



# CODER HOUSE

## TRABAJO FINAL

COMISIÓN 14075 2021

---

EQUIPO

SAVER - LELOUTRE - MONCHIERO

# ÍNDICE

---

<b>1  CONTEXTO.....</b>	<b>02</b>
<b>2  OBJETIVOS GENERALES.....</b>	<b>03</b>
<b>3  DATOS.....</b>	<b>04</b>
<b>4  ANÁLISIS DE VENTAS EN EL TIEMPO.....</b>	<b>12</b>
<b>5  ANÁLISIS DE SERIES DE TIEMPO.....</b>	<b>13</b>
<b>6  CONCLUSIONES.....</b>	<b>20</b>

# CONTEXTO

## MERCADO LIBRE



**Mercado Libre** es la empresa líder y pionera en comercio electrónico y fintech en América Latina, basada en una plataforma donde los usuarios compran, venden, publicitan, entregan, financian y pagan bienes y servicios a través de Internet.

Está construyendo un ecosistema emprendedor que está democratizando el comercio, el dinero y los pagos, empoderando a millones de personas en América Latina.

# OBJETIVOS GENERALES

## FORECASTING

Construir un modelo para **pronosticar las ventas** de artículos de los 30 días siguientes en base a los datos históricos de Mercado Libre. Todos los días, la unidad de envío de Mercadolibre (MercadoEnvíos) entrega miles de productos en toda Latinoamérica.

Para brindar a los clientes la mejor experiencia durante todo el proceso de envío, es fundamental contar con pronósticos de ventas precisos.

En el centro de este problema comercial, se encuentra la tarea de **Desafío de datos**: dada la serie temporal de ventas históricas para un subconjunto de los listados del mercado, **lo desafiamos a predecir las ventas del siguiente período de cada SKU**.



# DATOS

En este caso los datasets ya se proveen de manera separada.

Este conjunto de datos comprende **dos meses** (febrero y marzo de 2021) de datos de **ventas a nivel diario** para un subconjunto de SKU de Mercadolibre (unidades de mantenimiento de existencias). Cada fila corresponde a una combinación de fecha-SKU particular.

Un **SKU** es una **combinación de un artículo y una variación**. Un artículo podría ser, por ejemplo, una "camiseta de la marca X" y una variación podría ser "talla M, color negro". Por lo tanto, estamos interesados en predecir el stock a nivel de SKU porque podría quedarse sin stock para el color negro, pero podría tener un montón de la camiseta X para el color marrón. Además de SKU y fecha, para cada fila, están disponibles los siguientes campos:

## DESCRIPCIÓN DE LOS ATRIBUTOS

- **sold\_sold\_quantity**: número de unidades del SKU correspondiente que se vendieron en esa fecha en particular.
- **current\_price**: punto en el tiempo precio correcto del SKU.
- **currency**: en la que se expresa el precio.
- **Listing\_type**: tipo de listado que tenía el SKU para esa fecha en particular. Los valores posibles son clásicos o premium y se relacionan con la exposición que tienen los artículos y la tarifa que se cobra al vendedor como comisión de venta. Otra ventaja importante de un artículo que figura como premium es su capacidad para pagar en cuotas sin tasa de interés.

- **shipping\_logistic\_type:** tipo de método de envío que ofrece el SKU, para esa fecha en particular. Los valores posibles son cumplimiento, cross\_docking y drop\_off.
- **shipping\_payment:** si el envío del SKU ofrecido en esa fecha en particular fue gratuito o pagado, desde la perspectiva del comprador.
- **minutes\_active:** número de minutos que el SKU estuvo disponible para su compra en esa fecha en particular.

El Data Frame cuenta con **37M de filas** con ventas de Argentina, Brasil, México, etc. por sku, a los efectos entender como funciona un model de serie de tiempo nos vamos a quedar con la **base sólo de Argentina y las columnas sku, date y sold\_quantity**.

(37660279, 9)									
	sku	date	sold_quantity	current_price	currency	listing_type	shipping_logistic_type	shipping_payment	minutes_active
0	464801	2021-02-01	0	156.78	REA	classic	fulfillment	free_shipping	1440.0
1	464801	2021-02-02	0	156.78	REA	classic	fulfillment	free_shipping	1440.0
2	464801	2021-02-03	0	156.78	REA	classic	fulfillment	free_shipping	1440.0
3	464801	2021-02-04	0	156.78	REA	classic	fulfillment	free_shipping	1440.0
4	464801	2021-02-05	1	156.78	REA	classic	fulfillment	free_shipping	1440.0

# MAESTRO DE PRODUCTOS

---

En el archivo **items\_static\_metadata.jl** hay algunos datos adicionales relacionados con las características de los SKU. Comprende una lista de dictados donde cada dictado contiene metadatos de SKU específicos. Están disponibles los siguientes campos:

