UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Pallares Genaro, Judit. *Disseny i implementació d'un Data Warehouse per a l'anàlisi de vendes i rendibilitat empresarial*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337316

under the terms of the CC BY-SA license.

Disseny i implementació d'un Data Warehouse per a l'anàlisi de vendes i rendibilitat empresarial

Judit Pallarés Genaro

Resum– En moltes organitzacions, les dades de vendes es generen en sistemes transaccionals dissenyats principalment per registrar operacions, però no per facilitar l'anàlisi estratègica. Això dificulta respondre preguntes clau sobre el rendiment del negoci, com per exemple quins productes generen més benefici, com evoluciona la demanda amb el temps o quins segments de clients aporten més valor. Aquest treball presenta el disseny i la implementació d'un Data Warehouse orientat a l'anàlisi de vendes i de la rendibilitat empresarial, amb l'objectiu de facilitar la presa de decisions basada en dades. Es parteix d'un conjunt de dades públic de vendes del sector retail o e-commerce, que s'han transformat mitjançant un procés ETL (Extract, Transform, Load) per a construir una base de dades analítica basada en un model dimensional estrella. Aquest model organitza les dades en una taula de fets central que emmagatzema les mètriques de negoci (vendes, costos, marges), envoltada de taules de dimensions que aporten el context descriptiu (producte, client, temps, geografia), permetent analitzar les vendes des de múltiples perspectives de manera eficient. El sistema s'ha complementat amb dades de costos simulats i justificats, cosa que ha permès calcular indicadors de rendibilitat i analitzar marges de benefici. El Data Warehouse ha estat carregat a PostgreSQL i els resultats s'han explorat mitjançant nou indicadors clau de negoci (KPIs) i s'han materialitzat en quadres de comandament interactius desenvolupats amb Python (Streamlit) i Power BI.

Paraules clau– Data Warehouse, Business Intelligence, Model dimensional, ETL, Anàlisi de vendes, KPIs, PostgreSQL, Streamlit.

Abstract– In many organisations, sales data is generated by transactional systems designed primarily to record operations, not to support strategic analysis. This makes it difficult to answer key questions about business performance, such as which products generate the most profit, how demand evolves over time, or which customer segments contribute the most value. This work presents the design and implementation of a Data Warehouse oriented towards the analysis of sales and business profitability, with the aim of supporting data-driven decision making. It starts from a public sales dataset from the retail or e-commerce sector, which has been transformed through an ETL (Extract, Transform, Load) process to build an analytical database based on a dimensional star schema. This model organises the data into a central fact table that stores the business metrics (sales, costs, margins), surrounded by dimension tables that provide the descriptive context (product, customer, time, geography), allowing sales to be analysed from multiple perspectives efficiently. The system has been complemented with simulated and justified cost data, which has made it possible to compute profitability indicators and analyse profit margins. The Data Warehouse has been loaded into PostgreSQL and the results have been explored through nine key business indicators (KPIs) and materialised in interactive dashboards developed with Python (Streamlit) and Power BI.

Keywords– Data Warehouse, Business Intelligence, Dimensional model, ETL, Sales analysis, KPIs, PostgreSQL, Streamlit.

1 Introducció i Motivació

Actualment ens trobem immersos en l'anomenada era de la informació, en la qual la disponibilitat i l'ús adequat de les dades s'han convertit en un factor clau per a la competitivitat de les empreses. En pràcticament tots els sectors econòmics, la informació representa un recurs estratègic que permet comprendre millor el funcionament del negoci i prendre decisions informades.

En el context empresarial, la informació prové principalment dels sistemes operacionals de l'empresa (com els sistemes de vendes, producció o gestió de clients), que generen una gran quantitat de dades sobre les activitats diàries. Aquestes dades constitueixen una font de gran valor, ja que permeten analitzar el comportament del negoci i detectar patrons o tendències rellevants. A més d'aquestes fonts in-

• E-mail de contacte: Judit.PallaresG@autonoma.cat

• Treball tutoritzat per: Enric Martí Gòdia (Departament: Àrea de Ciències de la Computació)

• Curs 2025/26

ternes, moltes empreses també incorporen dades externes relacionades amb el mercat, la competència o el comportament dels consumidors.

Tanmateix, l'augment constant del volum de dades genera diversos problemes. Entre els més rellevants hi ha la sobreabundància d'informació, la dificultat per accedir-hi de manera eficient i la manca de selecció o estructuració adequada. En molts casos, els responsables de prendre decisions dediquen una part significativa del seu temps a buscar informació, preparar-la o transformar-la en un format útil per a l'anàlisi. Aquesta situació posa de manifest la necessitat de sistemes que permetin organitzar, integrar i explotar les dades de manera eficient.

En aquest context, els sistemes de Business Intelligence i, en particular, els Data Warehouse, permeten integrar dades provinents de diferents fonts i estructurar-les de manera que siguin fàcilment accessibles per a l'anàlisi. Aquests sistemes faciliten l'explotació de dades històriques i permeten analitzar el negoci des de diferents perspectives, com ara productes, clients o períodes temporals.

A més, en mercats cada vegada més competitius, les empreses necessiten analitzar constantment el seu rendiment i comparar-lo amb l'entorn en què operen. Disposar d'informació fiable, integrada i accessible permet anticipar tendències, identificar oportunitats i millorar la presa de decisions estratègiques.

Per aquest motiu, el desenvolupament de sistemes analítics basats en Data Warehouse s'ha convertit en una pràctica habitual en moltes organitzacions. Aquest treball s'inscriu en aquest context i aplica els principis de l'enginyeria de dades per construir una plataforma analítica que permeti analitzar les vendes i la rendibilitat empresarial a partir de dades històriques.

L'objectiu principal és, doncs, el disseny i la implementació d'un Data Warehouse orientat a aquesta anàlisi, que es concreta en les funcionalitats següents:

- **OE1:** Analitzar un dataset públic de vendes i identificar les seves característiques i limitacions, incloent-hi la caracterització del perfil empresarial.
- **OE2:** Dissenyar un model dimensional adequat per a l'anàlisi de dades de vendes mitjançant un esquema estrella.
- **OE3:** Desenvolupar un procés ETL per a l'extracció, neteja, transformació i càrrega de dades en el Data Warehouse.
- **OE4:** Definir i integrar un model de costos simulat que permeti analitzar la rendibilitat.
- **OE5:** Calcular indicadors clau de negoci (KPIs) relacionats amb ingressos, costos i marges mitjançant consultes SQL.
- **OE6:** Desenvolupar quadres de comandament interactius per a la visualització i exploració dels resultats.

El primer d'aquests objectius mereix una atenció especial, ja que l'anàlisi exploratòria ha permès caracteritzar l'empresa objecte d'estudi a partir dels patrons presents al dataset. Aquest perfil, que es resumeix a continuació, proporciona el context empresarial necessari per a la interpretació dels KPIs i dels resultats del dashboard:

- **Tipus i sector:** Majorista/distribuïdor B2B de productes de decoració de la llar, articles de regal i productes estacionals (sector retail/wholesale).
- **Mida:** Empresa mitjana (SME) amb un revenue aproximat de £10M anuals,un catàleg de prop de 5.000 referències de producte i prop de 5.900 clients registrats.
- **Mercat:** Seu al Regne Unit (85,5% del revenue). Exportació minoritària a Europa occidental, principalment Alemanya, França i Irlanda. Presència en 43 països.
- **Model de negoci B2B:** Justificat per la concentració d'activitat en dies laborables i horari d'oficina (pic a les 12h, activitat mínima en caps de setmana), els tickets mitjans elevats (£301 per factura) i la concentració del revenue en un 23% dels clients (principi de Pareto), característics d'una base de clients empresarials fidels i recurrents.
- **Canal de venda:** E-commerce B2B, sense evidència de presència física a les dades.
- **Estacionalitat:** Molt marcada, amb pics de novembre i desembre coherents amb un catàleg orientat a productes de regal i decoració nadalenca.

La resta de l'informe segueix l'ordre de les fases del treball, cadascuna recolzada en les eines pròpies de l'enginyeria de dades. La Secció 2 revisa l'estat de l'art i els fonaments teòrics (Business Intelligence, Data Warehouse, modelització dimensional i ETL), on les referències de Kimball [2] i Inmon [1] fonamenten les decisions de disseny. La Secció 3 descriu la metodologia: l'anàlisi exploratòria i la neteja del dataset, el procés ETL i el disseny del model dimensional, tot implementat en **Python** [9] amb la llibreria **pandas** [10] i documentat en un notebook **Jupyter** [11]; la càrrega del model al Data Warehouse es fa amb el connector **psycopg2** [12] sobre **PostgreSQL 17** [7], sistema sobre el qual s'executen també les consultes SQL. La Secció 4 presenta i discuteix els nou indicadors clau de negoci (KPIs) calculats sobre el Data Warehouse. Finalment, la Secció 5 mostra la visualització dels resultats, desenvolupada primer com a prototip amb **Streamlit** [8] i, després, com a dashboard final amb **Power BI Desktop** [13], connectat directament a PostgreSQL. La Secció 6 en recull les conclusions, les limitacions i les línies de treball futur.

