Universidad de San Carlos de Guatemala

Facultad de ingeniería

Escuela de ciencias y sistemas

Sistemas organizacionales y gerenciales 2

Ing. Mario José Bautista Fuentes

Aux. Jimmy Daniel Larios Martínez

PRACTICA 1 – DOCUMENTACION

José Luis Reynoso Tiu

201345126

N+

25 de septiembre de 2024, Guatemala, Guatemala

Planificación

- Herramientas y tecnología
    - Python
        - Python es un lenguaje de programación flexible y ampliamente utilizado para la manipulación de datos, desarrollo web y automatización de tareas.
        - Se conecta fácilmente con MySQL utilizando bibliotecas.
        - Permite la creación de scripts para realizar tareas ETL.
    - MySQL
        - Es un sistema de gestión de bases de datos relacional muy popular, robusto y de código abierto.
        - Ofrece un alto rendimiento para aplicaciones pequeñas y medianas.
        - Tiene un soporte excelente para operaciones transaccionales, lo que es esencial para manejar datos de ventas.
        - MySQL es compatible con SQL estándar y ofrece integración con muchas herramientas y lenguajes, incluyendo Python.
        - MySQL Workbench es una herramienta gráfica que facilita la creación y administración de bases de datos MySQL.
        - Permite visualizar el diseño del esquema dimensional y ejecutar consultas SQL para verificar que los datos estén correctamente almacenados.
    - Pandas
        - Pandas es una biblioteca de Python especializada en la manipulación de datos y análisis.
        - Nos permite cargar y trabajar con los datos de la tabla de ventas de manera eficiente, permitiendo seleccionar, agrupar y transformar los datos antes de insertarlos en las tablas de dimensiones y hechos.
    - Matplotlib y Seaborn
        - Estas bibliotecas de visualización de datos en Python son ideales para generar gráficos simples y visualizaciones personalizadas.
        - Son utilizadas para generar gráficos a partir de los datos agregados en el análisis exploratorio y las tendencias de ventas.
    - ETL
        - El proceso ETL es crucial en la creación de sistemas de análisis de datos. Python es ideal para implementar un ETL personalizado, ya que permite extraer los datos de la tabla venta, transformarlos en las dimensiones correspondientes y cargarlos en las tablas de hechos.

- Plazos de las fases del proyecto
    - **Fase de planificación y diseño del esquema (miércoles - jueves)**
        - Revisión de los requisitos del proyecto.
        - Diseño del modelo dimensional
        - Definición de las tablas de hechos y dimensiones.
    - **Fase de configuración de entorno y creación de tablas (viernes)**
        - Configuración del entorno de desarrollo (MySQL, Python).

        - Creación de las tablas en MySQL (venta, tablas de hechos y dimensiones).
        - Verificación del esquema creado mediante MySQL Workbench o una herramienta similar.
    - **Fase de desarrollo del proceso ETL (sábado)**
        - Desarrollo del script en Python para cargar datos desde la tabla venta a las tablas de dimensiones y hechos.
        - Asegurarse de manejar correctamente los duplicados o errores de inserción.
        - Realizar pruebas con los datos de ejemplo para verificar la carga en las tablas de hechos y dimensiones.
    - **Fase de análisis exploratorio y visualización (domingo)**
        - Implementar los análisis de datos usando las tablas de hechos y dimensiones.
        - Crear gráficos que muestren tendencias de ventas, análisis por categorías de productos y regiones, entre otros.
        - Validar los gráficos con los datos de ejemplo para verificar que representen correctamente la información.
    - **Fase de pruebas finales y corrección de errores (lunes)**
        - Realizar pruebas completas del sistema, incluyendo la validación de la correcta carga de datos y la generación de gráficos.
        - Corregir errores que puedan surgir en el proceso ETL o en las consultas de análisis.
        - Validar que los gráficos y análisis estén correctos.
    - **Fase de documentación y entrega (martes)**
        - Crear la documentación del proyecto, describiendo el modelo de datos, el proceso ETL y los análisis realizados.
        - Verificar que todo el código esté correctamente organizado.

