AGUACATES

Análisis y modelado predictivo del precio

Armen Hakobyan

ÍNDICE

Introducción

He decidido hacer este proyecto porque la verdad me encantan los aguacates, los utilizo siempre que puedo pero sin pasarme, por algo lo llaman el oro verde. Esto me llamó la atención y me fijé un poco y vi que los precios iban variando y me picó la curiosidad. Por este motivo he decidido hacer este proyecto sobre el análisis de los precios del aguacate.
La base de datos es de EEUU, esto se debe a que soy fanático de las peliculas de Hollywood y porque encontrar los datos decentemente estructurados ha sido más sencillo.
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 sacarle más provecho a los datos, exploración y corrección de los datos, etc.
También utilizaremos aplicaciones como NIFI, Cassandra para el tratamiento y la conservación de datos, Power BI y R para visualizar los datos.
Lenguaje SQL para realizar consultas...
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:

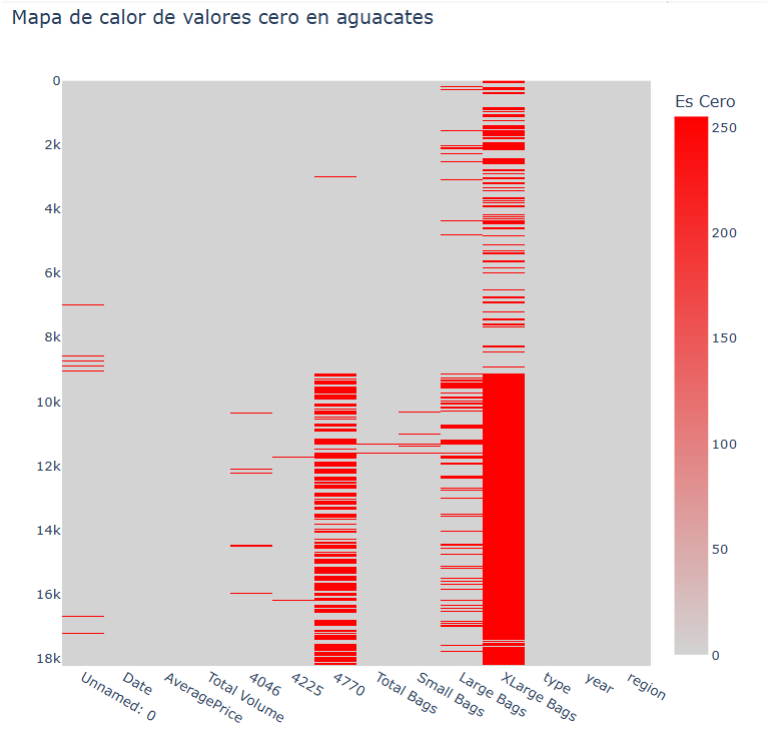
ORANGE	PYTHON	R
POWERBI	TABLEAU	GOOGLE COLLAB
GITHUB	VIRTUAL BOX	EXCEL

1. 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:



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.

```
df.head()
```

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. Esto puede dar lugar a confusión ya que no siempre que los datos sean 0 significa erróneo, podría significar también que la cantidad es tan insignificante que se redondea a 0.
He decidido comprobar si la suma de los diferentes tipos de bolsa da como resultado “Total Bags”, de este modo si hay una diferencia muy alta en los casos donde XLarge o cualquier otra sea 0 quiere decir que los datos que son 0 son datos faltantes, de lo contrario se da por hecho que no se ha comercializado con este producto en las zonas y no se tiene en cuenta.

