escuela británica de artes creativas y tecnología

Fase 1 de trabajo.

Cargarás los datos en Python utilizando Pandas, realizarás la limpieza y transformación de los datos, y consolidarás la información en un conjunto de datos listo para el análisis.

Paso a paso:

Cargar las tablas de datos en Pandas:

- Utiliza la función de Pandas para cargar cada tabla de datos desde el archivo csv o xlsx (DIM_CATEGORY, DIM_PRODUCT, DIM_SEGMENT, DIM_CALENDAR, FACT_SALES) en un DataFrame.

Revisar y entender los datos cargados:

- Realiza una revisión rápida de las primeras filas de cada DataFrame para entender la estructura de los datos y sus principales características. Utiliza métodos como `head()`, `info()` y `describe()` para tener una idea general del contenido de cada tabla.

Realizar la limpieza de datos:

- Corrige posibles inconsistencias en los datos, como errores tipográficos o formatos incorrectos.
- Identifica y maneja valores nulos en cada DataFrame (si es que los hay)
- Identifica si hay duplicados de los DataFrames y en caso de que los haya, eliminalos.

Unir los DataFrames relevantes:

- Realiza uniones entre los DataFrames para consolidar la información. Por ejemplo, une la tabla de productos con las ventas, las categorías y los segmentos.

Aplicar transformaciones necesarias:

- Estandariza los formatos de las columnas, como las fechas o categorías, para asegurar la consistencia en todo el conjunto de datos.
- Realiza cualquier transformación adicional requerida para preparar los datos para el análisis, como la creación de nuevas columnas calculadas o la agrupación de datos.

Guardar el conjunto de datos consolidado:

- Guarda el DataFrame consolidado en un nuevo archivo CSV o Excel para su uso posterior en el análisis.

Archivo final

Un archivo Jupyter Notebook (.ipynb) con los scripts utilizados para cargar, limpiar, transformar y consolidar los datos.

Anexos Fase 1

DIM_CATEGORY

DIM_PRODUCT

DIM_SEGMENT

DIM_CALENDAR

FACT_SALES

Fase 2 de trabajo

Genera una serie de visualizaciones que te ayudarán a entender mejor la distribución de los datos y las relaciones entre diferentes variables, utilizando las tablas proporcionadas.

Paso a paso:

Carga los datos:

- Carga la tabla consolidada de datos que se realizó en el avance de proyecto con empresa aliada anterior (en Módulo 7: Pandas: Dataframes, lectura y exploración de archivos) así como también los datos de alguna otra tabla que necesites usar para el EDA o la visualización y no fueron posibles consolidar en un solo archivo.

Identificar las variables clave para el análisis:

- Determina qué variables son más relevantes para responder a las preguntas de negocio.

Visualizar la distribución de ventas:

- Crea gráficos de distribución (histogramas o boxplots) para visualizar la distribución de ventas por diferentes categorías. Esto ayudará a identificar patrones de ventas y posibles outliers.

Analizar la tendencia de ventas en el tiempo:

- Genera gráficos de líneas que muestren las tendencias de ventas a lo largo del tiempo. Filtra las ventas por diferentes productos, regiones o segmentos para analizar cómo han cambiado las ventas a lo largo del tiempo.

Explorar la relación entre diferentes variables:

- Utiliza gráficos de dispersión (scatter plots) para investigar relaciones entre variables clave. Esto puede ayudar a identificar correlaciones significativas entre diferentes variables.

Visualizar la distribución geográfica de las ventas:

- Si es posible, utiliza mapas o gráficos de barras apiladas para visualizar las ventas por región. Esto puede ayudar a identificar qué regiones tienen el mayor o menor rendimiento.

Identificar posibles outliers y anomalías:

- Utiliza gráficos de caja (boxplots) para identificar posibles outliers en los datos de ventas por producto, categoría o región. Identificar outliers puede ayudar a entender mejor los patrones inusuales o anómalos en los datos.

Documentar los insights obtenidos:

- Anota los principales hallazgos que observes en cada visualización. Por ejemplo, si ciertas categorías tienen un desempeño consistentemente alto o si ciertas regiones muestran un crecimiento o declive inesperado en las ventas.