Proceso de análisis

- Enfoque para limpiar y preparar los datos
    - Comprender la estructura y naturaleza de los datos, identificar las posibles columnas que podían tener valores nulos, errores o inconsistencias, y decidir cómo dividir los datos entre las tablas de dimensiones y hechos.
    - Eliminar presencia de duplicados
    - Identificar las columnas con valores nulos se eliminan.
    - Normalización de los datos implica estandarizar los formatos de los datos, asegurando que los nombres, fechas y otros campos sigan un formato uniforme.
    - Una vez que los datos estaban limpios y preparados, el siguiente paso fue dividir la información en las tablas correspondientes de dimensiones y hechos según el modelo dimensional.
    - La última etapa consistió en realizar verificaciones finales para asegurarnos de que los datos estaban correctamente cargados y no presentaban inconsistencias.
- Decisiones durante el análisis exploratorio de datos
    - Calcular estadísticas básicas

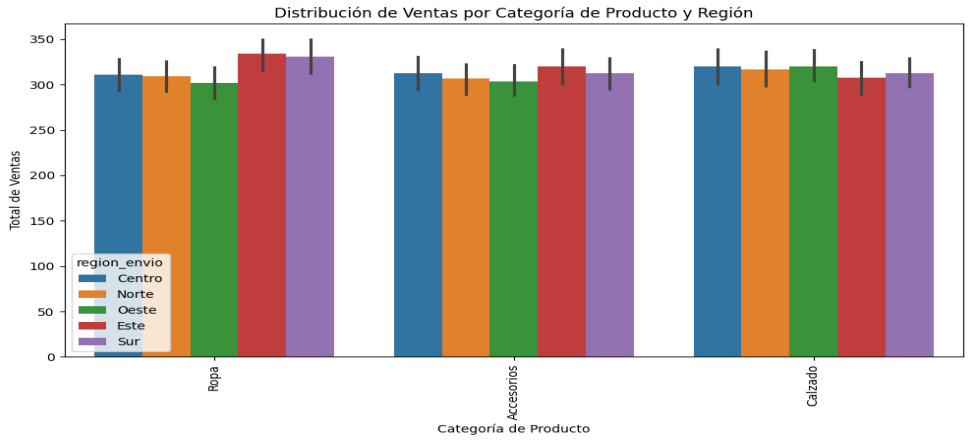
Promedio	
total_orden	314.675745
cantidad	3.000613
precio_producto	104.871070
edad_cliente	47.607697

Mediana	
total_orden	254.92
cantidad	3.00
precio_producto	104.82
edad_cliente	48.00

Moda	
total_orden	150.20
cantidad	2.00
precio_producto	50.78
edad_cliente	51.00

Correlación entre las variables				
	total_orden	cantidad	precio_producto	edad_cliente
total_orden	1.000000	0.632440	0.000909	-0.000396
cantidad	0.632440	1.000000	-0.000089	-0.000531
precio_producto	0.000909	-0.000089	1.000000	0.000221
edad_cliente	-0.000396	-0.000531	0.000221	1.000000

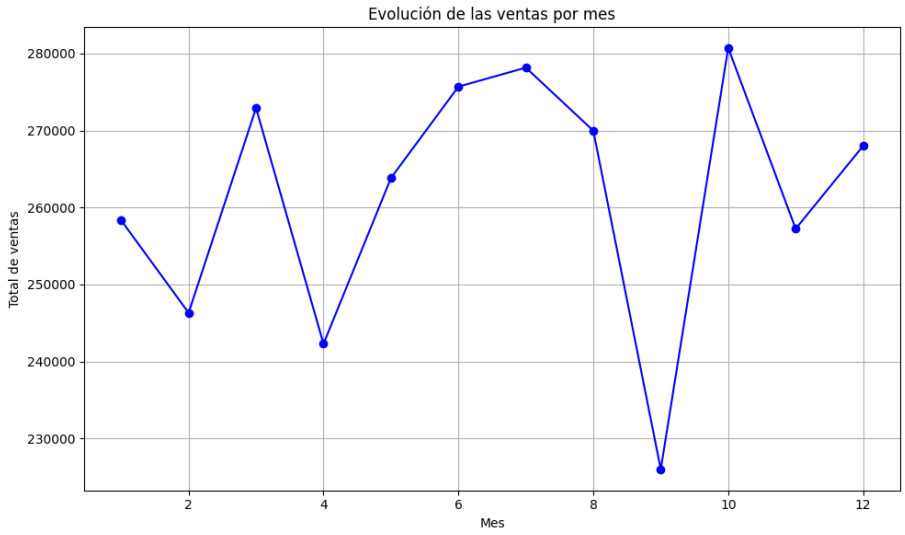
- La relación entre el **total de la orden** y la **edad del cliente** es prácticamente inexistente. Es decir, no hay una tendencia clara que indique que los clientes más jóvenes o mayores hagan compras significativamente mayores o menores.
-
    - Crear visualizaciones para mostrar la distribución de ventas por categoría de producto y región



