Proyecto de Ciencia de Datos: "¿Qué Deberíamos Hacer Sobre la Reducción de Nuestros Ingresos?"

Por Josue Medina

Fecha: Mayo 2025

Audiencia: Reclutadores de Analistas de Datos. NO es necesario conocimiento técnico

Sobre Este Proyecto

Este proyecto fue desarrollado en mayo de 2025 para demostrar mis habilidades como analista de datos. Está dirigido tanto a reclutadores técnicos como no técnicos. En él, abordo un escenario empresarial, usando una empresa ficticia llamada "Electronics and Office Things" (Electrónicos y cosas de oficina).
El proyecto comienza con un planteamiento del problema y sigue un proceso que simula un flujo de trabajo realista como analista de datos. Este documento está pensado para ser revisado junto con los archivos del proyecto. Para los lectores no técnicos, he incluido un apéndice para explicar términos y metodologías poco familiares.

¿Qué Encontrarás en Este Documento?

1. Resumen Ejecutivo: Presento una visión general del problema empresarial y resumo el enfoque que seguí para abordarlo.

2. Sobre Electronics and Office Things: Contexto breve sobre lo que hace la empresa y los desafíos que enfrenta.

3. Hallazgos Clave: Un resumen de los principales resultados obtenidos durante el análisis y las soluciones propuestas para abordar el problema.

4. Contenido de Este Repositorio de GitHub: Una descripción de los archivos y cómo cada uno contribuye al proyecto.

5.Planteamiento del problema: Primer contacto con mis supervisores y asignación de tareas

6. PASO 1 Entendiendo Nuestros Datos: Explico cómo utilicé SQL y un Jupyter Notebook para explorar la base de datos y obtener conocimientos iniciales, incluso sin un esquema de base de datos.

7. PASO 2 Resolución del problema: Describo cómo apliqué técnicas de análisis de datos utilizando la metodología CRISP-DM para investigar y comenzar a resolver el problema empresarial.

8. PASO 3 Discusión Interna con Supervisores y Creación del Panel: Después de revisar mi análisis con los supervisores, creé un panel en Power BI para comunicar los hallazgos a las partes interesadas.

9. PASO 4 Reunión: Una presentación grabada preparada para las partes interesadas, incluidos miembros de los equipos de Ventas y Marketing, explicando el problema, los hallazgos y las soluciones propuestas.

10. Apéndice: Conceptos técnicos como EDA y CRISP-DM, referencias utilizadas a lo largo del proyecto.

Resumen Ejecutivo

Planteamiento del Problema

La empresa ha experimentado una caída en los ingresos. Las preguntas clave son:

- *¿Esta reducción de ingresos afecta nuestras ganancias generales?*
- *Si es así, ¿qué podemos hacer para abordar el problema?*
- *¿Un préstamo podría resolver la situación o existen mejores alternativas?*

Enfoque

Para abordar este problema, comencé consultando los datos datos de la empresa para obtener información que respondiera a las preocupaciones de las partes interesadas. Utilicé SQL[1] para explorar los datos y creé un Notebook inicial en Jupyter (python)[2] para un Análisis Exploratorio de Datos (EDA)[3].

Durante este proceso, encontré inconsistencias en los datos[4] y las informé al equipo de Ingeniería de Datos. Después de que resolvieron los problemas y actualizaron la base de datos[5], reanudé el análisis.

Informé a mi gerente que presentaría los resultados utilizando Power BI[6] y le pregunté si prefería una solución por lotes[7] o en tiempo real[8]. Ella respondió: "Dado que aún estamos resolviendo problemas de arquitectura de base de datos[9], prefiero una solución por lotes."

Luego, desarrollé un segundo Notebook en Jupyter (python), siguiendo la metodología CRISP-DM[10], para guiar el proceso de resolución del problema. Tras completar mi análisis y validar los resultados, compartí la solución con la Gerente de Análisis y Tecnología. Tras su aprobación, creé un panel en Power BI para comunicar los hallazgos.

