**Taller Final Análisis Probabilístico**

Alexander Velez

Camilo Muñoz

Tecnológico De Antioquia

Facultad De Ingeniería

Julián Castelblanco Benítez

Mayo de 2025

Índice

1. Introducción .......................................................................... 1

2. Objetivos ............................................................................ 2

3. Análisis Descriptivo ............................................................... 3

3.1 Inspección inicial y limpieza de datos......................................................... 3

3.2 Estadísticas descriptivas ......................................................... 3

3.3 Análisis de ventas por mes .............................................. 4

3.4 Análisis de ventas por categoría .................................................. 4

3.5 Top 5 ciudades con más ventas.................................................. 5

4. Intervalos de Confianza por Categoría ...................................................... 6

5. Pruebas de Hipótesis ......................................................... 7

6. Conclusiones ...................................................................... 8

7. Referencias ......................................................................... 9

8. Anexos ................................................................................ 10

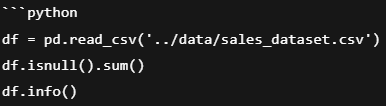
´

**3. Análisis Descriptivo**

Para iniciar el análisis probabilístico se trabajó con un conjunto de datos reales de ventas contenido en el archivo `sales\_dataset.csv`. En esta etapa se realizó una inspección preliminar del dataset, una limpieza básica, y un análisis descriptivo a través de estadísticas y visualizaciones.

**3.1 Inspección inicial y limpieza de datos**

El conjunto de datos se cargó con la librería `pandas`, y se revisaron los valores faltantes con `isnull().sum()`, identificando columnas potencialmente incompletas. Posteriormente, se verificó la estructura del dataset usando `info()`, lo cual permitió confirmar los tipos de datos de cada columna.



Esta revisión inicial permitió identificar si era necesario limpiar o transformar los datos antes del análisis. Aunque no se aplicaron técnicas avanzadas de imputación, se aseguró que las columnas clave (Amount, Profit, Quantity) estuvieran completas para el análisis.

Para comenzar el análisis, se cargó el archivo sales\_dataset.csv con la librería pandas y se inspeccionaron las primeras filas del dataset. Esto permitió verificar la estructura de los datos, identificar las variables disponibles y confirmar que se encuentran organizadas adecuadamente.

A continuación, se verificó la existencia de valores faltantes en las columnas. El resultado mostró que **no hay datos nulos en ninguna de las 12 columnas**, lo cual facilita un análisis sin necesidad de aplicar técnicas de imputación o limpieza adicional.

También se ejecutó el método info() para revisar la estructura del DataFrame. Se confirmó que el dataset contiene **1.194 registros** distribuidos en columnas de tipo numérico (int64) y categórico (object). No se detectaron anomalías en los tipos de datos ni en la cantidad de registros.

Este paso es fundamental para asegurar que los análisis posteriores se realicen sobre una base de datos íntegra y confiable.

**3.2 Estadísticas descriptivas**

Se calcularon estadísticas básicas de las columnas cuantitativas utilizando describe(), con foco en las variables Amount, Profit y Quantity:



Esto permitió obtener medidas como la media, mediana, desviación estándar, valores mínimos y máximos, fundamentales para entender la dispersión y tendencias generales de los datos. La variable Amount representa el monto total de cada venta, y Profit el beneficio correspondiente.

Se analizaron las variables numéricas clave: Amount (monto de venta), Profit (ganancia) y Quantity (cantidad vendida). Los estadísticos obtenidos fueron:

| **Variable** | **Mínimo** | **Máximo** | **Media** | **Desviación estándar** |
| --- | --- | --- | --- | --- |
| Amount | 508 | 9992 | 5178 | 2805 |
| Profit | 50 | 4930 | 1349 | 1118 |
| Quantity | 1 | 20 | 10.7 | 5.78 |

Se observa que el monto promedio de venta es de aproximadamente **5.178 unidades monetarias**, con una gran variabilidad (±2805). Lo mismo ocurre con la ganancia, que tiene una media de **1.349** pero una desviación alta, lo que sugiere que hay transacciones con ganancias muy dispares. La cantidad promedio vendida es de 11 unidades por transacción.