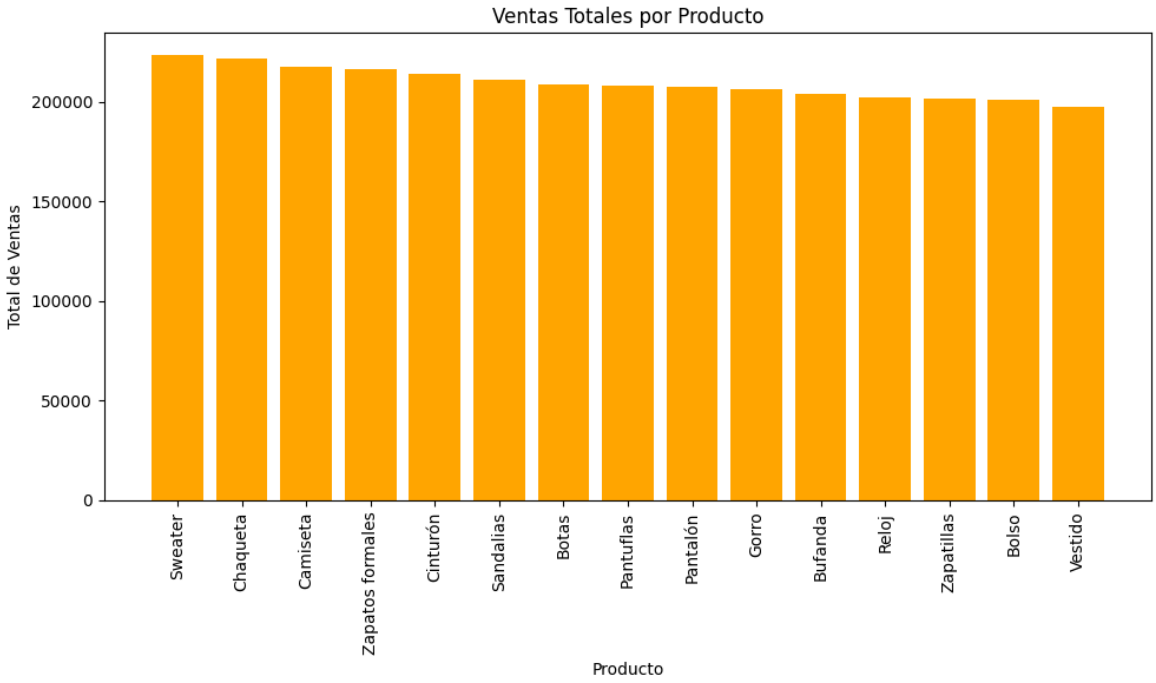
- Explicación las decisiones tomadas
    - Este análisis fue útil para entender el rango de valores en las transacciones, identificar posibles sesgos o anomalías (como ventas extremadamente altas o bajas) y detectar tendencias generales en los precios y cantidades.
    - La cantidad y el total de la orden tienen una correlación significativa (0.63), lo que es esperado, ya que la cantidad de productos vendidos impacta directamente en el total de la orden.
    - Sin embargo, otras correlaciones fueron prácticamente inexistentes, como entre precio_producto y cantidad.
    - Este tipo de gráfico permite ver claramente las diferencias en el volumen de ventas por cada categoría y cómo varía por región. En este caso, la distribución es bastante homogénea, lo que podría sugerir que no hay una gran diferencia en el desempeño por región para cada categoría.
- Desafíos durante el análisis
    - Durante la carga inicial de los datos, identificamos que había varias entradas duplicadas en la tabla venta, especialmente en el campo id_orden, lo que generaba inconsistencias al intentar insertar datos en las tablas de hechos y dimensiones.
        - Utilizamos la función drop_duplicates() de **Pandas** para eliminar los duplicados de las transacciones basándonos en el campo id_orden. Esto aseguró que solo se mantuvieran los registros únicos de cada transacción.
    - Algunas columnas críticas, como id_cliente, nombre_producto y total_orden, presentaban valores nulos en ciertas filas.
        - Se eliminaron los registros con valores nulos en las columnas críticas utilizando **Pandas**.
    -