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.
```

1.1 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.

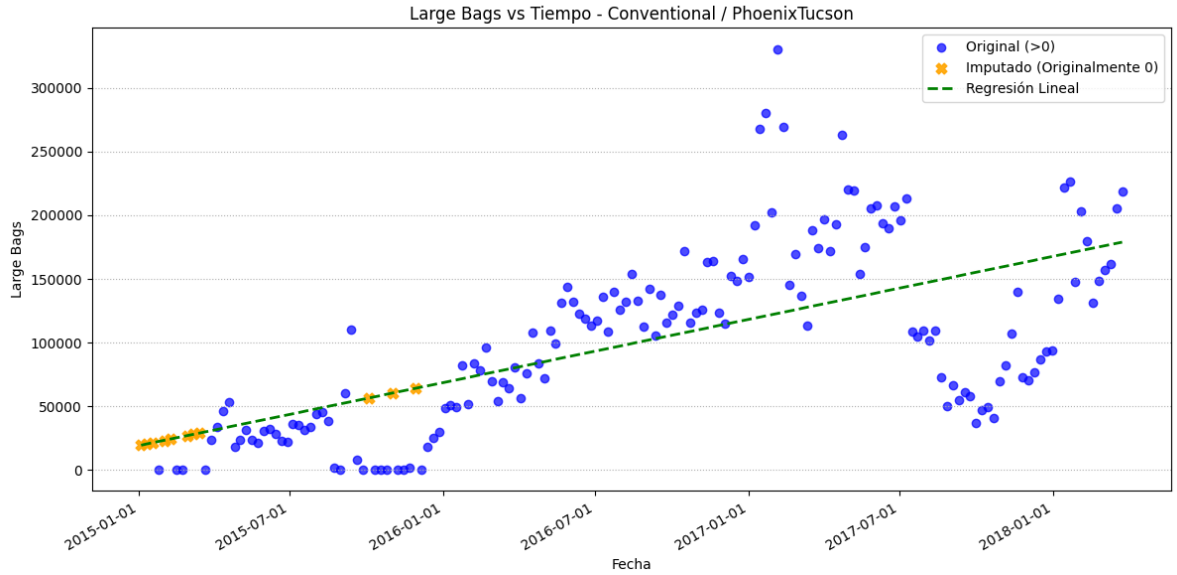
- De lo contrario, se generará un modelo de regresión lineal que servirá para rellenar los datos que son 0 con el valor predicho, aunque no sea muy preciso ni dependa de la estacionalidad nos será útil para hacer cálculos más aproximados.

- 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.

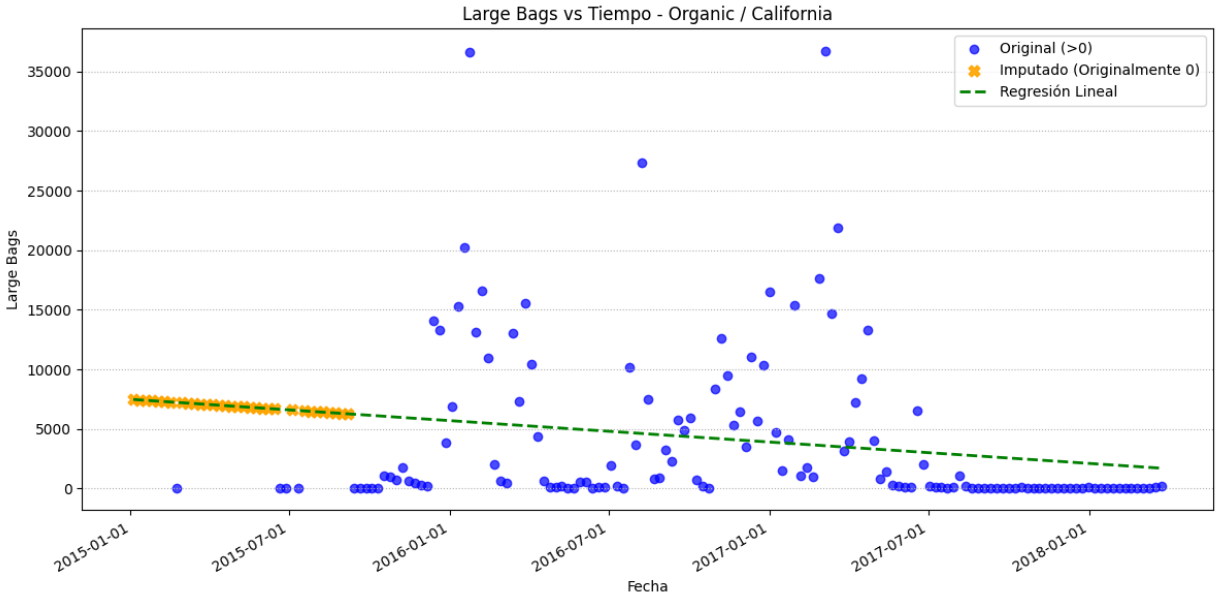
Una vez tenemos el modelo preparado, podemos indicarle de qué región región queremos graficar el modelo y las predicciones para los valores 0. Por ejemplo podemos ver el caso de PhoenixTucson.

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



O California, que podemos observar como con el paso del tiempo la gente ha perdido interés por las bolsas grandes del tipo orgánicos.

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

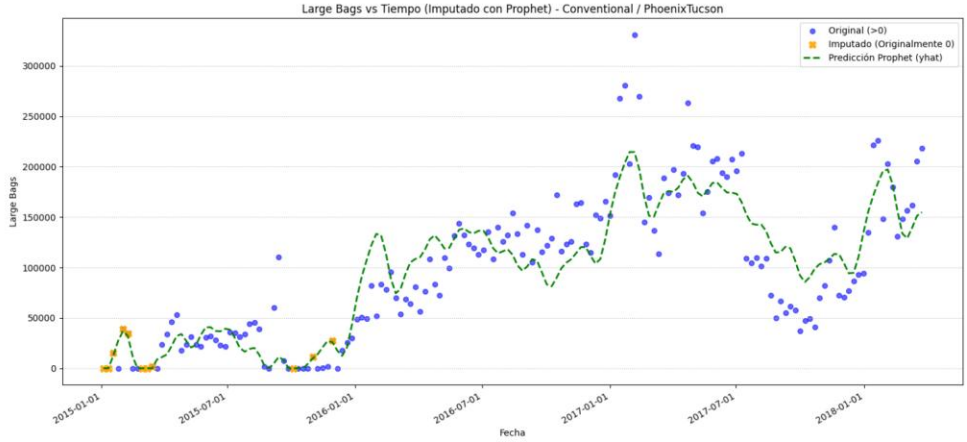


Al hacer esto, me doy cuenta que esto influirá en los datos de manera muy significativa, y no necesariamente a mejor. Si es cierto que en algunos casos la predicción no es tan mala, pero por ejemplo en california, al rellenar las bolsas vendidas con una regresión lineal nos da lugar a muchas confusiones y altera la veracidad de los datos.

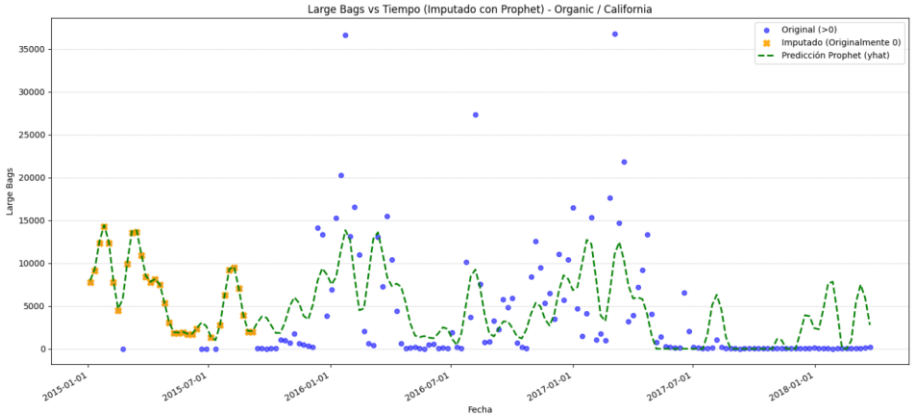
PROPHET Y KNN

Visto que las predicciones anteriores no han sido las más acertadas, decido probar un par de modelos para ver si vale la pena rellenar estos 0.
Aunque el volumen total si que aumenta estacionariamente, los valores separados como Large Bags no necesariamente sigue ese patrón. Por lo que las predicciones de Porphet no han sido muy útiles, ya que tergiversaron bastante los datos de manera probablemente errónea.

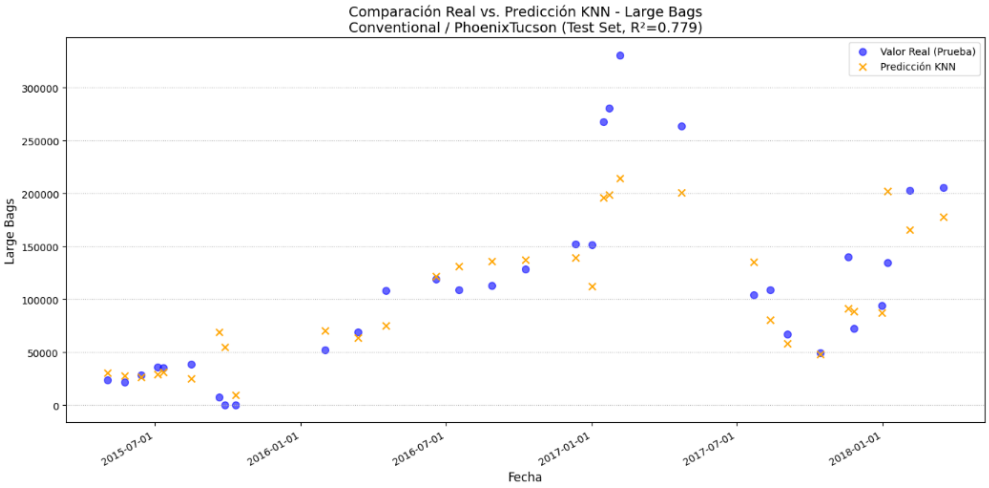
Siguiendo con los ejemplos del caso anterior, para PhoenixTucson vemos que prejudice mejor que el modelo de regresión lineal pero continua sin ser lo suficientemente útil, ya que nos tergiversa mucho los datos y nos puede perjudicar demasiado, ya que en casos de regiones en los que tengan muchos datos 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.



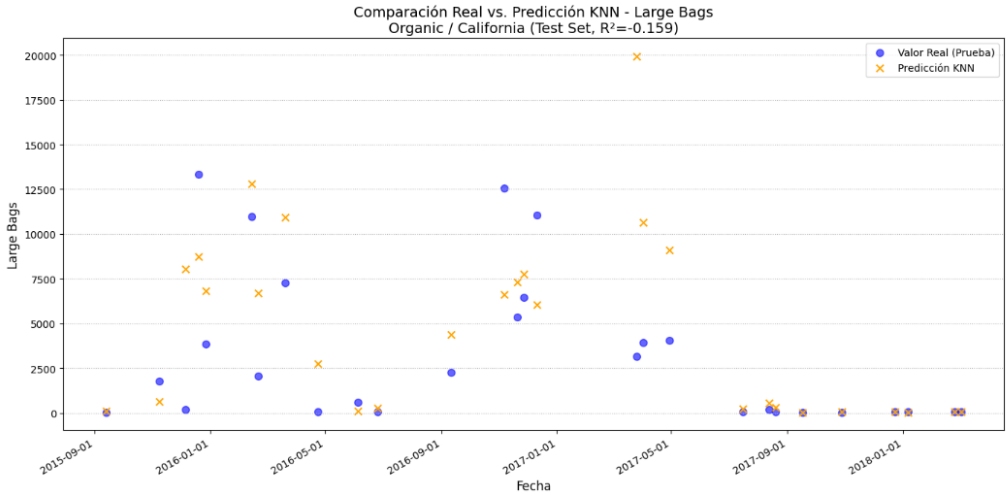
Observamos que aunque no es mal del todo, en muchas ocasiones nos causará más problemas que información aportada. Con los orgánicos de California vemos que sigue una tendencia algo más clara, y aunque en el final del periodo 2018 decae bastante, mi modelo sigue prediciendo el patrón estacional. Este 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, no es tampoco algo que se pueda implementar, 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:



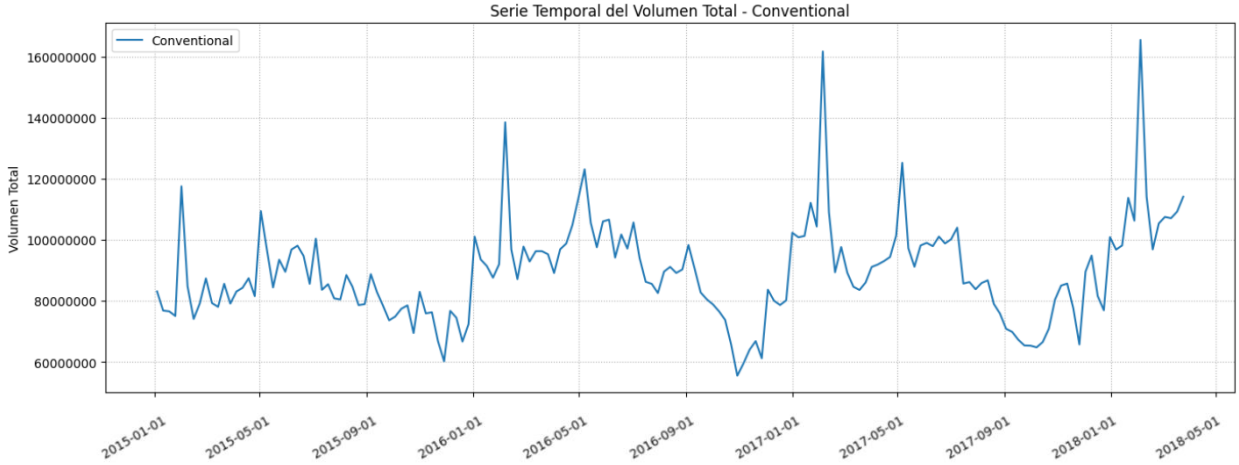
Pero de nuevo, estas predicciones no resultan muy útiles ya que aunque salieran bien, se sumaría X cantidad a cada región y todo se equilibraría de nuevo.

Por lo que después de realizar estas pruebas decido dejar los datos en 0 tal y como estan..

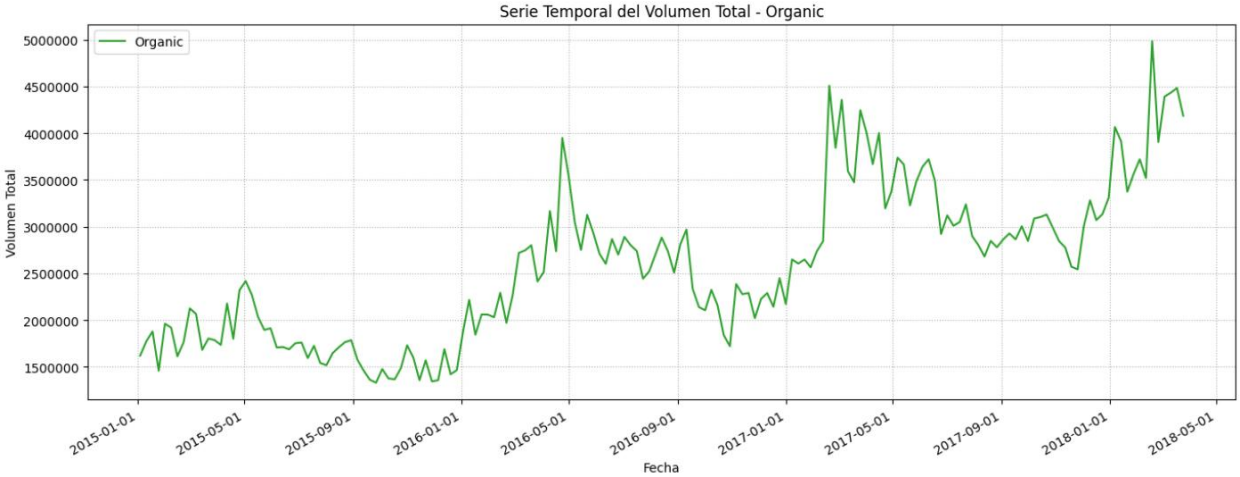
2. ANÁLISIS DE VOLUMEN

Antes de continuar rellenando los datos faltantes, es bastante oportuno analizar la tendencia de Total Volume y ver cómo actúa a lo largo del tiempo.

Los aguacates de tipo convencional se observa bastante bien 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.



