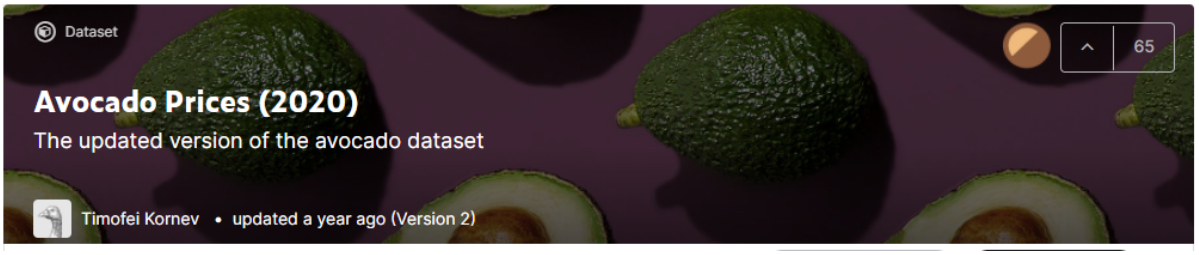
Proyecto: **Análisis de datos y Predicción de precios de Aguacate en USA**

En este proyecto se analizará la información obtenida del Dataset de Kaggle *"Avocado Prices (2020)"* con la finalidad de dar respuesta a ciertas preguntas planteadas sobre la información como tendencias de precios, preferencias de los consumidores y volúmenes de distribución y venta del producto en USA. Se obtendrán hallazgos conforme al análisis y visualización efectiva de los datos y se implementará un modelo de predicción para predecir el precio del aguacate en fechas futuras.

Obtención y carga de datos

En este proyecto vas a trabajar con un dataset de precios y volúmenes de venta de aguacate de los años 2015, 2016, 2017, 2018, 2019 y 2020 publicado en Kaggle: https://www.kaggle.com/timmate/avocado-prices-2020



Esta es una versión actualizada del conjunto de datos de aguacate compilado originalmente a partir de los datos de Hass Avocado Board (o HAB, para abreviar) y publicado en Kaggle por Justin Kiggins en 2018. El conjunto de datos presenta datos históricos sobre los precios del aguacate y el volumen de ventas en múltiples ciudades, estados , y regiones de los EE.UU.

Descripción de las columnas de acuerdo a la información extraída de la fuente:

- date: Fecha de la observación
- average_price: precio promedio por aguacate
- total_volume: volúmen total de aguacates vendido
- 4046: volúmen total de guacates PLU 4046 vendidos
- 4225: volúmen total de guacates PLU 4225 vendidos
- 4770: volúmen total de guacates PLU 4770 vendidos
- total_bags: Total de bolsas vendidas
- small_bags: Total de bolsas pequeñas vendidas
- large_bags: Total de bolsas grandes vendidas
- xlarge_bags: Total de bolsas extra grandes vendidas
- type: tipo de aguacate (convencional u orgánico)
- year: año
- geography: ciudad o región de observación

La carga de datos se realiza de la siguiente manera:

```
#Carga de dataset
data = pd.read_csv('avocado-2020.csv', sep =',')
data
```

Y posteriormente vemos una tabla como la siguiente:

#Visualización de las primeras líneas de datos que contiene el archivo
data.head()

	date	average_price	total_volume	4046	4225	4770	total_bags	small_bags	large_bags	xlarge_bags	type	year	geography
0	2015-01-04	1.22	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.0	conventional	2015	Albany
1	2015-01-04	1.79	1373.95	57.42	153.88	0.00	1162.65	1162.65	0.00	0.0	organic	2015	Albany
2	2015-01-04	1.00	435021.49	364302.39	23821.16	82.15	46815.79	16707.15	30108.64	0.0	conventional	2015	Atlanta
3	2015-01-04	1.76	3846.69	1500.15	938.35	0.00	1408.19	1071.35	336.84	0.0	organic	2015	Atlanta
4	2015-01-04	1.08	788025.06	53987.31	552906.04	39995.03	141136.68	137146.07	3990.61	0.0	conventional	2015	Baltimore/Washington

Exploración de los datos

El conjunto de datos contiene en total 33,405 filas y 13 columnas, las cuales no contienen datos nulos ni datos duplicados, se observa también en las siguientes líneas de código el tipo de dato de cada columna:

```
#Ver el tamaño del dataset
data.shape
```

```
(33045, 13)
```

```
# Primer acercamiento  a los datos tipos de dato
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33045 entries, 0 to 33044
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   date           33045 non-null  object
 1   average_price  33045 non-null  float64
 2   total_volume   33045 non-null  float64
 3   4046           33045 non-null  float64
 4   4225           33045 non-null  float64
 5   4770           33045 non-null  float64
 6   total_bags     33045 non-null  float64
 7   small_bags     33045 non-null  float64
 8   large_bags     33045 non-null  float64
 9   xlarge_bags    33045 non-null  float64
 10  type           33045 non-null  object
 11  year           33045 non-null  int64
 12  geography      33045 non-null  object
dtypes: float64(9), int64(1), object(3)
memory usage: 3.3+ MB
```

```
#Búsqueda de valores nulos
data.isnull().sum()
```

```
date            0
average_price   0
total_volume    0
4046            0
4225            0
4770            0
total_bags      0
small_bags      0
large_bags      0
xlarge_bags     0
type            0
year            0
geography       0
dtype: int64
```

```
#Busqueda de valores duplicados
data.duplicated().sum()
```

```
0
```