Guardar las visualizaciones:

- Guarda todas las visualizaciones generadas en archivos gráficos (PNG, JPEG) para usarlos posteriormente en el dashboard o en la presentación final.

Archivo final

Un archivo Jupyter Notebook (.ipynb) con todas las visualizaciones generadas, explicaciones y comentarios claros para cada paso, y los insights obtenidos de cada gráfico.

Fase 3 de trabajo

Aplicarás técnicas de clustering para identificar segmentos clave en los que la marca Vanish tiene presencia, y podrás identificar áreas en las que mejorar su estrategia.

Paso a paso:

1. **Cargar y preparar los datos:**
    - Carga el conjunto de datos relevante en un DataFrame de Pandas.
    - Revisa los datos para identificar las columnas más relevantes que se utilizarán
2. **Seleccionar las características para el clustering:**
    - Elige las características clave que se utilizarán para segmentar los productos o regiones. Por ejemplo, puedes basarte en:
        - Ventas totales
        - Categoría de producto
        - Región geográfica
        - Atributos específicos de los productos
3. **Estandarizar las características:**
    - Utiliza un método de escalado, como `StandardScaler` de Scikit-learn, para estandarizar las características seleccionadas. Esto es crucial, ya que K-Means se ve afectado por las diferentes escalas de los datos. Asegúrate de que todas las variables estén en la misma escala para obtener mejores resultados.
4. **Aplicar el algoritmo K-Means:**
    - Elige un número inicial de clusters (por ejemplo, 3 o 4) y aplica el algoritmo K-Means usando Scikit-learn.
    - Realiza varias iteraciones con diferentes números de clusters para determinar cuál es el óptimo. Puedes utilizar el **método del codo** para evaluar qué número de clusters minimiza la suma de las distancias al centroide de cada cluster.
5. **Evaluar los resultados del clustering:**

- Revisa los resultados obtenidos y asigna cada producto o región a un cluster. Cada cluster debe representar un grupo de productos o regiones con características similares.
- Analiza las características de cada cluster y observa si hay patrones o segmentos que destacan en las ventas, categoría o región.

6. **Visualizar los clusters:**
   - Crea gráficos de dispersión utilizando Matplotlib o Seaborn para visualizar cómo los productos o regiones se agrupan en función de las características seleccionadas. Usa colores para diferenciar los clusters y añade etiquetas para identificar los productos o regiones clave dentro de cada uno.
   - Genera gráficos adicionales, como gráficos de barras o diagramas de caja, para explorar más a fondo los resultados del clustering.
7. **Interpretar y analizar los clusters:**
   - Documenta los insights clave de los resultados. Por ejemplo, ¿hay segmentos donde la marca Vanish tiene un fuerte desempeño? ¿Qué clusters muestran un rendimiento bajo o áreas de mejora?
   - Identifica si existen patrones en los productos o regiones que puedan influir en la estrategia comercial. Por ejemplo, si un cluster contiene productos con ventas bajas, ¿hay alguna característica en común que podría explicar ese desempeño?
8. **Guardar los resultados y el modelo:**
   - Guarda los resultados del clustering (incluyendo la asignación de clusters a productos o regiones) en un nuevo DataFrame o archivo CSV.

Archivo final

Un archivo Jupyter Notebook (.ipynb) con los scripts utilizados para realizar el clustering, gráficos que visualizan los clusters y un análisis detallado de los insights obtenidos.

Fase 4 de trabajo

Crearás una base de datos, definirás la estructura de las tablas necesarias, cargarás los datos proporcionados y realizarás consultas básicas y avanzadas para analizar la información, todo esto usando las tablas del proyecto de la empresa aliada.

Paso a paso:

1. Crear una nueva base de datos en SSMS para almacenar todas las tablas de datos proporcionadas.

2. Definir la estructura de las tablas necesarias para los conjuntos de datos (DIM_CATEGORY, DIM_PRODUCT, DIM_SEGMENT, DIM_CALENDAR, FACT_SALES).

3. Importar los datos en cada tabla utilizando las funciones de importación.

4. Realizar consultas básicas para verificar que los datos se han cargado correctamente.

5. Ejecutar uniones entre las tablas para combinar la información y obtener insights clave sobre las ventas por categoría, región y periodo de tiempo.