Por lo contrario vemos una tendencia muy diferente. Si es cierto que los picos y los valles coinciden, pero los máximos y mínimos de los orgánicos han variado mucho desde que se tienen registros, lo cual indica un aumento de la popularidad de este producto.

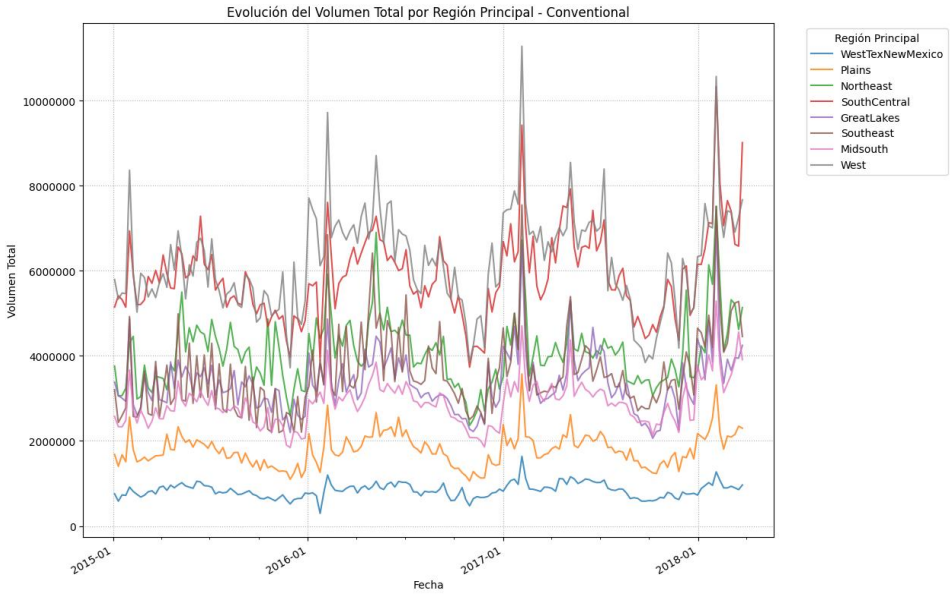


En ambos gráficos observamos una gran coincidencia en el patrón de comportamiento. A principios de año, aproximadamente en Febrero, vemos un pico. Volvemos a observar repuntes a medida que se acerca el mes de Mayo. Indagando la razón de este consumo elevado, me encuentro que en Febrero se celebra la Super Bowl, un evento de rugby muy popular en EEUU y es muy probable que sea debido a esto. 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.

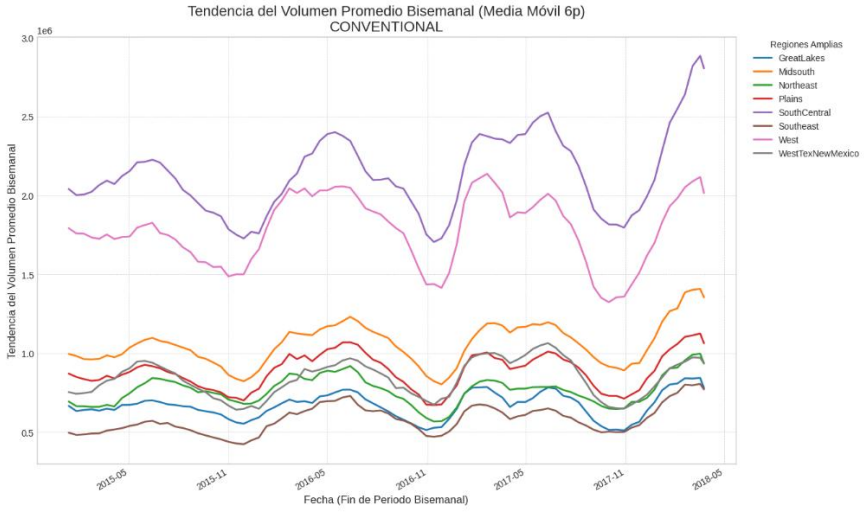
Para entender mejor cómo están separados los datos y poder hacer el análisis correctamente debemos observar nuestra característica Region, ya que no es lo que parece. Observamos una jerarquía, por lo que he creado una columna llamada clasificación en el que las ciudades que pertenecen a regiones amplias hereden el nombre de la región amplia, por comodidad.

Serie temporal Total Volume por región amplia / Convencional.

- Semanal



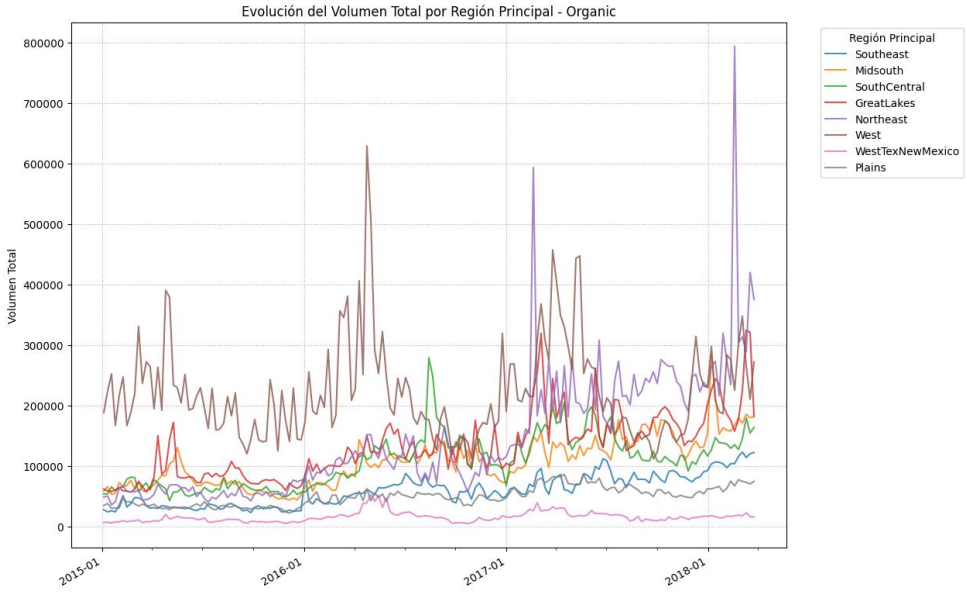
- Bisemanal, suavizado media movil 6p.



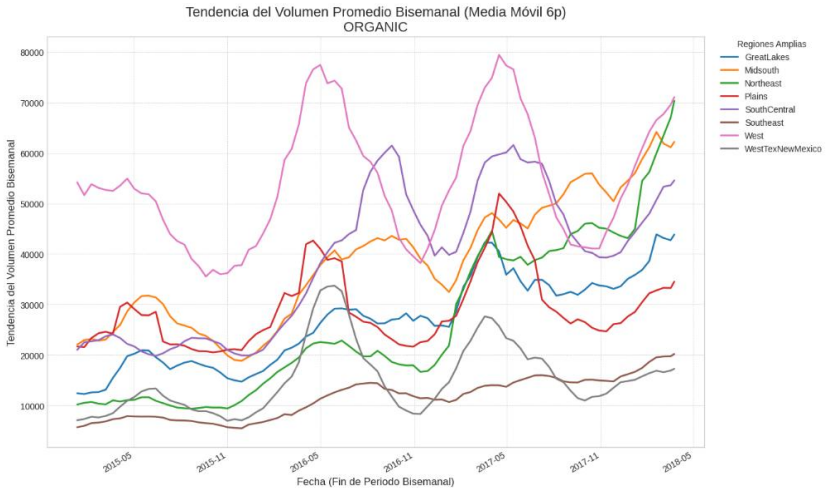
En el gráfico anterior se observa claramente cómo hay ciertos eventos anuales que acentúan el consumo de aguacate. Los volúmenes tan dispares entre ciertas zonas es debido a la población que las habitan, ya que en WEST por ejemplo hay más densidad de población que en PLAINS. Observamos también volúmenes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

Serie temporal Total Volume por región amplia / Organic.

- Semanal



- Bisemanal, grafico suavizado media movil 6p.

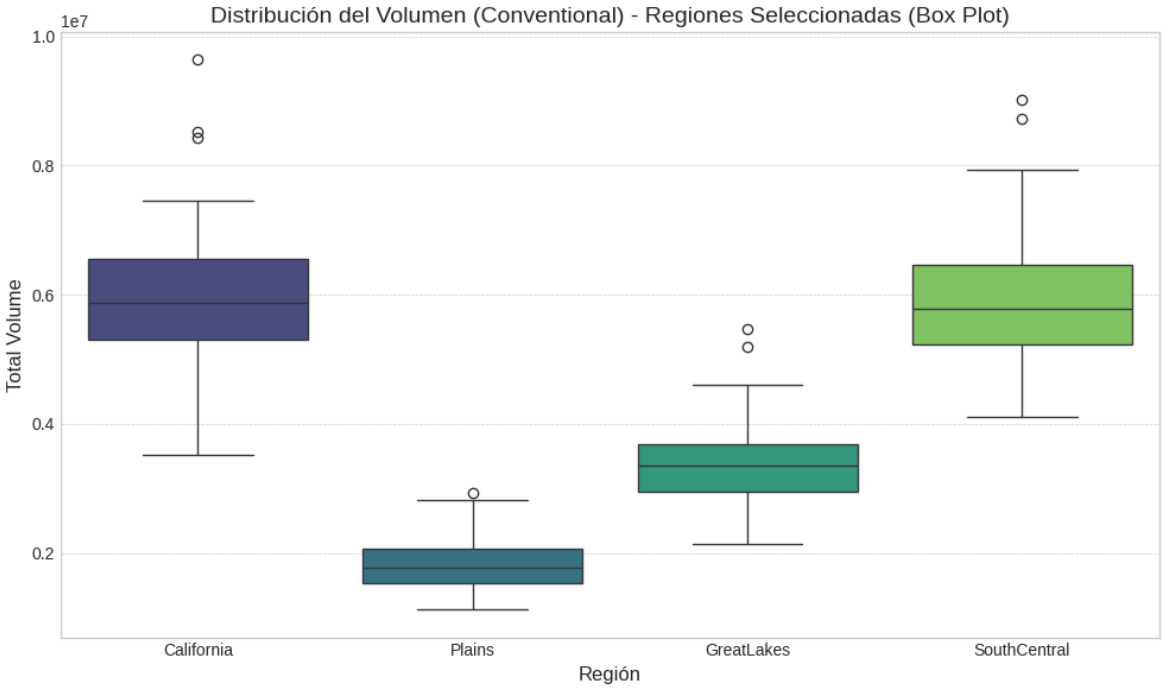


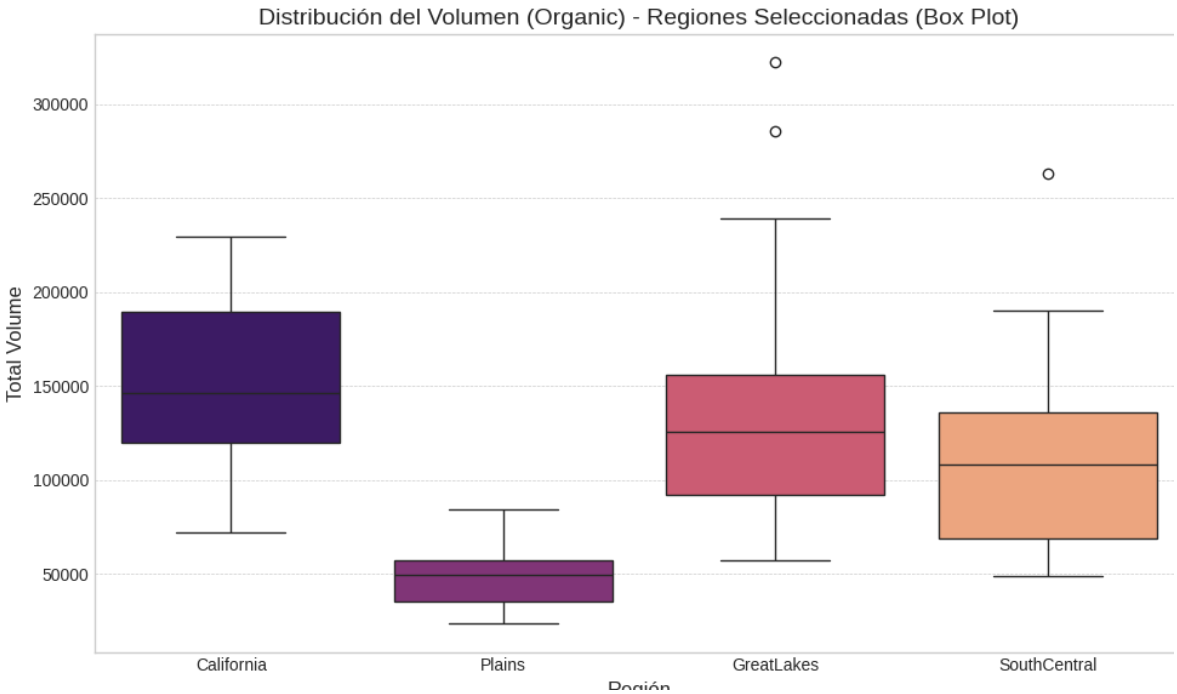
En cambio en los aguacates orgánicos se da un caso algo mas diferente. Observamos como en la mayoría de las regiones se ha ido popularizando el aguacate orgánico ya que el volumen total tiene una tendencia a la alza. Esto indica que la población se ha ido interesando acerca de la calidad de los alimentos y optan por las variedades más saludables, también porque no utilizan pesticidas, tienen más tacto con el medioambiente e incluso el sabor es diferente.

En ambos casos vemos que al agrupar los datos bisemanalmente no se pierde información de la tendencia y se ve más legible, por lo que utilizaremos esta agrupación para el análisis.

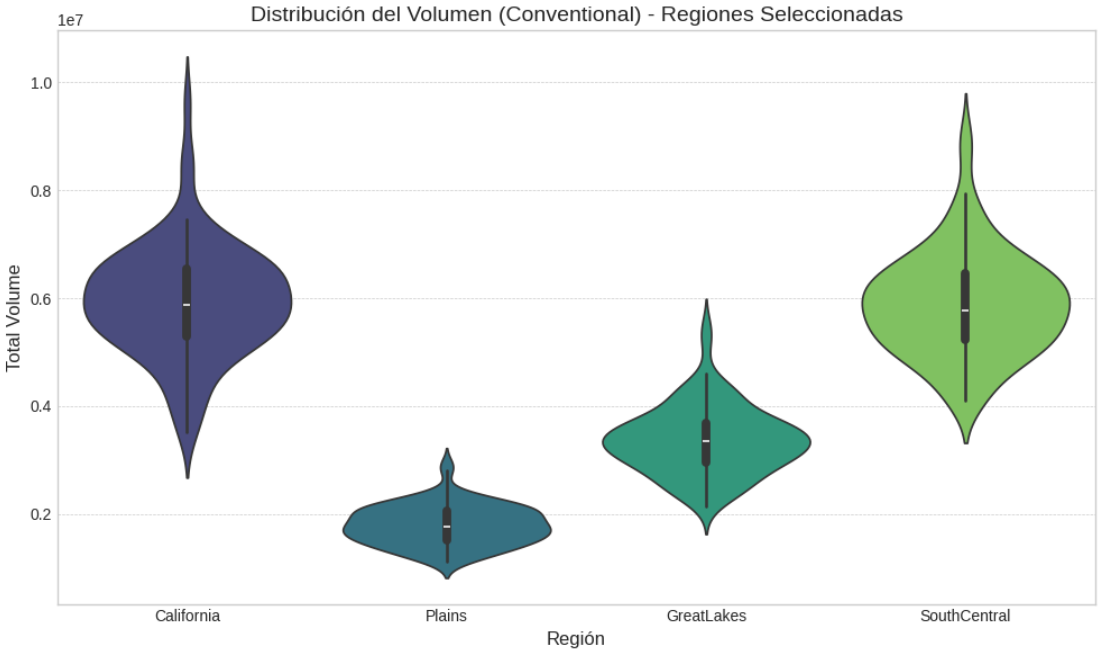
2.2 BOXPLOTS Y VIOLINS

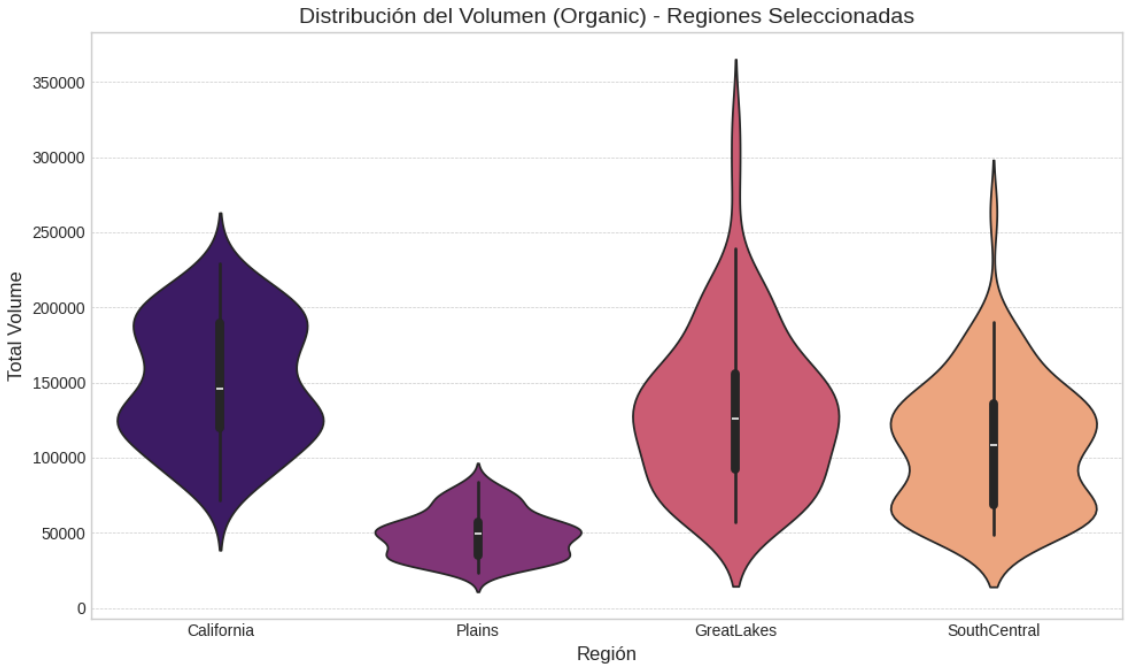
Con los siguientes boxplots vemos la distribución de los datos. En la zona rectangular es donde se encuentran la mayoría de los datos y con los bigotes nos indica un rango donde hay datos pero en menor medida. Esto quiere decir que a más largos los bigotes, mayor dispersión de los datos. Con los puntos blancos nos indica que hay valores que se salen del rango y los considera outliers.





Para estar seguros que realmente estos puntos son Outliers (casos específicos), lo haremos utilizando las gráficas de violines. Estas gráficas nos permiten entender de una forma más sencilla cómo están distribuidos los datos, ya que estos “Outliers” si son muchos, se tendrían que tener en cuenta.





Para tomar la decisión correcta, voy a sacar la fecha de estos Outliers 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 1.5 * IQR.

