AGUACATES

Análisis y modelado predictivo del precio

Armen Hakobyan

RESUMEN

El aguacate ha sido siempre un alimento muy valorado por sus consumidores. Desde hace no mucho que empecé a disfrutar de este fruto y cada día me gusta más. Este producto se caracteriza por un precio considerablemente elevado y cuando iba a los supermercados veía que variaban mucho dependiendo de la época del año y el tipo, si era “eco” o convencional. Durante una búsqueda en la plataforma Kaggle, se identificó un conjunto de datos que reflejaba el volumen y precio de los aguacates, pero de EEUU. Decidí analizarlos con el fin de encontrar patrones que se repiten a lo largo del año, así como tratar de predecir los acontecimientos futuros.

PALABRAS CLAVE

Estacionalidad: Hablamos de estacionalidad cuando nos referimos a un comportamiento que se repite en ciertas épocas del año.

Dataset: Un dataset es un conjunto de datos distribuidos en un archivo, ya sea un Excel, CSV o del tipo que sea.

Dataframe: Nos referimos con dataframe al conjunto de datos cargados en herramientas de transformación de datos cuando trabajamos con código.

Características: Se entiende como característica las variables predictoras de un dataframe.

Visualizaciones: Son gráficas para mostrar de manera más clara la información.

RMSE: Significa Root Mean Square y es un método de medición de la precisión de un modelo.

Interpolación: Se refiere a la capacidad de un modelo de IA para realizar predicciones o generar datos dentro del rango de los datos con los que fue entrenado. Esencialmente, estima valores desconocidos que se encuentran entre puntos de datos conocidos.

Extrapolación: Implica que un modelo de IA realice predicciones o genere datos más allá del rango de los datos originales de entrenamiento. Es decir, intenta predecir en situaciones o con valores que no ha visto explícitamente durante su aprendizaje.

ABSTRACT

The avocado has always been a highly valued food by its consumers. It wasn't long ago that I started to enjoy this fruit, and I like it more each day. This product, as we all know, has a fairly high price, and when I went to supermarkets, I saw that it varied a lot depending on the time of year and the type, whether it was "eco" (organic) or conventional. Browse Kaggle, I found a dataset that reflected the volume and price of avocados, but from the USA. I decided to analyze them in order to find patterns that repeat throughout the year, as well as to try to predict future events.

KEYWORDS

Seasonality: We talk about seasonality when we refer to a behavior that repeats at certain times of the year.

Dataset: A dataset is a collection of data distributed in a file, whether it's an Excel, CSV, or any other type.

Dataframe: We refer to a dataframe as the set of data loaded into data transformation tools when we work with code.

Features: Features are understood as the predictor variables of a dataframe.

Visualizations: These are graphs used to display information more clearly.

RMSE: Stands for Root Mean Squared Error and is a method for measuring the accuracy of a model.

Interpolation: Refers to the ability of an AI model to make predictions or generate data within the range of the data it was trained on. Essentially, it estimates unknown values that lie between known data points.

Extrapolation: Implies that an AI model makes predictions or generates data beyond the range of the original training data. That is, it attempts to predict in situations or with values it has not explicitly seen during its learning process.

Contenido

INTRODUCCIÓN

He decidido hacer este proyecto porque la verdad me encantan los aguacates. Los utilizo siempre que puedo, pero sin pasarme. Lo llaman el oro verde por el precio que tiene. Esta observación sobre la variabilidad de los precios despertó el interés por investigar sus causas. Por este motivo he decidido hacer este proyecto sobre el análisis de los precios del aguacate.

Se seleccionó una base de datos de EEUU debido a la mayor disponibilidad de datos estructurados adecuados para el análisis.

Mi objetivo con este proyecto es desarrollar los conocimientos aprendidos en clase para realizar un análisis de los datos y ver a qué se deben estos precios tan variados y hacer un modelo predictivo del precio por regiones de EEUU.

Utilizaremos distintos modelos de clasificación y regresión, diferentes técnicas de ingeniería de características para extraer un mayor valor de los datos, exploración y corrección de los datos, etc.

También Power BI y R para visualizar los datos.

El dataset lo podemos encontrar en Kaggle.

MARCO TEÓRICO

Este proyecto lo estoy haciendo con el fin de poder profundizar y familiarizarme en estos temas:

- Limpieza y corrección de datos: python junto a pandas, matplotlib, numpy… visualmente
- Normalización y estandarización: estas dos técnicas mejoran el rendimiento de los modelos.
- Detección de outliers: las gráficas de boxplot y violín nos ayudan en este aspecto
- Estadísticas descriptivas: nos ayuda a entender los datos.
- Visualización de datos: histogramas, diagramas, R, Power BI.
- Correlación de variables: entender y analizar correlaciones entre las características.
- Modelos predictivos y clasificadores: regresión lineal y polinómica, random forest, KNN, árboles de decisión y máquinas vectoriales.
    - Waka/Orange: Útiles para generar modelos de forma visual.

En resumen, las herramientas utilizadas en este proyecto son:

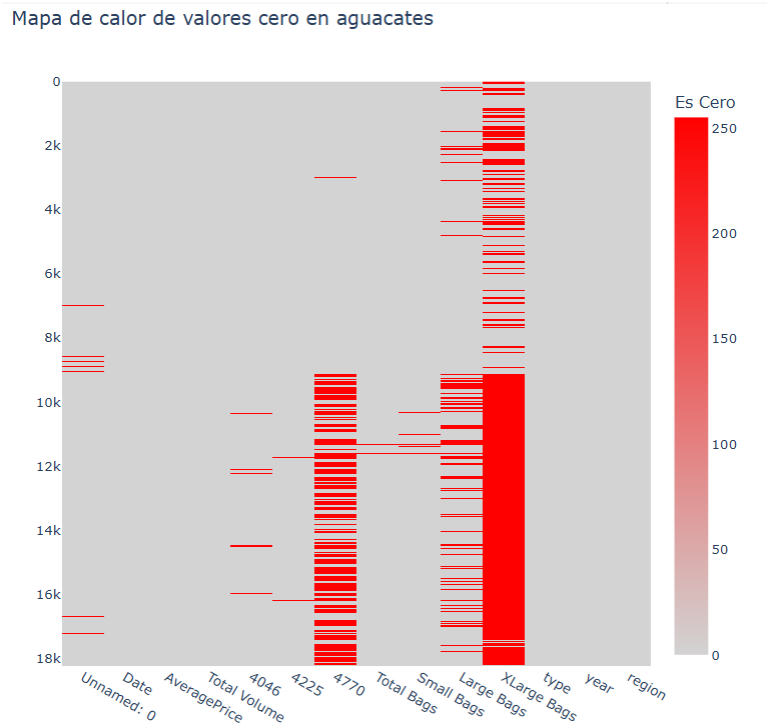
ORANGE	PYTHON	R
POWERBI	TABLEAU	GOOGLE COLLAB
GITHUB	VIRTUAL BOX	EXCEL

OBSERVACIÓN PREVIA

Como he mencionado anteriormente, la base de datos la he encontrado en Kaggle y contiene 18.250 registros y 13 características.
Estas características son:

- Date: Fecha del registro (semanal)
- AveragePrice: Precio promedio
- Total Volume: El volúmen total
- 4046: Tipo de aguacate Hass pequeño
- 4225: Tipo de aguacate Hass grande
- 4770: Tipo de aguacate Hass extragrande
- Total bags: Cantidad de bolsas vendidas totales
- Small Bags: Bolsas pequeñas
- XLarge Bags: Bolsas grandes
- type: Tipo
- year: Añi
- region: Lugar, ciudad.

El dataset no contiene datos vacíos, pero si tiene datos que son 0. Para poder ver con una visión más amplia de cuantos datos estamos hablando, he utilizado herramientas de visualización para generar la siguiente visualización: (documento AGUACATES ANALISIS DE DATA)



Se observa como en XLarge Bags nos encontramos con muchísimos datos que son 0. También participan, aunque en menor medida, las características 4046, 4225, 4770, Large Bags, Small Bags, y Total Bags.
Vemos también una columna llamada Unnamed.

Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany

Podemos ver que es un simple índice, por lo que podemos prescindir de él.

Llegados a este punto, primero vamos a comprobar que no hayan fechas repetidas. Es importante que esta columna sobre todo esté en correcto formato ya que es nuestro indicador del tiempo.
Para realizar esta comprobación, utilizaremos **Pandas** para aplicar una máscara booleana para comprobar filtrar los valores que sean duplicados. Debemos separar los datos por tipo y por región, ya que cada tipo tiene su propia fecha y entre regiones se van repitiendo también las fechas. Una vez agrupados, aplicaremos el método .transform() con una función lambda que llama al método duplicated que aplica sobre la serie Date **de cada grupo** y .transform() devuelve una serie booleana con el mismo índice.

```
mascara_duplicados_tipo_region = df.groupby(['type', 'region'])['Date'].transform(lambda x: x.duplicated(keep=False))

# Filtramos el DataFrame original 'df' usando la máscara booleana.
df_duplicados_tipo_region = df[mascara_duplicados_tipo_region]

print("--- Filas con fechas duplicadas DENTRO de cada combinación (type, region) ---")
if df_duplicados_tipo_region.empty:
  print("No se encontraron fechas duplicadas dentro de ninguna combinación (type, region).")
else:
  # Ordenamos por type, luego region y finalmente fecha para ver los duplicados agrupados.
  print(df_duplicados_tipo_region.sort_values(by=['type', 'region', 'Date']))
```

```
--- Filas con fechas duplicadas DENTRO de cada combinación (type, region) ---
No se encontraron fechas duplicadas dentro de ninguna combinación (type, region).
```

Para tratar los datos completados con 0 de las columnas mencionadas es necesario entender a que se deben estos valores, es decir, identificar porque son 0. Esta situación puede generar ambigüedad, dado que un valor de cero no implica necesariamente un error, sino que podría representar una cantidad tan reducida que se aproxima a cero tras el
Se procedió a verificar si la suma de los diferentes tipos de bolsa coincidía con el valor de “Total Bags”. Una discrepancia significativa en los casos donde 'XLarge Bags' u otra categoría similar fuese cero indicaría datos faltantes. En ausencia de tal discrepancia, se asumiría la no comercialización de dicho producto en esas zonas, excluyéndolo del ajuste.
Al ejecutar el código me encuentro con varios casos en los que no es igual, pero al aplicar un error +-1, no se encuentra ningún resultado erróneo. Aplicamos este error por tema de redondeos, que tenga un márgen.

```
--- Comprobando si 'Total Bags' está dentro de +/- 1 de la suma ('Small Bags' + 'Large Bags' + 'XLarge Bags') ---

¡Comprobación exitosa! En todas las filas la diferencia entre 'Total Bags' y la suma de bolsas individuales está dentro de la tolerancia de +/- 1.
```

COMPLETANDO LOS VALORES “0”

Comprobamos cuántos 0 hay de las Bags de orgánicos y cuantos en convencional.

```
--- Contando ceros en 'Small Bags', 'Large Bags', 'XLarge Bags' por 'type' ---

Número de veces que aparece un 0 en cada columna, agrupado por tipo:
              Small Bags  Large Bags  XLarge Bags
type
conventional           0         371         3070
organic              159        1999         8978
```

Para tratar estos datos lo voy a enfocar de la siguiente manera:

- Agrupar por tipo y por región

- Crear una nueva columna que será la Fecha Ordinal, que es una variable puramente numérica que indica los días que han pasado desde 01/01/0001. Esto ayudará a nuestro modelo a realizar mejores análisis.

```
df['Fecha_Ordinal'] = df['Date'].apply(lambda date: date.toordinal())
```

- Se entrenará un modelo por cada tipo de aguacate y región.

- Si del tipo de aguacate la región tiene un 50% de datos que son 0, no se generará ningún modelo y los datos se quedarán como están.

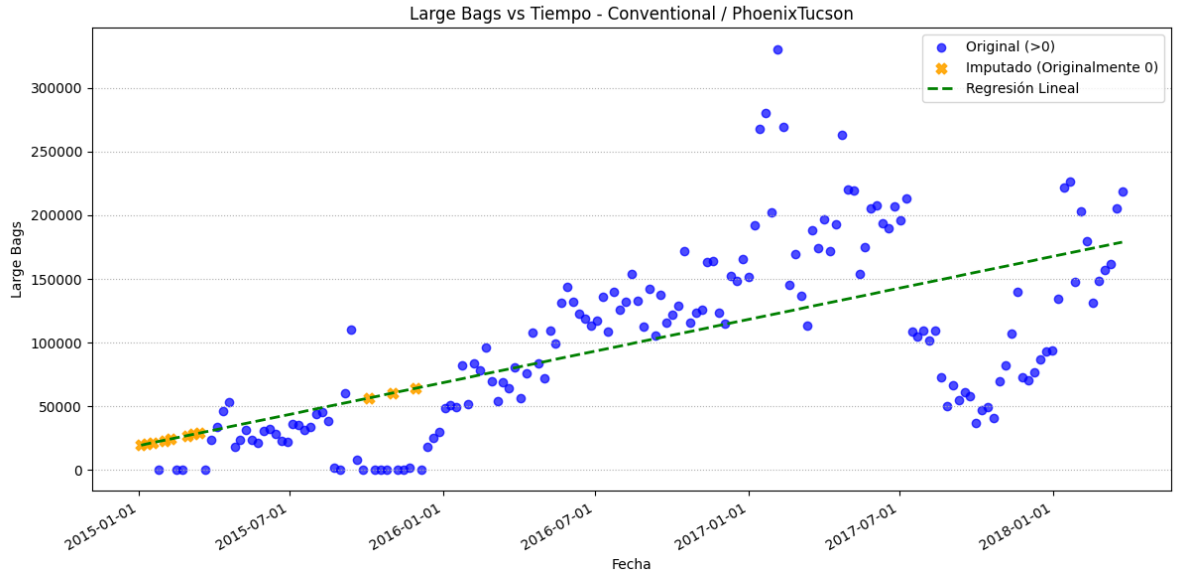
- De lo contrario, se generará un modelo de regresión lineal que servirá para rellenar los datos que son 0 con el valor predicho, resultará útil para obtener estimaciones más cercanas a la realidad, a pesar de su limitada precisión y su independencia de la estacionalidad.

- Los datos que se rellenen, se sumarán a Total Bags y a Total Volume.

Para este procedimiento no he separado los datos en entrenamiento y prueba puesto que es una regresión lineal simple y esto como mucho nos va a indicar la tendencia, son valores muy aproximados.