Se obtienen los valores únicos de las columnas de año, tipo de aguacate y ubicaciones:

```
#Ver los datos unicos de la columna año
print(data['year'].unique())
```

```
[2015 2016 2017 2018 2019 2020]
```

```
#El archivo contiene dos tipos de producto Convencional y Orgánico
print(data['type'].unique())
```

```
['conventional' 'organic']
```

```
#Ver todas las ubicaciones que contiene el archivo
print(data['geography'].unique())
```

```
['Albany' 'Atlanta' 'Baltimore/Washington' 'Boise' 'Boston'
 'Buffalo/Rochester' 'California' 'Charlotte' 'Chicago'
 'Cincinnati/Dayton' 'Columbus' 'Dallas/Ft. Worth' 'Denver' 'Detroit'
 'Grand Rapids' 'Great Lakes' 'Harrisburg/Scranton' 'Hartford/Springfield'
 'Houston' 'Indianapolis' 'Jacksonville' 'Las Vegas' 'Los Angeles'
 'Louisville' 'Miami/Ft. Lauderdale' 'Midsouth' 'Nashville'
 'New Orleans/Mobile' 'New York' 'Northeast' 'Northern New England'
 'Orlando' 'Philadelphia' 'Phoenix/Tucson' 'Pittsburgh' 'Plains'
 'Portland' 'Raleigh/Greensboro' 'Richmond/Norfolk' 'Roanoke' 'Sacramento'
 'San Diego' 'San Francisco' 'Seattle' 'South Carolina' 'South Central'
 'Southeast' 'Spokane' 'St. Louis' 'Syracuse' 'Tampa' 'Total U.S.' 'West'
 'West Tex/New Mexico']
```

```
#Ver todas la cantidad de ubicaciones diferentes que contiene el archivo
data['geography'].nunique()
```

```
54
```

Comprobamos que en la columna año se encuentran únicamente los años de 2015 a 2020, en la columna de tipo de aguacate observamos dos variantes: convencional u orgánico. Mientras que la columna de ubicación contiene en total 52 valores únicos.

Podemos observar que en la columna de ubicación (geography) contiene información de los estados de USA y además un resumen por región que contiene los subtotales: *total us, west, midsouth, northeast, south central y southeast.* Por lo cual se realizará la separación de estos datos, tendremos dos conjuntos de datos, uno por estados y otro por regiones para realizar nuestras visualizaciones.

```
#Se obtiene la lista de regiones a excluir
regiones = ['West','Midsouth', 'Southeast','South Central','Northeast']

#Se realiza el filtro del dataset por regiones

df_regiones = data[np.isin(data['geography'], regiones)]
df_estados = data[np.isin(data['geography'], regiones, invert=True)]

#Se elimina también la columna "total U.S"
df_estados = df_estados.loc[df_estados['geography'] != 'Total U.S.']
```

```
#Ver los valores únicos del conjunto de datos por region
print(df_regiones['geography'].unique())
```

```
['Midsouth' 'Northeast' 'South Central' 'Southeast' 'West']
```

```
#Ver los valores únicos del conjunto de datos por estado
print(df_estados['geography'].unique())
```

```
['Albany' 'Atlanta' 'Baltimore/Washington' 'Boise' 'Boston'
 'Buffalo/Rochester' 'California' 'Charlotte' 'Chicago'
 'Cincinnati/Dayton' 'Columbus' 'Dallas/Ft. Worth' 'Denver' 'Detroit'
 'Grand Rapids' 'Great Lakes' 'Harrisburg/Scranton' 'Hartford/Springfield'
 'Houston' 'Indianapolis' 'Jacksonville' 'Las Vegas' 'Los Angeles'
 'Louisville' 'Miami/Ft. Lauderdale' 'Nashville' 'New Orleans/Mobile'
 'New York' 'Northern New England' 'Orlando' 'Philadelphia'
 'Phoenix/Tucson' 'Pittsburgh' 'Plains' 'Portland' 'Raleigh/Greensboro'
 'Richmond/Norfolk' 'Roanoke' 'Sacramento' 'San Diego' 'San Francisco'
 'Seattle' 'South Carolina' 'Spokane' 'St. Louis' 'Syracuse' 'Tampa'
 'West Tex/New Mexico']
```

```
#Ver el tamaño del conjunto de datos por región
df_regiones.shape
```

```
(3060, 13)
```

```
#Ver el tamaño del conjunto de datos por estados
df_estados.shape
```

```
(29373, 13)
```

```
df_estados['geography'].nunique()
```

```
48
```

El tamaño del conjunto de datos por estados es de 29,373 filas y 13 columnas, mientras que los datos por regiones es de 3,060 filas por 13 columnas, obteniendo 48 ubicaciones diferentes.