2 ESTAT DE L'ART

El disseny de Data Warehouses es basa en uns fonaments teòrics ben establerts. Les dues referències principals del sector són Inmon [1], que defineix el Data Warehouse corporatiu i l'enfocament *top-down*, i Kimball [2], que proposa la modelització dimensional i l'enfocament *bottom-up*; totes dues continuen sent un punt de referència en projectes de Business Intelligence. Aquesta secció revisa els conceptes en què es fonamenta el treball (BI, Data Warehouse, modelització dimensional i ETL) i justifica les decisions de disseny preses.

2.1 Business Intelligence

La intel·ligència de negoci o, en anglès, Business Intelligence (BI), és el conjunt de processos, tecnologies i eines que transformen les dades brutes d'una organització en informació útil per a la presa de decisions. El concepte va ser popularitzat per Howard Dresner el 1989 [6], i des de llavors s'ha convertit en un àmbit central de l'enginyeria de dades i dels sistemes d'informació empresarial. Oracle el defineix com *"Una combinació d'estratègia i tecnologia*

per recopilar, analitzar i interpretar dades de fonts internes i externes." [3]. En altres paraules, les dades, per si soles, no aporten valor fins que s'organitzen, s'analitzen i es presenten de manera comprensible per als responsables de prendre decisions.

El BI es fonamenta en una jerarquia de tres elements: les **dades** (elements primaris que, per si mateixos, no guien l'acció), la **informació** (dades processades que adquireixen significat i redueixen la incertesa) i el **coneixement** (combinació d'experiència i informació que proporciona un marc per a l'acció). Les dades es transformen en informació, i aquesta, al seu torn, es converteix en coneixement, conformant el nucli del concepte de BI [5].

2.1.1 Cadena de valor del BI

La cadena de valor del BI descriu el recorregut que fan les dades des del seu origen fins a la presa de decisions. Curto [5] la descriu com un flux seqüencial de quatre etapes: les **dades brutes** s'extreuen dels sistemes operacionals i es transformen en un procés ETL; la informació resultant s'emmagatzema en un **repositori analític** (el Data Warehouse) [1]; sobre aquest repositori s'apliquen eines d'anàlisi i visualització per generar **informació estructurada** (KPIs, informes, dashboards) [2] i finalment aquesta informació permet prendre **decisions** basades en dades i no en intuïció [6]. Aquest treball concreta tota la cadena de manera verificable: del dataset Excel original, passant per un procés ETL en Python i un Data Warehouse a PostgreSQL, fins a la interpretació de KPIs i la seva visualització.

2.2 Data Warehouse

Un Data Warehouse (DW) és una base de dades dissenyada específicament per a l'anàlisi i la generació d'informes, no per al registre d'operacions. Bill Inmon [1] el defineix com *"una col·lecció de dades orientada a un tema, integrada, variable en el temps i no volàtil, que dóna suport al procés de presa de decisions"*: les dades s'organitzen per temes de negoci, s'unifiquen en un format consistent provinent de fonts diverses, conserven l'historial per analitzar-ne l'evolució temporal i no s'eliminen ni es modifiquen un cop carregades.

La distinció fonamental en l'àmbit dels sistemes d'informació empresarial és la que separa els sistemes **OLTP** (*Online Transaction Processing*) dels sistemes **OLAP** (*Online Analytical Processing*). Els sistemes OLTP són els operacionals del dia a dia (el sistema de vendes, el TPV, el CRM, la facturació), dissenyats per processar un gran volum de transaccions individuals de manera ràpida i concurrent. Els sistemes OLAP, en canvi, estan dissenyats per respondre preguntes analítiques complexes sobre grans volums de dades històriques, que agreguen milions de files. El Data Warehouse és la implementació tècnica d'un sistema OLAP: un repositori centralitzat on les dades dels sistemes OLTP s'integren, es transformen i s'estructuren de manera òptima per a l'anàlisi. La Taula 1 resumeix les diferències principals entre ambdós paradigmes.

TAULA 1: COMPARATIVA OLTP VS. DATA WAREHOUSE (OLAP)

Aspecte	OLTP	OLAP / DW
Objectiu	Registrar operacions del dia a dia	Analitzar dades històriques
Consultes	Simples i ràpides	Complexes, grans volums
Dades	Actuals, detallades	Històriques, agregades
Actualitz.	Contínua	Periòdica per càrregues
Orientació	Per procés o aplicació	Per tema de negoci

Tal i com es mostra a la taula 1, els dos paradigmes responen a necessitats oposades: mentre l'OLTP prioritza la rapidesa i la concurrència en operacions individuals sobre dades actuals, l'OLAP s'orienta a consultes complexes sobre grans volums de dades històriques agregades.

En aquest treball, el dataset **Online Retail II** [4] és la sortida d'un sistema OLTP (el sistema de vendes de l'empresa), i el Data Warehouse que es construeix és un sistema OLAP. Aquesta transformació d'OLTP a OLAP és precisament un dels punts clau del treball.

Quant a la metodologia de disseny, les dues principals referències del sector són **Inmon** [1], amb un enfocament *top-down* basat en un DW corporatiu normalitzat en tercera forma normal (3FN), i **Kimball** [2], amb un enfocament *bottom-up* basat en la construcció incremental mitjançant models dimensionals. La Taula 2 resumeix les diferències principals entre ambdues aproximacions.

TAULA 2: COMPARATIVA KIMBALL VS. INMON

Aspecte	Kimball	Inmon
Model	Dimensional (estrella)	Normalitzat (3FN)
Construcció	Incremental	Global des de l'inici
Velocitat	Ràpid, aviat visible	Lent, consistent
Idoneïtat	Mida mitjana	Grans organitzacions
TFG	**Escollit**	No aplicat

La Taula 2 mostra que la diferència essencial rau en l'estratègia de construcció: l'enfocament d'Inmon parteix d'un model corporatiu global i normalitzat, adequat per a grans organitzacions, mentre que el de Kimball construeix el sistema de manera incremental sobre models dimensionals, fet que el fa més ràpid de posar en marxa i més apropiat per a projectes de mida mitjana.

Per a aquest treball s'ha adoptat la metodologia de **Ralph Kimball** [2] precisament per aquesta capacitat de lliurament incremental sense una arquitectura global prèvia, perquè és la més habitual en treballs de BI de mida mitjana com el que es descriu a la Secció 1, i és la que millor s'adapta a les eines seleccionades (Power BI i PostgreSQL). El seu cicle de vida (*Business Dimensional Lifecycle*) consta de quatre etapes que el treball segueix en ordre: elegir el procés de negoci (vendes i rendibilitat), establir la granularitat (línia de factura), elegir les dimensions (data, producte, client, geografia) i identificar les mètriques de la taula de fets (revenue, cost, marge).

2.3 Modelització dimensional

La modelització dimensional és la tècnica de disseny sobre la qual es fonamenta la metodologia Kimball. L'objectiu és crear una estructura intuïtiva per als usuaris de negoci i eficient per a les consultes analítiques i, aquesta modelització, accepta un cert grau de redundància a canvi d'un guany significatiu en velocitat de consulta i facilitat d'ús. El resultat és un **esquema estrella** (*star schema*), representat mitjançant un **diagrama relacional** compost per una taula de fets central que emmagatzema les mètriques

numèriques, envoltada de taules de dimensions que aporten el context descriptiu. Quan les dimensions es descomponen en múltiples taules jeràrquiques, l'esquema es denomina **floc de neu** (*snowflake schema*). En aquest treball s'ha optat per l'esquema estrella per la seva simplicitat i perquè el rendiment és suficient per al volum de dades del dataset.

La **taula de fets** és l'element central: conté les mesures numèriques que es volen analitzar —els *fets*— i les claus externes que la connecten amb cada dimensió, on cada fila representa un esdeveniment de negoci a la granularitat escollida. La **granularitat** és una decisió fonamental, ja que determina el nivell de detall de cada registre i, per tant, quines preguntes analítiques es podran respondre: com més fina, més flexible serà el model però més gran serà la taula. Les **taules de dimensions** contenen els atributs descriptius (qui ha comprat, quan, quin producte i des d'on), centralitzen la informació descriptiva en un sol lloc, permeten enriquir el model amb atributs derivats que no existien al dataset original i fan el model autosuficient.

Finalment, un concepte clau és l'ús de **surrogate keys** (claus substitutes): en comptes d'usar els identificadors originals del sistema origen, es generen enters correlatius nous per a cada dimensió. Aquesta pràctica protegeix el Data Warehouse dels canvis als sistemes origen, millora el rendiment dels *joins* en treballar amb enters petits, i permet gestionar casos especials com els clients sense identificar mitjançant un registre específic de valor 0, sense necessitat de valors nuls a la taula de fets.

2.4 Procés ETL

ETL és l'acrònim d'**Extract, Transform, Load** (Extracció, Transformació i Càrrega). Fa referència al conjunt de processos que mouen les dades des dels sistemes origen fins al Data Warehouse, aplicant les transformacions necessàries per garantir-ne la qualitat i l'estructura adequada. La qualitat del Data Warehouse depèn directament de la qualitat del procés ETL: un ETL mal dissenyat genera un repositori amb dades incorrectes o inconsistents, la qual cosa invalida qualsevol anàlisi posterior per molt ben construïts que estiguin els dashboards.