Metodología

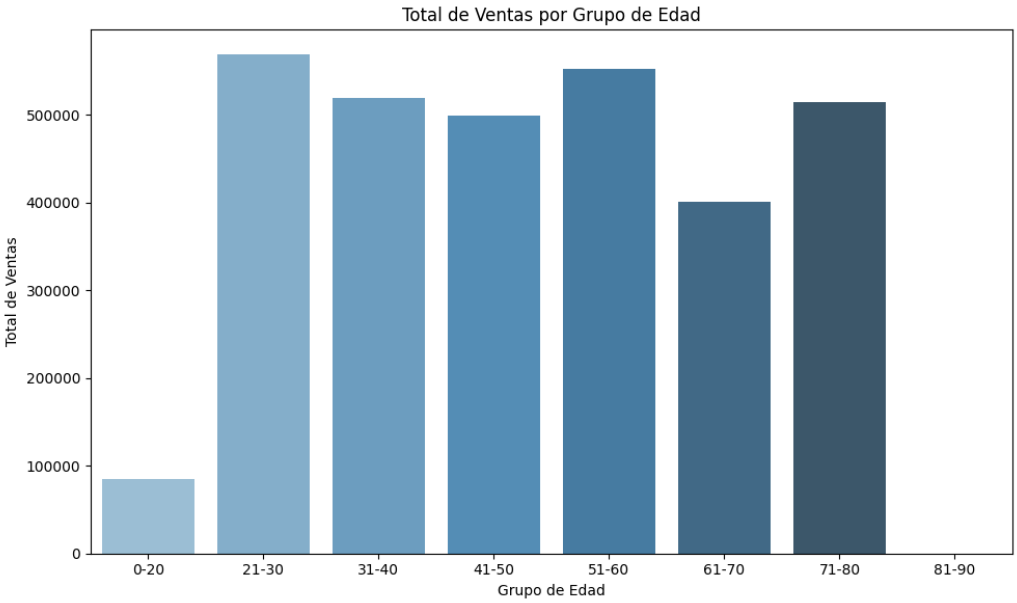
- Selección de las visualizaciones más apropiadas de los hallazgos.



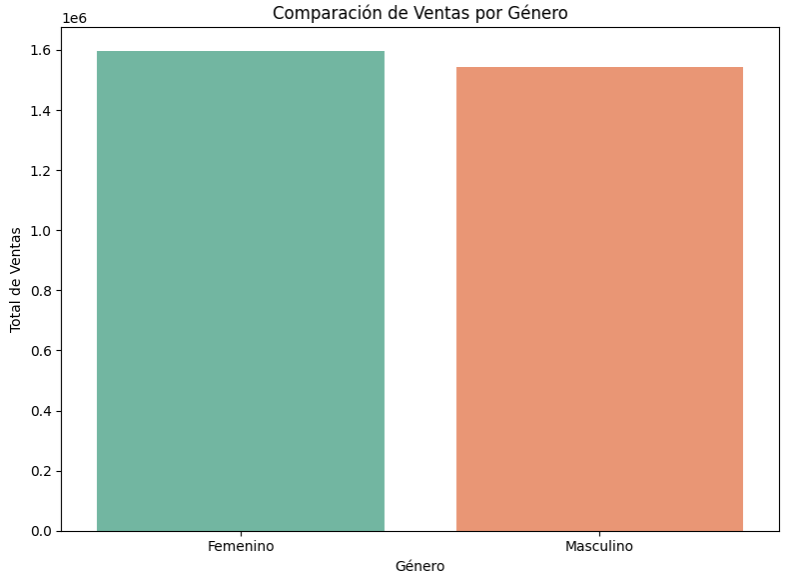
En la grafica se puede ver los meses con las ventas totales por mes, esta visualización ayuda a tener una proyección de ventas o bien una organización optima para atender la capacidad de ventas por mes.



