Laboratorio de Modelos Analíticos para la Toma de Decisiones

ISTEA

Trabajo Práctico Final

Alumno: Federico Gauna

Profesora: Stephanie Montiel Gallardo

Caso de Estudio

La empresa de retail nacional “Musimundo”, líder en la venta de productos electrónicos, ha decidido expandirse a nuevos mercados regionales. Actualmente opera en cinco ciudades principales y desea extender su presencia a tres nuevas ciudades donde aún no tiene operaciones. La dirección necesita una estrategia basada en datos para maximizar la rentabilidad y minimizar los riesgos asociados con esta expansión.

Estas cinco ciudades principales donde tiene mayor presencia son CABA y otras localidades dentro del conurbano. Por eso, lo que se está buscando es expandirse a ciudades del interior de Buenos Aires, y de esa forma ir armando una red de sucursales que abarque gran parte de la provincia.

Objetivos

El objetivo es desarrollar una propuesta integral para la expansión, que combine modelos analíticos para optimizar la toma de decisiones.

Dónde expandirse: Identificar las ciudades con mayor potencial basándose en el análisis de patrones históricos y características socioeconómicas.

Segmentación de mercado y pricing: Definir las estrategias de precios y segmentación que mejor se adapten a las nuevas ubicaciones.

Predicción de demanda: Estimar las ventas futuras para decidir niveles óptimos de inventario y recursos.

Optimización de la cadena de suministro: Planificar la logística para asegurar entregas eficientes sin incurrir en sobrecostos.

¿Qué conclusiones pueden sacar?

Exploración de los Datos

La empresa tiene datos históricos de ventas, comportamiento de clientes, campañas de marketing y operaciones logísticas de sus mercados actuales.

La database con la que se va a trabajar cuenta con 600 filas y 8 columnas. Los primeros 5 registros se ven así:

	City	Date	Sales	Marketing_Expense	Conversion_Rate	Avg_Basket_Size	Customer_Satisfaction	Logistic_Cost
0	City_1	2018-01-31	47410.59	10536.56	0.063255	168.47	100.000000	16263.58
1	City_1	2018-02-28	70701.37	7631.35	0.057410	85.94	83.910106	9325.61
2	City_1	2018-03-31	49616.69	9944.62	0.067591	224.15	91.575481	11029.27
3	City_1	2018-04-30	7271.86	16859.02	0.044563	243.88	63.204952	9613.87
4	City_1	2018-05-31	30579.78	20469.46	0.067412	76.17	79.296191	9304.71

Columnas

City	object
Date	object
Sales	float64
Marketing_Expense	float64
Conversion_Rate	float64
Avg_Basket_Size	float64
Customer_Satisfaction	float64
Logistic_Cost	float64

City: Nombre de la ciudad. Objeto.
Date: Fecha de cierre. Objeto.
Sales: Ventas generadas. Float64.
Marketing_Expense: Gastos en marketing. Float64.
Conversion_Rate: Tasa de conversión. Float64.
Avg_Basket_Size: Promedio de gasto por cliente. Float64.
Customer_Satisfaction: Índice de satisfacción de los clientes. Float64.
Logistic_Cost: Costos logísticos. Float64.

Con esta exploración ya tenemos una idea general de qué tipo de datos tenemos y para qué sirve cada uno de ellos. Antes de comenzar a analizarlos, se realizará un pequeño trabajo de limpieza y transformación.

La database no tiene datos nulos ni datos duplicados.

```
Datos faltantes:
City                    0
Date                    0
Sales                   0
Marketing_Expense       0
Conversion_Rate         0
Avg_Basket_Size         0
Customer_Satisfaction   0
Logistic_Cost           0
dtype: int64

 Datos duplicados:
0
```

Se transformará la columna “date” para pasarla de object a date.

```
df['Date'] = pd.to_datetime(df['Date'])
print("Tipo después de la conversión:", df['Date'].dtype)

Tipo después de la conversión: datetime64[ns]
```

Por último, se van a reemplazar los nombres de las ciudades de “City_X” a su nombre correspondiente.

```
Valores de "City":
['Zárate' 'Campana' 'Pilar' 'Escobar' 'Mercedes' 'Cañuelas'
 'Capilla del Señor' 'Luján' 'Gral. Rodríguez' 'Marcos Paz']
```

Dónde expandirse

Para abordar con este primer punto, se tiene que identificar las ciudades con mayor potencial basándose en el análisis de patrones históricos y características socioeconómicas.

Se analizarán las variables **"Sales"** (ventas) y **"Customer_Satisfaction"** (satisfacción del cliente). Estas variables son fundamentales porque, si bien las ventas son un buen indicador del desempeño, la satisfacción del cliente es la clave para asegurar la fidelidad de los clientes existentes y atraer nuevos en el futuro. Incluso en momentos en los que las ventas puedan bajar, tener una alta satisfacción te va a permitir tener una buena base de clientes sobra la cual se podrá aplicar las estrategias de marketing necesarias para volver a aumentar las ventas.