El procés s'estructura en tres fases. La **fase d'extracció** consisteix a obtenir les dades dels sistemes origen en el seu format original, sense transformacions, minimitzant el temps de connexió i el risc d'errors. La **fase de transformació** és la més extensa i crítica: inclou la neteja i validació (eliminació de registres no vàlids, tractament de valors nuls, verificació de tipus) i l'enriquiment i reestructuració (càlcul de mètriques derivades, extracció d'atributs addicionals i construcció de les taules dimensionals amb les seves surrogate keys), transformant el format transaccional original en l'esquema estrella. La **fase de càrrega** insereix les dades transformades al Data Warehouse respectant l'ordre d'integritat referencial: primer les dimensions i posteriorment la taula de fets. L'aplicació concreta d'aquestes tres fases sobre el dataset es descriu a la Secció 3.

3 DESENVOLUPAMENT DEL DATA WAREHOUSE

Aquesta secció descriu el procés complet de construcció del Data Warehouse: l'exploració i neteja del dataset (EDA), el procés ETL, el disseny del model dimensional i la seva implementació a PostgreSQL. L'objectiu és entendre l'estructura i qualitat de les dades, documentar les decisions de transformació preses i detallar com s'estructura el repositori analític resultant.

3.1 Anàlisi exploratòria i neteja del dataset

3.1.1 Descripció del dataset

Inicialment, la intenció del treball era tractar amb dades reals procedents d'una empresa, però no va ser possible obtenir-les per motius de confidencialitat. Davant d'aquesta situació, es va optar per un dataset públic que permetés desenvolupar el treball mantenint un nivell suficient de realisme i complexitat. Finalment s'ha seleccionat l'*Online Retail II Dataset* [4], àmpliament utilitzat en estudis de BI i anàlisi de vendes. Conté informació transaccional d'una empresa de venda al detall que comercialitza principalment productes decoratius i de regal, i cobreix el període entre l'1 de desembre de 2009 i el 9 de desembre de 2011, és a dir, aproximadament dos anys de vendes reals. El fitxer original és un Excel amb dues fulles (2009–2010 i 2010–2011); un cop concatenades, el dataset conté **1.067.371 registres** i 8 columnes. Cada fila representa una línia de factura: la venda d'un producte concret dins d'una transacció, que pot tenir múltiples files si inclou diversos productes. Es tracta en general de productes de preu baix (preu unitari mitjà de £3,38) venuts en grans quantitats per factura, coherent amb un model de distribució majorista B2B.

La informació del dataset s'estructura en vuit variables: `Invoice` (número de factura; si comença per "C" indica cancel·lació o devolució), `StockCode` (codi del producte), `Description` (descripció textual), `Quantity` (unitats venudes), `InvoiceDate` (data i hora), `Price` (preu unitari en £), `Customer ID` (identificador del client, pot ser nul si la compra és anònima) i `Country` (país de residència del client).

3.1.2 Qualitat de les dades

Abans de realitzar qualsevol anàlisi, s'ha avaluat la qualitat de les dades originals, identificant els problemes següents:

Valors nuls. La columna `Customer ID` té 243.007 registres buits (22,7% del total), probablement compres en mode convidat o al detall sense registre. La columna `Description` conté 4.382 valors nuls.

Cancel·lacions i devolucions. 19.104 registres de cancel·lació, identificats perquè el camp `Invoice` comença per 'C'. Si no s'eliminen, el revenue apareixeria inflat, doncs es comptaria la venda però no es descomptaria la devolució.

Anomalies numèriques. Aproximadament 21.000 registres amb l'atribut `Quantity` ≤ 0 i 1.744 amb `Price` ≤ 0, que no representen transaccions comercials vàlides.

StockCodes especials. 5.476 registres amb codis que no corresponen a articles reals sinó a conceptes interns (POST, D, M, BANK CHARGES, entre d'altres), que cal excloure de l'anàlisi de vendes.

Duplicats. Aproximadament 38.000 registres exactament duplicats, probablement originats per errors de càrrega del sistema transaccional original.

Les Fig. 6, 7 i 8 de l'Apèndix B il·lustren gràficament l'impacte d'aquesta neteja: els registres eliminats per motiu, els valors nuls originals i la proporció final de registres aptes per a l'anàlisi.

3.1.3 Neteja aplicada

A partir dels problemes identificats, s'ha aplicat el procés de neteja resumit a la Taula 3.

TAULA 3: RESUM DE LES DECISIONS DE NETEJA APLICADES AL DATASET

Motiu	Registres	Decisió
Cancel·lacions (`Invoice` 'C*')	19.104	Eliminats
`Quantity` o `Price` ≤ 0	∼21.000	Eliminats
StockCodes especials (POST, D, M...)	5.476	Eliminats
`Description` nul·la	4.382	Eliminats
Duplicats exactes	∼38.000	Eliminats
`Customer ID` nul	243.007	Substituïts per 'Unknown'
Registres nets resultants	**1.003.473**	**97,5% del total**

La decisió de substituir els registres amb `Customer ID` nuls per 'Unknown' en comptes d'eliminar-los respon al fet que aquestes transaccions contenen informació vàlida sobre productes, preus i dates. Excloure-les suposaria perdre el 22,7% del revenue total. A la dimensió de clients del Data Warehouse es crea un registre específic per a 'client desconegut', seguint la pràctica habitual en modelització dimensional.

3.1.4 Estadística descriptiva i anàlisi exploratòria

La Taula 4 resumeix els principals indicadors del dataset net. La quantitat mitjana per línia és d'11,15 unitats, el preu unitari mitjà de £3,38 i el revenue mitjà per línia de £19,62, aquest últim amb una desviació estàndard molt elevada que indica una distribució fortament asimètrica amb presència de comandes de gran volum.

TAULA 4: INDICADORS PRINCIPALS DEL DATASET NET

Concepte	Valor
Revenue total net	£19.687.139
Factures úniques	39.572
Productes únics	4.909
Clients identificats	5.861
Països	43
Rang temporal	01/12/2009–09/12/2011
Quantitat mitjana/línia	11,15 unitats
Preu unitari mitjà	£3,38
Revenue mitjà/línia	£19,62

L'anàlisi exploratòria (figures de l'Apèndix B) revela diversos patrons rellevants per al disseny del model:

- **Evolució temporal:** Patró estacional molt marcat, amb un pic pronunciat els mesos de novembre i desembre de cada any, coherent amb un negoci de productes de regal i decoració nadalenca (Fig. 9).
- **Distribució geogràfica:** Com es veu a la Fig. 1, el Regne Unit és el mercat principal amb el **85,5%** del revenue total (£16.832.304 sobre £19.687.139); la resta es reparteix entre 42 països, liderats per Alemanya, França i Irlanda. Això justifica la inclusió d'un atribut `is_domestic` a la dimensió geogràfica.

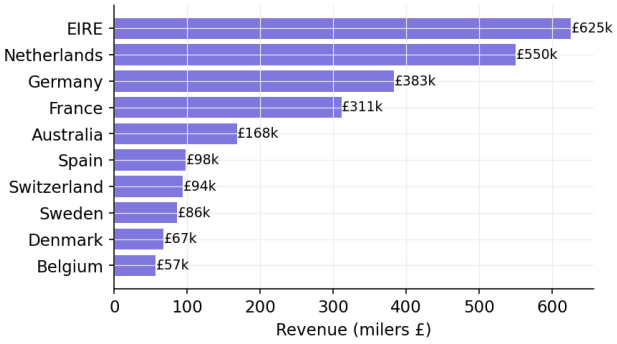


Fig. 1: Revenue per país (top 10 excloent UK)

- **Productes:** Dels 4.909 productes únics, els 15 primers concentren una proporció significativa del revenue. Els rànquings per revenue i per unitats venudes no coincideixen necessàriament (Fig. 2, 11).

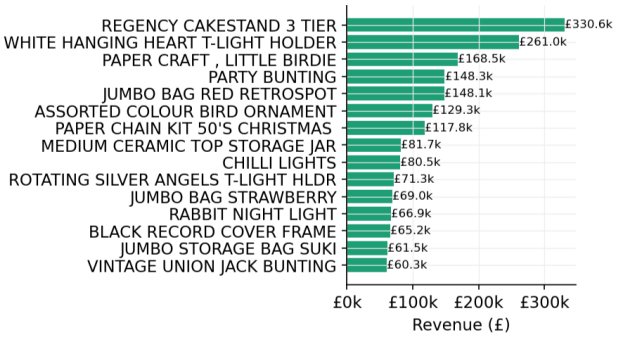


Fig. 2: Top 15 productes per revenue generat

- **Patrons temporals:** El dijous és el dia de major revenue i l'hora punta són les 12h; els caps de setmana presenten activitat molt reduïda, indicatiu d'una base de clients predominantment empresarial (Fig. 12, 13).
- **Clients:** Alta concentració del revenue. Pel principi de Pareto, el **23,1%** dels clients identificats (1.352 de 5.861) generen el **80%** del revenue (Fig. 3). El revenue medià per client és de £854,99 i la mitjana de £2.913, reflex d'uns pocs clients d'alt valor.