La información a la que se refieren las columnas “4064”, “4225” y “4770” es una etiqueta contiene el código PLU (Price Lookup, por sus siglas en inglés) y los supermercados utilizan facilitar el control del inventario de las frutas y verduras. La encargada de asignar estos códigos es la Federación Internacional para los Estándares de Productos (IFPS).
Infromación obtenida de: https://loveonetoday.com/how-to/identify-hass-avocados/

- Small/Medium Hass Avocado (~3-5oz) | #4046 Avocado
- Large Hass Avocado (~8-10oz) | #4225 Avocado
- Extra Large Hass Avocado (~10-15oz) | #4770 Avocado

Plantear preguntas a resolver

- ¿Existe una tendencia a la alza o baja en el precio de los aguacates a lo largo de los años?
- ¿Qué factores podrían influir en la alza o baja de precios de aguacates?
- ¿Durante qué periodo se vende la mayor cantidad de aguacates?
- ¿En qué región y/o estado se vende la mayor cantidad de aguacates?
- ¿El tipo de aguacate (convencional / orgánico) juega un papel importante en el precio o volúmen de venta del aguacate?

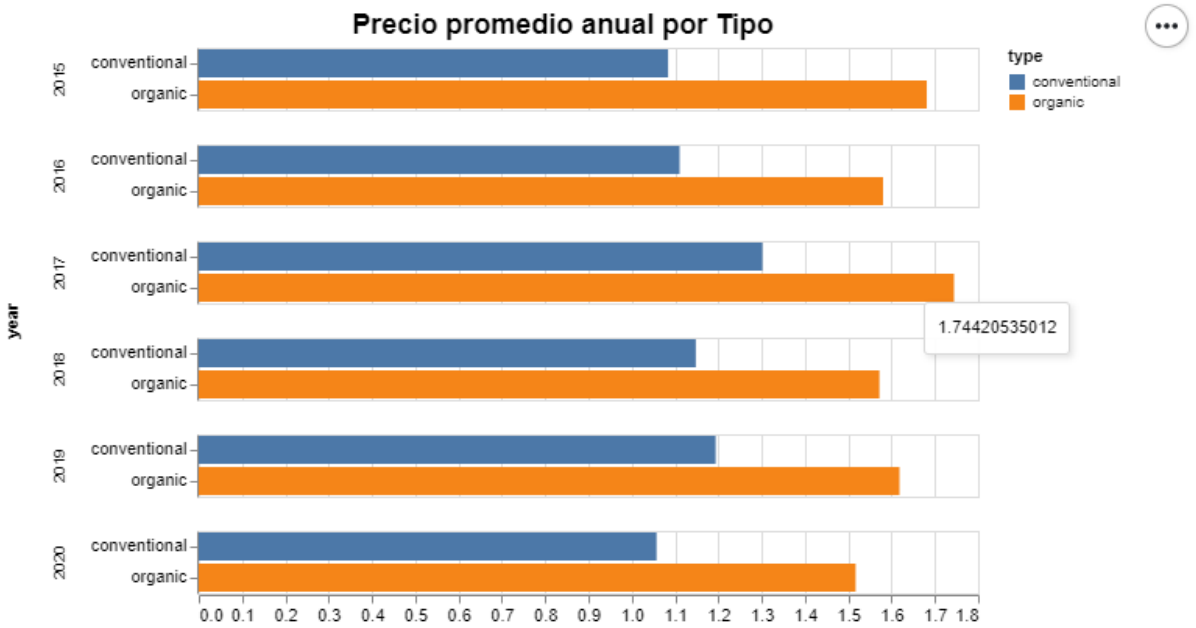
Visualizaciones de los datos con Altair

El código para las siguientes gráficas se puede encontrar en el notebook del repositorio en Github: https://github.com/amorgado13/Prediccion-precios-Aguacate

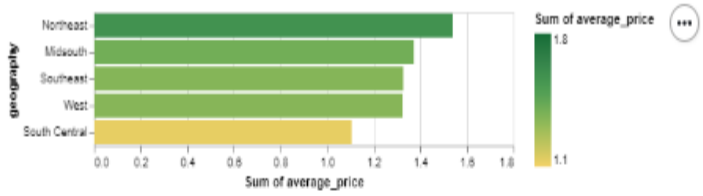
En la siguiente gráfica observamos que el volúmen de ventas por año ha incrementado de 2015 a 2020, en 2016 y 2017 se mantuvo igual, mientras que el mayor salto fue de 2019 a 2020.