```
--- 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 mes de agosto si que es extraños en el caso de South Central, por lo que lo regulamos al máximo permitido

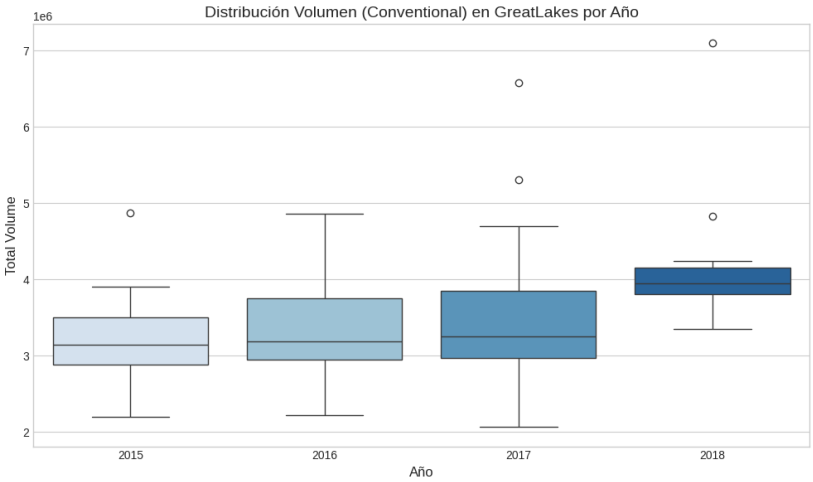
```
Se limitará el Total Volume a: 238019.0
Para: Region='SouthCentral', Tipo='organic', Fecha='2016-08-21'
Número de filas encontradas que coinciden: 1
Valor original de 'Total Volume': 263166.655
'Total Volume' modificado a: 238019.0
¡Modificación realizada con éxito!
```

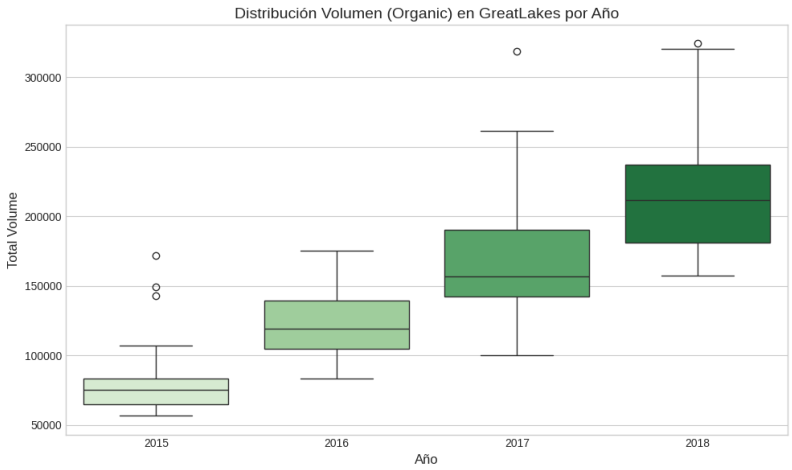
nprimir pot pantalla el registro de SouthCentral, organic y fecha 2016-08-21

```
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
```

Si observamos el boxplot de GreatLakes podemos ver como los orgánicos se han popularizado a lo largo de los años y esto ocurre en otras regiones también.

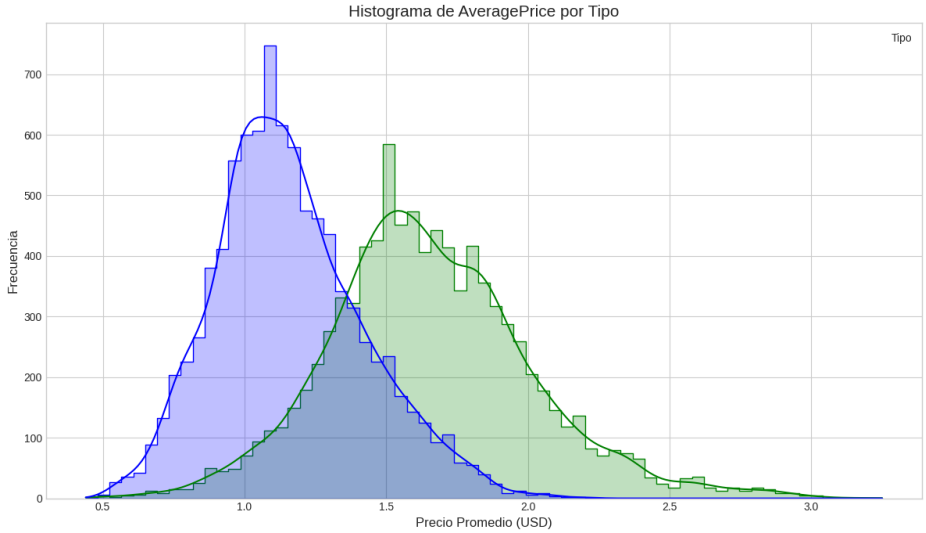




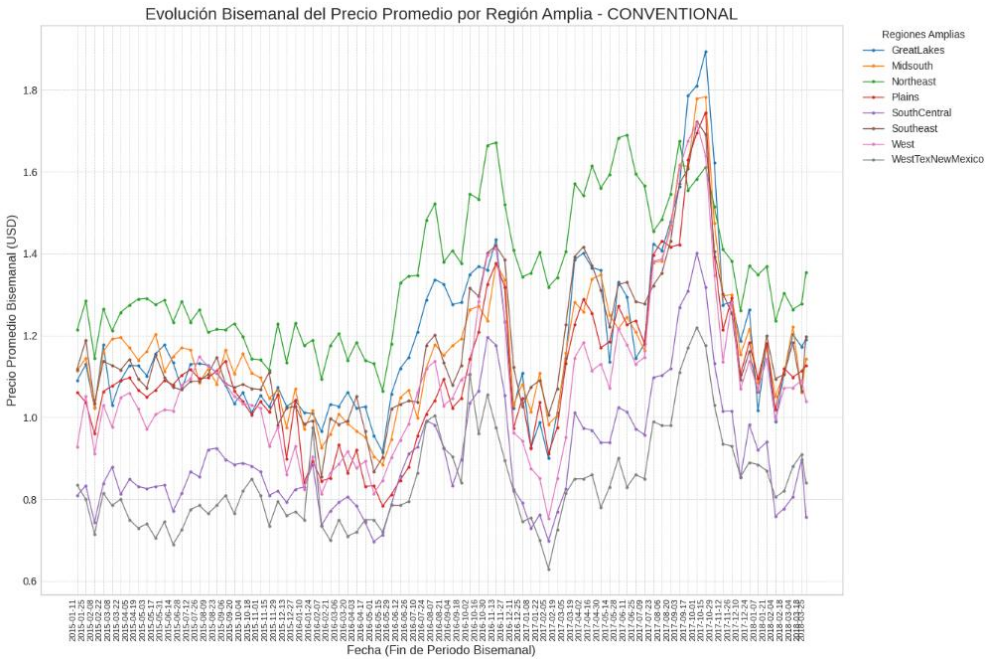
3. ANÁLISIS DE PRECIOS

HISTOGRAMA

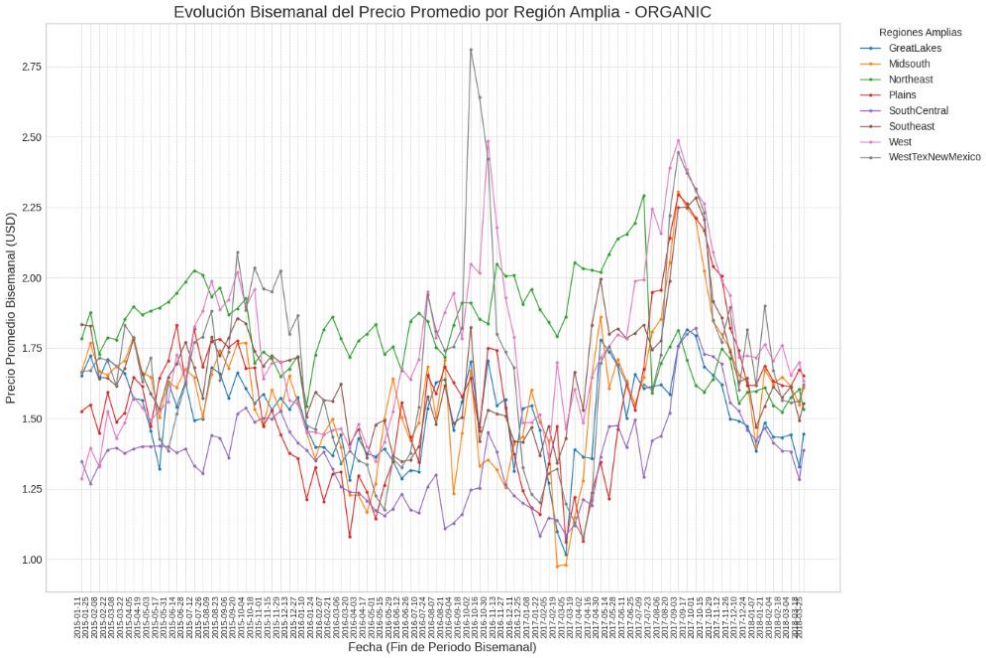
Se observa que el rango de precio de los aguacates orgánicos