También grabé una presentación en video para el equipo de Marketing, explicando los resultados. Luego de recibir confirmación y apoyo por correo electrónico tanto del equipo de Marketing como del de Análisis, desarrollé un plan de implementación para colaborar con el equipo de Inteligencia Artificial en los siguientes pasos.

La fecha establecida para este escenario es el 3 de agosto de 2020.

Sobre Electronics and Office Things

Electronics and Office Things fue fundada en 2010 con el objetivo de proporcionar suministros de oficina y productos electrónicos de alta calidad a organizaciones pequeñas, medianas y grandes. La empresa comenzó sus operaciones en

Mumbai, India, y gradualmente expandió su alcance por todo el país. En 2016, ingresó a mercados internacionales, comenzando por Francia y Estados Unidos. En 2017, la empresa comenzó a migrar sus datos a infraestructura moderna, adoptando plataformas en la nube y bases de datos estructuradas. A partir de 2020, la empresa continúa con su misión de ingresar a nuevos mercados, incluso a costa de pérdidas temporales en sus ganancias. Otro factor que contribuye a la disminución de sus ingresos es el aumento de la competencia con otras empresas. Para evitar perder clientes, se vieron obligados a ofrecer productos a precios extremadamente bajos, incluso si eso significaba incurrir en pérdidas. Aunque Electronics and Office Things no fabrica sus propios productos, actúa como un distribuidor de confianza, ofreciendo productos de otros fabricantes consolidados. Con el tiempo, la empresa ha construido relaciones sólidas y duraderas con sus clientes, quienes confían en ella por su calidad y servicio. Una caída sostenida en los ingresos durante los últimos años ha generado preocupación; de ahí se origina el problema empresarial.

Hallazgos Clave

• Un aumento en los ingresos no necesariamente conlleva un aumento en las ganancias. Por lo tanto, adquirir un préstamo por sí solo no resolverá el problema.
• Una disminución en los ingresos no daña directamente al negocio, pero una disminución en las ganancias sí lo hace. Nuestro enfoque debe estar en aumentar la ganancia promedio, no solo los ingresos.
• Nuestros principales clientes (por ingresos, ganancias y frecuencia) no siempre son los más rentables en términos de ganancia por unidad de ingreso.
• Necesitamos evaluar si la pérdida de ganancias por la expansión del mercado es justificable. Para ello, utilizaremos dos herramientas:

- Métrica de Ganancia por Unidad de Ingreso (PUR) [11] por su siglas en inglés
- Algoritmo de árbol de regresión basado en IA[12] que predice el margen de ganancia

• Para mejorar la rentabilidad, debemos enfocarnos en la métrica "ganancia por unidad de ingreso" como herramienta para la toma de decisiones.
• Establecimos los siguientes indicadores clave de rendimiento (KPIs) para monitorear el éxito:

- Pérdida de ganancias por número de nuevos clientes
- Ganancia por Unidad de Ingreso (PUR)
- Tasa de transacciones con ganancia negativa y positiva

Contenido de Este Repositorio de GitHub

Este repositorio[13] contiene todos los materiales clave utilizados a lo largo del

proyecto. Puedes seguir estos archivos paralelo a la lectura de esta guía, para comprender todo el proceso de ciencia de datos y la solución empresarial.