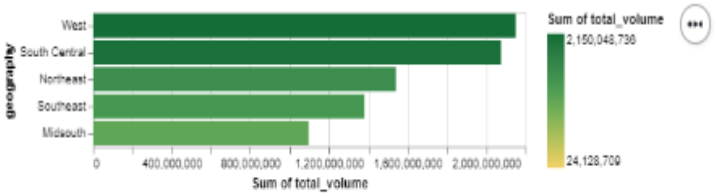
El precio promedio por año fue variando con los años, siendo el más alto en 2017 con $1.5 USD y los más bajos en 2016 y 2020 con $1.3 USD.

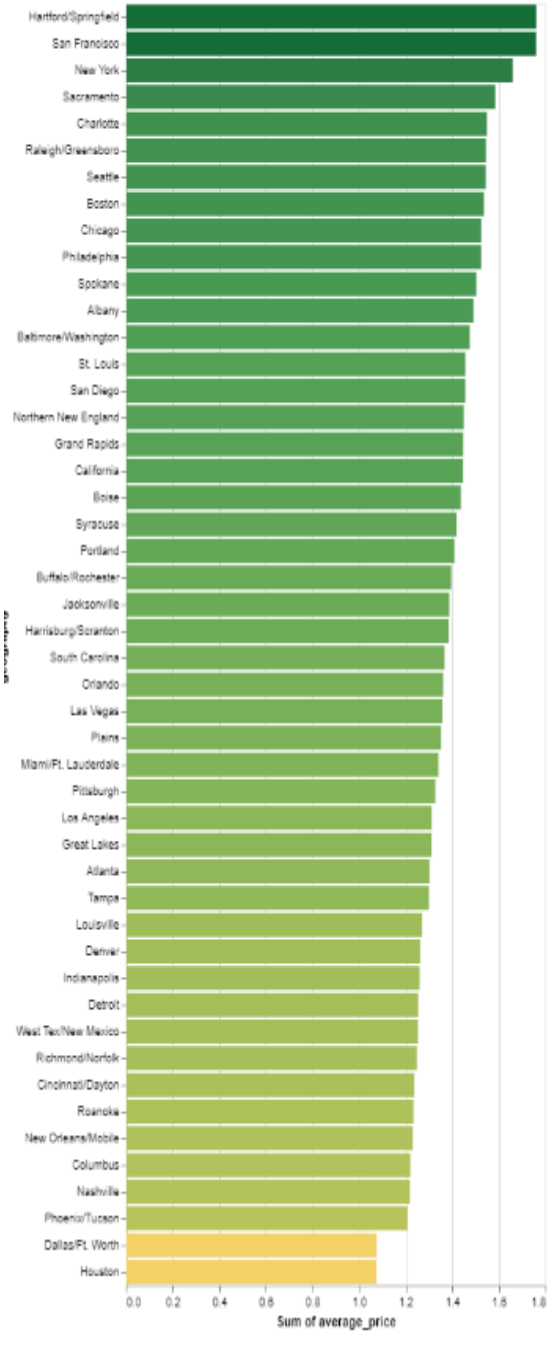
Sin embargo, en la gráfica anterior se está considerando la totalidad de los datos sin importar el tipo o la ubicación geográfica, por lo que en la siguiente gráfica se muestra el precio por tipo de aguacate (convencional u orgánico) y podemos observar que el precio del aguacate orgánico es mayor que el convencional para todos los años y que en 2017 en realidad el promedio de precio más alto es de $1.7 USD para el aguacate orgánico y $1.3 USD para el aguacate convencional.

