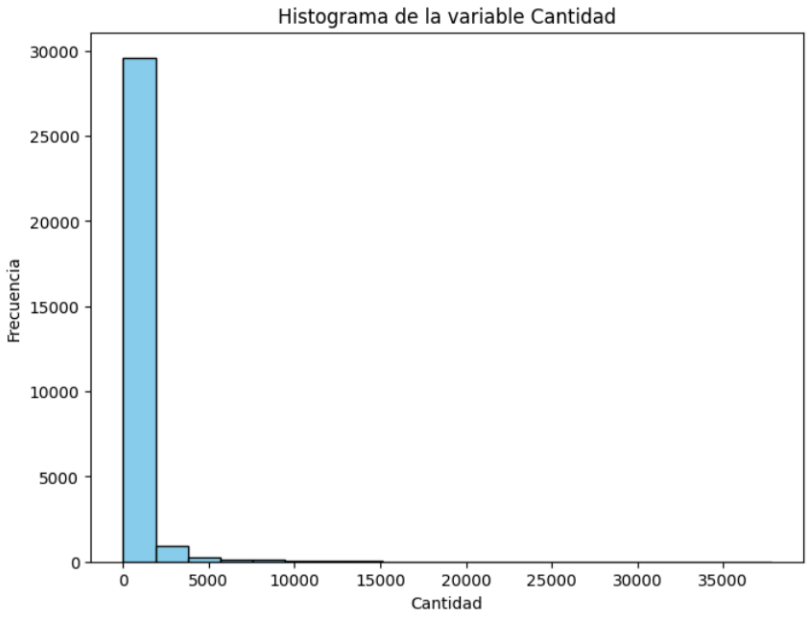
Hackathon Interlub

Jessica Dong Llauger A01638962
Sarah Gutiérrez Villalpando A01639343
Annete Cedillo Mariscal A01638984

En primera instancia, se realizó una limpieza del conjunto de datos: se eliminaron los valores duplicados, se realizó una búsqueda de valores faltantes para evaluar el método con el cual serían manejados, sin embargo, no se encontraron. También se ajustaron los datos capturados, ya que se redujeron los valores de la variable "Cantidad" que fueran un múltiplo de 1000.

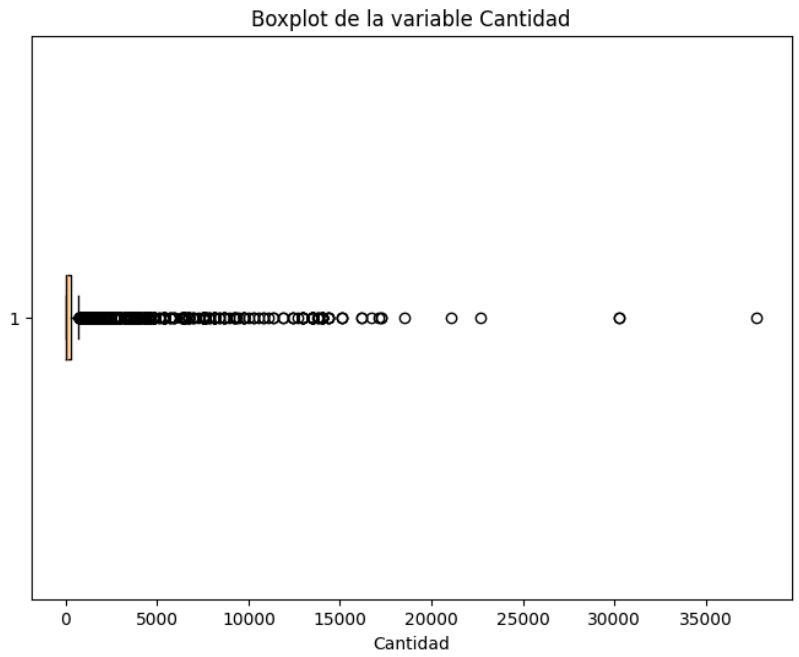
Análisis exploratorio de datos:

Después se realizó el análisis de los datos, en el cual se encontró que había solamente una variable numérica: "Cantidad", que sería la variable a predecir, por lo que se analizó su distribución mediante un histograma y gráfica de caja, y las métricas estadísticas de la variable:



Histograma de la variable Cantidad

En este histograma se puede apreciar como la frecuencia de cantidades entre 0 a 2500 es muy alta, el histograma muestra que los datos no tienen una distribución normal debido a valores atípicos muy alejados de las cantidades populares.



Boxplot de la variable Cantidad

Se puede observar que la mayoría de los datos son cantidades bajas pero se cuenta con pedidos de cantidades extremadamente altas como 37800 que son casos de pedidos masivos, estos distorsionan la distribución. Se consultó con Interlub y se afirmó que estos datos no son errores y si representan pedidos masivos reales.

Métricas descriptivas

Creación Orden de Venta	**Cantidad**	
count	31156	26199.000000
mean	2022-06-19 01:04:17.439979264	405.210695
min	2021-01-04 00:00:00	0.000000
25%	2021-09-15 00:00:00	22.000000
50%	2022-06-14 00:00:00	72.000000
75%	2023-03-14 00:00:00	300.000000
max	2023-12-29 00:00:00	37800.000000
std	NaN	1156.697059

Pruebas ANOVA para analizar las diferencias entre la variable numérica y las variables categóricas

- Hipótesis Nula (H0): No hay diferencias significativas en la 'Cantidad' vendida entre las diferentes categorías de la variable categórica.
- Hipótesis Alternativa (H1): Existen diferencias significativas en la 'Cantidad' vendida entre las diferentes categorías de la variable categórica.

Anova para la variable 'Articulo'

	sum_sq	df	F	PR(>F)
C(Articulo)	1.546747e+10	888.0	22.511002	0.0
Residual	1.958410e+10	25310.0	NaN	NaN

Interpretación: El tipo de artículo que se vende tiene un impacto significativo en la cantidad vendida. Los diferentes artículos que se venden no se venden en las mismas cantidades. Hay artículos que se venden significativamente más que otros. El p-value es mucho menor que el nivel de significancia común de 0.05. Esto significa que hay evidencia estadística muy fuerte para rechazar la hipótesis nula.