- **SKU**: Unidad de mantenimiento de existencias de SKU. Este es un identificador único para cada unidad física de inventario diferente.
- **item\_id**: identificador único del listado al que pertenece el SKU. La misma lista se puede asociar con más de un SKU, por ejemplo, los SKU "camiseta X, talla M color negro" y "camiseta X, talla M, color rojo" comparten el mismo item\_id que es "t- camisa X".
- **item\_domain\_id**: ID de dominio de la lista. Un dominio es una especie de agrupación de listados dentro de MercadoLibre. Por ejemplo, "camiseta X" podría estar en el dominio MLB\_SPORT\_TSHIRTS.
- **item\_title**: el título de la lista, el cual está a nivel de artículo. Entonces, en el ejemplo anterior, un posible título podría ser "camiseta X".
- **site\_id**: el sitio de MercadoLibre al que pertenece el listado. Las etiquetas MLB, MLA y MLM se refieren a Brasil, Argentina y México respectivamente.
- **product\_id**: lista de identificación del producto. El campo puede ser nulo para algunos listados. Debido a que un artículo está listado para el vendedor, a menudo es común que diferentes vendedores vendan lo mismo (el mismo producto). Este podría ser el caso de la "camiseta X" si se vende a muchos vendedores. Entonces, en un intento de catalogar el mismo producto, MercadoLibre asigna el mismo product\_id a todos esos artículos.