1. **Introducción Leer Primero.pdf**
   Un informe completo y estructurado del proyecto en formato PDF. Diseñado para ser leído junto con el resto de los materiales.
2. **Exploración Base de Datos.sql**
   Un script SQL utilizado para explorar la base de datos de la empresa. Este paso ayudó a identificar datos útiles y evaluar qué podía aprovecharse para resolver el problema empresarial.
3. **Analisis Inicial.ipynb**
   Un Notebook de Jupyter que contiene el primer análisis exploratorio de datos (EDA). Durante esta fase, se identificaron problemas de calidad de datos, se pausó el análisis para notificar al equipo de Ingeniería de Datos.
4. **CRISP_DM.ipynb**
   Un Notebook de Jupyter que sigue la metodología CRISP-DM. Incluye análisis de datos, modelado y la preparación de un boceto del panel para comunicar los resultados.
5. **Panel Solución Caida de Ingresos.pbix**
   Un archivo de Power BI utilizado para presentar la solución final a las partes interesadas. Comunica visualmente los hallazgos y las recomendaciones.
6. **Video Presentación.mp4**
   Una presentación en video grabada en la que explico los hallazgos y la solución propuesta a los equipos de marketing y análisis.
7. **Archivos del conjunto de datos[14]**
   Archivos SQL y CSV utilizados a lo largo del proyecto.

En la siguiente sección de este documento, encontrarás conversaciones por correo electrónico relacionadas con el proyecto, junto con una explicación más detallada del propósito y contexto de cada archivo. Recuerda que este proyecto fue realizado en mayo de 2025, pero la fecha establecida para este escenario es agosto de 2020.

Planteamiento del problema

De	**Laura, Representante de Ventas**
Para	**Josue Medina, Analista de Datos**
Fecha	**3 de Agosto, 2020**
Asunto	**Problema de Ingresos**

Hola Josue,
Espero que te encuentres bien. Me pongo en contacto contigo porque hemos notado una caída en los ingresos y necesitamos tu apoyo para entender mejor qué está ocurriendo. La líder del equipo de análisis, Carla, está actualmente ocupada con otros proyectos, así que tú estarás a cargo de este análisis. Una vez que tu trabajo esté completo, Carla revisará tu solución y será la responsable de aprobarla y realizar las correcciones técnicas o matemáticas necesarias antes de presentarla a las partes interesadas.
Tu audiencia está únicamente interesada en India.
En resumen, queremos entender cómo la reciente disminución en los ingresos podría estar afectando al negocio, si un préstamo podría ayudar o si existen mejores alternativas. Me mantendré informada sobre tu progreso, pero tu informe final deberá ser enviado a Carla para su validación. Avísame si necesitas algo para comenzar.

Laura
Representante de Ventas
Electronics and Office Things

De	**Josue Medina, Analista de Datos**
Para	**Laura, Representante de Ventas**
Fecha	**3 de Agosto, 2020**
Asunto	**Urgente! Llaves de acceso y detalles del problema**

Hola Laura,
Gracias por asignarme al proyecto de análisis de ingresos. Antes de comenzar, me gustaría aclarar algunos puntos para asegurarme de que mi trabajo esté alineado con las expectativas.
¿Podrías indicarme quiénes son las partes interesadas (stakeholders) de este proyecto? También sería útil saber si la audiencia principal tiene conocimientos técnicos o si la presentación debe adaptarse a un público no técnico.
Adicionalmente, por favor confírmame la fecha esperada de entrega del análisis final.

Por último, necesitaré con urgencia las credenciales de acceso[15] a las bases de datos relevantes para comenzar a explorar los datos. ¿Podrías ayudarme a obtener los permisos necesarios o indicarme con quién debo contactar?
Quedo atento a tu respuesta.
Saludos cordiales,
Josué Medina

De	**Carla, Lider de Analistas**
Para	**Josue, Analista de Datos**
Fecha	**3 de Agosto de 2020**
Asunto	**Accseso a Base de datos**

Hola Josue,
Te he concedido acceso a la base de datos que necesitarás para el análisis de ingresos. Ahora deberías poder consultar los datos directamente usando tus credenciales. Avísame si tienes algún problema para acceder.

Un aviso importante: debido al trabajo en curso sobre la arquitectura de la base de datos, la documentación del esquema[16] no está disponible por el momento. Por ahora, tendrás que explorar manualmente la estructura de la tabla. Te recomiendo comenzar con una inspección básica de las columnas y algunas consultas de ejemplo[17] para familiarizarte con los datos. Además, por esta razón, no implementes una solución en streaming.