Anova para la variable 'Codigo Cliente'

	sum_sq	df	F	PR(>F)
C(CodigoCliente)	8.965789e+09	906.0	9.594884	0.0
Residual	2.608578e+10	25292.0	NaN	NaN

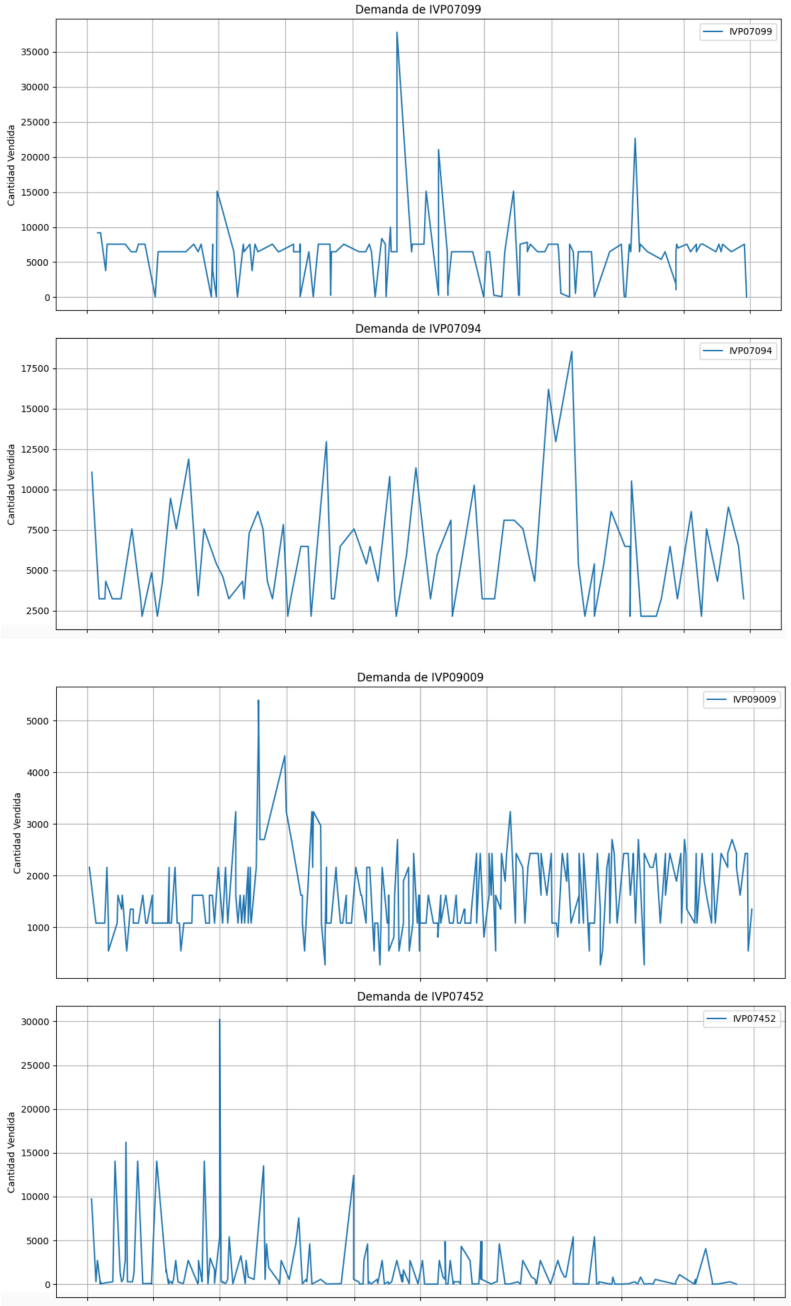
Interpretación: Existen diferencias estadísticamente significativas en la 'Cantidad' vendida entre los diferentes códigos de cliente: el código de cliente tiene un impacto significativo en la cantidad vendida. Los diferentes clientes no compran las mismas cantidades de productos. Hay clientes que compran significativamente más que otros. El código de cliente tiene un efecto mucho mayor en la cantidad vendida, que el artículo.

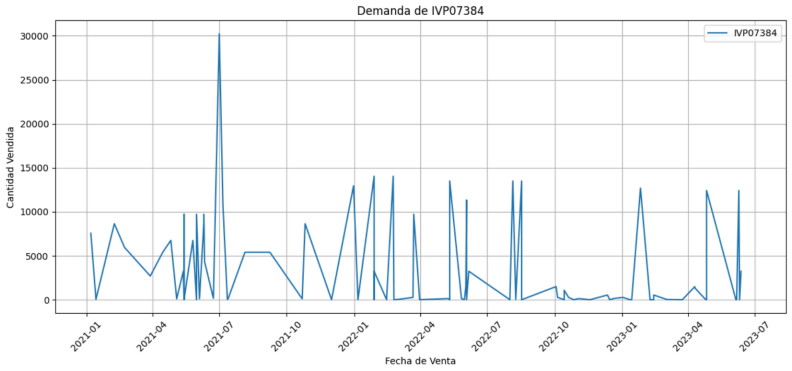
Anova para la variable 'Unidad de venta'

	sum_sq	df	F	PR(>F)
C(UnidadVenta)	1.291216e+09	2.0	500.952949	3.135064e-214
Residual	3.376035e+10	26196.0	NaN	NaN

