

EL CAFÉ EN COLOMBIA

APRENDIZ

JEAN PAUL BENITEZ DELGADO

LEIDY PAOLA RODALLEGA SANCHEZ

INSTRUCTOR

LUIS ARMANDO AMAYA QUINTERO

SERVICIO NACIONAL DE APRENDIZAJE

PROGRAMACIÓN PARA ANALITICA DE DATOS

SANTIAGO DE CALI, OCTUBRE/2020

Contenido

INTRODUCCIÓN	3
LA CALIDAD: ¿LA CLAVE DE SU ÉXITO?.....	4
¿QUÉ LO DIFERENCIA DEL RESTO DE CAFÉS?	4
ECONOMÍA CAFETERA Y DESARROLLO ECONÓMICO EN COLOMBIA	4
ANALISIS EN PYTHON NOTEBOOK	30

INTRODUCCIÓN

Este trabajo tiene como fin contextualizar, nos enseña la calidad del café colombiano, en que se diferencia del resto y su desarrollo económico en Colombia.

A continuación, se presentará de manera gráfica información respecto al cultivo del café, tanto su producción, recolección, hectáreas utilizadas, y peso en sacos, para dar conocimiento acerca del café, se hizo un análisis en graficas estadísticas que nos representan dichos porcentajes.

Este informe fue hecho con ayuda de muchos sitios que representan la producción no solo del café si no de otros cultivos, algunos actualizados y otros que representan años pasados, en gráficas y textos se elaboró de manera dinámica para una mejor comprensión.

LA CALIDAD: ¿LA CLAVE DE SU ÉXITO?

Podríamos pensar entonces que la fama del Café Colombiano está motivada por su calidad. Lo cierto es que el café de origen colombiano es muy apreciado en el resto del mundo. Sus denominaciones de origen nacionales (Cauca, Nariño, Huila, Santander, Tolima y Sierra Nevada.) y desde 2005 también con una denominación de origen otorgada por la Unión Europea.

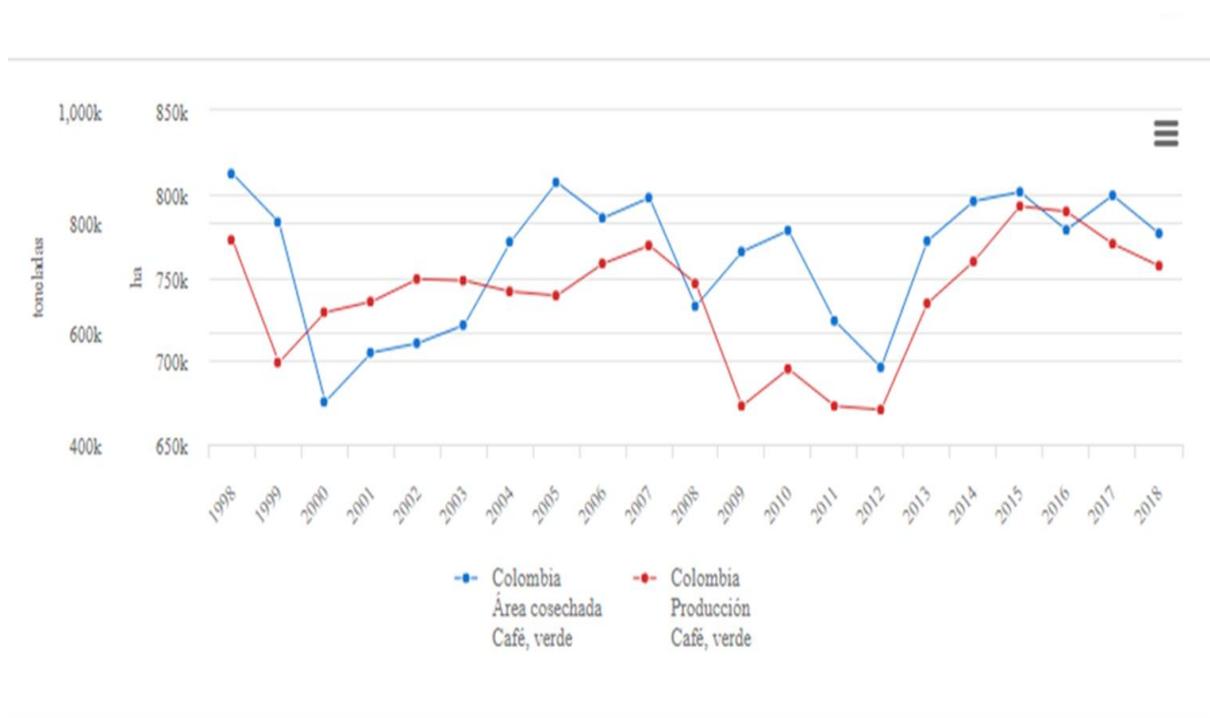
¿QUÉ LO DIFERENCIA DEL RESTO DE CAFÉS?

- **La variedad:** cultivan sólo café arábico, el más apreciado por su sabor y aroma.
- **Las condiciones geográficas:** su clima tropical y las altas montañas son ideales para el cultivo de café.
- **Los procesos:** mientras que Brasil apuesta por una recolección automatizada, en Colombia predomina la recolección manual que mejora la calidad del grano.

ECONOMÍA CAFETERA Y DESARROLLO ECONÓMICO EN COLOMBIA

Los colombianos no podemos olvidar que el café ha sido uno de nuestros productos de exportación más importantes. Su nivel de producción es tan alto que compromete a 590 municipios y los departamentos andinos del país. El área disponible para el cultivo del café es de cerca de 7,3 millones de hectáreas y se cultiva en 970 mil hectáreas, empleando a las familias propietarias de los predios cafeteros, y a miles de recolectores de café, que conforman el grueso de los trabajadores indirectos e indirectos, situación que determina que ésta sea nuestra industria emblemática.

PRODUCCIÓN/RENDIMIENTO DE CAFÉ, VERDE EN COLOMBIA.



FUENTE: FAOSTAT

URL: <http://www.fao.org/faostat/es/#data/QC/visualize>

COLOMBIA AREA COSECHADA

CAFÉ VERDE:

VALOR MAXIMO.

El pico mayor de café en área cosechada de 1998 al 2018 fue entre 800k y 850k Toneladas

El menor número de café en área cosechada de 1998 al 2018 fue entre 600k y 650k Toneladas

COLOMBIA PRODUCCIÓN

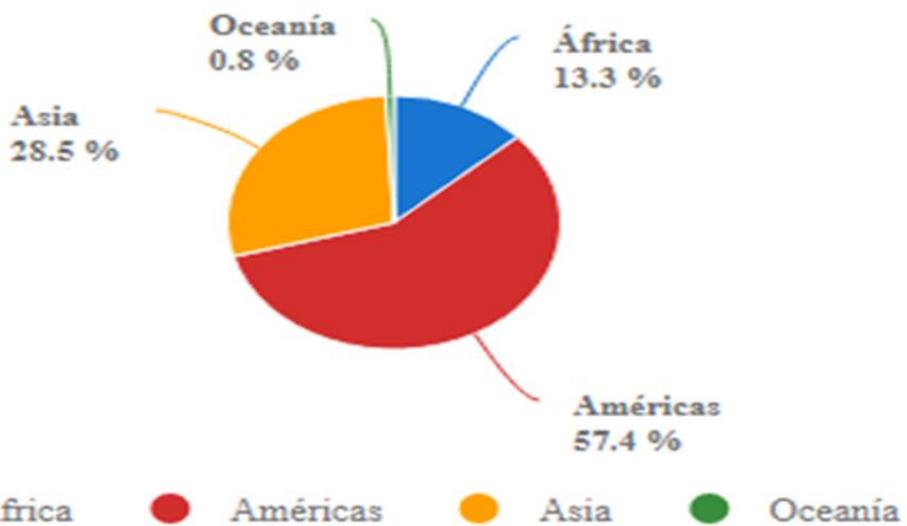
CAFÉ VERDE:

VALOR MINIMO.

El pico mayor de café en producción de café verde del 1998 al 2018 fue entre 800k y 1000k Toneladas

El menor número de café en producción de café verde del 1998 al 2018 fue entre 800k y 1000k Toneladas

PROPORCIÓN DE PRODUCCIÓN DE CAFÉ, VERDE POR REGIÓN.



FUENTE: FAOSTAT

URL: <http://www.fao.org/faostat/es/#data/QC/visualize>

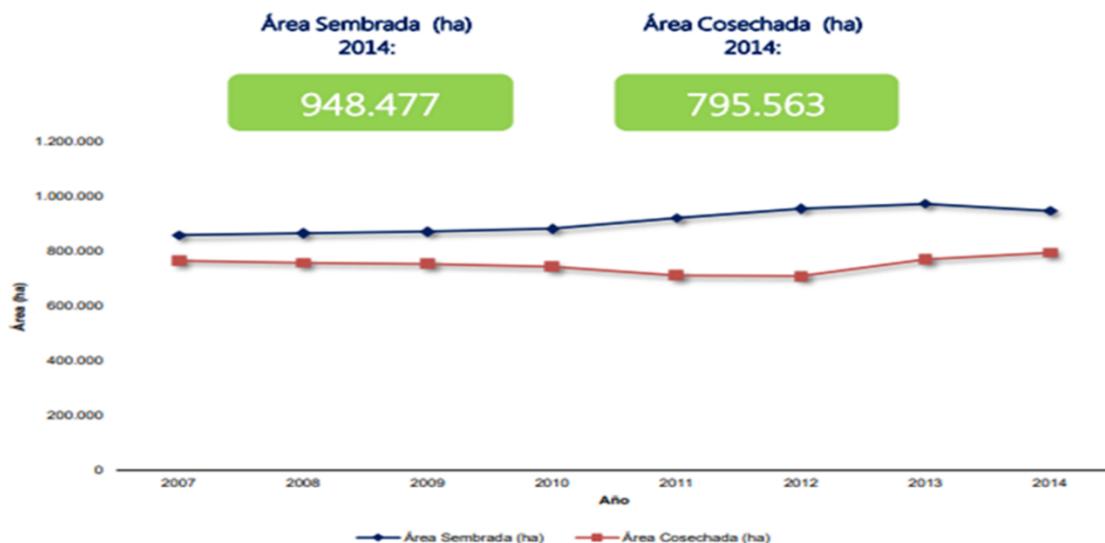
VALOR MAXIMO:

Proporción de producción de café por región, mayor porcentaje en Américas: fue de 57.4%

VALOR MINIMO:

Proporción de producción de café por región, menor porcentaje en Oceanía: fue de 0,8%

ÁREA SEMBRADA Y ÁREA COSECHADA DEL CULTIVO DE CAFÉ 2007-2014



FUENTE: FAOSTAT

URL: : <http://www.fao.org/faostat/es/#data/QC/visualize>

AREA SEMBRADA:

VALOR MAXIMO.

Área sembrada de café en (ha) menor número de (A.s en ha) en 2007: fue entre 800.000ha y 1.000.000ha Área cosechada en (ha)

VALOR MINIMO.

Área sembrada de café en (ha) pico mayor en 2013 al 2014: fue de 948.477 Área sembrada en (ha)

AREA COSECHADA:

VALOR MAXIMO.

Área cosechada de café en (ha) menor número de (A.c en ha) en 2012: fue entre 600.000ha y 800.000ha Área cosechada en (ha)

VALOR MINIMO.