No dudes en escribirme si necesitas ayuda para interpretar alguno de los campos o si encuentras algún bloqueo de seguridad.

Saludos,
Carla
Analista Principal
Electronics and Office Things

De	**Laura, Representante de ventas**
Para	**Josue Medina, Analista de Datos**
Fecha	**3 de Agosto, 2020**
Asunto	**Resolviendo tus dudas**

Gracias por tu mensaje. Tus principales partes interesadas en este proyecto serán los equipos de Ventas y Marketing. Ten en cuenta que la audiencia no tiene formación técnica. Recuerda que no necesito un informe escrito, sino que prepares una reunión por Zoom con la solución.
Necesitaremos tu análisis completado para el viernes 7 de agosto, de modo que podamos revisarlo y prepararnos para las discusiones internas.

Laura
Representante de Ventas
Electronics and Office Things

PASO 1 Entendiendo Nuestros Datos

Para comenzar a abordar el problema, se crearon dos archivos. Después de leer la descripción de cada uno, puedes abrirlos si te interesa. Si no tienes conocimientos técnicos, no te preocupes, simplemente continúa leyendo este documento y espera la presentación en video más adelante.

Archivo número 2: Database_Exploration.sql
Creado para la Exploración de Datos el 3 de agosto de 2020 *(recuerda que esta es la fecha del escenario, no la fecha real)*
Este script de SQL[18] contiene el análisis exploratorio inicial de la base de datos de la empresa, un paso crítico debido a la falta de un esquema de base de datos. Incluye:
• Consultas para inspeccionar la estructura de la base de datos.
• Detección de la tabla de hechos[19] y las tablas de dimensiones asociadas[20], esenciales para construir una vista completa de los datos.
• Validación de tipos de datos e identificación de posibles relaciones.
• Extracción de estadísticas descriptivas básicas para evaluar la calidad y utilidad de cada tabla.
• Una investigación exhaustiva sobre qué tablas son más relevantes para responder al problema del negocio.

Esta exploración estableció que la tabla de transacciones sería central para el análisis.

Archivo número 3: Initial_Data_Analysis.ipynb
Creado para obtener Perspectivas Iniciales el 3 de agosto de 2020
Este Jupyter Notebook inicia el Análisis Exploratorio de Datos (EDA) utilizando los datos identificados en el script SQL. El descubrimiento clave durante esta fase fue que la tabla de transacciones contiene la mayor parte de la información necesaria para responder a las preguntas del negocio. La tabla incluye alrededor de 150,000 filas y los siguientes campos:
• sales_amount (Ingresos)
• profit_margin (Margén de ganancia)
• profit_margin_percentage (Margén de ganancia porcentual)
• order_date (Fecha de la venta)
• quantity (Cantidad)
• currency (Moneda)
• client (Cliente)
• cost_price (Costo)

La transacción más antigua registrada es del 4 de octubre de 2017.
Además de la tabla de hechos, el análisis incorporará tablas de dimensiones (dim tables), que proporcionan contexto descriptivo a los datos numéricos en la tabla de transacciones.

Sin embargo, el análisis fue pausado debido a varias inconsistencias en los datos, las cuales fueron compartidas con el equipo de Ingeniería de Datos:
• ¿Por qué el cliente 005 es el único que paga en USD?
• ¿Todos los campos monetarios están realmente en INR aunque el campo currency diga USD?
• Nombres de columnas que no coinciden: *market_code* (en tabla de hechos) vs. *markets_code* (en tabla de mercados).
• Error tipográfico en la tabla de clientes: *custmer_name* debería ser *customer_name*.
• ¿Por qué solo hay 4 ventas que no se realizaron en India?
• ¿Por qué no hay nombre del producto, solo el código del producto?