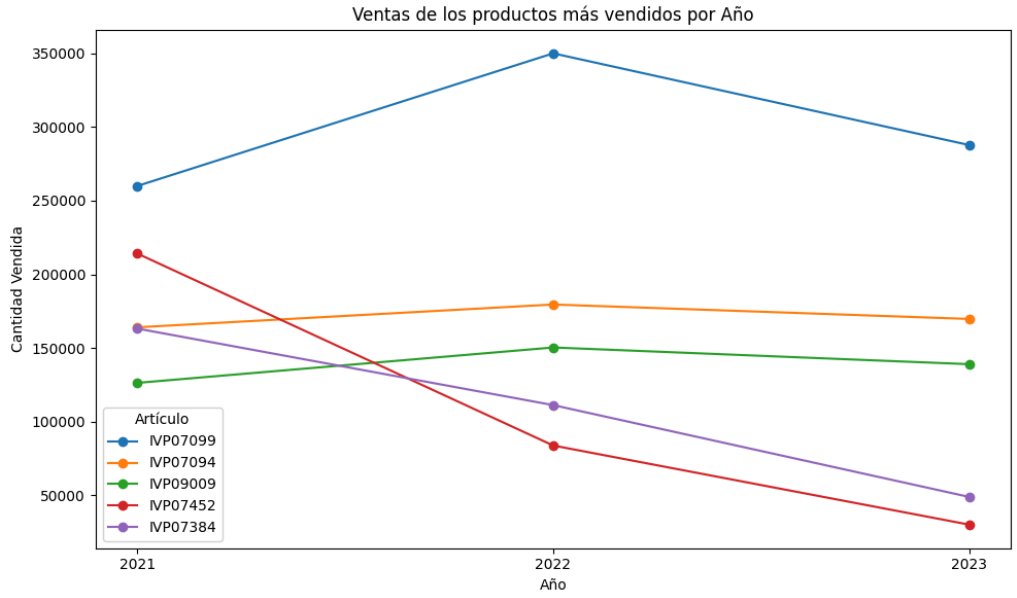
Interpretación: El p-value es muy pequeño, mucho menor que 0.05. Esto indica que hay evidencia estadística extremadamente fuerte para rechazar la hipótesis nula.Existen diferencias significativas en la 'Cantidad' vendida entre las diferentes unidades de venta: la unidad de venta tiene un impacto significativo en la cantidad vendida. Las diferentes unidades de venta influyen significativamente en la cantidad de producto vendido.

5 productos más vendidos

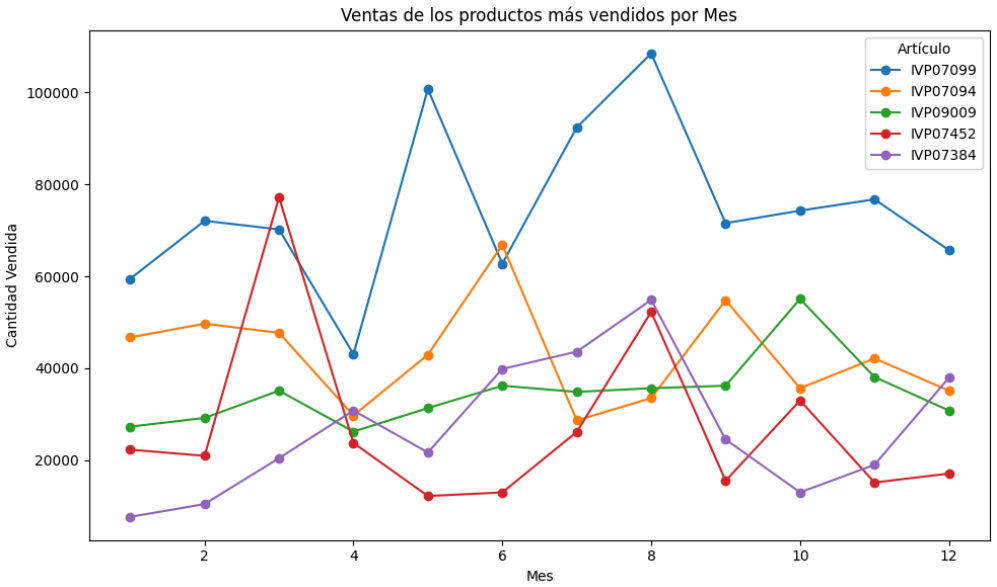




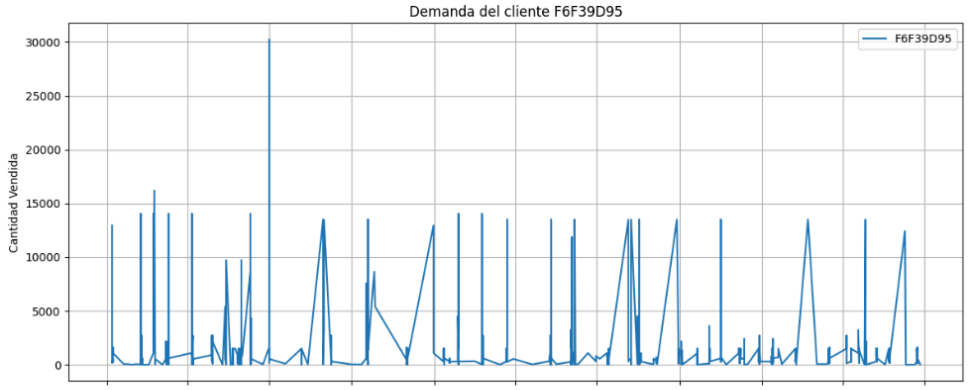
Productos más vendidos: IVP07099, IVP07094, IVP09009, IVP07452, IVP07384

