¿Qué haríamos sin ellos en el presente y futuro?

"Evolución de Infusiones y otros productos en el mercado interno argentino"

Data Science
Comisión 29795
Profesor: Marcos Rojo
Tutor: Fernando Pareja
Equipo: Silvana Tomsig
Sebastián de León

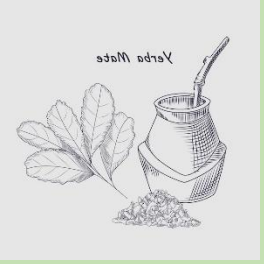
Índice de Contenidos

Presentación

Todo comenzó con una propiedad ganadera llamada Vuelta del Ombú, Departamento de Santo Tomé - Corrientes. La familia y sus descendientes realizaron las primeras plantaciones de yerba mate. Posteriormente, se avanzó con la industrialización del producto elaborado En el siglo XXI la empresa realiza convenios con varias empresas internacionales para comenzar a ser distribuidora nacional de sus productos, con lo cual además de productos nacionales de primera necesidad comienza la comercialización de otros productos.

¿Qué es la Yerba Mate? **La yerba mate, es una especie arbórea neotropical (Ilex paraguariensis) originaria de América del Sur. De las hojas y ramas, secas y molidas se prepara el mate, una infusión originaria de su zona**

¿Qué es el té? El té es la infusión de las hojas y brotes de la planta del té (Camellia sinensis). Su sabor es fresco, ligeramente amargo y astringente. Las hojas del arbusto, si no son secadas apenas se recolectan, comienzan a oxidarse. Para prevenir este proceso de oxidación, se calientan con el objetivo de quitarles la humedad.



Definición de temática

Hipótesis

La empresa posee rendimientos por la comercialización 4 tipos de productos.

La empresa buscará maximizar el nivel de facturación y rendimiento.

Objetivos Principales

Predecir que facturación en pesos.

Predecir que los tipos de rendimientos en pesos.

Objetivos Secundarios

Determinar relación entre facturación y costos

Determinar relación entre facturación y kilos

Determinar los costos por línea y cadena de distribución

Calcular rentabilidad

Determinar períodos de mayor facturación.

Usuarios finales

El alcance del presente es solamente en el ámbito del directorio y gerencias de la empresa, poniendo al alcance de la mano una herramienta diseñada específicamente para la toma futura de decisiones.

Línea futura

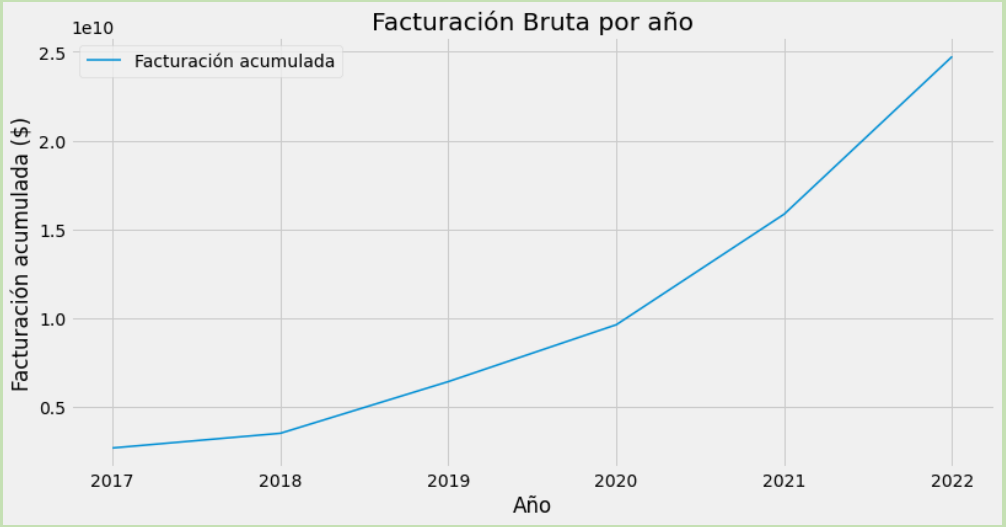
Este informe se puede ampliar con nuevos criterios de medida, se debe actualizar con nuevos datos, abrir por producto o canal de distribución y ampliar el espectro hacia otros rubros económicos desarrollados por la empresa para así tener un panorama más amplio de la actividad desarrollada.