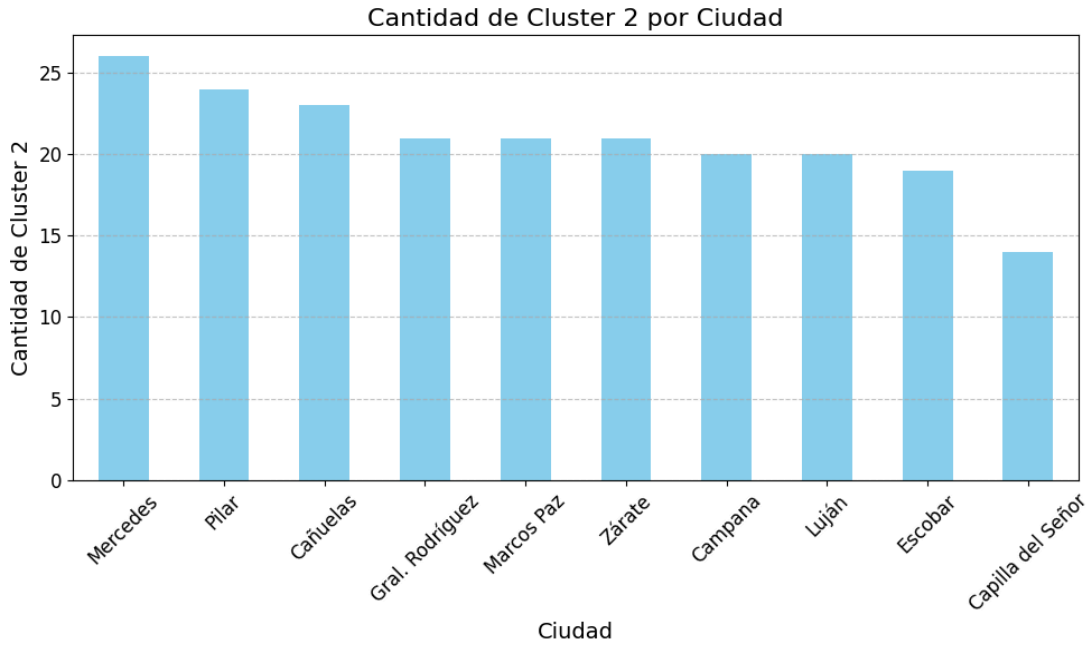
El método seleccionado será un modelo de aprendizaje no supervisado de clustering con K-Means. Este modelo agrupa los registros en clústeres basados en similitudes de los valores de las variables elegidas. Esto permite segmentar las observaciones según patrones comunes en ventas y satisfacción del cliente. Una vez aplicado el modelo, se comparan los cluster analizándolos mediante ciertas métricas, en este caso, el **promedio de ventas** y el **promedio de satisfacción del cliente** dentro de cada clúster. Por último, se seleccionarán las ciudades pertenecientes a los clústeres con mejor desempeño, priorizando aquellos que representan tanto un alto nivel de ventas como una elevada satisfacción del cliente.

```
Promedio de ventas por cluster:
 Cluster
0    65838.329478
1    31993.311810
2    56698.826651
3    42646.019375
```

```
Promedio de índice de satisfacción al cliente por cluster:
 Cluster
0    67.394867
1    89.319717
2    91.376148
3    67.619058
```

Resulta que el cluster 2 es el más performante, ya que es el segundo con mejor promedio de ventas y es el que mejor índice de satisfacción tiene.

Ahora lo que toca es contabilizar la cantidad de clusters que tiene cada ciudad.



Vemos la cantidad de “cluster 2” que hay en cada ciudad y las ordenamos de forma descendente para elegir las tres mejores.

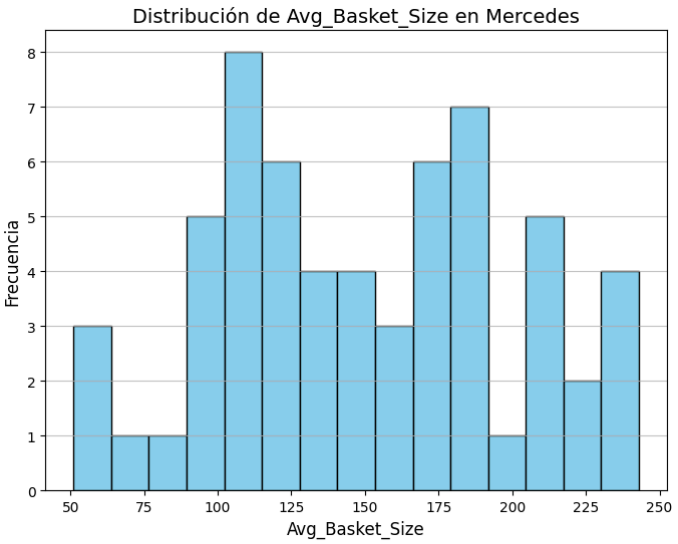
Como resultado obtenemos que “Mercedes”, “Pilar” y “Cañuelas” son las ciudades con más “cluster 2”. Esto significa que van a tener una cantidad de ventas alta en el cierre de mes y que tienden a tener una alta satisfacción al cliente, características que serán de utilidad para asegurar una sólida base de clientes en estas nuevas localidades.