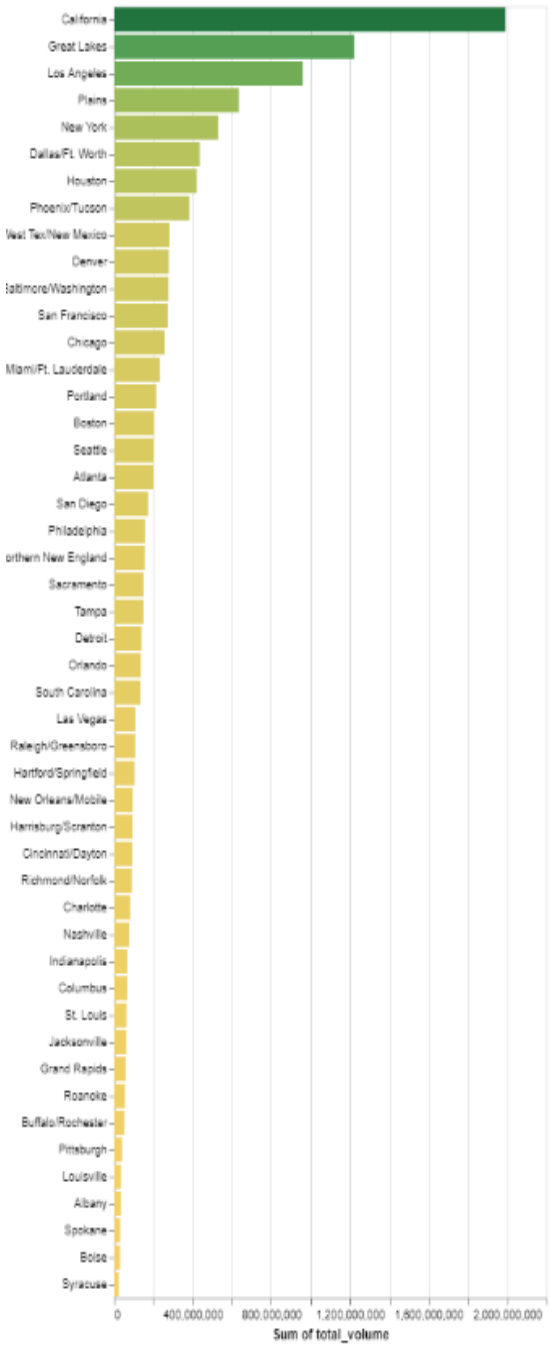


Posteriormente evaluamos el volúmen de ventas por ubicación:







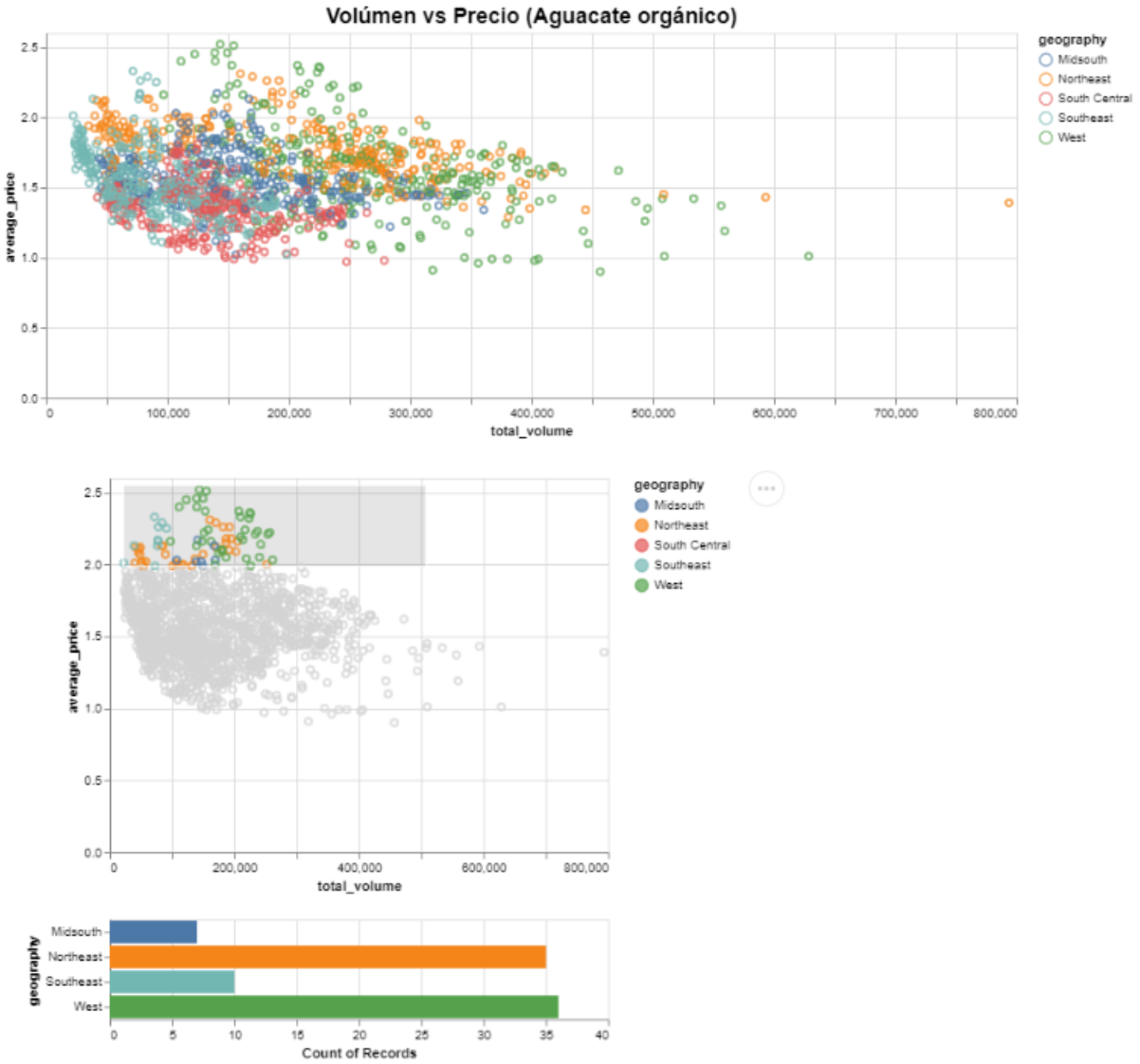


En estas gráficas observamos que las regiones con mayor volúmen de ventas son West y South Central, las cuales son también las regiones con el menor promedio de precio, por lo que en este sentido el precio si podría estar influyendo en el volúmen de ventas realizadas. También observamos que el mayor volúmen de ventas se encuentra en California y Great Lakes, pero estos no son los

estados con menor promedio de precio, se encuentran en los lugares de el centro en cuanto a precio. Los estados con menor precio son Dallas y Houston, quienes se encuentran en el lugar 6° y 7° de mayor cantidad de volúmen de ventas.

Con las gráficas que se muestran a continuación podemos dar una mejor conclusión en cuanto a la influencia del precio en el volúmen de venta:

En la gráfica anterior del aguacate convencional observamos que existe una tendencia de entre menor precio, es mayor el volúmen de venta, además de que se puede ver a que las regiones se encuentran bien segmentadas, South Central, South East y West tienen las observaciones con el menor precio en el rango de $0.6 a $0.8 USD.

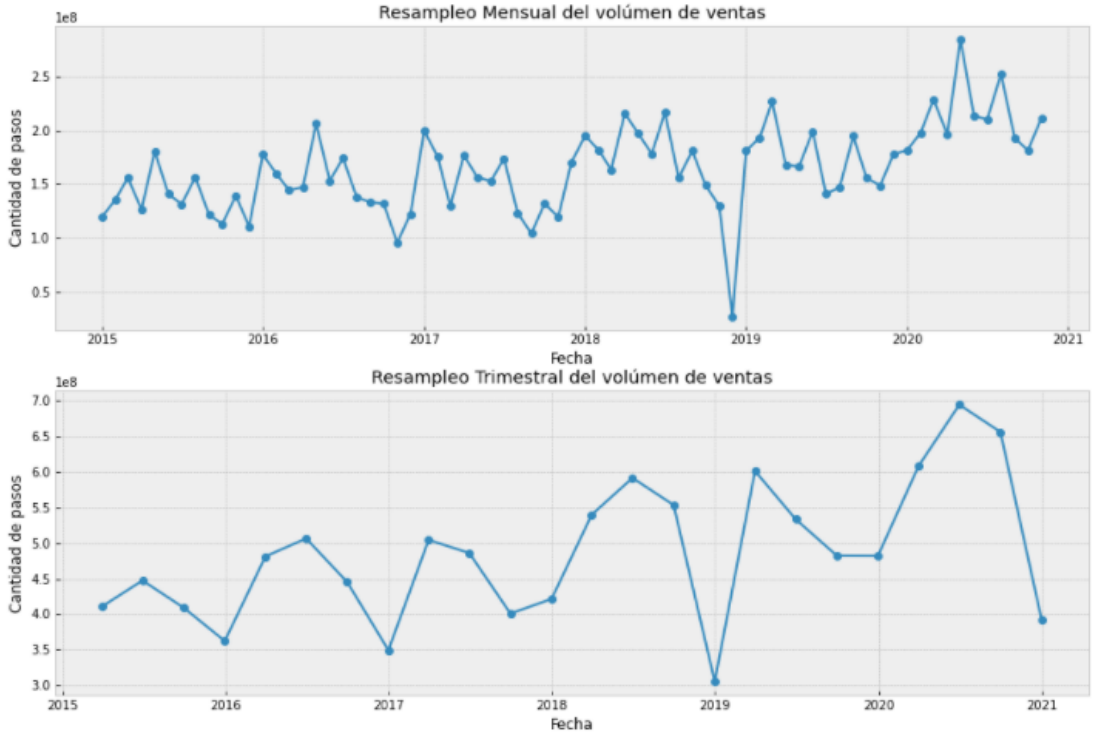


El gráfico de dispersión del aguacate orgánico existe una tendencia muy ligera, se encuentran más agrupados los puntos al inicio de la gráfica, ya que el volúmen de venta de este tipo de alimento es menor y el precio oscila entre $1 y $2.5 USD promedio, de igual manera West y North west se encuentran en el rango de mayor precio de $2 a $2.5 USD.