Tabla de versiones

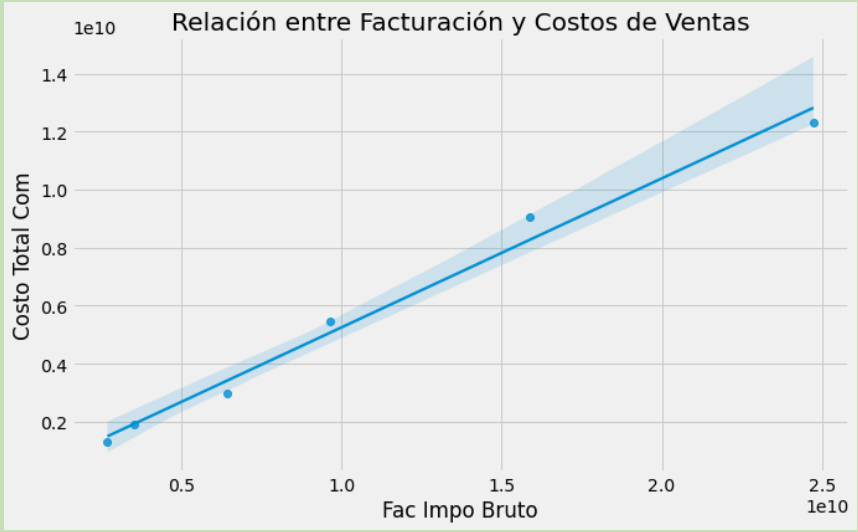
Fecha	Versión	Entrega	Alcance
03/08/22	V.1	1° Pre entrega	Elección data set, EDA, análisis de series temporales. Base de datos actualizada al 08/02/2022
17/10/22	V.2	2° Pre entrega	V.1, limpieza data set, obtención de insights. Base de datos actualizada al 03/08/2022
27/12/22	V.3	Entrega final	V.2, entrenamiento ML, búsqueda de parámetros, elección de modelo. Base actualizada al 19/11/22

Análisis Exploratorio de Datos

- Presentación de las variables de interés
- Las principales variables, a tener en cuenta son: Facturación y Costos de los productos los cuales se encuentran medidos en la moneda AR$