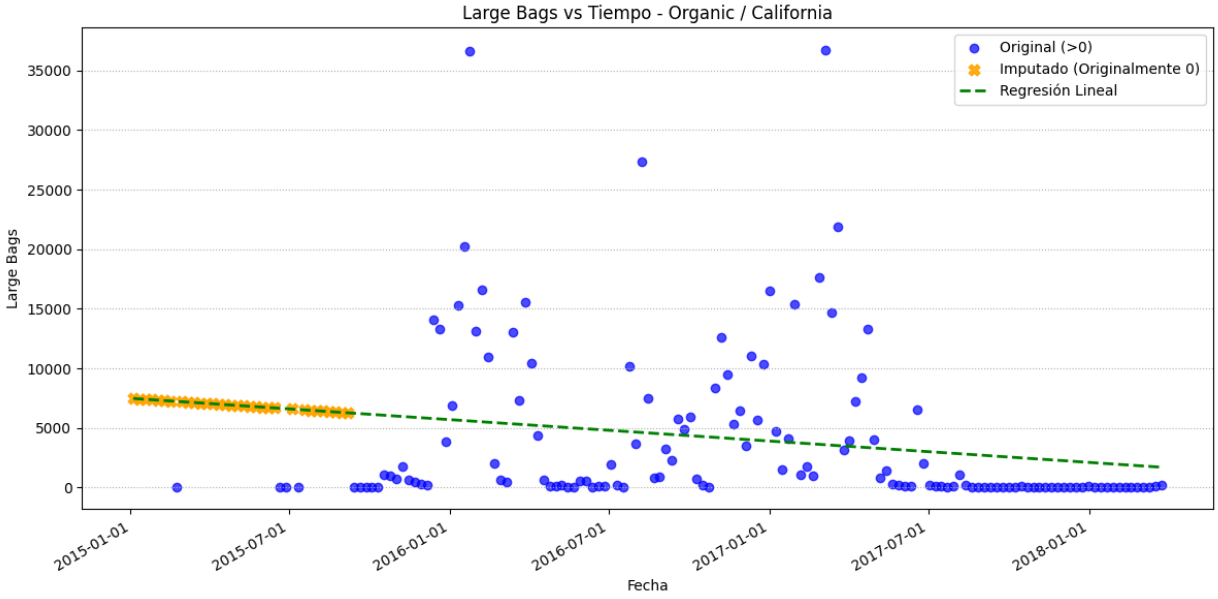
Una vez tenemos el modelo preparado, podemos indicarle de qué región región queremos graficar el modelo y las predicciones para los valores 0. Por ejemplo podemos ver el caso de PhoenixTucson. (documento AGUACATES ANALISIS DE DATA)

```
Generando gráfico para: type='conventional', region='PhoenixTucson'
   -> Intentando re-entrenar modelo con 158 puntos originales (>0).
```



O California, que se observa una disminución en la demanda de bolsas grandes de tipo orgánico con el transcurso del tiempo.

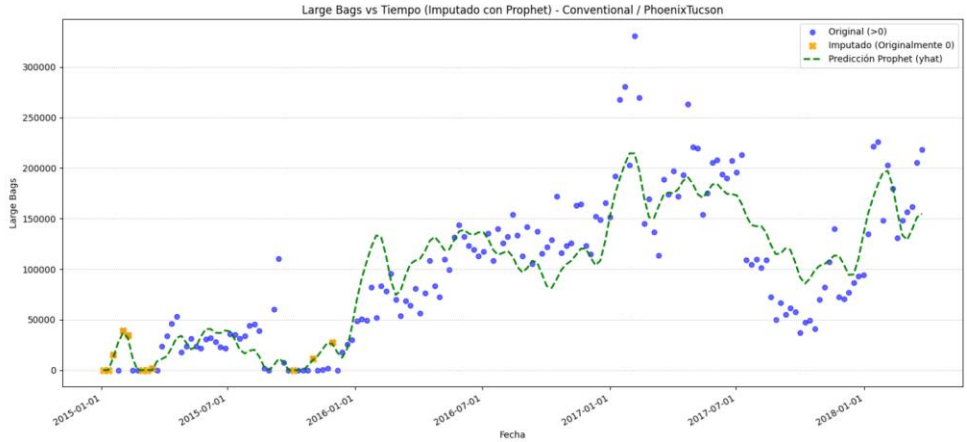
```
Generando gráfico para: type='organic', region='California'
   -> Intentando re-entrenar modelo con 137 puntos originales (>0).
```



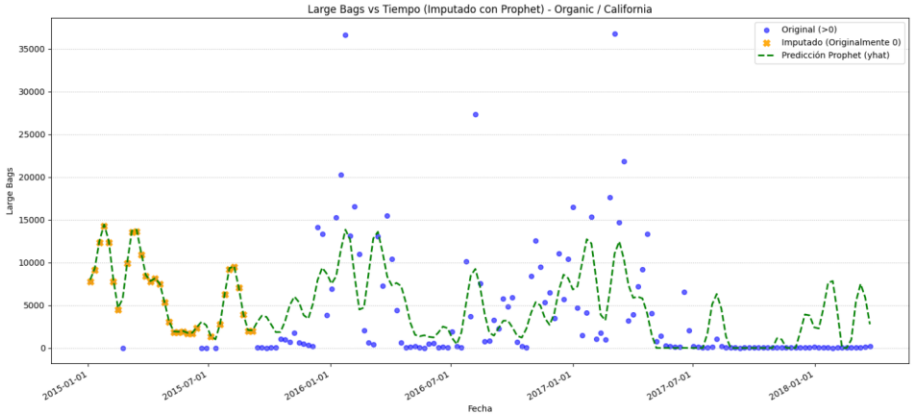
Se constata que esta imputación podría influir en los datos de manera significativa, sin garantizar una mejora en la calidad de los mismos . Si es cierto que en algunos casos la predicción no es tan mala, pero por ejemplo en california, al rellenar las bolsas vendidas con una regresión lineal nos da lugar a muchas confusiones y altera la veracidad de los datos.

PROPHET Y KNN

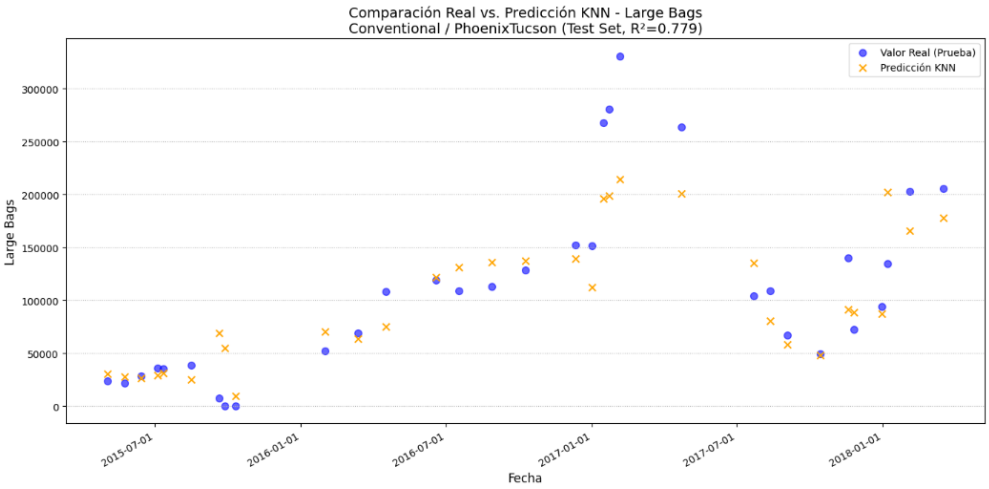
Considerando que las predicciones previas no resultaron óptimas, se procedió a evaluar modelos alternativos para determinar la conveniencia de imputar estos valores cero.
Aunque el volumen total si que aumenta estacionariamente, los valores separados como Large Bags no necesariamente sigue ese patrón. Consecuentemente, las predicciones del modelo Prophet resultaron poco efectivas, introduciendo una distorsión considerable y potencialmente errónea en los datos.
Siguiendo con los ejemplos del caso anterior, para PhoenixTucson vemos que prejudice mejor que el modelo de regresión lineal pero sigue presentando limitaciones en su utilidad, dado que distorsiona significativamente los datos, lo cual podría tener un impacto negativo considerable en el análisis, ya que en casos de regiones en los que tengan muchos datos vacíos, se completarán no de manera muy precisa y esto nos puede llevar a conclusiones erróneas. Este modelo tiene un R^2 de 0.75, lo que no es malo del todo pero de nuevo, tampoco óptimo. (documento AGUACATES ANALISIS DE DATA)



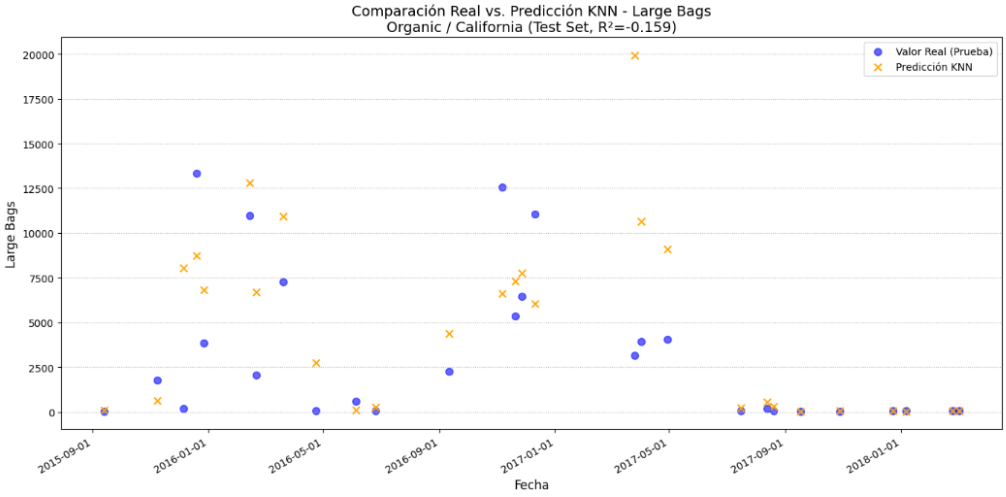
Se observa que, si bien el modelo presenta ciertos aspectos positivos, frecuentemente podría generar más inconsistencias que beneficios informativos. Con los orgánicos de California vemos que sigue una tendencia algo más clara, y aunque en el final del periodo 2018 decae bastante, mi modelo sigue prediciendo el patrón estacional. Este modelo tiene un R^2 calculado a partir de todos los datos existentes en Large Bags, no está separado para que el modelo se nutra de todos los datos existentes para rellenar los faltantes. Este modelo tiene R^2 de 0.35.



Dado a este problema, decidí hacerlo con KNN estacional, que compara los vecinos más cercanos también de las otras fechas parecidas y aunque pueda parecer mejor, tampoco resulta una solución implementable de forma generalizada, al menos no para todas las regiones, para dependiendo que regiones funciona mejor o peor.
Continuando con los ejemplos anteriores, podemos ver PhoenixTucson:



Para California: (documento KNN- ANÁLISIS DE DATA 2)



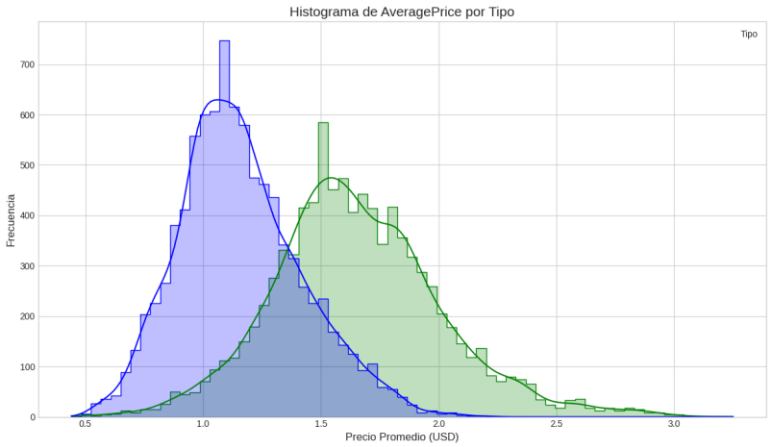
No obstante, estas predicciones continúan ofreciendo una utilidad limitada, puesto que una imputación exitosa simplemente añadiría una cantidad X a cada región, neutralizando el efecto en análisis comparativos.

Consecuentemente, tras estas evaluaciones, se optó por mantener los valores cero originales en el conjunto de datos.

ANÁLISIS DE PRECIOS

HISTOGRAMA

Los histogramas nos permiten hacer recuento para ver las veces que se repite un valor de una variable. En el siguiente histograma observamos el recuento de los valores del precio del tipo convencional de color azul y los valores de los orgánicos en verde. Los precios de los orgánicos está entre un rango más elevado y el precio que más se repite es mayor que la variante convencional. Aunque este tipo de aguacates no tiene tanta comercialización como los aguacates convencionales, el método de obtención es más complejo y por ello su precio es mayor, alcanzando máximos de hasta 3 dólares. (documento AGUACATES ANÁLISIS DE DATA)

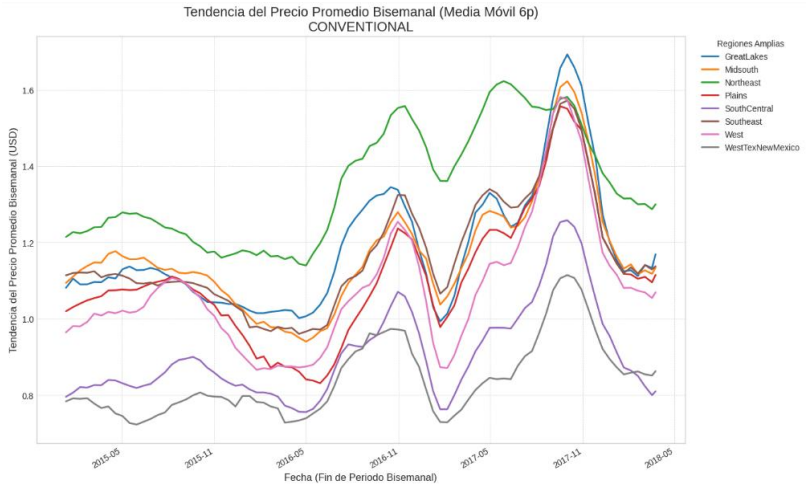


SERIES TEMPORALES – REGIONES ÁMPLIAS

La siguiente gráfica nos muestra la tendencia que toman los datos de volumen total respecto al tiempo. He agrupado los datos bisemanalmente, haciendo la media entre las dos primeras semanas y las dos últimas. Para visualizar la tendencia de manera más clara y sin tanto ruido, he aplicado la técnica de la media móvil. Este método se basa en definir un periodo determinado y calcular la media con los datos de ese periodo.

