Documentación

Laboratorio Sistemas Organizacionales y Gerenciales 2

Carlos Javier Martinez Polanco - 201709282

Kevin Josue Calderon Peraza - 201902714

Planificación

- **¿Cómo se dividieron las tareas entre los miembros del equipo?**

  El trabajo se estructuró en tres fases principales, para optimizar el tema del tiempo. En la primera fase, que correspondió a la infraestructura, un miembro del equipo se encargó de configurar todos los recursos necesarios, incluyendo la creación de entornos de desarrollo y la implementación de servicios en la nube, como AWS.

  En la segunda fase, que involucró el análisis y el código, se decidió que ambos miembros trabajaran de forma consultando y realizando el desarrollo correspondiente. La razón de esta colaboración fue que tanto el análisis de datos como la implementación del código. Ambos integrantes realizaron la parte analítica y desarrollo del código, asegurándose de que las soluciones implementadas fueran técnicamente viables y alineadas con los requisitos del proyecto.

  Finalmente, la documentación fue responsabilidad de un solo miembro del equipo, quien se encargó de detallar cada parte del proceso, asegurando que todos los pasos y decisiones tomadas estuvieran bien documentados, que también permitió tener una referencia clara de todo el trabajo realizado, lo que era esencial para la revisión y posible expansión del proyecto en el futuro.

- **¿Qué herramientas y tecnologías decidieron utilizar y por qué?**

  Seleccionamos una serie de herramientas y tecnologías con el objetivo de aprovechar nuestras fortalezas previas y facilitar el desarrollo del proyecto de manera eficiente. Python fue nuestra elección principal como lenguaje de programación debido a su versatilidad, sintaxis sencilla y su amplia disponibilidad de librerías y frameworks que facilitan tareas como el análisis de datos, el procesamiento de información y la implementación de soluciones.

  La elección de AWS (Amazon Web Services) para la infraestructura en la nube se basó en nuestra experiencia previa con esta plataforma. AWS ofrece una gran escalabilidad y flexibilidad, lo que nos permitió ajustar los recursos de manera rápida y económica. Además, su fiabilidad y la amplia gama de servicios disponibles (como instancias de computación, almacenamiento, bases de datos) facilitaron la creación de un entorno adecuado para desarrollar, probar y desplegar nuestras soluciones.

  Para el análisis de datos, utilizamos Jupyter Notebook como entorno de desarrollo interactivo, lo que nos permitió documentar y compartir el análisis de manera sencilla, facilitando la revisión y colaboración. Las librerías Pandas y Matplotlib fueron seleccionadas debido a su eficiencia y facilidad de uso en proyectos previos. Pandas es ideal para manejar grandes volúmenes de datos, realizar transformaciones y análisis complejos, mientras que Matplotlib nos permitió crear visualizaciones claras y efectivas para comunicar los resultados de nuestro análisis.

- **¿Cómo establecieron los plazos para cada fase del proyecto?**

  Al definir los plazos, tomamos en cuenta no solo la carga de trabajo individual de cada miembro del equipo, sino también la complejidad de las tareas en cada fase. Esto permitió asignar tiempos realistas para cada actividad, sin comprometer algún atraso. Además, consideramos que la flexibilidad era importante para adaptarnos a cambios y avances significativos que pudieran surgir en el camino.

Proceso de análisis

- **Describa el enfoque paso a paso que siguieron para limpiar y preparar los datos.**

  El proceso de limpieza y preparación de los datos es fundamental para asegurar que el análisis posterior sea preciso y confiable. A continuación, se describe cómo se llevó a cabo este proceso:

  - **Carga de datos:** En primer lugar, se cargó el conjunto de datos desde las fuentes disponibles. Esto implicó importar los archivos a un entorno de trabajo adecuado (en este caso, utilizando Python y las librerías necesarias como Pandas). Una vez cargados, los datos se almacenaron en una tabla temporal para facilitar su manipulación y revisión inicial. Esta tabla sirvió como un primer vistazo a los datos y permitió identificar posibles problemas de estructura o calidad.
  - **Limpieza de datos:** Después de la carga inicial, se procedió a la limpieza de los datos. Para esto, se aplicaron tres criterios principales:
    - Conversión de valores numéricos: En muchas ocasiones, los datos pueden venir en formatos incorrectos o inconsistentes, como cadenas de texto que deberían ser números. Para garantizar la integridad de los cálculos y análisis posteriores, se convirtió cada valor numérico a su tipo adecuado (por ejemplo, convertir cadenas con números en valores enteros o flotantes). Si alguna conversión fallaba, ese valor se marcaba como un error o se transformaba en un valor nulo (NaN).
    - Forzar errores a NaN: Durante la limpieza, se identificaron ciertos valores que eran evidentes errores en los datos (por ejemplo, texto o símbolos donde deberían haber existido valores numéricos). Para evitar que estos valores distorsionaran el análisis, se convirtieron en NaN (Not a Number), lo que los hacía fácilmente identificables y manejables más adelante.
    - Eliminación de valores inválidos: Además de forzar errores a NaN, se eliminaron valores completamente inválidos o que no cumplían con los requisitos del análisis. Esto incluyó la eliminación de filas que contenían datos fuera de rango o que no representaban una muestra válida dentro del contexto del proyecto.
    - Manejo de valores nulos: En muchos casos, los datos contenían valores nulos o ausentes. Para abordar esto, se tomó la decisión de asignar un

valor predeterminado (como 0 o el valor medio) a estos casos cuando fuera apropiado. Este enfoque permitió que el análisis continuara sin interrupciones, pero siempre se documentó que estos valores no eran observaciones originales y que eran imputaciones para evitar distorsiones.

- Inserción en las tablas definitivas: Una vez que los datos fueron limpiados y preparados, se insertaron en sus tablas respectivas. Esto implicó organizar los datos de acuerdo con las estructuras necesarias para el análisis, asegurándose de que cada tabla estuviera bien definida y sin inconsistencias. El paso final fue revisar que los datos estuvieran listos para ser utilizados en el análisis exploratorio y en las etapas posteriores del proyecto.

- **Explique las decisiones tomadas durante el análisis exploratorio de datos.**

Identificación de patrones y tendencias: Al comenzar con el analisis, el primer paso fue explorar visualmente los datos para detectar patrones y tendencias generales. Se utilizaron herramientas como Matplotlib **y** Seaborn para crear gráficos de dispersión, histogramas y diagramas de caja, con el objetivo de obtener una visión general de las distribuciones de los datos. Las decisiones clave en este punto fueron centrarse en las relaciones más relevantes entre las variables y excluir cualquier variable que no aportara valor significativo al análisis.

Revisión de la calidad de los datos: Una de las decisiones más importantes fue verificar la calidad de los datos antes de realizar un análisis más profundo. Esto implicó buscar outliers, valores atípicos, y cualquier inconsistencia que pudiera afectar las conclusiones. Los valores erróneos o inconsistentes fueron tratados adecuadamente mediante la conversión a NaN o la eliminación, como mencioné en la fase de limpieza.

Selección de variables relevantes: Durante el AED, se tomaron decisiones sobre qué variables serían las más relevantes para el análisis. En este sentido, se aplicaron filtros para seleccionar solo aquellas que aportaran información útil y eliminar variables que no tenían una relación directa con el objetivo del proyecto. Esto ayudó a reducir la complejidad del análisis y enfocarse en lo esencial.

- **Detalle los desafíos encontrados durante el análisis y cómo los superaron.**

  Valores sin sentido o fuera de lugar: Durante la limpieza de los datos, se descubrió que muchos valores no tenían sentido dentro del contexto del conjunto de datos, como valores extremadamente grandes o pequeños, que evidentemente eran errores. Por ejemplo, algunos registros mostraban fechas en el futuro o números imposibles (como salarios de 0 o de millones de dólares en un contexto donde los salarios eran más bajos). Estos valores fueron rápidamente identificados y eliminados o reemplazados por valores nulos (NaN) para evitar que afectaran los cálculos posteriores.