Facturación bruta acumulada por año desde 2017 a 2022

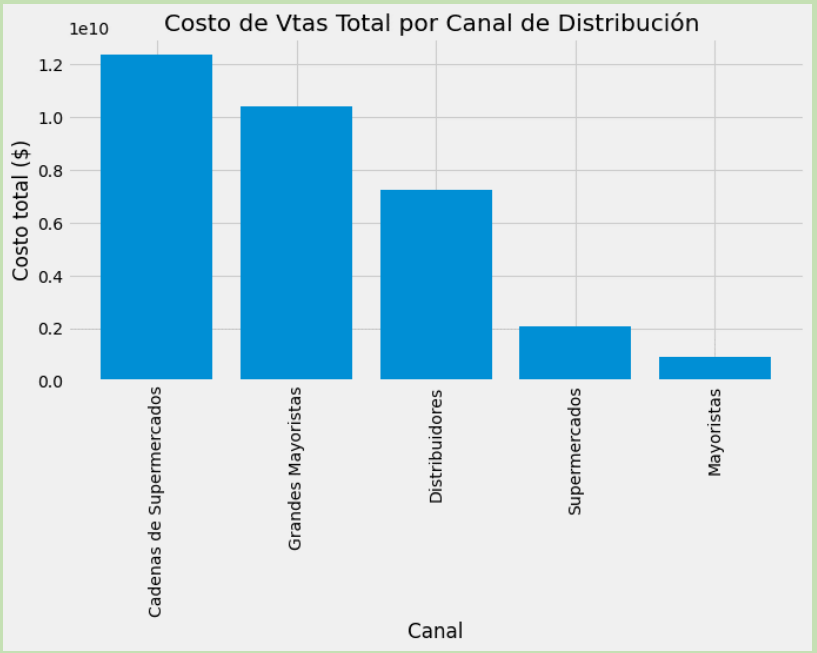


Relación entre facturación bruta y costo

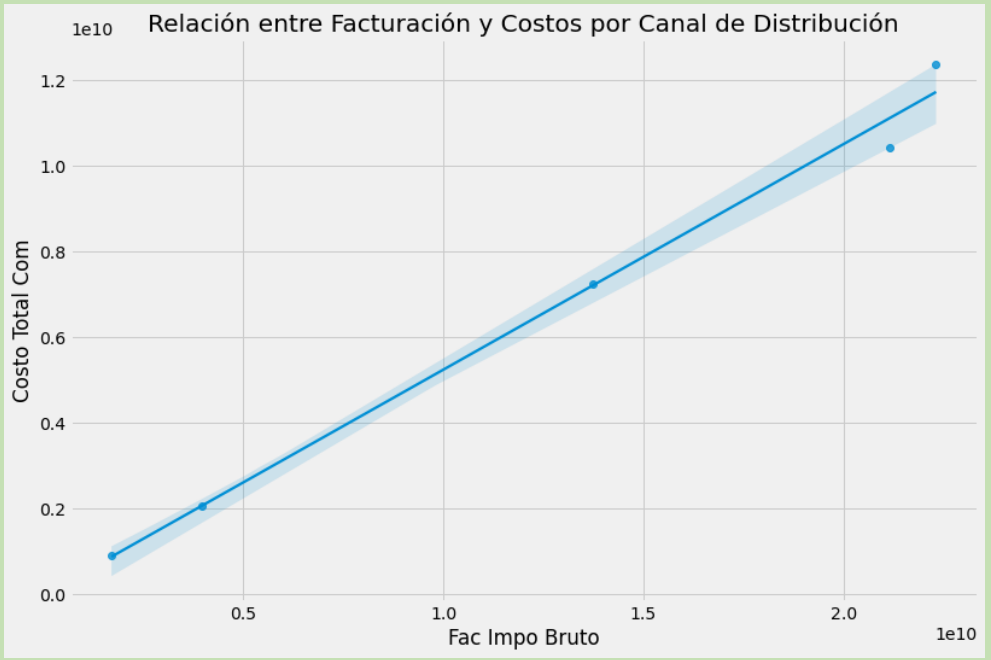
Al observar esta relación se evidencia cómo bien menciona la teoría, que ambas variables poseen una relación directamente proporcional, al aumentar una variable automáticamente la otra aumenta.

Análisis Exploratorio de Datos (EDA)

- *“Es preferible una respuesta aproximada a la pregunta correcta, que frecuentemente es formulada de manera imprecisa, que una respuesta exacta a la pregunta incorrecta, que siempre puede ser formulada de manera precisa.” — John Tukey*



En este gráfico podemos observar, como las grandes cadenas de supermercados y grandes mayoristas, son los principales distribuidores de los productos

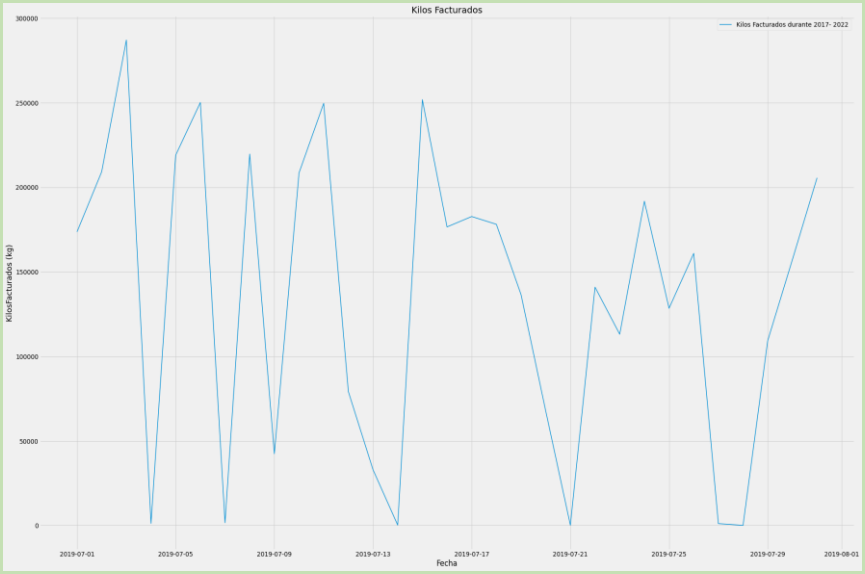


Para este caso las variables de estudio muestran una relación directamente proporcional.
Visualmente se registra que entre ambas variables existe una diferencia, que puede deberse a diferentes razones.