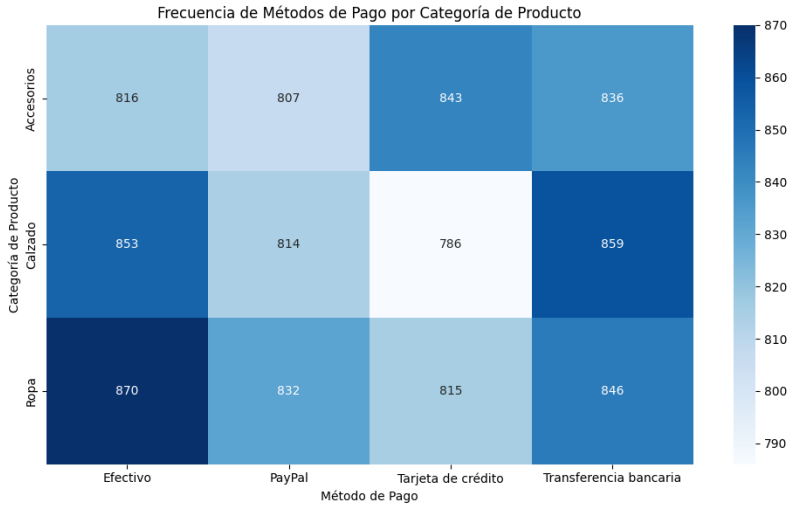
Con grafica anterior podemos visualizar las ventas en orden de mayor a menor por lo que podemos identificar los productos más vendidos y menos vendidos que nos pueden ayudar a determinar la cantidad de stock que el negocio necesita. Apoyándonos con la grafica de ventas mensuales podemos incluso mejorar las decisiones a tomar.



En esta grafica podemos apreciar que las personas entre 21 y 30 años son los mayores consumidores al igual que los de 51 a 60 años. Con ello podemos crear estrategias para acaparar mayor numero de clientes en ese rango de edad y tener productos exclusivamente atrayentes para jóvenes o bien analizar y fomentar nuevas estrategias para aumentar el consumo en los demás rangos de edad.