El trabajo continúa con un correo electrónico dirigido al equipo de ingeniería de datos.

DE	**Josue Medina**
Para	**Laura, Carla**
Fecha	**August 3, 2020**
Asunto	**Pausa debido a inconsistencias de datos – Progreso hasta ahora**
Estoy escribiendo para informarles que he tenido que pausar temporalmente mi análisis debido a algunos problemas relacionados con la arquitectura de datos que identifiqué en los datos de transacciones. El equipo de Ingeniería de Datos confirmó que están trabajando activamente en resolver estos problemas y me avisarán cuando sea seguro continuar. Mientras tanto, estoy compartiendo el trabajo que he completado hasta ahora, que incluye la exploración inicial de la base de datos y los primeros pasos del análisis exploratorio de datos. Adjunto a este correo los archivos para su referencia. Reanudaré el proyecto tan pronto como reciba la confirmación del equipo de ingeniería. Por favor, no duden en escribirme si tienen alguna pregunta o comentario mientras tanto. Archivos adjuntos: •Descripción de los archivos.pdf • Database_Exploration.sql • Initial_Data_Analysis.ipynb	

PASO 2: RESOLVIENDO EL PROBLEMA

CRISP_DM_Modeling.ipynb
Creado para encontrar una solución – 4 de agosto de 2020

Una vez que me informaron que los problemas de arquitectura de datos habían sido resueltos, retomé el proyecto utilizando un enfoque basado en el flujo de trabajo CRISP-DM. Creé el cuaderno CRISP_DM_Modeling.ipynb para analizar el conjunto de datos previamente descrito.
Desde el 4 hasta el 6 de agosto, compartí actualizaciones diarias del progreso con mis supervisores. Al finalizar este período, ya había completado el análisis e identificado dos posibles soluciones para abordar el problema del negocio.
Si tienes conocimientos técnicos, puedes revisar el archivo directamente. De lo contrario, a continuación encontrarás una descripción completa del trabajo y los resultados.
Nota: Dado que este documento está dirigido tanto a públicos técnicos como no técnicos, no abrumaré con cada detalle o hallazgo. En su lugar, me centraré en los hallazgos más relevantes y sus implicaciones para el negocio.

Hallazgos clave de este archivo

- Aumentar los ingresos no necesariamente lleva a un aumento en las ganancias.
- Para seguir enfrentando las pérdidas de rentabilidad, podemos apoyarnos en el indicador Ganancia por Unidad de Ingreso (PUR, por sus siglas en inglés) para identificar transacciones que ofrezcan mejores retornos y reducir gastos ineficientes.
- Implementé un modelo de árbol de decisión que predice la rentabilidad. Este modelo evalúa combinaciones de mercado, producto, cliente, zona, cantidad y precio de costo para determinar si una transacción probablemente será rentable. Funciona como una herramienta de apoyo para la toma de decisiones que ayuda al negocio a identificar qué oportunidades vale la pena perseguir.
- El análisis reveló que, aunque una caída en los ingresos no perjudica directamente al negocio, una caída en las ganancias sí lo hace. Por lo tanto, nuestro enfoque debe cambiar de aumentar ingresos a mejorar la rentabilidad.

Sección 1: Introducción

4 de agosto de 2020
Escribí una breve descripción general del proyecto y los problemas que quiero resolver.

Sección 2: Problema del Negocio

4 de agosto de 2020
En esta sección, analizo cómo han evolucionado las ganancias y los ingresos a lo largo del tiempo para comprender mejor el desafío del negocio.
Para visualizar estas tendencias, creé funciones personalizadas que grafican cualquier métrica seleccionada a lo largo de los años.

Sección 3: Comprensión de los Datos

4 de agosto de 2020
En esta sección, exploro los datos para entender su estructura y calidad. Uno la tabla de hechos con las tablas de dimensiones, limpio y estandarizo los datos, y verifico valores faltantes.
Además, creo funciones útiles para extraer información y generar visualizaciones iniciales que me ayuden a comprender mejor los patrones en los datos.