Análisis de Series Temporales

Una serie temporal es cualquier conjunto de datos donde los valores se miden en diferentes puntos en el tiempo. Muchas series temporales son medidas uniformemente con una frecuencia específica

Evolución de la Facturación



En la evolución se puede observar el comportamiento de la serie en el tiempo, demostrando que no es estática.

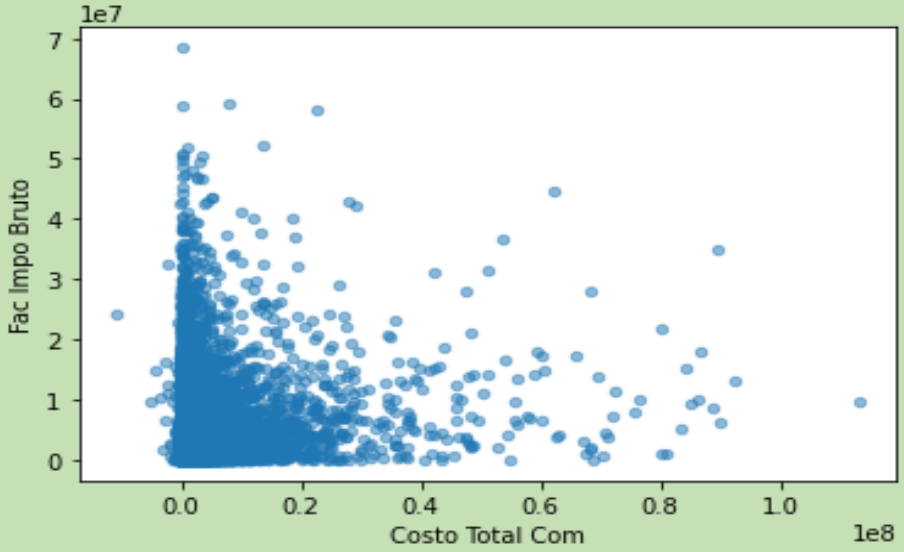
Estacionalidad

date	
2021-03-31	48.300000
2021-06-30	52.219780
2021-09-30	50.934783
2021-12-31	46.065217

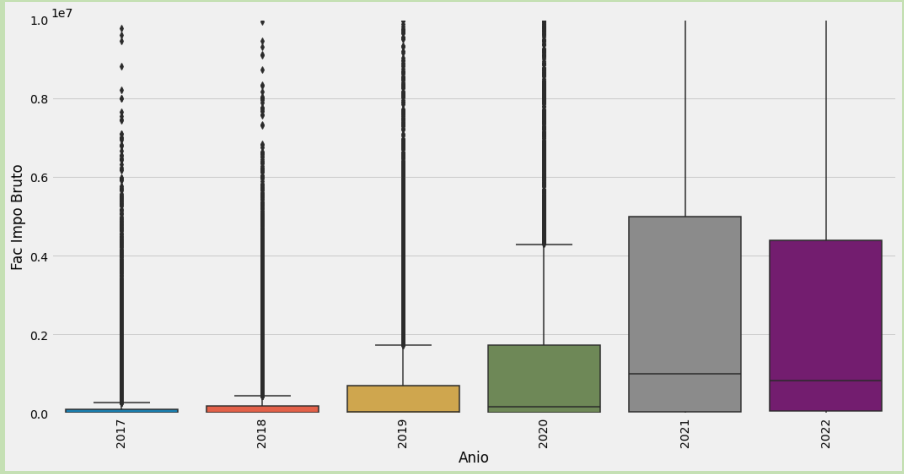
En un análisis de estacionalidad, observamos que el segundo y tercer trimestre registran los mayores volúmenes de facturación en pesos, quedando el primer lugar el cuarto trimestre.

Análisis de Series Temporales

[Relación con variables numéricas]



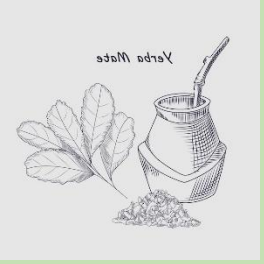
[Relación con variables categóricas]



Resumiendo:

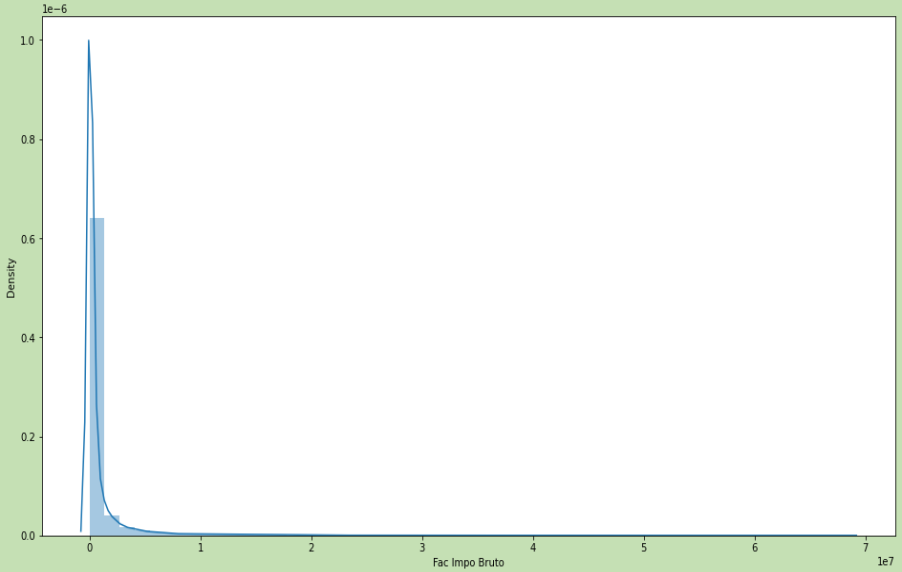
La relación entre facturación y costo, poseen una estructura atomizada dada su relación directa.

En el diagrama de caja se puede observar la variabilidad con respecto al transcurso del tiempo