Segmentación de mercado y pricing

Para esta parte del análisis, se tiene que definir las estrategias de precios y segmentación que mejor se adapten a las nuevas ubicaciones.

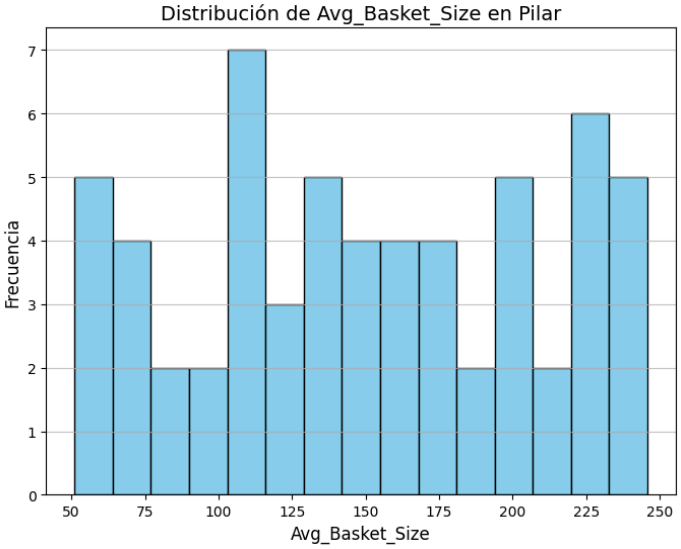
Ya contamos con las 3 ciudades separadas cada una en un dataframe para poder realizar sobre ellas el trabajo de análisis correspondiente. Las columnas a trabajar ahora serán “Avg_Basket_Size”, “Conversion_Rate” y “Marketing_Expense”. Con estas, lo que se busca es hacer una segmentación entre las ciudades, medirlas a través de distintas métricas con el objetivo de ponderarlas, encontrar patrones y decidir qué tipo de estrategias de pricing se podrían aplicar.

Primero vamos a visualizar en histogramas la distribución del “Avg_Basket_Size” en cada ciudad para ir conociendo el comportamiento de los clientes . En cada ciudad también se utilizó estadística descriptiva para conocer los cuartiles y etiquetar a cada registro con un mote de “Bajo”, “Medio” y “Alto” según su valor.



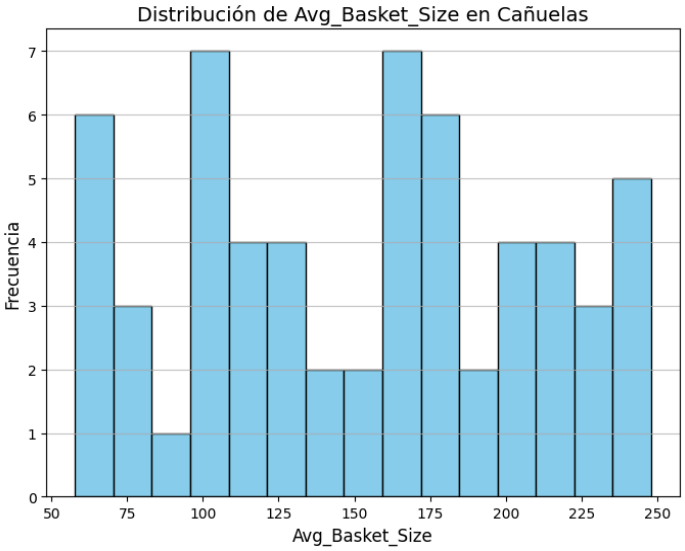
	indice_gasto	Count
0	Alto	15
1	Bajo	15
2	Medio	30