Área cosechada de café en (ha) pico mayor en 2014: fue de 795.536 Área cosechada en (ha)

PRODUCCIÓN Y RENDIMIENTO DEL CULTIVO DE CAFÉ 2007-2014



FUENTE: AGRONET

URL: <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=3>

PRODUCCIÓN:

VALOR MAXIMO.

Producción-pico de café más alto en 2007 y 2008 fue entre 800.000(t) y 900.000(t)

VALOR MINIMO.

Producción-pico de café más bajo en 2011 y 2012 fue entre 600.000(t) y 700.000(t)

RENDIMIENTO:

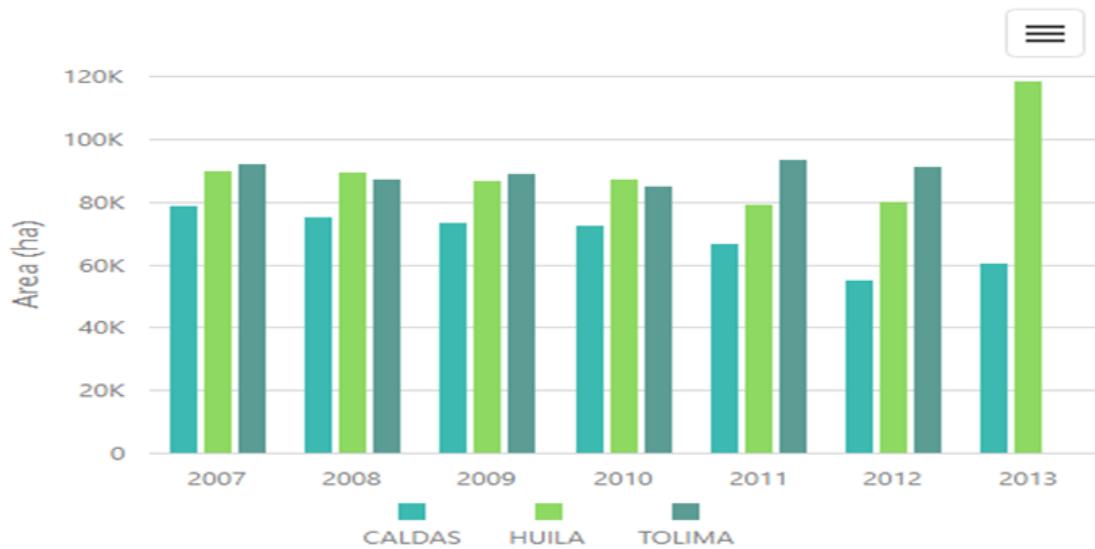
VALOR MAXIMO.

Rendimiento-pico de café más alto en 2008 fue entre 1.00(t/ha) y 1.20(t/ha)

VALOR MINIMO.

Rendimiento-pico de café más bajo en 2012 fue entre 0.80(t/ha) y 1.00(t/ha)

Área Cosechada por Departamento



FUENTE: AGRONET

URL: <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=3>

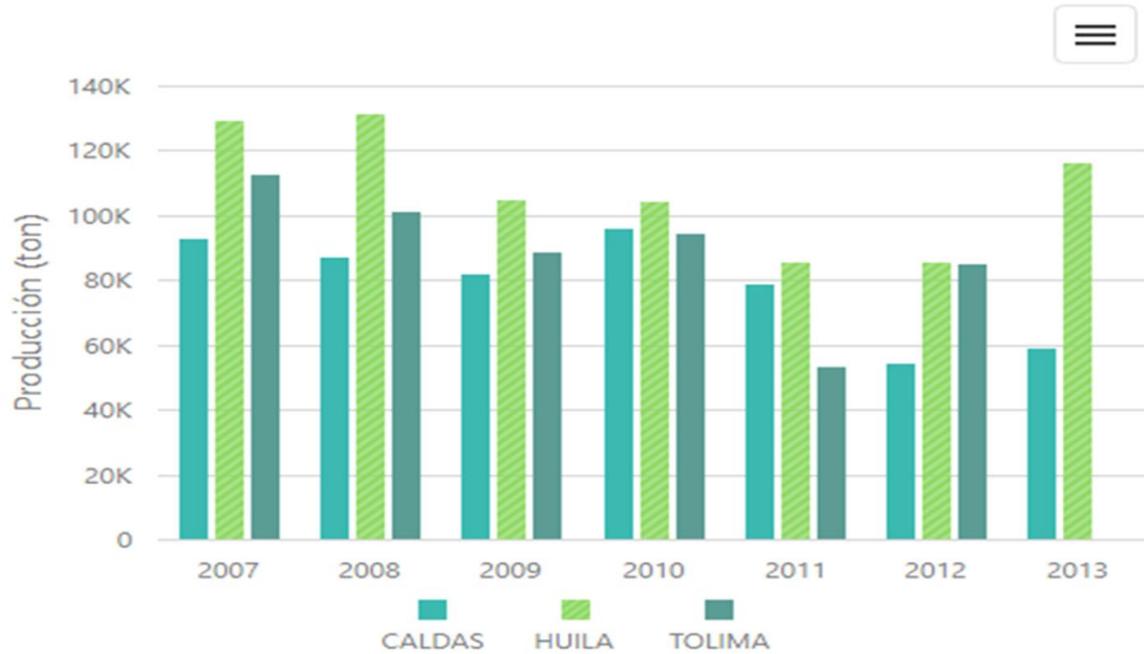
VALOR MAXIMO:

Mayor Área cosechada de café en 2013 por el departamento del huila fue entre 100k y 120k área (ha)

VALOR MINIMO:

Menor Área cosechada de café en 2012 por el departamento de caldas fue entre 40k y 60k área (ha)

Producción por Departamento



FUENTE: AGRONET

URL: <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=3>

VALOR MAXIMO:

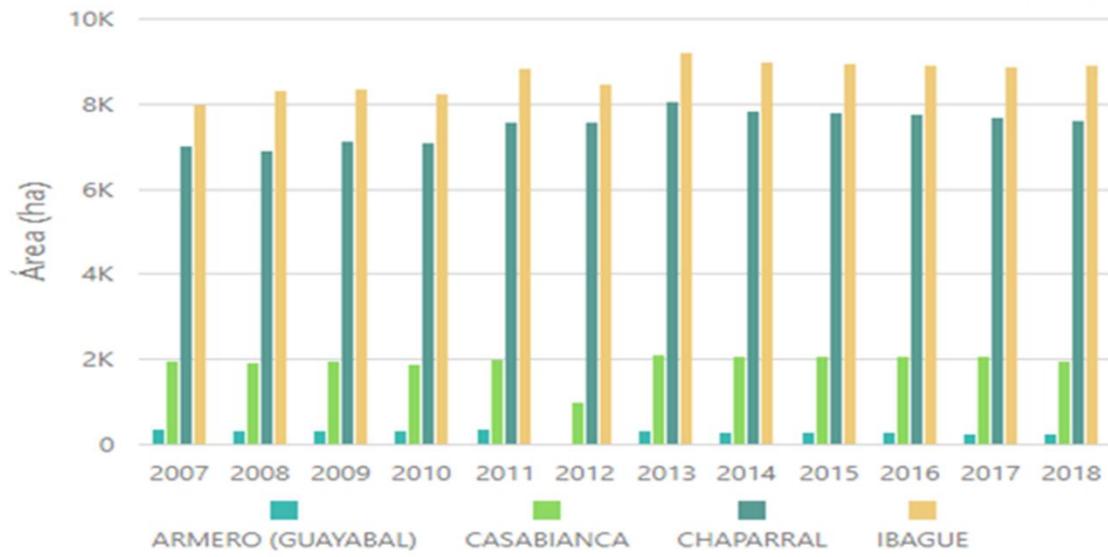
Mayor producción de café en 2008 por el departamento del huila fue entre 120k y 140k (ton)

CV

VALOR MINIMO:

Menor producción de café en 2011 por el departamento de Tolima fue entre 40k y 60k (ton)

Área Sembrada por Municipio



FUENTE: AGRONET

URL: <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=4>

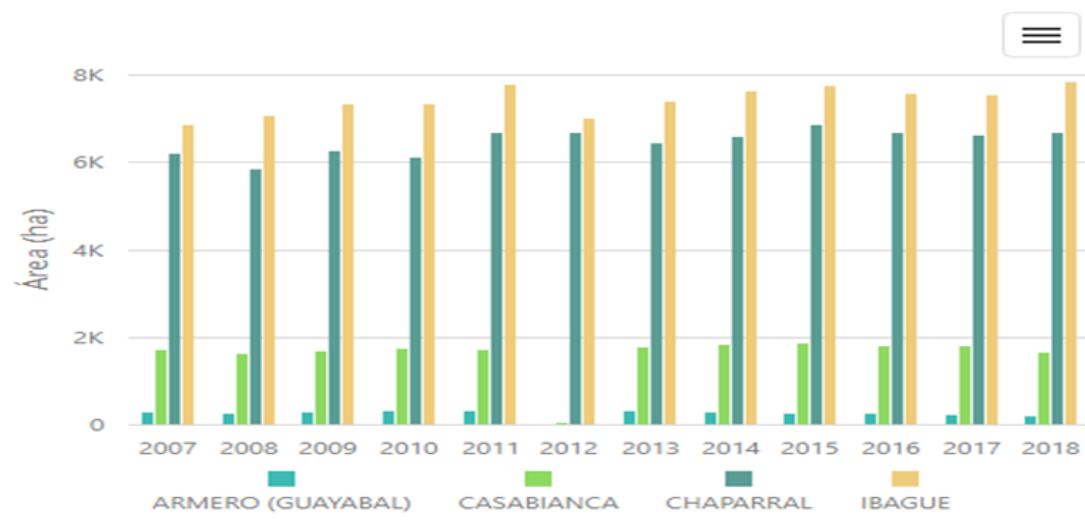
VALOR MAXIMO:

Mayor Área sembrada de café en 2013 por el municipio de Ibagué fue de 8k(ha) a 10k(ha)

VALOR MINIMO:

Menor Área sembrada de café en 2018 por el municipio de Armero fue de 0k(ha) a 2k(ha)

Área Cosechada por Municipio



FUENTE: AGRONET

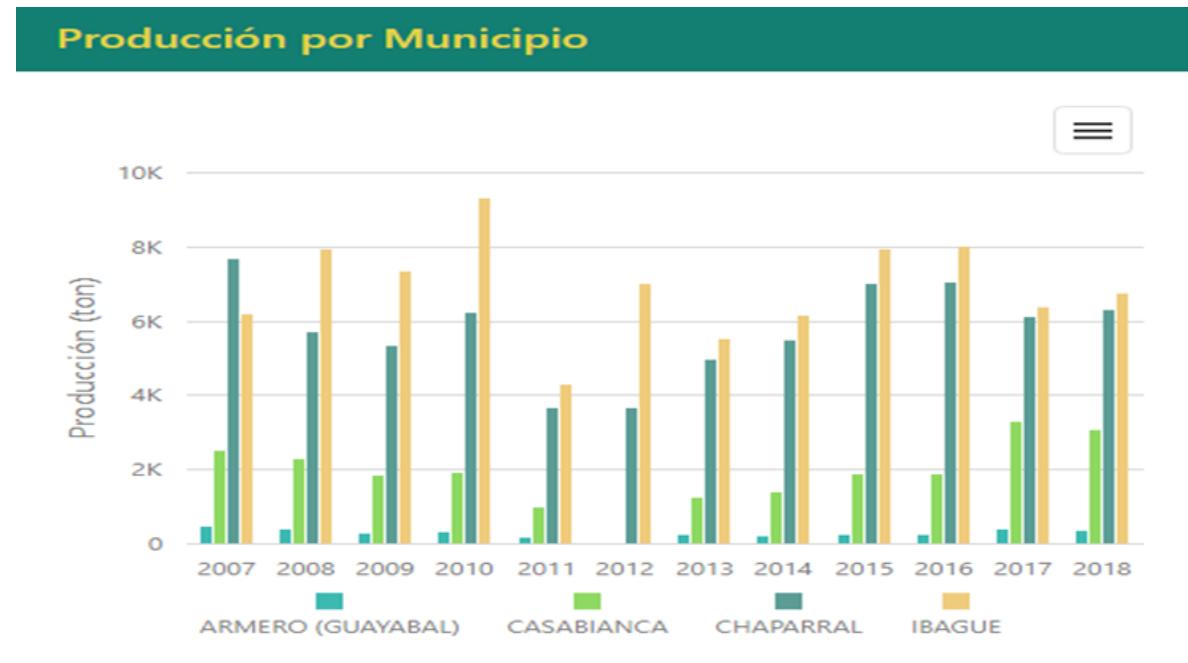
URL: <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=3>

