

BI-Housing

Reporte de ventas para Abarrotes Lugo S. A. de C. V.

Administración de Big Data

Reto 4

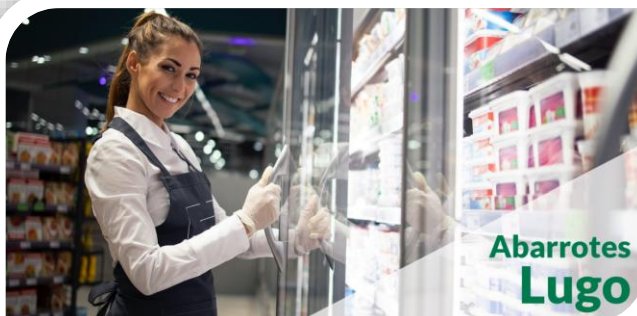
Calidad de la información y reportes para las organizaciones

Nombre: [Josué Contreras Calderón](#)

Matrícula: [21037074](#)

Nombre del Asesor: [Deysy Galeana Pérez](#)

Fecha de elaboración: [12 de septiembre de 2026](#)



Título: **Reporte de ventas por categoría de productos**

Para: Ing. Juan Pérez, gerente general, y Departamento de Ventas de Abarrotes Lugo S. A. de C. V.

De: 12 de enero de 2026 **A:** 12 de abril de 2026

Estaciones: las 20 sucursales (TlaxTL, ZacaZA, ColimaCOL, GroOA, MonteNL, DuraDGO, TepicNA, VallaJAL, LeonGTO, SilaoGTO, PuebPUE, TijuBJC, ToluEMX, CancQR, GuadaJAL, QroQRO, ZacaMI, JuareCHIH, ZapoJAL y NezaEMX).

Objetivo

Identificar qué categoría de productos vende más y cuál vende menos en las 20 sucursales, de enero a abril de 2026, para orientar las decisiones de inversión de Abarrotes Lugo: surtido, espacio en anaquel y compras. Las cifras son unidades vendidas (las fuentes no incluyen precio ni moneda).

Metodología

Las fuentes son las bases de datos de punto de venta de las 20 sucursales; sus datos se concentraron en el Data Warehouse (almacén central de datos) creado en enero. Sobre esos datos se aplicaron los tres grupos de herramientas de Data Management, la administración de los datos (UVEG, s. f.-f):

1. Consulta de datos: con herramientas de reportes se extrajeron las unidades vendidas por categoría y sucursal.
2. Data Mining (la extracción de datos de calidad): se removieron registros erróneos o duplicados, se unificaron nombres de categoría que variaban entre sucursales y se validó que cada sucursal tuviera sus ocho categorías completas.
3. EIS y DSS (sistemas de información ejecutiva y de soporte de decisión): con el EIS se monitorearon totales, promedios y porcentajes por categoría y sucursal; con el DSS se armaron las recomendaciones de inversión.

Los datos ya limpios pasaron al Data Mart de ventas, el subconjunto que usa el Departamento de Ventas. Las tablas y gráficas se generaron con código en Python y HTML/SVG (gráficas vectoriales responsivas que no pierden calidad al cambiar su tamaño), para repetir el reporte cada periodo.

Resultados

Sucursal	Enlatado	Botanas	Confitería	Harinas	Frutas y verduras	Bebidas	Bebidas alcohólicas	Automedicación	Total
GuadaJAL	517	567	448	546	500	544	428	477	4,027
TijuBJC	440	562	562	490	521	595	375	459	4,004
CancQR	547	543	532	524	533	357	536	337	3,909
ColimaCOL	467	361	564	569	574	557	415	383	3,890
ToluEMX	327	442	525	470	459	598	480	545	3,846
LeonGTO	518	420	501	389	391	551	418	575	3,763
ZacaZA	577	430	318	484	310	469	557	587	3,732
ZacaMI	303	469	490	480	568	442	408	571	3,731
GroOA	406	546	458	391	355	493	583	489	3,721
NezaEMX	403	545	537	501	333	370	596	388	3,673
JuareCHIH	588	426	432	462	350	476	423	442	3,599
SilaoGTO	558	570	346	579	307	387	350	471	3,568
DuraDGO	335	304	573	491	392	557	327	577	3,556
QroQRO	589	372	494	512	495	326	316	423	3,527
VallaJAL	563	370	363	321	449	542	453	387	3,448
PuebPUE	462	510	370	395	550	375	376	394	3,432
TepicNA	536	583	363	318	363	421	423	390	3,397
TlaxTL	565	458	513	413	424	375	309	301	3,358
MonteNL	438	303	424	590	356	447	417	373	3,348
ZapoJAL	359	459	300	395	389	343	452	395	3,092
Total	9,498	9,240	9,113	9,320	8,619	9,225	8,642	8,964	72,621

Tabla 1. Unidades vendidas por sucursal y categoría, ordenadas de mayor a menor venta total (verde: máximo de cada categoría; rojo: mínimo).

