FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

[7506-9558] Organización/Ciencia de Datos
Cátedra Martinelli
Segundo cuatrimestre de 2025

Alumno	Padrón	Email
Giannantonio, Maurizio	108119	mgiannantonio@fi.uba.ar

Índice

1. Introducción

En el presente trabajo se lleva a cabo un análisis exploratorio de un conjunto de datos compuesto por los siguientes *datasets*: `orders`, `order_items`, `categories`, `customers`, `reviews`, `products` e `inventory_logs`. Estos archivos, provistos por la cátedra en formato `.csv`, fueron cargados en `pandas` como *dataframes* para su posterior tratamiento. En una primera instancia se realizó un proceso de limpieza y la asignación adecuada de tipos de datos en cada columna. Los distintos *datasets* se encuentran relacionados, de manera total o parcial, mediante identificadores que permiten integrar y vincular la información.

El objetivo principal de este trabajo práctico es que el estudiante formule y resuelva consultas sobre las relaciones y patrones presentes en los datos, utilizando la librería `pandas` para su procesamiento y análisis, y complementando los resultados con visualizaciones que faciliten su interpretación.

Para cada consulta se expone la hipótesis inicial, un resumen del análisis realizado y la conclusión correspondiente. Cabe destacar que el código completo empleado se encuentra disponible en el *notebook* de Google Colab adjunto; por lo tanto, el presente informe se centra en el razonamiento seguido y en los hallazgos obtenidos a partir del análisis.

2. Consultas de la cátedra

2.1. Estado con más descuentos

Hipótesis: Se espera que los estados con mayor volumen de órdenes presenten también la mayor cantidad de descuentos aplicados, dado que un mayor nivel de actividad comercial suele estar asociado con más promociones y rebajas.

Conclusión: El análisis muestra que el estado FM es el que concentra la mayor cantidad de descuentos en términos absolutos. Sin embargo, al considerar el promedio de descuentos por orden, el estado NC lidera el ranking, lo que indica que no necesariamente los estados con mayor volumen de ventas son los que otorgan los descuentos más altos en promedio.

Nota: El código completo y las salidas de esta consulta se encuentran disponibles en el *notebook* de Google Colab adjunto.

2.2. ¿Cuáles son los 5 códigos postales más comunes para las órdenes con estado 'Refunded'? ¿Y cuál es el nombre más frecuente entre los clientes de esas direcciones?

Hipótesis: Se espera que los códigos postales con mayor concentración de devoluciones coincidan con zonas de alta densidad de clientes. Asimismo, es probable que los nombres más frecuentes en esas direcciones correspondan a los más comunes en la población general.

Conclusión: El análisis muestra que los cinco códigos postales con mayor número de órdenes reembolsadas son: 07266, 27351, 44223, 46959 y 70823. Dentro de ellos, los nombres más frecuentes de clientes asociados son Antonio, Amanda, Amber, Alicia y Jennifer. lo cual refleja patrones generales de la distribución de nombres en la base de clientes, sin evidenciar un sesgo particular hacia los reembolsos.

Nota: El código completo y las salidas de esta consulta se encuentran disponibles en el *notebook* de Google Colab adjunto.

2.3. Para cada tipo de pago y segmento de cliente, devolver la suma y el promedio expresado como porcentaje, de clientes activos y de consentimiento de marketing.

Hipótesis: Se espera que la proporción de clientes activos sea elevada en todos los segmentos y métodos de pago, dado que la mayoría de los usuarios de la plataforma mantienen su cuenta en uso.Y que el consentimiento de marketing muestre variaciones mayores: posiblemente más alto en clientes premium, que suelen estar más

expuestos a campañas personalizadas, y más bajo en segmentos budget u otros.

Conclusión: El porcentaje de clientes activos es efectivamente muy alto (alrededor del 89–90 %) y se mantiene estable en todos los segmentos y métodos de pago, sin diferencias significativas. En cuanto al consentimiento de marketing, los valores también son bastante homogéneos (≈ 69 - 70 %), con ligeras variaciones positivas en el segmento Regular (≈ 70,1 %) y Premium (≈ 70,0 %), mientras que Budget y Others rondan el ≈ 69,6 %.Esto indica que ni el método de pago ni el segmento de cliente generan diferencias sustanciales en la actividad o en la aceptación del marketing.