**3.3 Análisis de ventas por mes**

Se agruparon las ventas por mes utilizando la columna Year-Month, sumando los montos (Amount). Esto permite observar estacionalidades o picos de ventas:

Texto

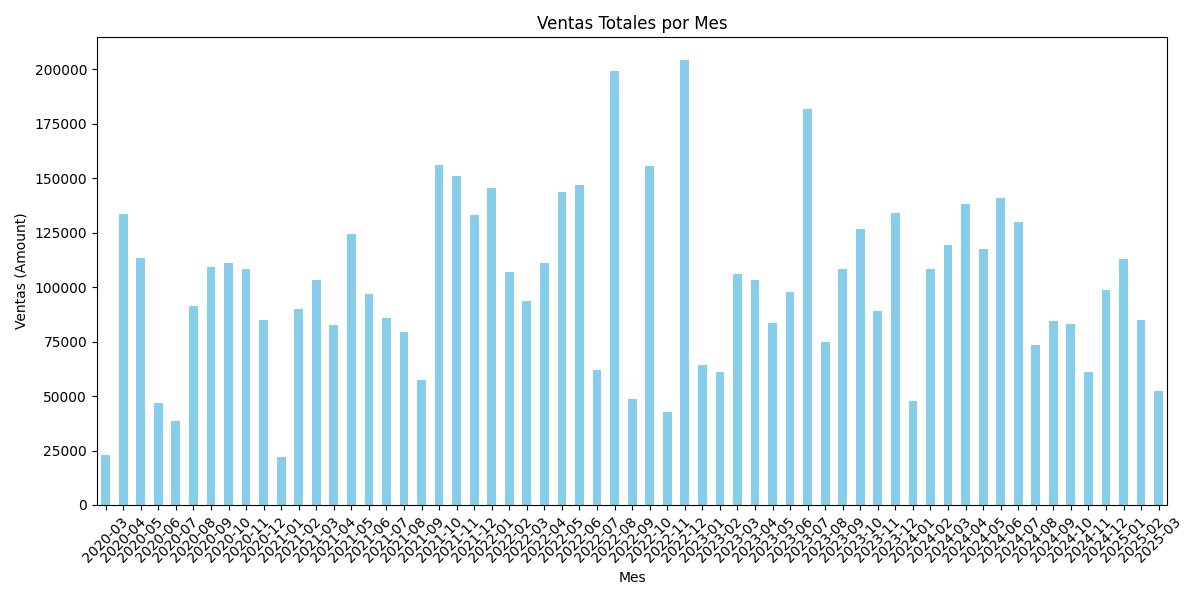
El contenido generado por IA puede ser incorrecto.

Para identificar patrones temporales, se agruparon los datos por la columna Year-Month y se sumaron los valores de Amount.

El análisis reveló que los meses con mayores ventas fueron **abril de 2020, enero de 2025 y diciembre de 2024**, con montos superiores a los 100.000. Estos picos pueden asociarse a campañas comerciales, temporadas de alta demanda o cierre de año fiscal.

Este tipo de información permite planificar mejor la producción, el inventario y las campañas de marketing a futuro.

**Figura 1**

****  
Ventas totales por mes (agrupadas por columna Year-Month).  
*Se observa una variación considerable entre meses, lo cual podría estar relacionado con campañas promocionales, temporadas altas o decisiones estratégicas internas de la empresa.*

**3.4 Análisis de ventas por categoría**

Se exploró el comportamiento de las ventas según la categoría del producto (Category). Los resultados se visualizaron mediante un gráfico circular (pie chart):

Texto

El contenido generado por IA puede ser incorrecto.