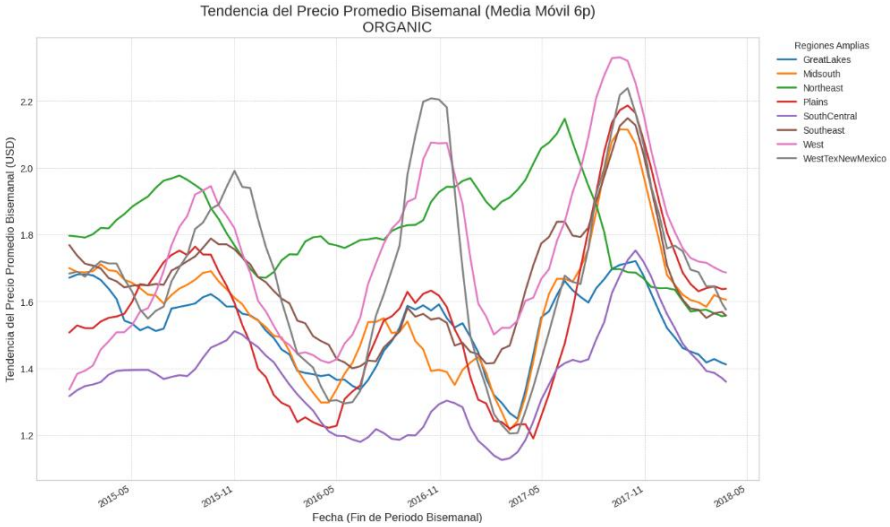
```
# --- Calcular Tendencia con Media Móvil ---

# Definir la ventana para la media móvil (6 periodos = ~3 meses)
window_size = 6
# min_periods: mínimo de observaciones en la ventana para producir un valor (útil al principio/final)
# center=True: la etiqueta de la media móvil se alinea con el centro de la ventana
print(f"\nCalculando tendencia con media móvil (ventana={window_size} periodos)...")

trend_conv = regions_conv_to_plot.rolling(window=window_size, center=True, min_periods=window_size // 2).mean()
trend_org = regions_org_to_plot.rolling(window=window_size, center=True, min_periods=window_size // 2).mean()
```

CONVENCIONAL (documento AGUACATES ANÁLISIS DE DATA 2- P TV models)



Respecto a la tendencia, observamos que es a la alza, los picos cada vez son más pronunciados y los valles son más elevados.

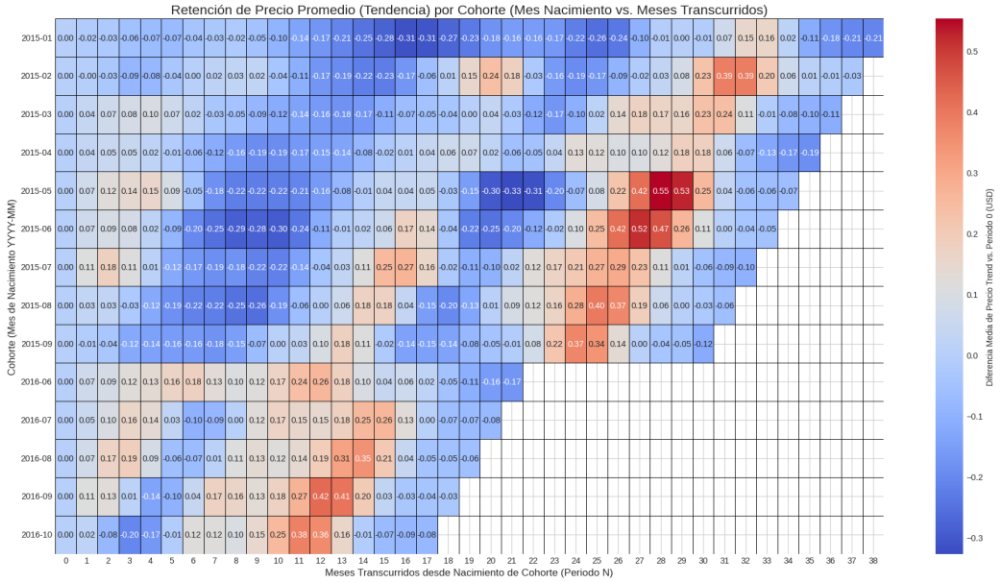
ORGÁNICO (documento AGUACATES ANÁLISIS DE DATA 2- P TV models)



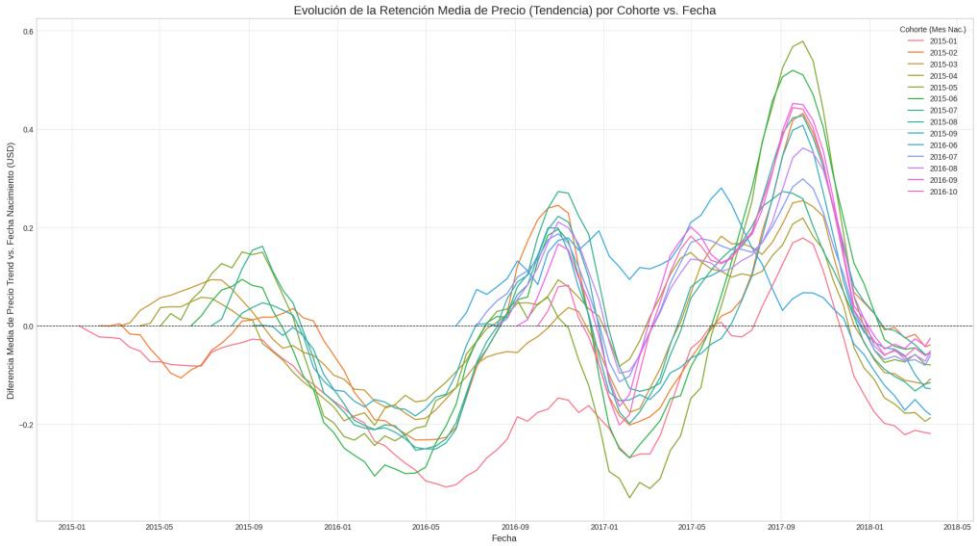
En cuanto a los orgánicos nos fijamos que los precios se mantienen bastante estables. Podemos notar como los picos si que son más altos a medida que avanza el tiempo pero por lo general se vuelve a normalizar. Estas pequeñas diferencias se pueden achacar a la economía de EEUU.

COHORTE

En el siguiente gráfico se plasma la retención del precio de las regiones. Una cohorte consiste en agrupar registros (en este caso las regiones) por grupos en el que todos ellos compartan una cualidad comuna y observar cómo evolucionan.
Esta cohorte agrupa el primer mes que las regiones que hayan superado la media de su precio. A partir de aquí, para los siguientes meses lo que hace es comparar el precio promedio de dicho mes con el primero y así ver si el precio se ha mantenido, subido o bajado.
Cuanto más fuerte sea el azul de la casilla, más ha bajado el precio, y lo mismo para el color rojo.
(documento AGUACATES ANÁLISIS DE DATA 2- P TV models)



En la siguiente gráfica podemos observar de manera más clara la evolución.

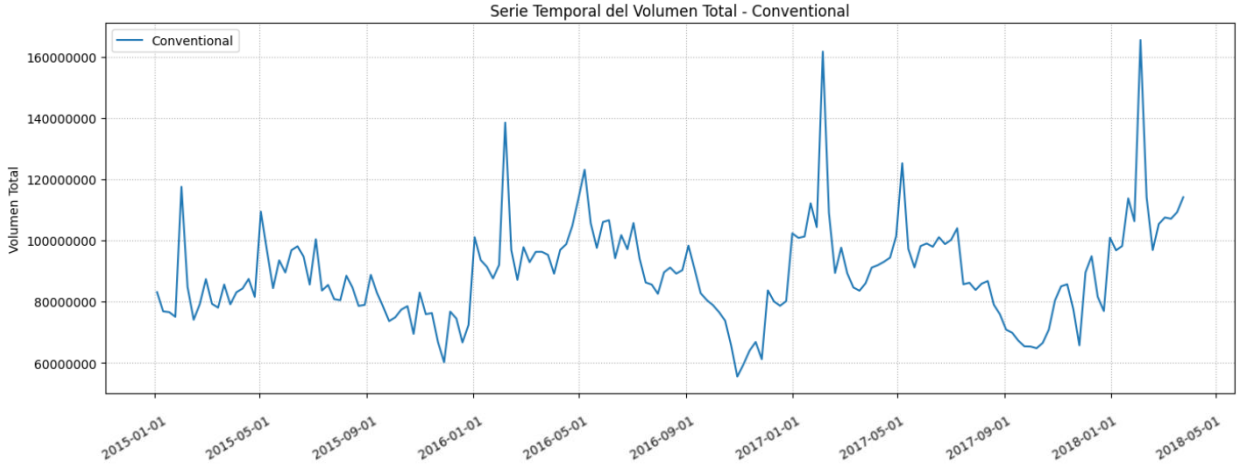


Vemos que en líneas generales, cuando una cohorte supera por primera vez su media el precio tiende a subir y luego caer por debajo. Esto se debe a que la subida de precio está relacionado con la comercialización en esas épocas del año. A medida que se acercan las etapas de mayor consumo, observamos un crecimiento mayor en el precio (son épocas de mucha demanda por lo que el precio sube)y bajan de nuevo y vuelven a subir. Vemos que las cohortes suelen

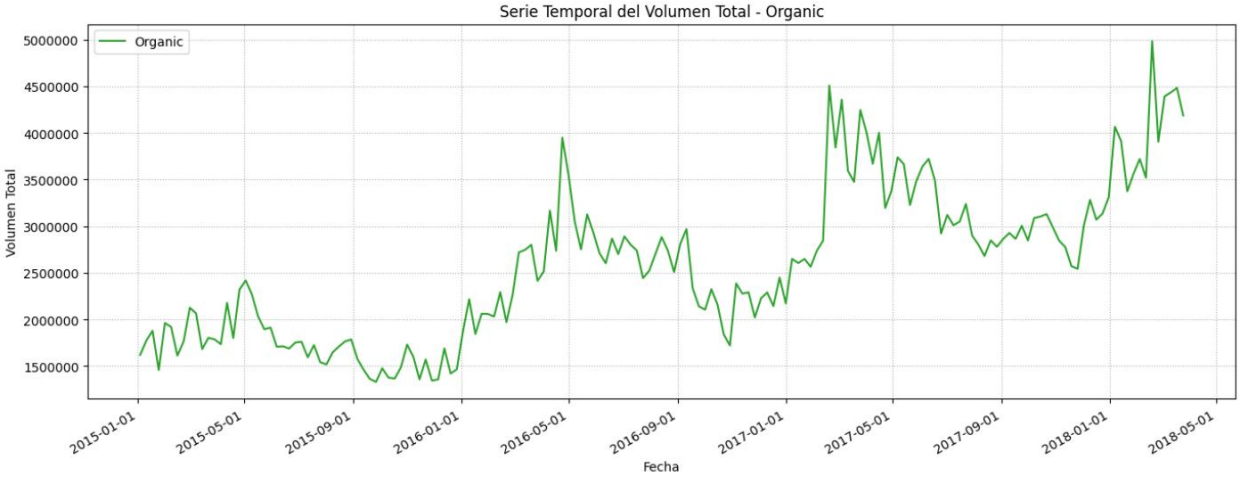
generarse en momentos en los que el consumo era debido a que se acercaba alguna fecha señalada, por lo que su tendencia indica aumento en el momento en el que superó su media por primera vez. Una vez pasado este evento observamos una gran disminución en el precio, pero de nuevo volvemos a tener picos elevados durante la misma etapa.
Si nos fijamos en el grupo 2015-05 podemos ver que su pico es el más alto, seguido de su vecino del mes 06. Del grupo nacido en 2016-06 se observa que tiene una retención extremadamente positiva, pero a finales de 2017 /inicios de 2018 vemos una caída.

ANÁLISIS DE VOLUMEN

En los aguacates de tipo convencional, se observa claramente un comportamiento estacionario, tiene picos muy definidos en ciertos momentos y valles muy definidos también. Aunque el volumen en sí varía mucho, es un patrón constante a lo largo de los años y se mueve entre los mismos máximos y mínimos. (documento AGUACATES ANÁLISIS DE DATA)



Por lo contrario vemos una tendencia muy diferente. Aunque los picos y valles coinciden temporalmente, los valores máximos y mínimos para los aguacates orgánicos han experimentado una variación considerable desde que se tienen registros, lo cual indica un aumento de la popularidad de este producto.



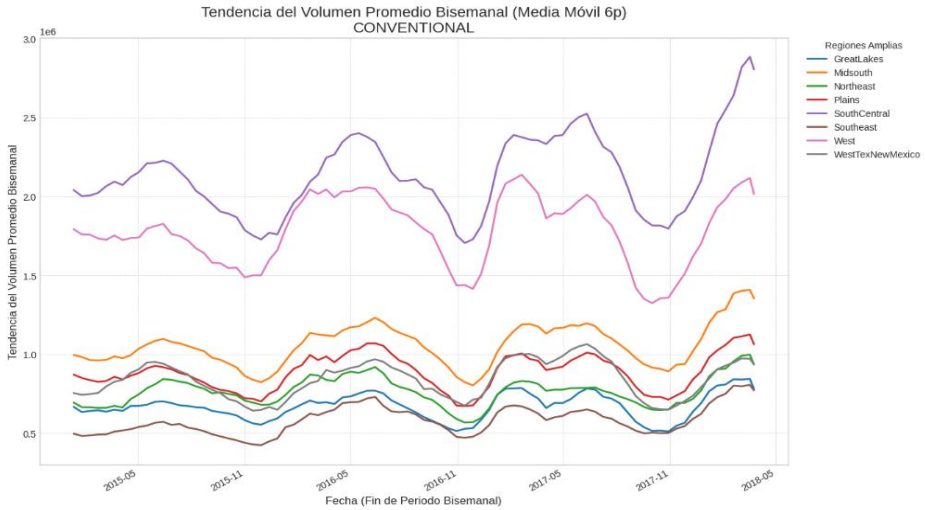
En ambos gráficos observamos una gran coincidencia en el patrón de comportamiento. A principios de año, aproximadamente en Febrero, vemos un pico. Volvemos a observar repuntes a medida que se acerca el mes de Mayo. Indagando la razón de este consumo elevado, me encuentro que en Febrero se celebra la Super Bowl, un evento de rugby de gran popularidad en EEUU, lo que probablemente explique este incremento. En Mayo observamos otro evento y es que es el día de la cultura mexicoamericana, por lo que sería lógico deducir que el consumo es debido a esta celebración, ya que es México es el principal proveedor de aguacates de EEUU.

Para entender mejor cómo están separados los datos y poder hacer el análisis correctamente debemos observar nuestra característica 'Region', cuya estructura requiere un análisis más detallado. Observamos una jerarquía, por lo que he creado una columna llamada clasificación en el que las ciudades que pertenecen a regiones amplias hereden el nombre de la región amplia, por comodidad.