- **product\_id\_family**: lista de identificación de familia de productos. El campo puede ser nulo para algunos listados. Es el mismo que el anterior, pero con un producto de catálogo de alta jerarquía.

	item_domain_id	item_id	item_title	site_id	sku	product_id	product_family_id
0	MLB-SNEAKERS	492155	Tênis Masculino Olympikus Cyber Barato Promoçao	MLB	0	None	MLB15832732
1	MLB-SURFBOARD_RACKS	300279	Suporte Rack Prancha Parede C/ Regulagem Horiz...	MLB	1	None	None
2	MLM-NECKLACES	69847	5 Collares Plateados Dama Gargantilla Choker -...	MLM	2	None	None
3	MLM-RINGS	298603	Lindo Anillo De Bella Crepusculo Twilight Prom...	MLM	3	None	None
4	MLB-WEBCAMS	345949	Webcam Com Microfone Hd 720p Knup Youtube Pc V...	MLB	4	None	None

## SET DE PRUEBAS

Para las pruebas, se proporciona el siguiente archivo **test\_data.csv**. Este archivo contiene solo dos columnas:

- **SKU**: indica el SKU para el que tiene que hacer su predicción.
- **target\_stock**: nivel de inventario (también conocido como número de unidades del SKU correspondiente para el que debe proporcionar su estimación de días de inventario).

# EDA

Utilizamos distintas técnicas de exploración para observar los datos del dataset y la relación que existe entre ellos.

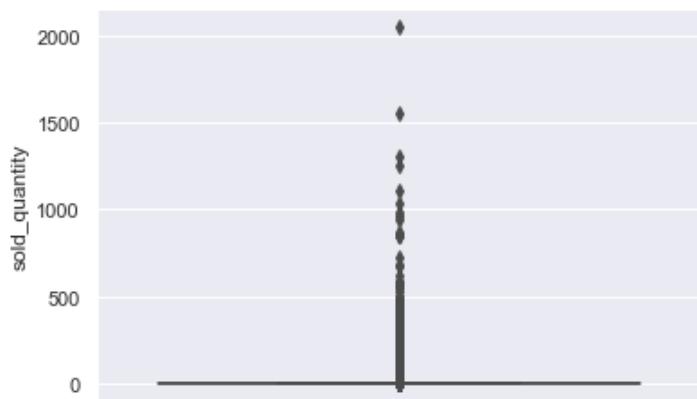
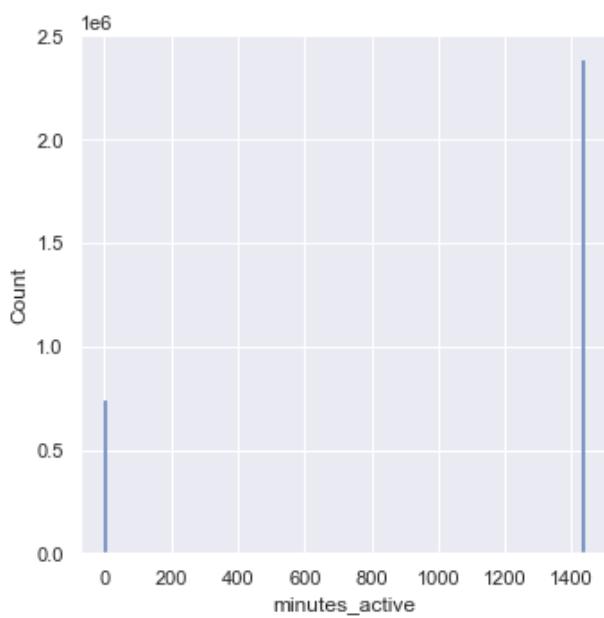
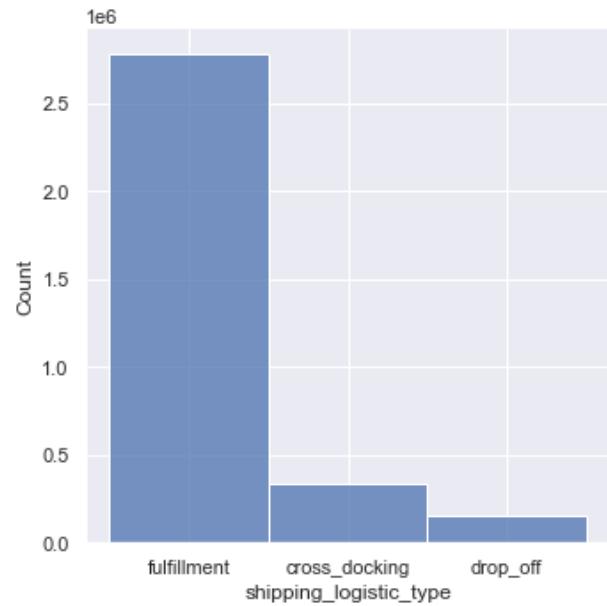
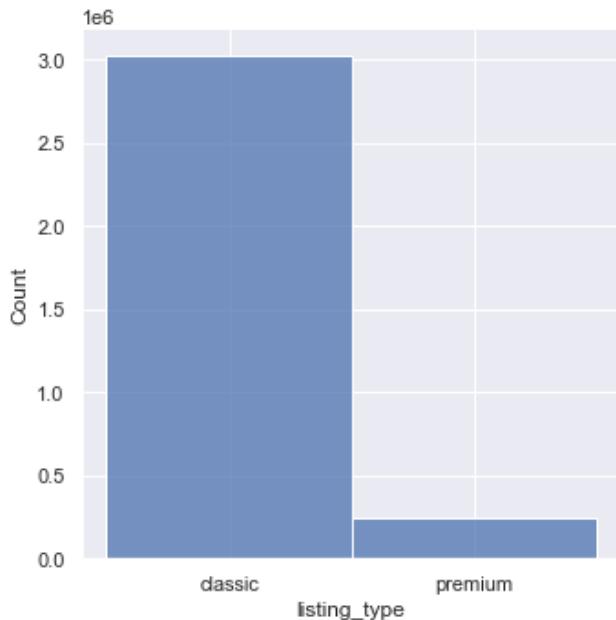
```
1 # Realizamos primeras tareas de análisis y exploración del dataset
2 train_data.describe().T      # Resumen de medidas estadísticas
```