VALOR MAXIMO:

Mayor Área cosechada de café en 2018 por el municipio de Ibagué fue de 6k(ha) a 8k(ha)
Área en (ha)

VALOR MINIMO:

Menor Área cosechada de café en 2012 por el municipio de Armero fue de 0k(ha) a 2k(ha)
Área en (ha)



FUENTE: AGRONET

URL: <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=3>

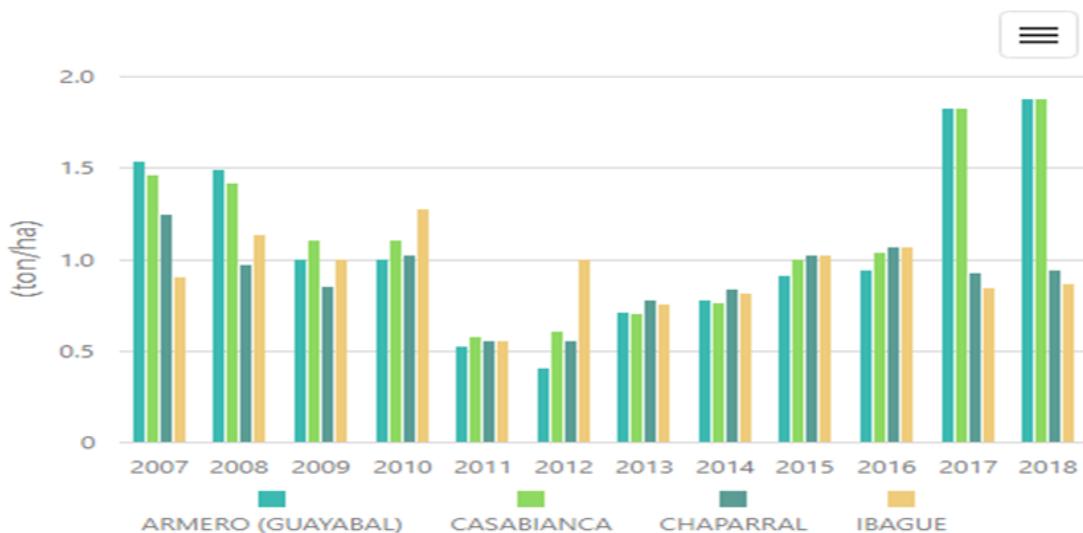
VALOR MAXIMO:

Mayor producción de café en 2010 por el municipio de Ibagué fue de 8k(tn) a 10k(tn)
Toneladas

VALOR MINIMO:

Menor producción de café en 2012 por el municipio de Armero fue de 0k(tn) a 2k(tn)
Toneladas

Rendimiento por Municipio



FUENTE: AGRONET

URL: <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=3>

VALOR MAXIMO:

Mayor rendimiento de café en 2018 por el municipio de Armero y Casablanca fue de 1.5(tn/ha) a 2.0(tn/ha)

VALOR MINIMO:

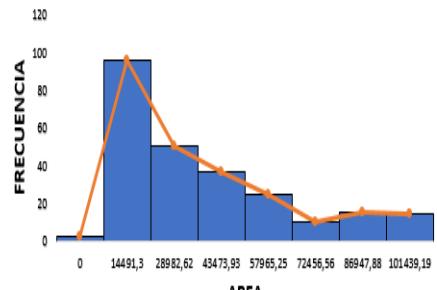
Menor rendimiento de café en 2012 por el municipio de Armero fue de 0(tn/ha) a 0.5(tn/ha)

Columna1
Media
34167,005
Error típico
2124,789295
Mediana
23309
Moda
30171,84
Desviación estándar
34654,26467
Varianza de la muestra
1200918060
Curtosis
0,017410554
Coeficiente de asimetría
1,010413484
Rango
130452,4
Mínimo
0
Máximo
130452,4
Suma
9088423,33
Cuenta
266

Número de clases	9
Tamaño de Clase	14491,31

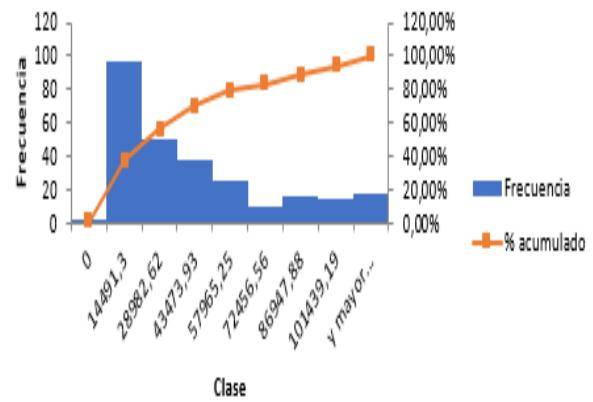
INTERVALOS		GRUPOS	FRECUENCIA
Lí	Ls		
0	14491,31	14491,3	96
14491,31	28982,63	28982,62	50
28982,63	43473,94	43473,93	37
43473,94	57965,26	57965,25	25
57965,26	72456,57	72456,56	10
72456,57	86947,89	86947,88	15
86947,89	101439,20	101439,19	14

HISTOGRAMA Y POLIGONO DE FRECUENCIA

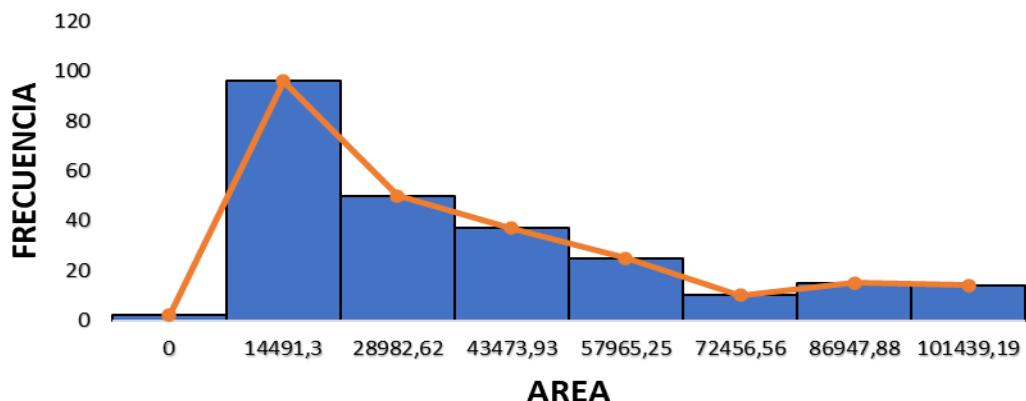


Clase	Frecuencia	% acumulado
0	2	0,75%
14491,3	96	36,84%
28982,62	50	55,64%
43473,93	37	69,55%
57965,25	25	78,95%
72456,56	10	82,71%
86947,88	15	88,35%
101439,19	14	93,61%
y mayor...	17	100,00%

Histograma



HISTOGRAMA Y POLIGONO DE FRECUENCIA



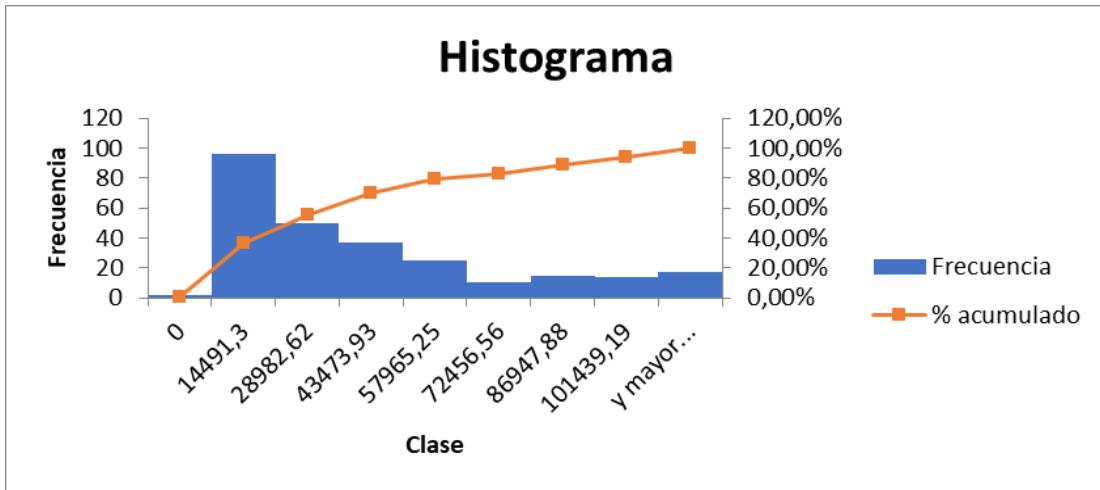
FUENTE: EXCEL

VALOR MAXIMO:

En el polígono de frecuencia, el pico mayor del 14491,3 del área fue de 100 (fr).

VALOR MINIMO:

En el polígono de frecuencia, el menor valor de 0 del área fue de 1(fr).



FUENTE: EXCEL

FRECUENCIA:

VALOR MAXIMO.

El valor máximo del 14491,3 de clase fue de 100(fr)

VALOR MINIMO.

El valor mínimo del 0 de clase fue de 0(fr)

% ACUMULADO:

VALOR MAXIMO.

El valor máximo del mayor porcentaje acumulado fue de 100.00%

VALOR MINIMO.