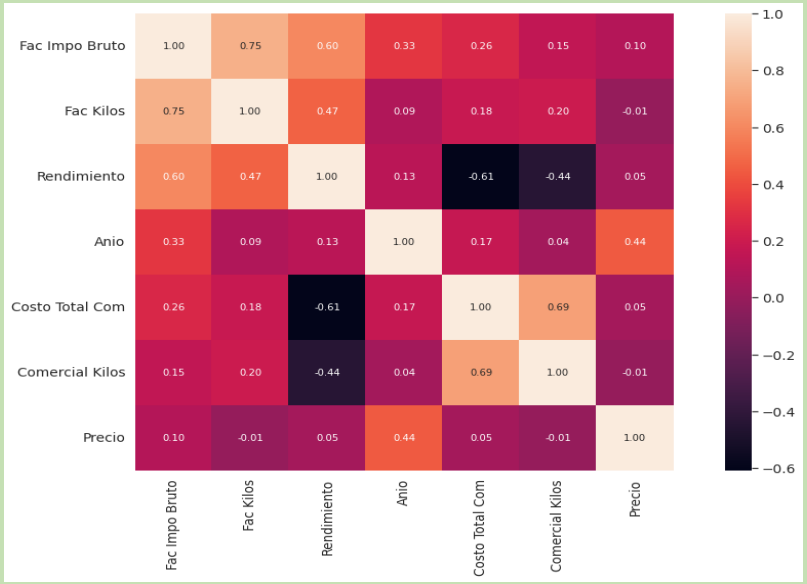


Análisis de Series Temporales

Análisis Univariable- facturación

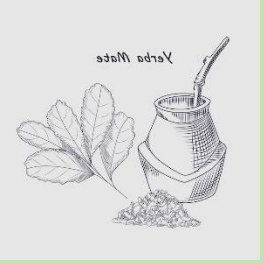


Matriz de correlación



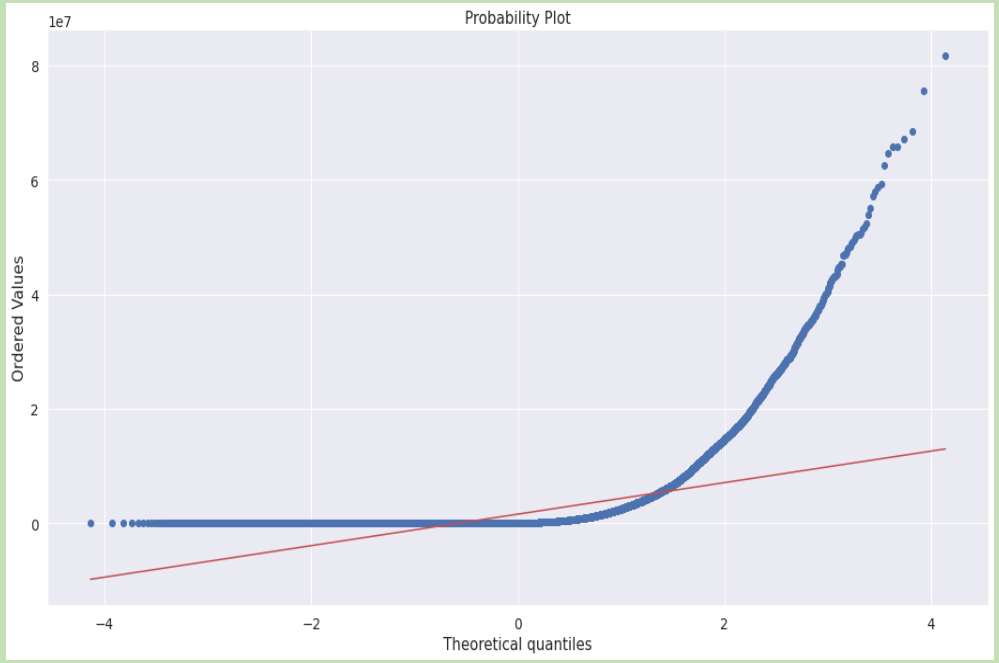
Resumiendo:

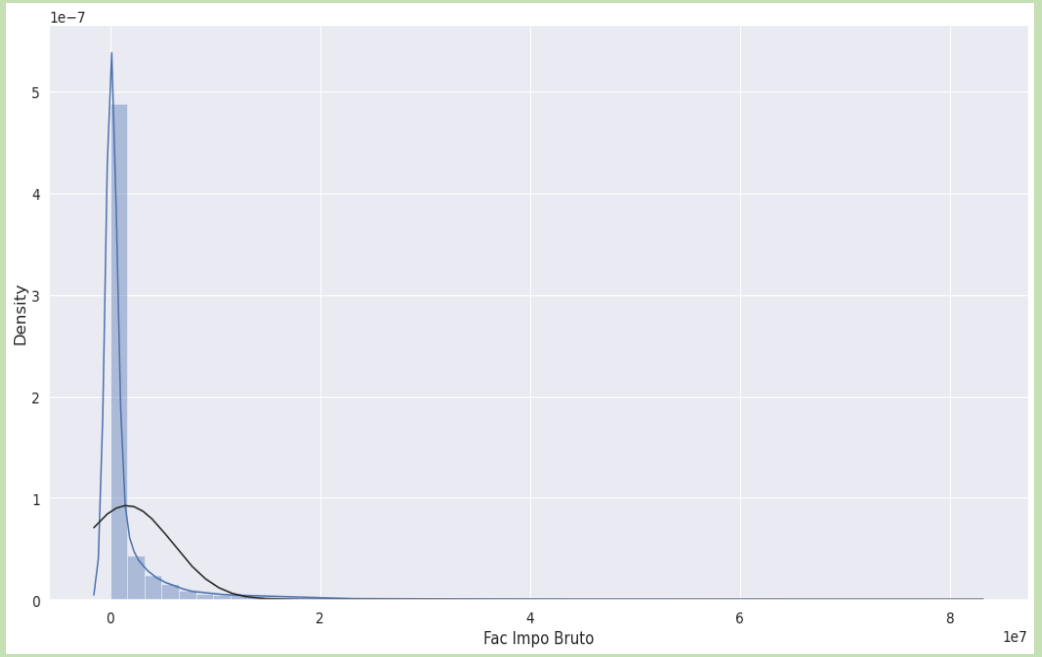
El mapa de calor es una forma visual muy útil para para conocer las variables y sus relaciones. A primera vista, se muestra una correlación uno a uno con su propia variable, y a medida que nos alejamos de una variable su correlación va disminuyendo.



Análisis de Series Temporales

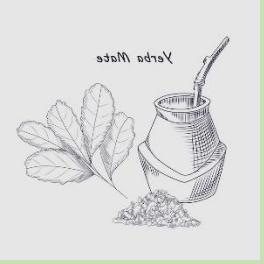
[Análisis de normalidad- facturación]





Resumiendo:

De estos gráficos se desprende que 'Fac Impo Bruto' no conforma una distribución normal, presenta similitud a una Chi Cuadrado. Muestra picos, asimetría positiva y no sigue la línea diagonal.