Nota: El código completo y las salidas de esta consulta se encuentran disponibles en el *notebook* de Google Colab adjunto.

2.4. Peso total de inventario por marca para productos con 'stuff'

Hipótesis: Se espera que las marcas con mayor volumen de productos en inventario sean también las que acumulen el mayor peso total. Es probable que marcas reconocidas como Nike, Adidas o 3M figuren entre las primeras posiciones.

Conclusión: El análisis muestra que las cinco marcas con mayor peso total de inventario asociado a productos que contienen la palabra "stuff" en su descripción son: 3M, Adidas, Nike, Hasbro y Wayfair. Estas concentran una parte significativa del stock medido en kilos y toneladas, lo que confirma la hipótesis de que las marcas con mayor variedad de productos y reconocimiento internacional tienden a dominar el inventario en términos de peso total.

Nota: El código completo y las salidas de esta consulta se encuentran disponibles en el *notebook* de Google Colab adjunto.

3. Consultas exploratorias complementarias

3.1. ¿Cuál es el patrón de ventas a lo largo de los meses del año?

Hipótesis: Se espera identificar el comportamiento general de las ventas completadas a lo largo de los meses del año.
Conclusión: El análisis muestra que durante el primer semestre las ventas siguen una tendencia creciente, alcanzando su máximo en el mes de junio. Posteriormente, se observa una caída abrupta al inicio del segundo semestre, particularmente en julio, con una leve recuperación a partir de agosto en adelante.

Nota: La visualización correspondiente a esta consulta se encuentra en la sección de Visualizaciones, Continua con linea de tiempo, figura 1.

3.2. Top 10 promedio de rating por categoría de productos activos

Hipótesis: Se espera que las categorías relacionadas con productos de uso cotidiano o de mayor valoración subjetiva, como indumentaria, libros y calzado, presenten un puntaje promedio de reseñas más alto. Asimismo, se anticipa que las categorías con productos de consumo masivo o de menor diferenciación muestren valoraciones más cercanas al promedio general.

Conclusión: El análisis de las categorías de productos activos muestra que el Top 10 de puntaje promedio de reseñas está encabezado por Handmade, Garden & Outdoor y Clothing, todas con valores cercanos a 4.0. Estos resultados confirman parcialmente la hipótesis: efectivamente, categorías asociadas a experiencias personales o de consumo cotidiano tienden a recibir mejores valoraciones. Sin embargo, también destacan rubros menos esperados, como Pet Supplies o Musical Instruments, lo que sugiere que la satisfacción del cliente no depende únicamente del tipo de producto, sino también de factores como la calidad percibida y la expectativa del consumidor.

Nota: La visualización correspondiente a esta consulta se encuentra en la sección de Visualizaciones, Visualizaciones a eleccion N°1, figura 6.

3.3. ¿Existe relación entre el precio del producto y la cantidad de votos útiles de sus reseñas?

Hipótesis: Se espera que los productos de mayor precio reciban más votos útiles en sus reseñas, ya que los clientes suelen estar más motivados a evaluar y valorar la utilidad de opiniones en compras de mayor inversión económica.

Conclusión: El análisis muestra que no existe relación entre el precio del producto y la cantidad de votos útiles de sus reseñas. La regresión lineal presenta una pendiente prácticamente nula y el coeficiente de correlación (aproximadamente - 0.002) confirma la ausencia de asociación significativa. Esto refuta la hipótesis inicial, indicando que el precio no influye en la utilidad percibida de las reseñas.

Nota: La visualización correspondiente a esta consulta se encuentra en la sección de Visualizaciones, Continua con Continua, figura 4.

3.4. ¿Qué segmento de clientes realiza más compras?

Hipótesis: Se espera que los clientes del segmento Premium concentren la mayor cantidad de compras, dado que su mayor poder adquisitivo les permitiría realizar transacciones con mayor frecuencia.