SERIES TEMPORALES – REGIONES AMPLIAS

TIPO CONVENCIONAL

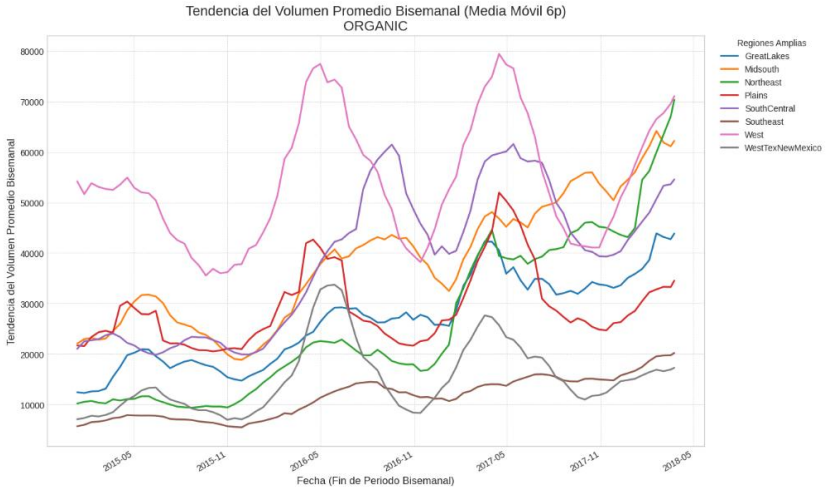
Utilizamos el mismo método de la media móvil de 6 periodos. (documento AGUACATES ANÁLISIS DE DATA 2-Volumen)



En el gráfico anterior se observa claramente cómo hay ciertos eventos anuales que acentúan el consumo de aguacate. Los volúmenes tan dispares entre ciertas zonas es debido a la población que las habitan, ya que en WEST, por ejemplo, hay más densidad de población que en PLAINS. No se observa un mayor consumo de un año a otro, en algunos casos los picos son más acentuados pero generalmente se mantiene estable.

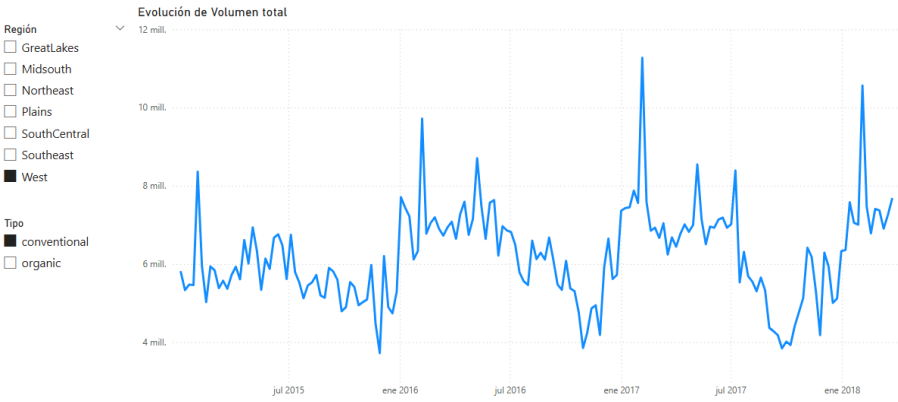
TIPO ORGÁNICO

De igual manera que antes, en el siguiente gráfico se observa la tendencia de volumen, pero nos ofrece una información distinta.



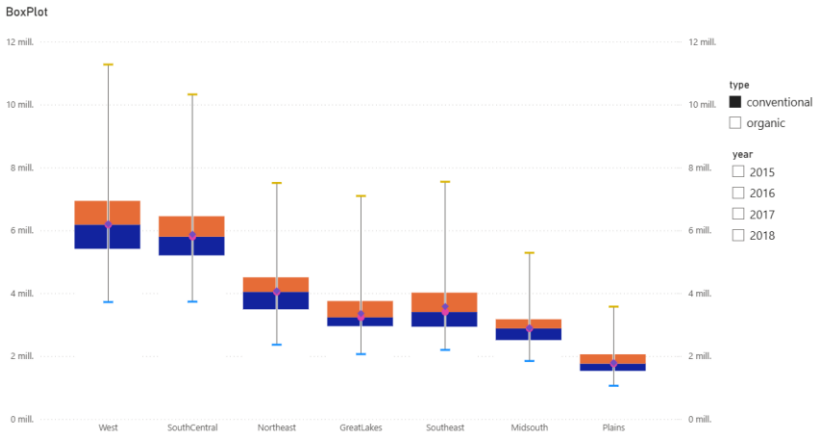
En el caso de los aguacates orgánicos, se observa un comportamiento distinto. Observamos como en la mayoría de las regiones se ha popularizado el aguacate orgánico ya que el volumen total tiene una tendencia alcista. Esto indica que la población se ha interesado más acerca de la calidad de los alimentos y optan por las variedades más saludables. Otra razón de su popularización se debe al aumento de publicaciones en redes sociales promocionando alimentos ecológicos.

En Power BI he generado una gráfica en la que se puede observar de manera más clara cuando se dispara el consumo de los aguacates. Se contempla



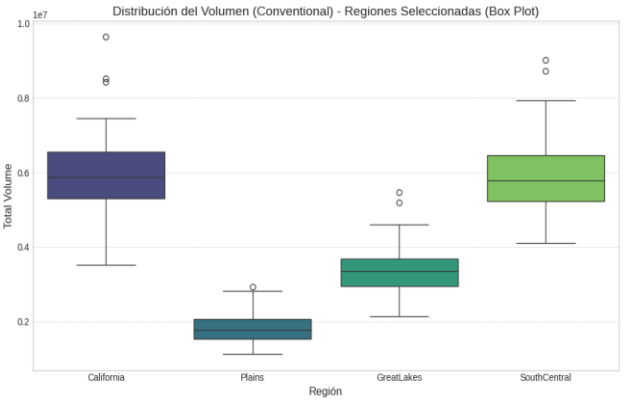
BOXPLOTS Y VIOLINS

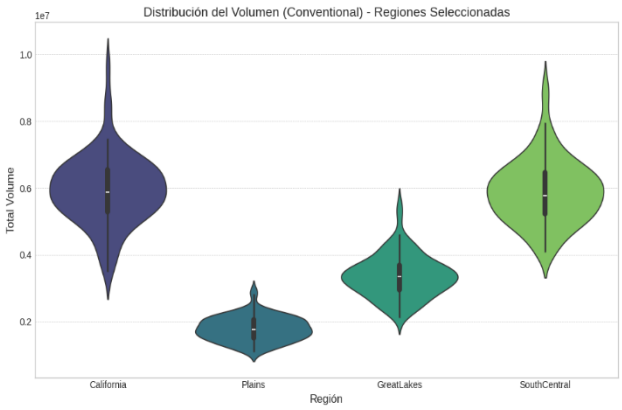
Las gráficas de cajas y de violines nos sirven para ver el rango que ocupan nuestros datos y donde se agrupa la mayor densidad. En Power BI he generado los gráficos de cajas, pero estos no permiten ver los outliers.



Con los siguientes boxplots, con Python, podemos observar mejor la distribución. En la zona rectangular es donde se encuentran la mayoría de los datos y con los bigotes nos indica un rango donde están la mayoría de datos pero en menor medida. Esto quiere decir que a más largos los bigotes, mayor dispersión de los datos. Los puntos blancos nos indica que hay valores que se salen del rango y se consideran outliers.
Outlier es un término utilizado para describir valores atípicos, valores que están muy alejados del resto de datos. Pueden ser tanto positivos como negativos.
Con boxplots localizamos estos valores, y las de violín se utilizan para saber si estos valores son casos aislados o si realmente estos valores son significativos. Cuanto más ancha sea la gráfica de violín, mayor cantidad de registros alberga.

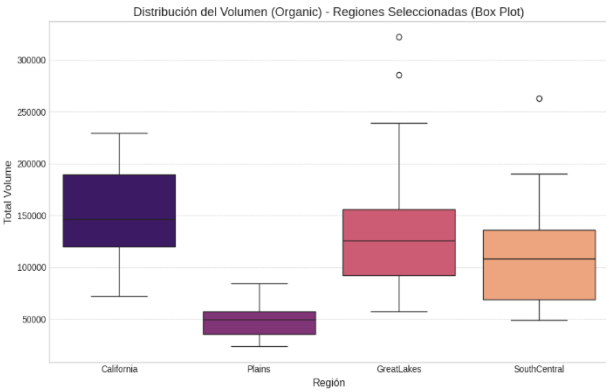
CONVENCIONALES

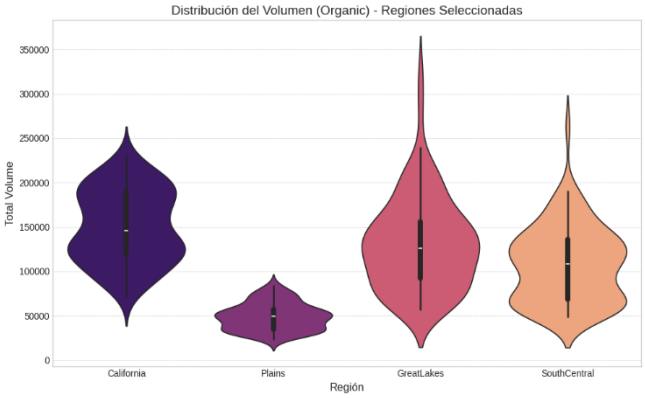




(documento AGUACATES ANÁLISIS DE DATA 2- Volumen)

ORGÁNICOS





(documento AGUACATES ANÁLISIS DE DATA 2- Volumen)

IQR

Para tomar la decisión correcta sobre estos outliers, buscaré la fecha de los registros para ver si coincide con festividades. Los outliers se calculan con IQR, restamos el tercer cuartil (el valor en el que por debajo están el 75% de los datos) al cuartil 1 (25%) Q3-Q1= IQR. El valor máximo es 1.5 * IQR, si supera esta cantidad se considera valor atípico.

```
--- Outliers Identificados (según regla 1.5 * IQR) --
```

	region	type	Date	Total Volume
0	California	conventional	2016-02-07	8434186.065
1	California	conventional	2017-02-05	9634539.180
2	California	conventional	2018-02-04	8514359.175
3	GreatLakes	conventional	2017-02-05	5187170.430
4	GreatLakes	conventional	2018-02-04	5476013.435
5	Plains	conventional	2018-02-04	2925216.160
6	SouthCentral	conventional	2018-02-04	8717007.790
7	SouthCentral	conventional	2018-03-25	9010588.320

```
--- Outliers Identificados (según regla 1.5 * IQR
```

	region	type	Date	Total Volume
0	GreatLakes	organic	2017-03-05	286030.070
1	GreatLakes	organic	2018-03-18	322246.650
2	SouthCentral	organic	2016-08-21	263166.655

En la mayoría de los casos coinciden con fechas señaladas.
En febrero se celebra la Super Bowl, un evento que dispara el consumo de aguacates. Los meses de marzo se consideran efecto de la Super Bowl también.
El comportamiento en el mes de agosto para South Central resulta anómalo, por lo que se procedió a ajustar este valor al máximo permitido según el criterio IQR.

```
Se limitará el Total Volume a: 238019.0
Para: Region='SouthCentral', Tipo='organic', Fecha='2016-08-21'
Número de filas encontradas que coinciden: 1
Valor original de 'Total Volume': 263166.655
'Total Volume' modificado a: 238019.0
¡Modificación realizada con éxito!
```

nprimir pot pantalla el registro de SouthCentral, organic y fecha 2016-08-21

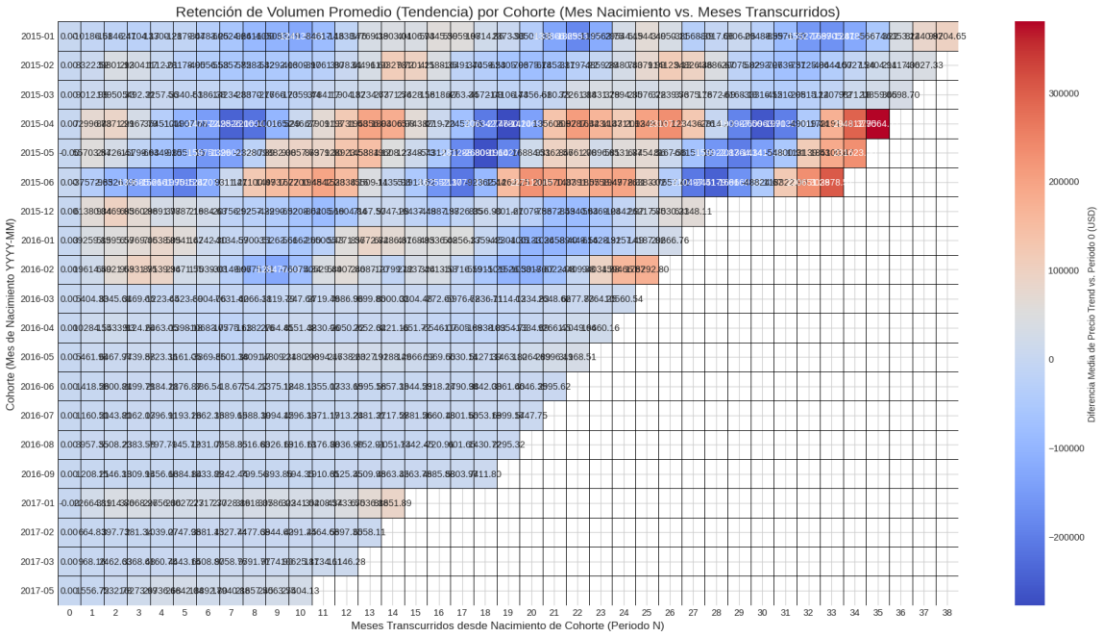
```
print(final_df_manual[(final_df_manual['region'] == 'SouthCentral') & (final_df_manual['type'] == 'organic') & (final_df_manual['Date'] == '2016-08-21')])
```

```
          type       region Clasificación      Date  AveragePrice  \
8457  organic  SouthCentral  SouthCentral 2016-08-21         0.975

      Total Volume      4046     4225  4770  Total Bags  Small Bags  \
8457      238019.0  71631.74  4703.31   0.0  186831.605   179916.78

      Large Bags  XLarge Bags
8457    6914.825          0.0
```