Serie temporal bisemanal / Convencional

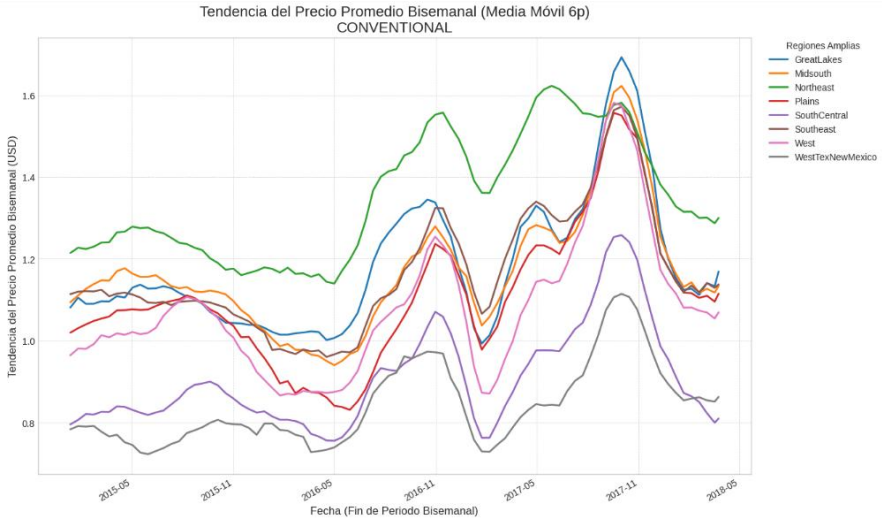


Serie temporal bisemanal / Organico



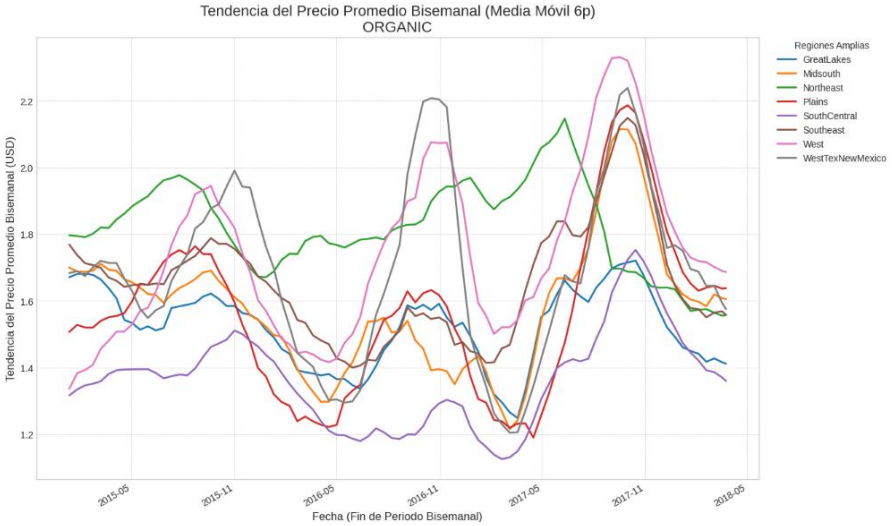
Veamos las tendencias para entenderlo mejor

Convencional



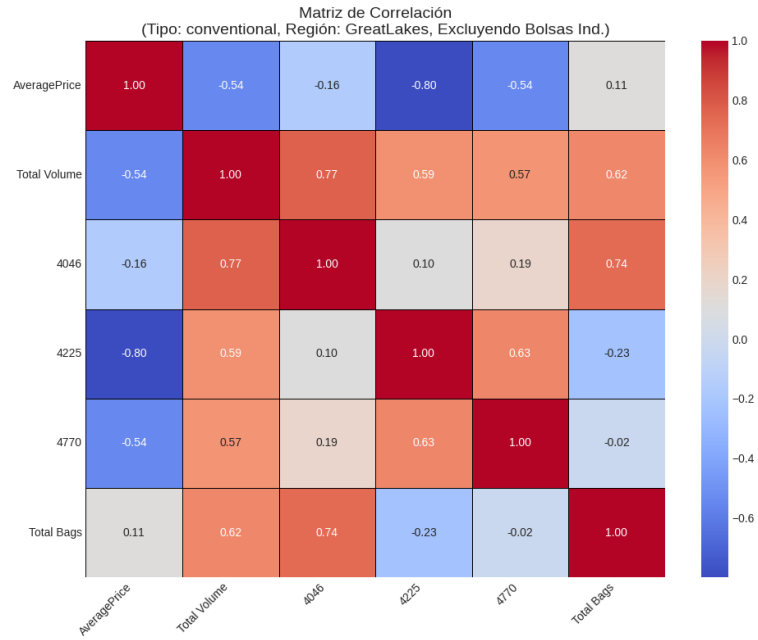
Respecto a la tendencia, observamos que es a la alza, los picos cada vez son más pronunciados y los valles son más elevados.

Orgánico

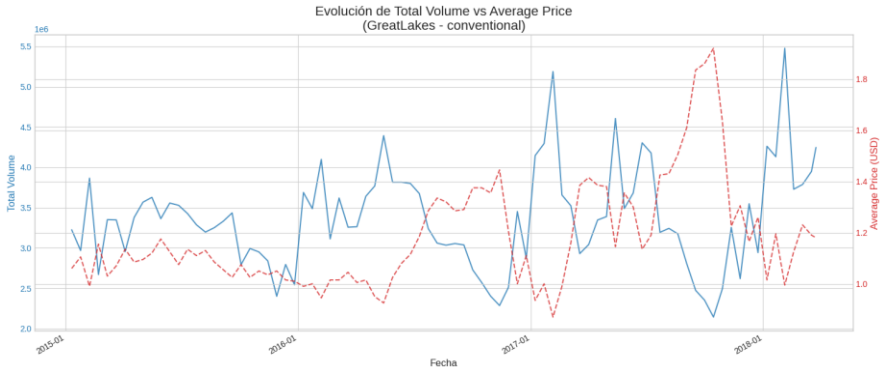


En cuanto a los orgánicos nos fijamos que los precios se mantienen bastante estables. Podemos notar como los picos si que son más altos a medida que avanza el tiempo pero por lo general se vuelve a normalizar.

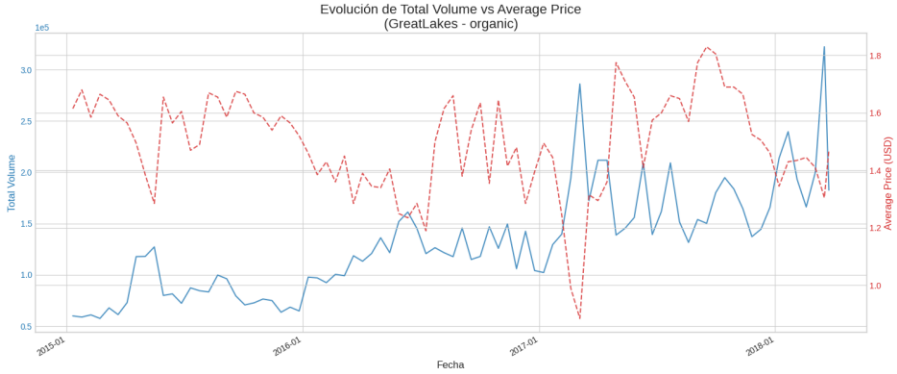
En la siguiente matriz de correlación podemos observar la correlación entre las características del dataframe. Para GreatLakes y el tipo convencional, se observa una correlación negativa entre Total Volume y Average Price.



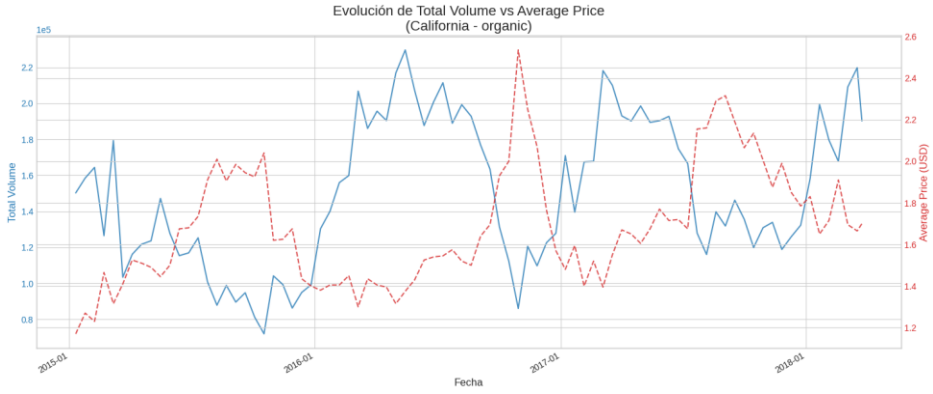
Para demostrar esta correlación negativa, lo podemos graficar para ver mejor su evolución.



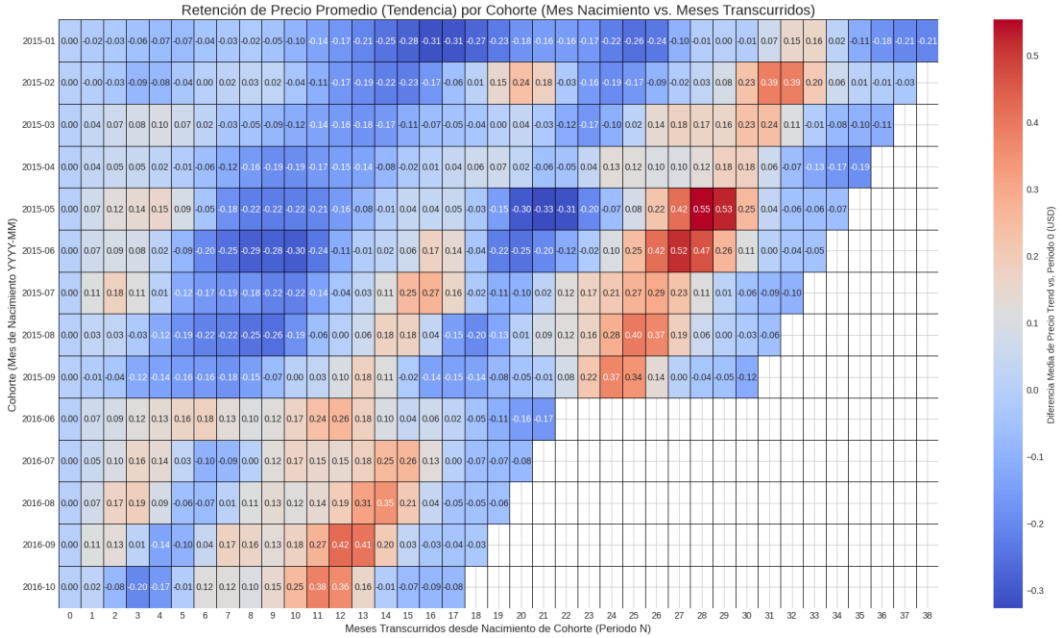
Observamos claramente cuándo Total Volume sube, AveragePrice baja.
Lo mismo ocurre para los aguacates orgánicos, pero vemos como el volumen total poco a poco tiene bajadas menos pronunciadas y cada vez están mas asentados en el mercado.

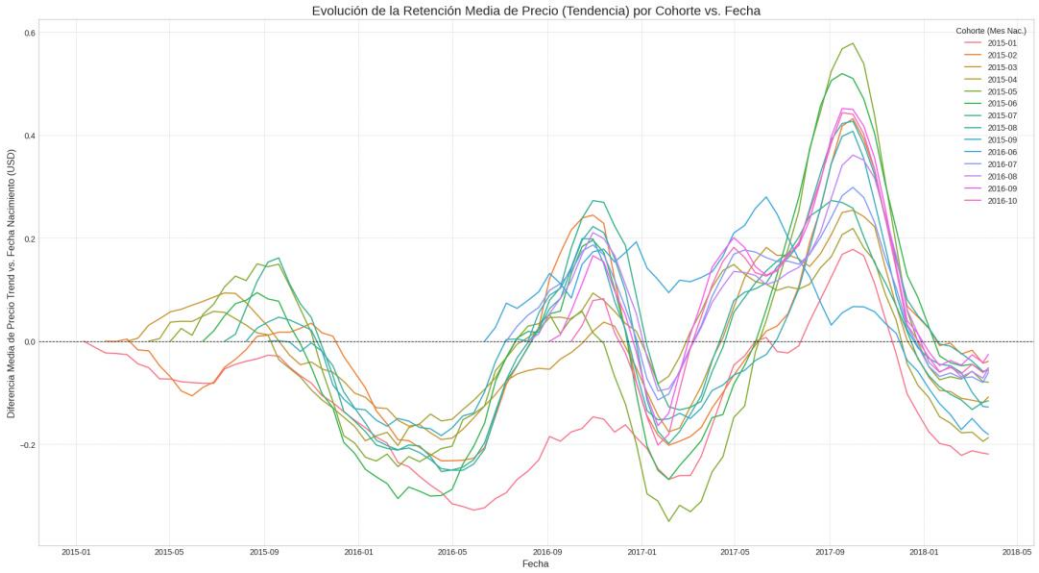


Podemos ver otro ejemplo, California organic, donde podemos observar un patrón muy similar.