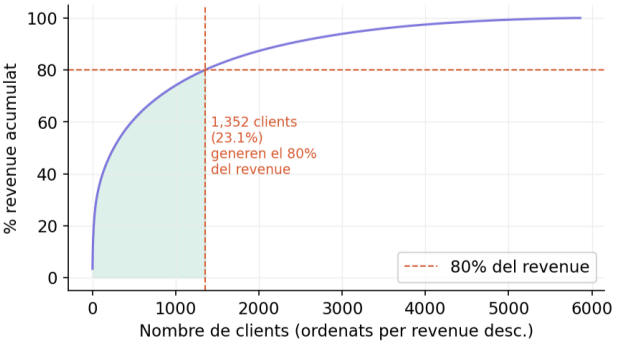


Fig. 3: Corba de Pareto — concentració del revenue

- **Ticket per factura:** El ticket medià és de £301,12 i la mitjana de £497,50, amb un màxim de £168.469,60, confirmant l'asimetria de la distribució (Fig. 15, 16).

En conjunt, l'EDA confirma una qualitat acceptable de les dades (el 97,5% dels registres són vàlids un cop netejats), justifica les dimensions geogràfica i temporal del model, i mostra un patró de compra majoritàriament en horari laboral que suggereix una base de clients B2B. A més,

l'absència de dades de costos al dataset original ha requerit la construcció d'un model de costos simulat, descrit a la Secció 3.3.

3.2 Procés ETL aplicat

El procés ETL s'ha implementat en `Python` amb la llibreria `pandas`, documentat en un `notebook Jupyter` que actua alhora com a codi executable i documentació tècnica. Consta de tres passos:

1. **Extracció.** Les dues fulles del fitxer Excel (2009–2010 i 2010–2011) es llegeixen i es concatenen en un únic dataframe mitjançant `pd.read_excel()` amb `sheet_name=None`, que llegeix totes les fulles alhora, i `pd.concat()` amb `ignore_index=True` per evitar duplicitats d'índex. El resultat és un dataframe de 1.067.371 files i 8 columnes.
2. **Transformació.** S'apliquen totes les decisions de neteja documentades a la Taula 3, seguides de l'enriquiment: càlcul del `Revenue`, extracció de camps temporals (`Year`, `Month`, `DayOfWeek`, `Hour`), assignació de categories als productes per paraules clau de la `Description`, generació del model de costos simulat (`cost_ratio` per categoria), i construcció de les cinc taules dimensionals amb surrogate keys correlatives.
3. **Càrrega.** Les cinc taules generades es carreguen a `PostgreSQL` mitjançant el connector `psycopg2`, respectant l'ordre d'integritat referencial: primer les dimensions i posteriorment la taula de fets. Es verifica que totes les claus foranes de `Fact_Sales` tenen correspondència a les dimensions abans de donar per completat el procés.

3.3 Disseny del model dimensional

El dataset net (`df_clean`) conté 1.003.473 registres en format transaccional, adequat per registrar operacions però poc eficient per respondre preguntes analítiques com el marge brut per país o les vendes per trimestre i categoria. Per aquest motiu, i seguint la metodologia Kimball (Secció 2), l'esquema estrella consta d'una taula de fets central (`Fact_Sales`) i quatre taules de dimensions (`Dim_Date`, `Dim_Product`, `Dim_Customer` i `Dim_Geography`), cadascuna connectada mitjançant una clau forana i identificada per una surrogate key. La Fig. 4 mostra el diagrama relacional resultant.

Les quatre taules de dimensions responen a les preguntes *quan*, *quin producte*, *qui* i *des d'on*, i cadascuna s'ha enriquit amb atributs que no existien al dataset original:

- `Dim_Date` (739 files): generada a partir de `InvoiceDate` amb `pd.date_range()`, una fila per dia del rang. Precalcula atributs temporals (any, mes, trimestre, dia de la setmana, `season`) que eviten càlculs repetits a cada consulta i permeten identificar l'estacionalitat amb un simple filtre.
- `Dim_Product` (4.909 files): una fila per producte únic, amb el preu unitari medià i la `category` assignada per paraules clau de la descripció. Incorpora el model de costos simulat (`unit_cost`) que fa possible el càlcul de marges.
- `Dim_Customer` (5.862 files): una fila per client (5.861 identificats més un registre 'Unknown' amb `customer_key = 0` per als 243.007 registres sense identificador), amb data de primera compra, nombre de factures i revenue acumulat.

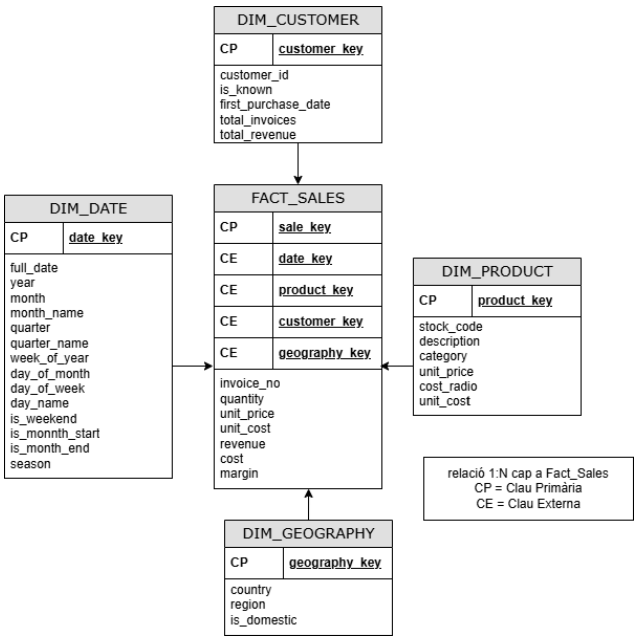


Fig. 4: Diagrama de l'esquema estrella

- `Dim_Geography` (43 files): una fila per país, amb la `region` geogràfica i l'indicador `is_domestic` assignats manualment.

L'estructura detallada d'atributs de cada dimensió i de la taula de fets es recull a l'Apèndix A (Taules 9–13), on s'especifiquen els tipus de dades, les claus primàries i foranes i la descripció de cada camp, completant així la visió general presentada en aquesta secció. Aquesta informació permet entendre amb precisió com es relacionen les taules entre si i com flueixen les dades des de les dimensions cap a la taula de fets a través de les claus externes.

3.3.1 Model de costos simulat

Atès que el dataset no inclou dades de costos, s'ha construït un model simulat assignant a cada categoria de producte l'atribut `cost_ratio` (proporció del cost sobre el preu de venda), a partir del qual es calcula el cost unitari com $\texttt{unit_cost} = \texttt{unit_price} \times \texttt{cost_ratio}$. Aquests ràtios són **estimacions raonades**, no valors extrets directament del dataset, que no en conté. Per fixar-ne l'ordre de magnitud s'han pres com a referència les dades de marge brut per sector publicades per Damodaran (NYU Stern) [14], que situen el retail entre el 30% i el 35% (*Retail General* i *Special Lines*) i els fabricants de productes de consum i llar en franges superiors (*Household Products*, 51%; *Apparel*, 57%). Atès que l'empresa estudiada opera com a majorista de productes de regal i decoració de baix cost de producció, s'ha situat el model a la franja del 45–55% de marge brut, diferenciant després cada categoria de manera relativa segons criteris qualitatius (percepció de valor, pressió competitiva i estacionalitat). Aquesta pràctica de simular mètriques absents al sistema origen és habitual en modelització dimensional [2]. Com que el cost de cada línia es calcula sobre el preu real de venda i no sobre un cost fix, el marge percentual de cada categoria és constant i igual a $1 - \texttt{cost_ratio}$, fet que garanteix la coherència interna de tots els indicadors de rendibilitat. La Taula 5 en resumeix l'assignació i la justificació.

TAULA 5: COST_RATIO PER CATEGORIA, MARGE BRUT I JUSTIFICACIÓ

Categoria	ratio	Marge	Justificació
Seasonal (Nadal...)	0,55	45%	Alta demanda estacional, marge comprimit per competència
Lighting & Candles	0,50	50%	Marge estàndard del sector
Kitchen & Dining	0,48	52%	Producte funcional, bona percepció de valor
Bags & Accessories	0,45	55%	Alta percepció de valor, baix cost de producció
Home Decor	0,50	50%	Marge estàndard del sector
Children & Toys	0,52	48%	Mercat competitiu, marge menor
Stationery & Giftwrap	0,45	55%	Cost de producció baix, bon marge
Storage & Organisation	0,50	50%	Marge estàndard del sector
Other	0,52	48%	Valor conservador per defecte

3.3.2 La taula de fets

`Fact_Sales` és la taula central de l'esquema, amb una fila per línia de factura (1.003.473 registres). Emmagatzema les quatre claus externes cap a les dimensions i sis mètriques: `quantity`, `unit_price`, `unit_cost`, `revenue`, `cost` i `margin`, on les tres últimes es calculen com:

$$\texttt{revenue} = \texttt{quantity} \times \texttt{unit_price} \quad (1)$$

$$\texttt{cost} = \texttt{quantity} \times \texttt{unit_cost} \quad (2)$$

$$\texttt{margin} = \texttt{revenue} - \texttt{cost} \quad (3)$$

Gràcies a la columna `margin`, el Data Warehouse permet calcular no només indicadors de vendes sinó també de rendibilitat (marge brut per producte, client, país o període), un dels objectius principals del treball. La taula es genera a partir del `df_clean` mitjançant merges successius amb les quatre dimensions per incorporar-ne les claus foranes, i finalment es verifica que totes tinguin correspondència a les dimensions respectives (cap clau externa sense referència). La taula 6 resumeix les cinc taules del model.