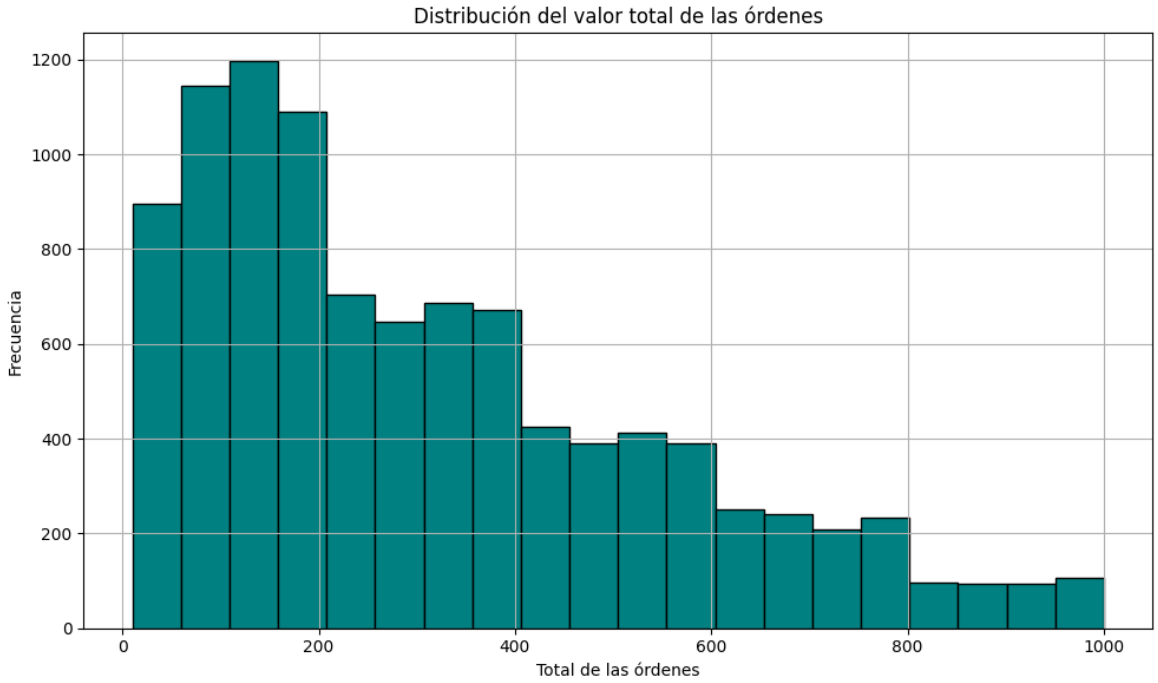


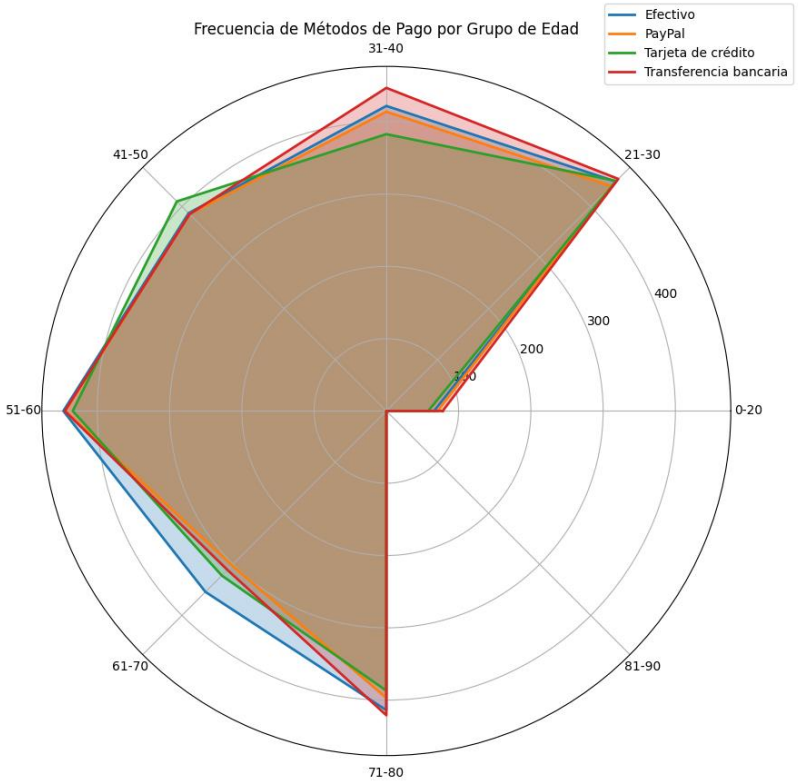
Con la comparación de ventas por genero podemos observar que no varía significativamente, esto quiere decir que el consumo por genero es balanceado y que los productos están enforcados para cualquier género.

