

UNIVERSIDAD NACIONAL AUTÓNOMA DE  
MÉXICO  
FES ACATLÁN



# ANÁLISIS DE VENTA DE AGUACATE HASS EN EUA

PRECIADO RIVERA MARÍA FERNANDA

MÓDULO 1. INTRODUCCIÓN, MANIPULACIÓN, EXPLORACIÓN Y  
VISUALIZACIÓN DE DATOS



JUNIO 2021

# Índice general

<b>1. Introducción, Manipulación, Exploración y Visualización de Datos</b>	<b>1</b>
1.1. Objetivo/Motivación . . . . .	1
1.2. Data Set . . . . .	3
1.3. Contenido y Clasificación de variables . . . . .	4
1.4. Calidad de Datos . . . . .	5
1.4.1. Completitud . . . . .	5
1.4.2. Conformidad . . . . .	5
1.4.3. Duplicación . . . . .	5
1.4.4. Consistencia . . . . .	6
1.5. Tratamiento de Outliers . . . . .	7
1.6. Ingeniería de Variables . . . . .	9
1.7. Análisis exploratorio . . . . .	15
1.8. Reducción de variables:	24
1.8.1. Filtro de alta correlación . . . . .	24
1.8.2. Correlación con objetivo . . . . .	25
<b>A. Referencias</b>	<b>26</b>

# Capítulo 1

## Introducción, Manipulación, Exploración y Visualización de Datos

### 1.1. Objetivo/Motivación

Al aguacate se le ha denominado como **oro verde**, esta fruta es uno de los principales productos de exportación de México y **Estados Unidos de América** destaca como el principal consumidor del aguacate mexicano.

El impulso en el consumo de esta fruta se atribuye a factores como la creciente demanda de la comunidad latina en Estados Unidos de América por este producto, una mayor difusión de los beneficios a la salud por el consumo de esta fruta, lo que ha llevado a que diversas cadenas de comida rápida la incluyan en sus menús.

El interés de seleccionar este tema viene de la curiosidad de cómo otra cultura ha adaptado un producto tan “mexicano” que es simplemente una fruta a su alimentación no solo en la presentación de “guacamole” sino que incluso se ha adaptado a platillos dulces, salados y hasta bebidas.

## 2 Introducción, Manipulación, Exploración y Visualización de Datos



Dada la motivación de explorar los niveles de consumo de esta fruta en EUA, el conjunto de datos que se utilizó en el análisis que se muestra a continuación se obtuvo de HASS AVOCADO BOARD que es “la única organización de aguacates que equipa a toda la industria global para el éxito al recolectar, enfocar y distribuir inversiones para mantener y expandir la demanda de aguacates en los Estados Unidos.”



Como objetivo principal es conocer la distribución de aguacates por zonas/mercados de EUA así como la exploración de la base que podría ayudar a estimar el volumen o precio de ciertos tipos de aguacates así como encontrar una posible relación de estos datos con nuestro país.

## 1.2. Data Set

El conjunto de datos se obtuvo de la página de **HASS AVOCADO BOARD** y la única manera de descargar los datos era por año, por lo que se hicieron 6 descargas, es decir desde el año 2015 al año 2020. Aunque los datos venían de la misma fuente, por calidad de datos aprendimos que había que revisar la **conformidad**, es decir que tuviera un formato estándar y legible por lo que se exploró el número de filas y columnas de cada año, de esta manera se descartó utilizar la información del año 2015 ya que incluía solo 3 meses y el resto de los años (2016 - 2020) eran años con información de todos los meses y al obtener la primer fecha de cada uno comenzaba en los primeros días de enero y la última semana de información coincidía con fechas de diciembre. Además que para estos años el número de columnas era el mismo: **12 variables** y el número de registros en promedio era de **5365 filas**. Se revisó el nombre de columnas año a año y se revisó el tipo de datos que incluían y con eso se detectó que para los últimos 3 años (2018 a 2020) se podían unir las tablas ya que nombre de columnas y tipo de datos coincidian perfectamente. Sin embargo para el año 2016 y 2017 el nombre de las columnas variaba ligeramente en redacción por lo que se renombraron como las de los últimos años y otra de las diferencias que se encontró es que las variables de texto también se conformaban de guiones y símbolos pero se validó que fueran las mismas categorías que en los años 2018 a 2020, es decir, el tipo de datos y la información era compatible entre años pero el formato de 2016 y 2017 variaba usando caracteres especiales por lo que se juntaron las tablas de los 5 años sabiendo que posteriormente sería necesaria la limpieza de texto.



## 4 Introducción, Manipulación, Exploración y Visualización de Datos

### 1.3. Contenido y Clasificación de variables

Después de juntas las bases de los años 2016 a 2020 la tabla inicial quedó con una dimensión de (**27430 filas , 12 columnas**), a continuación se muestra un corte de la tabla:

Geography	Type	ASP Current Year	Total Bulk and Bags Units	4046 Units	4225 Units	4770 Units	TotalBagged Units	SmlBagged Units	LrgBagged Units	X_LrgBagged Units	Week_Ending
Albany	Conventional	1.47	113514.42	2622.70	101135.53	20.25	9735.94	5556.98	4178.96	0.00	2018-01-01
Los Angeles	Organic	1.74	78872.75	15369.81	20722.83	0.00	42780.11	42693.41	86.70	0.00	2018-01-01
Las Vegas	Organic	1.55	9413.05	1479.66	1733.93	0.00	6199.46	6190.31	9.15	0.00	2018-01-01
Jacksonville	Organic	1.65	2023.95	44.23	496.69	5.06	1477.97	1431.11	46.86	0.00	2018-01-01
Indianapolis	Organic	1.45	9437.04	200.76	1231.56	0.00	8004.72	5566.32	2438.40	0.00	2018-01-01
...	...	...	...	...	...	...	...	...	...	...	...
BuffaloRochester	organic	1.15	10568.46	176.50	82.85	0.00	10309.11	5004.55	5304.56	0.00	2017-12-31
Boston	organic	1.86	25275.08	9.19	2796.62	0.00	22469.27	20109.15	2360.12	0.00	2017-12-31
Boise	organic	1.72	2017.17	78.28	646.17	0.00	1292.72	213.18	1079.54	0.00	2017-12-31
GreatLakes	conventional	0.92	4398887.92	1092011.22	1740608.60	364049.45	1202218.65	927628.08	231172.74	43417.83	2017-12-31
Chicago	organic	1.79	26040.03	142.00	16241.00	0.00	9657.03	9650.37	6.66	0.00	2017-12-31