Se vendieron 72,621 unidades. Cada categoría se mueve entre 300 y 598 unidades según la sucursal, y ninguna domina en todas. QroQRO, por ejemplo, tiene el máximo de Enlatado (589) y el mínimo de Bebidas (326): la venta depende tanto de la sucursal como del producto.

Unidades vendidas por categoría, enero a abril de 2026

Suma de las 20 sucursales. Total general: 72,621 unidades.

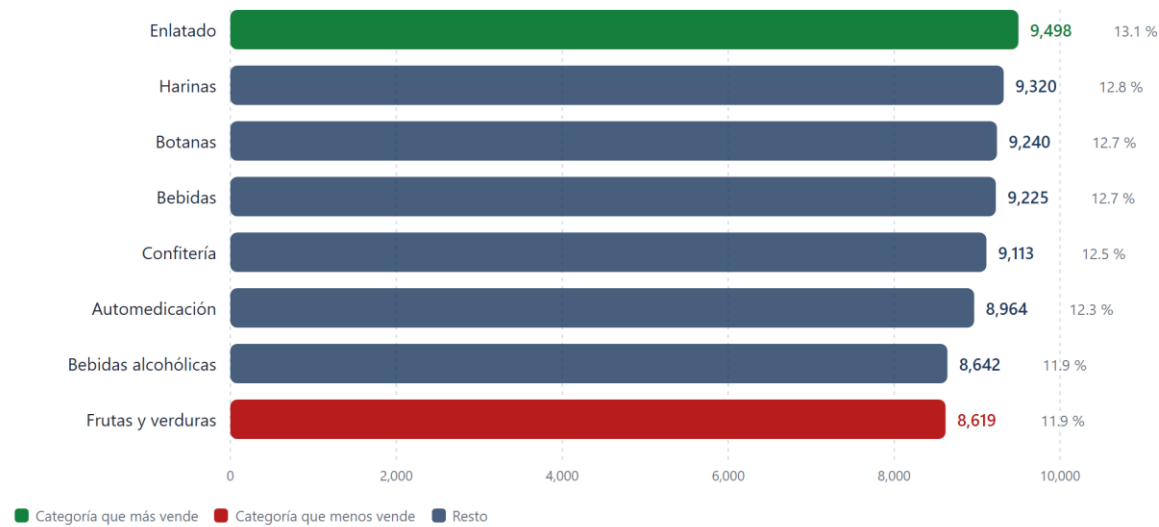


Figura 1. Total por categoría, de mayor a menor (verde: la que más vende; rojo: la que menos).

Enlatado encabeza con 9,498 unidades y Frutas y verduras cierra con 8,619. Las barras son casi del mismo largo: la diferencia es de 879 unidades, apenas 10.2 % de la peor.

#	Categoría	Total (unidades)	Promedio por sucursal	% del total	Sucursal con más venta	Sucursal con menos venta
1	Enlatado	9,498	474.9	13.1 %	QroQRO (589)	ZacaMI (303)
2	Harinas	9,320	466.0	12.8 %	MonteNL (590)	TepicNA (318)
3	Botanas	9,240	462.0	12.7 %	TepicNA (583)	MonteNL (303)
4	Bebidas	9,225	461.3	12.7 %	ToluEMX (598)	QroQRO (326)
5	Confitería	9,113	455.7	12.5 %	DuraDGO (573)	ZapoJAL (300)
6	Automedicación	8,964	448.2	12.3 %	ZacaZA (587)	TlaxTL (301)
7	Bebidas alcohólicas	8,642	432.1	11.9 %	NezaEMX (596)	TlaxTL (309)
8	Frutas y verduras	8,619	431.0	11.9 %	ColimaCOL (574)	SilaoGTO (307)

Tabla 2. Clasificación de categorías: total, promedio por sucursal, porcentaje y sucursales con más y menos venta.

Las ocho categorías quedan entre 11.9 % y 13.1 % del total, con promedios de 431.0 a 474.9 unidades por sucursal. Las columnas de máximo y mínimo muestran la variación real: en cada categoría, la mejor sucursal vende casi el doble que la peor.

Mapa de calor: unidades vendidas por sucursal y categoría

Sucursales ordenadas por venta total. Cuanto más oscuro, más unidades.