TAULA 6: RESUM DE LES TAULES DEL MODEL DIMENSIONAL

Taula	Files	Col.	Font
`Dim_Date`	739	15	InvoiceDate
`Dim_Product`	4.909	7	StockCode + costos simulats
`Dim_Customer`	5.862	6	Customer ID
`Dim_Geography`	43	4	Country
`Fact_Sales`	1.003.473	12	df_clean + merges

3.4 Implementació a PostgreSQL

Un cop construït el model dimensional en Python, on les taules resideixen només en memòria, s'ha traslladat a un sistema gestor de bases de dades relacional per emmagatzemar-lo de manera permanent, cosa que permet consultes SQL analítiques eficients i la connexió posterior amb les eines de visualització. S'ha instal·lat **PostgreSQL 17** per a Windows, juntament amb **pgAdmin 4** com a eina d'administració, i s'ha creat la base de dades `tfg_dw`, amb connexió des del notebook Jupyter mitjançant `psycopg2-binary`.

Les cinc taules s'han creat en SQL des de Python definint claus primàries, claus foranes i tipus de dades, respectant l'ordre d'integritat referencial (primer les dimensions, després la taula de fets). La Taula 7 mostra el recompte final de files carregades, que coincideix exactament amb el generat durant l'ETL.

TAULA 7: FILES CARREGADES A POSTGRESQL PER TAULA

Taula	Files carregades
`dim_date`	739
`dim_product`	4.909
`dim_customer`	5.862
`dim_geography`	43
`fact_sales`	1.003.473
Total	**1.015.026**

Finalment, s'ha verificat la integritat mitjançant consultes de recompte sobre cada taula i la comprovació que totes les claus foranes de `fact_sales` tenen correspondència a les dimensions respectives. El resultat ha estat satisfactori, sense cap clau externa orfena: el Data Warehouse està operatiu i llest per a les consultes analítiques.

Com a validació empírica de l'eficiència del model, una consulta KPI representativa (el càlcul de revenue i marge per categoria, que agrega tot el milió de registres de la taula de fets i el creua amb la dimensió de producte) s'executa en aproximadament 550 ms, confirmant que l'esquema estrella ofereix un rendiment adequat per a l'anàlisi OLAP malgrat el volum de dades.

4 OBTENCIÓ DE RESULTATS: KPIS

Un cop el Data Warehouse es troba operatiu a PostgreSQL, s'han definit i calculat nou indicadors clau de negoci (KPIs) mitjançant consultes SQL que creuen la taula de fets amb les dimensions corresponents. Els KPIs cobreixen les quatre dimensions analítiques principals del treball: geografia, producte, temps i client.

Seguidament s'expliquen cadascun dels KPIs:

4.1 KPI 1: Revenue i quota per país

Resultat i interpretació: El Regne Unit concentra el 85,55% del revenue total (£16,8M de £19,7M). Irlanda (EIRE) és el segon mercat amb £625k (3,17%), seguida de Països Baixos (£550k) i Alemanya (£383k). Aquesta concentració extrema és coherent amb el perfil d'una SME britànica en fase de consolidació del mercat domèstic, on l'exportació és incipient però creixent. La presència en 43 països indica potencial d'expansió internacional no explotada.

4.2 KPI 2: Revenue i marge per categoria de producte

Resultat i interpretació: La categoria "Other" lidera en revenue (£6,4M) per volum, però Bags & Accessories i Stationery & Giftwrap són les categories més rendibles, totes dues amb un marge del 55%, seguides de Kitchen & Dining (52%). Totes les categories se situen entre el 45% i el 55% de marge brut, coherent amb el model de costos definit per categoria, cosa que confirma que l'empresa opera en un segment de producte amb bona percepció de valor i marges saludables. Per a decisions estratègiques, les categories Bags & Accessories i Stationery haurien de ser prioritzades en accions de creixement per la seva rendibilitat superior.

4.3 KPI 3: Evolució mensual del revenue

Resultat i interpretació: El patró estacional és molt pronunciat i consistent en els dos anys del dataset. El revenue mínim es registra al febrer (£508–538k) i el màxim al novembre (£1.436–1.452k), una variació de gairebé 3 vegades. Novembre supera desembre, indicant que els clients B2B fan les seves comandes nadalenques amb un mes d'antelació. Aquest patró té implicacions directes per a la gestió d'estoc i la planificació financera de l'empresa.

4.4 KPI 4: Top 10 clients per revenue

Resultat i interpretació: El client top (ID 18102) ha generat £580.987 en 145 factures, amb un ticket mitjà de £558,64 per factura, molt superior a la mitjana global. Els deu primers clients concentren aproximadament el 14% del revenue total, amb una freqüència de compra molt alta (alguns amb més de 150 factures en 2 anys), la qual cosa indica relacions comercials estables i recurrents característiques del B2B. La dependència d'uns pocs clients d'alt valor representa tant una oportunitat (fidelització) com un risc de concentració.

4.5 KPI 5: Basket size (unitats per factura)

Resultat i interpretació: La mediana del basket size és de 148 unitats per factura, mentre que la mitjana és de 282 unitats. La diferència entre ambdós valors (91%) confirma l'existència de comandes de gran volum que distorsionen la distribució. Aquests valors són molt superiors als d'un retailer B2C típic (on el basket size seria d'1–5 unitats), la qual cosa reforça definitivament la hipòtesi B2B: els clients compren en quantitats de majorista per a revenda o decoració de locals.

4.6 KPI 6 i 7: Marge brut per mes i rendibilitat global

Resultat i interpretació: El marge brut mensual es manté molt estable al voltant del 50% (entre el 49,7% i el 51,2%) al llarg de tot el període. La rendibilitat global és d'un revenue de £19,7M, un cost de £9,8M i un marge brut de £9,87M (50,14%). Aquesta estabilitat del marge al llarg de l'any és conseqüència directa del model de costos, que aplica un percentatge de cost proporcional al preu de venda de cada categoria: el marge global, doncs, reflecteix la barreja de categories venudes en cada període, que es manté força constant.

4.7 KPI 8: Marge brut per país (top 10)

Resultat i interpretació: Els marges per país són molt homogenis, tots situats en una franja estreta entre el 50,4% i el 51,5%. Els mercats amb marge més alt són Sud-àfrica (RSA, 51,46%) i Portugal (51,31%). Aquesta homogeneïtat és coherent amb el model de costos aplicat: com que el marge de cada venda depèn del ràtio de cost de la seva categoria i no de polítiques de preus diferenciades per país, les diferències entre mercats són petites i reflecteixen únicament la composició del catàleg comprat a cada destinació. Així, un mercat amb més pes de categories d'alt marge (com Bags & Accessories) presenta un marge global lleugerament superior.

4.8 KPI 9: Top 10 productes per marge

Resultat i interpretació: El producte REGENCY CAKESTAND 3 TIER encapçala el rànquing amb £171.905 de marge brut i un 52% de marge percentual. Els productes de les categories Bags & Accessories i Stationery (com JUMBO BAG RED WHITE SPOTTY o PAPER CRAFT LITTLE BIRDIE) assoleixen el marge percentual més alt (55%), confirmant que aquestes categories ofereixen la rendibilitat relativa més alta del catàleg. De manera general, el top 10 de productes per marge absolut coincideix en gran part amb el top per revenue, indicant que els productes més venuts també són rendibles, sense que hi hagi una tensió evident entre volum i marge.

4.9 Síntesi dels resultats

La taula 8 resumeix els nou KPIs i els resultats obtinguts.

TAULA 8: RESUM DELS 9 KPIS IMPLEMENTATS

KPI	Àmbit	Resultat clau
1	Revenue per país	UK = 85,55%
2	Marge per categoria	Bags & Acc. = 55%
3	Evolució mensual	Pic nov–des ($\times$3 sobre mínim)
4	Top 10 clients	Client top = £580.987
5	Basket size	Mediana 148 u./factura
6	Marge per mes	Estable ($\sim$50%)
7	Rendibilitat global	Rev. £19,7M · Marge 50,14%
8	Marge per país	Homogeni ($\sim$50%)
9	Productes rendibles	REGENCY CAKESTAND encapçala

5 DASHBOARD ANALÍTIC

La visualització dels resultats s'ha desenvolupat en dues fases: un prototip ràpid en **Streamlit**, per validar el disseny i la selecció dels KPIs, i el **dashboard final en Power BI**, com a producte de visualització definitiu connectat directament al Data Warehouse.