La descripción de las 12 variables se muestra a continuación además de su clasificación:

Variable	Tipo de variable	Descripción
Geography	Categórica	Ciudad o region (mercados = reportes regionales de HASS AVOCADO BOARD)
Type	Categórica	Tipo de aguacate (Convencional u Orgánico)
ASP Current Year	Numérica (continua)	Average Selling Price (Dolares /Unidades)
Total Bulk and Bags Units	Numérica (continua)	Número total de aguacates vendidos
4046 Units	Numérica (continua)	Número total de "Small/Medium Hass Avocado(~3-5oz avocado)" vendidos
4225 Units	Numérica (continua)	Número total de "Large Hass Avocado (~8-10oz avocado)" vendidos
4770 Units	Numérica (continua)	Número total de "Extra Large Hass Avocado (~10-15oz avocados)" vendidos
TotalBagged Units	Numérica (continua)	Total de bolsas de aguacate Hass vendidas
SmlBagged Units	Numérica (continua)	Total de bolsas pequeñas de aguacate Hass vendidas
LrgBagged Units	Numérica (continua)	Total de bolsas grandes de aguacate Hass vendidas
X_LrgBagged Units	Numérica (continua)	Total de bolsas extra grandes de aguacate Hass vendidas
Week_Ending	Fecha	Semana de ventas/producción

Posteriormente se les asignó el etiquetado de variables como se aplicó a lo largo del módulo.

## 1.4. Calidad de Datos

### 1.4.1. Completitud

Se revisaron los valores ausentes de todas las variables iniciales de la tabla y se encontró que cada una de ellas tenía 100 % de completitud:

	nombre_columna	total_nulos	completitud
0	v_Geography	0	100.0
1	v_Type	0	100.0
2	c_ASP Current Year	0	100.0
3	c_Total Bulk and Bags Units	0	100.0
4	c_4046 Units	0	100.0
5	c_4225 Units	0	100.0
6	c_4770 Units	0	100.0
7	c_TotalBagged Units	0	100.0
8	c_SmlBagged Units	0	100.0
9	c_LrgBagged Units	0	100.0
10	c_X-LrgBagged Units	0	100.0
11	d_Week_Ending	0	100.0

### 1.4.2. Conformidad

Esta revisión de calidad (formato estándar y legible) se realizó al concatenar las bases de cada uno de los años (2016 a 2020), las variables que hay que limpiar son Type y Geografía porque no tienen el mismo formato o codificación, pero esto limpieza se realizará más adelante.

### 1.4.3. Duplicación

La tabla tiene duplicados de fecha pero esto es lógico ya que una fecha se repite para todos los mercados que tiene la base ya que la información viene desglosada a nivel geografía por semana de producción. En cuanto a duplicados generales no se encontró alguno como era de esperarse por la calidad de data que promete la página de [HASS AVOCADO BOARD](#)

## 6 Introducción, Manipulación, Exploración y Visualización de Datos

### 1.4.4. Consistencia

- Ninguna de las variables numéricas, por ser volumen y precio deberían contener valores negativos lo cual se revisó obteniendo el mínimo de cada una de dichas variables:

c_ASP Current Year	0.44
c_Total Bulk and Bags Units	253.45
c_4046 Units	0.00
c_4225 Units	0.00
c_4770 Units	0.00
c_TotalBagged Units	7.02
c_SmlBagged Units	0.00
c_LrgBagged Units	0.00
c_X-LrgBagged Units	0.00

- Se revisó que la variable c\_TotalBagged Units fuera la suma de los 3 tipos de bolsas lo cual se cumple:



=

$$\text{c\_SmlBagged Units} + \text{c\_LrgBagged Units} + \text{c\_X-LrgBagged Units}$$

- Se validó que el total del volumen c\_Total Bulk and Bags Units fuera la suma de los totales de bolsa más los aguacates a granel por tamaño:


$$= \text{Bag Icon} + 4046 + 4525 + 4770$$


Lo cual no se cumplió para el 4% de los registros, en estos casos en lugar de eliminar estos registros se recalcularó con la suma del total de bolsas y el volumen de cada uno de los aguacates por tamaño.

## 1.5. Tratamiento de Outliers

Los valores outliers son aquellos que tienen un comportamiento subyacente, diferente al del resto de los datos. En esto caso sabíamos de posibles outliers de tipo contextual ya que por investigación en varias noticias se encontró que el año 2017 había sido un año “atípico” debido a que la producción no fue tan buena como en otros años y la demanda aumentó.

Los algoritmos de aprendizaje automático son sensibles al rango y distribución de atributos, por esto se aplicaron 2 métodos matemáticos que ayudan a la detección de outliers, estos métodos o funciones matemáticas se explican a continuación:

1. IQR.- Esta opción se define como rango intercuartil (IQR) la cual se basa en la diferencia entre el tercer y el primer cuartil de una distribución. Dicha diferencia es una medida de dispersión estadística. A partir de este rango se define una cota inferior y otra superior:

$$\text{INF} = Q1 - 1,5(\text{IQR})$$

$$\text{SUP} = Q3 + 1,5(\text{IQR})$$

Con las cotas anteriores se determinan como valores outliers aquellos menores a INF y mayores a SUP.