Metodología

- **Explique cómo seleccionaron las visualizaciones más apropiadas para sus hallazgos. Se seleccionaron**

  Se seleccionaron tres tipos de graficas para visualizar de mejor manera los resultados.

  - Gráfico de barras (Distribución de ventas por categoría de producto)

  Comparación clara: Permite comparar fácilmente las ventas entre diferentes categorías de productos (Ropa, Accesorios, Calzado).

  Visualización de tendencias: Muestra cuáles categorías tienen mayores ventas de manera rápida.

  Facilidad de interpretación: Es intuitivo y útil para comunicar datos a audiencias no técnicas.

  Inclusión de intervalos de confianza: Indica la variabilidad en los datos, ayudando en la toma de decisiones.

  - Gráfico de líneas (Patrones de compra por edad)

  Muestra tendencias a lo largo de una variable continua: En este caso, permite analizar cómo varían las compras en función de la edad.

  Identificación de patrones y anomalías: Se observa que el volumen de compras aumenta drásticamente después de los 50 años, lo cual es un insight valioso.

  Facilita la predicción de tendencias: Puede ser útil para modelos de predicción basados en datos históricos.

  Conexión fluida entre puntos de datos: Ayuda a visualizar cómo evoluciona el comportamiento de los consumidores a lo largo del tiempo o con el cambio de una variable.

- Gráfico de correlación

  Estas gráficas son esenciales en la toma de decisiones basada en datos, ya que ayudan a visualizar la fuerza y dirección de una relación, permitiendo anticipar comportamientos y validar hipótesis en diversas áreas, como finanzas, ciencia y negocios.

Respuestas a las preguntas planteadas

a. ¿Cómo podrían los insights obtenidos ayudar a diferenciarse de la competencia?

Respuesta1: Los insights podrían permitir a la empresa ofrecer productos que son altamente demandados por ciertos grupos demográficos, lo que la haría más competitiva.

Respuesta2: Analizar la distribución de ventas por categoría de producto ayuda a identificar qué áreas generan mayores ingresos y cuáles necesitan ajustes. Si una categoría como Ropa, por ejemplo, tiene mayores ventas, se pueden destinar más recursos a su promoción, mejorar su oferta o ampliar la variedad de productos.

b. ¿Qué decisiones estratégicas podrían tomarse basándose en este análisis para aumentar las ventas y la satisfacción del cliente?

Respuesta1: Aumentar la presencia en regiones con mayor demanda y ofrecer descuentos en productos menos vendidos.

Respuesta2: El análisis de los patrones de compra por edad también ofrece ventajas competitivas al permitir una segmentación más efectiva del mercado. Si los datos muestran que los clientes mayores de 50 años realizan más compras, se pueden diseñar campañas específicas para este grupo, ajustando la publicidad, los canales de venta y la experiencia de usuario para adaptarse mejor a sus necesidades.

c. ¿Cómo podría este análisis de datos ayudar a la empresa a ahorrar costos o mejorar la eficiencia operativa?

Respuesta1: Al identificar los productos menos rentables, la empresa podría optimizar su inventario y reducir los costos de almacenamiento.

Respuesta2: La optimización de inventarios, el diseño de campañas de marketing basadas en datos y la mejora de la experiencia del cliente pueden marcar una diferencia significativa en la rentabilidad del negocio. Así, las empresas que aprovechan estos análisis pueden posicionarse de manera más efectiva en el mercado y ofrecer un valor superior en comparación con aquellas que no utilizan un enfoque basado en datos.

d. ¿Qué datos adicionales recomendarían recopilar para obtener insights aún más valiosos en el futuro?

Respuesta1: Datos sobre las preferencias de los clientes, análisis de comportamiento en el sitio web y feedback sobre productos y servicios.

Respuesta2: Podría ser el historial de compras individual, lo que permitiría identificar patrones de recompra, preferencias específicas y niveles de lealtad del cliente, para promocionar descuentos, y ciertos beneficios que puede ayudar al cliente al volver.