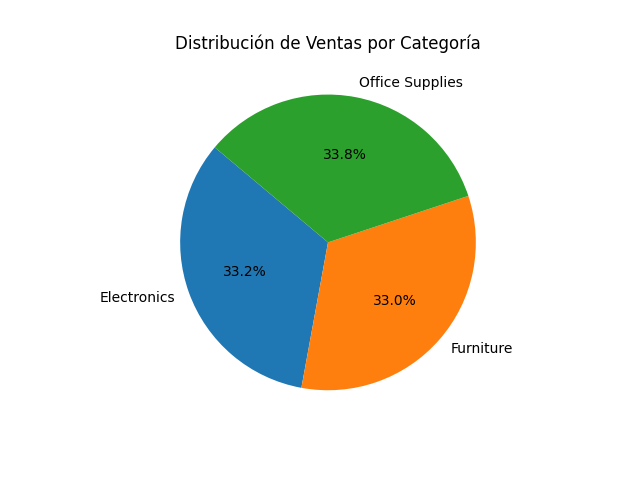
Se agruparon las ventas por categoría de producto, obteniendo los siguientes totales:

* **Office Supplies**: 2.089.510
* **Electronics**: 2.054.456
* **Furniture**: 2.038.673

Aunque las tres categorías están relativamente equilibradas, se destaca que **“Office Supplies”** lidera con una leve ventaja. Esto podría deberse a una mayor rotación de productos o a una mayor cantidad de ventas de menor valor unitario.

Este análisis es útil para priorizar categorías en promociones o reestructurar inventarios.

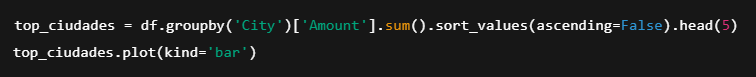
**Figura 2**

****

Distribución de ventas por categoría.  
*Este análisis ayuda a identificar qué tipo de productos representan la mayor parte del ingreso total. Puede usarse para tomar decisiones sobre inventario o marketing.*

**3.5 Top 5 ciudades con más ventas**

Finalmente, se identificaron las cinco ciudades con mayor volumen de ventas, lo cual permite conocer dónde se concentran los clientes más activos o rentables:

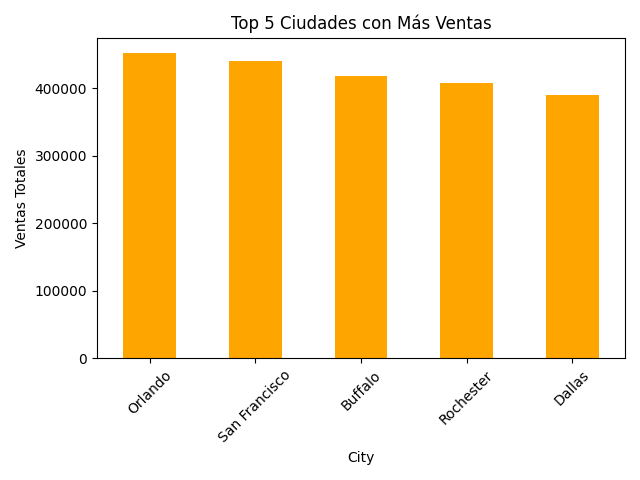


Finalmente, se agruparon las ventas por ciudad y se identificaron las cinco con mayores ingresos:

1. **Orlando** – 452.158
2. **San Francisco** – 440.000
3. **Buffalo** – 418.514
4. **Rochester** – 407.291
5. **Dallas** – 390.144

Estas ciudades representan los principales focos de ingreso para la empresa. Este tipo de hallazgo es crucial para la planificación geográfica de estrategias comerciales, expansión de tiendas físicas o enfoque de marketing.

**Figura 3**



Top 5 ciudades con mayores ventas totales.  
*Este tipo de análisis geográfico puede orientar la toma de decisiones logísticas o comerciales, como la ubicación de nuevos puntos de venta o campañas regionales.*

**4. – Intervalos de Confianza por Categoría**

**Descripción del Análisis**

En este apartado se calcularon intervalos de confianza del 95% para tres variables cuantitativas del dataset: Amount, Profit y Quantity, considerando cada una de las categorías de producto existentes en la columna Category: Furniture, Electronics y Office Supplies.  
  