	count	mean	std	min	25%	50%	75%	max
sku	714100.0	330129.571491	190259.010004	10.0	164838.0	331483.0	494720.0	660909.0
sold_quantity	714100.0	3.562266	10.079143	1.0	1.0	2.0	3.0	2049.0
current_price	714100.0	3477.457636	10364.793955	8.0	630.0	1485.0	3500.0	1799999.0
minutes_active	714100.0	1396.491831	174.134803	0.0	1440.0	1440.0	1440.0	1440.0
target_stock	714100.0	29.122804	89.708863	1.0	3.0	8.0	23.0	4335.0

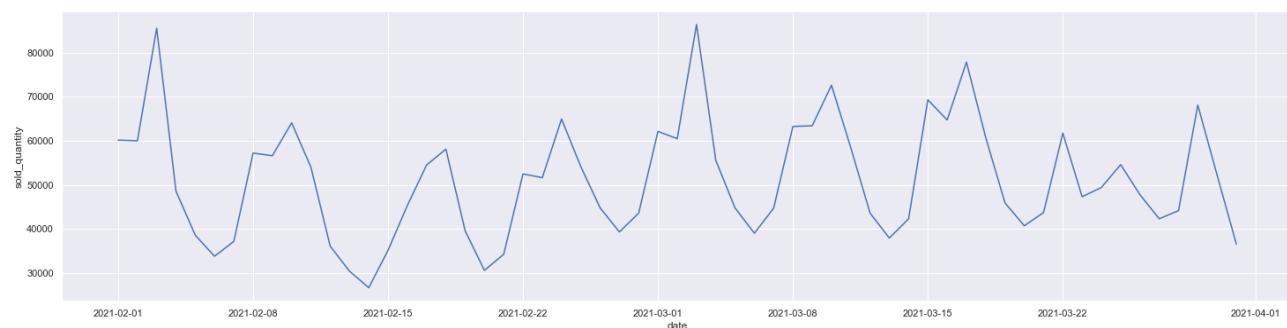
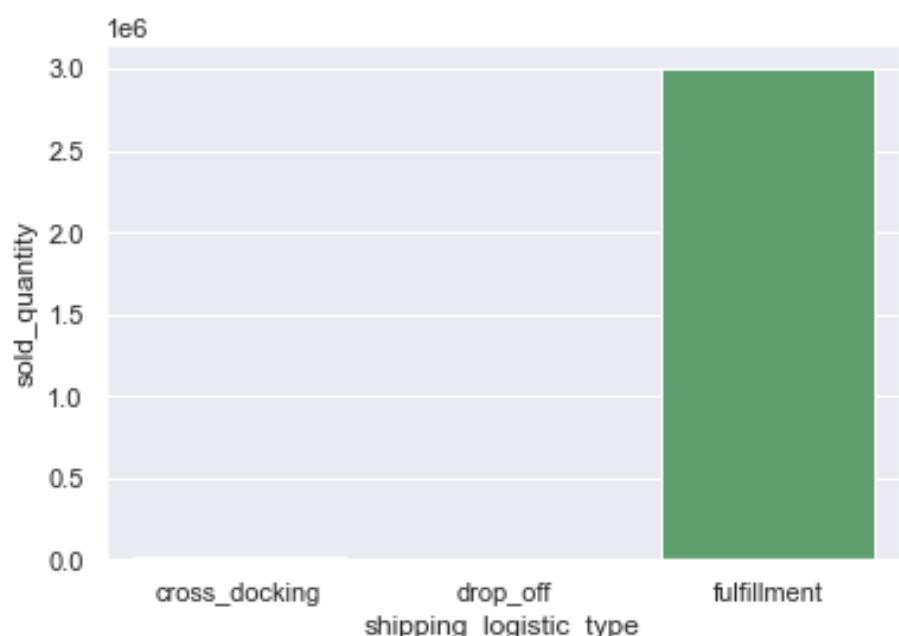
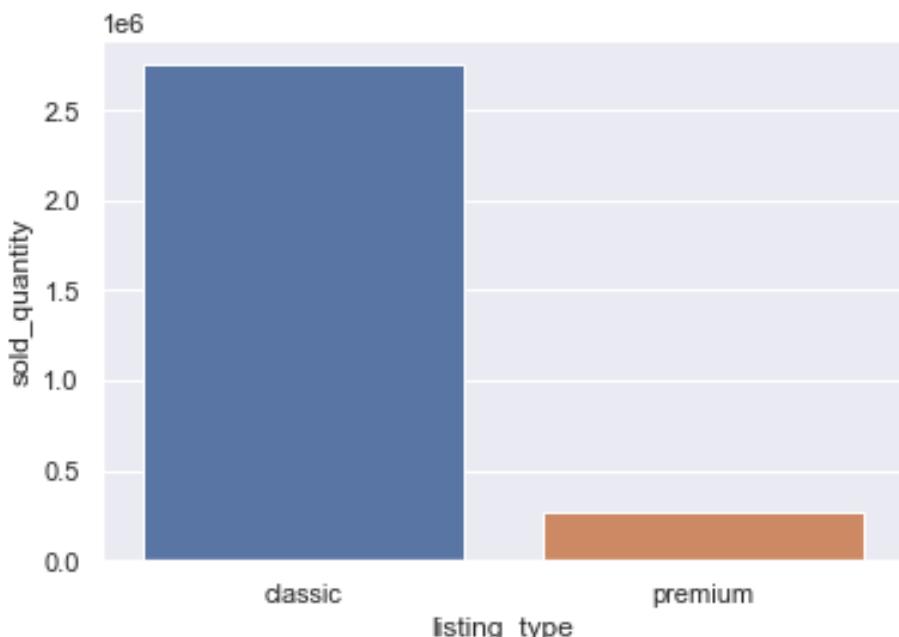
```
1 train_data.dtypes    # Resumen del tipo de datos del dataframe
```

```
sku                      int64
date                     object
sold_quantity            int64
current_price           float64
listing_type             object
shipping_logistic_type  object
shipping_payment         object
minutes_active          float64
target_stock             int64
dtype: object
```