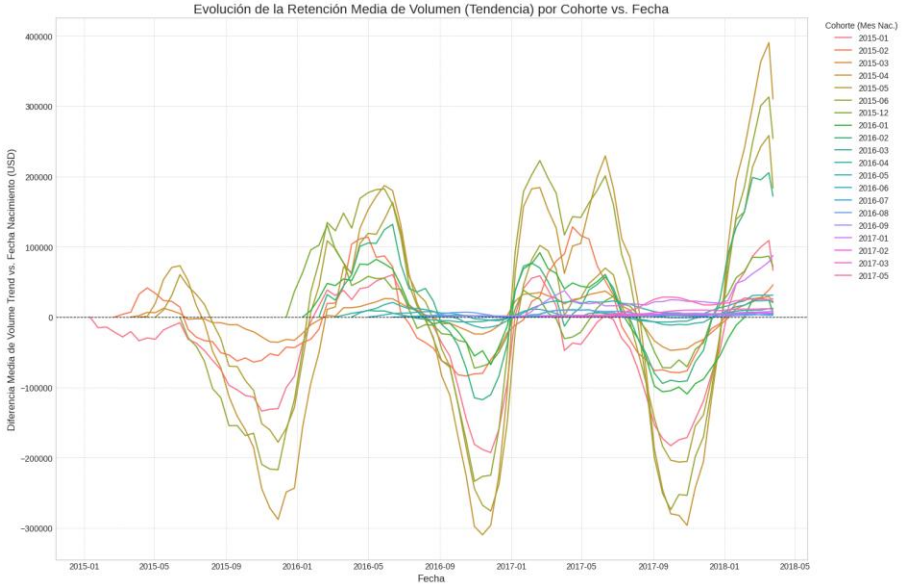
COHORTE

Para estas cohortes he utilizado el mismo criterio del Average Price pero con Total Volume. Observamos como en los grupos de cohortes que nacen antes de 2016 tienen muchas variaciones en comparación a los grupos más recientes, en los que se observa una clara estabilidad

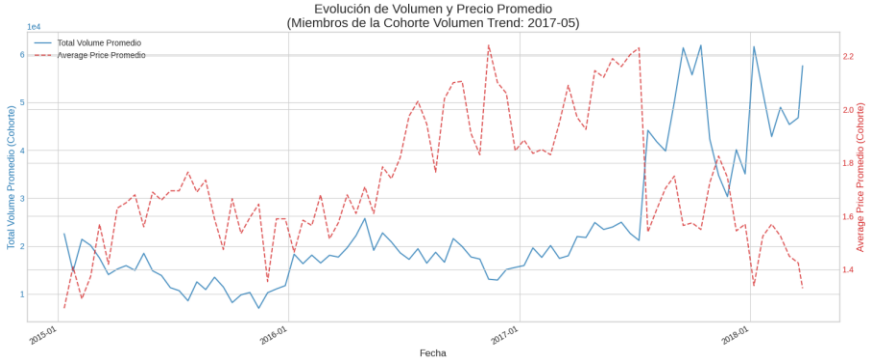
(documento AGUACATES ANÁLISIS DE DATA 2- P TV models)



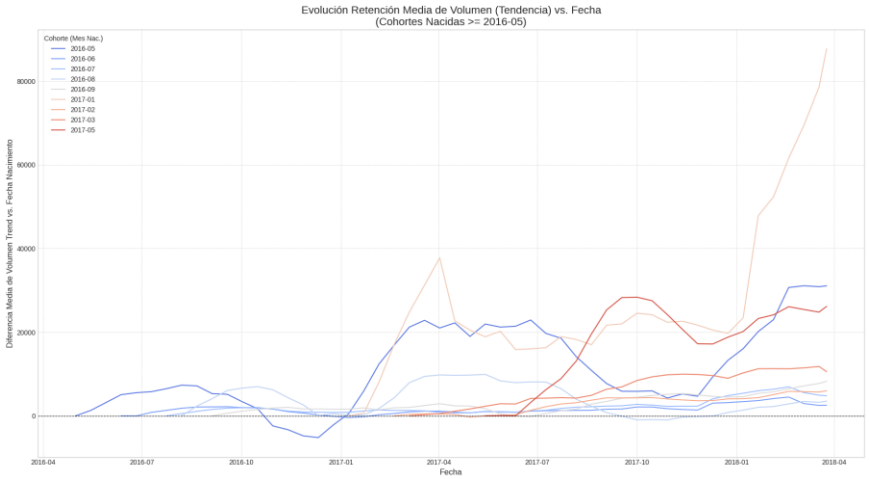
Si miramos el gráfico vemos esto bastante más claro.



La región perteneciente a la cohorte 2017-05, BaltimoreWashington, tiene un volumen bastante constante, si observamos la gráfica Volumen- Precio podemos observar cómo a partir de 2017-05 los valores de volumen sube, baja el precio.

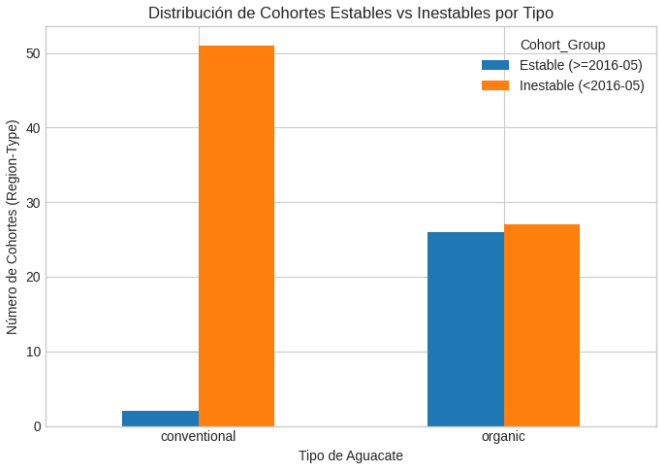


En este otro gráfico he filtrado las cohortes más recientes para observar con más claridad su estabilidad.



Podemos hacer recuento de la cantidad de cohortes que hay en cada grupo y por tipo.

Para comenzar podemos ver que cuando se habla de estabilidad positiva nos encontramos muchos grupos más que son orgánicos que convencionales.



Para verlo más claro, aquí extraigo las regiones y el tipo de los registros en los que las cohortes resultaron más estables. Se observa claramente un mayor asentamiento en el mercado por parte de los aguacates orgánicos, parece que los orgánicos que superan su media más tarde superan un umbral, lo que indica que su mercado ha tardado en madurar.

	region	type	Cohort_Month_Trend_Str	Geo_Group
53	Albany	organic	2016-08	Northeast
54	Atlanta	organic	2016-05	Southeast
55	BaltimoreWashington	organic	2017-05	Southeast
57	Boston	organic	2016-05	Northeast
60	Charlotte	organic	2016-07	Southeast
9	CincinnatiDayton	conventional	2016-05	Midwest
62	CincinnatiDayton	organic	2016-05	Midwest
66	Detroit	organic	2016-05	Midwest
67	GrandRapids	organic	2017-02	Midwest
68	GreatLakes	organic	2016-05	Midwest
69	HarrisburgScranton	organic	2017-03	Northeast
70	HartfordSpringfield	organic	2017-01	Northeast
71	Houston	organic	2016-08	SouthCentral
72	Indianapolis	organic	2017-02	Midwest
76	Louisville	organic	2016-06	Midsouth
78	Midsouth	organic	2016-05	Unknown
26	Nashville	conventional	2016-05	Southeast
81	NewYork	organic	2017-01	Northeast
82	Northeast	organic	2017-01	Unknown
83	NorthernNewEngland	organic	2017-03	Northeast
84	Orlando	organic	2016-05	Southeast
85	Philadelphia	organic	2017-01	Northeast
87	Pittsburgh	organic	2016-09	Northeast
88	Plains	organic	2016-05	Plains
90	RaleighGreensboro	organic	2016-06	Southeast
97	SouthCarolina	organic	2016-05	Southeast
98	SouthCentral	organic	2016-05	Unknown
99	Southeast	organic	2016-05	Unknown

De la cohorte 2017-01, vemos como más allá de estabilizarse, el volumen consumido aumenta significativamente. Observamos que las regiones de aumento son pertenecientes a Northeast, como Philadelphia, Nueva York o HartfordSpringfield. De la siguiente cohorte, 2017-02, se

observa que las regiones que predominan son Indianápolis, y GrandRapids, en Midwest. En 2017-03 volvemos a ver regiones de Northeast, como NorthernNewEngland o HarrisburgScranton. Y por último en 2017-05 observamos BaltimoreWashington, de Southeast.

Para ver entender mejor que está sucediendo en Northeast, podemos ver cuantas de sus regiones están dentro de las cohortes estables:

	region	type	Cohort_Month_Trend_Str	Geo_Group
0	Albany	conventional	2015-06	Northeast
53	Albany	organic	2016-08	Northeast
4	Boston	conventional	2015-06	Northeast
57	Boston	organic	2016-05	Northeast
5	BuffaloRochester	conventional	2015-05	Northeast
58	BuffaloRochester	organic	2016-04	Northeast
16	HarrisburgScranton	conventional	2015-05	Northeast
69	HarrisburgScranton	organic	2017-03	Northeast
17	HartfordSpringfield	conventional	2015-05	Northeast
70	HartfordSpringfield	organic	2017-01	Northeast
28	NewYork	conventional	2015-05	Northeast
81	NewYork	organic	2017-01	Northeast
29	Northeast	conventional	2015-05	Northeast
82	Northeast	organic	2017-01	Northeast
30	NorthernNewEngland	conventional	2015-05	Northeast
83	NorthernNewEngland	organic	2017-03	Northeast
32	Philadelphia	conventional	2015-05	Northeast
85	Philadelphia	organic	2017-01	Northeast
34	Pittsburgh	conventional	2015-06	Northeast
87	Pittsburgh	organic	2016-09	Northeast
49	Syracuse	conventional	2015-06	Northeast
102	Syracuse	organic	2016-04	Northeast

Observamos que todas las regiones con tipo orgánico están dentro de las cohortes estables, lo que deducimos que es un mercado que se ha consolidado muy bien en esta zona. Los convencionales destacan por lo pronto que superan su media y por lo variable que es su consumo a lo largo del año.

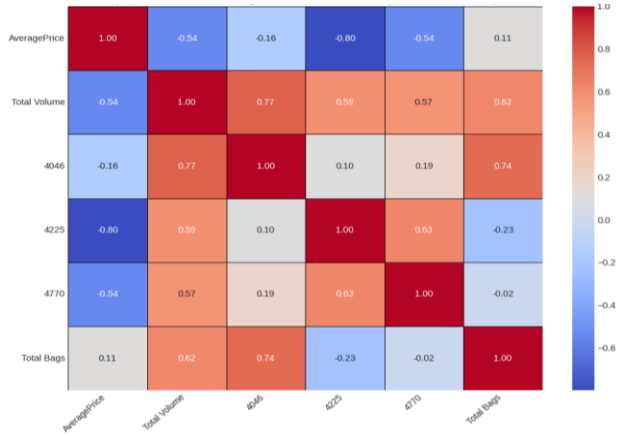
Con las regiones de MidSouth y Southeast tenemos casos similares, aunque regiones como Roanoke o RichmondNorfolk son excepciones.

	region	type	Geo_Group	Cohort_Month_Trend_Str
23	Louisville	conventional	Midsouth	2015-01
76	Louisville	organic	Midsouth	2016-06
25	Midsouth	conventional	Midsouth	2015-05
78	Midsouth	organic	Midsouth	2016-05
1	Atlanta	conventional	Southeast	2016-04
54	Atlanta	organic	Southeast	2016-05
2	BaltimoreWashington	conventional	Southeast	2015-01
55	BaltimoreWashington	organic	Southeast	2017-05
7	Charlotte	conventional	Southeast	2016-02
60	Charlotte	organic	Southeast	2016-07
20	Jacksonville	conventional	Southeast	2016-02
73	Jacksonville	organic	Southeast	2016-04
24	MiamiFtLauderdale	conventional	Southeast	2016-02
77	MiamiFtLauderdale	organic	Southeast	2016-02
26	Nashville	conventional	Southeast	2016-05
79	Nashville	organic	Southeast	2016-03
27	NewOrleansMobile	conventional	Southeast	2015-01
80	NewOrleansMobile	organic	Southeast	2016-04
31	Orlando	conventional	Southeast	2016-02
84	Orlando	organic	Southeast	2016-05
37	RaleighGreensboro	conventional	Southeast	2016-02
90	RaleighGreensboro	organic	Southeast	2016-06
38	RichmondNorfolk	conventional	Southeast	2015-05
91	RichmondNorfolk	organic	Southeast	2015-05
39	Roanoke	conventional	Southeast	2016-02
92	Roanoke	organic	Southeast	2015-04
44	SouthCarolina	conventional	Southeast	2015-05
97	SouthCarolina	organic	Southeast	2016-05
46	Southeast	conventional	Southeast	2016-02
99	Southeast	organic	Southeast	2016-05
50	Tampa	conventional	Southeast	2016-01
103	Tampa	organic	Southeast	2016-04

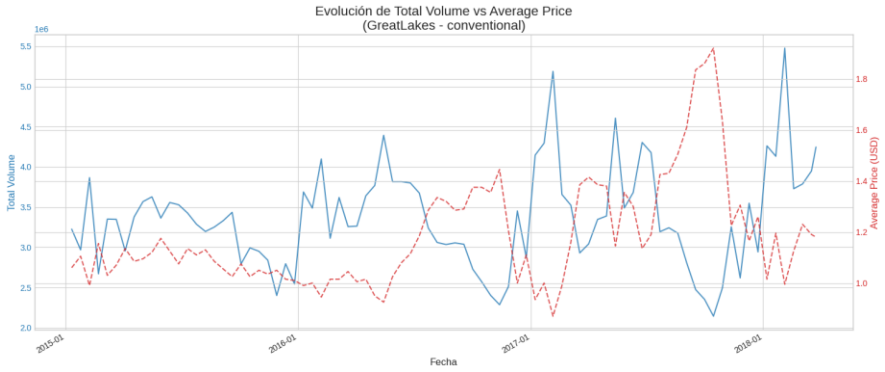
PRECIO VS VOLUMEN

Prestando atención a las gráficas generadas tanto de volumen como de precio, se observa un comportamiento similar pero inverso.