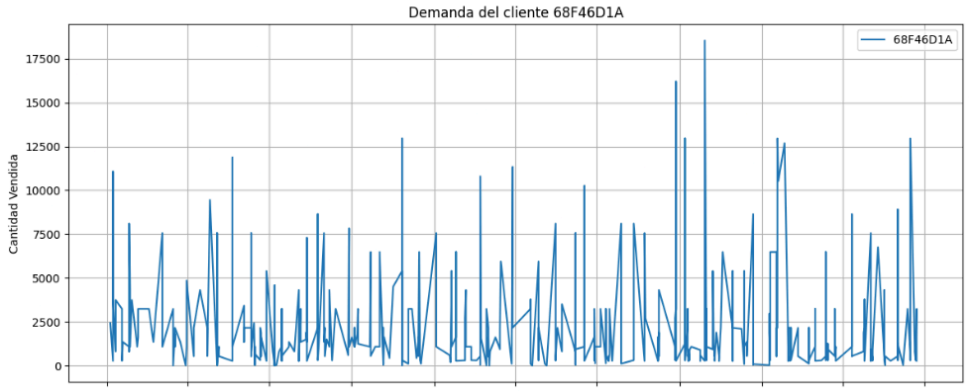


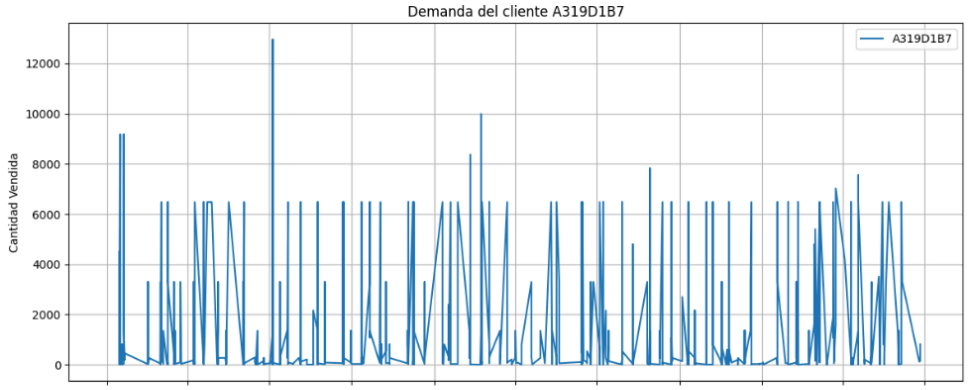
Esta gráfica presenta la cantidad más vendida de cada producto, dividida por año. Solo toma en cuenta los 5 productos más vendidos.

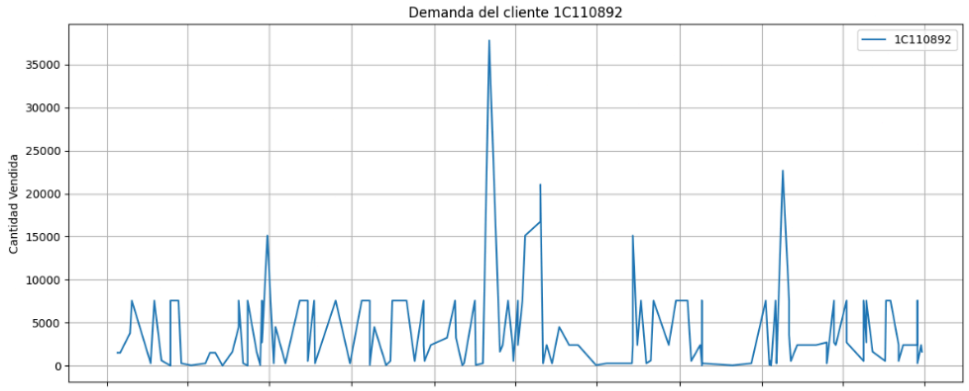


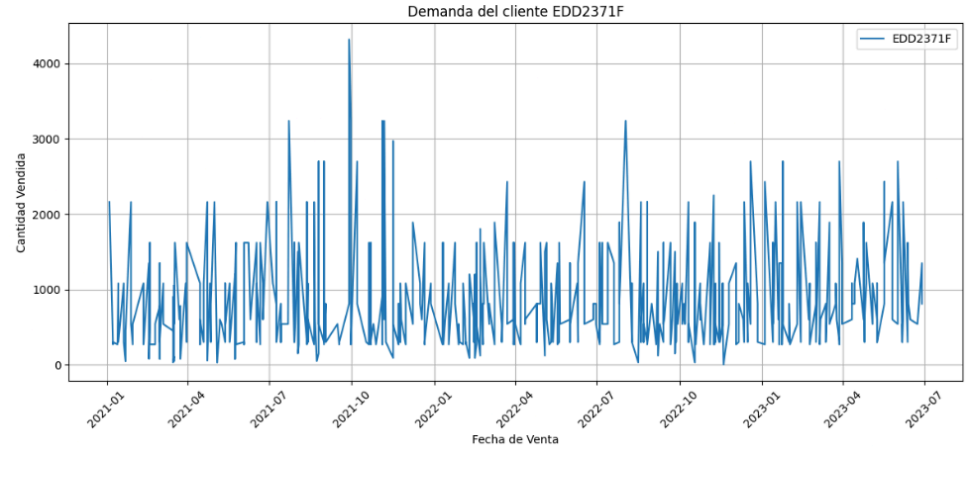
Esta gráfica muestra la cantidad de ventas de los productos más vendidos, dividido por meses.


Demanda del cliente F6F39D95
F6F39D95
Cantidad Vendida
30000
25000
20000
15000
10000
5000
0


Demanda del cliente 68F46D1A
68F46D1A
Cantidad Vendida
17500
15000
12500
10000
7500
5000
2500
0


Demanda del cliente A319D1B7
A319D1B7
Cantidad Vendida
12000
10000
8000
6000
4000
2000
0


Demanda del cliente 1C110892
1C110892
Cantidad Vendida
35000
30000
25000
20000
15000
10000
5000
0


Demanda del cliente EDD2371F
EDD2371F
Cantidad Vendida
4000
3000
2000
1000
0
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
2022-07
2022-10
2023-01
2023-04
2023-07
Fecha de Venta

A continuación los clientes que compran más y los artículos que más compran:

Código Cliente	Artículo
1C110892	IVP07099
68F46D1A	IVP07094
A319D1B7	IVP07099
EDD2371F	IVP09009
F6F39D95	IVP07384

