del conjunt de dades seleccionat, el projecte busca donar resposta a un conjunt de preguntes rellevants per a la gestió operativa d'una empresa de transport. Aquestes qüestions s'han definit tenint en compte:

- la naturalesa relacional i complexa del dataset
- la combinació de dades operacionals, geogràfiques i de rendiment
- i l'objectiu de construir un dashboard empresarial útil i accionable (a priori pensat per fer en PowerBI)

Tot i que el conjunt de dades és molt ampli i abasta múltiples àrees de l'operativa d'una empresa de transport (entregues, conductors, camions, rutes, costos, incidències, carburant, etc.), el projecte se centrarà en un subconjunt ben definit de preguntes relacionades amb l'eficiència operativa i logística.

Aquesta acotació és necessària per garantir una anàlisi clara i unes visualitzacions útils, però es mantindrà una visió oberta durant l'exploració de les dades, de manera que, si s'identifiquen relacions o patrons inesperats, el dashboard es podrà adaptar per incorporar visualitzacions addicionals que aportin valor.

Així, algunes de les moltes preguntes que potencialment es podrien respondre amb aquest dataset són:

- Com varien les entregues puntuals (on-time deliveries) segons la ubicació geogràfica?

  Aquesta pregunta permet detectar ciutats o regions amb problemes operatius, com retards freqüents o colls d'ampolla logístics. Combina dades de delivery_events (hora programada vs real), facilities (ubicació de les instal·lacions) i drivers (per identificar responsabilitats). La visualització ideal seria un bubble map o un heatmap geogràfic, on el color indiqui el percentatge de puntualitat i la mida representi el nombre de lliuraments.

- Quin és el rendiment dels conductors en termes d'eficiència (MPG, temps en ralentí, puntualitat)?

  Aquesta pregunta avalua els patrons de rendiment dels conductors, permetent identificar aquells més eficients i els que necessiten millora. Utilitza les variables de driver_monthly_metrics i drivers, incloent característiques com gènere o antiguitat. Es pot visualitzar amb histogrames (distribució d'eficiència o temps en ralentí), scatter plots (comparant puntualitat i MPG) o radar charts centrats en els conductors amb menor rendiment.

- Quines rutes o loads generen més costos de combustible?

Aquesta pregunta ajuda a identificar rutes menys eficients energèticament i a planificar millors assignacions de vehicles i rutes. Integra fuel_purchases, routes i trips per calcular el cost de combustible per ruta o per càrrega. La visualització podria ser un scatter plot (cost vs distància o volum) o un treemap per veure quines rutes representen la major despesa energètica.

- Quines rutes són més profitables?

Permet identificar quines rutes generen més marge i quines poden ser candidates per expandir o eliminar. Això ajuda a prendre decisions estratègiques sobre on assignar més capacitat, negociar tarifes i maximitzar beneficis. Es poden combinar dades de loads, trips i routes per calcular marge total, marge per unitat i volum transportat. Una bubble chart seria ideal, on la mida representi el volum i el color indiqui el marge.

- Quines són les principals causes d'incidents de seguretat i on es concentren geogràficament?

Explora safety_incidents per identificar els tipus d'incidents més comuns i les ubicacions on es produeixen amb més freqüència. Això permet prioritzar mesures de seguretat. Es pot visualitzar amb heatmaps geogràfics, bar charts per tipus d'incident o stacked charts per veure combinacions de tipus i ubicació.

- Hi ha relació entre l'antiguitat dels camions i els seus costos de manteniment?

Aquesta anàlisi integra trucks i maintenance_records per comprovar si els vehicles més antics tenen costos més alts o més temps de fora de servei. La visualització pot ser un scatter plot (anys vs cost), un heatmap per categories de camions o un violin/boxplot per mostrar distribució de costos segons antiguitat.

- Quines ciutats generen més trajectes buits?

Permet identificar oportunitats per reduir empty miles i costos de reposicionament, millorant l'eficiència de la xarxa de rutes. Es poden combinar trips i routes per calcular trajectes amb càrrega nul·la. La visualització podria ser un mapa geogràfic o un bar chart per ciutat.