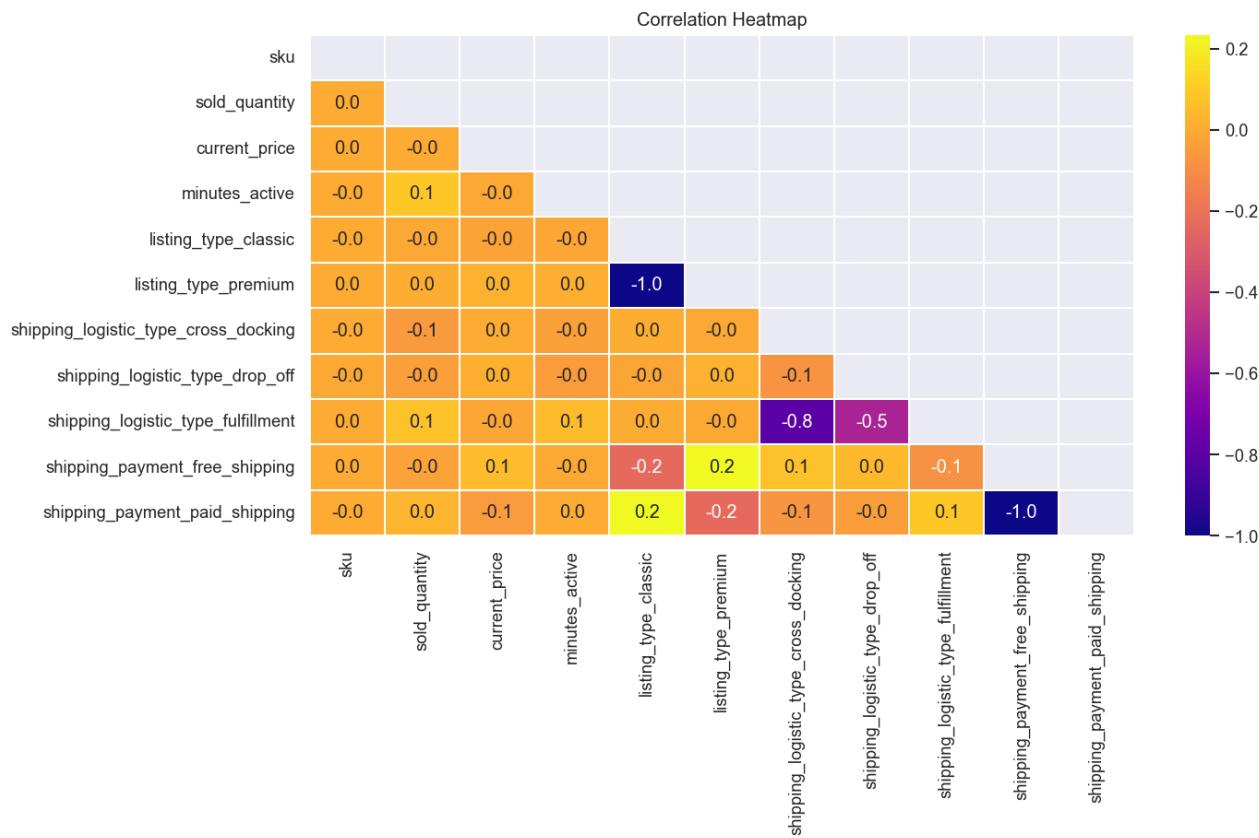
## ANÁLISIS UNIVARIADO



## ANÁLISIS BIVARIADO



## ANÁLISIS MULTIVARIADO

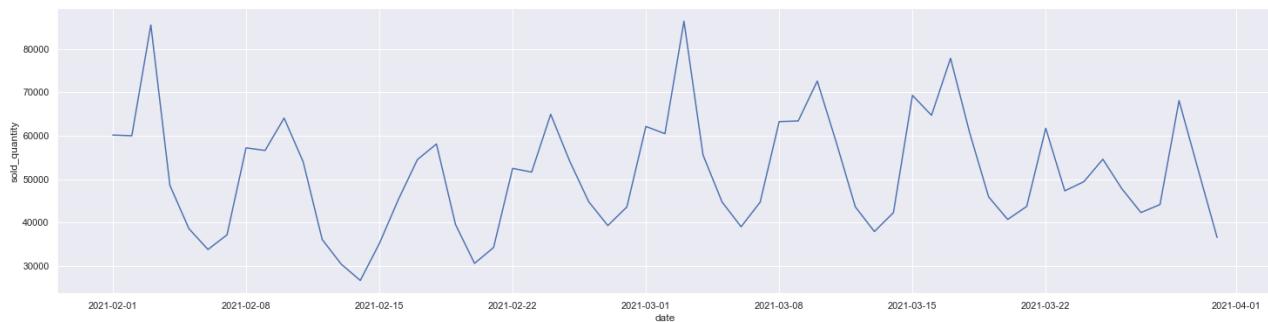


El análisis exploratorio muestra un **dataset muy desbalanceado** en todas sus variables y sin correlación entre ellas y la variable a predecir sold\_quantity.

Por tal motivo, desestimamos las mismas para el análisis de predicción de ventas en el tiempo.

# ANÁLISIS DE VENTAS EN EL TIEMPO

En base a los dos meses de información histórica brindados por Mercadolibre analizamos el comportamiento en un gráfico:



Del mismo, podemos observar que existe una estacionalidad asociada a los datos. Cada semana tiene un día de **ventas pico** y un dia de **ventas mínimas**.

# ANÁLISIS DE SERIES DE TIEMPO

El análisis de series de tiempo es una de las habilidades centrales de cualquier científico de datos y cualquier persona que trabaje en el campo de la analítica a menudo se encuentra con la situación para pronosticar el futuro con los datos pasados y presentes.

## ¿Qué es una serie temporal?

Es una **serie de observaciones tomadas en momentos específicos** básicamente a intervalos iguales. Se utiliza para **predecir valores futuros basados en valores pasados** observados. Los componentes que puede observar en el análisis de series de tiempo son **Tendencia, Estacional, Irregular y Ciclicidad**.

En el caso de estos conjuntos de datos, donde solo se observa una variable en cada momento, se denomina "**Serie de tiempo univariante**" y si se observan dos o más variables en cada momento, se denomina "**Serie de tiempo multivariante**".

En este artículo, nos centraremos en la serie de tiempo univariante para pronosticar las ventas con la funcionalidad **Auto ARIMA en Python**, que es casi similar a Auto ARIMA en R.