Cohorte de retención el primer mes es el que superan la media de su precio promedio, a los meses siguientes se le resta el valor del primer mes para ver la retención.



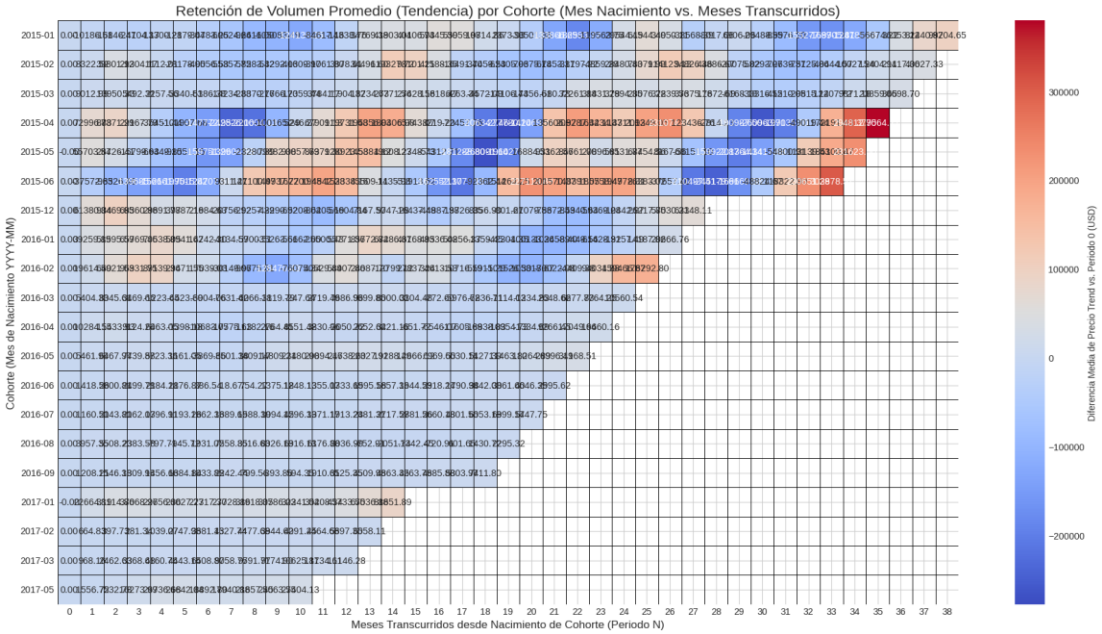


Vemos que en líneas generales, cuando una cohorte supera por primera vez su media el precio tiende a subir y luego caer por debajo. Esto se debe a que la subida de precio está relacionado con la comercialización en esas épocas del año. A medida que se acercan las etapas de mayor consumo, observamos un 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.

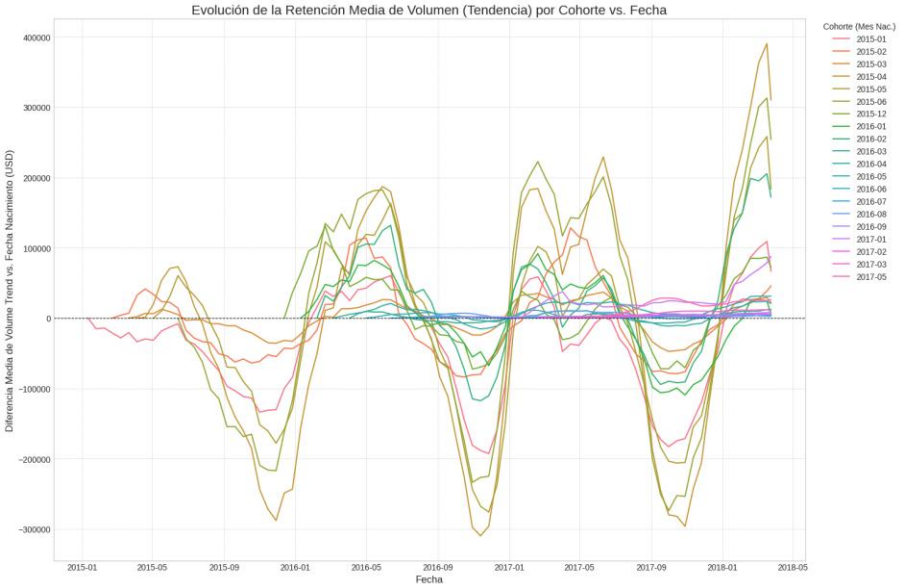
Si nos fijamos en el grupo 2015-05 podemos ver que su pico es el más alto, seguido de su vecino del mes 06. Del grupo nacido en 2016-06 se observa que tiene una retención extremadamente positiva, pero a finales de 2017 /inicios de 2018 vemos una caída.

4. TOTAL VOLUME

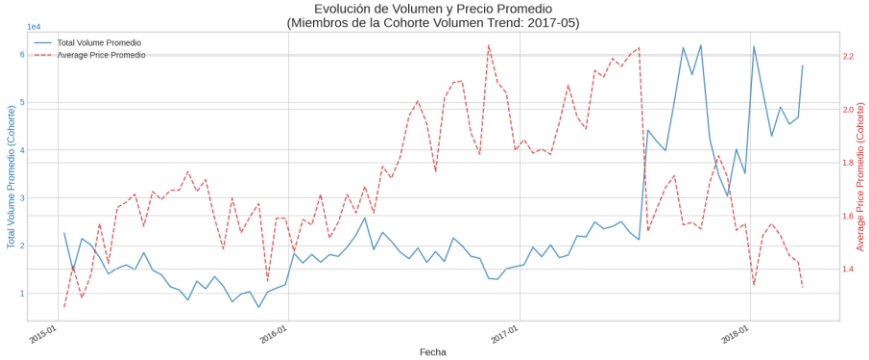
Estas son unas nuevas cohortes pero con Total Volume. Observamos como en los grupos de cohortes que nacen antes de 2016 tienen muchas variaciones en comparación a los grupos más recientes, en los que se observa una clara estabilidad



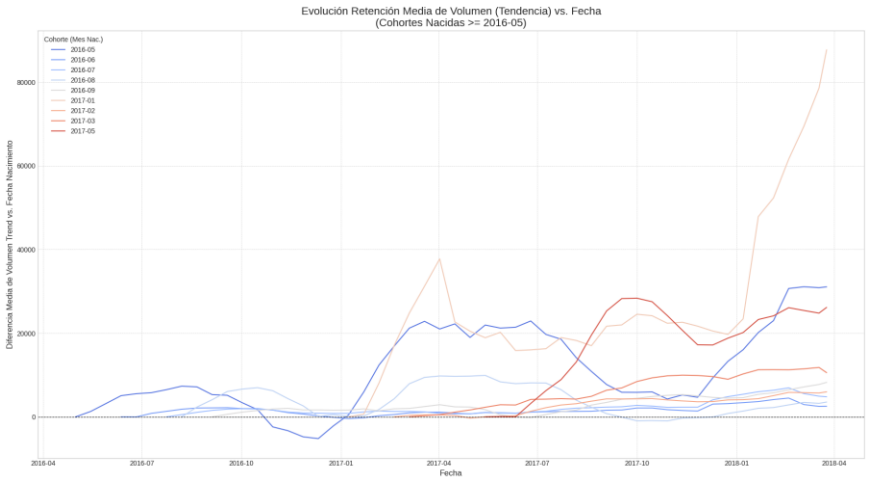
Si miramos el gráfico vemos esto bastante más claro.
Observamos que cuando hay más volumen de aguacates el precio tiende a bajar, ya que la demanda no es la suficiente.



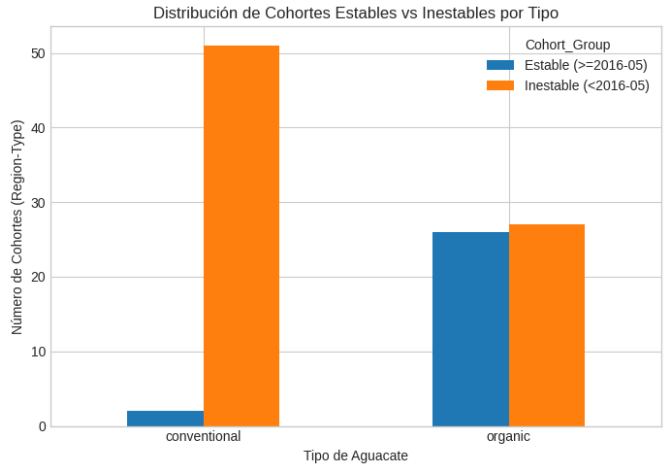
La región perteneciente a la cohorte 2017-05, BaltimoreWashington, tiene un volumen bastante constante, si observamos la gráfica Volumen- Precio podemos observar cómo a partir de 2017-05 los valores de volumen sube, baja el precio.



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

5. ANÁLISIS POR REGIONES ESPECÍFICAS

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)

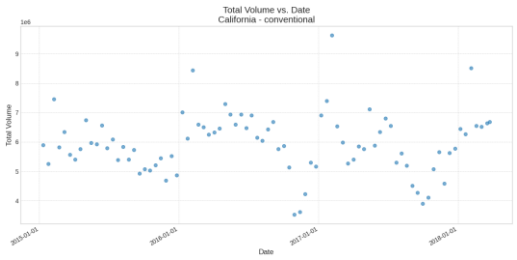
West: California
Es un mercado enorme, a menudo un referente, y sus dinámicas pueden ser complejas y distintas a otras zonas. Es bueno tenerlo para representar la costa Oeste y un alto volumen.

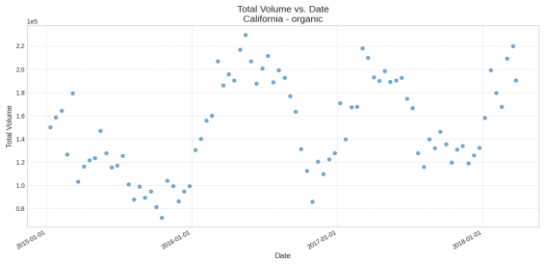
Southeast: Roanoke
Es de los pocos orgánicos que supera su media en cohortes tempranas.

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.

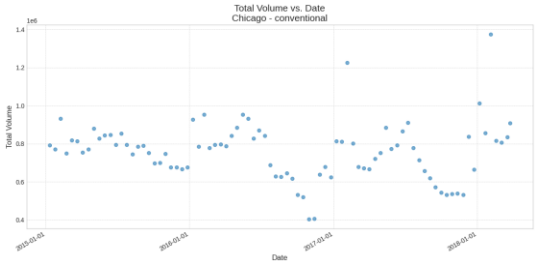
Midwest: Chicago
Excelente para tener un gran mercado del Medio Oeste, con su propia dinámica.