6. Guardar y documentar el proceso, incluyendo capturas de pantalla de las consultas y los resultados obtenidos.

Archivo final

Documento de PDF que contenga capturas de pantalla del proceso así como también el archivo .sql con los queries realizados.

Fase 5 de trabajo

Crearás un dashboard interactivo en Power BI que permitirá a los stakeholders explorar los insights obtenidos de manera clara y concisa, visualizando resultados clave sobre las ventas, segmentos, productos y escenarios predictivos.

Paso a paso:

Configurar Power BI Desktop:

- Abre Power BI Desktop e inicia un nuevo proyecto.
- Familiarízate con la interfaz principal, donde podrás conectar tus datos, crear visualizaciones y construir tu dashboard.

Conectar las fuentes de datos:

- Conecta Power BI a los conjuntos de datos preparados. Puedes importar datos desde archivos CSV, Excel o conectarte directamente a bases de datos como SQL Server.
- Asegúrate de que los datos estén correctamente estructurados y listos para ser usados en las visualizaciones.

Limpiar y transformar los datos:

- Usa el editor de consultas de Power BI para realizar cualquier transformación de los datos que sea necesaria.
- Revisa que todos los datos relevantes, como ventas por producto, región, categoría, etc., estén listos para su análisis.

Crear la estructura del dashboard:

- Planifica la estructura general de tu dashboard, dividiendo la información en secciones clave como Resumen de ventas, Desempeño por categoría, Análisis geográfico.
- Define qué tipos de visualizaciones son las más adecuadas para cada sección (gráficos de barras, líneas, mapas, matrices, etc.).

Diseñar las visualizaciones principales:

- Crea visualizaciones clave para representar los insights obtenidos, como:
    - **Gráfico de barras:** para comparar ventas por categorías o productos.
    - **Gráfico de líneas:** para mostrar tendencias de ventas a lo largo del tiempo.
    - **Mapa interactivo:** para visualizar la distribución geográfica de las ventas.
    - **Matriz o gráfico de tablas:** para analizar relaciones y segmentación de productos.
- Asegúrate de que cada visualización esté correctamente etiquetada y tenga un formato claro.

Agregar interactividad al dashboard:

- Utiliza los filtros y segmentadores (slicers) de Power BI para permitir que los usuarios interactúen con el dashboard. Estos filtros pueden basarse en variables como fecha, categoría de producto, región, o segmento.
- Asegúrate de que las visualizaciones se actualicen de forma dinámica cuando los usuarios interactúen con los filtros.

Optimizar el diseño del dashboard:

- Revisa el diseño del dashboard para asegurarte de que sea profesional y fácil de entender. Utiliza una paleta de colores coherente, fuentes legibles, y mantén una estructura clara.
- Organiza las visualizaciones para que las más importantes estén en la parte superior y que las secciones tengan un flujo lógico.

Agregar secciones de texto y explicaciones:

- Incluye notas, títulos y explicaciones breves en el dashboard para guiar a los usuarios y destacar los insights más relevantes.
- Asegúrate de que el dashboard no solo muestre datos, sino que cuente una historia basada en los insights obtenidos.

Probar la interactividad del dashboard:

- Navega por el dashboard usando los filtros interactivos para asegurarte de que todo funcione correctamente. Verifica que las visualizaciones cambien dinámicamente y que la experiencia del usuario sea fluida.

Publicar y compartir el dashboard:

- Publica el dashboard en el servicio de Power BI (Power BI Service) para compartirlo.
- Ajusta las configuraciones de acceso y visibilidad, y genera un enlace para que los usuarios puedan acceder al dashboard de manera segura.

Archivo final

Dashboard publicado en Power BI, con un enlace compartido y una breve guía de navegación que explique cómo interactuar con el dashboard.

Fase 6 de trabajo

Desarrollarás un modelo predictivo para prever las ventas futuras de productos clave como Vanish y Lysol. Utilizarás técnicas de regresión o análisis de series de tiempo, y asegurarás que el modelo esté correctamente entrenado y validado.

Paso a paso:

Cargar y preparar los datos:

- Carga el conjunto de datos consolidado que contiene las ventas históricas

Seleccionar la técnica de modelado:

- Decide si utilizarás un modelo de regresión (por ejemplo, regresión lineal múltiple) o un modelo de series de tiempo (por ejemplo, ARIMA, SARIMA).
- Justifica tu elección en base a los patrones observados en los datos.