5.1 Prototip en Streamlit

Com a primera fase es va construir un prototip amb **Streamlit** [8], escollit per la seva integració nativa amb Python, que permet reutilitzar directament el codi i les estructures de dades del procés ETL, i per la velocitat de prototipatge. Aquest prototip [15] va servir per iterar sobre el disseny de les visualitzacions i validar la distribució dels KPIs en quatre pàgines (visió general, rendibilitat, clients i geografia) abans d'implementar el dashboard definitiu. Utilitza **Plotly** per a les visualitzacions interactives i estableix l'estructura de pàgines i la paleta de colors que després es reprodueix a Power BI.

5.2 Dashboard final en Power BI

El dashboard definitiu s'ha implementat en **Power BI Desktop** [13], connectat directament a la base de dades `tfg_dw` de PostgreSQL mitjançant el connector natiu en mode *Import*. Aquesta connexió permet que el dashboard reflecteixi l'estat del Data Warehouse sense necessitat d'exportacions intermèdies: n'hi ha prou amb actualitzar el model perquè reimporti les dades actualitzades.

Un cop importades les cinc taules, Power BI detecta automàticament les quatre relacions entre la taula de fets i les dimensions a partir de les claus externes, reproduint l'esquema estrella en la vista de model amb cardinalitat *molts-a-un* i filtre unidireccional des de cada dimensió cap als

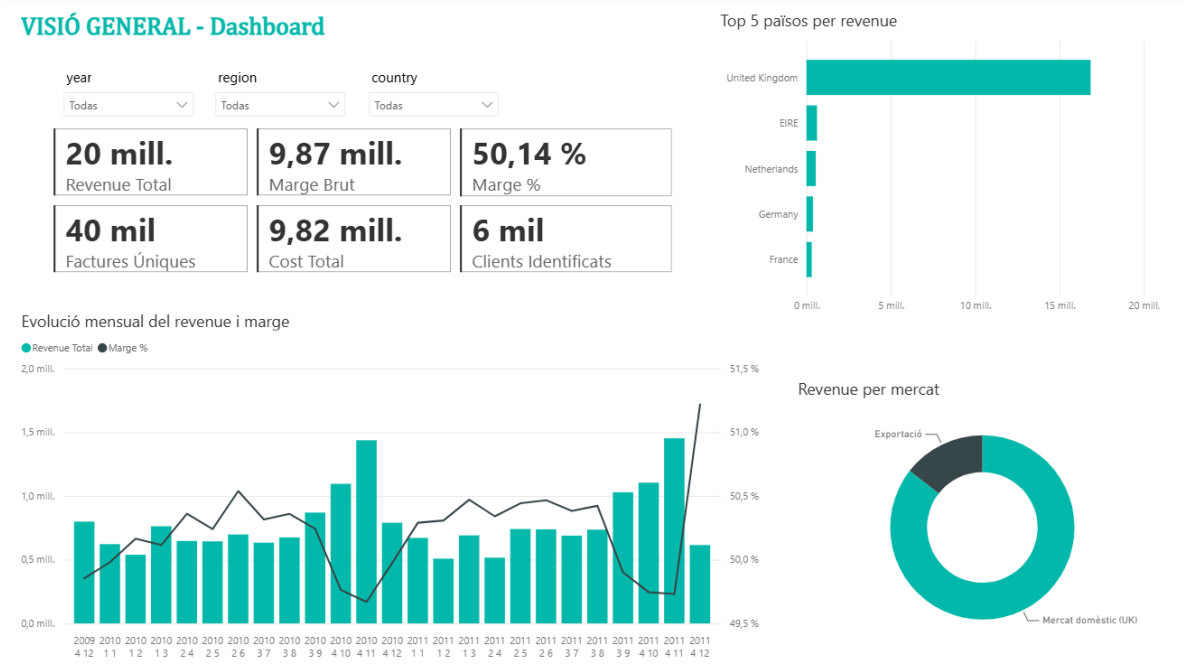


Fig. 5: Pàgina de visió general del dashboard de Power BI

fets. Sobre aquest model s'han definit sis mesures en llenguatge **DAX** (*Data Analysis Expressions*) que es recalculen dinàmicament segons els filtres aplicats: `Revenue Total`, `Cost Total`, `Marge Brut` i `Marge %` per als indicadors de rendibilitat, i `Factures Úniques` i `Clients Identificats` per a les mètriques de volum i client. El marge percentual fa servir la funció `DIVIDE` per gestionar amb seguretat la divisió per zero, i el recompte de clients exclou el registre `Unknown` (`customer_key = 0`) per comptar únicament els clients identificats.

A partir d'aquestes mesures, el dashboard organitza els resultats en quatre pàgines alineades amb el prototip i amb la secció de KPIs:

- **Pàgina 1: Visió general.** Targetes amb les mètriques principals del negoci, gràfic combinat de l'evolució mensual del revenue i el marge, distribució entre mercat domèstic i exportació, i rànquing dels cinc primers mercats per revenue (Fig. 5).
- **Pàgina 2: Rendibilitat per categoria i producte.** Marge brut per categoria amb la línia de referència de la mitjana global, comparativa de revenue, cost i marge per categoria, i taula dels deu productes més rendibles (Fig. 17).
- **Pàgina 3: Anàlisi de clients.** Top de clients per revenue, distribució del revenue entre clients identificats i anònims, basket size mitjà per mes i indicadors de comportament de compra (revenue medià i basket size) (Fig. 18).
- **Pàgina 4: Anàlisi geogràfica.** Rànquing de països per revenue, marge brut per país i distribució del revenue per regió geogràfica (Fig. 19).

A diferència del prototip, el dashboard final incorpora les capacitats interactives natives de Power BI: **segmentadors** (*slicers*) per dimensió temporal, geogràfica i de categoria, que filtren totes les visualitzacions de la pàgina de manera coordinada, i **drill-down** sobre la jerarquia temporal (any → trimestre → mes) i geogràfica (regió → país), que permet aprofundir en el detall sense canviar de visual. El resultat és una eina d'exploració que permet a un usuari de negoci respondre de manera autònoma les preguntes analítiques plantejades a l'inici del treball.

6 Conclusions i millores

S'han completat totes les etapes previstes per a la construcció d'un Data Warehouse orientat a l'anàlisi de vendes i rendibilitat: el disseny i la implementació del model dimensional estrella, el procés ETL complet sobre 1.003.473 registres de vendes, la càrrega i verificació a PostgreSQL, el càlcul de nou KPIs que responen les preguntes de negoci plantejades a l'inici, i el desenvolupament de quadres de comandament interactius en Streamlit i Power BI. El sistema construït converteix un conjunt de dades transaccional en una plataforma analítica capaç de respondre, de manera eficient i des de múltiples perspectives, les preguntes estratègiques sobre productes, clients, mercats i períodes.

Els resultats permeten caracteritzar amb precisió l'empresa estudiada: un majorista B2B britànic de decoració i regal, amb £19,7M de revenue en dos anys, un marge brut global del 50,14%, una base de clients altament concentrada (el 23% dels clients generen el 80% del revenue) i una estacionalitat nadalenca molt pronunciada. Cal subratllar que les conclusions més sòlides del treball, com ara la concentració de clients segons Pareto, el pic estacional de novembre, la quota del 85% del mercat domèstic i els patrons de compra característics del B2B, es deriven **directament de les dades reals**, sense dependre de cap supòsit del model de costos, i demostren que l'arquitectura dimensional respon les preguntes de negoci de forma fiable i eficient.

Quant a les limitacions, la principal és l'ús d'un **model de costos simulat**: com que el dataset no inclou costos reals, el marge s'ha calculat a partir de ràtios per categoria justificats amb benchmarks sectorials. Per construcció, el marge de cada categoria és igual a un menys el seu ràtio de cost, de manera que els KPIs de rendibilitat tenen **valor**

metodològic (demostren que el sistema calcula rendibilitat correctament) més que de descobriment analític, i els seus valors absoluts no poden validar-se contra la realitat de l'empresa. En segon lloc, la classificació de productes per paraules clau de la descripció és una aproximació pragmàtica però amb error inherent, evidenciat per la mida de la categoria "Other". Finalment, el treball s'ha desenvolupat sobre un **dataset públic**; l'Online Retail II, però, és prou realista i complet per demostrar totes les fases del treball.

A partir de la base construïda, s'identifiquen diverses línies de treball futur:

- Incorporació de dades de costos reals per validar el model simulat i obtenir marges contrastables amb la realitat de l'empresa.
- Implementació d'un model de segmentació RFM (Recència, Freqüència, Valor Monetari) sobre la dimensió de clients per identificar segments estratègics.
- Anàlisi de la cistella de la compra (*market basket analysis*) per detectar productes comprats conjuntament, candidats a campanyes de *cross-selling*.

AGRAÏMENTS

Vull agrair a l'Escola d'Enginyeria de la Universitat Autònoma de Barcelona i al professorat del grau d'Enginyeria de Dades la formació rebuda al llarg d'aquests anys, que m'ha donat les bases per afrontar un treball d'aquesta naturalesa. Vull agrair també al tutor Enric Martí Gòdia per el seguiment durant el desenvolupament del treball.

A la família i amics, pel seu suport constant i la seva paciència durant aquest camí. I, per acabar, a mi mateixa, per no rendir-me i fer realitat aquest treball.

REFERÈNCIES

[1] W. H. Inmon, *Building the Data Warehouse*, 4th ed. Wiley, 2005.