Pruebas de estacionalidad y estacionariedad

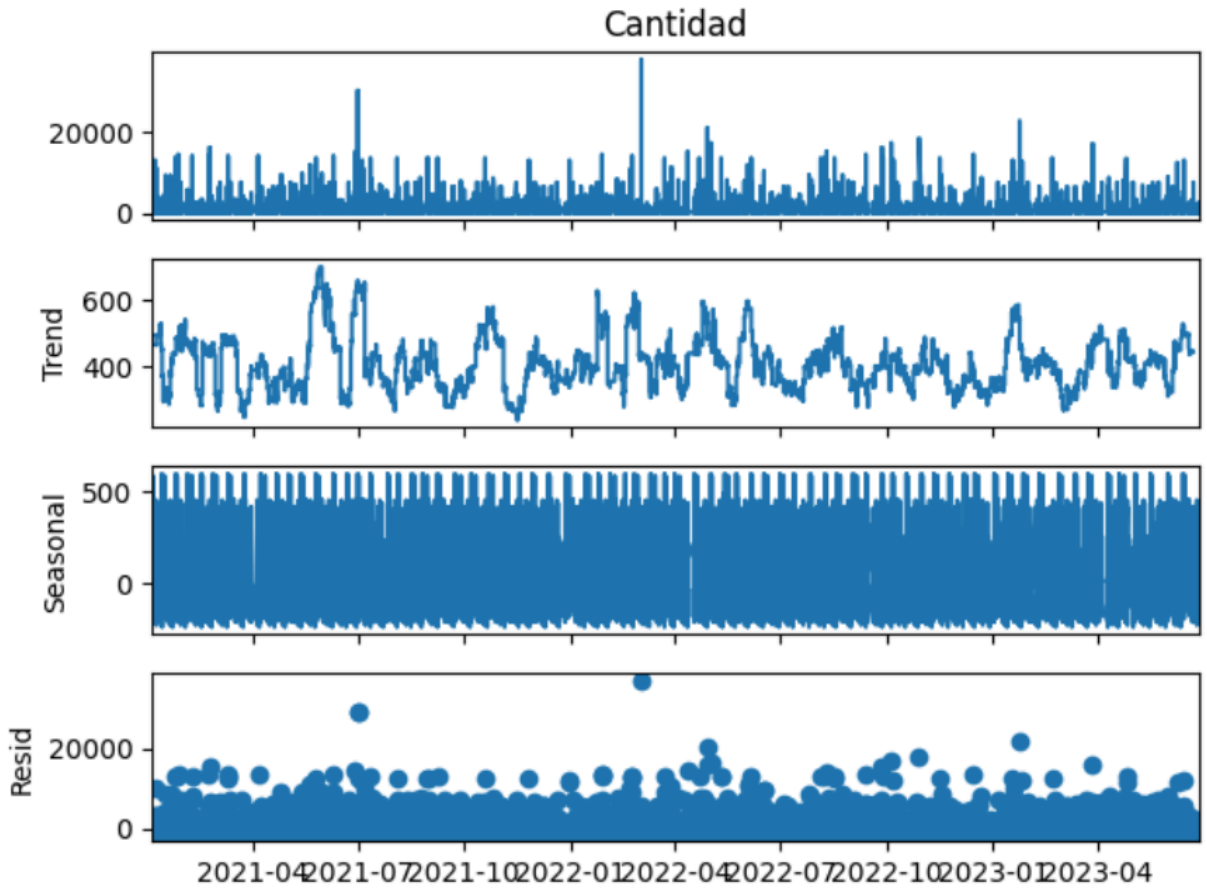
Prueba Dickey-Fuller Aumentada.

- H0: Tiene propiedades estadísticas constantes a lo largo del tiempo, su media y varianza no cambian.
- H1: Tiene características de media, varianza y covarianza que cambian a lo largo del tiempo.

```
Resultados de la prueba ADF para el artículo: IVP07099
Test Statistic                -9.488095e+00
p-value                        3.715640e-16
#Lags Used                     0.000000e+00
Number of Observations Used    1.130000e+02
Critical Value (1%)           -3.489590e+00
Critical Value (5%)           -2.887477e+00
Critical Value (10%)          -2.580604e+00
dtype: float64
Resultados de la prueba ADF para el artículo: IVP07094
Test Statistic                -8.319708e+00
p-value                        3.600139e-13
#Lags Used                     0.000000e+00
Number of Observations Used    8.500000e+01
Critical Value (1%)           -3.509736e+00
Critical Value (5%)           -2.896195e+00
Critical Value (10%)          -2.585258e+00
dtype: float64
Resultados de la prueba ADF para el artículo: IVP09009
Test Statistic                  -4.955656
p-value                          0.000027
```

```
#Lags Used                        3.000000
Number of Observations Used    223.000000
Critical Value (1%)             -3.460019
Critical Value (5%)             -2.874590
Critical Value (10%)            -2.573725
dtype: float64
Resultados de la prueba ADF para el artículo: IVP07452
Test Statistic                  -4.772594
p-value                          0.000061
#Lags Used                       3.000000
Number of Observations Used    162.000000
Critical Value (1%)             -3.471374
Critical Value (5%)             -2.879552
Critical Value (10%)            -2.576373
dtype: float64
Resultados de la prueba ADF para el artículo: IVP07384
Test Statistic                 -2.971683
p-value                         0.037625
#Lags Used                      2.000000
Number of Observations Used    66.000000
Critical Value (1%)            -3.533560
Critical Value (5%)            -2.906444
Critical Value (10%)           -2.590724
dtype: float64
```

Descomposición estacional



En la primera gráfica se observan fluctuaciones significativas en la cantidad, con picos y valles pronunciados (pueden ser considerados valores atípicos). Se observa que la serie no tiene una media y varianza constante, por lo que no es estacionaria.

Gráfica de tendencia: representa la tendencia subyacente de la serie temporal, además de que muestra una fluctuación gradual a lo largo del tiempo, con algunos aumentos y disminuciones.

Gráfica de estacionalidad: se observa un patrón repetitivo muy claro a lo largo del tiempo que indica la presencia de estacionalidad en la serie temporal.