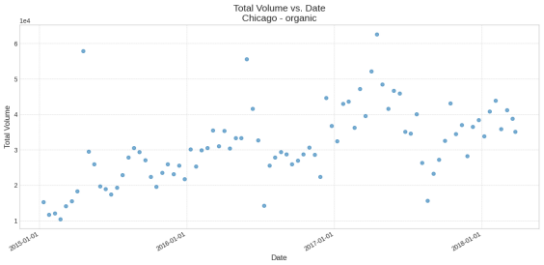
VOLUMEN - CALIFORNIA



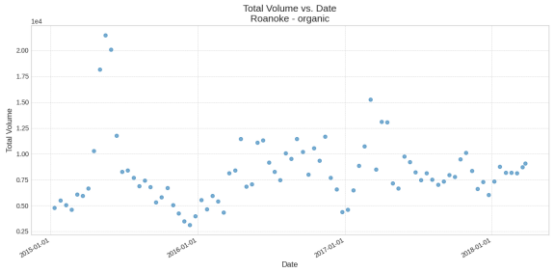


VOLUMEN - CHICAGO

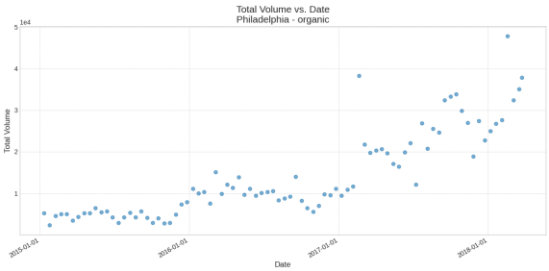




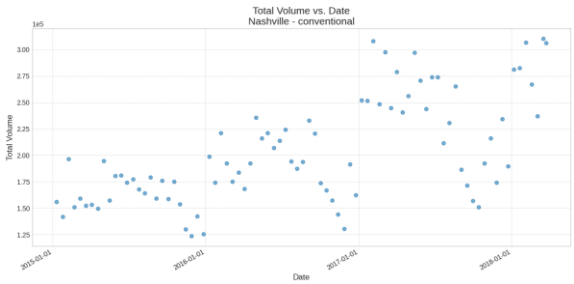
ROANOKE



PHILADELPHIA



NASHVILLE



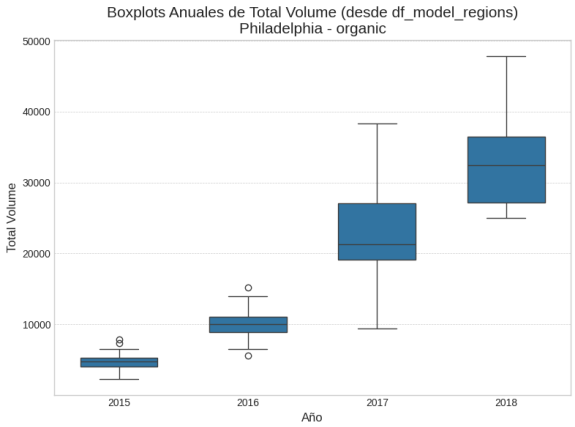
Observamos bastantes outliers. En Roanoke en 2015 vimos un repunte muy alto sobre los productos orgánicos. Esto es muy probable que se deba a que la fundación de Roanoke Natural Foods Co-op celebró su 40 aniversario, realizando campañas de márketing para popularizar este tipo de productos. Al tratarse 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

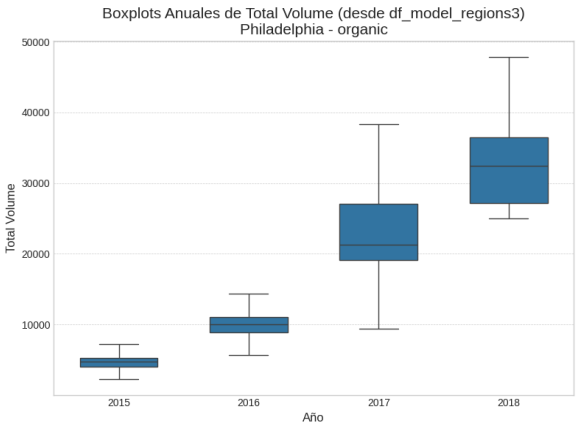
Si bien es cierto que en muchas ocasiones caen en eventos especiales, estos outliers no nos benefician en absoluto a la hora de generar el modelo. Con la misma técnica de IQR, cambiaremos estos valores para toparlos al Q3 + 1.5 *IQR.

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

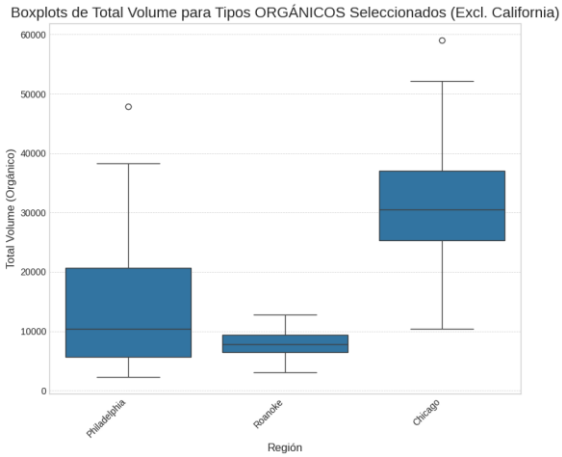
DATAFRAME SIN TOPAR



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.

6. 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

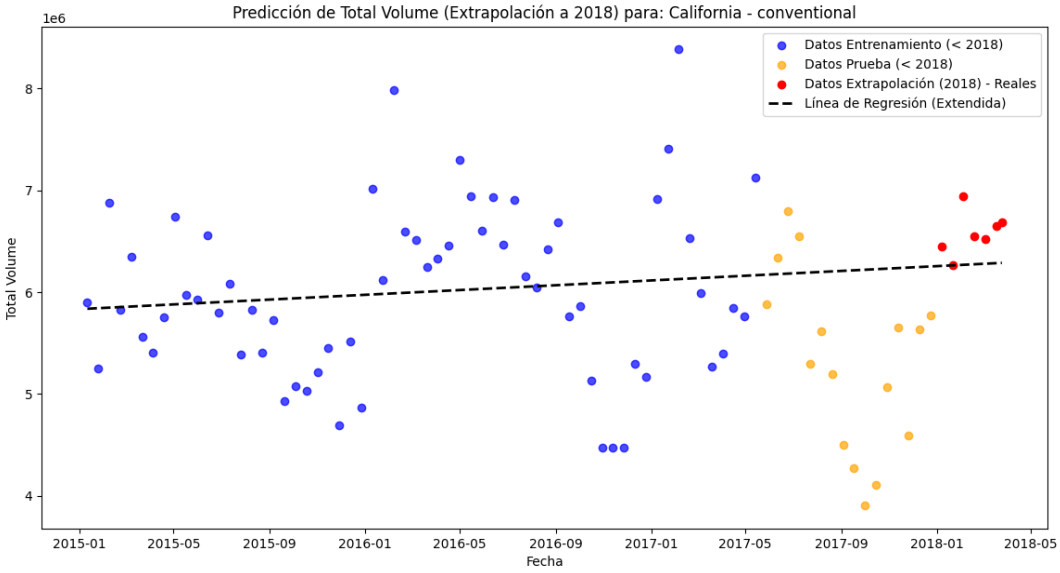
Observamos Lag 26 de Total Volume que nos indica la cantidad de volumen de hace 26 periodos. Lag 1 de Price nos indica el precio del periodo anterior. En los Near, se indica si se acerca esta festividad, es 1 mientras 4 periodos antes de la fecha y en los 2 siguientes.

Se aplican estos cambios al data frame original para que todos los datos sean uniformes.

7. MODELOS PREDICTIVOS

7.1 REGRESIÓN LINEAL

Generalmente utilizado con datos con tendencias claras y poca variabilidad en datos. No nos será de gran utilidad debido a la naturaleza del dataset.
Modelos bastante inútiles para estos casos, ya que los datos tienen comportamientos variados.



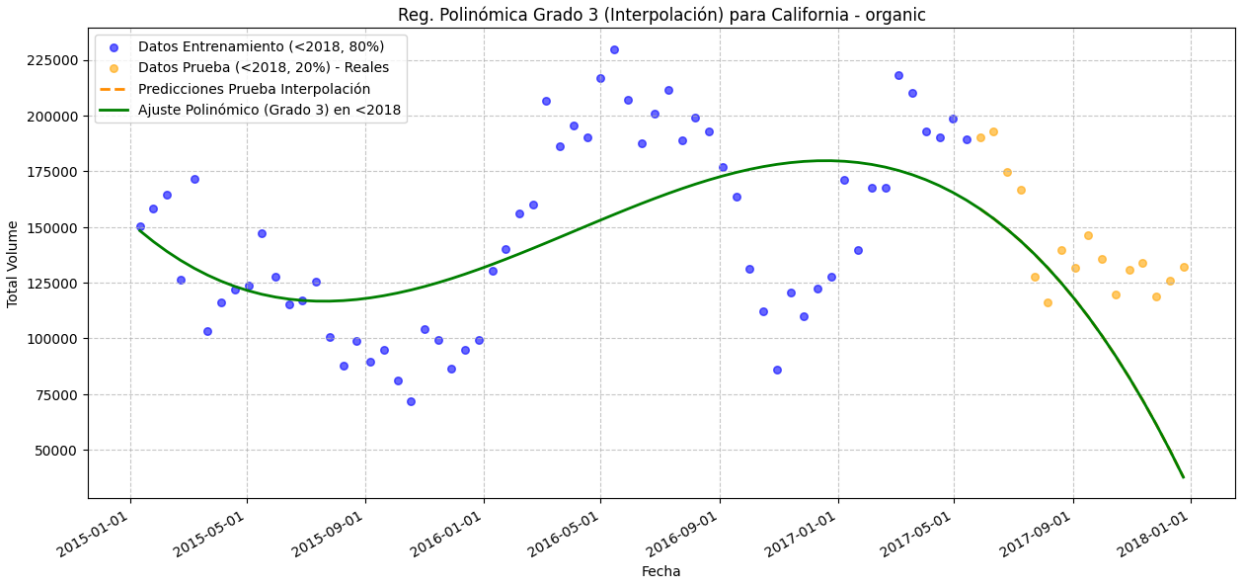
7.2 REGRESIÓN POLINÓMICA

Para la regresión polinómica 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
```

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

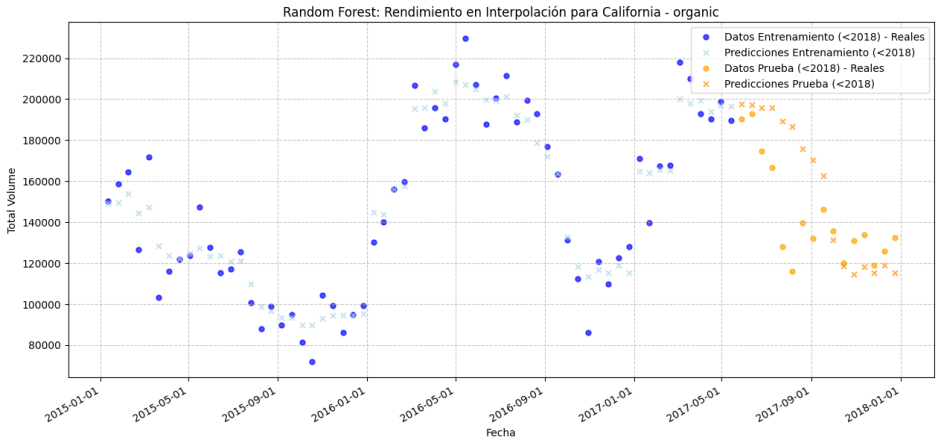


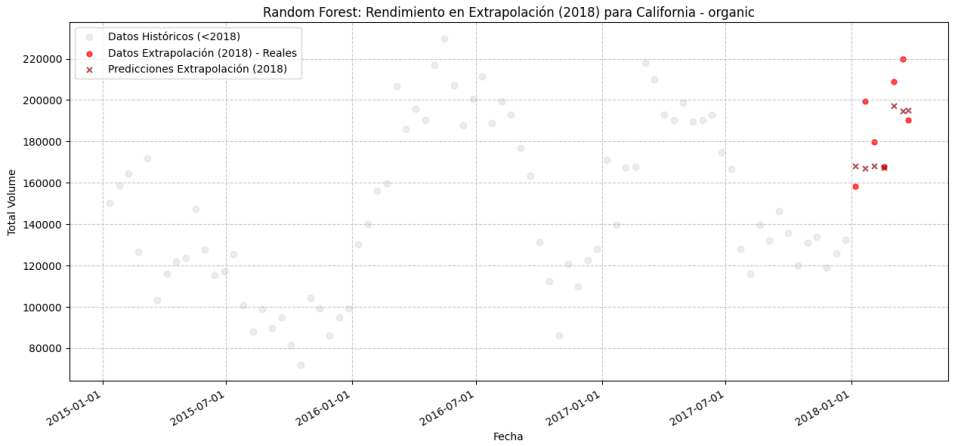
Obtenemos como resultados