	Enlatado	Botanas	Confitería	Harinas	Frutas y verduras	Bebidas	Bebidas alcohólicas	Automedicación	Total
GuadaJAL	517	567	448	546	500	544	428	477	4,027
TijuBJC	440	562	562	490	521	595	375	459	4,004
CancQR	547	543	532	524	533	357	536	337	3,909
ColimaCOL	467	361	564	569	574	557	415	383	3,890
ToluEMX	327	442	525	470	459	598	480	545	3,846
LeonGTO	518	420	501	389	391	551	418	575	3,763
ZacaZA	577	430	318	484	310	469	557	587	3,732
ZacaMI	303	469	490	480	568	442	408	571	3,731
GroOA	406	546	458	391	355	493	583	489	3,721
NezaEMX	403	545	537	501	333	370	596	388	3,673
JuareCHIH	588	426	432	462	350	476	423	442	3,599
SilaoGTO	558	570	346	579	307	387	350	471	3,568
DuraDGO	335	304	573	491	392	557	327	577	3,556
QroQRO	589	372	494	512	495	326	316	423	3,527
VallaJAL	563	370	363	321	449	542	453	387	3,448
PuebPUE	462	510	370	395	550	375	376	394	3,432
TepicNA	536	583	363	318	363	421	423	390	3,397
TlaxTL	565	458	513	413	424	375	309	301	3,358
MonteNL	438	303	424	590	356	447	417	373	3,348
ZapoJAL	359	459	300	395	389	343	452	395	3,092

300 unidades 598 unidades Máximo de la categoría Mínimo de la categoría

Figura 2. Mapa de calor por sucursal y categoría (más oscuro, mayor venta).

Ningún renglón ni columna es uniforme: Enlatado va de 303 en ZacaMI a 589 en QroQRO, y Frutas y verduras de 307 en SilaoGTO a 574 en ColimaCOL. Las celdas muy claras, como Confitería en ZapoJAL (300) o Botanas en MonteNL (303), son los puntos que conviene revisar.

Unidades vendidas por sucursal, enero a abril de 2026

Suma de las ocho categorías.

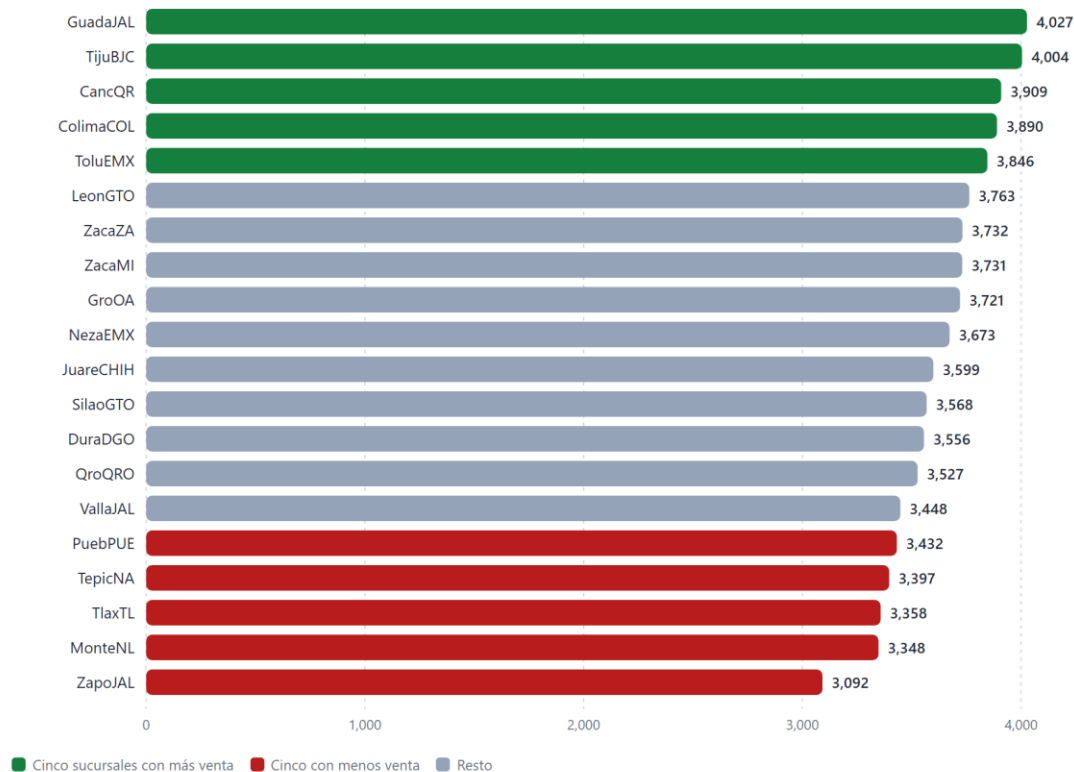


Figura 3. Total por sucursal (verde: las cinco con más venta; rojo: las cinco con menos).