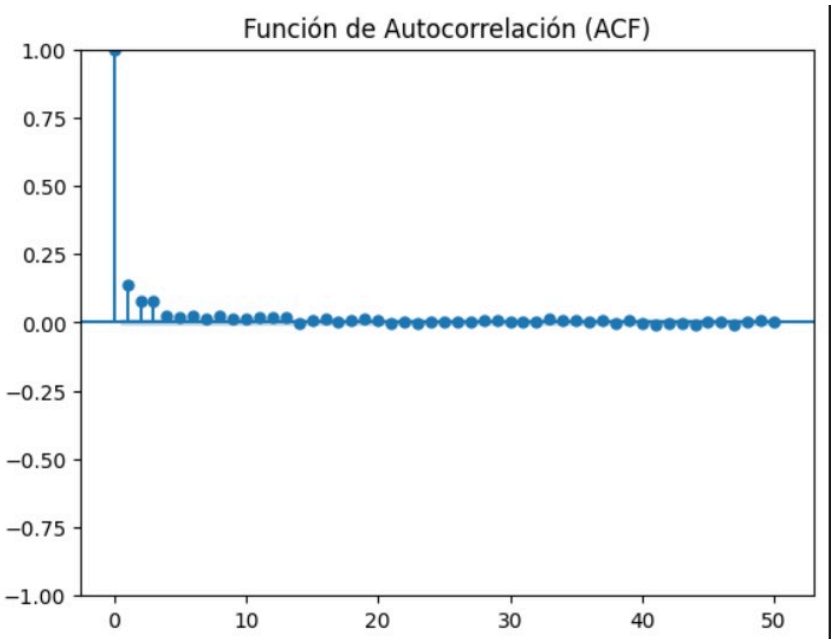
Gráfica de residuos: representa la parte de la serie temporal que no se explica por la tendencia o la estacionalidad. Muestra una distribución aleatoria de puntos, con algunos valores atípicos, lo cual indica que los residuos deben ser aleatorios y no mostrar patrones.

Test de estacionariedad Dickey-Fuller

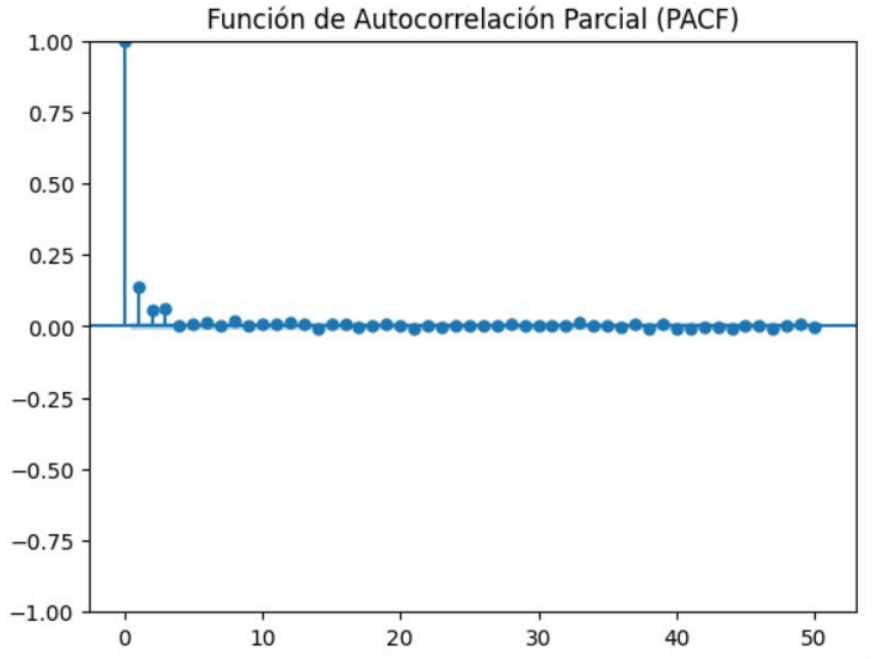
```
Estadístico de prueba ADF: -40.37544908777359
Valor p: 0.0 -> la serie no es estacionaria
```

Estadístico de prueba ADF: -40.37544908777359

Este es el valor del estadístico de prueba calculado por la prueba ADF. Cuanto más negativo sea este valor, más fuerte es la evidencia contra la hipótesis nula de que la serie tiene una raíz unitaria (es decir, que no es estacionaria).



La ACF muestra la correlación entre la demanda actual y sus valores pasados. En la gráfica observada, la ACF cae rápidamente a cero, lo que indica que la demanda reciente influye más que la de largo plazo.



La PACF mide la relación directa entre la demanda actual y sus rezagos, eliminando efectos intermedios. La gráfica PACF indica que solo los primeros rezagos son significativos.

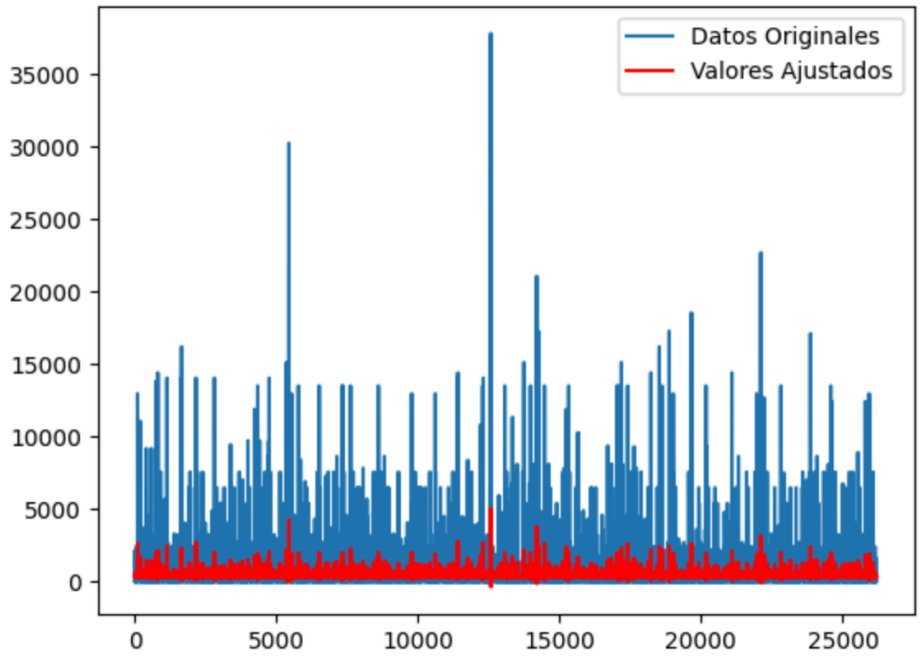
Desarrollo de modelo predictivo:

A partir del análisis exploratorio, el primer modelo que se realizó fue el modelo Media Móvil de orden 3 para modelar los valores actuales de la serie temporal como una combinación lineal de los errores de los valores pasados.

```
==============================================================
Dep. Variable:               Cantidad   No. Observations:                26199
Model:                 ARIMA(0, 0, 3)   Log Likelihood             -221632.529
Date:                Sat, 22 Mar 2025   AIC                         443275.059
Time:                        22:27:35   BIC                         443315.926
Sample:                             0   HQIC                        443288.256
                                - 26199
Covariance Type:                  opg
==============================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------
const        405.2107     13.977     28.991      0.000     377.816     432.605
ma.L1          0.1233      0.002     61.310      0.000       0.119       0.127
ma.L2          0.0642      0.004     17.464      0.000       0.057       0.071
ma.L3          0.0716      0.002     41.512      0.000       0.068       0.075
sigma2      1.306e+06   2178.224    599.342      0.000     1.3e+06    1.31e+06
==============================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):          21427443.91
Prob(Q):                              0.83   Prob(JB):                         0.00
Heteroskedasticity (H):               0.77   Skew:                             9.03
Prob(H) (two-sided):                  0.00   Kurtosis:                       141.94
==============================================================
```

Interpretación:

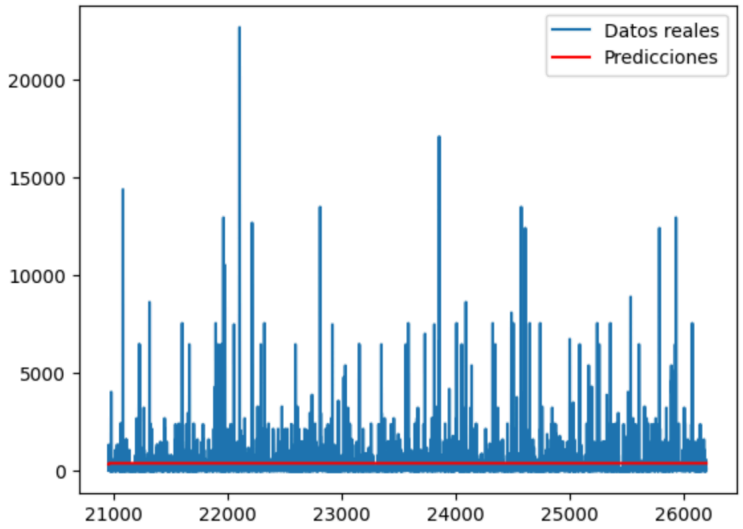
El modelo MA(3) se ajusta bien a la serie, con coeficientes significativos y sin autocorrelación en los residuos, según la prueba de Ljung-Box. Sin embargo, las métricas de Jarque-Bera, skewness y kurtosis indican que los residuos no siguen una distribución normal, lo que podría afectar la validez de algunas inferencias. Además, la presencia de heterocedasticidad sugiere que la varianza de los errores no es constante, lo que podría mejorarse con transformaciones o modelos alternativos como ARIMA.



Se observa que el modelo subestima la variabilidad de los datos, ya que los valores ajustados se mantienen cerca de la media y no capturan los picos extremos de la serie. Esto sugiere que el modelo no es suficiente para representar la estructura de la serie, debido a la alta volatilidad y los valores atípicos. Además, la heterocedasticidad detectada en las métricas del modelo se refleja en la gráfica, donde los datos originales muestran variaciones amplias mientras que los valores ajustados permanecen relativamente estables.

Modelo ARIMA

A partir de estos resultados, se realizó un grid search para encontrar los mejores parámetros para la realización de un modelo ARIMA, ya que al realizar un modelo autorregresivo es posible capturar la volatilidad de los datos. Los resultados fueron: (0, 1, 2) con MSE: 1122549.8867922942. Por lo cual, se realizaron predicciones con los nuevos parámetros:



Mean Squared Error (MSE): 1304435.3822828631

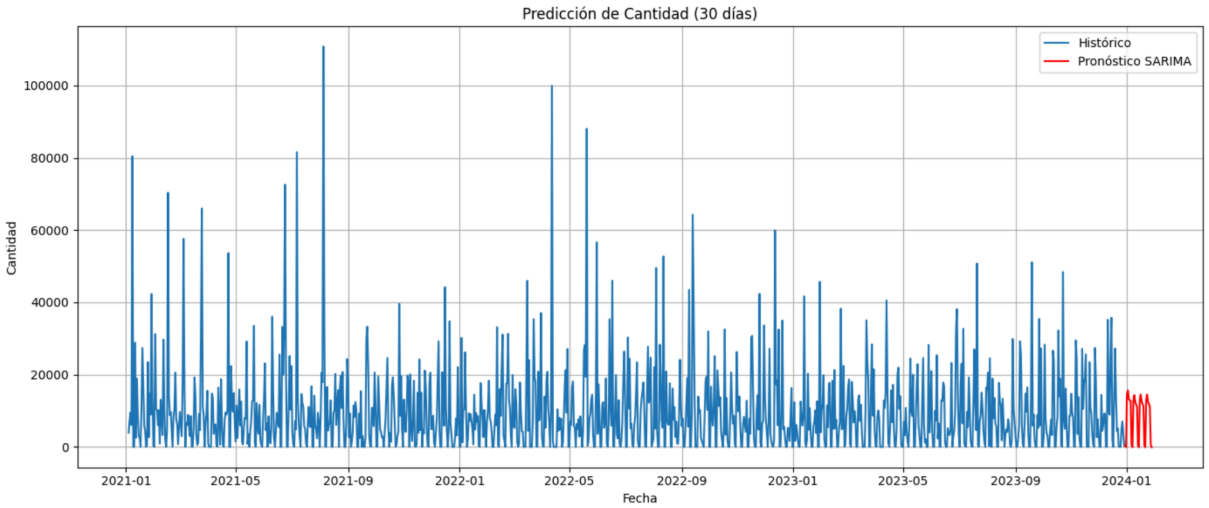
Aunque el modelo mejoró ligeramente al incluir diferenciación (d=1), lo que ayuda a capturar tendencias en los datos, siguió sin adaptarse bien a la variabilidad de la serie. Se observa que las predicciones permanecen cerca de un valor medio, sin reflejar los picos y fluctuaciones presentes en los datos reales. Esto sugiere que el modelo sigue sin capturar completamente la dinámica de la serie, por lo que se decidió realizar un modelo SARIMA, ya que permite capturar patrones estacionales en los datos, como se encontraron en el análisis exploratorio.