Mediana de Avg_Basket_Size 151.01999999999998



	indice_gasto	Count
0	Alto	15
1	Bajo	15
2	Medio	30

Mediana de Avg_Basket_Size 146.115



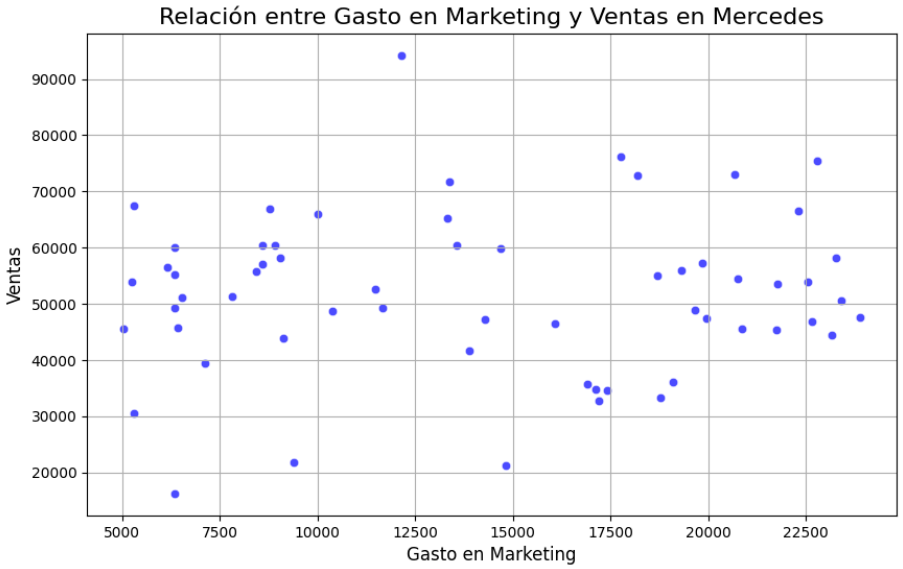
	indice_gasto	Count
0	Alto	15
1	Bajo	15
2	Medio	30

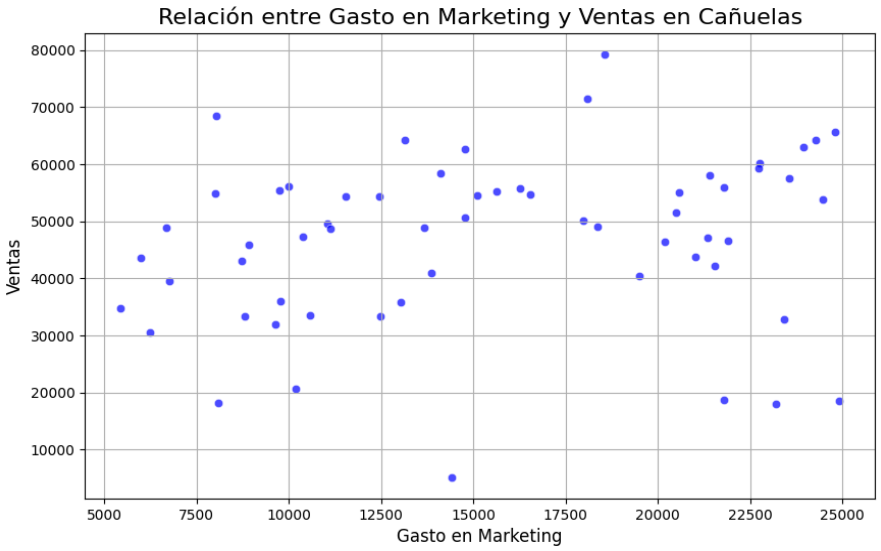
Mediana de Avg_Basket_Size 163.0

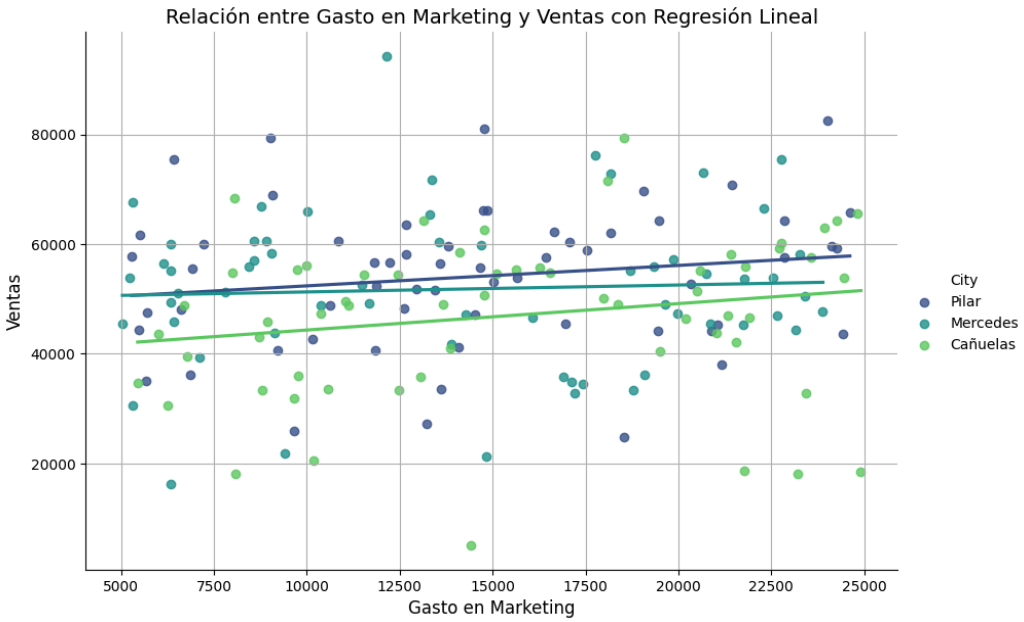
Con estas visualizaciones, se puede concluir que las tres ciudades tienen una distribución de gastos por cliente bastante similar. Los puntos a destacar que nos permiten diferenciar estas ciudades son:

- Pilar es la ciudad con una distribución más uniforme entre los valores 100 y 180, los cuales se pueden considerar como valores altos.
- Tanto Mercedes como Cañuelas tienen una distribución más variable entre este rango de valores, estando la mayoría concentrados alrededor de 100 y 180.
- Cañuelas tiene un detalle que las otras no, y es que tiene muchas ocurrencias en valores bajos. También acotar que Pilar tiene mayor cantidad de ocurrencias en valores pasados los 200.