Conclusión: El análisis muestra que el segmento Regular es el que realiza la mayor cantidad de compras, seguido por Premium y Budget, mientras que Others representa la menor proporción. Esto contradice la hipótesis inicial, indicando que la mayor actividad de compra no proviene de los segmentos de mayor poder adquisitivo, sino de los clientes regulares, probablemente debido a su mayor volumen poblacional dentro de la base de clientes.

Nota: La visualización correspondiente a esta consulta se encuentra en la sección de Visualizaciones, Visualizaciones a eleccion N°2, figura 7.

3.5. ¿Cuál es el país con mayor cantidad de clientes activos y cuánto representan en el total?

Hipótesis: Se espera que Estados Unidos concentre la mayor cantidad de clientes activos, dado su tamaño de mercado y relevancia en el comercio electrónico a nivel global.

Conclusión: El análisis muestra que Brasil es el país con mayor cantidad de clientes activos (41.285), representando aproximadamente un 10,1 % del total. Le siguen Canadá e India con cifras muy similares. Esto contradice la hipótesis inicial, ya que Estados Unidos, a pesar de su peso económico, se ubica recién en la décima posición en términos de clientes activos dentro del dataset.

Nota: La visualización correspondiente a esta consulta se encuentra en la sección de Visualizaciones, Visualizaciones especial, figura 8.

3.6. ¿Cuál es la cantidad de clientes activos por genero por país?

Hipótesis: Se espera que en la mayoría de los países la distribución de clientes activos esté relativamente balanceada entre hombres y mujeres, con una proporción menor en otros generos.

Conclusión: El análisis confirma la hipótesis: en todos los países analizados, la cantidad de clientes activos masculinos y femeninos es muy similar (diferencias menores a 500 clientes), mientras que las personas que no se identifican en algunos de los dos generos principales, representa un volumen consistentemente menor (alrededor de 5.000 clientes por país). Esto indica que la distribución por género es homogénea a nivel global, sin grandes sesgos hacia un género específico, y que los clientes que no se identifican dentro de uno de estos dos generos tienen un peso significativamente inferior frente a femenino y masculino.

Nota: La visualización correspondiente a esta consulta se encuentra en la sección de Visualizaciones, Discreta con discreta, figura 3.

3.7. ¿Cuál es la cantidad de clientes por genero que realizaron compras por país?

Hipótesis: Se espera que la cantidad de compradores por género esté relativamente equilibrada entre hombres y mujeres en todos los países, mientras que la categoría "OTHER" represente un porcentaje considerablemente menor.

Conclusión: El análisis confirma la hipótesis: en todos los países la distribución entre hombres y mujeres es casi idéntica, con diferencias mínimas (generalmente menores a 100 compradores). La categoría "OTHER" mantiene una proporción menor, en torno al 12–14 % de los clientes, lo que refuerza la homogeneidad de la distribución de género en las compras a nivel global. Esto indica que el género no es un factor diferenciador en el volumen de compras dentro de los distintos países del dataset.

Nota: La visualización correspondiente a esta consulta se encuentra en la sección de Visualizaciones, Visualizacion extra N°2, figura 10.

3.8. Distribución de reseñas por calificación

Hipótesis: Se espera que la mayoría de las reseñas se concentren en puntajes altos (4 y 5), reflejando una tendencia positiva en la valoración de los productos, mientras que los puntajes bajos (1 y 2) representen una minoría.

Conclusión: El análisis confirma la hipótesis: la distribución de reseñas por puntaje está sesgada hacia valores altos, con predominio de 4 estrellas (49.757 reseñas) y 5 estrellas (43.073 reseñas). Los puntajes bajos son poco frecuentes: apenas 3.707 reseñas de 1 estrella y 8.603 de 2 estrellas. Esto evidencia que los usuarios, en general, manifiestan una percepción positiva de los productos, aunque las reseñas críticas de menor puntaje también aportan información relevante para detectar insatisfacción.

Nota: La visualización correspondiente a esta consulta se encuentra en la sección de Visualizaciones, Visualizacion extra N°3, figura 11.