Modelo SARIMA

Se realizó el modelo SARIMA con los siguientes parámetros de orden:

Orden no estacional (1,1,1): captura tendencias a corto plazo y fluctuaciones en los datos.

Orden estacional (1,1,1,7): el modelo asume una estacionalidad semanal.



Después, se realizó una validación cruzada en series de tiempo para evaluar el desempeño del modelo, con la cual se obtuvieron los siguientes resultados:

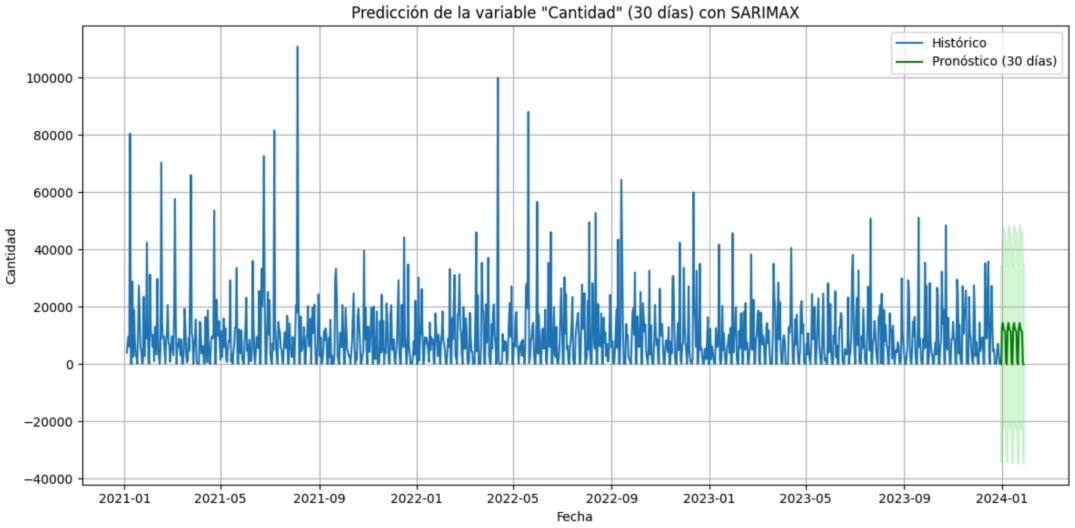
```
([26278.684741709483,
  11582.86997531515,
  10091.363036880899,
  7176.130435526607,
  12576.283825799213],
 np.float64(13541.06640304627))
```

Interpretación: el modelo SARIMA muestra una tendencia estable en sus pronósticos, aunque con valores significativamente más bajos y menos fluctuantes en comparación con la serie histórica, lo que sugiere que el modelo suaviza los picos extremos presentes en los datos reales. En la validación, se obtuvieron errores RMSE para los 5 folds de [26278.68, 11582.88, 10091.36, 7213.04, 12556.47], con un RMSE promedio de 13,544.49, lo que indica que, en general, el modelo tiene un error aproximado de 13,500 unidades en sus predicciones semanales. Se observa que el error disminuye en los últimos folds, lo que sugiere que el modelo mejora conforme avanza en el tiempo, probablemente porque dispone de más datos para entrenarse en los pliegues posteriores. Además, es posible que en los primeros períodos de la serie haya habido más ruido, irregularidades o cambios en el comportamiento de los datos, lo que afecta la precisión de las primeras predicciones.

Modelo SARIMA ajustado

En última instancia, se decidió ajustar los parámetros del modelo SARIMA. Mediante el método grid-search utilizado anteriormente, se buscaron los parámetros que mejor se ajustaran a los datos de la serie. Se obtuvieron los siguientes resultados:

```
order = (1, 1, 1)
seasonal_order = (0, 1, 1, 7)
```



El modelo SARIMAX ajustó la estacionalidad semanal y reprodujo ese patrón en la predicción. El modelo predice que la variable "Cantidad" se mantendrá dentro de rangos similares al pasado reciente, pero con cierta incertidumbre (amplitud de las bandas de confianza). En el pasado hay picos extremos, pero el modelo no los proyecta en el futuro, ya que no está considerando valores atípicos como recurrentes, prioriza la tendencia y estacionalidad promedio, no los outliers. En la validación cruzada se encontró un RMSE de 8,912.93, lo cual indica un mucho mejor ajuste a comparación de los modelos probados anteriormente.

Recomendaciones futuras

- Probar enfoques como búsqueda bayesiana o validación cruzada más extensa podría ayudar a encontrar combinaciones más óptimas para la optimización de parámetros.
- Explorar modelos más sofisticados de predicción para series temporales como Prophet (para series con estacionalidad fuerte) o modelos de aprendizaje profundo como LSTMs y Transformers para mejorar la precisión de las predicciones.
- Entrenar modelos separados para distintos productos para mejorar la predicción en contextos específicos en lugar de aplicar un único modelo global.
- Aplicar técnicas de suavizado como medias móviles o filtrado de ruido para reducir la sensibilidad a valores atípicos.
- Incorporar datos de eventos especiales si las ventas están influenciadas por estas condiciones.
- Incorporar variables estacionales (festividades, temporadas altas y bajas) o de mercado (competencia, promociones, tendencias de consumo), ya que puede mejorar la capacidad del modelo para capturar patrones más complejos en la demanda de productos.

Link al repositorio de Github con los modelos:

https://github.com/sarx-oh/Interlub_Hackaton

Link al video:

https://drive.google.com/file/d/1GJLA0CrewjsPc19KqYgxKf3qQlpRpHnN/view?usp=drivesdk