Sección 4: Preparación de los Datos

5 de agosto de 2020
En esta sección, desarrollo algunos hallazgos para construir un panel interactivo. Descubrí que el monto de ventas y el precio de costo no están fuertemente correlacionados con la ganancia, por lo que no puedo usarlos para segmentación. Luego intenté segmentar los datos en función de ganancias negativas versus positivas, pero noté que algunos datos no son representativos. Por lo tanto, en la siguiente sección implementaré un clasificador de árbol de decisión para realizar una segmentación adecuada.
Sin embargo, extraje algunas ideas de esta segmentación (ganancia positiva vs. negativa) para ayudar a explicar qué es un árbol de decisión al equipo no técnico. Adicionalmente, planeo construir un modelo de regresión para predecir el margen de ganancia, con el fin de apoyar la toma de decisiones.

Sección 5: Modelado

5 de agosto de 2020

Para construir el modelo de árbol de decisión, seguí un proceso paso a paso:

1. Preparación de los Datos: Primero, organicé y limpié los datos para asegurarme de que estuvieran listos para el análisis. Esto incluyó

seleccionar columnas relevantes como ganancia, producto, mercado, cliente y detalles de ventas.

2. Entrenamiento del Modelo: Usando los datos limpios, “entrené” el árbol de decisión. Esto significa que el modelo aprendió patrones de ventas pasadas, identificando qué factores tienden a conducir a ganancias positivas o negativas.
3. Prueba del Modelo: Después del entrenamiento, probé el modelo con una parte de los datos que no había visto antes. Esto ayuda a verificar qué tan preciso es el modelo al predecir ganancias.
4. Evaluación de Resultados: Revisé el rendimiento del modelo para asegurarme de que hiciera predicciones confiables. Si fue necesario, ajusté el modelo para mejorar su precisión.
5. Uso del Modelo: Finalmente, el árbol de decisión ahora puede clasificar futuras ventas como probablemente rentables o no, según los factores de entrada. Esto ayuda a orientar las decisiones sobre dónde enfocar esfuerzos y recursos.

Sección 6: De la lluvia de ideas a Power BI

5 de agosto de 2020

Después de completar el análisis de datos y construir el modelo de árbol de decisión, usaré todos estos hallazgos para diseñar un panel interactivo en Power BI.
En esta sección, hago un boceto de cómo se verá el panel.

Este panel reúne los hallazgos clave en un solo lugar, facilitando la comprensión de la situación actual y del impacto de los diferentes factores sobre la rentabilidad. Permite a los interesados explorar visualmente los datos, observar tendencias a lo largo del tiempo y comparar mercados, productos y clientes.

PASO 3 – Discusión Interna con Supervisores y Creación del Panel

Archivo: Revenue_Drop_Solution_Dashboard.pbix
Del 5 al 6 de agosto de 2020
Después de completar mi análisis, tuve una discusión interna por ZOOM con mis supervisores para alinear los hallazgos clave y decidir cómo comunicarlos de la mejor manera. Las conclusiones fueron revisadas y aprobadas, y se me pidió

crear un panel de Power BI dirigido a una audiencia no técnica, en preparación para la próxima reunión con los equipos de Ventas y Marketing.

Tengas o no conocimientos técnicos, puedes explorar el archivo de Power BI (.pbix) para. Los contenidos del panel son:

1. Introducción al Problema Empresarial
   El panel comienza con una explicación clara del desafío empresarial: la disminución de ingresos y beneficios a lo largo del tiempo. Los gráficos muestran cómo han evolucionado estas métricas, enmarcando las preguntas clave que necesitamos responder.

2. Resumen del Negocio
   Un resumen del desempeño general del negocio, respaldado por estadísticas descriptivas básicas, brinda contexto a los interesados antes de pasar a análisis más complejos.