Modelo y parámetros óptimos

- **Entrenando un Algoritmo de ML**

El proceso de forecasting consiste en predecir el valor futuro de una serie de tiempo, bien modelando la serie únicamente en función de su comportamiento pasado (autorregresivo) o empleando otras variables externas.
Avanzando con la funcionalidad del modelo, se toma como base la variable Facturación en pesos, sin discriminar producto o cadena de distribución.

1- *Fechas*: *Se realiza un group by para agrupar la información por fecha*

```
DatetimeIndex(['2017-01-01', '2017-02-01', '2017-03-01', '2017-04-01',
               '2017-05-01', '2017-06-01', '2017-09-01', '2017-10-01',
               '2017-11-01', '2017-12-01',
               ...
               '2022-11-17', '2022-11-17', '2022-11-17', '2022-11-17',
               '2022-11-18', '2022-11-18', '2022-11-18', '2022-11-18',
               '2022-11-18', '2022-11-19'],
              dtype='datetime64[ns]', name='Fecha', length=39334, freq=None)
```

Modelo y parámetros óptimos

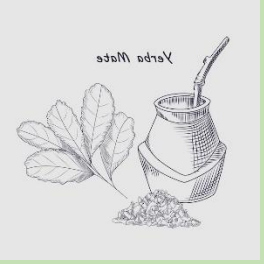
2- Train y Test: En este punto se define la cantidad de datos con el cual se armará el modelo y se entrenará el ML.

```
[ ]  y_train.tail()
```

```
Fecha
2022-11-03    105852085
2022-11-04    101767815
2022-11-05     79884097
2022-11-06     47654107
2022-11-07    114493700
Name: Fac Impo Bruto, dtype: int64
```

```
[ ]  y_test.head()
```

```
Fecha
2022-11-08    257388344
2022-11-09     41564389
2022-11-10    134375014
2022-11-11    148301090
2022-11-13     31029537
Name: Fac Impo Bruto, dtype: int64
```



Modelo y parámetros óptimos

3- Modelo: En este punto se define la cantidad de datos con el cual se armará el modelo y se entrenará el ML.
A- *Modelo ARIMA* para la predicción de series de tiempo ARIMA significa modelo de promedio móvil integrado autorregresivo y se especifica mediante tres parámetros de orden: (p, d, q).
B- *Modelo SARIMAX* El modelo básico de ARIMA se puede extender más allá incorporando la estacionalidad de la serie y variables exógenas. En este caso estaríamos hablado del modelo SARIMAX representado por (p, d, q) x (P, D, Q) S: donde los parámetros (P, D, Q) representan la misma idea que los (p, d, q) pero tratan sobre la parte estacional de la serie.

Vamos a utilizar el modelo SARIMAX, porque en su implementación existen herramientas adicionales que nos facilitan el análisis y que no están disponibles en la implementación de ARIMA.