## ¿Por qué utilizar Auto ARIMA?

Por lo general, en el modelo ARIMA básico, **necesitamos proporcionar los valores p, d y q** que son esenciales. Usamos técnicas estadísticas para generar estos valores realizando la diferencia para eliminar la no estacionariedad y trazando gráficos ACF y PACF. En Auto ARIMA, **el modelo mismo generará los valores óptimos de p, d y q** que serían adecuados para que el conjunto de datos proporcione un mejor pronóstico.

## Crear la serie de tiempo

Si bien el dataset muestra las ventas por día por sku, a los efectos de ejecutar un primer análisis **agrupamos las ventas por día**.

## PRUEBA DE ESTACIONARIEDAD

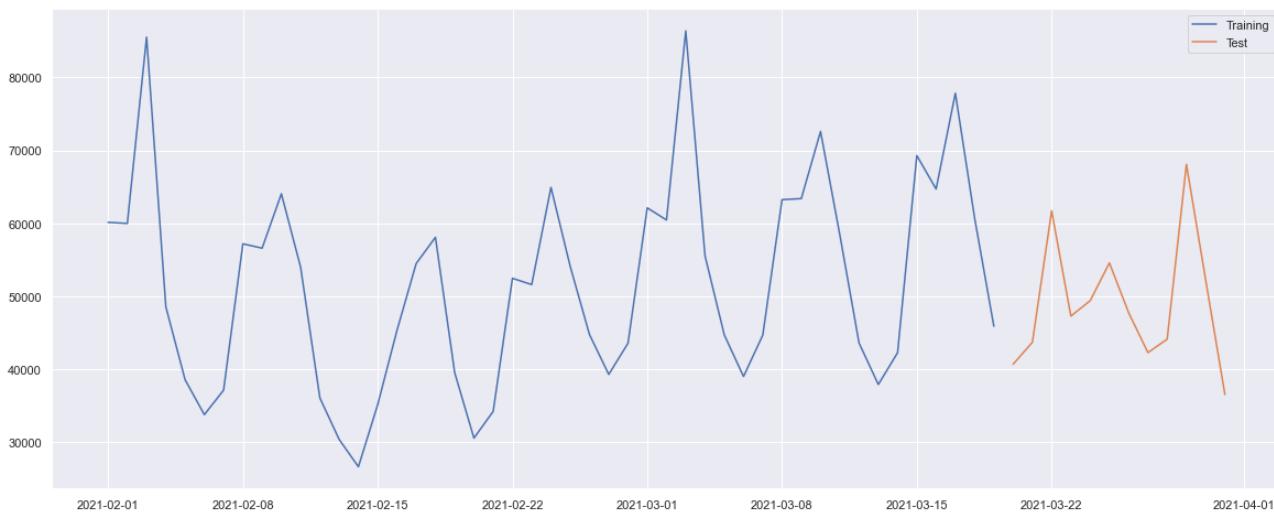
La estacionariedad es un concepto importante en las series de tiempo y **cualquier dato de la serie de tiempo debe someterse a una prueba de estacionariedad** antes de proceder con un modelo.

Usamos la "**Prueba Dickey-Fuller aumentada**" para verificar si los datos son estacionarios o no, lo cual está disponible en el paquete "**pmdarima**".

De lo anterior, podemos concluir que **los datos no son estacionarios**. Por lo tanto, necesitaríamos usar el concepto "**Integrado (I)**", denotado por el valor "d" en series de tiempo para hacer que los datos sean estacionarios mientras se construye el modelo Auto ARIMA.

## Train and Test split

Dividir el dataset en entrenamiento y test.



## Construir un modelo Auto Arima

- **Auto-regresivo (p):** Número de términos autorregresivos.
- **Integrado (d):** Número de diferencias no estacionales necesarias para la estacionariedad.
- **Media móvil (q):** Número de errores de pronóstico retrasados en la ecuación de predicción.

En el modelo Auto ARIMA, hay que tener en cuenta que los **valores pequeños de p, d, q representan componentes no estacionales**, y los **valores P, D, Q mayúsculos representan componentes estacionales**. Funciona de manera similar a las técnicas de hiper tuning para encontrar el valor óptimo de p, d y q con diferentes combinaciones y los valores finales se determinarían teniendo en cuenta los parámetros de AIC más bajos, BIC.

Aquí, **estamos probando con los valores p, d, q que van de 0 a 10** para obtener mejores valores óptimos del modelo.