3. Crecimiento y Retos con Nuevos Clientes
   Destaco que nuestro equipo ha hecho un excelente trabajo incorporando 19 nuevos clientes en los últimos tres años. Sin embargo, muchos de estos clientes compran con poca frecuencia, lo cual afecta la rentabilidad. Presento la métrica de Ganancia por Unidad de Ingreso (PUR), una de las dos soluciones que propongo para resolver el problema. También demuestro que nuestros mejores clientes en términos de ganancia o ingreso total no son necesariamente los más eficientes (en términos de PUR).
   Una caída en los ingresos no necesariamente perjudica las ganancias, pero una caída en las ganancias sí lo hace. También muestro cuántas transacciones con ganancia negativa hemos tenido que realizar para captar a estos nuevos clientes, agrupándolas por año.

4. Falsa Creencia sobre los Préstamos
   Usando los datos, dejo claro que tomar un préstamo no resolvería el problema, ya que tener ingresos altos no implica tener ganancias altas. De hecho, tenemos ventas que generan millones en ingresos mientras producen cientos de miles en pérdidas.

5. Modelado Predictivo para Planificación de Campañas
   Presento la segunda solución clave: un clasificador tipo árbol de decisión. Este modelo nos ayuda a predecir si una transacción potencial generará ganancia o pérdida, ofreciendo una valiosa guía para planificar futuras campañas de marketing y ventas.

6. Indicadores Clave y Próximos Pasos
   Finalmente, el panel presenta los principales KPIs (Indicadores Clave de Desempeño) y los valores objetivos que deben monitorearse después de la

reunión. También propongo acciones de seguimiento y próximos pasos basados en los hallazgos discutidos.

PASO 4 – Reunión

7 de agosto de 2020
Finalmente, tengo una reunión con mis partes interesadas. Puedes verla en el archivo Stakeholder_Presentation_Video.mp4.

Apéndice

1. **SQL (Lenguaje de Consulta Estructurado)** es un lenguaje usado para comunicarse con bases de datos. Nos ayuda a hacer preguntas como: "¿Cuántos productos se vendieron el año pasado?" o "¿Cuál fue la ganancia más alta?"
   Melton, J., & Simon, A. R. (2002). Understanding the New SQL: A Complete Guide. Morgan Kaufmann.

2. **Jupyter Notebook** es una herramienta que nos permite escribir y ejecutar código paso a paso. Es útil para el análisis de datos porque también puedes agregar explicaciones y gráficos junto al código.
   Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows.

3. **EDA (Análisis Exploratorio de Datos)** es el proceso de explorar y comprender los datos antes de tomar decisiones. Implica buscar patrones, errores o valores faltantes.
   Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

4. **Inconsistencias en los Datos**: son problemas en los datos, como formatos distintos, nombres incorrectos o valores no coincidentes, que pueden causar resultados erróneos.
   Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. Communications of the ACM, 41(2), 79–82.

5. **Base de Datos**: es como un archivador digital. Almacena mucha información que se puede organizar y acceder fácilmente.
   Date, C. J. (2004). An Introduction to Database Systems (8th ed.). Pearson Education.

6. **Power BI** es una herramienta que permite crear gráficos y paneles a partir de datos. Ayuda a las personas no técnicas a entender los datos mediante visualizaciones.
   Microsoft. (2020). Power BI documentation.

7. **Procesamiento por Lotes (Batch Processing)** se refiere a trabajar con datos procesándolos en grandes grupos o “lotes” en horarios programados, en lugar de analizarlos en tiempo real. Es eficiente cuando solo necesitas resultados de vez en cuando (por ejemplo, una vez al día o por semana). En este proyecto, se usó procesamiento por lotes para analizar los datos después de exportarlos de la base.
   Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113.