GuadaJAL (4,027), TijuBJC (4,004), CancQR, ColimaCOL y ToluEMX encabezan la lista; PuebPUE, TepicNA, TlaxTL, MonteNL (3,348) y ZapoJAL (3,092) la cierran. La distancia entre la primera y la última (935 unidades) supera la que hay entre categorías (879): en el total pesa más la sucursal.

Vista 3D: unidades vendidas por sucursal y categoría

Sucursales de izquierda a derecha, de mayor a menor venta total: categorías en profundidad. Altura = unidades.

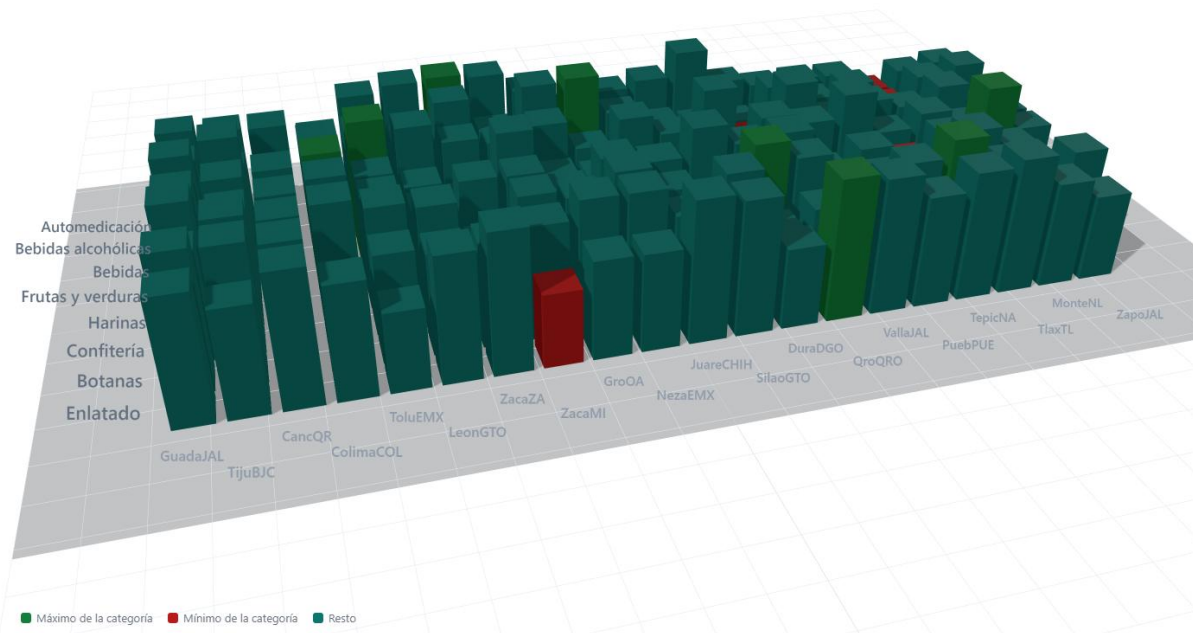


Figura 4. Vista tridimensional por sucursal y categoría, como apoyo visual.

Los picos, como Bebidas en ToluEMX (598), y los valles, como Confitería en ZapoJAL (300), están repartidos por todo el mapa. Esta vista sirve para presentar el panorama; las decisiones salen de las tablas 1 y 2.

Análisis de resultados

Enlatado es la categoría que más vende (9,498 unidades, 13.1 %) y Frutas y verduras la que menos (8,619, 11.9 %). Con solo 879 unidades de diferencia, ninguna categoría sobra. Lo que sí varía es cada categoría por sucursal (de unas 300 a unas 600 unidades); por eso la inversión debe **decidirse por sucursal y categoría**. Se recomienda **reforzar Enlatado, Harinas y Botanas** donde ya rotan bien. También conviene **revisar la cadena de frío y la merma de Frutas y verduras** en SilaoGTO (307), ZacaZA (310) y NezaEMX (333). Por último, se sugiere estudiar por qué GuadaJAL y TijuBJC venden más y ZapoJAL menos.

Conclusiones

El siguiente paso es cruzar las unidades vendidas con el precio, el margen y la merma de cada categoría, porque vender más no siempre es ganar más.

Se recomienda al Departamento de Ventas incorporar al almacén el precio de venta y el costo de cada producto, ya que son datos que hoy el punto de venta no entrega.

Una ventaja es que, como los datos ya viven en el Data Warehouse, el reporte puede repetirse cada trimestre y alimentar un tablero para el gerente de forma automática.