## Comparar lo Real vs. lo Predecido

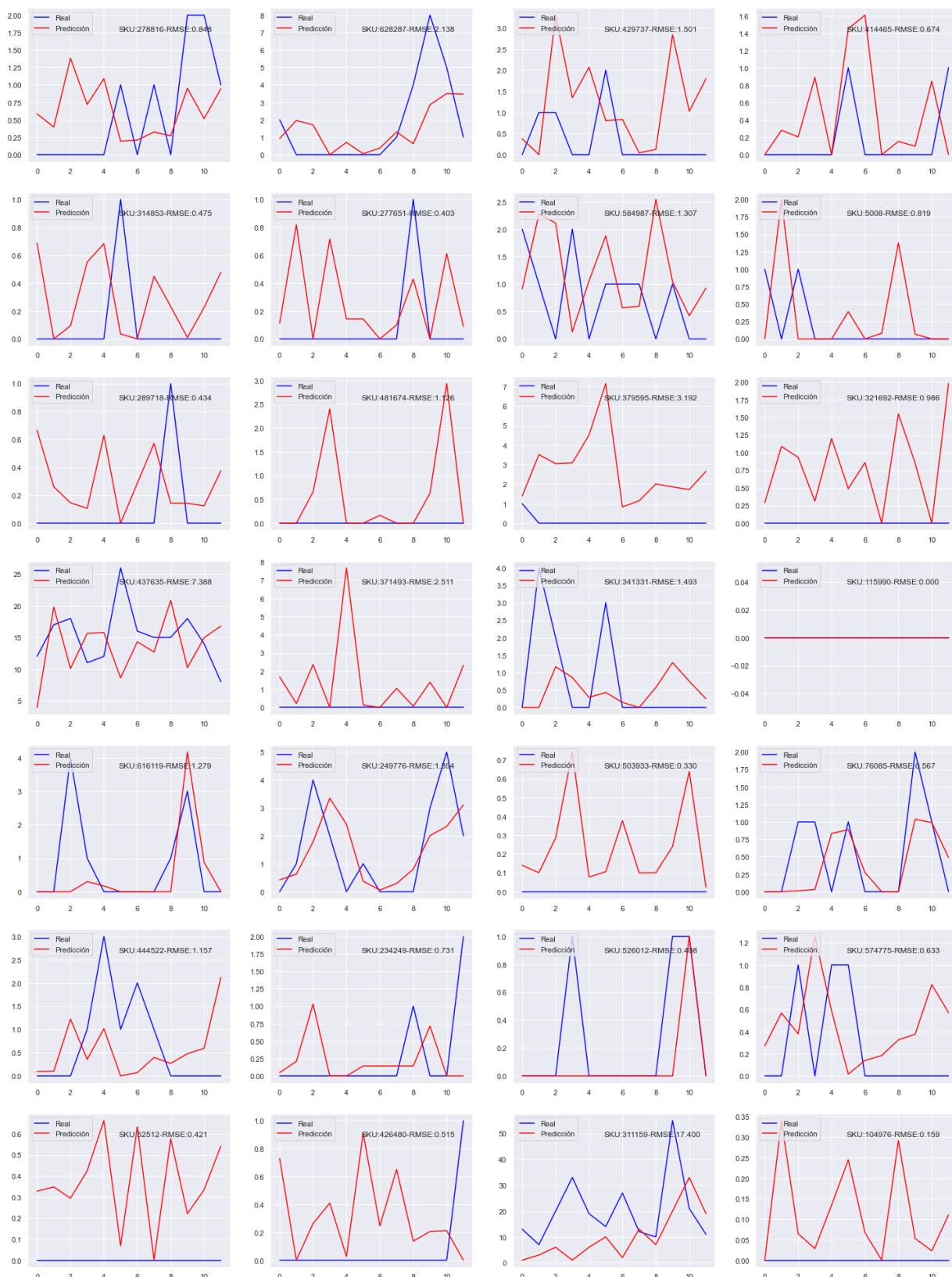


Utilizando la misma lógica anterior vamos a realizar las **predicciones de ventas del mes de Abril**.

Como cada sku tiene su comportamiento buscamos los mejores parametros p,d,q y P,D,Q para cada uno de ellos. Estos valores los vamos a grabar en el archivo de resultados para luego estudiar si tienen algún comportamiento.

Graficaremos para cada sku las comparación entre lo real y lo predecido.

## ALGUNAS PREDICCIONES



## CLUSTERIZACIÓN DE SERIES DE TIEMPO

Este formato de trabajo nos llevaría a realizar modelos para tantas series de tiempo como sku existen en el dataset.