8. **Streaming (Transmisión de Datos)** significa procesar datos en tiempo real, a medida que llegan, como monitorear ventas mientras ocurren.
   Kreps, J., Narkhede, N., & Rao, J. (2011). Kafka: A Distributed Messaging System for Log Processing. NetDB.

9. **Problemas en la Arquitectura de Bases de Datos**: son problemas con la estructura o conexión de una base de datos, que pueden bloquear el análisis o causar errores de datos.
   Coronel, C., & Morris, S. (2016). Database Systems: Design, Implementation, & Management (12th ed.). Cengage Learning.

10. **CRISP-DM**: es un proceso estándar que guía los proyectos de análisis de datos en 6 pasos: Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelado, Evaluación y Despliegue.
    Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.

11. **Ganancia por Unidad de Ingreso (PUR)**: es una métrica empresarial calculada dividiendo la ganancia entre el ingreso. Muestra cuánta ganancia se obtiene por cada unidad de ingreso.
    Horngren, C. T., Datar, S. M., & Rajan, M. (2014). Cost Accounting: A Managerial Emphasis (15th ed.). Pearson.

12. **Algoritmo de Árbol de Regresión Basado en IA**: es un tipo de modelo de aprendizaje automático que predice un número (como ganancia) dividiendo los datos en grupos más pequeños según reglas.
    Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. Wadsworth.

13. **Repositorio de GitHub**: es una carpeta compartida en línea donde las personas almacenan y gestionan archivos de proyectos y código, con control de versiones.
    Chacon, S., & Straub, B. (2014). Pro Git (2nd ed.). Apress.

14. **Archivos de Conjunto de Datos (Datasets)**: son los datos usados para el análisis. Los obtuve de Codebasics.

Codebasics. (n.d.). Codebasics: AI Courses. Recuperado en mayo de 2025, de https://codebasics.io/

15. **Credenciales de Acceso**: nombres de usuario y contraseñas que permiten a alguien ingresar y usar una base de datos o sistema.
Pfleeger, C. P., & Pfleeger, S. L. (2006). Security in Computing (4th ed.). Prentice Hall.

16. **Esquema de Base de Datos**: el plano de una base de datos, que muestra cómo se conectan las tablas y qué tipo de datos contiene cada una.
Elmasri, R., & Navathe, S. B. (2015). Fundamentals of Database Systems (7th ed.). Pearson.

17. **Consultas a la Base de Datos**: preguntas escritas en SQL que le piden a la base de datos que devuelva información específica.
Oppel, A. J. (2009). Databases Demystified (2nd ed.). McGraw-Hill Education.

18. **Script SQL**: archivo que contiene una lista de comandos SQL para ejecutar varias consultas de una vez.
Pratt, P. J., & Last, M. Z. (2013). A Guide to SQL (9th ed.). Cengage Learning.

19. **Tabla de Hechos**: una tabla central en una base de datos que registra eventos de negocio (como ventas o transacciones) y contiene cifras (como ganancia o cantidad).
Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit (3rd ed.). Wiley.

20. **Tablas de Dimensión**: tablas que describen los hechos con más información, como nombres de clientes, detalles de productos o ubicaciones.
Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit.

21. **Función en Python**: un bloque reutilizable de código en Python que realiza una tarea. Las funciones ayudan a que el código sea más limpio y fácil de gestionar.
Lutz, M. (2013). Learning Python (5th ed.). O'Reilly Media.

22. **Unir Tablas (Merge)**: proceso de combinar dos o más tablas usando una columna en común, como vincular la tabla de clientes con sus ventas.
McKinney, W. (2017). Python for Data Analysis (2nd ed.). O'Reilly Media.

23. **Limpiar y Estandarizar Datos**: corregir errores en los datos y asegurarse de que todo esté en el mismo formato (por ejemplo, fechas, nombres, monedas) para garantizar un análisis preciso.

Dasu, T., & Johnson, T. (2003). Exploratory Data Mining and Data Cleaning. Wiley-Interscience.