2. Percentiles.- Generalmente tomamos el percentil 5 % y el 95 % como cotas inferior y superior, respectivamente. Todos aquellos datos que queden por fuera de estas cotas serán catalogados como outliers.

Para este análisis se aplicaron los 2 métodos anteriores y aquellos datos que resultaron identificados como outliers en la intersección de las dos funciones matemáticas se eliminan sin embargo por el contexto antes mencionado se revisó la proporción de años de los registros que se clasificaron como outliers con el fin de detectar que no estuvieran concentrados en su mayoría en el año 2017:

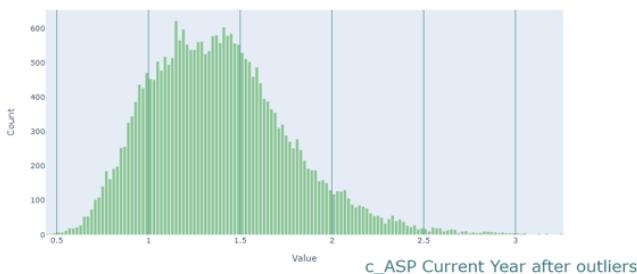
## 8 Introducción, Manipulación, Exploración y Visualización de Datos



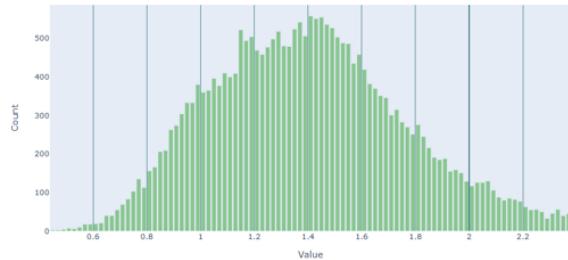
Como se observó que los outliers no estaba concentrados en el año 2017 se eliminaron los 3,230 registros que representaban aproximadamente el 12 % de las filas iniciales de la tabla.

### ■ c\_ASP Current Year

c\_ASP Current Year



c\_ASP Current Year after outliers



No se muestran todas los histogramas generados (antes y después) de las 9 variables numéricas a las que se le eliminaron outliers, estas pueden encontrarse en el código.

## 1.6. Ingeniería de Variables

### ■ v\_Geography

Como se mencionó antes, se detectó que la variables “Geography” necesitaría tratamiento de limpieza de texto ya que algunas categorías se encontraban con caracteres especiales o escritas ligeramente diferente por lo que se tenían 78 categorías/mercados y no las 54 categorías en total que se encuentran en la página de HASS AVOCADO BOARD:

	region	Frecuencia
3	Albany	508
13	Atlanta	501
37	Baltimore/Washington	293
50	BaltimoreWashington	210
14	Boise	493
...	...	...
75	Total U.S.	106
77	TotalUS	103
48	West	247
38	West Tex/New Mexico	286
65	WestTexNewMexico	198

Una vez se aplicó la limpieza de texto se confirmó que las categorías fueran 54 que son los mercados que maneja la fuente oficial:

```
df["v_Geography"].unique()
array(['pittsburgh', 'louisville', 'greatlakes', 'tampa', 'syracuse',
       'philadelphia', 'houston', 'stlouis', 'lasvegas', 'spokane',
       'atlanta', 'boise', 'plains', 'denver', 'cincinnatidayton',
       'sanfrancisco', 'columbus', 'chicago', 'phoenixtucson', 'albany',
       'orlando', 'westtexnewmexico', 'southcentral', 'jacksonville',
       'indianapolis', 'southcarolina', 'dallasftworth', 'southeast',
       'grandrapids', 'buffalorochester', 'baltimorewashington',
       'harrisburgscranton', 'nashville', 'northernnnewengland', 'west',
       'hartfordspringfield', 'miamiflaurerdale', 'totalus',
       'neworleansmobile', 'sandiego', 'losangeles', 'roanoke',
       'raleighgreensboro', 'portland', 'boston', 'seattle', 'california',
       'detroit', 'richmondnorfolk', 'northeast', 'midsouth', 'newyork',
       'sacramento', 'charlotte'], dtype=object)
```

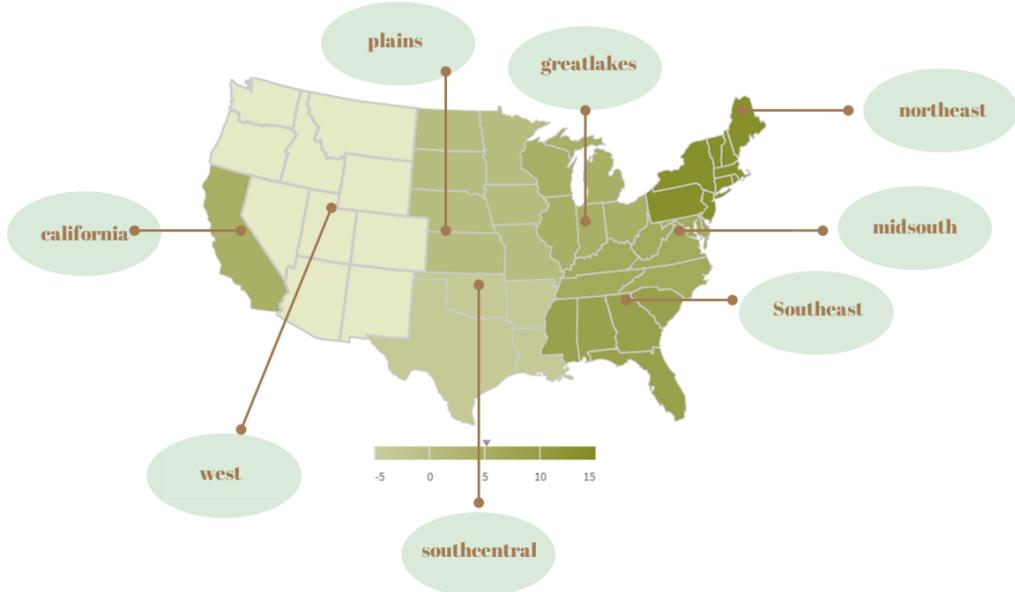
## 10Introducción, Manipulación, Exploración y Visualización de Datos