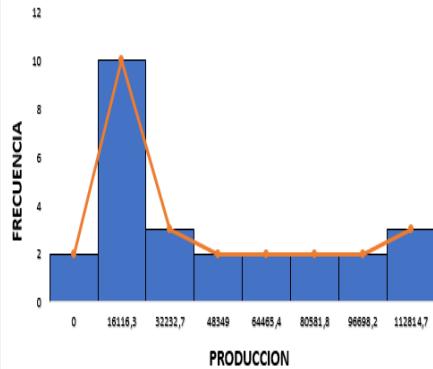
El valor mínimo de 0 fue de 0.00%

	Column1
2	
3	
4	Media 34605,88297
5	Error típico 2356,830387
6	Mediana 20706,69
7	Moda 26,7
8	Desviación estándar 38438,74035
9	Varianza de la muestra 1477536760
10	Curtosis 0,378679023
11	Coeficiente de asimetría 1,175037959
12	Rango 145168,1
13	Mínimo 0
14	Máximo 145168,1
15	Suma 9205164,87
16	Cuenta 266
17	
18	
19	
20	

Numero de Clases	9
Tamaño de Clase	16116,38

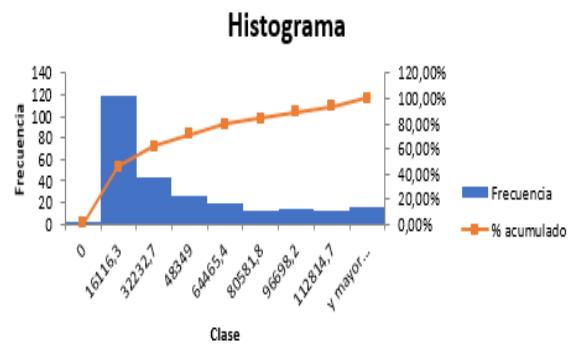
	INTERVALOS		GRUPOS	FRECUENCIA
	Li	Ls		
0		0	0	2
1		16116,4	16116,3	10
2	16116,4	32232,8	32232,7	3
3	32232,8	48349,1	48349	2
4	48349,1	64465,5	64465,4	2
5	64465,5	80581,9	80581,8	2
6	80581,9	96698,3	96698,2	2
7	96698,3	112814,7	112814,7	3

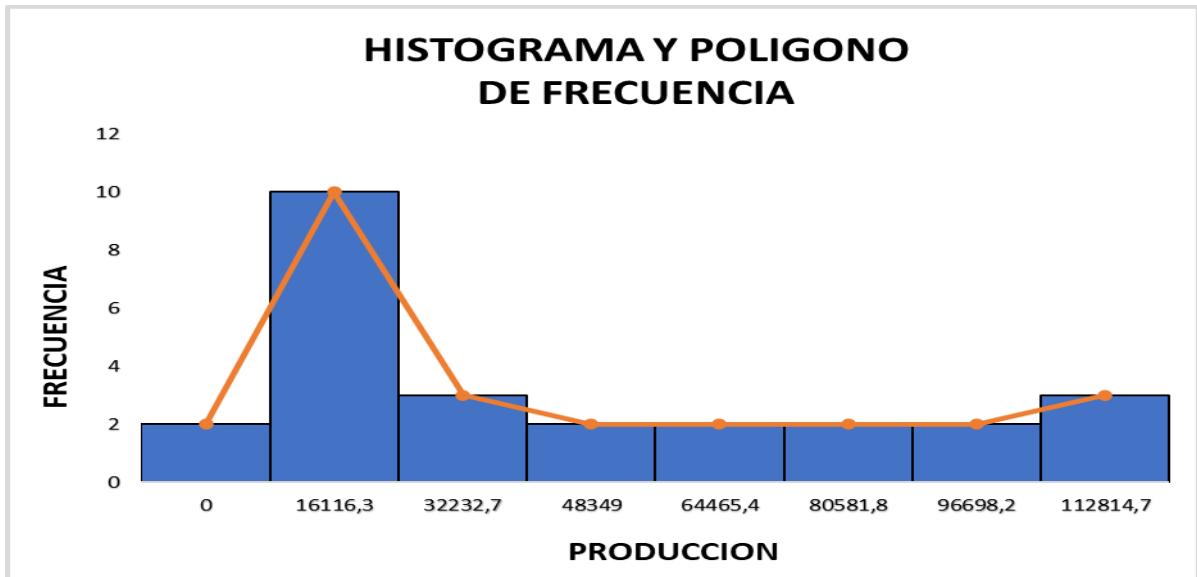
HISTOGRAMA Y POLIGONO DE FRECUENCIA



PRODUCCION

Close	Frecuencia	% acumulado
0	2	0,75%
16116,3	120	45,86%
32232,7	43	62,03%
48349	26	71,80%
64465,4	20	79,32%
80581,8	13	84,21%
96698,2	14	89,47%
112814,7	12	93,98%
y mayor...	16	100,00%





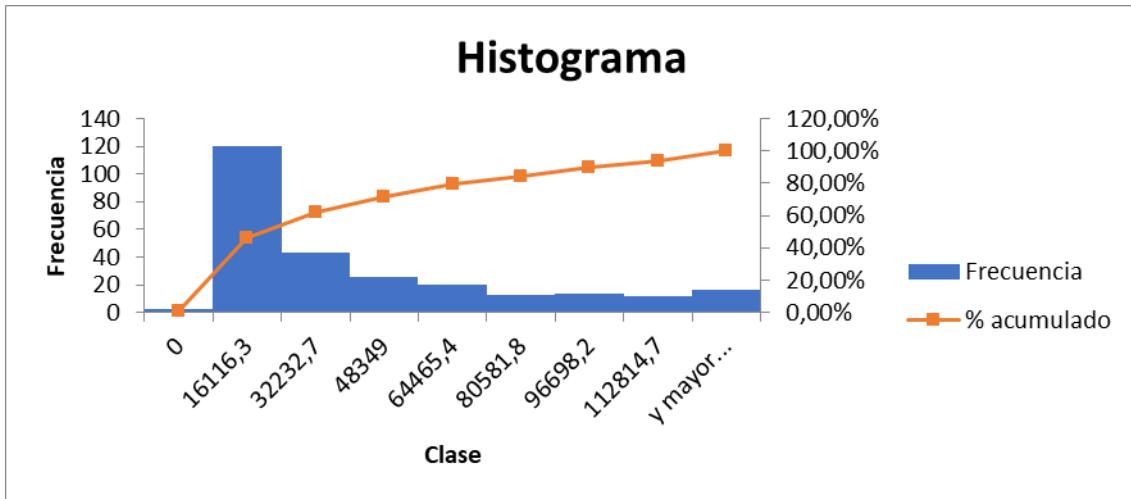
FUENTE: EXCEL

VALOR MAXIMO:

En el polígono de frecuencia, el pico mayor del 16116,3 de la producción fue de 10(fr).

VALOR MINIMO:

En el polígono de frecuencia, el menor valor de 0 de la producción fue de 2(fr).



FUENTE: EXCEL

FRECUENCIA:

VALOR MAXIMO.

El valor máximo del 1,1 de clase fue de 85(fr)

VALOR MINIMO.

El valor mínimo del 0 de clase fue de 0(fr)

% ACUMULADO:

VALOR MAXIMO.

El valor máximo del mayor porcentaje acumulado fue de 120.00%

VALOR MINIMO.

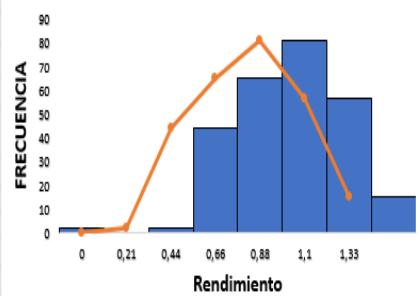
El valor mínimo de 0 fue de 0.00%

Columna1	
Media	0,936428571
Error típico	0,016378757
Mediana	0,94
Moda	0,6
Desviación estándar	0,267129445
Varianza de la	0,07135814
Curtosis	1,194773469
Coeficiente de asimetría	0,092323793
Rango	2
Mínimo	0
Máximo	2
Suma	249,09
Cuenta	266

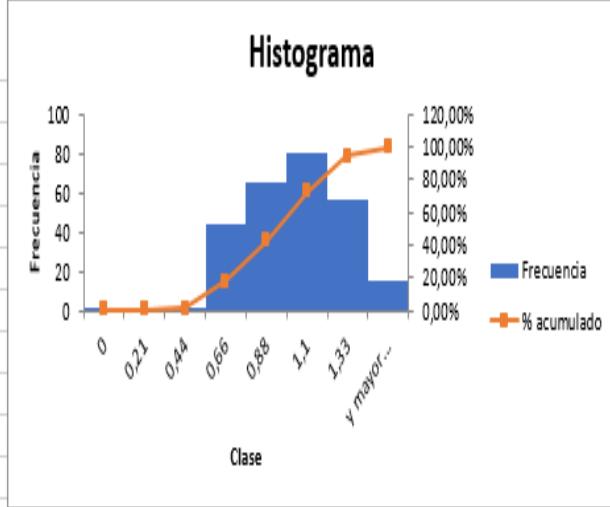
Numero de Clases	9
Tamaño de Clase	0,222170

INTERVALOS		GRUPOS	FRECUENCI
Li	Ls		2
	0	0	0
0	0,22	0,21	2
0,22	0,44	0,44	44
0,44	0,67	0,66	65
0,67	0,89	0,88	81
0,89	1,11	1,1	57
1,11	1,33	1,33	15

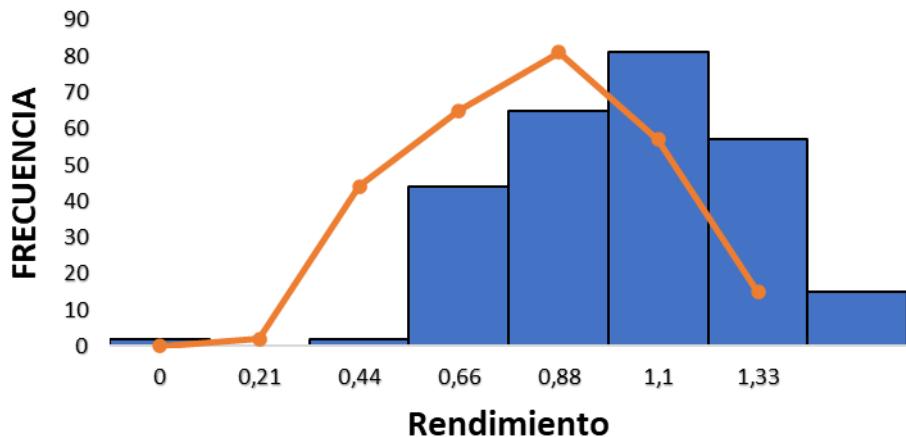
HISTOGRAMA Y POLIGONO DE FRECUENCIA



Clase	Frecuencia	% acumulado
0	2	0,75%
0,21	0	0,75%
0,44	2	1,50%
0,66	44	18,05%
0,88	65	42,48%
1,1	81	72,93%
1,33	57	94,36%
y mayor...	15	100,00%



HISTOGRAMA Y POLIGONO DE FRECUENCIA



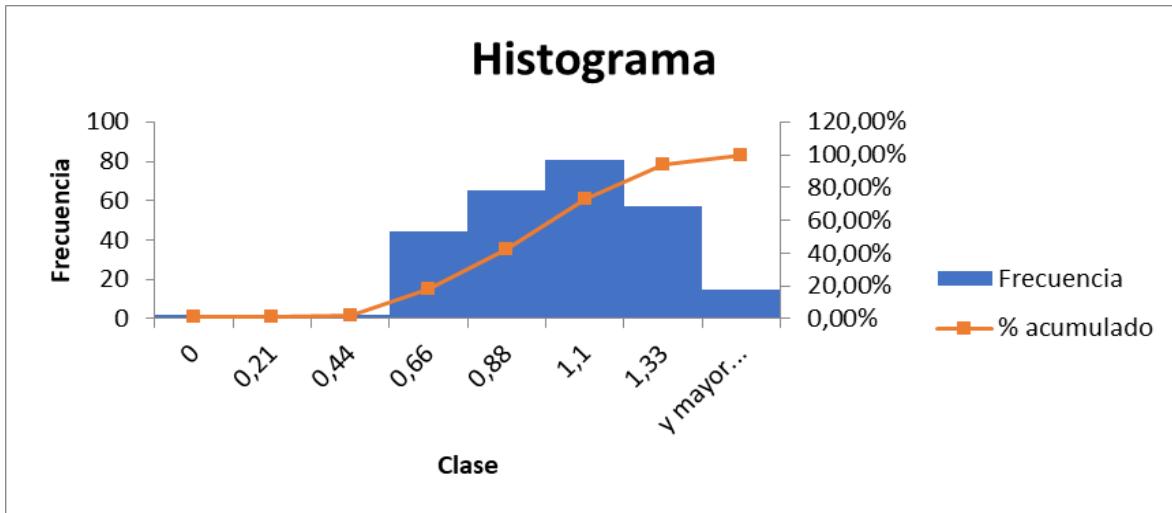
FUENTE: EXCEL

VALOR MAXIMO:

En el polígono de frecuencia, el pico mayor del 0,88 del rendimiento fue de 85 (fr).