Tendencias

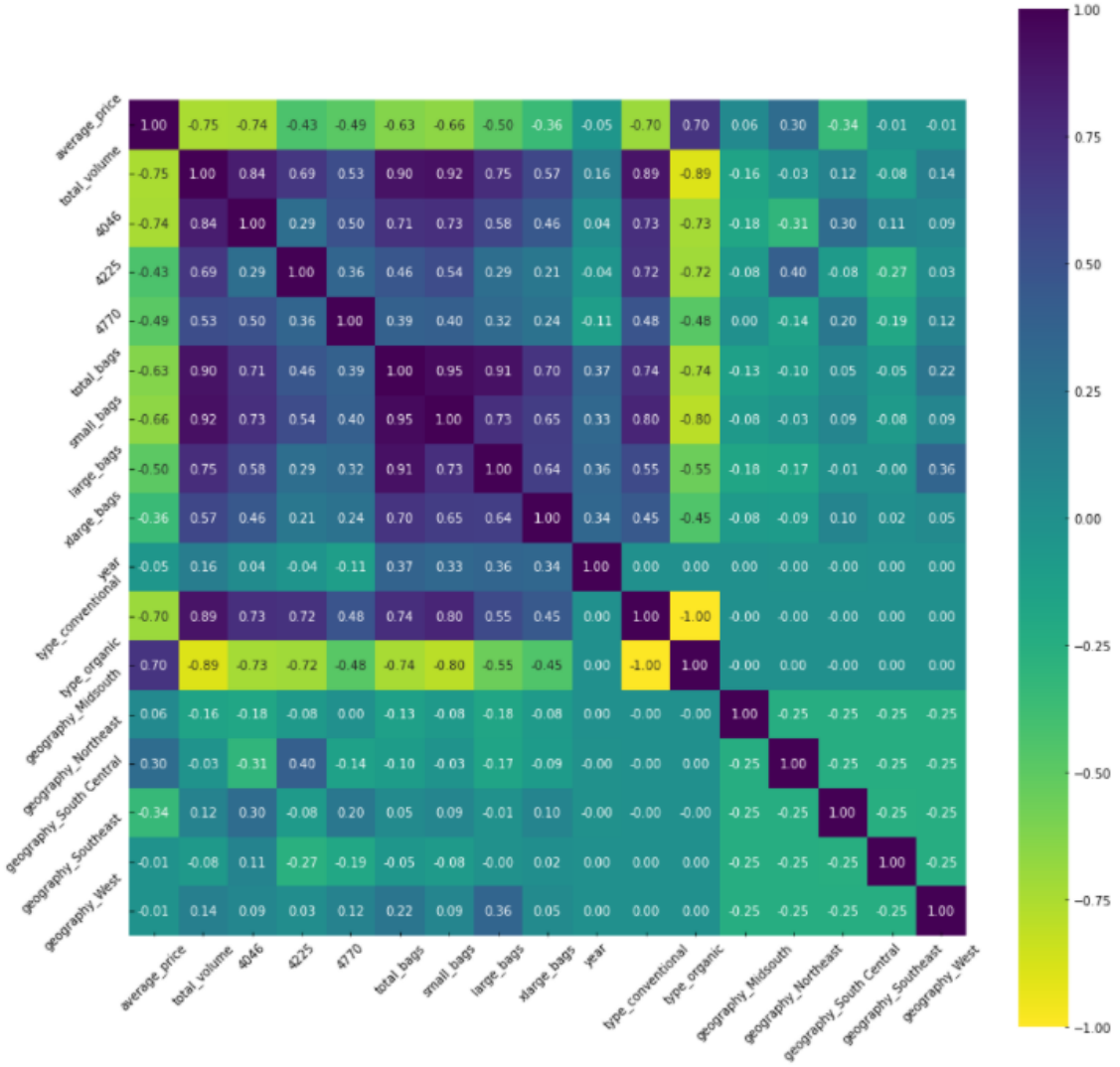
Se realiza un resampleo de fechas para evaluar las tendencias de forma mensual y trimestral.

Se observa una tendencia en el comportamiento del precio en la que disminuye en los últimos 3 a 4 meses en la gráfica mensual, excepto por 2018 en el que hay dos picos en la segunda mitad del año.



En el volúmen de ventas observamos un comportamiento variable en las ventas por mes con altas y bajas en el año, mientras que en el resampleo trimestral se observa que en el primer trimestre se eleva la venta los dos primeros trimestres del año y disminuye el último periodo del año.

Correlaciones



Como podemos apreciar en el gráfico anterior las variables con mayor correlación con el precio son, el volúmen y el código de producto "4046". Las variables de tipo Convencional u orgánico están correlacionadas el 70% una de forma negativa y otra positiva, es decir que el tipo orgánico incrementa el precio mientras que el convencional lo disminuye. Las variables geográficas por región no tienen una correlación importante.