- Com varien els carregues al llarg de l'any?

Aquesta pregunta permet detectar pics estacionals i planificar la capacitat amb antelació. Utilitza dades de loads per agrupar-les per mes o trimestre i calcular volums i ingressos. Es podria visualitzar amb un line chart de tendència mensual o un variance bar chart que mostri la diferència respecte a la mitjana mensual.

A continuació es mostra un diccionari amb les principals variables que utilitzarien les visualitzacions que respondrien les preguntes anteriors:

Variable	**Taula / Font**	**Tipus de dada**	**Fet / Dimensió**	**Descripció**
load_id	loads	Numèric	Dimensió	Identificador únic de cada càrrega
load_date	loads	Data	Dimensió	Data de la càrrega / dispatch, permet analitzar temporalment
route_id	routes / loads	Numèric	Dimensió	Identificador de la ruta associada
origin_city	routes	Text	Dimensió	Ciutat d'origen de la ruta
destination_city	routes	Text	Dimensió	Ciutat de destinació de la ruta
actual_distance_miles	trips	Numèric	Fet	Distància de la ruta en milles
revenue	loads	Numèric	Fet	Ingressos generats per la càrrega
fuel_gallons_used	trips	Numèric	Fet	Gallons de combustible consumits
fuel_cost	trips (s'ha de calcular)	Numèric	Fet	Cost del combustible (fuel_gallons_used × preu unitari)
gross_margin	trips (s'ha de calcular)	Numèric	Fet	Ingressos menys costos de combustible
margin_pct	trips (s'ha de calcular)	Percentatge	Fet	Percentatge de marge (gross_margin ÷ revenue × 100)
event_id	delivery_events	Numèric	Dimensió	Identificador de l'esdeveniment de lliurament
event_type	delivery_events	Numèric	Dimensió	Tipus d'esdeveniment: Pickup / Delivery

scheduled_datetime	delivery_events	Data/Hora	Dimensió	Hora programada de l'esdeveniment
actual_datetime	delivery_events	Data/Hora	Dimensió	Hora real de l'esdeveniment
on_time_flag	delivery_events	Boolean	Fet	True si l'entrega és puntual
facility_id	facilities	Numèric	Dimensió	Identificador de la instal·lació
city	facilities	Text	Dimensió	Ciutat de la instal·lació
driver_id	drivers / driver_monthly_metrics	Numèric	Dimensió	Identificador del conductor
first_name	drivers	Text	Dimensió	Nom del conductor
gender	drivers (afegida manualment)	Text	Dimensió	Sexe del conductor
average_mpg	driver_monthly_metrics	Numèric	Fet	Eficiència de combustible per conductor
average_idle_hours	driver_monthly_metrics	Numèric	Fet	Temps en ralentí del vehicle
on_time_pct	driver_monthly_metrics (s'ha de calcular)	Percentatge	Fet	Percentatge d'entregues puntuals per conductor (a partir de on_time_delivery_rate)
truck_id	trucks / maintenance_records	Numèric	Dimensió	Identificador del camió
truck_age	Trucks (s'ha de calcular)	Numèric	Dimensió	Antiguitat del camió en anys, per segmentar manteniment (calcular a partir de acquisition_date i acquisition_mileage)
total_cost	maintenance_records	Numèric	Fet	Cost de manteniment del camió
downtime_hours	maintenance_records	Numèric	Fet	Temps fora de servei del camió per manteniment

fuel_purchase_id	fuel_purchases	Numèric	Dimensió	Identificador de compra de combustible
total_cost	fuel_purchases	Numèric	Fet	Cost de la compra de combustible
incident_id	safety_incidents	Numèric	Dimensió	Identificador de l'incident de seguretat
incident_type	safety_incidents	Text	Dimensió	Tipus d'incident o avaria
location_city	safety_incidents	Text	Dimensió	Ciutat on es va produir l'incident
empty_miles	trips (s'ha de calcular)	Numèric	Fet	Milles recorregudes sense càrrega
month	Loads (s'ha de calcular)	Període	Dimensió	Mes de la càrrega (a partir de load_date)