VALOR MINIMO:

En el polígono de frecuencia, el menor valor de 0 del área rendimiento fue de 0(fr).



FUENTE: EXCEL

FRECUENCIA:

VALOR MAXIMO.

El valor máximo del 1,1 de clase fue de 85(fr)

VALOR MINIMO.

El valor mínimo del 0 de clase fue de 0(fr)

% ACUMULADO:

VALOR MAXIMO.

El valor máximo del mayor porcentaje acumulado fue de 120.00%

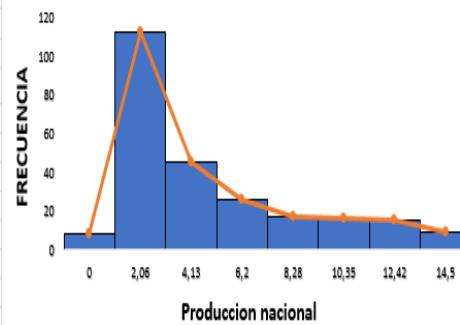
VALOR MINIMO.

El valor mínimo de 0 fue de 0.00%

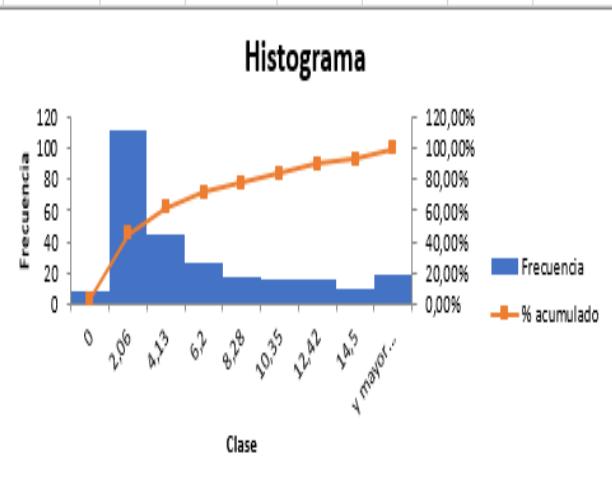
Columna1		Numero de Clases	9
		Tamaño de Clase	2,07148801
Media	4,511315789		
Error típico	0,303538783		
Mediana	2,72		
Moda	0,02		
Desviación estándar	4,950567735		
Varianza de la muestra	24,5081209		
Curtosis	0,1252144		
Coeficiente de	1,107890468		
Rango	18,67		
Mínimo	0		
Máximo	18,67		
Suma	1200,01		
Cuenta	266		

INTERVALOS		GRUPOS	FRECUENCIA
Lí	Ls		
	0	0	8
	0	2,07	2,06
	2,07	4,14	4,13
	4,14	6,21	6,2
	6,21	8,29	8,28
	8,29	10,36	10,35
	10,36	12,43	12,42
	12,43	14,50	14,5
			9

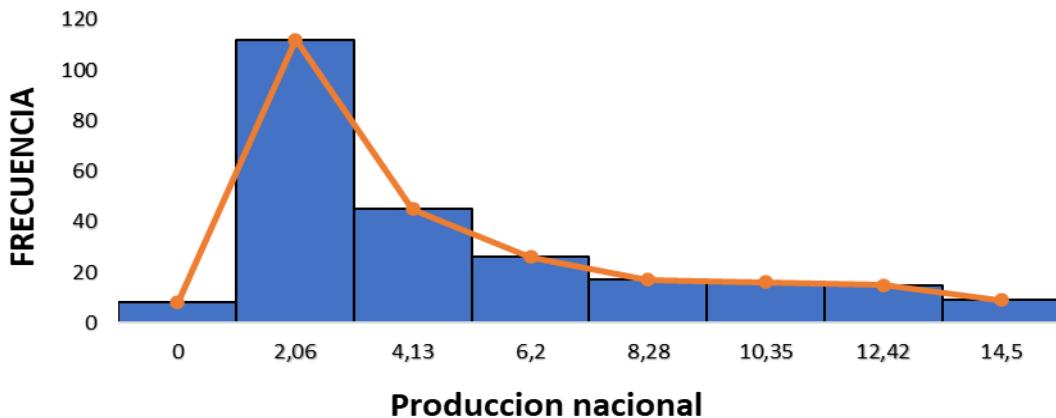
HISTOGRAMA Y POLIGONO DE FRECUENCIA



Clase	Frecuencia	% acumulado
0	8	3,01%
2,06	112	45,11%
4,13	45	62,03%
6,2	26	71,80%
8,28	17	78,20%
10,35	16	84,21%
12,42	15	89,85%
14,5	9	93,23%
y mayor...	18	100,00%



HISTOGRAMA Y POLIGONO DE FRECUENCIA



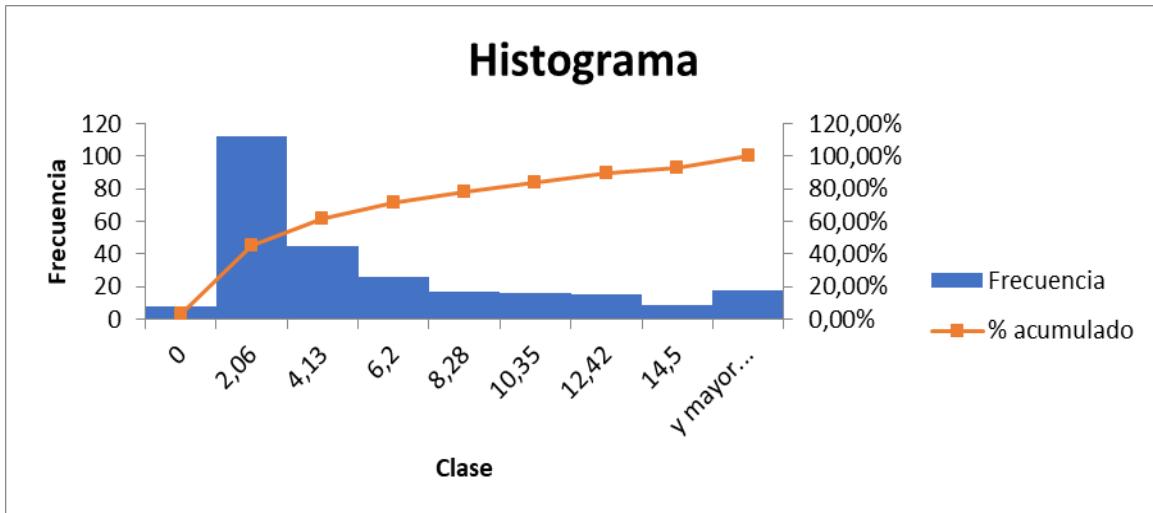
FUENTE: EXCEL

VALOR MAXIMO:

En el polígono de frecuencia, el valor máximo del 2,06 de producción nacional fue de 118(fr)

VALOR MINIMO:

En el polígono de frecuencia, el valor mínimo del 0 de producción nacional fue de 4(fr)



FUENTE: EXCEL

FRECUENCIA:

VALOR MÁXIMO.

En el histograma, el valor máximo del 2,06 de clase fue de 118(fr)

VALOR MÍNIMO.

El valor mínimo del 0 de clase fue de 10(fr)

% ACUMULADO:

VALOR MAXIMO.

El valor máximo del mayor porcentaje acumulado fue de 100.00%

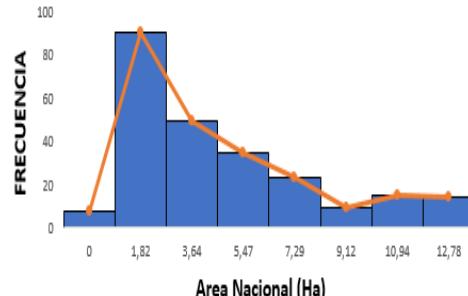
VALOR MINIMO.

El valor mínimo de 0 fue de 0.00%

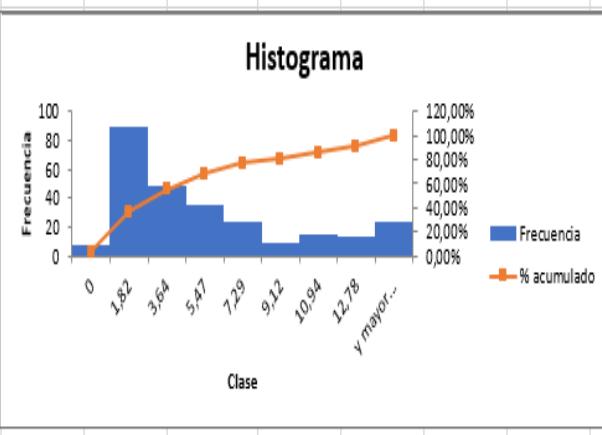
Columna1
Media
4,511203008
Error típico
0,279951139
Mediana
3,12
Moda
0,02
Desviación estándar
4,565864904
Varianza de la muestra
20,84712232
Curtosis
-0,057837939
Coeficiente de
0,992148635
Rango
16,43
Mínimo
0
Máximo
16,43
Suma
1199,98
Cuenta
266

Numero de Clases	9
Tamaño de Clase	1,82512779
INTERVALOS	
<i>Li</i>	<i>Ls</i>
0	0
0	1,82
1,82	3,65
3,65	5,48
5,48	7,30
7,30	9,13
9,13	10,95
10,95	12,78
12,78	12,78
FRECUENCIA	
8	
90	
49	
35	
23	
9	
15	
14	

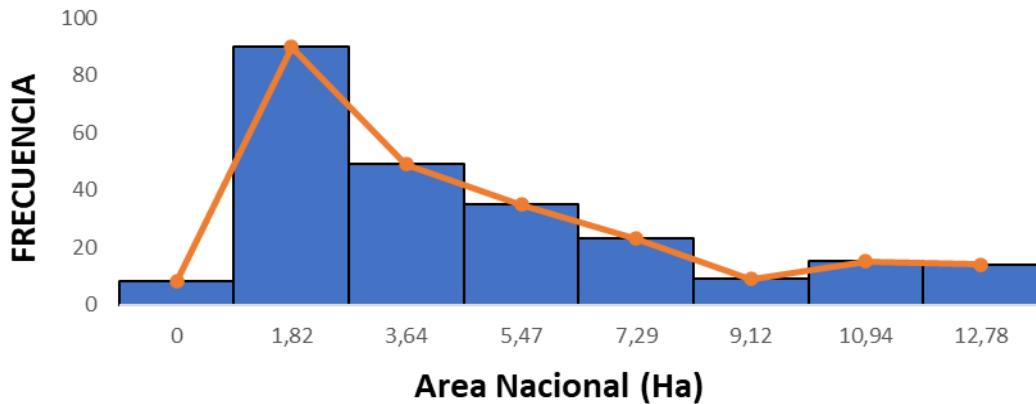
HISTOGRAMA Y POLIGONO DE FRECUENCIA



Clase	Frecuencia	%
0	8	3,01%
1,82	90	36,84%
3,64	49	55,26%
5,47	35	68,42%
7,29	23	77,07%
9,12	9	80,45%
10,94	15	86,09%
12,78	14	91,35%
y mayor...	23	100,00%