2. La importancia de asegurar la calidad de la información

Un reporte vale lo que valen sus datos. Con estas tablas y gráficas, el gerente de Abarrotes Lugo decidirá en qué categorías invertir. Si una sucursal reportó mal una categoría, o si dos llamaron distinto al mismo producto, el orden cambia y la inversión se va al lugar equivocado. Entre la mejor y la peor categoría hay solo 879 unidades, y entre las dos últimas (Bebidas alcohólicas, 8,642, y Frutas y verduras, 8,619) apenas 23. Un error de captura de ese tamaño en una sola sucursal cambiaría cuál aparece como la que menos vende. Por eso la Lección 7 señala que el Data Mining se aplica para remover los datos erróneos o irrelevantes (UVEG, s. f.-e). La Lección 8 lo resume como la extracción de datos de calidad (UVEG, s. f.-f).

Entendí la importancia de asegurar la calidad de la información y su flujo, en la clínica donde trabajo en el área de estadística. Dirección me encomendó investigar por qué unos indicadores (Como el DM 01 de diabetes, que estaba en 7.31 % contra una meta de 14.3 %) no subían y me proporcionaron acceso a múltiples fuentes de datos (bases de datos del SIAIS en SQL Server, diversos reportes en Excel como CP02, CP03 e IN26, RLC (Red Local de Consulta), MIMM 2018-2025, catálogos de claves UP, hojas RAIS y cuatro manuales de Estadística); algunas ya las tenía. Y tras semanas de investigación y pruebas se concluyó que el procedimiento que generaba el indicador no contaba claves, sino el valor de glucosa en servicios preventivos, por lo que unas 40 mil claves se grababan sin contar. Presenté los resultados como aquí (Sección 5, Figuras 9 a 11), con tablas, gráficos, videos, página web y recomendaciones para la jefatura clínica y a nivel central. Nadie decidía mal a propósito; el problema era estructural, de la calidad del dato que llegaba al cálculo global.

También como dice la Lección 7, junto con la calidad, hay que asegurar la privacidad (UVEG, s. f.-e). Son siete pasos: identificar y clasificar los datos, cuantificar su valor, identificar su vulnerabilidad y sus costos de protección, seleccionar medidas de seguridad y evaluar su efectividad. En este caso validar que cada sucursal tenga sus ocho categorías completas y proteger la información comercial. Por lo que podemos afirmar que, tanto en Abarrotes Lugo como en cualquier otra industria, un Data Warehouse solo sirve si quien lo consulta puede confiar en sus datos para decidir.

3. Qué ocurriría si no se aseguran los datos

Si BI-Housing no usara las herramientas de Data Management para el Data Mining, estaría expuesto a estos riesgos:

1. Datos duplicados y dobles gastos. Extraer de cada fuente por separado, en vez de concentrar todo en el Data Warehouse, repite datos y duplica el gasto de tiempo y dinero (UVEG, s. f.-e). Una sucursal contada dos veces inflaría sus ocho categorías y movería el orden final. (En el caso de mi trabajo, sucede cuando se capturan las acciones médicas y preventivas en hojas clínicas físicas llamadas RAIS y a la vez en el sistema digital SIMF, también valido y limpio esos duplicados.)
2. Registros erróneos que nadie remueve. Sin la limpieza del Data Mining, un cero de más o una venta en la categoría equivocada se sumaría al total, ensuciando los datos. (En nuestra oficina esto se hace corriendo búsquedas diarias para detectar continuamente

datos erróneos o antiguos, que pueden ser, por ejemplo, porque un doctor puso una clave mal o desactualizada, o por error del sistema, entre otros; esto se llama monitoreo y genera reportes como el OC34).

3. Categorías que no se suman bien. Si cada sucursal escribe “Botanas”, “botana” o “Snacks”, la consulta las trataría como tres categorías distintas. Esto lo experimenté al hacer diversas aplicaciones RPA para ponernos al corriente con el caso anterior, donde tenía que coincidir exactamente el nombre del encabezado (NSS, matrícula, etc.); lo solucioné tomando las primeras letras e identificando y clasificando por contenido (por ejemplo, un NSS tiene 10 dígitos, no puede ser fecha; un nombre, solo letras).
4. Decisiones de inversión equivocadas. Como hay diferencia tan pequeña entre categorías, cualquiera de estos errores marcaría como “peor” a una que vende bien.
5. Pérdida de privacidad y credibilidad. Sin clasificar ni proteger los datos, la información de ventas y clientes queda expuesta, y una cifra que no cuadre hace que el gerente desconfíe de todo el reporte aunque no sea el caso.

4. Reflexión

a) ¿Sabías la importancia de asegurar la calidad de datos de cualquier proceso?

Sí, aunque lo aprendí en la práctica antes que en la teoría. De hecho en mi trabajo, en estadística médica, básicamente es asegurar la integridad de la información generada en toda la clínica así como en campañas externas en empresas, escuelas y organizaciones, mediante distintas herramientas, procedimientos y manuales. A veces la información no cuadra e incluso hay indicadores que no sabíamos que medían. Personalmente al crear la herramienta de automatización de captura, le puse verificación en cada paso y no hubo errores. Lo que la lección me aportó fue el orden de los siete pasos para asegurar datos y la idea de que la calidad se cuida desde la extracción, no al final.

b) ¿Conocías las herramientas de Data Management antes de completar esta lección?