Inicialmente con esta información de geografía se pensó buscar coordenadas para poder generar mapas, pero al analizarla a detalle encontramos que en algunos casos los mercados eran ciudades y otros casos el estado completo o unión de ciudades por lo que los tipos de mercados que maneja la fuente no están en el mismo nivel geográfico por lo que se decidió que en lugar de asignar coordenadas se generaría una nueva variable “**v\_Geography\_zonas**” donde se agruparía los mercados en las 8 zonas o regiones que maneja la fuente de información.

En la página oficial de nuestros datos se encuentra un mapa interactivo donde al poner el cursor en algun estado indica los nombres de los mercados que están dentro de la zona (i) a continuación un ejemplo:



De esta manera se “**normalizó** la variable de geografía y el total país queda distribuido de la siguiente manera:



#### ■ v\_Geography\_zonas\_clas

Tambien se detectó que los mercados venían por 44 (mercados más pequeños), 8 (zonas o regiones totalizadores) y 1 total país por lo que se generó esta columna identificadora donde:

0 = mercado pequeño

1 = zona total (suma de los mercados que la integran)

2 = total país

Es importante destacar en este punto que en la página es claro que la suma de zonas es menor a total país ya que aunque la página permite una suscripción que es la única manera de obtener la data, por temas de confidencialidad se reservan la información de algunas zonas y mercados.

- **v\_Type**

Para esta variable tambien fue necesaria la limpieza de texto para homologar las categorías y mantener las 2 únicas opciones (organico y convencional)



v_Type	
Organic	7920
organic	5376
conventional	4622
Conventional	3909
Conventional	2373

v_Type	
organic	13296
conventional	10904

- **c\_Total\_Units**

A este punto sabemos que el volumen total viene de la suma de tipos de aguacates con el total de bolsas que a su vez esta última es la suma de bolsas chicas, grades y extra grandes por lo que parece importante tener una columna con la suma de unidades de aguacates  $c\_4046\text{ Units} + c\_4225\text{ Units} + c\_4770\text{ Units}$  y posiblemente hacer análisis del tipo de aguacates a granel únicamente.

$$\begin{aligned} \text{C\_Total Units} = & + 4046 \\ & + 4525 \\ & + 4770 \end{aligned}$$


- **c\_precio\_dolares**

Como ASP (Average Selling Price) es el precio de venta promedio es decir (Dolares /Unidades) y en la tabla tengo unidades (c\_Total Bulk and Bags Units ) y el ASP (c\_ASP Current Year) puedo calcular los precios en dolares al multiplicar las dos variables de la ecuación con las que cuento. Por la construcción de esta variable fue que se revisaron antes los outliers de las columnas iniciales.

**■ d\_Week\_Ending**

Esta variable inicialmente contiene fecha con el formato yyyy-mm-dd por lo que se convirtió a datetime y posteriormente se obtuvo:

**d\_año:** El único propósito de generar esta nueva columna es para hacer análisis exploratorio por año pues como se mencionó en la sección de outliers, sabíamos que posiblemente en 2017 el precio o el volumen variaría. Esta variable posteriormente en los modulos de predicción no aportaría información importante.

**d\_mes:** Se obtuvo el mes con el fin de analizar la producción de aguacate por temporalidad de esta manera se decidió también generar la columna **v\_estaciones:**



**d\_dia:** Finalmente se obtuvo el día del corte de la producción. Los cortes vienen por semana donde en algunos años el corte se hace los domingos y otros años el corte fue el día viernes.

## 14Introducción, Manipulación, Exploración y Visualización de Datos

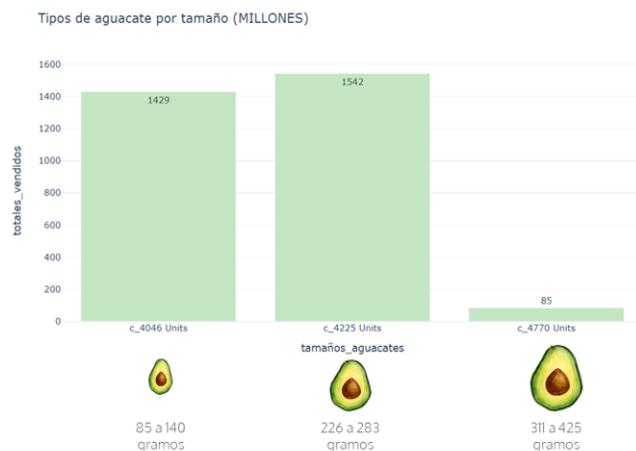
De esta manera las columnas de la tabla quedan como se muestra a continuación:

Variable	Tipo de variable	Descripción
v_Geography	Categórica	Ciudad o region (mercados = reportes regionales de HASS AVOCADO BOARD)
v_Type	Categórica	Tipo de aguacate (Convencional u Orgánico)
c_ASP Current Year	Numérica (continua)	Average Selling Price (Dolares /Unidades)
c_Total Bulk and Bags Units	Numérica (continua)	Número total de aguacates vendidos
c_4046 Units	Numérica (continua)	Número total de "Small/Medium Hass Avocado(~3-5oz avocado)" vendidos
c_4225 Units	Numérica (continua)	Número total de "Large Hass Avocado (~8-10oz avocado)" vendidos
c_4770 Units	Numérica (continua)	Número total de "Extra Large Hass Avocado (~10-15oz avocados)" vendidos
c_TotalBagged Units	Numérica (continua)	Total de bolsas de aguacate Hass vendidas
c_SmlBagged Units	Numérica (continua)	Total de bolsas pequeñas de aguacate Hass vendidas
c_LrgBagged Units	Numérica (continua)	Total de bolsas grandes de aguacate Hass vendidas
c_X_LrgBagged Units	Numérica (continua)	Total de bolsas extra grandes de aguacate Hass vendidas
d_Week_Ending	Fecha	Semana de ventas/producción
<b>Ingeniería de variables</b>		
v_Geography_zonas	Categórica	Zona o regiones de HASS AVOCADO BOARD
vGeographyzonasclas	Categórica	0 = mercado, 1 = zona , 2 = total país
c_Total_Units	Numérica (continua)	Número total de aguacates a granel (4046 U + 4225 U + 4770 U)
c_precio_dolares	Numérica (continua)	Precio de venta en dólares
d_año	Fecha	Año de la producción
d_dia	Fecha	Día de la producción
d_mes	Fecha	Mes de la producción
v_MES_x	Categórica	Estaciones (string) para análisis exploratorio
v_estaciones	Categórica	Estacion del año de la producción

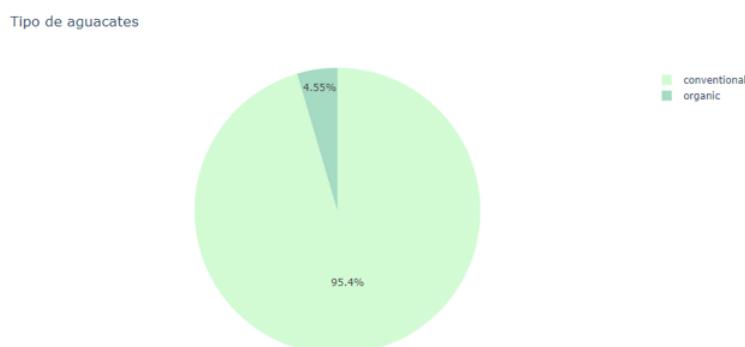
Dando como resultado final un conjunto de datos con la siguiente dimensión: **(24200 filas , 21 columnas)**

## 1.7. Análisis exploratorio

- En EUA 47 % de las unidades de aguacate Hass que se venden son los aguacates más pequeños y 50 % el Hass tamaño mediano... los aguacates que pesan casi medio kilo equivalen solo al 3 % de las ventas de las unidades.



- La proporción de ventas de aguacates orgánicos es muy baja, solo 5 % de las ventas van para esta opción ecológicamente más sustentable.



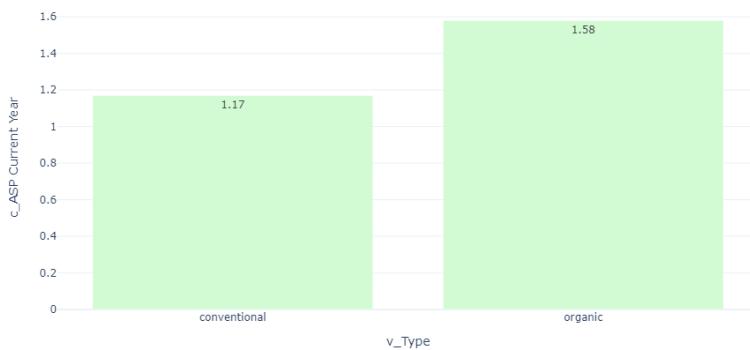
## 16Introducción, Manipulación, Exploración y Visualización de Datos

A nivel año podemos ver como el aumento en la venta de tipo de aguacate orgánico tiene una tendencia creciente aunque la proporción sigue siendo muy baja:

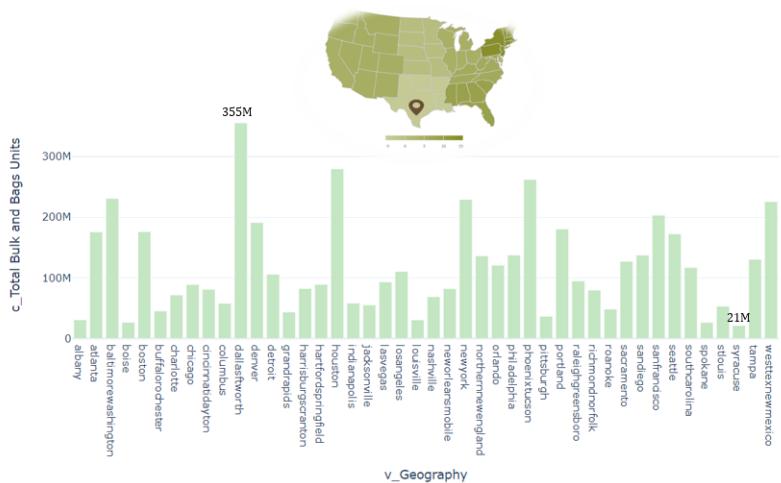
		c_Total Bulk and Bags Units					
d_año	v_Type	2016	2017	2018	2019	2020	
		conventional	97%	96%	95%	95%	94%
	organic	3%	4%	5%	5%	6%	

Como se esperaba, el ASP (Average Selling Price) en los aguacates de tipo orgánico es mayor que el del tipo convencional:

ASP promedio tipo de aguacates



- Se revisó a nivel mercado (44 mercados) el total de volumen y se encontró que el mercado de “**Hass Avocado Board**” con más ventas es “**Dallas ft Worth**” el cual pertenece a la zona **South central** esta zona colinda con México en 4 estados (Chihuahua, Coahuila, Nuevo Leon y Tamaulipas)



Esto motivó a revisar el volumen de ventas en las 8 regiones y se encontró que en este nivel la zona con más ventas es “**West**” la cual incluye lugares como “Las Vegas” y “Seattle” seguida de la zona “**North East**” donde se encuentran “New York” y “Philadelphia”.