4. Visualizaciones

4.1. Continua con línea de tiempo

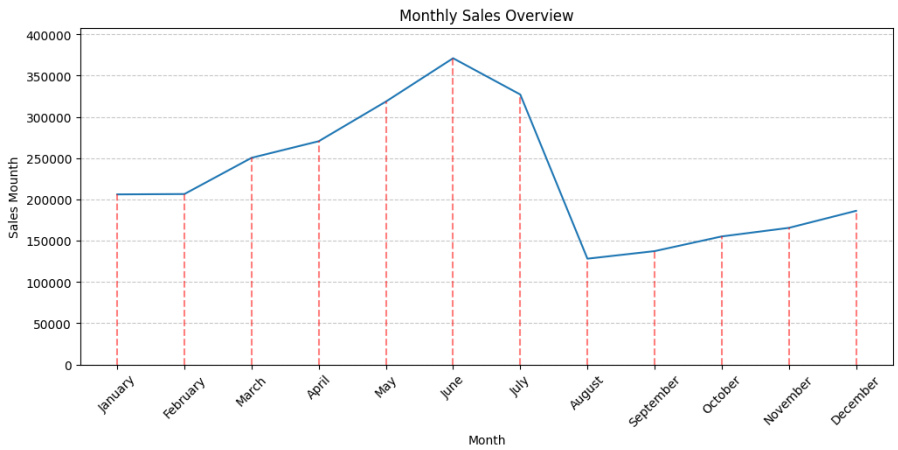

Figura 1: Ventas mensuales agregadas por mes.

4.2. Una discreta con una continua.

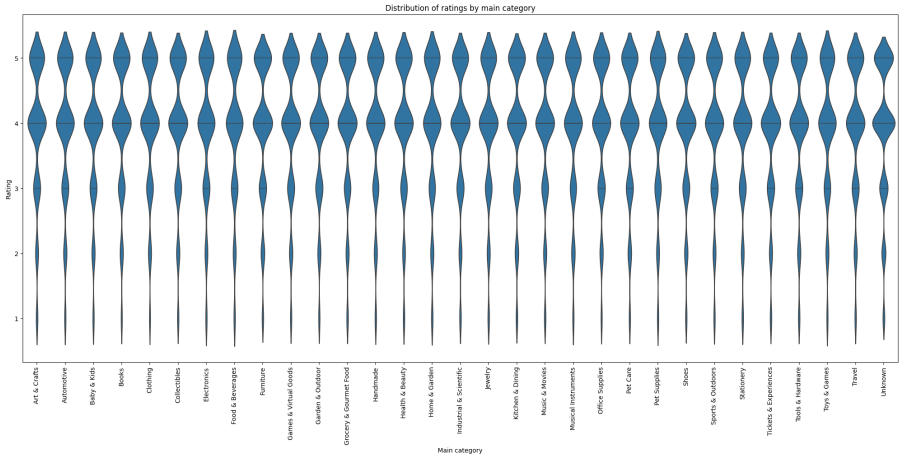

Figura 2: Distribución de calificaciones por categoría principal.

4.3. Discreta con discreta

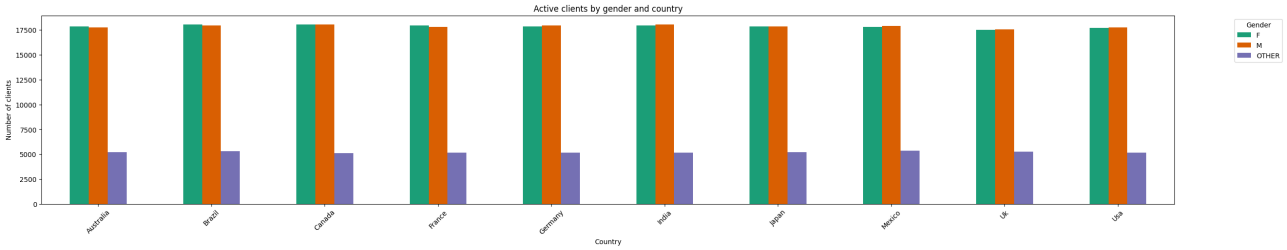

Figura 3: Clientes activos por género y país.