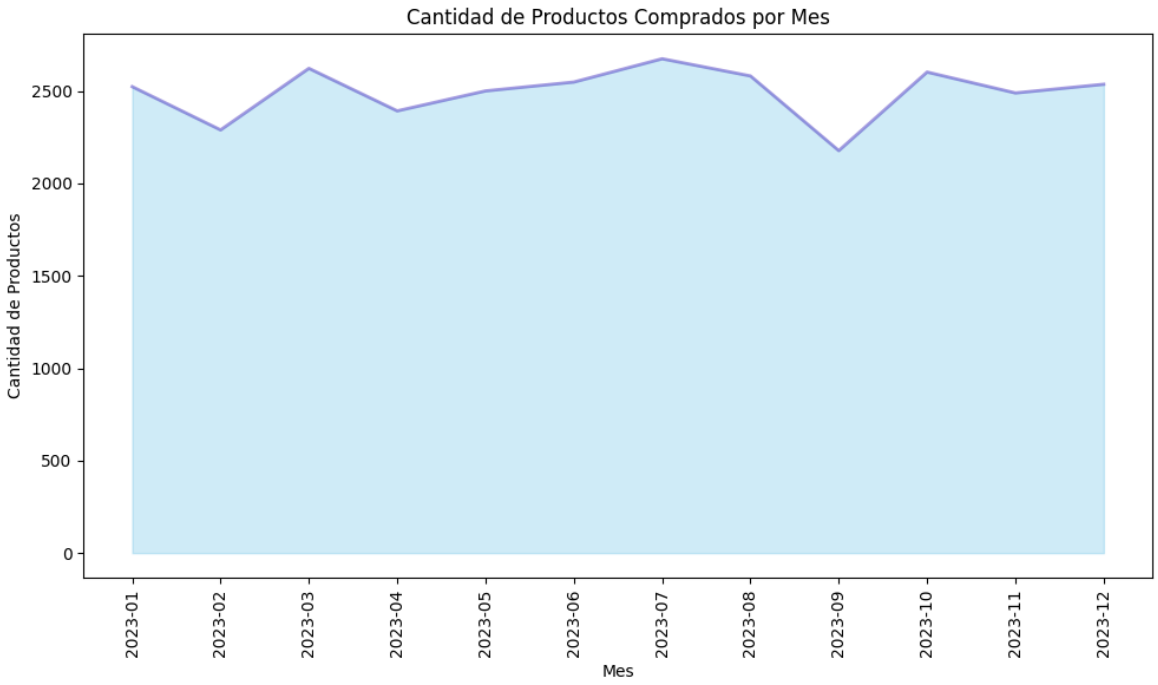


En esta grafica se observa que el efectivo es el método de pago más utilizado y cuando se compra ropa y el menos utilizado es la tarjeta de crédito al momento de comprar calzado.

Graficas extras







Las graficas se seleccionaron dependiendo de como poder acomodar los datos y que sea de fácil comprensión. Se necesita comprender fácilmente ya que el propósito de las mismas es facilitar la toma de decisiones basándose en los datos que presenta.

Preguntas y respuestas

- ¿Cómo podrían los insights obtenidos ayudar a diferenciarse de la competencia?
    - Los insights obtenidos del análisis de métodos de pago y patrones de compra por edad ofrecen una ventaja clave como la personalización. Si la empresa entiende qué grupos de edad prefieren ciertos métodos de pago
        - Ofrecer promociones específicas para los métodos de pago más populares como descuentos en el uso de tarjetas de crédito para clientes jóvenes o programas de lealtad para aquellos que prefieren pagos en efectivo.
        - Implementar las opciones de pago más demandadas en su interfaz y en los canales de ventas, asegurando que el proceso de compra sea más rápido y cómodo que el de la competencia.
- ¿Qué decisiones estratégicas podrían tomarse basándose en este análisis para aumentar las ventas y la satisfacción del cliente?
    - Usar los datos para crear campañas dirigidas específicamente a los segmentos más activos. Si, por ejemplo, los clientes entre 30-40 años utilizan más las tarjetas de crédito, la empresa podría lanzar promociones específicas para ellos.
    - Si los datos muestran que un método de pago como la **transferencia bancaria** es especialmente popular entre clientes mayores, la empresa puede centrarse en mejorar la experiencia de este método de pago, ofreciendo opciones de financiación o descuentos asociados.

    - Conociendo los patrones de compra según la edad, la empresa puede recomendar productos más relevantes para ciertos grupos. Esto puede aumentar tanto las ventas como la satisfacción del cliente al ofrecer una experiencia de compra más personalizada.
- ¿Cómo podría este análisis de datos ayudar a la empresa a ahorrar costos o mejorar la eficiencia operativa?
    - Saber qué métodos de pago son más populares entre los clientes puede ayudar a reducir costos. Como cuando en ciertos métodos de pago resultan ser más económicos en términos de comisiones y procesamiento, la empresa podría impulsar su uso mediante incentivos.
    - Al comprender qué productos son preferidos por cada grupo de edad, la empresa puede ajustar sus inventarios para evitar sobre-stock o faltantes de productos clave, optimizando el flujo de efectivo y reduciendo costos de almacenamiento.
    - Si se observa que ciertos métodos de pago son más propensos a fraudes o devoluciones en algunos grupos de clientes, la empresa puede implementar controles más estrictos solo en esos segmentos, mejorando la seguridad y reduciendo pérdidas.
- ¿Qué datos adicionales recomendarían recopilar para obtener insights aún más valiosos en el futuro?
    - Recopilación de datos como
        - El tiempo pasan los clientes en el sitio web antes de hacer una compra
        - Qué productos consultan antes de tomar una decisión
        - Estos datos pueden ayudar a mejorar la experiencia de usuario.
    - Encuestas de satisfacción pueden proporcionar un contexto importante sobre cómo mejorar la experiencia de compra.
    - El análisis por región puede revelar diferencias en preferencias de métodos de pago o productos, lo que permitiría adaptar las estrategias de marketing a nivel local.
    - Recopilar datos sobre la frecuencia con la que los clientes compran y qué productos prefieren repetidamente puede ayudar a generar programas de fidelización más efectivos.
    - Saber si los clientes prefieren realizar compras a través de la web, la app o en tiendas físicas permitirá optimizar los esfuerzos de marketing y distribución de recursos entre los diferentes canales.