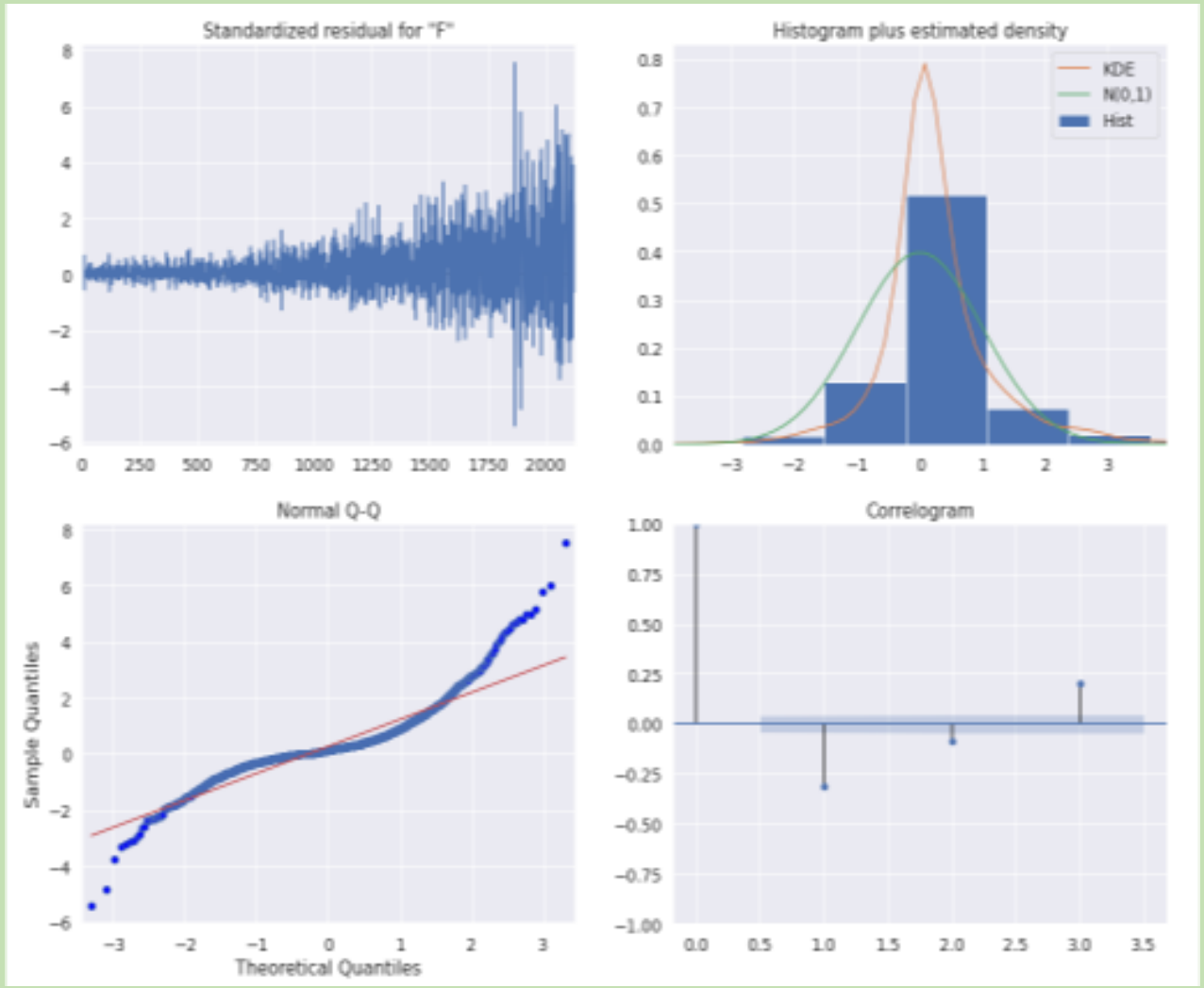
4- Interpretación del modelo:

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6925	0.009	75.107	0.000	0.674	0.711
sigma2	1.059e+15	3.71e-19	2.86e+33	0.000	1.06e+15	1.06e+15

Se busca por medio de los coeficientes estadísticos si el modelo seleccionado se ajusta a la serie de datos que se posee.
De acuerdo a estos coeficientes el modelo SARIMAX se ajusta a la serie de datos

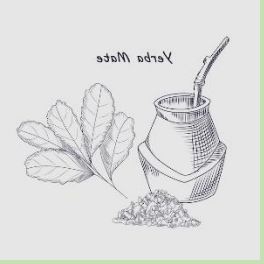
Modelo y parámetros óptimos

4- Interpretación del modelo:



Interpretando los gráficos podemos observar lo siguiente:

- Arriba a la izquierda: los residuos del modelo parece que siguen un proceso de Ruido Blanco (White Noise) y son predecibles. Esto implica que nuestro modelo ha extraído la mayor cantidad de información de los datos.
- Arriba a la derecha: vemos que la distribución de los residuos sigue una distribución próxima a la Normal (0, 1).
- Abajo a la derecha: vemos que la auto correlación parcial entre los residuos y residuos - k, dan lugar a valores significativos. Esto implica que el modelo ha sido capaz de reproducir el patrón de comportamiento sistemático de la serie.
- Abajo a la izquierda: la distribución ordenada de los residuos sigue una Normal.



Conclusiones

- Relación entre variables: Facturación y Costos tienen una relación directa, dado que esta segunda variable forma parte de la determinación de la primera.
- Series temporales: Poca estacionalidad en las variables.
- La estacionalidad de los productos se registra en períodos fríos que se registran entre Mayo a Septiembre.
- Económicos: La relación de las variables es directa, además se debe identificar el impacto de inflación en las precios.
- El rendimiento de las distintas líneas en ocasiones es negativa por la estacionalidad y el incremento en los costos.
- Modelo ML: Se logro optimizar un modelo SARIMAX para una empresa de productos de consumo masivo.