4.4. Continua con continua

Figura 4: Precio vs. votos útiles (con regresión lineal)

4.5. Heatmap

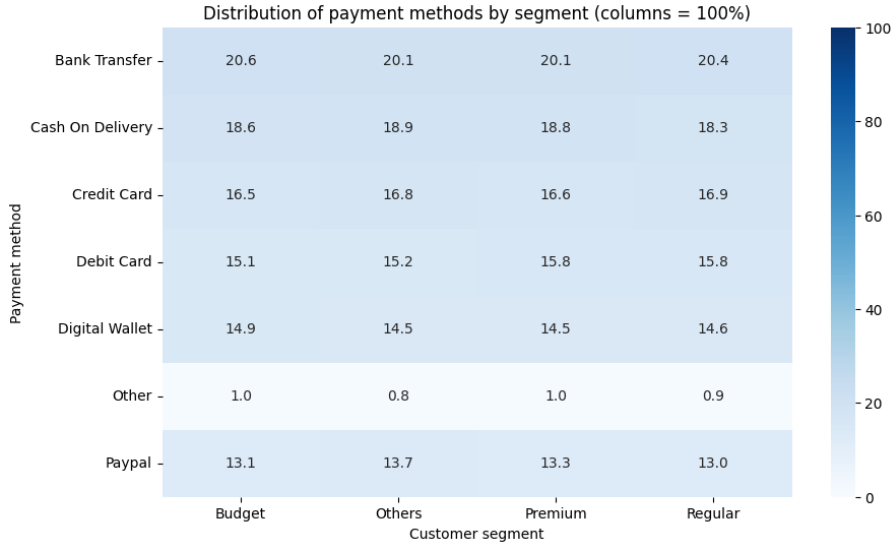

Figura 5: Distribución de métodos de pago por segmento .