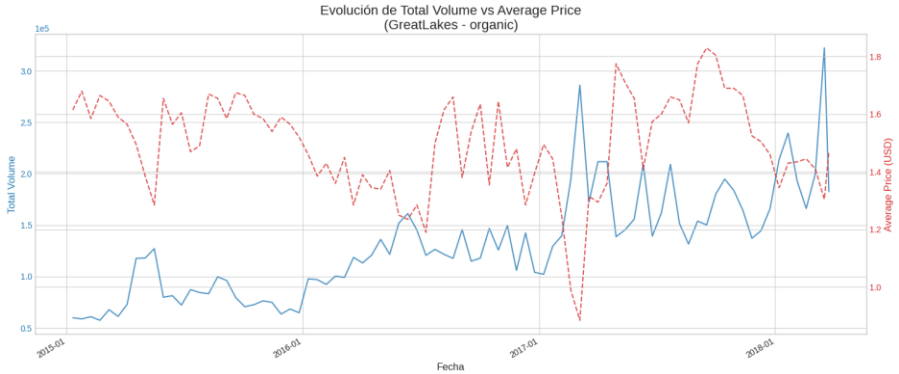
Primero de todo observaremos la matriz de correlación. Podemos observar la correlación entre las características del dataframe. Se observa una correlación negativa entre Total Volume y Average Price. (documento AGUACATES ANÁLISIS DE DATA 2- P TV models)



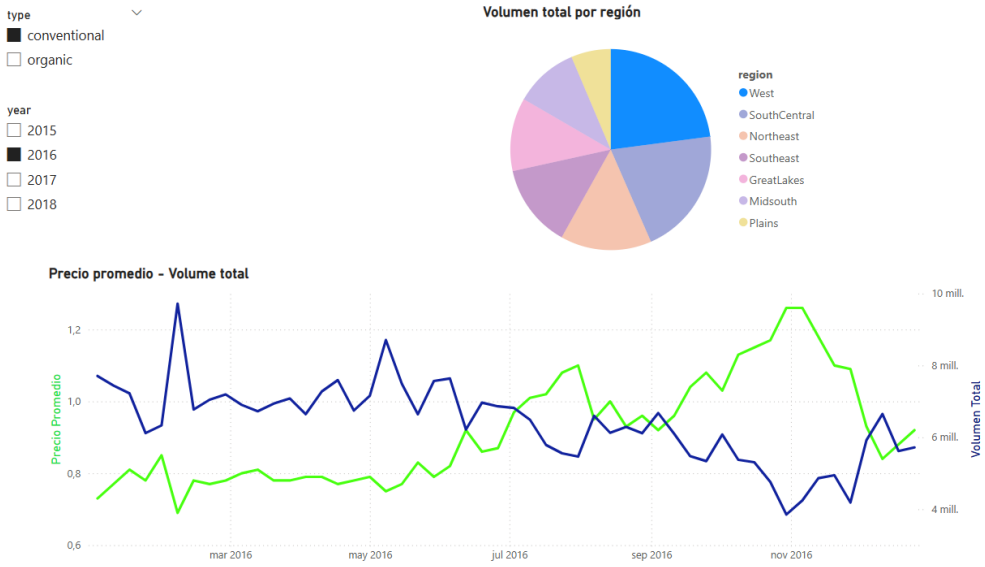
Para demostrar esta correlación negativa, lo podemos graficar para ver mejor su evolución, por ejemplo para GreatLakes de tipo convencional.



Observamos claramente cuándo Total Volume sube, AveragePrice baja.
Lo mismo ocurre para los aguacates orgánicos, pero vemos como el volumen total poco a poco tiene bajadas menos pronunciadas. Esto indica un asentamiento en el mercado de este producto.



He generado una visualización en Power BI que nos permite ver esta correlación, por años y regiones amplias para verlo más claro.



SELECCIÓN DE REGIONES

Para realizar análisis más específicos y modelos más precisos, es importante seleccionar regiones que sean representativas y que tengan comportamientos interesantes, intentando cubrir zonas diversas de EEUU.

Northeast: Philadelphia
Pertenece a un grupo de cohorte que tiene un comportamiento interesante, su total volume aumenta drásticamente después de superar su media (tipo orgánico)

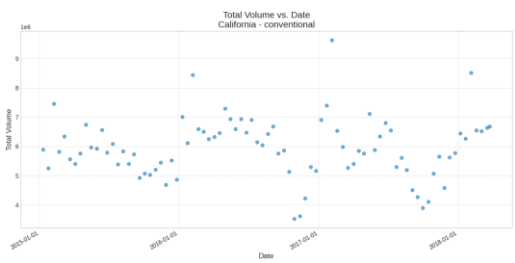
West: California
Es un mercado enorme, a menudo un referente, y sus dinámicas pueden ser complejas y distintas a otras zonas. Es bueno tenerlo para representar la costa Oeste y un alto volumen.

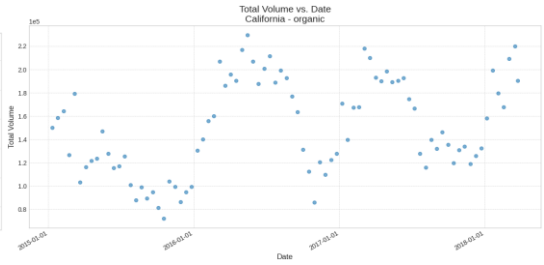
Southeast: Roanoke
Es de los pocos orgánicos que supera su media en cohortes tempranas.

Midsouth: Nashville
Buena elección para representar la zona Midsouth. Además presenta un comportamiento interesante y es que su tipo convencional está dentro de las cohortes más recientes, es decir las que mantienen o aumentan su volumen.

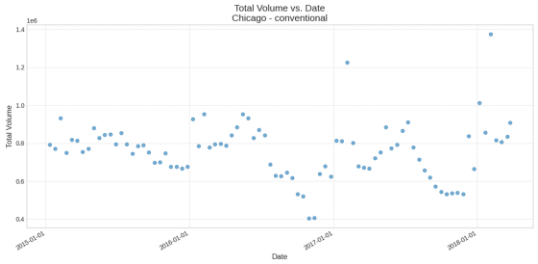
Midwest: Chicago
Excelente para tener un gran mercado del Medio Oeste, con su propia dinámica.

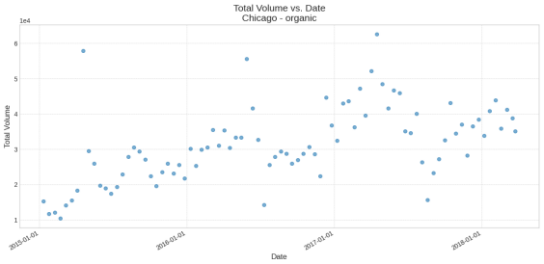
VOLUMEN - CALIFORNIA



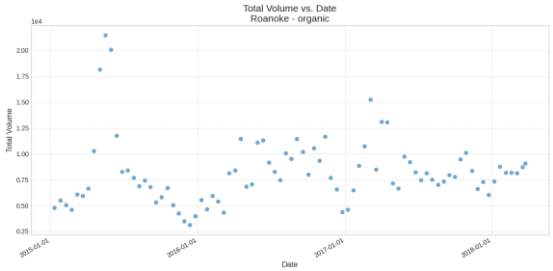


VOLUMEN - CHICAGO

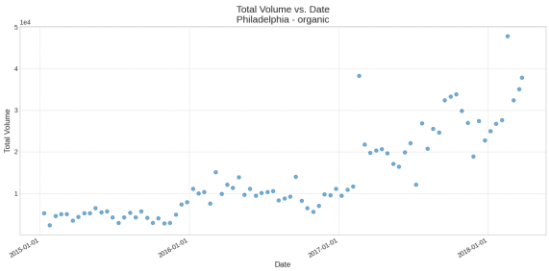




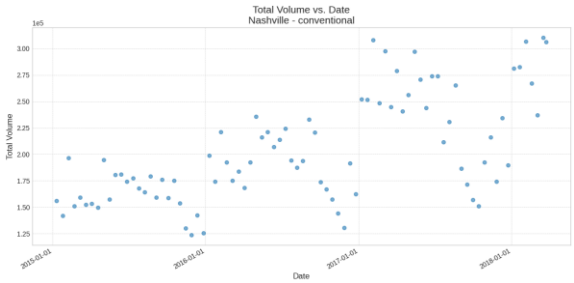
ROANOKE



PHILADELPHIA



NASHVILLE



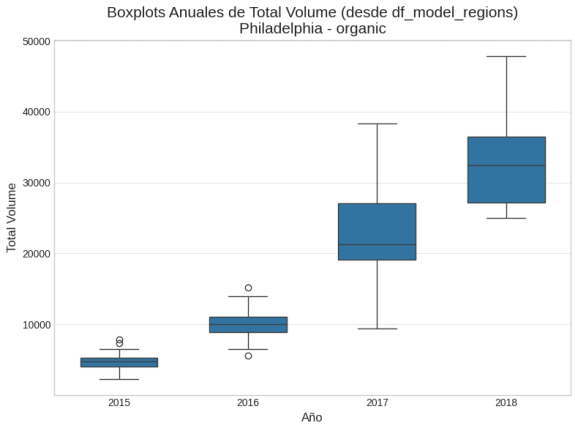
Observamos bastantes outliers. En Roanoke en 2015 vimos un repunte muy alto sobre los productos orgánicos. Es altamente probable que este fenómeno se deba a la celebración del 40 aniversario de la fundación de Roanoke Natural Foods Co-op, realizando campañas de márketing para popularizar este tipo de productos. Al tratarse de una situación especial, también se considerarán como outliers y los toparemos según IQR.

	region	type	Date	Total Volume	Q1_Volume	Q3_Volume	IQR_Volume	Lower_Bound_Volume	Upper_Bound_Volume
12	California	conventional	2016-02-07	8,434,186.06	5,296,836.72	6,548,788.96	1,251,952.24	3,418,908.36	8,426,717.32
13	California	conventional	2017-02-05	9,634,539.18	5,296,836.72	6,548,788.96	1,251,952.24	3,418,908.36	8,426,717.32
14	California	conventional	2018-02-04	8,514,359.18	5,296,836.72	6,548,788.96	1,251,952.24	3,418,908.36	8,426,717.32
5	Chicago	conventional	2016-10-30	403,243.89	672,131.36	838,092.14	165,960.77	423,190.21	1,087,033.29
6	Chicago	conventional	2016-11-13	406,593.81	672,131.36	838,092.14	165,960.77	423,190.21	1,087,033.29
7	Chicago	conventional	2017-02-05	1,225,876.68	672,131.36	838,092.14	165,960.77	423,190.21	1,087,033.29
8	Chicago	conventional	2018-02-04	1,375,307.41	672,131.36	838,092.14	165,960.77	423,190.21	1,087,033.29
9	Chicago	organic	2015-04-19	57,917.97	25,289.22	37,016.40	11,727.18	7,698.44	54,607.18
10	Chicago	organic	2016-05-29	55,546.96	25,289.22	37,016.40	11,727.18	7,698.44	54,607.18
11	Chicago	organic	2017-04-16	62,609.75	25,289.22	37,016.40	11,727.18	7,698.44	54,607.18
0	Philadelphia	organic	2018-02-18	47,883.94	5,556.05	20,632.78	15,076.74	-17,059.07	43,247.90
1	Roanoke	organic	2015-05-03	18,143.94	6,462.43	9,356.05	2,893.61	2,122.01	13,696.47
2	Roanoke	organic	2015-05-17	21,481.15	6,462.43	9,356.05	2,893.61	2,122.01	13,696.47
3	Roanoke	organic	2015-05-31	20,075.69	6,462.43	9,356.05	2,893.61	2,122.01	13,696.47
4	Roanoke	organic	2017-03-05	15,255.94	6,462.43	9,356.05	2,893.61	2,122.01	13,696.47

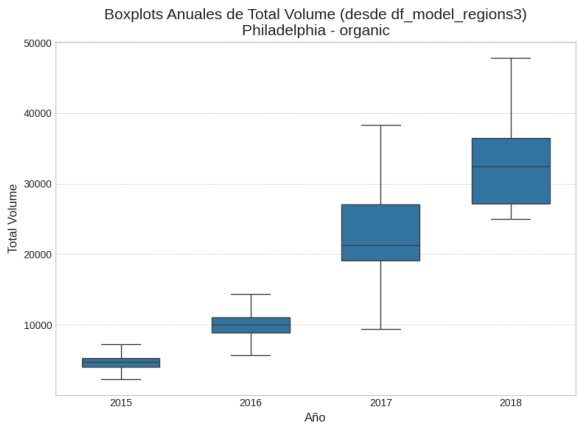
Aunque en muchas ocasiones estos valores atípicos coinciden con eventos especiales, su presencia no resulta beneficiosa para la generación del modelo. Con la misma técnica de IQR, cambiaremos estos valores para toparlos al Q3 + 1.5 *IQR.

Una vez topados, podemos observar con boxplot para ver si se ha realizado correctamente.

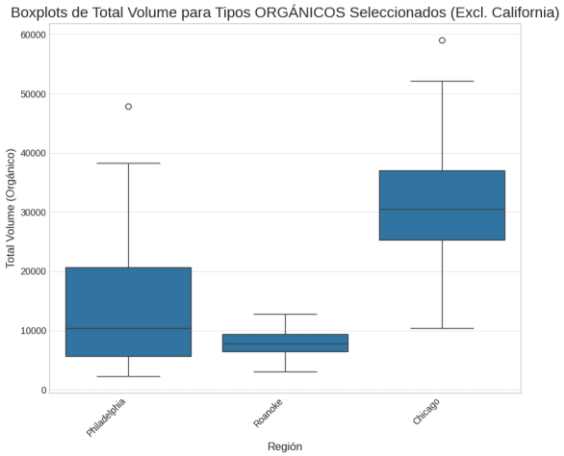
DATAFRAME SIN TOPAR



DATAFRAME MODIFICADO



Estos outliers se han calculado de manera anualmente puesto que hacerlo de manera total la variación de los otros años puede influir mucho en los datos.
Pongo de ejemplo únicamente esta región, pero se ha hecho lo mismo para todas las combinaciones. Al haber hecho los outliers por años, no podemos agruparlos todos en un único boxplot general, ya que datos de 2018 pueden ser outliers para los de 2015, entonces aparecen outliers sin serlo en realidad. Ejemplo:



Vemos que en Philadelphia si aparecen outliers si agrupamos los años, aunque no sea el caso si hacemos una separación distinta.

INGENIERÍA DE CARACTERÍSTICAS

Para poder entrenar de manera más eficiente los modelos, primeramente he decidido agregar algunas características que puedan facilitarles identificar patrones.

