



Santiago Gabian Perez A01658280

Grupo 222

Mayo 2022

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Ciudad de México

Herramientas computacionales: el arte de la analítica

TC1002.S

Carga de datos:

```
In [23]: import pandas as pd
df = pd.read_csv("avocado.csv")
df.head()
```

```
Out[23]:
```

	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	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	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	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	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	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

Cantidad de Datos:

```
In [24]: df.count()
```

```
Out[24]:
```

Unnamed: 0	18249
Date	18249
AveragePrice	18249
Total Volume	18249
4046	18249
4225	18249
4770	18249
Total Bags	18249
Small Bags	18249
Large Bags	18249
XLarge Bags	18249
type	18249
year	18249
region	18249
dtype: int64	

En el archivo de datos de avocado se encuentran 18249 datos de la cual todos excepto región y tipo son enteros y región y tipo son strings.

Análisis de variables:

```
In [13]: df.describe()
```

```
Out[13]:
```

	Unnamed: 0	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags		
count	18249.000000	18249.000000	1.824900e+04	1.824900e+04	1.824900e+04	1.824900e+04	1.824900e+04	1.824900e+04	1.824900e+04	18249.000000	18249.0	
mean	24.232232	1.405978	8.506440e+05	2.930084e+05	2.951546e+05	2.283974e+04	2.396392e+05	1.821947e+05	5.433809e+04	3106.428507	2016.1	
std	15.481045	0.402677	3.453545e+06	1.264989e+06	1.204120e+06	1.074641e+05	9.862424e+05	7.461785e+05	2.439660e+05	17692.894652	0.9	
min	0.000000	0.440000	8.456000e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	2015.0	
25%	10.000000	1.100000	1.083858e+04	8.540700e+02	3.008780e+03	0.000000e+00	5.088640e+03	2.849420e+03	1.274700e+02	0.000000	2015.0	
50%	24.000000	1.370000	1.073768e+05	8.645300e+03	2.906102e+04	1.849900e+02	3.974383e+04	2.636282e+04	2.647710e+03	0.000000	2016.0	
75%	38.000000	1.660000	4.329623e+05	1.110202e+05	1.502069e+05	6.243420e+03	1.107834e+05	8.333767e+04	2.202925e+04	132.500000	2017.0	
max	52.000000	3.250000	6.250565e+07	2.274362e+07	2.047057e+07	2.546439e+06	1.937313e+07	1.338459e+07	5.719097e+06	551693.650000	2018.0	

Media, Mediana y Desviación estándar:

```
In [22]: df.std(axis = 0)
```

C:\Users\santi\anaconda3\envs\Semana tec\lib\site-packages\ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
 """Entry point for launching an IPython kernel.

```
Out[22]: Unnamed: 0      1.548104e+01
AveragePrice      4.026766e-01
Total Volume      3.453545e+06
4046              1.264989e+06
4225              1.204120e+06
4770              1.074641e+05
Total Bags        9.862424e+05
Small Bags        7.461785e+05
Large Bags        2.439660e+05
XLarge Bags       1.769289e+04
year              9.399385e-01
dtype: float64
```

```
In [20]: df.mean(axis = 0)
```

C:\Users\santi\anaconda3\envs\Semana tec\lib\site-packages\ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
 """Entry point for launching an IPython kernel.

```
Out[20]: Unnamed: 0      24.232232
AveragePrice      1.405978
Total Volume      850644.013009
4046              293008.424531
4225              295154.568356
4770              22839.735993
Total Bags        239639.202060
Small Bags        182194.686696
Large Bags        54338.088145
XLarge Bags       3106.426507
year              2016.147899
dtype: float64
```

```
In [21]: df.median(axis = 0)
```

C:\Users\santi\anaconda3\envs\Semana tec\lib\site-packages\ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
 """Entry point for launching an IPython kernel.

```
Out[21]: Unnamed: 0      24.00
AveragePrice      1.37
Total Volume      107376.76
4046              8645.30
4225              29061.02
4770              184.99
Total Bags        39743.83
Small Bags        26362.82
Large Bags        2647.71
XLarge Bags       0.00
year              2016.00
dtype: float64
```

Esto nos dice que el precio promedio es alrededor de 1.4 y una desviación estándar de 0.402 lo cual nos dice que el precio no varía tanto. Con el precio promedio mínimo siendo 0.44 y el máximo 3.25. La poca varianza tiene que ver que ya es un precio promedio y los promedios varían menos que si fueran todos los datos sin promediar.

Luego en el Volumen total se puede ver que hay una gran desviación estándar dando a entender la gran varianza que hay de volúmenes totales.

También se puede ver que varía mucho el número total de bolsas incluyendo grandes y chicas. Según los datos podemos concluir que normalmente se usan mucho más bolsas pequeñas que grandes. De que en promedio hay 239,639 bolsas totales y de esas 182,194 son chicas, 54,338 son grandes y 3,106 son extra grandes.

Así que en conclusión el precio promedio total no suele variar mucho estando en promedio en 1.4 mientras que el volumen total si. También podemos concluir de que de todas las bolsas compradas se compran mucho más que las grandes y extragrandes.