[2] R. Kimball i M. Ross, *The Data Warehouse Toolkit*, 2nd ed. Wiley, 2002.

[3] Oracle España, "¿Qué es la inteligencia de negocio?" [En línia]. Disponible: https://www.oracle.com/es/what-is-business-intelligence/

[4] Kaggle, "Online Retail Dataset." [En línia]. Disponible: https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset

[5] J. Curto Díaz, *Introducción al Business Intelligence*. Editorial UOC, 2016.

[6] H. J. Dresner, "Business Intelligence: A term coined to describe a set of concepts and methods to improve business decision making by using fact-based support systems," Gartner Group, 1989. Citat a: H. J. Watson i B. H. Wixom, "The Current State of Business Intelligence," *Computer*, vol. 40, no. 9, pp. 96–99, 2007.

[7] The PostgreSQL Global Development Group, *PostgreSQL 17 Documentation*. [En línia]. Disponible: https://www.postgresql.org/docs/17/

[8] Streamlit Inc., *Streamlit Documentation*. [En línia]. Disponible: https://docs.streamlit.io/

[9] Python Software Foundation, *Python Language Reference, version 3.10*. [En línia]. Disponible: https://www.python.org/

[10] W. McKinney, "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference*, 2010. [En línia]. Disponible: https://pandas.pydata.org/

[11] Project Jupyter, *Jupyter Notebook Documentation*. [En línia]. Disponible: https://jupyter.org/

[12] F. Di Gregorio i D. Varrazzo, *Psycopg – PostgreSQL adapter for Python*. [En línia]. Disponible: https://www.psycopg.org/

[13] Microsoft Corporation, *Power BI Desktop Documentation*. [En línia]. Disponible: https://learn.microsoft.com/power-bi/

[14] A. Damodaran, "Margins by Sector (US) — Profit Margins," NYU Stern School of Business, gen. 2026. [En línia]. Disponible: https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html

[15] J. Pallarés Genaro, *Dashboard analític del treball (Streamlit)*, 2026. [En línia]. Disponible: https://drive.google.com/file/d/14C4p7g3IHjdAzzipPGj2NZ08mEsR0e8h/view?usp=drive_link

A ESTRUCTURA DETALLADA DE LES TAULES DEL MODEL DIMENSIONAL

TAULA 9: ATRIBUTS DE LA TAULA DIM_DATE (739 FILES, 15 COLUMNES)

Atribut	Tipus	Descripció
date_key	PK INT	Format YYYYMMDD. FK a Fact_Sales
full_date	DATE	Data completa. Ex: 2010-12-15
year	INT	Any: 2009, 2010, 2011
month	INT	Mes numèric: 1–12
month_name	VARCHAR	Nom del mes. Ex: 'December'
quarter	INT	Trimestre: 1, 2, 3 o 4
quarter_name	VARCHAR	Nom. Ex: 'Q4'
week_of_year	INT	Setmana ISO: 1–52
day_of_month	INT	Dia del mes: 1–31
day_of_week	INT	Dia setmana: 1=Dll, 7=Dg
day_name	VARCHAR	Nom del dia. Ex: 'Thursday'
is_weekend	BOOL	True si dissabte o diumenge
is_month_start	BOOL	True si 1r dia del mes
is_month_end	BOOL	True si últim dia del mes
season	VARCHAR	Winter/Spring/Summer/Autumn

TAULA 10: ATRIBUTS DE LA TAULA DIM_PRODUCT (4.909 FILES, 7 COLUMNES)

Atribut	Tipus	Descripció
product_key	PK INT	Surrogate key correlatiu
StockCode	VARCHAR	Codi original del dataset
description	VARCHAR	Nom del producte
category	VARCHAR	Seasonal, Home Decor, etc.
unit_price	NUMERIC	Preu unitari medià del dataset
cost_ratio	NUMERIC	% cost sobre el preu per categoria
unit_cost	NUMERIC	unit_price × cost_ratio

TAULA 11: ATRIBUTS DE LA TAULA DIM_CUSTOMER (5.862 FILES, 6 COLUMNES)

Atribut	Tipus	Descripció
customer_key	PK INT	0 reservat per 'Unknown'
customer_id	VARCHAR	ID original o 'Unknown'
is_known	BOOL	True si el client té ID
first_purchase_date	DATE	Mínim InvoiceDate per client
total_invoices	INT	Nombre total de factures
total_revenue	NUMERIC	Revenue total acumulat

TAULA 12: ATRIBUTS DE LA TAULA DIM_GEOGRAPHY (43 FILES, 4 COLUMNES)

Atribut	Tipus	Descripció
geography_key	PK INT	Surrogate key (1–43)
country	VARCHAR	Nom del país original
region	VARCHAR	British Isles, W. Europe, etc.
is_domestic	BOOL	True només per a UK

TAULA 13: ATRIBUTS DE LA TAULA FACT_SALES (1.003.473 FILES, 12 COLUMNES)

Atribut	Tipus	Descripció
sale_key	PK INT	Surrogate key
date_key	FK INT	→ Dim_Date
product_key	FK INT	→ Dim_Product
customer_key	FK INT	→ Dim_Customer (0=Unknown)
geography_key	FK INT	→ Dim_Geography
invoice_no	VARCHAR	Traçabilitat al sistema origen
quantity	INT	Unitats venudes
unit_price	NUMERIC	Preu unitari en £
unit_cost	NUMERIC	Cost unitari simulat
revenue	NUMERIC	qty × unit_price
cost	NUMERIC	qty × unit_cost
margin	NUMERIC	revenue − cost

B FIGURES DE L'ANÀLISI EXPLORATÒRIA

Aquest apèndix recull les figures generades durant l'anàlisi exploratòria (EDA) del dataset, referenciades al llarg de la Secció 3. S'organitzen en quatre blocs: qualitat de les dades, distribució geogràfica i de productes, patrons temporals i anàlisi de clients i tiquet.

Qualitat de les dades

Les tres figures següents quantifiquen l'impacte del procés de neteja descrit a la Taula 3. La Fig. 6 desglossa els registres eliminats segons el motiu, la Fig. 7 mostra la presència de valors nuls al dataset original i la Fig. 8 resumeix la proporció final de registres vàlids per a l'anàlisi (97,5%).

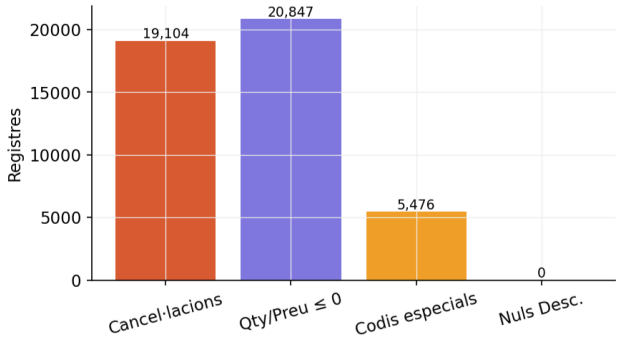


Fig. 6: Registres eliminats per motiu

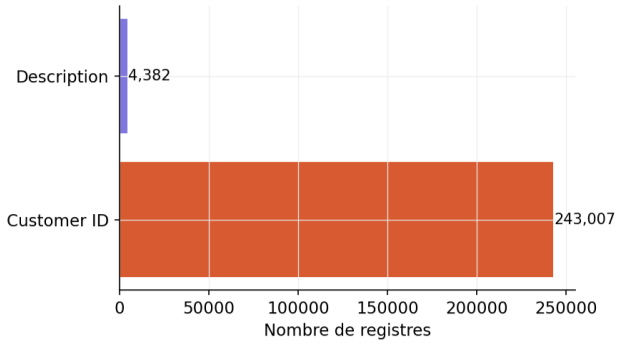


Fig. 7: Valors nuls originals

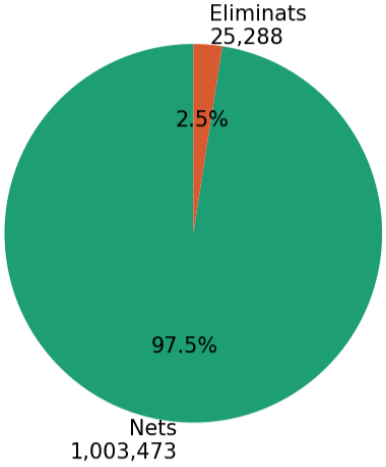


Fig. 8: Registres aptes per a l'anàlisi

Evolució temporal i distribució geogràfica

La Fig. 9 mostra l'evolució mensual del revenue i del nombre de factures al llarg dels dos anys, on s'aprecia el fort pic estacional de novembre i desembre. La Fig. 10 presenta el rànquing de països per revenue incloent-hi el Regne Unit, que evidencia la concentració del mercat domèstic (85,5%) que justifica l'atribut `is_domestic` de la dimensió geogràfica.