Teniendo esto en cuenta, se aplicará un modelo de regresión lineal para ver la correlación entre “Marketing_Expense” y “Sales. El objetivo es ver que tanto afecta a las ventas los gastos en marketing que se hacen en cada ciudad.







En los tres casos ocurre lo mismo, obtenemos una recta bastante cercana a ser horizontal. Que quiere decir esto, que estas variables están poco correlacionadas. Un alto nivel de marketing no hace una gran diferencia en los montos de venta. Esto podría significar que hay posibilidad de ahorro y que se busquen estrategias donde el monto a gastar esté más cerca del promedio.

Con estos gráficos se puede concluir que ciudades como Pilar pueden aguantar estrategia de precios más agresivas, mientras que Mercedes y Cañuelas necesitan estrategias más reservadas, especialmente Cañuelas en donde se puede buscar fomentar la venta de productos de bajo precio.
Por otro lado, se tiene que revisar los gastos de marketing ya que valores altos no están generando una gran diferencia en las ventas.

Predicción de demanda

Estimar las ventas futuras para decidir niveles óptimos de inventario y recursos.

Para hacer la predicción de la demanda se van a combinar dos técnicas: un modelo predictivo de árbol de regresión y un análisis gráfico de tendencias temporales. Para el modelo de regresión, las variables que se usaran son “Marketing_Expense”, “Conversion_Rate”, “Avg_Basket_Size”, “Customer_Satisfaction” y “Logistic_Cost” como variables independientes, y “Sales” como variable dependiente.

Mercedes

valores reales	valores predichos
60482.43	57104.48
73005.92	55100.77
39356.30	51040.17
46955.29	44442.38
32776.05	48992.33
75499.36	53591.07
45297.05	53591.07
21851.71	55156.82
30506.28	45573.65
72789.26	47384.69
49357.65	65899.05
59772.83	46553.01