Date	type	region	Clasificación	AveragePrice	Total Volume	Geo_Group	year	WeekOfYear	Lag_26_Total_Volume	Lag_1_AveragePrice	Near_SuperBowl	Near_CincoDeMayo	Near_July4th	Near_Thanksgiving	Near_ChristmasNewYear
2015-01-11	organic	Philadelphia	Northeast	1.675	5251.89	Northeast	2015	2	NaN	NaN	0	0	0	0	0
2015-01-25	organic	Philadelphia	Northeast	1.875	2293.81	Northeast	2015	4	NaN	1.675	1	0	0	0	0
2015-02-08	organic	Philadelphia	Northeast	1.725	4540.08	Northeast	2015	6	NaN	1.875	1	0	0	0	0
2015-02-22	organic	Philadelphia	Northeast	1.715	5014.26	Northeast	2015	8	NaN	1.725	0	0	0	0	0
2015-03-08	organic	Philadelphia	Northeast	1.730	4978.06	Northeast	2015	10	NaN	1.715	0	0	0	0	0

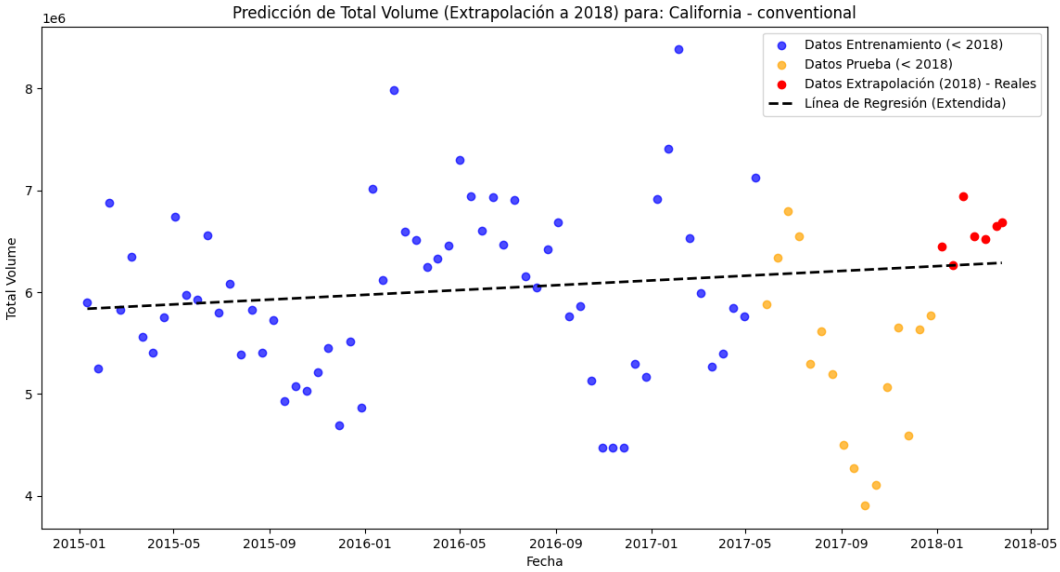
Observamos Lag 26 de Total Volume que nos indica la cantidad de volumen de hace 26 periodos. Lag 1 de Price nos indica el precio del periodo anterior. En los Near, se indica si se acerca esta festividad, es 1 mientras 4 periodos antes de la fecha y en los 2 siguientes.

Se aplican estos cambios al data frame original para que todos los datos sean uniformes y contengan la misma información. Con estas variables será suficiente para ayudar al modelo ya que al estar buscando un origen estacional es importante marcar fechas señaladas.

MODELOS PREDICTIVOS

REGRESIÓN LINEAL

Generalmente utilizado con datos con tendencias claras y poca variabilidad en datos. Este modelo presenta una utilidad limitada debido a la naturaleza del conjunto de datos.
Estos modelos resultan poco efectivos para estos casos, dada la variabilidad en el comportamiento de los datos. (documento Creación de modelos)



Métricas para el conjunto de prueba de interpolación (datos < 2018):
RMSE: 1233797.30
R^2: -1.13
Métricas para el conjunto de extrapolación (datos de 2018):
RMSE: 360318.33
R^2: -2.43

Como se puede observar la predicción que hace para el conjunto de datos es una línea recta, casi como si delimitara la media. Se observa que en caso de california tiene una tendencia ligeramente alcista, lo que implica un pequeño crecimiento en el mercado.
El modelo resulta ineficaz para la interpolación y, consecuentemente, también para la extrapolación..

REGRESIÓN POLINÓMICA

La regresión polinómica es una técnica de aprendizaje supervisado utilizada en inteligencia artificial. Modela la relación entre una variable independiente y una dependiente como un polinomio de grado 'n'. A diferencia de la regresión lineal simple, puede capturar relaciones no lineales en los datos. El objetivo es encontrar la curva polinómica que mejor se ajusta a los puntos de datos observados. Se usa para predicciones cuando la relación entre variables no es una línea recta. Aumentar el grado del polinomio puede mejorar el ajuste, pero también puede llevar al sobreajuste (overfitting).
Para la regresión polinómica es necesario realizar una búsqueda de hiperparámetros y la búsqueda de hiper parámetros he utilizado GridSearch. Se observa como el mejor resultado me lo proporciona un grado 1, lo que es una regresión lineal. La valoración del modelo se basa en el error cuadrático medio, que es el cálculo del error en las predicciones.
Resultados:

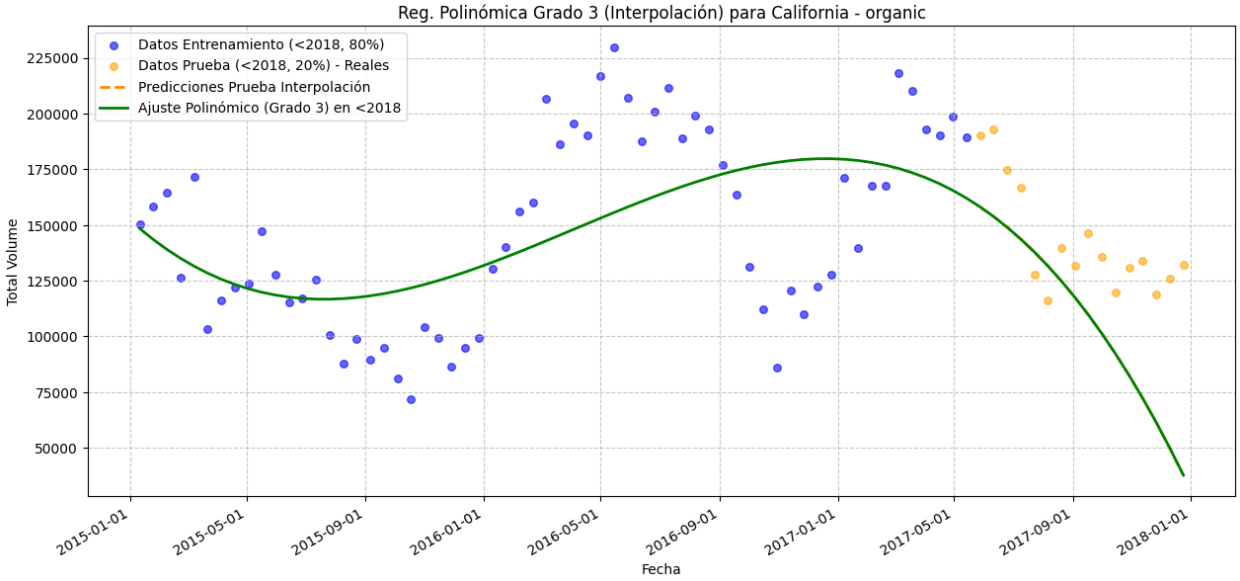
```
Mejor grado polinómico encontrado: 1

Métricas en el CONJUNTO DE PRUEBA DE INTERPOLACIÓN (<2018, 20%):
  RMSE: 55819.19
  R^2: -4.42

Métricas en el CONJUNTO DE EXTRAPOLACIÓN (2018):
  RMSE: 24165.40
  R^2: -0.39
```

En estos modelos es posible aplicar dos configuraciones.
Lasso: Añade una penalización a la función de pérdida igual a la suma de los valores absolutos de los coeficientes, multiplicada por alfa. Esto también encoge los coeficientes y puede reducirlos exactamente a cero, realizando así una selección de características.
Ridge: Añade una penalización a la función de pérdida igual a la suma de los cuadrados de los coeficientes, multiplicada por alfa. Esto encoge los coeficientes, reduciendo la complejidad del modelo y el sobreajuste, pero raramente los hace exactamente cero.
He aplicado Ridge, de nuevo utilizando GridSearch para buscar los mejores parámetros basado en las métricas RMSE y obtengo el siguiente resultado.

```
Métricas en el CONJUNTO DE PRUEBA DE INTERPOLACIÓN (<2018, 20%):
  Grado Polinómico: 5, Alpha Lasso: 0.1
  RMSE: 55855.78
  R^2: -4.43

Métricas en el CONJUNTO DE EXTRAPOLACIÓN (2018):
  Grado Polinómico: 5, Alpha Lasso: 0.1
  RMSE: 24202.85
  R^2: -0.39
```

Persiste un rendimiento deficiente del modelo." o "Se sigue observando un comportamiento subóptimo.

De todos modos, podemos aplicar un polinomio de grado 3 para observar el comportamiento:

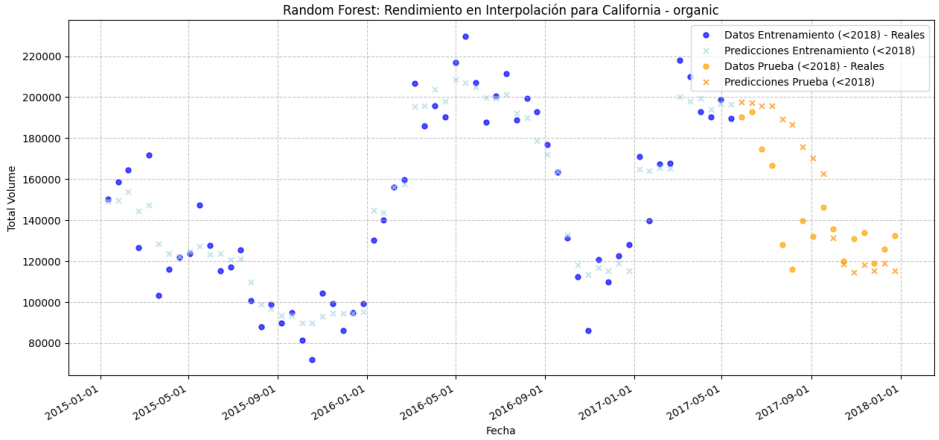


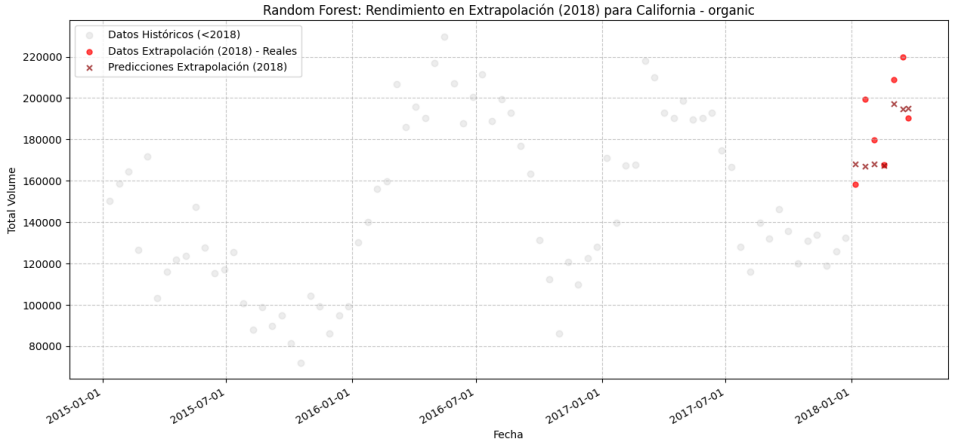
Obtenemos como resultados

```
Métricas en CONJUNTO DE PRUEBA DE INTERPOLACIÓN (<2018, 20%):
  RMSE: 44741.20
  R^2: -2.48

Métricas en CONJUNTO DE EXTRAPOLACIÓN (2018):
  RMSE: 211885.59
  R^2: -105.54
```

RANDOM FOREST

Son modelos utilizados mayoritariamente para clasificación, por lo que no suele tener en cuenta tendencias. Este modelo trabaja creando varios árboles de decisión y estos árboles hacen cada una sus predicciones sobre el target. Una vez hecho esto, la predicción que más haya salido es la que Random Forest toma como válida.






```
Métricas de ajuste en CONJUNTO DE ENTRENAMIENTO DE INTERPOLACIÓN (<2018, 80%):
  RMSE (ajuste entrenamiento): 10683.64
  R^2 (ajuste entrenamiento): 0.94

Métricas en CONJUNTO DE PRUEBA DE INTERPOLACIÓN (<2018, 20%):
  RMSE: 29653.37
  R^2: -0.53

Métricas en CONJUNTO DE EXTRAPOLACIÓN (2018):
  RMSE: 17173.70
  R^2: 0.30
```

En el conjunto de datos de entrenamiento vemos que tiene un sobreajuste del 94%, pero en predicciones a futuro no se comporta muy bien.

Realizando GridSearch para reducir el sobreajuste con algunos parámetros esenciales como Max_depth (profundidad de cada árbol), min_samples_leaf (mínimo de muestras requeridas para separar un nodo) y 'min_samples_split' (mínimo de muestras requeridas en un nodo hoja.)
He aplicado un rango de 5 datos para cada parámetro y el que mejor encuentra es el siguiente.

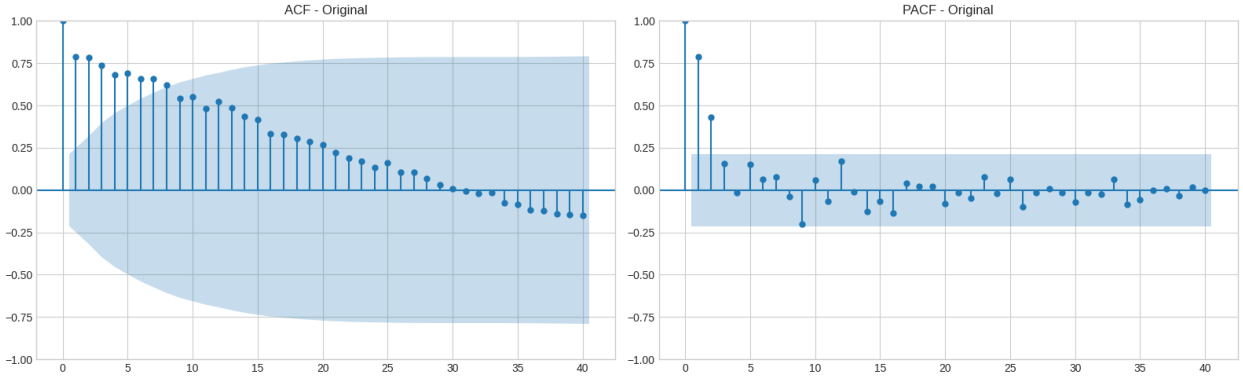
```
Iniciando GridSearchCV para Random Forest (con 3-fold TimeSeriesSplit)...
Fitting 3 folds for each of 36 candidates, totalling 108 fits
Mejores parámetros encontrados:
{'max_depth': 3, 'min_samples_leaf': 5, 'min_samples_split': 40}

Métricas de ajuste del MEJOR MODELO en CONJUNTO DE ENTRENAMIENTO DE INTERPOLACIÓN:
  RMSE (ajuste entrenamiento): 31491.47
  R^2 (ajuste entrenamiento): 0.46

Métricas del MEJOR MODELO en CONJUNTO DE PRUEBA DE INTERPOLACIÓN (<2018, 20%):
  RMSE: 22490.08
  R^2: 0.12

Métricas del MEJOR MODELO en CONJUNTO DE EXTRAPOLACIÓN (2018):
  RMSE: 31565.13
  R^2: -1.36
```

Se evidencia una reducción del sobreajuste y una ligera mejora en el rendimiento con los datos de prueba; no obstante, el modelo continúa siendo marcadamente deficiente para la predicción.