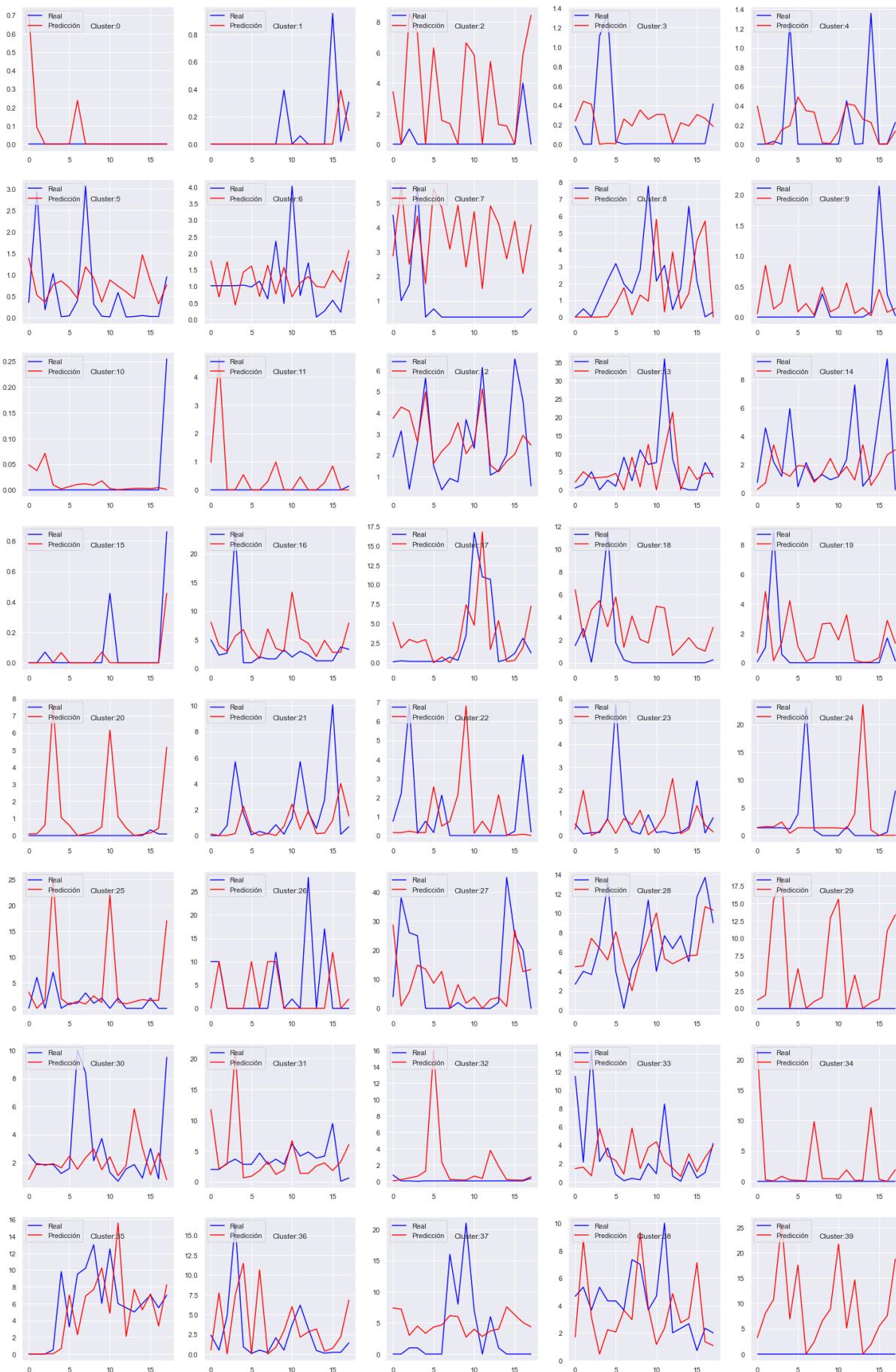
Por tal motivo realizamos una clusterización del dataset por sku para agrupar los mismos y realizar predicciones grupales.

La agrupación o clustering es uno de los **métodos de extracción de datos más populares**, no sólo por su poder de exploración, sino también como paso previo al procesamiento de otras técnicas.



\*adjuntamos clusterización completa

## PREDICCIONES EN BASE A LOS CENTROIDES



\*adjuntamos predicción completa

# CONCLUSIONES

---

En base a los datos recibidos por MercadoLibre, sobre sesenta días de ventas para distintos artículos en tres países diferentes, dentro de los meses de febrero y marzo, pudimos aplicar una **estimación del comportamiento de las ventas a nivel global** (filtrando la información sobre Argentina).

En esta última, aplicamos en base a las ventas de los primeros treinta días una estimación para los próximos treinta, y ese resultado lo comparamos con los datos reales, y, en consecuencia, **obtuvimos una estimación global muy acertada respecto al comportamiento de las ventas en el país para esos períodos.**

No obstante, cabe aclarar que los datos brindados por MercadoLibre son muy acotados, con muy pocas variables a considerar por cada SKU, motivo por el cual **es muy poco probable que con este imput las estimaciones realmente sean válidas a la hora de llevarlo a una práctica real**, aunque la metodología consideramos es correcta.

Por consiguiente, luego del análisis global de la predicción de la demanda, lo llevamos a un nivel más micro, es decir, por cada SKU, pero notamos que, al tener un nivel de inventario tan grande, esa metodología era incorrecta. Por lo tanto, **establecimos distintos clusters sobre los SKU, en base al comportamiento de cada uno a lo largo del tiempo.** En ese sentido pudimos recabar que dentro de esas categorías **hay muchos SKU con las mismas tendencias de ventas**, y que en algunos casos la predicción coincide o es muy parecida con respecto al comportamiento real observado y en otros casos no.

Finalmente, podemos decir luego de agrupar los SKU que, **si bien en la globalidad la tendencia es muy marcada y podemos tener una predicción**, también podemos obtener la misma, sobre algunos clusters de SKU, no obstante, **no podemos afirmar con certeza estas definiciones dado que no tenemos todas las variables para analizar por cada SKU**, y a la vez, **nuestra base de datos en función del tiempo es muy acotada** como para tener una certeza real sobre estos comportamientos.