Parcialmente. Conocía las herramientas de consulta, porque uso SQL, y, por el reto anterior, las de tableros como Power BI y Metabase. No conocía la clasificación en tres grupos de la Lección 8 (UVEG, s. f.-f). Me faltaba la diferencia entre EIS, que monitorea variables, y DSS, que orienta a tomar decisiones, y no sabía que las herramientas para Data Mining son un grupo propio. Tampoco había relacionado las técnicas de la Lección 7 (UVEG, s. f.-e), muestreo estadístico, análisis de grupos y redes neuronales, con este reporte. Ahora sé que un tablero es un EIS, que la limpieza que hago con Python es Data Mining y que agrupar ventas por sucursal y categoría es un análisis de grupos. En nuestro trabajo usamos principalmente sistemas propietarios del IMSS altamente personalizados y de código cerrado (SIMF, SIAIS, SIMO, etc.). Pero también usamos servicios en la nube y, claro, los conceptos que se vieron en este curso, como el Data Warehouse y el Data Mart, entre otros.

c) ¿Conoces algún otro método para el desarrollo de un reporte en un ambiente empresarial?

Sí. En el módulo anterior vi la metodología CRISP-DM (Chapman et al., 2000), que va desde entender el negocio hasta desplegar el resultado. Otro método es el proceso de seis fases del certificado de Google Data Analytics: preguntar, preparar, procesar, analizar, compartir y actuar (Google, s. f.). También conozco los tableros en tiempo real con Tableau, Data Studio de Google, Power BI o Metabase, y los reportes que se regeneran solos con Python y su librería pandas. Para todos sirve la estructura de introducción, metodología, resultados, análisis y conclusiones. Como menciona la Lección 8, un reporte influye en las decisiones de la empresa, y definir quién lo lee y qué contiene importa tanto como su presentación (UVEG, s. f.-f). Esto puede ser directo en la nube, o en mi caso, diariamente en LAN interna con servidor cerrado, sin acceso desde afuera, por seguridad, y solo ciertos indicadores sincronizados con Excel ya sea mediante Google Sheets, Azure y otros proveedores de terceros. Reportes semanales, mensuales, trimestrales, semestrales y anuales. Nos basamos mucho en el Manual metodológico de indicadores médicos (Instituto Mexicano del Seguro Social, 2018) y en la CIE-10 (Organización Panamericana de la Salud, s. f.), además de otros recursos de la intranet institucional, documentos y manuales internos de Estadística.

5. Anexo: el reporte publicado y el almacén de datos en Azure

Para que el reporte no se quede en un documento, lo llevé a dos servicios de Azure. Primero, los datos de las 20 sucursales se cargaron en un Data Warehouse real: una base de datos Azure SQL Database, sin costo, con un esquema en estrella: una tabla central con las 160 combinaciones de sucursal y categoría, ligada a los catálogos de categoría y de sucursal. El ranking de la Tabla 2 sale de una sola consulta SQL sobre esa base, y devuelve exactamente los mismos totales (Figuras 5 a 7).

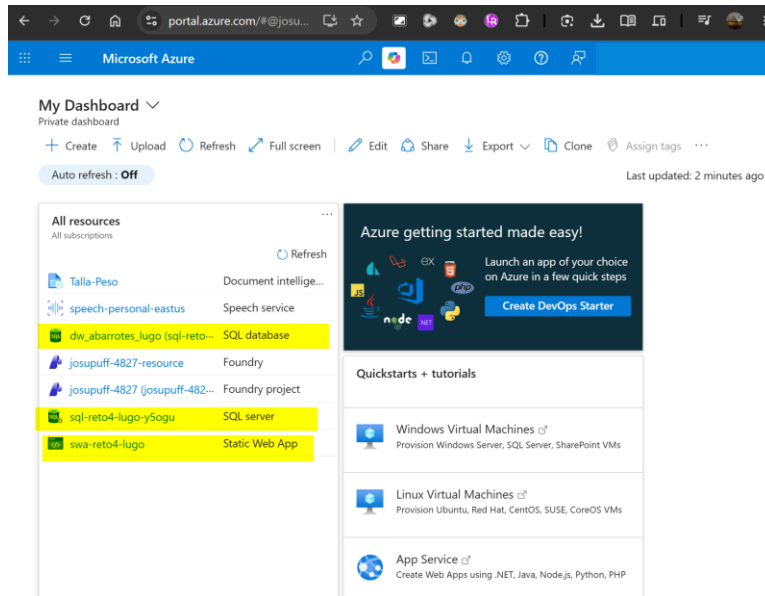


Figura 5. Recursos creados en Azure para el reporte: la base de datos `dw_abarotes_lugo`, su servidor SQL y la Static Web App `swa-reto4-lugo`.