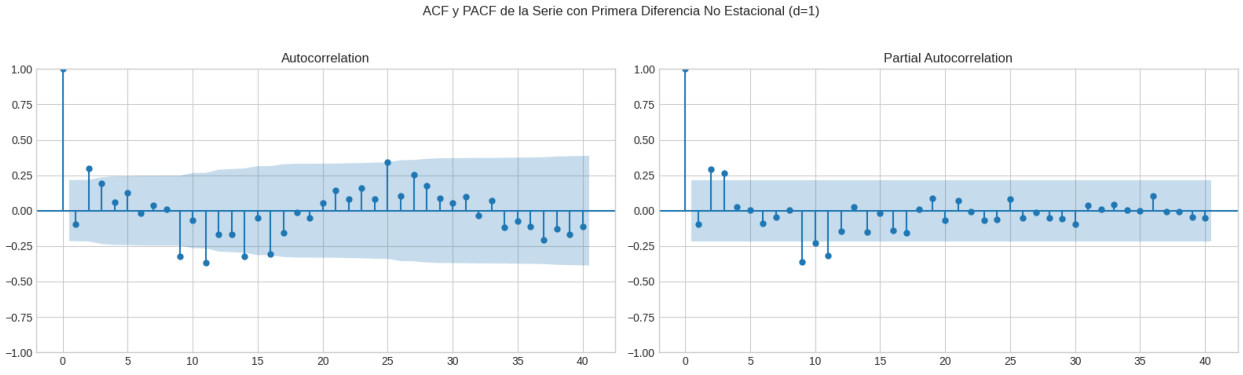
SARIMA

Para la selección mencionada anteriormente utilizaré el modelo SARIMA, una extensión del modelo ARIMA que tiene en cuenta la estacionalidad.

Para estos modelos estacionales es importante hacer la separación de datos de manera cronológica, ya que trabajan mejor de esta manera.
SARIMA necesita algunos parámetros de configuración para desempeñar el modelo. Para calcular dichos valores debemos observar las gráficas de Autocorrelación y la autocorrelación parcial.
Autocorrelación se utiliza para ver el impacto del dato anterior sobre el actual. No se observa ninguna estacionalidad, vemos que decrece lentamente lo que no indica estacionalidad.
Estas gráficas se deben hacer para cada region-type ya que cada uno de ellos se comporta de manera diferente. En este caso he utilizado Philadelphia- Organic para llevar a cabo el estudio.



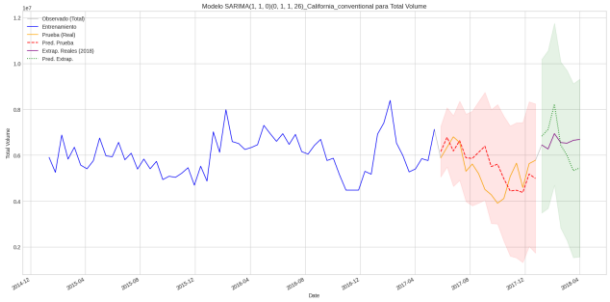
Aplicando una diferencia no estacionaria “d=1” si que podemos observar una estacionalidad, lo que esto hace es, en vez de tener en cuenta el dato, tiene en cuenta la diferencia del dato actual con el dato anterior. En este caso sí que se observa una estacionalidad clara.



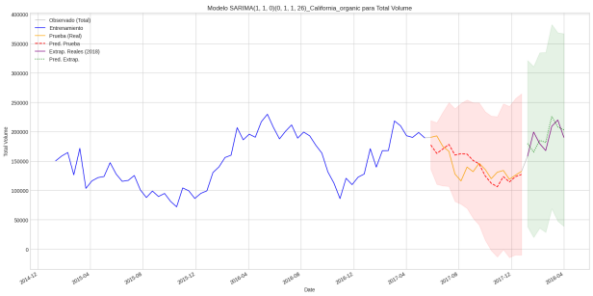
Aplicaré los parámetros encontrados sobre las combinaciones mencionadas para ver el comportamiento de SARIMA.

CALIFORNIA

Observamos los datos de entrenamiento de color azul, los de prueba en naranja y las líneas discontinuas son las predicciones del modelo.

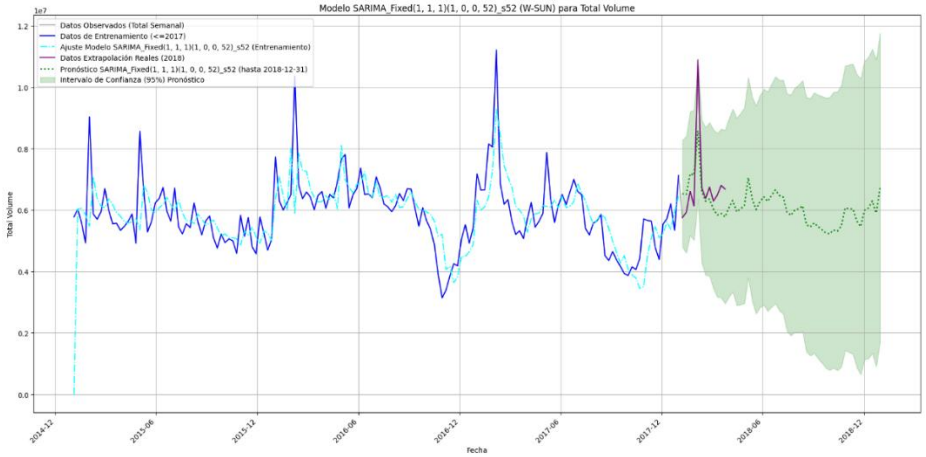


Combinación: California – conventional
RMSE Entrenamiento: 1106253.75, R² Entrenamiento: -0.79
RMSE Prueba: 1062416.01, R² Prueba: -0.58
RMSE Extrapolación: 992388.72, R² Extrapolación: -25.04



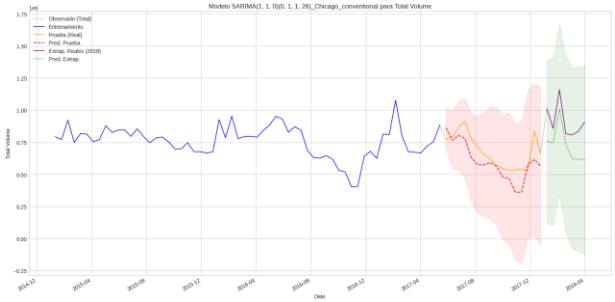
Combinación: California - organic
RMSE Entrenamiento: 32235.92, R² Entrenamiento: 0.43
RMSE Prueba: 27085.06, R² Prueba: -0.28
RMSE Extrapolación: 26090.66, R² Extrapolación: -0.62

Como no actua muy correctamente, haremos el análisis de parámetros para california-conventional e intentar predecir de manera más correcta su comportamiento.
Utilizaremos KFOLD para california convencional y seleccionará los mejores parámetros basado en R^2. He extendido las predicciones hasta final de año de 2018. Este modelo lo he generado utilizando los datos semanalmente, por lo que en vez de “26” utilizaremos “52” para definir de manera correcta los datos

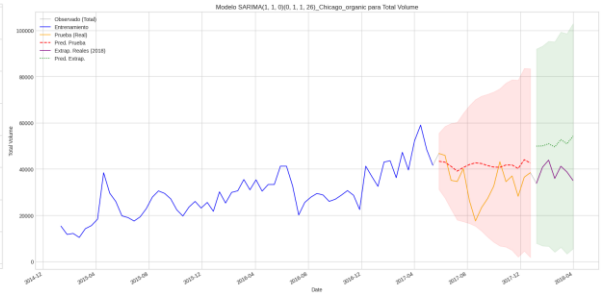


Combinación: California – conventional
RMSE Entrenamiento: 1237536.98, R² Entrenamiento: 0.46
RMSE Prueba: 1398465.99, R² Prueba: 0.54

CHICAGO

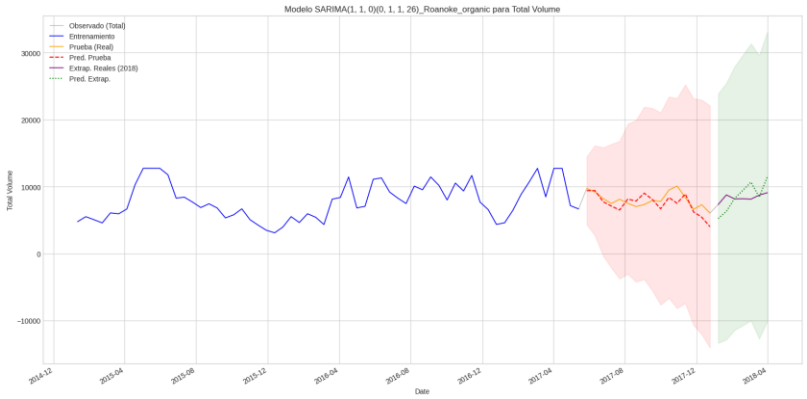


Combinación: Chicago - conventional
RMSE Entrenamiento: 148008.94, R² Entrenamiento: -0.47
RMSE Prueba: 83917.39, R² Prueba: 0.57
RMSE Extrapolación: 269014.83, R² Extrapolación: -4.02



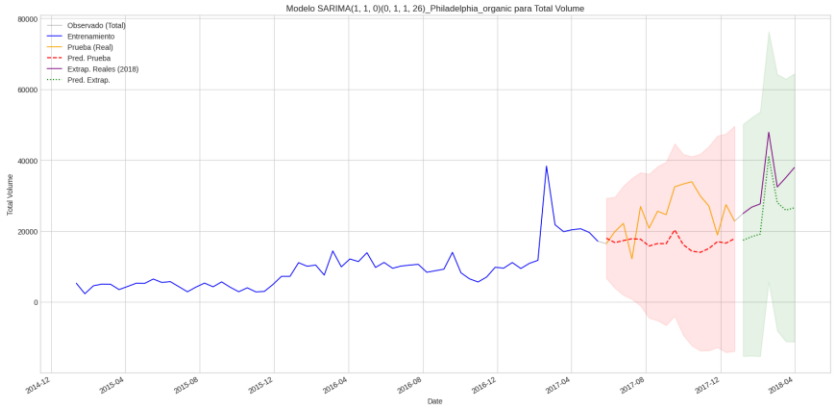
Combinación: Chicago - organic
RMSE Entrenamiento: 6402.94, R² Entrenamiento: 0.59
RMSE Prueba: 11380.49, R² Prueba: -1.10
RMSE Extrapolación: 12173.28, R² Extrapolación: -11.67

ROANOKE – organic



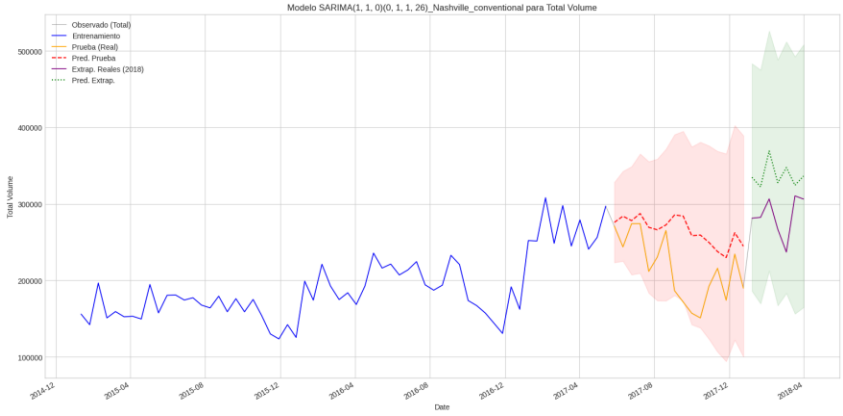
Combinación: Roanoke - organic
RMSE Entrenamiento: 2243.42, R² Entrenamiento: 0.33
RMSE Prueba: 1299.11, R² Prueba: -0.39
RMSE Extrapolación: 1244.34, R² Extrapolación: -4.31

PHILADELPHIA- organic



Combinación: Philadelphia - organic
RMSE Entrenamiento: 3785.52, R² Entrenamiento: 0.61
RMSE Prueba: 13001.26, R² Prueba: -3.72
RMSE Extrapolación: 8564.59, R² Extrapolación: -0.34

NASHVILLE - conventional



Combinación: Nashville - conventional
RMSE Entrenamiento: 33468.97, R² Entrenamiento: 0.40
RMSE Prueba: 52019.88, R² Prueba: -0.60
RMSE Extrapolación: 37944.20, R² Extrapolación: -1.40

En algunos casos el modelo captura bien el comportamiento de los datos, mientras que en otros únicamente los mantiene en el rango de confianza.

CONCLUSIONES

El objetivo principal fue identificar los factores que subyacen a las fluctuaciones de precios y volúmenes, así como evaluar la viabilidad de desarrollar modelos predictivos robustos.

Impacto Estacional y Eventos: El consumo aumenta notablemente con eventos como la Super Bowl y festividades de mayo, afectando volumen y precios.

Auge de Aguacates Orgánicos: Su popularidad y volumen de ventas crecen constantemente, a diferencia de los convencionales, más estables.

Relación Volumen-Precio Inversa: A mayor volumen de aguacates en el mercado, el precio tiende a bajar, y viceversa.

Mercado Regional Heterogéneo: Existen diferencias significativas en consumo, precios y preferencia por orgánicos entre regiones de EE. UU..

Modelos Clásicos Limitados: Regresión lineal y polinómica ofrecieron poca utilidad predictiva para este mercado. Random Forest tendió al sobreajuste en extrapolaciones.

SARIMA, Mejor Opción Estacional: Este modelo fue el más apto para capturar la estacionalidad en las series de volumen, requiriendo una cuidadosa parametrización.

Importancia de Ingeniería de Características: Crear variables como lags de volumen/precio y marcadores de festividades fue relevante para los modelos.

Análisis de Cohortes y Madurez del Mercado: Este análisis reveló una mayor estabilidad en cohortes recientes, especialmente de orgánicos, sugiriendo una consolidación de estos mercados.