Dividir los datos en conjuntos de entrenamiento y prueba:

- Divide los datos en conjuntos de entrenamiento (training) y prueba (testing) para validar el modelo.
- Utiliza un método como la división temporal o la validación cruzada si trabajas con series de tiempo.

Construir y entrenar el modelo predictivo:

- Si eliges una regresión, entrena el modelo utilizando las variables independientes seleccionadas y la variable dependiente (ventas).
- Si eliges un modelo de series de tiempo, ajusta el modelo ARIMA o SARIMA a las ventas históricas, seleccionando los parámetros óptimos mediante métodos como AIC o BIC.

Validar el modelo:

- Evalúa el desempeño del modelo utilizando métricas de error comunes como el error cuadrático medio (MSE), el error absoluto medio (MAE) o el porcentaje de error absoluto medio (MAPE).
- Realiza un gráfico de las predicciones frente a los datos reales para visualizar el ajuste del modelo.

Ajustar y optimizar el modelo:

- Realiza ajustes al modelo para mejorar su precisión (por ejemplo, ajustando hiperparámetros, eliminando variables no significativas o probando diferentes métodos de modelado).
- Documenta los cambios realizados y su impacto en el rendimiento del modelo.

Generar las predicciones futuras:

- Utiliza el modelo optimizado para predecir las ventas futuras para los próximos meses.

- Presenta las predicciones en un formato claro, como gráficos de líneas o tablas comparativas.

Documentar el proceso y los resultados:

- Documenta todos los pasos del proceso de construcción del modelo, incluyendo la justificación de la técnica de modelado elegida, los resultados de la validación y los ajustes realizados.
- Explica las predicciones obtenidas y proporciona recomendaciones basadas en los resultados del modelo.

Archivo final

Un archivo Jupyter Notebook (.ipynb) que contenga el código, los gráficos y las explicaciones de cada paso del proceso de construcción y validación del modelo.

Fase final de trabajo

Asumirás el rol de un **Científico de Datos** en una compañía de productos de consumo, donde tu objetivo principal es proporcionar insights clave sobre el rendimiento de ventas de productos y realizar una predicción de ventas futuras para ayudar a la compañía a tomar decisiones estratégicas. Durante este proyecto, has trabajado con herramientas como **SQL, Python y PowerBI** para llevar a cabo un análisis exhaustivo de los datos. Ahora, deberás presentar los hallazgos obtenidos y las predicciones generadas en una presentación estructurada que muestre tanto tu capacidad técnica como la habilidad para comunicar resultados de manera efectiva.

Paso a Paso:

1. Introducción del Proyecto

En la introducción, debes proporcionar una descripción clara y concisa del proyecto que realizarás. Asegúrate de incluir los siguientes puntos:

1. **Contexto del Proyecto**: Explica el propósito general del análisis.
2. **Relevancia del Análisis**: Menciona por qué este análisis es importante para la estrategia de la empresa.
3. **Herramientas Utilizadas**: Indica las herramientas y tecnologías que usarás durante el proyecto, como SQL, Python, PowerBI, etc.

2. Limpieza y Transformación de Datos con Python

- **Carga y Limpieza de Datos**: Los datos fueron cargados en **Pandas** desde archivos CSV y Excel. Se realizó una limpieza exhaustiva para manejar valores nulos, eliminar duplicados y corregir inconsistencias en los datos.
- **Transformaciones**: Se aplicaron transformaciones para normalizar los datos y estandarizar formatos de fechas y categorías.
- **Capturas y Código**: Incluir fragmentos de código que demuestran el proceso de limpieza y consolidación de los datos en **Python**.

3. Análisis Exploratorio de Datos (EDA) y Visualizaciones

- **Visualización de la Distribución de Ventas**: Se crearon histogramas y boxplots para visualizar la distribución de ventas por categoría de producto y región.
- **Análisis de Tendencias**: Gráficos de líneas permitieron observar tendencias de ventas en el tiempo, filtrando por categorías clave.
- **Identificación de Outliers**: Se usaron gráficos de dispersión y de caja para identificar posibles valores atípicos que puedan requerir un análisis más detallado.
- **Capturas de Visualizaciones**: Incluir ejemplos de las visualizaciones generadas con **Matplotlib** y **Seaborn** en Python.