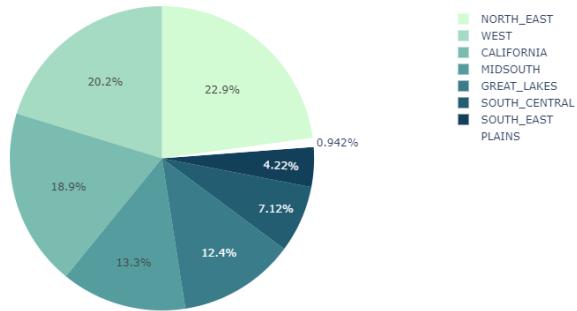


## 18Introducción, Manipulación, Exploración y Visualización de Datos

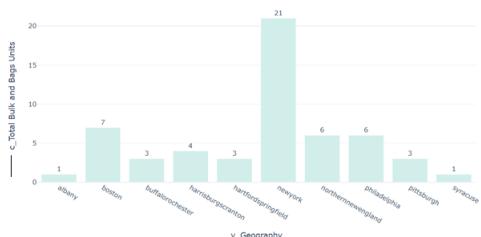
### ■ Distribución de aguacates orgánicos por zona:

De los aguacates orgánicos en la zona “[North East](#)” el mercado predominante es por mucho [New York](#) con ventas reportadas de 21 millones en volumen. Sin embargo en la 3er zona con más ventas en este tipo de aguacate “[Californornia](#)” se encontró que “[Los Ángeles](#)” tiene ventas de hasta 28 millones de volumen, lo cual es una de las zonas que ha destacado en los últimos años en los mapas de revolución de fitness en Estados Unidos y hace sentido que sea el mercado donde más aguacates orgánicos sean consumidos.

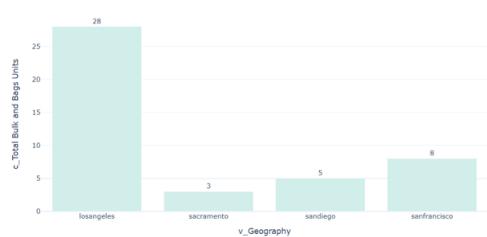
Volumen total (ORGANICOS) por zona



NORTH\_EAST\_Millones\_ORGANICOS



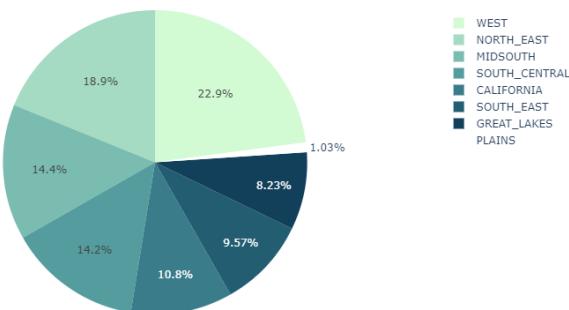
CALIFORNIA\_Millones\_ORGANICOS



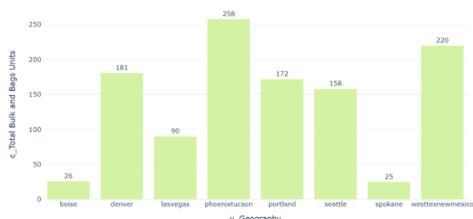
- Distribución de aguacates convencionales por zona:

En cuanto a los aguacates convencionales en la zona “West” que es la zona no. 1 es **Phoenix** el mercado que reporta más ventas (258M volumen). Sin embargo en la 4ta zona con más ventas en este tipo de aguacate “South Central” se encontró que “Dallas” tiene ventas de hasta 347M de volumen.

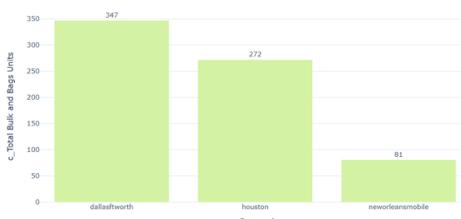
Volumen total (CONVENCIONALES) por zona



WEST\_Millones\_CONVENCIONALES



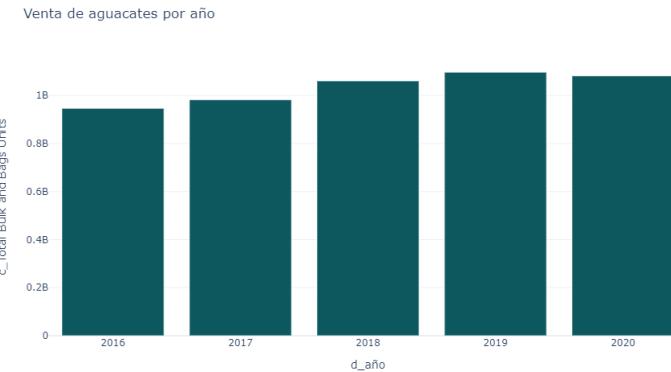
SOUTH\_CENTRAL\_Millones\_CONVENCIONALES