El propósito de este análisis es estimar con qué grado de confianza podemos afirmar los valores promedio (medias) de estas variables en cada categoría y, a partir de ello, determinar cuál categoría es más rentable en términos de ganancia (Profit).

**Explicación del Código**

El código implementa una función general llamada calcular\_intervalo\_confianza que permite calcular el intervalo de confianza para cualquier serie de datos numéricos. Posteriormente, se aplica esta función a cada combinación de variable y categoría para mostrar los resultados.  
  
Paso a paso:

1. Limpieza de datos (dropna): Se eliminan valores nulos para evitar distorsiones en el cálculo.  
2. Estadísticas básicas:  
 - n: número de datos válidos.  
 - media: promedio muestral.  
 - std: desviación estándar muestral (ddof=1).  
3. Valor crítico t: Se obtiene a partir de la distribución t de Student para un nivel de confianza del 95%.  
4. Cálculo del margen de error: usando la fórmula t \* (std / sqrt(n)).  
5. Intervalo de confianza: se retorna como (media, media - ME, media + ME).

Luego, se aplica esta función a cada categoría y cada variable del dataset, imprimiendo la media y los límites inferior y superior del intervalo.

**Resultados Obtenidos**

Intervalos de confianza calculados al 95%:

Para Amount:

• Electronics: Media = 5294.99, IC 95% = [5017.36, 5572.62]

• Office Supplies: Media = 5236.87, IC 95% = [4954.42, 5519.31]

• Furniture: Media = 5009.02, IC 95% = [4739.95, 5278.10]

Para Profit:

• Electronics: Media = 1336.55, IC 95% = [1228.98, 1444.11]

• Office Supplies: Media = 1382.39, IC 95% = [1268.17, 1496.62]

• Furniture: Media = 1328.11, IC 95% = [1219.45, 1436.78]

Para Quantity:

• Electronics: Media = 10.97, IC 95% = [10.40, 11.55]

• Office Supplies: Media = 10.14, IC 95% = [9.56, 10.72]

• Furniture: Media = 10.91, IC 95% = [10.36, 11.46]

**Interpretación de Resultados**

En la variable Profit, que es el indicador principal de rentabilidad, se observa lo siguiente:  
  
• La categoría más rentable en promedio es Office Supplies, con una media de ganancias de 1382.39.  
• Aunque las diferencias no son muy grandes entre categorías, el intervalo de Office Supplies está ligeramente por encima de los de Electronics y Furniture.  
• No hay una separación drástica entre intervalos, lo que sugiere que no hay una diferencia estadísticamente significativa muy marcada, pero se puede tomar a Office Supplies como la opción más rentable hasta este punto del análisis.

Punto 2.2 Prueba de Hipótesis — Estado más rentable  
**Paso 1: Determinar la categoría más rentable**

Calcularemos la rentabilidad total (Profit) por categoría para elegir la más rentable.

Luego, en esa categoría, compararemos las ganancias (Profit) entre los estados **Illinois**, **New York** y **California**.

**Categoría más rentable: Office Supplies**

Es la que tiene mayor ganancia total (Profit = 551,575).

**Paso 2: Preparar la prueba de hipótesis**

Vamos a comparar las ganancias (Profit) de la categoría **Office Supplies** entre los estados:

* **Illinois**
* **New York**
* **California**

**Hipótesis:**

* **Hipótesis nula (H₀):** No hay diferencia significativa en las ganancias medias de "Office Supplies" entre los tres estados.
* **Hipótesis alternativa (H₁):** Al menos uno de los estados tiene una media de ganancias significativamente diferente.

**Nivel de significancia:**

* α = 0.05

**Método de prueba:**

Como se comparan más de dos grupos, aplicaremos un **ANOVA de un factor (One-Way ANOVA)**.

**Resultados del ANOVA:**

* **Estadístico F:** 1.157
* **Valor p:** 0.316

**Conclusión:**

Dado que el **valor p = 0.316** es **mayor** que el nivel de significancia α = 0.05:

✅ **No se rechaza la hipótesis nula**.