Modelo de predicción

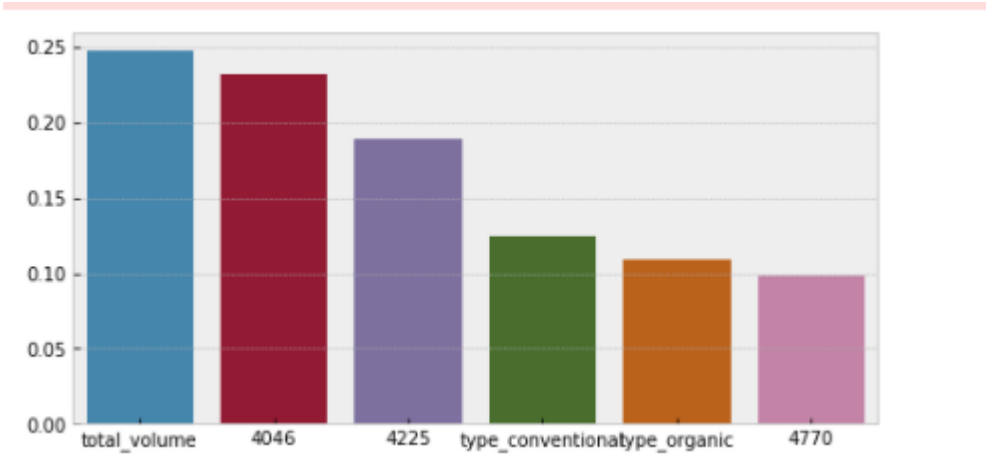
Se implementó el modelo de predicción de Random Forest, en el cual como paso principal se realizó un escalado de datos para las columnas que son más representativas del conjunto de datos de acuerdo a la gráfica anterior (volumen, tipo de aguacate, código PLU). Se consideraron los datos por estado ya que son los que contienen más observaciones.

El conjunto de datos final utilizado para el modelo fue el siguiente:

	total_volume	4046	4225	4770	type_conventional	type_organic
0	0.003602	0.000588	0.006903	0.000063	1	0
1	0.000114	0.000012	0.000038	0.000000	0	1
2	0.038406	0.075989	0.005813	0.000103	1	0
3	0.000332	0.000313	0.000229	0.000000	0	1
4	0.069578	0.011261	0.134934	0.050324	1	0
...	...	...	...	...	...	...
33036	0.000274	0.000035	0.000022	0.000000	0	1
33037	0.049658	0.048073	0.003731	0.000056	1	0
33038	0.001042	0.000021	0.000000	0.000000	0	1
33043	0.072650	0.048953	0.019574	0.013267	1	0
33044	0.002121	0.000258	0.000151	0.001969	0	1

29373 rows × 6 columns

De acuerdo a la gráfica, el atributo de mayor importancia para predecir el precio es el volúmen total, seguido de código PLU 4046, 4225 y el tipo de aguacate.



Como resultado este modelo nos arroja una evaluación de 93% de efectividad en el conjunto de entrenamiento y 66% en el conjunto de prueba, es un resultado que se puede mejorar ya sea implementando otro tipo de modelos o evaluando si requerimos modificar la muestra o agregar más atributos.

```
#evaluar modelo con r2
print(metrics.r2_score(y_train, y_train_pred))
print(metrics.r2_score(y_test, y_test_pred))
```

```
0.937121326422565
0.6652302736172591
```

Conclusiones y Hallazgos

Como conclusión este conjunto de datos nos permitió explorar y analizar la venta de aguacates obteniendo respuestas a las preguntas planteadas inicialmente, existe una tendencia a la alza del precio de los aguacates a lo largo de los años, esto se puede deber a que también hay mayor cantidad de demanda (volúmen vendido) o también a la inflación de precios. Los factores que influyen en la alza o baja son la cantidad vendida y a que tipo de aguacate pertenece, como podemos ver los convencionales son los más demandados y también los más baratos, mientras que los orgánicos se venden en menor volúmen y tienen un precio más alto. EL periodo del año en el que se venden a mayor cantidad de aguacates es al inicio del año, los dos primeros trimestres van a la alza y en el último comienza a bajar la demanda. La región en que se vende la mayor cantidad de aguacates es en California y la zona Este del país de USA, por lo cual desde el punto de vista del productor/vendedor de aguacates sería un punto de venta a considerar y desde el punto de vista del comprador podemos observamos que en Houston y Dallas es donde se encuentra el menor precio.

Este análisis es un buen primer acercamiento a la evaluación del negocio de venta de aguacates en USA, ya sea como productor o consumidor da un panorama de posibles oportunidades y áreas de interés para iniciar o incrementar la rentabilidad de un negocio.