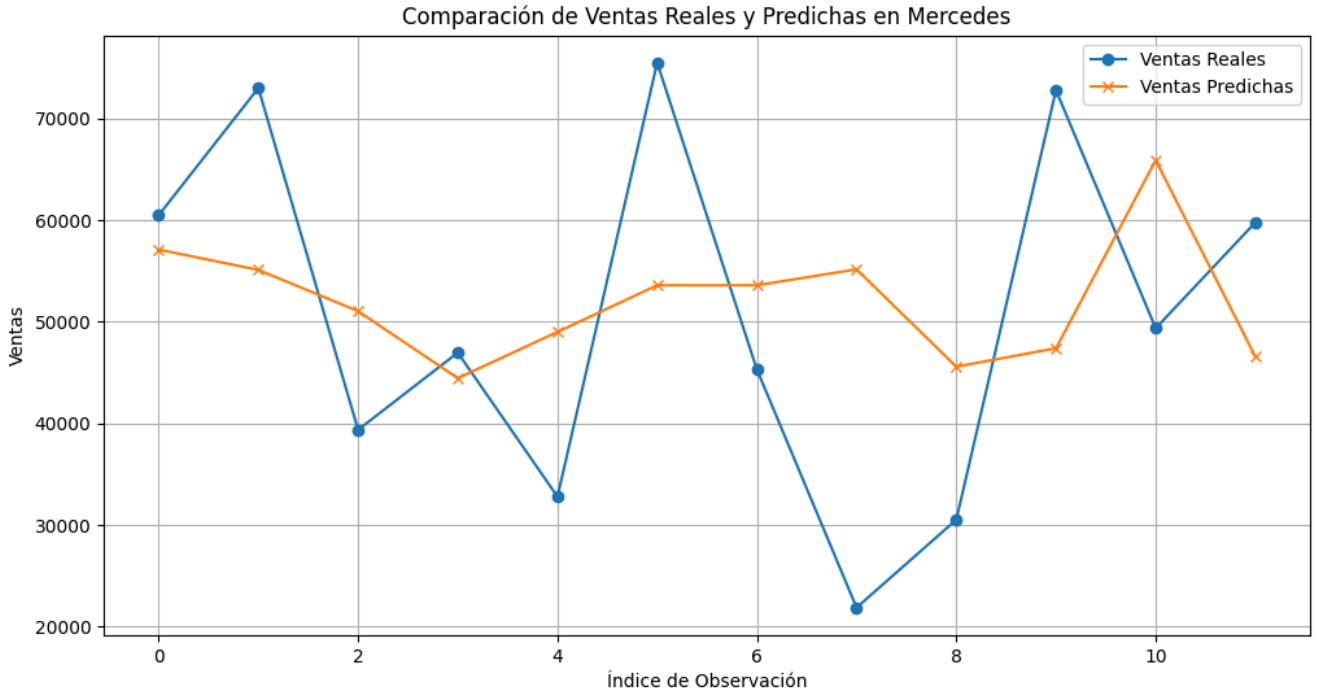
Pilar

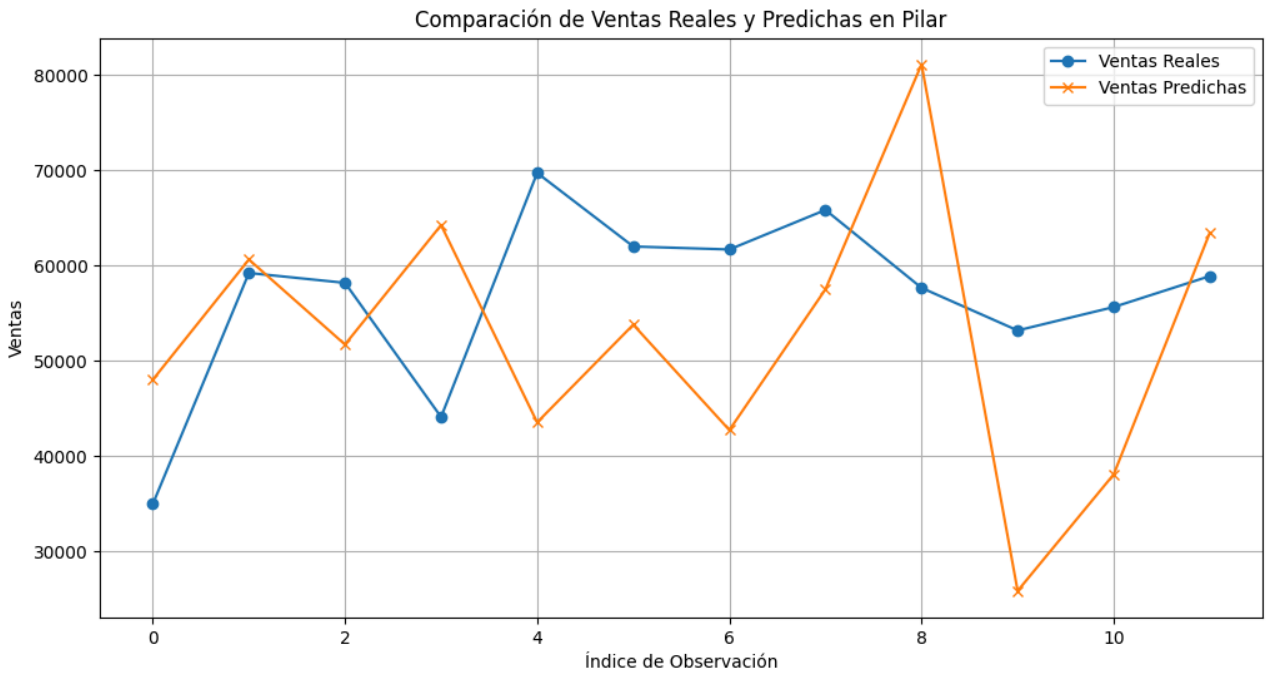
valores reales	valores predichos
34995.03	48031.15
59230.51	60641.78
58194.26	51725.39
44125.27	64254.61
69723.72	43560.47
62006.14	53806.09
61694.91	42754.08
65840.85	57516.41
57668.04	81111.24
53199.41	25886.60
55656.51	38113.70
58903.35	63452.59