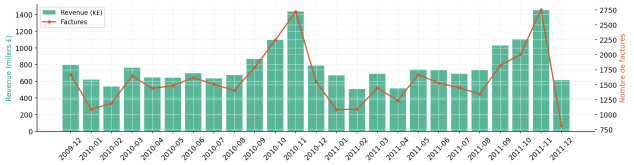


Fig. 9: Evolució mensual del revenue i les factures (2009–2011)

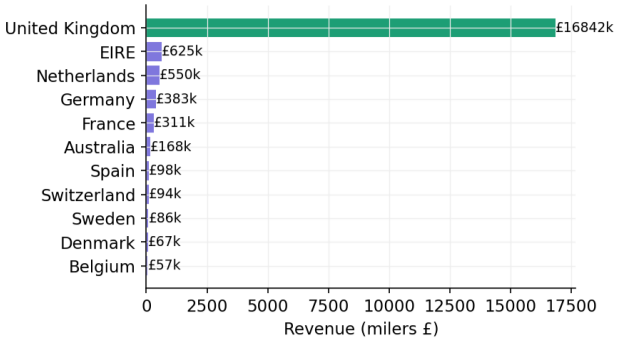


Fig. 10: Revenue per país (top 10 incloent UK)

Productes

La Fig. 11 ordena els quinze productes amb més unitats venudes. El rànquing no coincideix amb el dels productes amb més revenue (Fig. 2), ja que productes de preu baix poden vendre molt volum sense liderar en facturació.

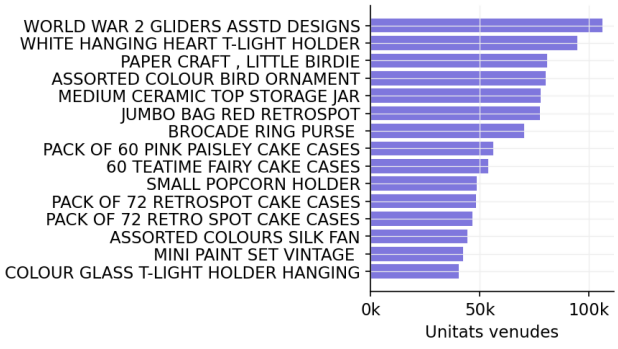


Fig. 11: Top 15 productes per unitats venudes

Patrons temporals de compra

Les dues figures següents reforcen la hipòtesi d'una base de clients B2B. La Fig. 12 mostra que el revenue es concentra en dies laborables, amb el dijous com a dia de major activitat, mentre que la Fig. 13 situa l'hora punta de compra al migdia (12h), patró propi d'una activitat empresarial en horari d'oficina.

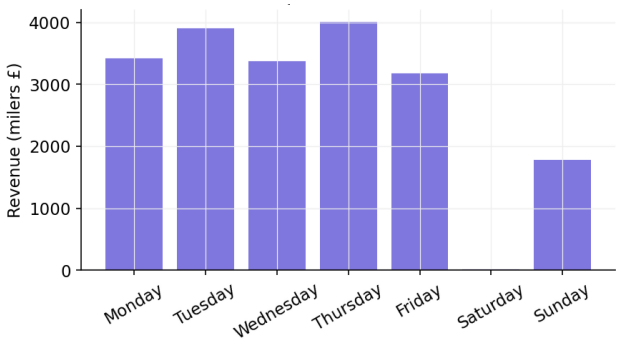


Fig. 12: Revenue per dia de la setmana

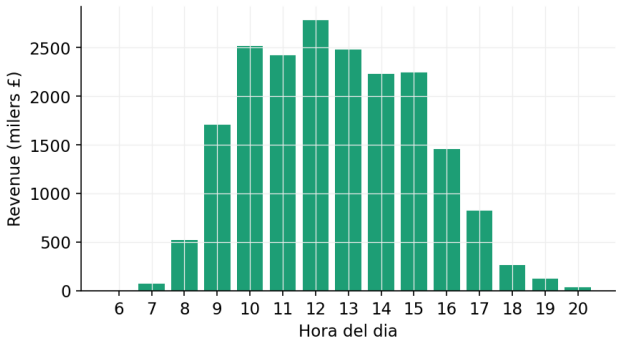


Fig. 13: Revenue per hora del dia

Anàlisi de clients i tiquet per factura

Aquest darrer bloc il·lustra l'elevada concentració del revenue. La Fig. 14 mostra la distribució del revenue per client, fortament asimètrica. Les Fig. 15 i 16 mostren, respectivament, la distribució del valor per factura i la seva evolució per any, confirmant la presència de comandes de gran volum que distorsionen la mitjana respecte de la mediana.

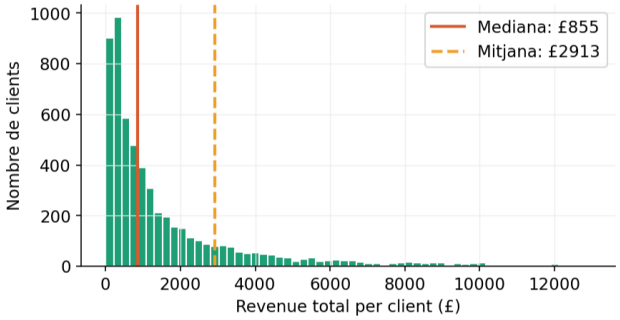


Fig. 14: Distribució del revenue per client

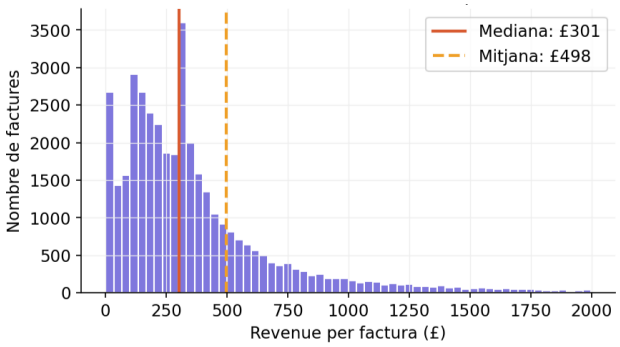


Fig. 15: Distribució del valor per factura

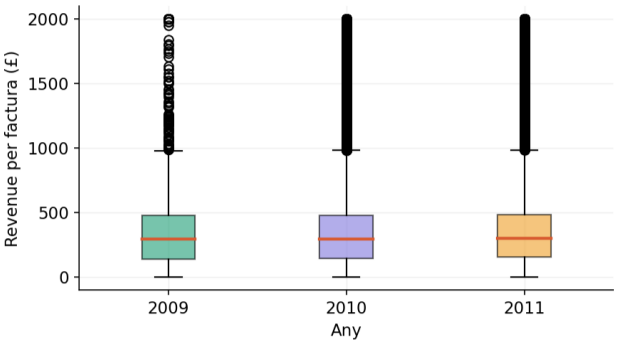


Fig. 16: Distribució del ticket per any

C PÀGINES ADDICIONALS DEL DASHBOARD DE POWER BI

Aquest apèndix recull les tres pàgines restants del dashboard de Power BI, complementant la pàgina de visió general presentada a la Fig. 5 de la Secció 5.

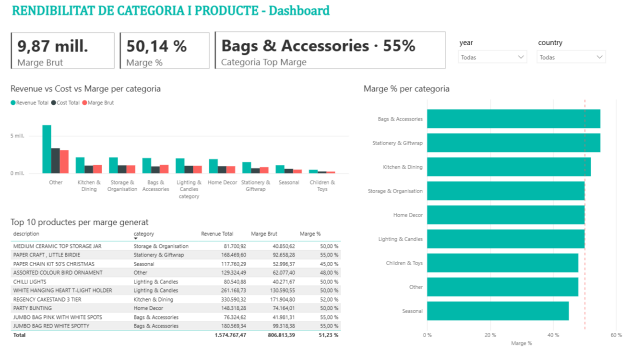


Fig. 17: Pàgina de rendibilitat per categoria i producte

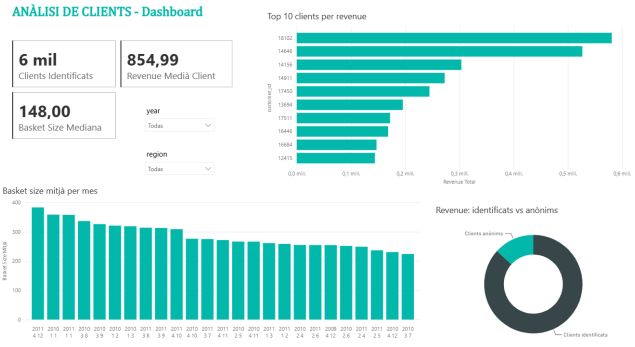


Fig. 18: Pàgina d'anàlisi de clients

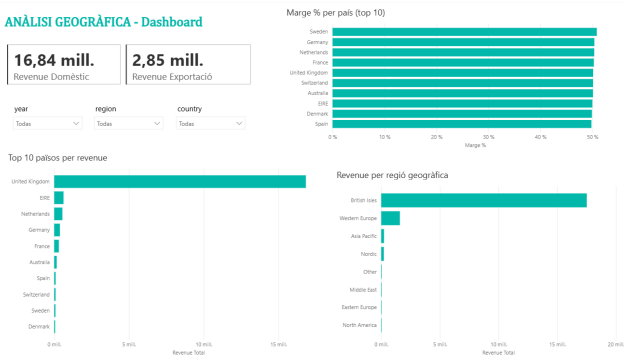


Fig. 19: Pàgina d'anàlisi geogràfica