```
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
```

7.3 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.

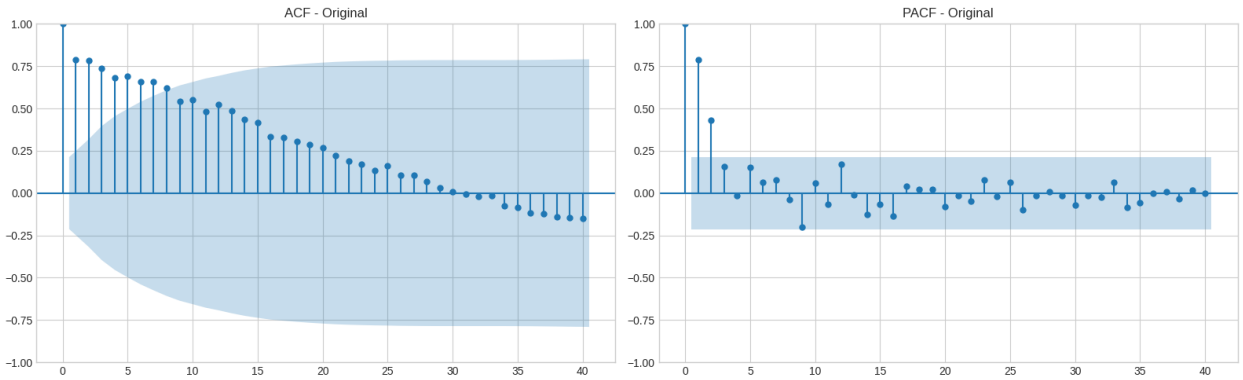
```
Iniciando GridSearchCV para Random Forest (con 3-fold TimeSeriesSplit)...
Fitting 3 folds for each of 36 candidates, totalling 108 fits
Mejores parámetros encontrados:
{'max_depth': 3, 'min_samples_leaf': 5, 'min_samples_split': 40}

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 ve claramente que reduce el sobreajuste, funciona algo mejor con los datos de prueba pero sigue siendo muy deficiente para predecir.

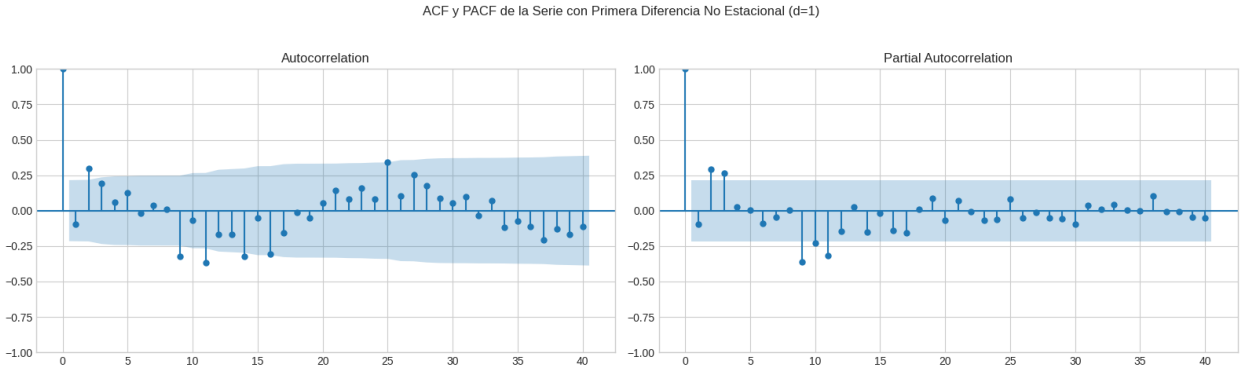
7.4 SARIMA

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 estos modelos trabajan mejor de esta manera. Observamos los datos de entrenamiento de color azul, los de prueba en naranja y las lineas discontinuas son las predicciones del modelo.
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.

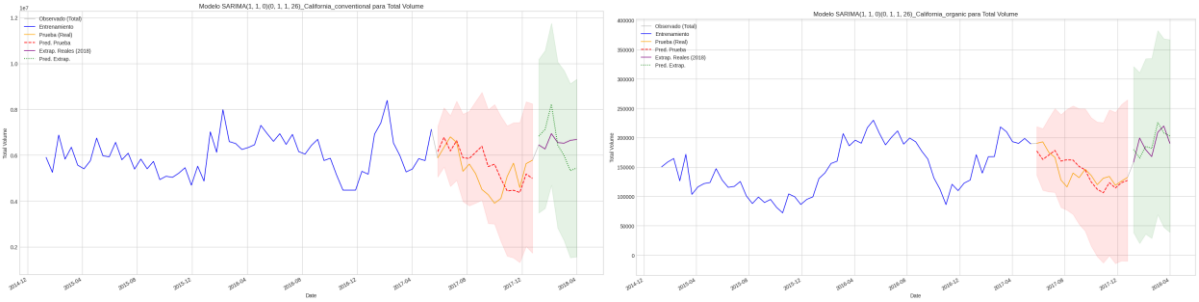


Aplicando una diferencia no estacionaria “d=1” si que podemos observar una estacionalidad, lo que esto hace es, en vez de tener en cuenta el dato, tiene en cuenta la diferencia del dato actual con el dato anterior. En este caso sí que se observa una estacionalidad clara.



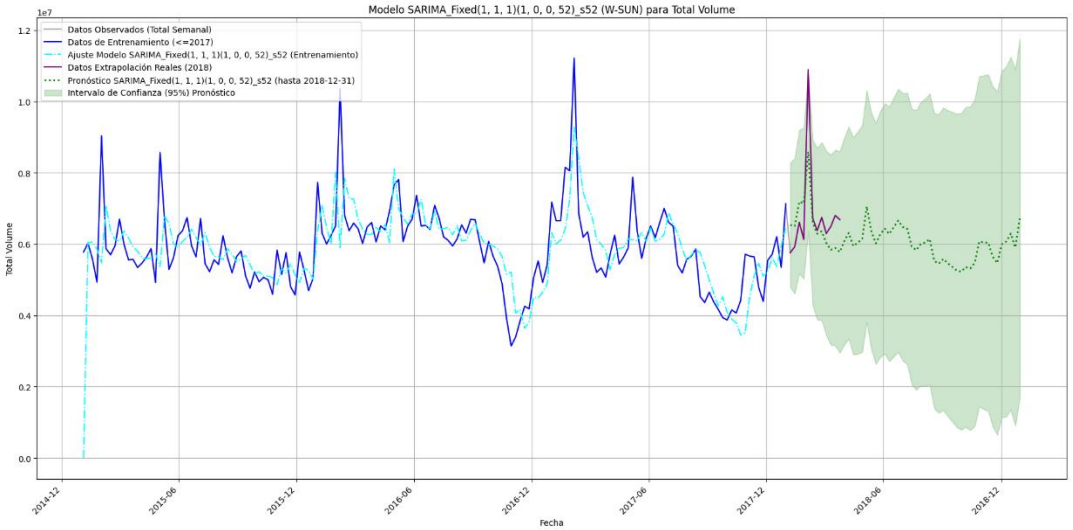
Aplicaré los parámetros encontrados sobre las combinaciones mencionadas para ver el comportamiento de SARIMA.

CALIFORNIA

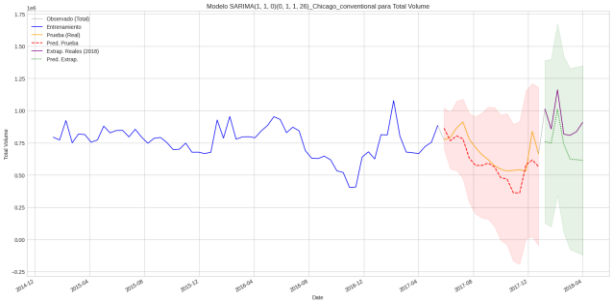


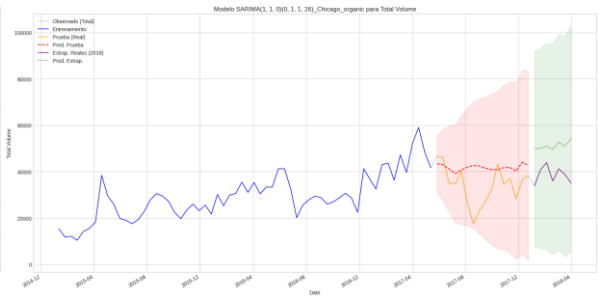
Como no actua muy correctamente, haremos el análisis de parámetros para california-conventional e intentar predecir de manera más correcta su comportamiento.

Utilizaremos KFOLD para california convencional y seleccionará los mejores parámetros basado en R^2. He extendido las predicciones hasta final de año de 2018.

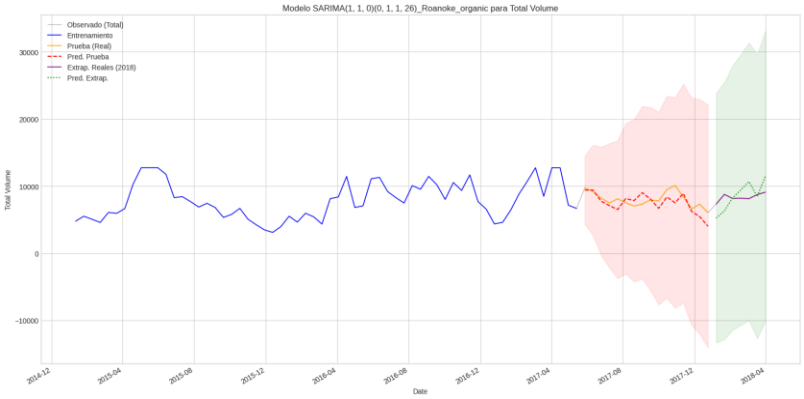


CHICAGO

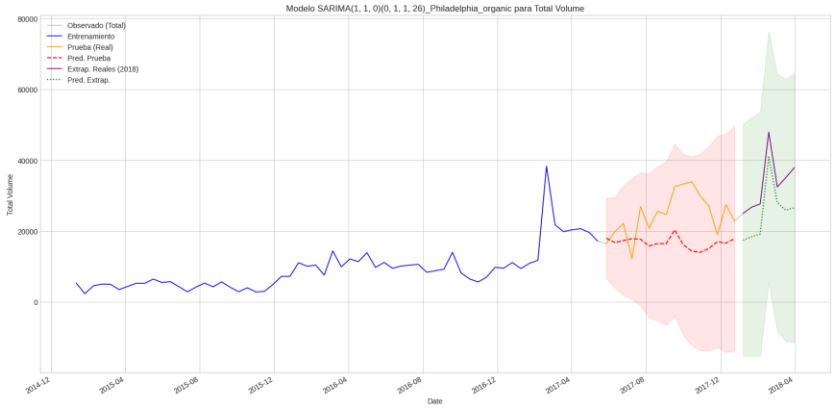




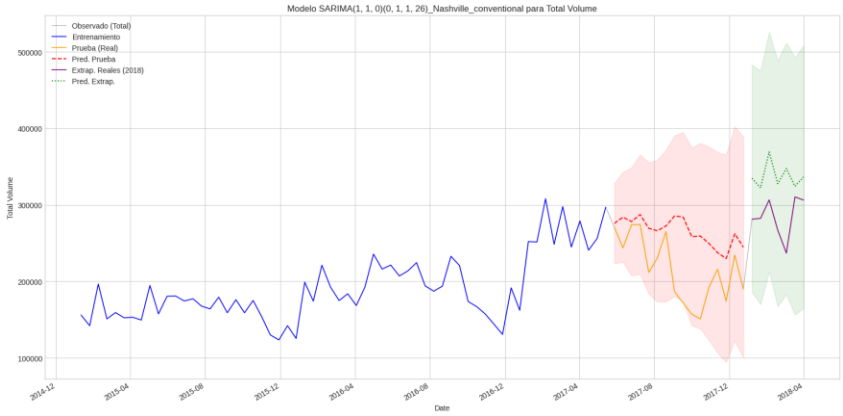
ROANOKE - organic



PHILADELPHIA- organic



NASHVILLE - conventional



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

8. 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.