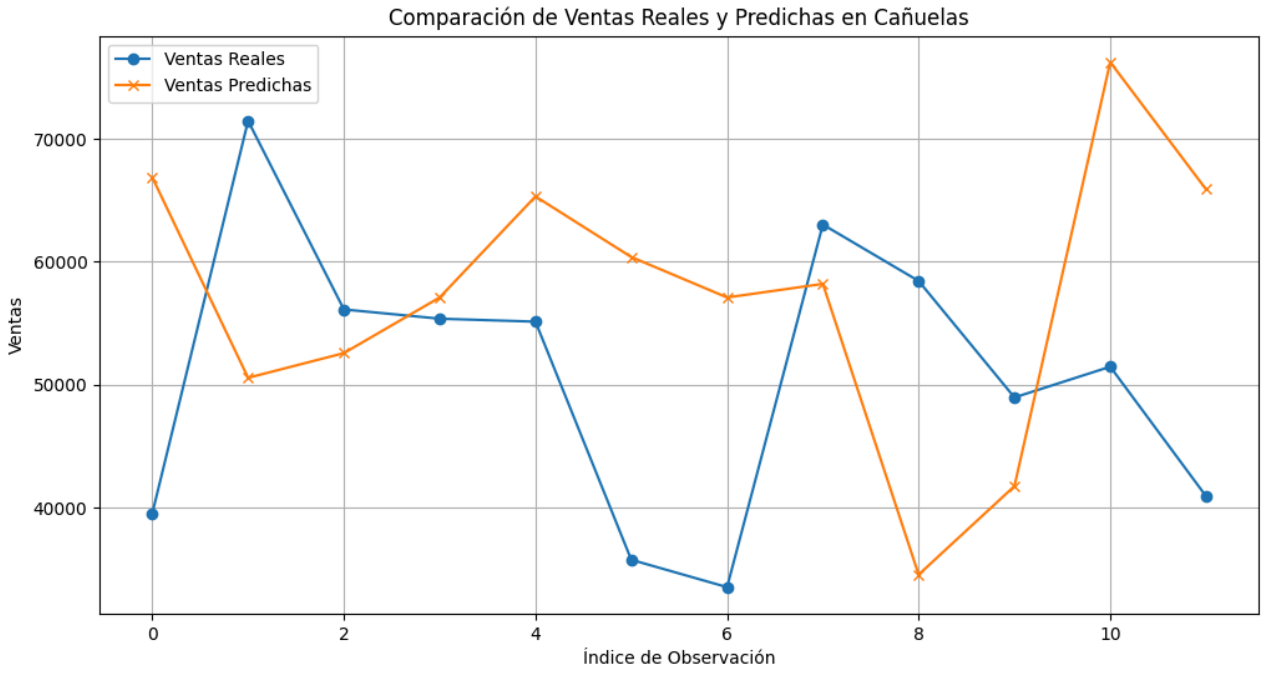
Cañuelas

valores reales	valores predichos
39548.42	66865.30
71470.50	50569.07
56129.11	52577.59
55376.80	57104.48
55135.07	65320.75
35766.25	60393.78
33540.46	57104.48
63037.32	58208.97
58443.21	34559.43
48970.49	41756.64
51471.01	76243.76
40940.29	65899.05

Con los datos predichos, se procede a hacer el análisis de tendencias temporales a través de un gráfico de líneas.







Optimización de la cadena de suministro

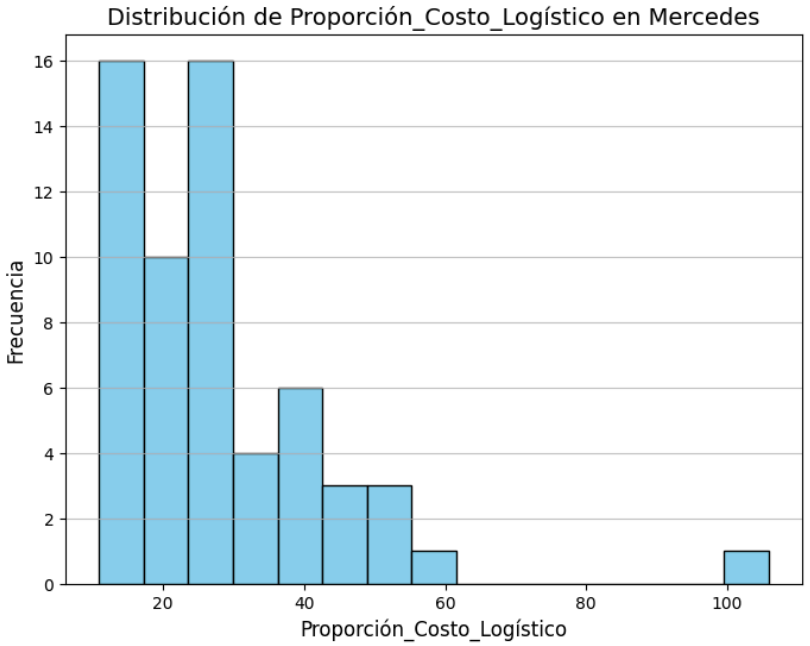
Planificar la logística para asegurar entregas eficientes sin incurrir en sobrecostos.

Para planificar esto utilizamos las variables “Logistic_Cost” y “Sales” para averiguar cual es la ciudad con menor gastos en logística.

Lo primero que hacemos es crear una nueva métrica llamada “Proporción_Costo_Logístico”, cuyo cálculo está basado en el concepto estadístico de” frecuencia relativa de ocurrencia”. Lo que se busca medir es qué porcentaje de proporción abarca el costo logístico con respecto a la venta de ese mes.