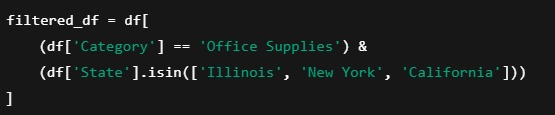
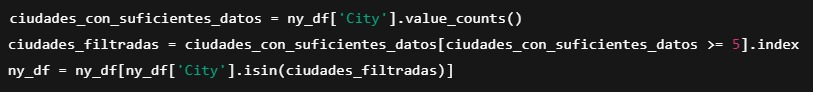
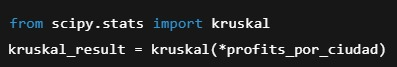
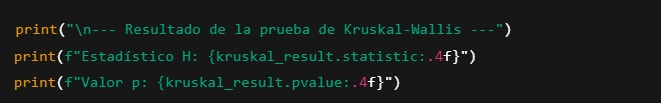
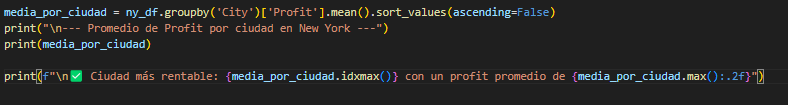
🔍 **Interpretación:** No hay evidencia estadísticamente significativa para afirmar que existen diferencias en las ganancias medias de la categoría **"Office Supplies"** entre los estados de **Illinois, New York y California**.

Punto 2.3 – Análisis de rentabilidad por ciudad (sin asumir normalidad)

Objetivo del análisis:

El objetivo es determinar qué ciudad dentro del estado más rentable (identificado previamente como New York) genera el mayor beneficio ("Profit") promedio, sin asumir que los datos siguen una distribución normal. Para ello, se aplica una prueba de hipótesis no paramétrica (Kruskal-Wallis) para comparar las ciudades entre sí.

paso a paso del análisis con explicación del código:

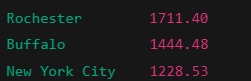
Paso 1: Filtrado de los datos relevantes  
  
Se filtran solo las filas que corresponden a la categoría más rentable (Office Supplies) y los tres estados más rentables identificados previamente: New York, Illinois y California.  
Paso 2: Cálculo del profit promedio por estado  
  
Calculamos la ganancia promedio (Profit) por estado para confirmar cuál es el más rentable. El resultado indica que New York es el estado con mayor profit promedio.  
Paso 3: Filtrar datos solo del estado más rentable  
  
Se filtran las filas correspondientes únicamente al estado New York y a la categoría seleccionada.  
Paso 4: Filtrar ciudades con suficientes datos  
  
Solo se consideran las ciudades con al menos 5 registros, para asegurar que la comparación entre ciudades sea estadísticamente significativa.  
Paso 5: Preparación de datos para la prueba de hipótesis  
  
Se agrupan los profits de cada ciudad en listas separadas, para pasarlas a la prueba de Kruskal-Wallis.  
Paso 6: Prueba de Kruskal-Wallis  
  
Se utiliza la prueba Kruskal-Wallis, una alternativa no paramétrica a la ANOVA, ideal para comparar múltiples grupos sin asumir normalidad  
  
Se imprimen los resultados de la prueba. Si el valor-p < 0.05, existen diferencias significativas entre las ciudades en cuanto al profit.  
Paso 7: Determinar ciudad más rentable  
  
Resultados obtenidos

Estado más rentable: New York

(Profit promedio: $1,485.60)

Estadístico H de Kruskal-Wallis: 0.5868

Valor p: 0.7457

Media de profit por ciudad en New York:  
  
Ciudad más rentable: Rochester

(Profit promedio: $1,711.40)  
Conclusión

Dado que el valor p (0.7457) es mayor al nivel de significancia típico de 0.05, no se rechaza la hipótesis nula. Es decir, no hay evidencia estadísticamente significativa para afirmar que hay diferencias reales en el profit entre las ciudades analizadas.

Sin embargo, Rochester presenta el mayor promedio de ganancia, por lo que podría considerarse la ciudad más rentable dentro del estado de New York para la categoría Office Supplies, desde un punto de vista descriptivo.