HISTOGRAMA Y POLIGONO DE FRECUENCIA



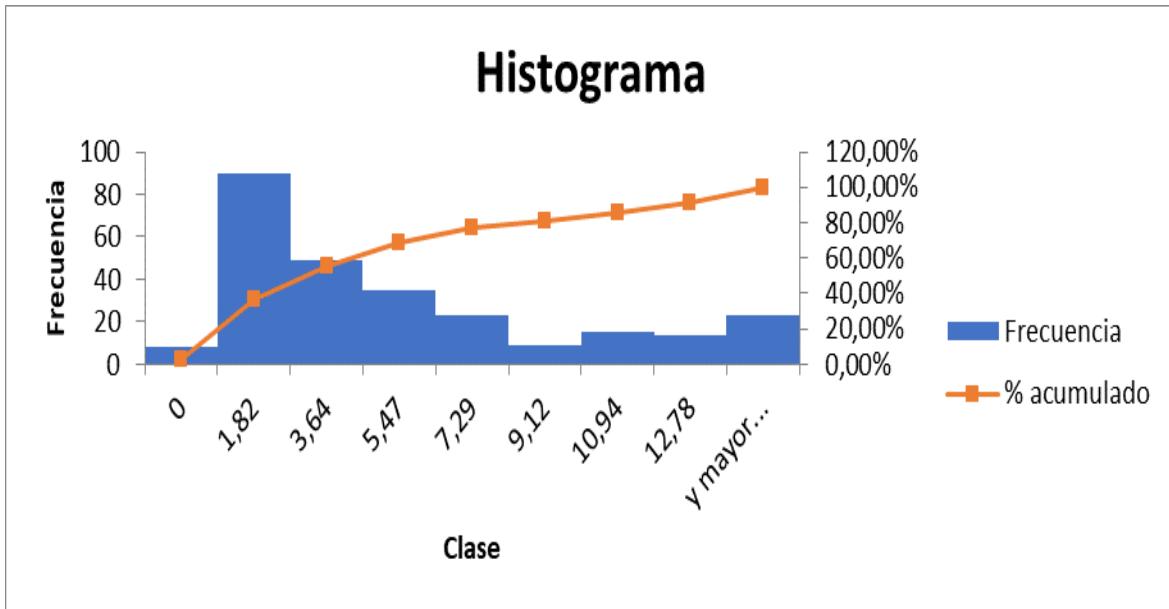
FUENTE: EXCEL

VALOR MAXIMO:

En el polígono de frecuencia, el pico mayor del 1,82 del área nacional fue de 85 (fr).

VALOR MINIMO:

En el polígono de frecuencia, el menor valor de 0 del área nacional fue de 10 (fr).



FUENTE: EXCEL

FRECUENCIA:

VALOR MAXIMO.

El valor máximo del 1,82 de clase fue de 85(fr)

VALOR MINIMO.

El valor mínimo del 0 de clase fue de 9(fr)

% ACUMULADO:

VALOR MAXIMO.

El valor máximo del mayor porcentaje acumulado fue de 100.00%

VALOR MINIMO.

El valor mínimo de 0 fue de 0.00%

ANALISIS EN PYTHON NOTEBOOK

```
In [1]: import pandas as pd  
# Importar la Libreria PANDAS
```

```
In [2]: pd.read_csv("PRODUCCION.csv")  
#Lectura del Dataframe
```

Out[2]:

	Anio	Departamento	Producto	Area (ha)	Produccion (ton)	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
0	2007	ANTIOQUIA	CAFE	112,343.60	120,500.80	1.07	14.54	14.66
1	2007	BOLIVAR	CAFE	502.00	446.00	0.89	0.05	0.07
2	2007	BOYACA	CAFE	11,374.50	9,683.10	0.85	1.17	1.48
3	2007	CALDAS	CAFE	78,393.65	92,815.00	1.18	11.20	10.23
4	2007	CAQUETA	CAFE	2,295.00	2,134.00	0.93	0.26	0.30
...
261	2018	QUINDIO	CAFE	16,374.73	17,739.03	1.08	2.07	2.21
262	2018	RISARALDA	CAFE	35,874.73	45,918.75	1.28	5.37	4.83
263	2018	SANTANDER	CAFE	42,269.07	55,918.71	1.32	6.53	5.69
264	2018	TOLIMA	CAFE	97,304.04	97,451.31	1.00	11.39	13.11
265	2018	VALLE DEL CAUCA	CAFE	48,305.31	49,667.88	1.03	5.80	6.51

266 rows × 8 columns

```
In [5]: produccion_df=pd.read_csv("PRODUCCION.csv")  
# Asignación del nombre del Dataframe
```

29/9/2020

Untitled - Jupyter Notebook

```
In [6]: produccion_df  
# Listado general del Dataframe produccion
```

Out[6]:

	Anio	Departamento	Producto	Area (ha)	Produccion (ton)	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
0	2007	ANTIOQUIA	CAFE	112,343.60	120,500.80	1.07	14.54	14.66
1	2007	BOLIVAR	CAFE	502.00	446.00	0.89	0.05	0.07
2	2007	BOYACA	CAFE	11,374.50	9,683.10	0.85	1.17	1.48
3	2007	CALDAS	CAFE	78,393.65	92,815.00	1.18	11.20	10.23
4	2007	CAQUETA	CAFE	2,295.00	2,134.00	0.93	0.26	0.30
...
261	2018	QUINDIO	CAFE	16,374.73	17,739.03	1.08	2.07	2.21
262	2018	RISARALDA	CAFE	35,874.73	45,918.75	1.28	5.37	4.83
263	2018	SANTANDER	CAFE	42,269.07	55,918.71	1.32	6.53	5.69
264	2018	TOLIMA	CAFE	97,304.04	97,451.31	1.00	11.39	13.11
265	2018	VALLE DEL CAUCA	CAFE	48,305.31	49,667.88	1.03	5.80	6.51

266 rows × 8 columns

```
In [7]: type(produccion_df)
Out[7]: pandas.core.frame.DataFrame

In [8]: produccion_df.dtypes
Out[8]: Anio          int64
Departamento    object
Producto        object
Area (ha)       object
Produccion (ton)   object
Rendimiento (ha/ton) float64
Produccion Nacional (ton) float64
Area Nacional (ha)  float64
dtype: object

In [9]: pd.unique(produccion_df['Anio'])
#indica los valores de los años en el dataframe
Out[9]: array([2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017,
2018], dtype=int64)
```