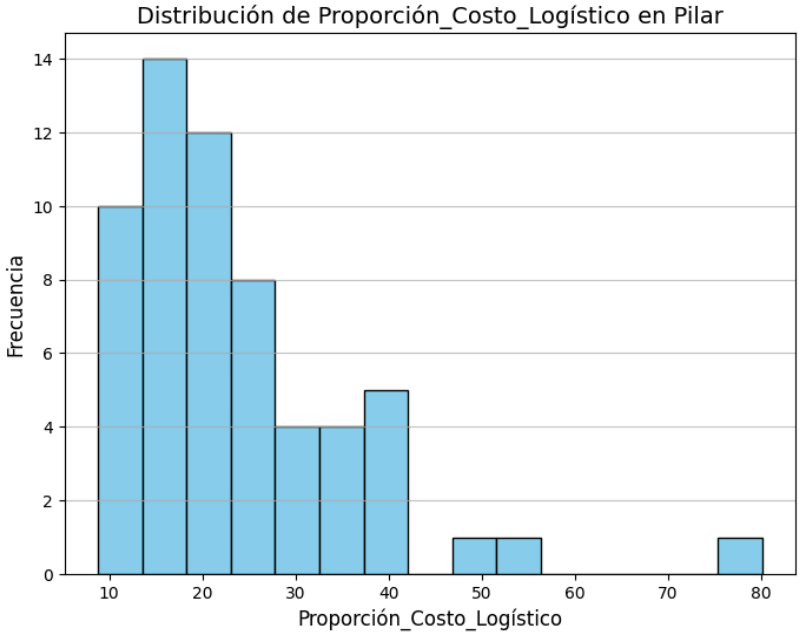
$$\text{Proporción\_Costo\_Logístico} = \frac{\text{Costo\_Logístico}}{\text{Ventas}} \text{ X } 100$$

Lo que nos interesa es encontrar los valores bajos porque significa que ese mes hubo una gran diferencia entre lo que se vendió con respecto a lo que se gastó en logística.



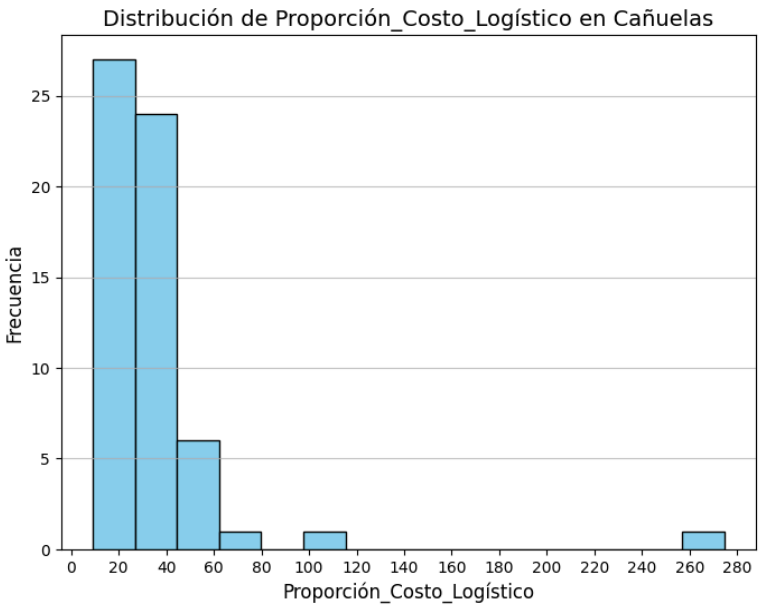

```
Mediana de Costo Logístico 13094.075

Mediana de Ventas 51917.494999999995
```




```
Mediana de Costo Logístico 11390.165

Mediana de Ventas 56032.215
```




```
Mediana de Costo Logístico 12817.755000000001

Mediana de Ventas 49003.835
```

Como conclusión se puede decir que Pilar es la ciudad con menor costo logístico, seguido de Mercedes y Cañuelas.

Conclusión general

Haciendo un resumen de todo lo analizado hasta ahora, el trabajo de análisis concluyó en que las tres ciudades donde mayor potencial de expansión existe son Mercedes, Pilar y Cañuelas, con cada ciudad teniendo características distintas que exigen estratégias específicas.

- Con respecto a la segmentación y estratégias de pricing, Pilar se destacó por tener los mejores indicadores y calidad de clientes, lo que permite aplicar estrategias de pricing más “agresivas” o que fomenten la compra de productos más caros o de alta calidad. En cambio, Mercedes y Cañuelas requieren un enfoque más conservador, necesitando de estrategias más reservadas, que busquen priorizar la oferta de productos más accesibles junto con descuentos u ofertas cruzadas con los productos más caros, para impulsar la venta de estos.
- Con respecto al marketing, el análisis evidenció una relación débil entre gastos en marketing y los resultados en ventas. Se recomienda reducir la inversión a niveles promedio, o a un rango de valores promedios, y de esta forma redirigir los recursos a otras áreas.
- En la parte de la predicción de la demanda, Mercedes demostró una demanda bastante estable, mientras que tanto Pilar como Cañuelas se mostraron un poco más variantes, con picos altos y bajos en la segunda mitad de año. Sería prudente estar atentos y monitorear las métricas clave durante esos meses para ajustar los recursos e inventario de manera preventiva.
- Por último, con respecto a la logística, Pilar es la que destacó con menores costos logísticos, después de Mercedes y Cañuelas. También se pudieron identificar casos atípicos en las tres ciudades, que deben ser revisados e investigados con mayor profundidad para determinar cuáles fueron las causas de valores extremos.