4.6. Visualizaciones a elección N°1

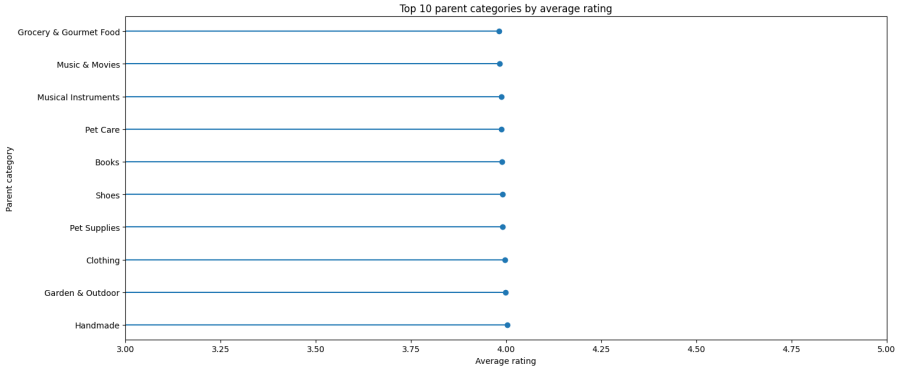

Figura 6: Las 10 categorías principales por calificación promedio

4.7. Visualizaciones a elección N°2

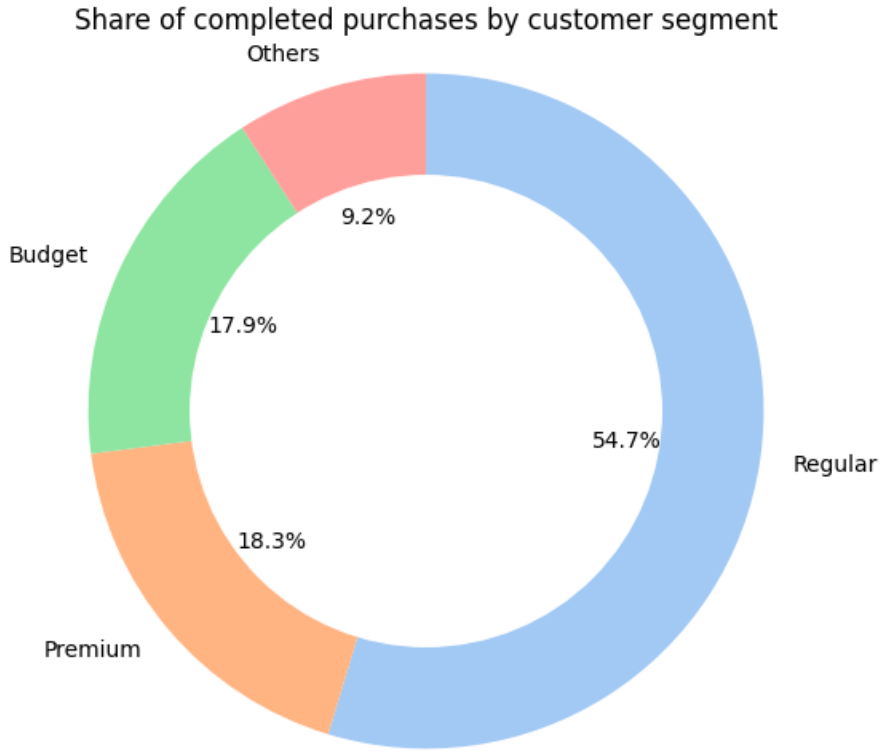

Figura 7: Porcentaje de compras completadas por segmento de clientes

4.8. Visualización especial

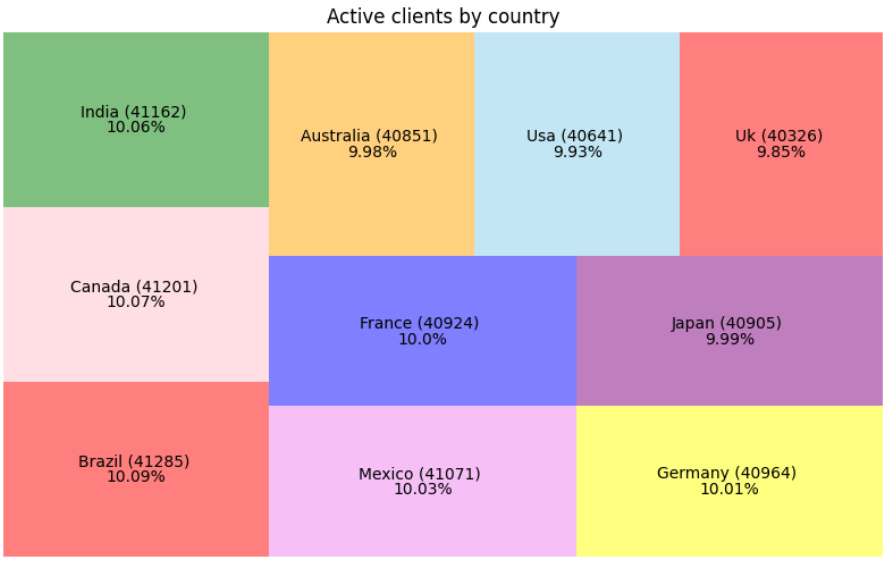

Figura 8: Clientes activos por país

4.9. Visualización extra N°1

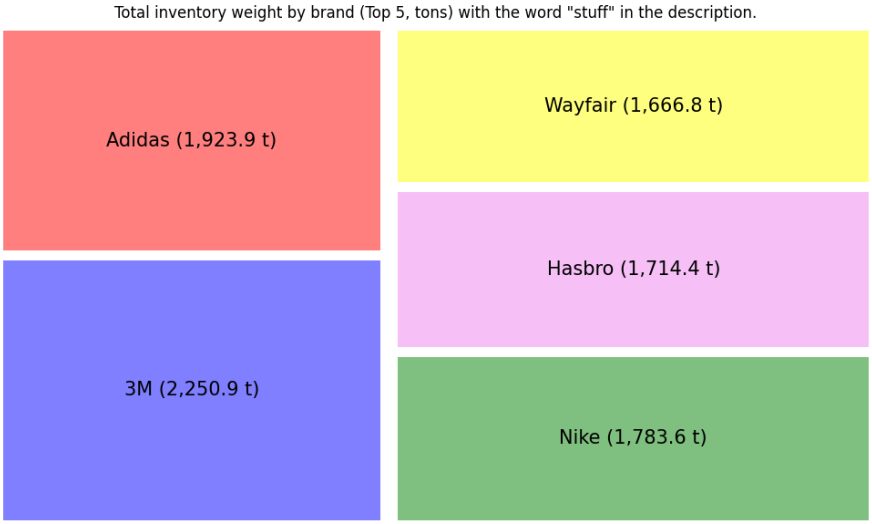

Figura 9: Peso total del inventario por marca (Top 5, toneladas) con la palabra `stuff` en la descripción.