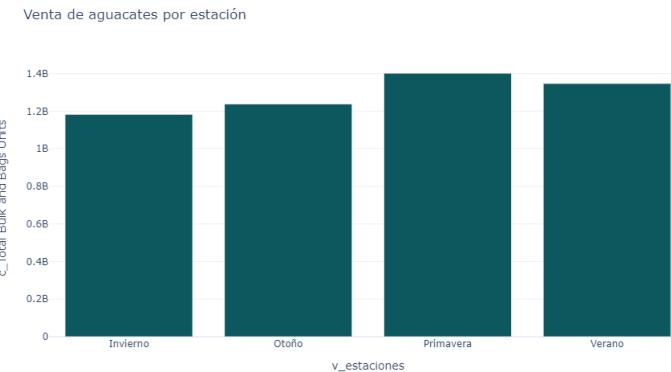
## 20Introducción, Manipulación, Exploración y Visualización de Datos

### ■ Temporalidad:

En los últimos 5 años, **2019** es el año con más reporte de ventas:



La estación donde más se venden aguacates es **primavera**: marzo a mayo



A nivel mensual, pese a que se esperaba que febrero fuera el mes con más ventas debido al Super Bowl y la alta demanda del producto motivo de este análisis en este evento, podemos observar que **mayo** es el mes con más ventas mientras que **diciembre** es el mes donde menos se compran aguacates Hass.

### Exportación récord de aguacate mexicano para el Super Bowl LV

Javier Rosas 3 Feb, 2021



**La Asociación de Productores y Empacadores Exportadores de Aguacate de México aumentará su envío de producto a Estados Unidos**

El aguacate volverá a estar presente en el **Super Bowl**.

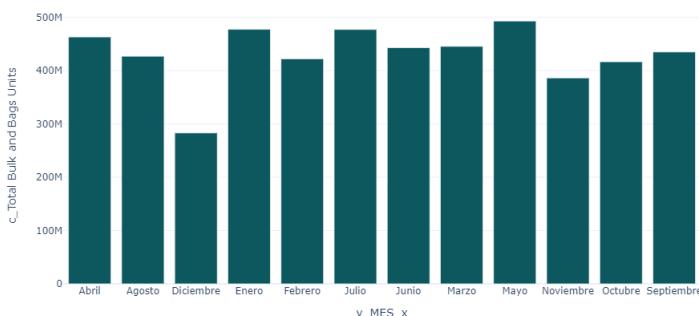
De acuerdo a datos proporcionados a **ESPN Digital** por la Asociación de Productores, Empacadores y Exportadores de Aguacate de México (APEAM), se espera que la exportación del producto a Estados Unidos sea una cifra récord con un aumento estimado de 4 por ciento en comparación al año anterior, en el que se mandaron 120 mil toneladas previo al Super Bowl LIV.



**ESPN**

La pandemia del COVID-19 no afectó las ventas de aguacate para los productores mexicanos. ESPN Digital

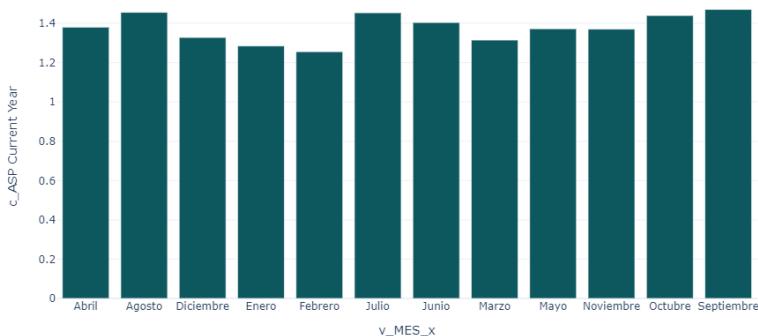
Venta de aguacates por mes



## 22Introducción, Manipulación, Exploración y Visualización de Datos

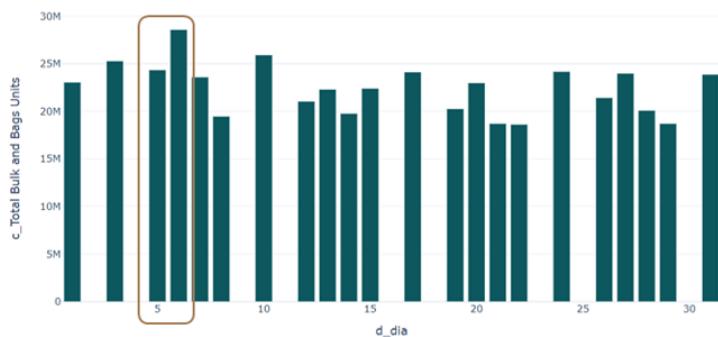
Al obtener el ASP promedio por mes podemos observar que **diciembre**, mes en donde menos aguacates se venden, no es el mes con el ASP más alto sino que en realidad es **septiembre**:

ASP de aguacates por mes



Como observamos que **mayo** es el mes con más ventas se revisaron los días de este mes y encontramos que las semanas dentro del periodo de tiempo 2016-2020 donde coincide el **5 de mayo** es donde más movientos de venta se encontraron:

Mayo - Ventas por días (Week Ending)



Y es que precisamente el **5 de mayo** ha sido usado como una celebración cultural para reconocer a los mexicanos y latinos en Estados Unidos, su cultura y aportes a la diversidad americana, celebración que se ha vuelto más importante que la Independencia de México en ese país. “Y no es raro encontrar a estadounidenses -incluso latinos- que piensan que el Cinco de Mayo, como le llaman, es el día de la Independencia mexicana.”

## 5 de mayo: claves para entender la relación de "amor y odio" entre México y Estados Unidos

Dario Brooks  
BBC News Mundo

5 mayo 2021



Cada 5 de mayo, los estadounidenses celebran a México.