4. Segmentación de Productos o Regiones con Clustering

- **Aplicación de Clustering con K-Means:** Se aplicó el algoritmo K-Means para segmentar los productos o regiones en función de características clave como ventas totales, categorías y regiones.
- **Número Óptimo de Clusters:** Se utilizó el método del codo para determinar el número óptimo de clusters, logrando identificar patrones de rendimiento.
- **Interpretación de los Clusters:** Se analizó cada cluster para identificar áreas de alto rendimiento y segmentos con oportunidades de mejora.
- **Visualizaciones de Clusters:** Se generaron gráficos de dispersión para visualizar los clusters y entender mejor las relaciones entre las diferentes variables.
- **Capturas de Pantalla y Código:** Incluir capturas de pantalla de los gráficos y el código utilizado para implementar el modelo de clustering en Python.

5. Análisis de Datos con SQL

- **Creación de la Base de Datos**: La base de datos fue diseñada e implementada en **SQL Server** utilizando las tablas de ventas proporcionadas (**DIM_CATEGORY, DIM_PRODUCT, DIM_SEGMENT, DIM_CALENDAR, FACT_SALES**).

- **Consultas Clave**: Se ejecutaron consultas básicas para verificar la correcta carga de datos y se realizaron uniones entre las tablas para obtener insights sobre las ventas por categoría y región.
- **Capturas de Pantalla y Código**: Incluir capturas de pantalla del proceso de creación de la base de datos y ejemplos de queries ejecutados en **SQL** para obtener insights clave.

6. Creación de un Dashboard en PowerBI

- **Estructura del Dashboard:** El tablero incluye secciones clave como resumen de ventas, desempeño por categoría, y análisis geográfico. Las visualizaciones fueron diseñadas para ser interactivas, permitiendo a los usuarios explorar los datos mediante filtros.
- **Visualizaciones Interactivas:** Se utilizaron gráficos de barras, líneas, mapas y tablas dinámicas para representar los insights obtenidos. Los usuarios pueden filtrar los resultados por región, segmento y fechas.
- **Capturas de Pantalla y Enlace:** Incluir capturas de pantalla del dashboard en Power BI, junto con un enlace al tablero publicado si es necesario.

7. Predicción de Ventas con Machine Learning

- **Selección del Modelo**: Se utilizó un modelo de **regresión lineal múltiple** y un modelo de **series de tiempo ARIMA** para predecir las ventas futuras. La selección del modelo se basó en los patrones de ventas observados en los datos históricos.
- **Validación del Modelo**: El modelo fue validado utilizando métricas como el **Error Absoluto Medio (MAE)** y el **Error Cuadrático Medio (MSE)**. Se realizaron ajustes para optimizar la precisión del modelo.
- **Predicciones Futuros**: Las predicciones de ventas para los próximos meses se presentaron en forma de gráficos de líneas comparativos entre los datos reales y las predicciones del modelo.
- **Capturas y Explicaciones**: Incluir gráficos comparativos y fragmentos de código que explican la implementación del modelo en **Python**.

8. Conclusiones y Recomendaciones

En las conclusiones, debes resumir tus hallazgos más importantes y proporcionar recomendaciones basadas en tu análisis. Aquí tienes lo que debes incluir:

1. **Resumen de los Hallazgos Clave**: Ofrece un resumen de los principales insights obtenidos a lo largo del proyecto.

2. **Impacto del Análisis en la Estrategia Empresarial**: Explica cómo los resultados pueden influir en las decisiones futuras de la compañía.
3. **Recomendaciones**: Proporciona sugerencias específicas para mejorar la estrategia de la empresa.
4. **Futuras Mejoras**: Indica posibles mejoras o pasos adicionales que podrían tomarse en el futuro.

Archivo Final

- **Formato de Entrega**: La presentación debe entregarse en un archivo **PowerPoint (.pptx)** o mediante un enlace a **Google Slides**. El archivo debe incluir todas las capturas de pantalla, gráficos y visualizaciones mencionadas.

¿Cómo presentar su entrega?

Un archivo Jupyter Notebook (.ipynb) que contenga el código, los gráficos y las explicaciones de cada paso del proceso de construcción y validación del modelo.