4.10. Visualización extra N°2

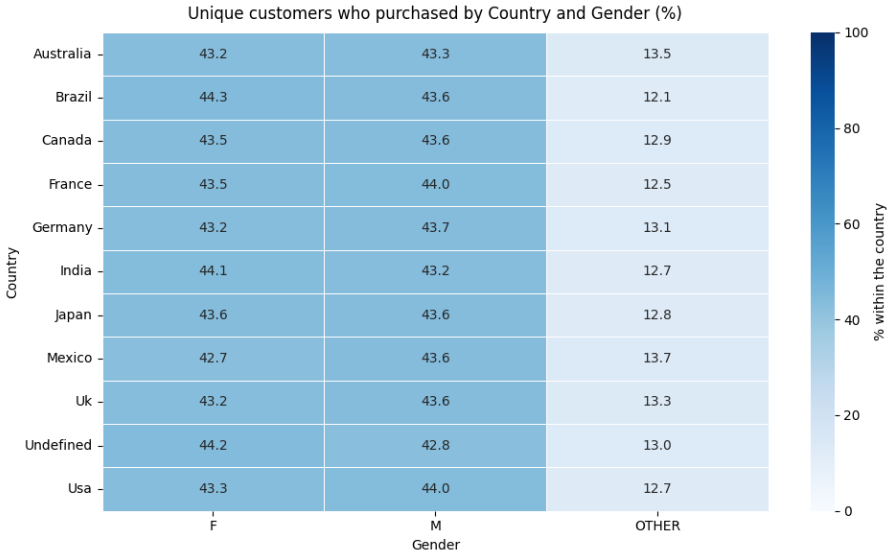

Figura 10: Clientes únicos que compraron por país y género (%)

4.11. Visualización extra N°3

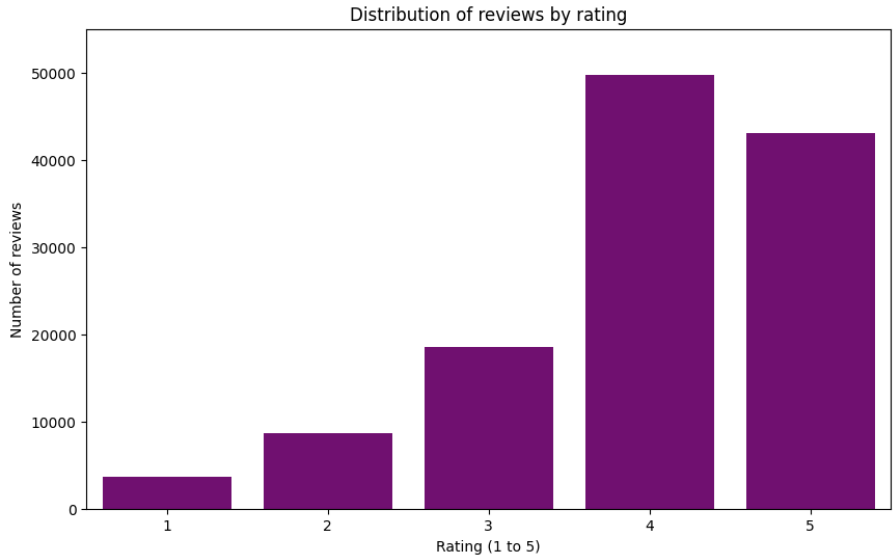

Figura 11: Distribución de reseñas por calificación

Herramientas y librerias utilizadas

- `Python 3`, `Google Colab`.
- Librerías: `pandas`, `matplotlib`, `seaborn`, `numpy`, `plotly.express`, `squarify` `calendar`.
- Documento elaborado con Overleaf.

Enlace al repositorio: https://github.com/MaurizioG28/TP_CIENCIA_DATOS.git