Una fecha que en el país latinoamericano **no causa expectación**, e incluso pasa desapercibida, en Estados Unidos es motivo para deleitarse con comida mexicana, ponerse sombrero y bigotes falsos. Y con una margarita en mano brindan por su vecino del sur.

La fecha marca el triunfo del ejército mexicano sobre los invasores franceses en la Batalla de Puebla, el 5 de mayo de 1862.

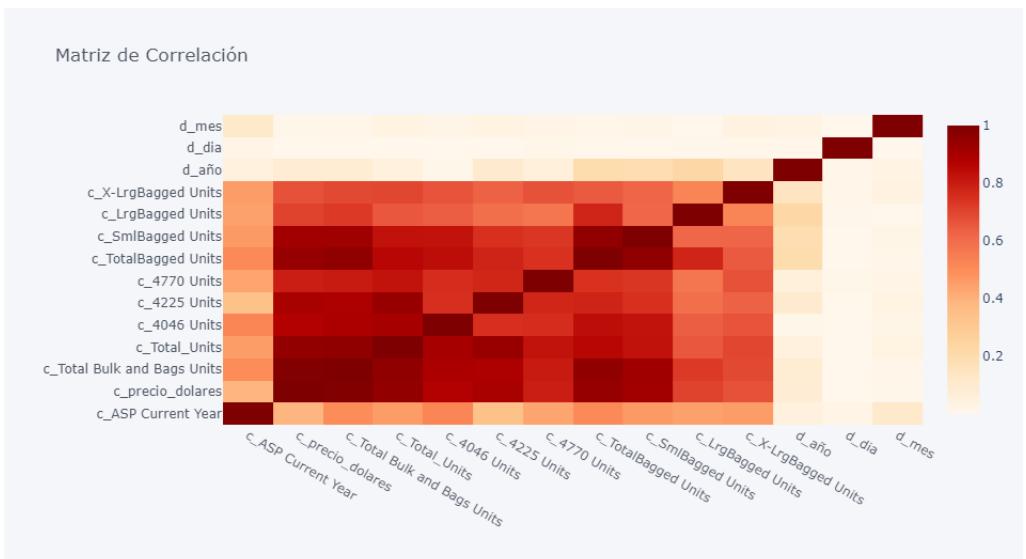
Y no es raro encontrar a estadounidenses -incluso latinos- que piensan que el **Cinco de Mayo**, como le llaman, es el día de la Independencia mexicana (en realidad es el 16 de septiembre).

## 1.8. Reducción de variables:

### 1.8.1. Filtro de alta correlación

Sabemos que la correlación nos indica qué tanta relación existe entre una variable y otra, entonces podemos hacer el análisis de correlación de cada una de nuestras variables con el resto de ellas, (relación uno a uno) y encontraremos de acuerdo a los valores de correlación aquellas que en pares sean muy cercanas, es decir que nos aporten información similar y las vuelvan redundantes y por lo tanto podrían aportar colinealidad al modelo, de esta manera podríamos eliminar las que estén altamente correlacionadas con nuestra variable objetivo.

A continuación se muestra un mapa de calor:



Como podemos observar, existe alta colinealidad en nuestra tabla, si fijaramos como variable target “c\_4025\_Units” que es el tipo de aguacate a granel con mayores ventas y quisieramos estimar su volumen entonces veríamos lo siguiente:

c_4225 Units	
c_precio_dolares	0.900199
c_Total Bulk and Bags Units	0.890363
c_Total_Units	0.938972
c_4046 Units	0.751479
c_4225 Units	1.000000
c_4770 Units	0.776818
c_TotalBagged Units	0.780459
c_SmlBagged Units	0.750090

Es decir, como podríamos intuir por colinealidad podríamos prescindir de las variables “c\_precio\_dolares”, “c\_Total Bulk and Bags Units” y la “c\_Total\_Units”.

### 1.8.2. Correlación con objetivo

Contrario al punto anterior en este filtro se identifican las variables que tienen correlación mínima o muy baja con nuestra variable objetivo ya que posiblemente no nos aporte información útil al modelo. En este caso si fijamos la variable target como “c\_4025\_Units” tendríamos lo siguiente:

c_4225 Units	
d_dia	0.003412
d_mes	0.035385

Con la tabla de datos que quedaría al eliminar las 5 variables anteriores podríamos entrenar un modelo y así cumplir con el objetivo inicial del siguiente modulo sabiendo que los datos están limpios, su calidad es aceptable y sin valores atípicos, además de conocer el contexto de nuestro análisis.

## Apéndice A

# Referencias

DATOS:

- <https://hassavocadoboard.com/category-data/>
- <https://hassavocadoboard.com/inside-hab/>
- <https://hassavocadoboard.com/research/holidays-events/>
- <https://loveonetoday.com/how-to/identify-hass-avocados/>

NOTAS :

- <https://www.elfinanciero.com.mx/rankings/la-importancia-del-aguacate-para-mexico-en-graficas/>
- <https://www.visittheusa.mx/info/tiempo-y-estaciones-del-ano>
- [https://www.espn.com.mx/futbol-americano/nota/\\_/id/8126068/super-bowl-lv-aguacate-mexicano-estima-exportacion-record](https://www.espn.com.mx/futbol-americano/nota/_/id/8126068/super-bowl-lv-aguacate-mexicano-estima-exportacion-record)
- <https://www.umnews.org/es/news/por-que-el-5-de-mayo-es-celebrado-en-estados-unidos>
- <https://www.bbc.com/mundo/noticias-56938589>
- <https://www.univision.com/noticias/citylab-salud/en-mapas-la-revolucion-del-fitness-en-estados-unidos>