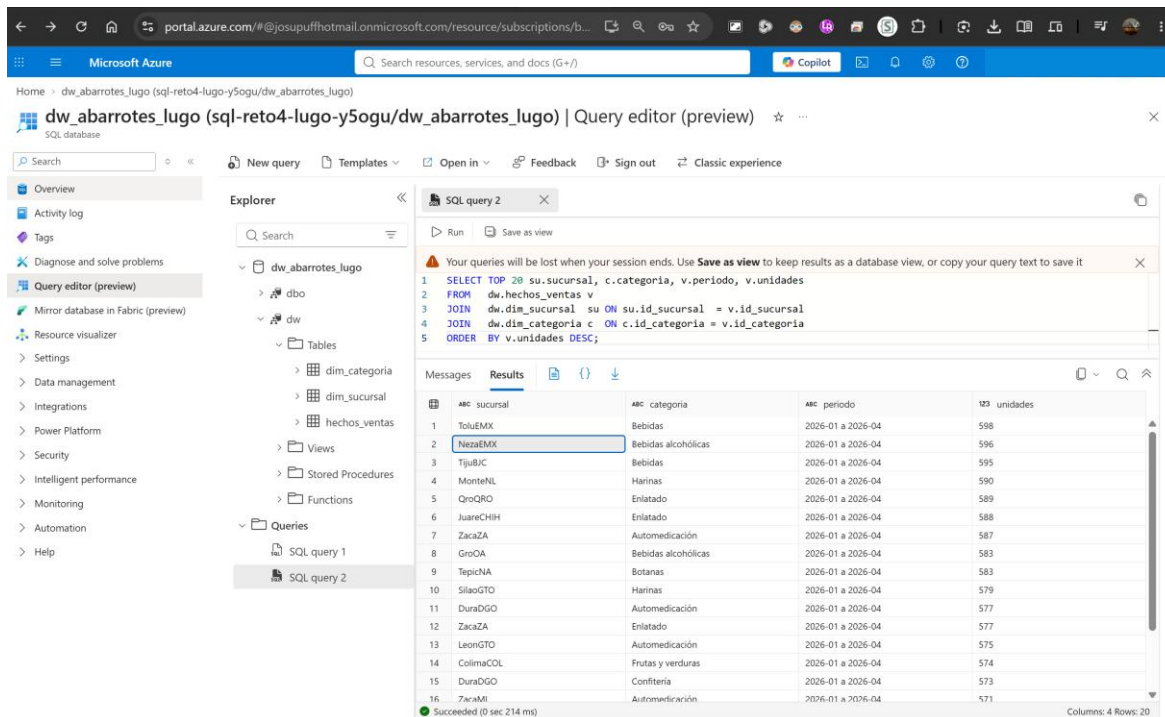


Figura 6. Esquema en estrella en el editor de consultas: tablas `dim_categoria`, `dim_sucursal` y `hechos_ventas`, con una consulta de detalle por sucursal y categoría.

