Informe de Planificación y Proceso de Análisis de Datos

a) Planificación

División de tareas en el equipo

El equipo se conformó por **2 integrantes**, y las responsabilidades se distribuyeron de la siguiente manera:

- **Douglas Rivera (Líder de proyecto):** Coordinación general, establecimiento de plazos e incisos 1-4, y despliegue en nube.
- **Steven Gonzalez:** Insicos de 5-8, conclusiones y documentacion.

Herramientas y tecnologías utilizadas

- **Python (pandas, seaborn, matplotlib):** para la manipulación, análisis y visualización de datos.
- **MySQL en Railway:** almacenamiento y consulta inicial de datos de ventas.
- **GitHub:** control de versiones y trabajo colaborativo en ramas.

Estas herramientas se eligieron porque permitían **colaboración remota, reproducibilidad del análisis y escalabilidad** del proyecto.

Establecimiento de plazos

El proyecto se organizó en **4 fases**, cada una con plazos definidos:

1. **Recolección y limpieza de datos:** 1 dia.
2. **Normalización y construcción de DataFrames de análisis:** 2 dias.
3. **Análisis exploratorio y visualizaciones:** 3 dias.
4. **Documentación y presentación final:** 3 dias.

La duración se estimó considerando la **complejidad de cada tarea y la experiencia previa** de los miembros del equipo.

---

b) Proceso de análisis

Enfoque paso a paso para limpiar y preparar los datos

1. **Extracción inicial:** se descargaron 106,281 registros desde la base de datos en Railway todo en una sola tabla debido a que por limitaciones de railway no podiamos almacenar mas entidades, tambien no usamos aws por no tener una cuenta disponible inmediata.
2. **Revisión de duplicados:** se eliminaron entradas repetidas en órdenes y productos.
3. **Normalización:** se crearon tablas separadas para clientes, productos, órdenes y detalles de compra.
4. **Renombrado de columnas:** se estandarizaron a nombres en español para mayor claridad (ej. `customer_id` → `ID del cliente`).
5. **Conversión de tipos:** las fechas se transformaron a formato `datetime` y los precios a `float`.

6. **Validación:** se verificó la consistencia de claves primarias y foráneas entre tablas.

Decisiones tomadas en el análisis exploratorio de datos

- **Agrupación de edades:** se definieron rangos (18-24, 25-34, etc.) para simplificar la segmentación.
- **Selección de métricas clave:** ticket promedio, ventas por categoría y frecuencia de compra por cliente.
- **Visualización prioritaria:** se decidió usar gráficos de barras para categorías y regiones, y boxplots para comparar gasto por grupo de edad.
- **Descartar correlaciones irrelevantes:** se comprobó que la edad no tenía correlación significativa con el monto de compra, por lo que no se incluyó como factor principal en las conclusiones.

Desafios

- Principalmente no tener una cuenta a la mano de gcp o de aws nos limitó por lo que optamos por usar railway y luego se nos presentó la situacion de que railway en capa gratuita no nos permitia almacenar varias tablas y hacer bulk inserts de los csv por lo tanto se opto por dejar una sola tabla en la nube y luego con los datos en el notebook normalizar.

- limitaciones de tiempo tambien ya que la practica duro aproximadament semana y media y al haber mas entregables y trabajo se complicó un poco.

---

c) Metodologia

Selección de visualizaciones

La elección de visualizaciones se basó en el tipo de información que queríamos resaltar y en la claridad con la que los hallazgos podían ser interpretados por la audiencia:

1. **Gráficos de barras:**
   Utilizados para comparar **ventas totales por categoría de producto y por región**.
   - Motivo: permiten ver rápidamente cuáles son las categorías o regiones más rentables.
   - Ejemplo: la categoría *Ropa* destacó como la principal fuente de ingresos.

2. **Gráficos de líneas:**
   Aplicados al análisis de la **tendencia mensual de ventas**.
   - Motivo: son ideales para identificar variaciones en el tiempo y detectar patrones de crecimiento o estacionalidad.
   - Ejemplo: se observaron picos en determinados meses que podrían corresponder a temporadas de rebajas.

3. **Boxplots (diagramas de caja):**
   Usados para analizar los **patrones de compra por grupo de edad**.
   - Motivo: muestran la dispersión de los montos de compra, la mediana y la presencia de valores atípicos.

- Ejemplo: permitió visualizar que, a pesar de que los jóvenes (18-24) concentran más órdenes, los tickets altos aparecen en rangos de mayor edad.

4. **Gráficos de pastel:**
   Empleados para la **participación de ventas por región**.
   - Motivo: útiles para resaltar la proporción relativa de cada zona dentro del total de ventas.
   - Ejemplo: la región Este representó la mayor contribución en ingresos.

5. **Histogramas:**
   Utilizados para la **distribución de precios de productos**.
   - Motivo: ayudan a identificar rangos de precios más comunes y la dispersión de los valores.
   - Ejemplo: se evidenció que la mayoría de productos tienen un precio promedio cercano a $100.

d) Respuestas

1. ¿Cómo podrían los insights obtenidos ayudar a diferenciarse de la competencia?
   - Al identificar que los clientes jóvenes (18-24) concentran la mayoría de órdenes, la empresa puede enfocar campañas digitales dirigidas exclusivamente a este público, logrando un marketing más efectivo que sus competidores.
   - Identificar el equilibrio de ventas entre ropa, calzado y accesorios permite diseñar colecciones integradas o "combos" exclusivos, ofreciendo propuestas de valor diferenciadas frente a tiendas que solo destacan en una categoría.
2. ¿Qué decisiones estratégicas podrían tomarse basándose en este análisis para aumentar las ventas y la satisfacción del cliente?
   - Implementar descuentos por metas o rachas, membresías o beneficios por volumen de compra que mantengan el alto nivel de recurrencia en el segmento 18-24.
   - Desarrollar nuevas líneas de productos premium y campañas específicas para clientes con mayor poder adquisitivo, aumentando así el monto de venta promedio.
3. ¿Cómo podría este análisis de datos ayudar a la empresa a ahorrar costos o mejorar la eficiencia operativa?
   - Al conocer los productos más y menos vendidos, se puede reducir el stock de baja rotación y enfocar la inversión en los más demandados.
   - Los datos por región de envío permiten ajustar rutas, negociar con transportistas y abrir puntos de distribución estratégicos para reducir costos y tiempos de entrega.
4. ¿Qué datos adicionales recomendarían recopilar para obtener insights aún más valiosos en el futuro?
   - Encuestas post-compra, tiempos de entrega reales, devoluciones y reseñas para correlacionar satisfacción con ventas.

- Navegación en paginas web, carritos abandonados, fuentes de tráfico (redes sociales, publicidad, búsqueda orgánica) y tasas de conversión, para optimizar campañas de marketing.