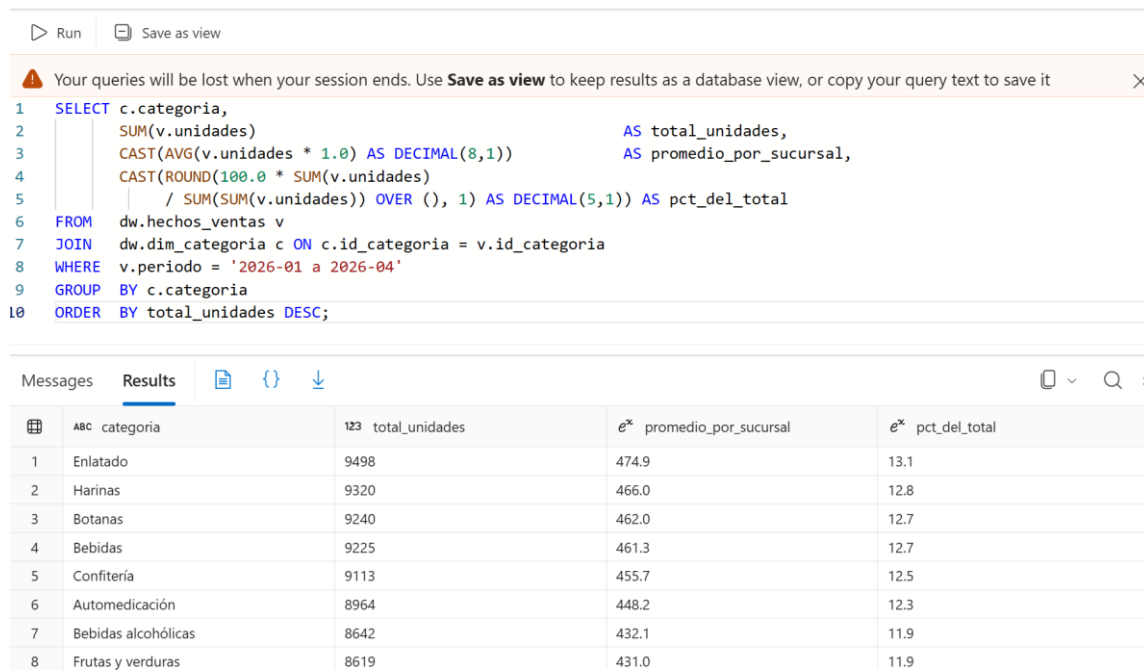


Figura 7. Consulta de ranking ejecutada en Azure SQL Database y su resultado.

Arriba está la consulta: agrupa las unidades por categoría, calcula el promedio por sucursal y el porcentaje del total, y ordena de mayor a menor. Abajo, la respuesta del servidor: Enlatado arriba con 9,498 unidades y Frutas y verduras abajo con 8,619, igual que en el reporte. Esto es lo que la Lección 7 describe como concentrar todas las fuentes en un solo lugar para hacer el Data Mining una vez y no en cada sucursal.

Segundo, publiqué el reporte como sitio interactivo en la siguiente página web con certificados SSL:

<https://reto4.mission-deck.cc>

Tiene los indicadores clave, un breve video explicativo, la gráfica, tabla, una vista en tres dimensiones que se puede girar, las tres recomendaciones y la consulta SQL que produce los números (Figura 8).

En futuros reportes también se le podría agregar autenticación y seguridad con perfiles y contraseñas; integración con servicios de IA, como visión artificial para digitalización automática de documentos, tickets o facturas, y sistemas LLM con arquitecturas RAG para que los usuarios interactúen con el reporte, manuales y bases internas mediante chatbots (algo ya implementado

en muchas empresas); e incluso migrar las propias bases de datos locales conectadas para una mayor disponibilidad, y así eliminar el riesgo de pérdida de información con respaldo de datos y VMs de mayor capacidad. Otra opción muy buena también es el servicio AWS de Amazon.



Figura 8. Portada del sitio público del reporte en Azure Static Web Apps.

Con esto, el gerente no depende de que alguien le mande el archivo: entra al sitio, ve el resumen en un minuto y, si quiere el detalle, descarga el PDF. Y como los datos viven en el almacén, repetir el reporte el próximo trimestre es volver a correr la misma consulta.

Este mismo proceso (almacén de datos en Azure, diseño digital y presentación de resultados) es el que he aplicado el último año en la clínica para videos educativos dirigidos a pacientes, instructivos para el personal y la presentación de investigaciones internas, como la del indicador DM 01 descrita en la sección 2 (Figuras 9 a 11).

Dos ejemplos públicos: el canal Vive Bien con Diabetes en YouTube, donde publico los videos educativos del departamento de Diabetes y el demo del Turnero de Medicina Preventiva, una aplicación en Azure App Service con acceso por contraseña (clave de prueba: 2026).

<https://www.youtube.com/@ViveBienconDiabetes>

<https://turnero-preventiva-demo.azurewebsites.net/entrar> pass: 2026

Referencias

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.

Google. (s. f.). *Google Data Analytics Professional Certificate* [programa de certificación en línea]. Coursera. <https://www.coursera.org/professional-certificates/google-data-analytics>

Instituto Mexicano del Seguro Social. (2018). *Manual metodológico de indicadores médicos 2018*. Dirección de Prestaciones Médicas. <https://www.imss.gob.mx/sites/all/statics/profesionalesSalud/investigacionSalud/normatividadInst/manualmetodologico2018.pdf>

Organización Panamericana de la Salud. (s. f.). *Clasificación estadística internacional de enfermedades y problemas relacionados con la salud: CIE-10* (10.^a rev., Vol. 1: Lista tabular). <https://ais.paho.org/classifications/chapters/pdf/volume1.pdf>

Universidad Virtual del Estado de Guanajuato. (s. f.-e). *Lección 7. Asegurar la calidad de la información* [material del módulo Administración de Big Data]. <https://view.genially.com/606dd6c4660bea0d73aebdab>

Universidad Virtual del Estado de Guanajuato. (s. f.-f). *Lección 8. Diseño y generación de reportes* [material del módulo Administración de Big Data]. <https://view.genially.com/606dd7dd660bea0d73aebdc1>