29/9/2020

Untitled - Jupyter Notebook

```
In [10]: pd.unique(produccion_df['Departamento'])
#indica Los departamentos en el dataframe
Out[10]: array(['ANTIOQUIA', 'BOLIVAR', 'BOYACA', 'CALDAS', 'CAQUETA', 'CASANARE',
'CAUCA', 'CESAR', 'CHOCO', 'CUNDINAMARCA', 'HUILA', 'LA GUAJIRA',
'MAGDALENA', 'META', 'NARIÑO', 'NORTE DE SANTANDER', 'PUTUMAYO',
'QUINDIO', 'RISARALDA', 'SANTANDER', 'TOLIMA', 'VALLE DEL CAUCA',
'ARAUCA', 'GUAVIARE'], dtype=object)

In [11]: pd.unique(produccion_df['Producto'])
#indica Los productos que están en el dataframe
Out[11]: array(['CAFE'], dtype=object)
```

```
In [12]: pd.unique(produccion_df['Area (ha)'])
```

```
Out[12]: array(['112,343.60', '502.00', '11,374.50', '78,393.65', '2,295.00',
       '2,605.00', '53,471.00', '23,172.00', '290.00', '43,017.30',
       '89,661.56', '4,785.00', '17,506.00', '2,048.00', '24,458.50',
       '30,171.84', '35.00', '19,904.00', '47,689.25', '34,406.67',
       '91,679.10', '76,667.80', '114,694.00', '572.00', '10,778.50',
       '74,897.00', '2,735.00', '2,149.00', '56,208.00', '23,198.00',
       '90.00', '43,633.35', '89,131.20', '4,553.00', '17,521.00',
       '2,146.00', '25,582.00', '31.00', '19,571.00', '47,227.00',
       '34,169.37', '86,829.20', '72,419.00', '112,420.20', '770.00',
       '10,672.50', '73,083.00', '2,332.00', '1,904.00', '57,860.00',
       '23,420.00', '70.00', '43,475.84', '86,726.78', '4,488.00',
       '17,036.00', '2,216.00', '26,467.20', '33,552.58', '23.00',
       '19,052.00', '45,428.00', '37,985.90', '88,667.00', '67,001.30',
       '111,602.71', '0.00', '850.00', '9,427.00', '72,240.58',
       '2,536.00', '2,198.00', '55,162.00', '22,489.50', '157.50',
       '44,264.16', '87,139.53', '4,207.00', '17,000.00', '2,326.00',
       '23,504.05', '30,731.96', '24.00', '18,159.00', '47,308.00',
       '39,000.64', '84,658.70', '69,332.10', '106,419.57', '10.00',
       '8,441.74', '66,331.61', '2,810.00', '2,081.50', '54,246.42',
       '22,350.00', '37,478.87', '78,792.21', '4,100.00', '16,577.00',
       '2,578.00', '24,263.80', '21,520.45', '40.00', '20,139.30',
       '44,733.64', '37,282.04', '93,145.35', '68,038.40', '112,221.14',
       '870.00', '6,698.20', '54,871.88', '2,882.50', '2,322.00',
       ...])
```

```
In [13]: pd.unique(produccion_df['Produccion (ton)'])
```

```
Out[13]: array(['120,500.80', '446.00', '9,683.10', '92,815.00', '2,134.00',
       '2,048.40', '51,348.00', '13,278.50', '205.90', '33,729.14',
       '129,052.51', '2,958.70', '14,005.00', '1,617.20', '31,770.05',
       '13,593.24', '34.00', '25,426.00', '72,842.55', '29,469.52',
       '112,322.38', '69,618.24', '113,505.20', '711.00', '9,547.30',
       '86,884.00', '2,469.00', '1,388.13', '48,073.00', '13,841.45',
       '68.00', '78,254.77', '131,316.47', '2,328.90', '14,017.00',
       '1,656.96', '31,262.50', '13,593.25', '35.60', '23,669.00',
       '60,079.00', '29,016.75', '101,201.88', '65,666.43', '103,703.00',
       '292.60', '8,567.97', '81,668.22', '2,332.00', '2,079.70',
       '47,221.00', '12,770.00', '78.75', '37,118.07', '104,609.42',
       '2,340.40', '13,412.80', '1,672.60', '27,487.71', '10,221.69',
       '26.70', '21,985.00', '53,648.00', '26,311.61', '88,633.10',
       '62,711.08', '121,253.38', '0.00', '510.00', '7,083.07',
       '95,957.90', '2,902.50', '2,564.86', '45,113.00', '13,276.08',
       '98.00', '37,214.80', '104,336.56', '2,393.00', '13,600.00',
       '2,221.90', '24,594.10', '22,111.65', '21,065.00', '72,091.00',
       '27,094.16', '94,230.20', '69,496.65', '115,267.98', '12.00',
       '5,643.39', '78,805.87', '2,528.40', '2,023.50', '41,645.39',
       '11,035.85', '32,780.35', '85,150.66', '1,933.00', '13,301.60',
       '2,533.75', '24,073.95', '12,332.00', '45.80', '20,814.11',
       '49,042.31', '22,089.82', '53,288.42', '65,475.63', '91,621.30',
       '652.50', '4,981.59', '54,115.96', '2,446.38', '1,718.25',
       '50,588.14', '19,994.35', '140.00', '30,786.41', '85,212.64',
       '3,434.30', '14,096.05', '2,133.10', '28,077.94', '12,214.54',
       '48.40', '18,030.13', '36,989.43', '23,271.89', '85,027.49',
       '61,190.55', '102,403.24', '395.07', '5,591.05', '58,634.19',
       ...])
```

```
In [14]: pd.unique(produccion_df['Rendimiento (ha/ton)'])
```

```
Out[14]: array([1.07, 0.89, 0.85, 1.18, 0.93, 0.79, 0.96, 0.57, 0.71, 0.78, 1.44,
   0.62, 0.8 , 1.3 , 0.45, 0.97, 1.28, 1.53, 0.86, 1.23, 0.91, 0.99,
   1.24, 1.16, 0.9 , 0.65, 0.6 , 0.76, 1.79, 1.47, 0.51, 0.77, 1.22,
   1.15, 1.21, 1.27, 1.17, 0.92, 0.38, 1.12, 1. , 1.09, 0.82, 0.55,
   1.13, 0.52, 0.75, 1.04, 0.3 , 0.69, 0.94, 0. , 1.33, 1.14, 0.59,
   0.84, 1.2 , 1.05, 0.72, 1.11, 1.52, 1.08, 0.67, 1.19, 0.49, 0.87,
   0.47, 0.98, 1.03, 1.1 , 0.74, 2. , 0.83, 1.01, 0.63, 0.81, 0.88,
   0.66, 0.7 , 1.06, 0.64, 1.02, 0.95, 1.41, 1.32, 1.5 , 1.26, 1.37,
   1.35, 1.25, 1.45, 1.29, 1.4 , 1.38])
```

```
In [15]: pd.unique(produccion_df['Produccion Nacional (ton)'])
```

```
Out[15]: array([1.454e+01, 5.000e-02, 1.170e+00, 1.120e+01, 2.600e-01, 2.500e-01,
   6.190e+00, 1.600e+00, 2.000e-02, 4.070e+00, 1.557e+01, 3.600e-01,
   1.690e+00, 2.000e-01, 3.830e+00, 1.640e+00, 0.000e+00, 3.070e+00,
   8.790e+00, 3.560e+00, 1.355e+01, 8.400e+00, 1.370e+01, 9.000e-02,
   1.150e+00, 1.049e+01, 3.000e-01, 1.700e-01, 5.800e+00, 1.670e+00,
   1.000e-02, 9.440e+00, 1.585e+01, 2.800e-01, 3.770e+00, 2.860e+00,
   7.250e+00, 3.500e+00, 1.221e+01, 7.930e+00, 1.463e+01, 4.000e-02,
   1.210e+00, 1.152e+01, 3.300e-01, 2.900e-01, 6.660e+00, 1.800e+00,
   5.240e+00, 1.476e+01, 1.890e+00, 2.400e-01, 3.880e+00, 1.440e+00,
   3.100e+00, 7.570e+00, 3.710e+00, 1.250e+01, 8.850e+00, 1.556e+01,
   7.000e-02, 9.100e-01, 1.231e+01, 3.700e-01, 5.790e+00, 1.700e+00,
   4.780e+00, 1.339e+01, 3.100e-01, 1.750e+00, 3.160e+00, 2.840e+00,
   2.700e+00, 9.250e+00, 3.480e+00, 1.209e+01, 8.920e+00, 1.800e+01,
   8.000e-02, 8.800e-01, 3.900e-01, 3.200e-01, 6.500e+00, 1.720e+00,
   5.120e+00, 1.330e+01, 2.080e+00, 4.000e-01, 3.760e+00, 1.930e+00,
   3.250e+00, 7.660e+00, 3.450e+00, 8.320e+00, 1.022e+01, 1.462e+01,
   1.000e-01, 7.900e-01, 8.630e+00, 2.700e-01, 8.070e+00, 3.190e+00,
```

```
In [16]: pd.unique(produccion_df['Area Nacional (ha)'])
```

```
Out[16]: array([1.466e+01, 7.000e-02, 1.480e+00, 1.023e+01, 3.000e-01, 3.400e-01,
   6.980e+00, 3.020e+00, 4.000e-02, 5.610e+00, 1.170e+01, 6.200e-01,
   2.280e+00, 2.700e-01, 3.190e+00, 3.940e+00, 0.000e+00, 2.600e+00,
   6.220e+00, 4.490e+00, 1.196e+01, 1.000e+01, 1.513e+01, 8.000e-02,
   1.420e+00, 9.880e+00, 3.600e-01, 2.800e-01, 7.410e+00, 3.060e+00,
   1.000e-02, 5.750e+00, 1.175e+01, 6.000e-01, 2.310e+00, 3.370e+00,
   3.980e+00, 2.580e+00, 6.230e+00, 4.510e+00, 1.145e+01, 9.550e+00,
   1.490e+01, 1.000e-01, 1.410e+00, 9.680e+00, 3.100e-01, 2.500e-01,
   7.670e+00, 3.100e+00, 5.760e+00, 1.149e+01, 5.900e-01, 2.260e+00,
   2.900e-01, 3.510e+00, 4.450e+00, 2.520e+00, 6.020e+00, 5.030e+00,
   8.880e+00, 1.499e+01, 1.100e-01, 1.270e+00, 9.710e+00, 2.000e-02,
   5.950e+00, 1.171e+01, 5.700e-01, 3.160e+00, 4.130e+00, 2.440e+00,
   6.360e+00, 5.240e+00, 1.137e+01, 9.310e+00, 1.494e+01, 1.200e-01,
   1.180e+00, 3.900e-01, 7.610e+00, 3.140e+00, 5.260e+00, 1.106e+01,
   5.800e-01, 2.330e+00, 3.410e+00, 2.830e+00, 6.280e+00, 5.230e+00,
   1.308e+01, 1.580e+01, 9.400e-01, 7.720e+00, 4.100e-01, 3.300e-01,
   8.000e+00, 3.220e+00, 1.123e+01, 7.200e-01, 2.490e+00, 3.910e+00,
   2.720e+00, 2.970e+00, 6.420e+00, 4.780e+00, 1.280e+01, 9.780e+00,
   1.422e+01, 9.000e-02, 1.200e+00, 7.810e+00, 3.800e-01, 9.600e+00,
   3.250e+00, 4.690e+00, 1.531e+01, 7.500e-01, 2.200e+00, 3.200e-01,
   4.160e+00, 3.280e+00, 2.750e+00, 5.130e+00, 5.000e+00, 1.261e+01,
   6.930e+00, 1.384e+01, 1.240e+00, 7.510e+00, 9.690e+00, 3.290e+00,
   4.230e+00, 1.612e+01, 7.600e-01, 4.220e+00, 2.980e+00, 2.700e+00,
   5.050e+00, 5.120e+00, 1.267e+01, 7.040e+00, 1.369e+01, 1.300e-01,
   1.310e+00, 7.290e+00, 4.300e-01, 9.660e+00, 3.240e+00, 4.260e+00,
   1.628e+01, 7.000e-01, 2.250e+00, 4.180e+00, 2.860e+00, 2.680e+00,
   5.210e+00, 5.330e+00, 1.290e+01, 6.860e+00, 1.359e+01, 1.400e-01,
   7.200e+00, 4.400e-01, 1.008e+01, 4.270e+00, 1.621e+01, 7.100e-01,
   4.210e+00, 2.770e+00, 5.200e+00, 5.320e+00, 6.770e+00, 1.318e+01,
   1.500e-01, 6.880e+00, 4.500e-01, 1.066e+01, 3.340e+00, 4.100e+00,
   1.627e+01, 2.410e+00, 4.470e+00, 2.840e+00, 3.000e-02, 2.350e+00,
```

```
In [17]: produccion_df['Anio'].min()
Out[17]: 2007

In [18]: produccion_df['Anio'].max()
Out[18]: 2018

In [19]: produccion_df['Area (ha)'].min()
Out[19]: '0.00'

In [20]: produccion_df['Area (ha)'].max()+" Hectarea"
Out[20]: '99,311.53 Hectarea'
```

29/9/2020

Untitled - Jupyter Notebook

```
In [21]: produccion_df['Rendimiento (ha/ton)'].min()
Out[21]: 0.0
```

```
In [22]: produccion_df['Rendimiento (ha/ton)'].max()
Out[22]: 2.0
```

```
In [23]: produccion_df['Anio'].isnull()
Out[23]: 0      False
         1      False
         2      False
         3      False
         4      False
         ...
        261     False
        262     False
        263     False
        264     False
        265     False
Name: Anio, Length: 266, dtype: bool
```

```
In [24]: produccion_df['Area (ha)'].isnull()
Out[24]: 0      False
         1      False
         2      False
         3      False
         4      False
         ...
        261     False
        262     False
        263     False
        264     False
        265     False
Name: Area (ha), Length: 266, dtype: bool
```

```
In [25]: produccion_df['Rendimiento (ha/ton)'].isnull()

Out[25]: 0      False
1      False
2      False
3      False
4      False
...
261     False
262     False
263     False
264     False
265     False
Name: Rendimiento (ha/ton), Length: 266, dtype: bool
```

29/9/2020

Untitled - Jupyter Notebook

```
In [26]: produccion_df['Rendimiento (ha/ton)'].isnull().sum()
```

```
Out[26]: 0
```

```
In [27]: produccion_df['Area (ha)'].isnull().sum()
```

```
Out[27]: 0
```

```
In [28]: produccion_grouped_Anio=produccion_df.groupby("Anio").sum()
produccion_grouped_Anio
```

```
Out[28]:
          Rendimiento (ha/ton)  Produccion Nacional (ton)  Area Nacional (ha)
Anio
```

Anio	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
2007	20.91	100.01	100.00
2008	21.62	100.00	99.99
2009	19.39	100.00	99.98
2010	20.84	100.01	100.00
2011	19.65	100.02	100.00
2012	19.75	99.99	100.00
2013	16.71	100.00	99.99
2014	18.09	100.00	100.00
2015	22.54	99.98	100.00
2016	22.34	99.99	100.00
2017	23.50	100.01	100.00
2018	23.75	100.00	100.02

29/9/2020

Untitled - Jupyter Notebook

```
In [29]: produccion_grouped_Anio2=produccion_df.groupby("Anio").sum()
produccion_grouped_Anio2
```

Out[29]:

	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
Anio			
2007	20.91	100.01	100.00
2008	21.62	100.00	99.99
2009	19.39	100.00	99.98
2010	20.84	100.01	100.00
2011	19.65	100.02	100.00
2012	19.75	99.99	100.00
2013	16.71	100.00	99.99
2014	18.09	100.00	100.00
2015	22.54	99.98	100.00
2016	22.34	99.99	100.00
2017	23.50	100.01	100.00
2018	23.75	100.00	100.02

```
In [35]: produccion_df['Produccion (ton)'].count()
# cuenta el numero de registros en el dataframe para el campo de la Producción
```

Out[35]: 266

```
In [36]: produccion_df['Anio'].count()
# cuenta el numero de registros en el dataframe para el campo del año
```

Out[36]: 266

29/9/2020

Untitled - Jupyter Notebook

```
In [41]: produccion_Anio=produccion_df.groupby(["Anio"]).describe()
produccion_Anio
```

Out[41]:

	Rendimiento (ha/ton)							Produccion Nacional (ton)						
	count	mean	std	min	25%	50%	75%	max	count	mean	75%	m	
Anio														
2007	22.0	0.950455	0.279566	0.45	0.7900	0.900	1.1525	1.53	22.0	4.545909	...	7.8475	1:	
2008	22.0	0.982727	0.322670	0.45	0.7775	0.905	1.2000	1.79	22.0	4.545455	...	7.7600	1:	
2009	22.0	0.881364	0.264652	0.30	0.7600	0.930	1.1125	1.21	22.0	4.545455	...	7.3425	1:	
2010	23.0	0.906087	0.324692	0.00	0.7050	0.960	1.1250	1.52	23.0	4.348261	...	7.3550	1:	
2011	23.0	0.854348	0.238305	0.47	0.6100	0.900	1.0550	1.20	23.0	4.348696	...	7.0800	1:	
2012	23.0	0.858696	0.329618	0.00	0.7450	0.830	0.9150	2.00	23.0	4.347391	...	6.9850	1:	
2013	22.0	0.759545	0.145421	0.60	0.6000	0.755	0.8800	0.99	22.0	4.545455	...	6.4400	1:	
2014	22.0	0.822273	0.157629	0.64	0.6500	0.815	0.9500	1.06	22.0	4.545455	...	6.5950	1:	
2015	22.0	1.024545	0.110096	0.77	0.9350	1.065	1.1075	1.15	22.0	4.544545	...	6.4675	1:	
2016	21.0	1.063810	0.116725	0.79	0.9600	1.120	1.1500	1.19	21.0	4.761429	...	6.6800	1:	
2017	22.0	1.068182	0.272443	0.66	0.8450	1.090	1.2900	1.50	22.0	4.545909	...	6.3550	1:	
2018	22.0	1.079545	0.296672	0.62	0.8575	1.120	1.3125	1.52	22.0	4.545455	...	6.3475	1:	

12 rows × 24 columns

```
In [43]: produccion_df.describe()  
# Indica datos estadísticos generales del dataframe produccion desde el año 2007
```

Out[43]:

	Anio	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
count	266.000000	266.000000	266.000000	266.000000
mean	2012.469925	0.936429	4.511316	4.511203
std	3.443484	0.267129	4.950568	4.565865
min	2007.000000	0.000000	0.000000	0.000000
25%	2010.000000	0.750000	0.352500	0.390000
50%	2012.000000	0.940000	2.720000	3.120000
75%	2015.000000	1.120000	7.147500	6.875000
max	2018.000000	2.000000	18.670000	16.430000

29/9/2020

Untitled - Jupyter Notebook

```
In [45]: produccion_df["Produccion Nacional (ton)"].describe()  
# Indica datos estadísticos generales para la Producción nacional del dataframe producc
```

Out[45]:

```
count    266.000000  
mean      4.511316  
std       4.950568  
min       0.000000  
25%       0.352500  
50%       2.720000  
75%       7.147500  
max      18.670000  
Name: Produccion Nacional (ton), dtype: float64
```

```
In [46]: produccion_df["Area Nacional (ha)"].describe()  
# Indica datos estadísticos generales para el Area Nacional del dataframe producc
```

Out[46]:

```
count    266.000000  
mean      4.511203  
std       4.565865  
min       0.000000  
25%       0.390000  
50%       3.120000  
75%       6.875000  
max      16.430000  
Name: Area Nacional (ha), dtype: float64
```

```
In [47]: produccion_df.describe()
produccion_df.mean()
# Indica el promedio del dataframe produccion para Rendimiento, Produccion y el Area Nacional
```

```
Out[47]: Anio          2012.469925
Rendimiento (ha/ton)    0.936429
Produccion Nacional (ton) 4.511316
Area Nacional (ha)      4.511203
dtype: float64
```

```
In [49]: produccion_df.duplicated().sum()
#Registros que esten duplicado
```

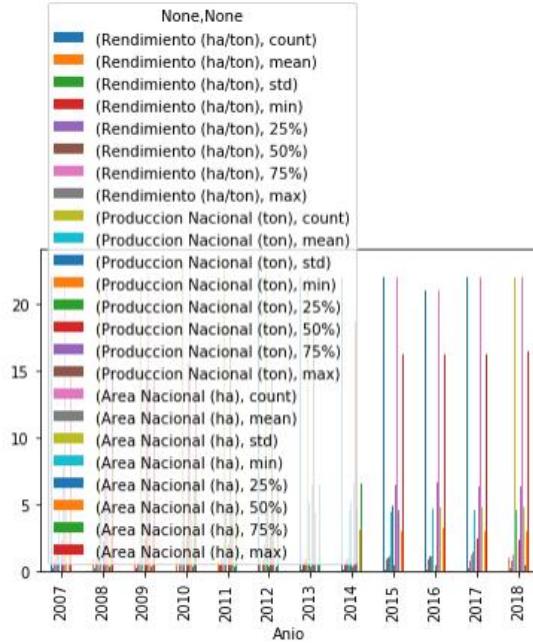
```
Out[49]: 0
```

29/9/2020

Untitled - Jupyter Notebook

```
In [50]: # Construcción del gráfico produccion por año tipo Lineas
import numpy as np
import re
import sys
%matplotlib inline
produccion_Anio.plot(kind='bar')
```

```
Out[50]: <matplotlib.axes._subplots.AxesSubplot at 0x28ad594b948>
```



29/9/2020

Untitled - Jupyter Notebook

```
In [57]: grouped_data = produccion_df.groupby("Departamento")
z=grouped_data.describe().mean()
print (z)
```

```
Anio          count    11.083333
              mean     2012.382576
              std      3.479313
              min     2007.333333
              25%     2009.854167
              50%     2012.375000
              75%     2014.895833
              max     2017.458333
Rendimiento (ha/ton)  count    11.083333
                      mean     0.889467
                      std      0.216119
                      min     0.620833
                      25%     0.769167
                      50%     0.863750
                      75%     0.986771
                      max     1.235417
Produccion Nacional (ton)  count    11.083333
                      mean     4.166733
                      std      0.719931
                      min     3.261250
                      25%     3.687812
                      50%     4.031667
                      75%     4.614271
                      max     5.387500
Area Nacional (ha)        count    11.083333
                      mean     4.166632
                      std      0.511340
                      min     3.537500
                      25%     3.758229
                      50%     4.136042
                      75%     4.588854
                      max     4.838333
dtype: float64
```

29/9/2020

Untitled - Jupyter Notebook

```
In [58]: departamentos_counts = produccion_df.groupby("Departamento")["Producto"].count()
print(departamentos_counts)
# Permite verificar y contar para cada uno de Los Departamentos Las distintas var
# Se encuentra que algunos departamentos tienen otros valores diferentes a Los 12
```

```
Departamento
ANTIOQUIA      12
ARAUCA         2
BOLIVAR        12
BOYACA         12
CALDAS         12
CAQUETA        12
CASANARE       12
CAUCA          12
CESAR           12
CHOCO           12
CUNDINAMARCA   12
GUAVIARE        1
HUILA           12
LA GUAJIRA      12
MAGDALENA       12
META            12
NARIÑO          12
NORTE DE SANTANDER 12
PUTUMAYO        11
QUINDIO         12
RISARALDA       12
SANTANDER       12
TOLIMA          12
VALLE DEL CAUCA 12
Name: Producto, dtype: int64
```

```
In [59]: Grupos_Departamentos=produccion_df.groupby("Anio")["Departamento"].count()
print (Grupos_Departamentos)
# Indica la cantidad de Departamentos incluidos o analizados en cada uno de Los Años
```

Anio	Count
2007	22
2008	22
2009	22
2010	23
2011	23
2012	23
2013	22
2014	22
2015	22
2016	21
2017	22
2018	22

Name: Departamento, dtype: int64

29/9/2020

Untitled - Jupyter Notebook

```
In [60]: Departamento_Meta=produccion_df.loc[produccion_df["Departamento"]=="META"]
print (Departamento_Meta)
# Indica Los resultados estadisticos por año para el Departamento Seleccionado
```

Anio	Departamento	Producto	Area (ha)	Produccion (ton)
13	2007	META	CAFE 2,048.00	1,617.20
35	2008	META	CAFE 2,146.00	1,656.96
57	2009	META	CAFE 2,216.00	1,672.60
80	2010	META	CAFE 2,326.00	2,221.90
103	2011	META	CAFE 2,578.00	2,533.75
126	2012	META	CAFE 2,783.00	2,133.10
148	2013	META	CAFE 2,483.43	1,650.41
170	2014	META	CAFE 2,739.71	1,950.84
192	2015	META	CAFE 2,922.21	3,206.35
214	2016	META	CAFE 2,924.89	3,322.42
235	2017	META	CAFE 2,926.85	4,013.11
257	2018	META	CAFE 2,761.01	3,877.62

Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
13	0.79	0.20
35	0.77	0.20
57	0.75	0.24
80	0.96	0.29
103	0.98	0.40
126	0.77	0.34
148	0.66	0.25
170	0.71	0.27
192	1.10	0.38
214	1.14	0.39
235	1.37	0.47
257	1.40	0.45

29/9/2020

Untitled - Jupyter Notebook

```
In [61]: Departamento_QUINDIO=produccion_df.loc[produccion_df["Departamento"]=="QUINDIO"]
print (Departamento_QUINDIO)
# Indica Los resultados estadisticos por año para el Departamento Seleccionado
```

Anio	Departamento	Producto	Area (ha)	Produccion (ton)	\
17	2007	QUINDIO	CAFE	19,904.00	25,426.00
39	2008	QUINDIO	CAFE	19,571.00	23,669.00
61	2009	QUINDIO	CAFE	19,052.00	21,985.00
84	2010	QUINDIO	CAFE	18,159.00	21,065.00
107	2011	QUINDIO	CAFE	20,139.30	20,814.11
130	2012	QUINDIO	CAFE	21,109.83	18,030.13
152	2013	QUINDIO	CAFE	21,203.03	20,599.27
174	2014	QUINDIO	CAFE	21,462.81	22,518.42
196	2015	QUINDIO	CAFE	21,491.21	24,694.56
217	2016	QUINDIO	CAFE	20,041.70	23,791.30
239	2017	QUINDIO	CAFE	17,699.67	18,792.05
261	2018	QUINDIO	CAFE	16,374.73	17,739.03
		Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)	
17		1.28	3.07	2.60	
39		1.21	2.86	2.58	
61		1.15	3.10	2.52	
84		1.16	2.70	2.44	
107		1.03	3.25	2.83	
130		0.85	2.88	2.97	
152		0.97	3.16	2.75	
174		1.05	3.09	2.70	
196		1.15	2.90	2.68	
217		1.19	2.79	2.58	
239		1.06	2.21	2.35	
261		1.08	2.07	2.21	

29/9/2020

Untitled - Jupyter Notebook

```
In [62]: Estadistica_Anio2015=produccion_df.loc[produccion_df["Anio"]== 2015]
print (Estadistica_Anio2015)
# Indica Los resultados estadisticos por departamento para el año 2015
```

Anio	Departamento	Producto	Area (ha)	Produccion (ton)	\
179	2015	ANTIOQUIA	CAFE	109,649.61	120,365.77
180	2015	BOLIVAR	CAFE	1,065.07	1,089.74
181	2015	BOYACA	CAFE	10,461.85	9,501.54
182	2015	CALDAS	CAFE	58,376.40	67,231.37
183	2015	CAQUETA	CAFE	3,410.56	3,749.27
184	2015	CASANARE	CAFE	2,752.31	2,626.73
185	2015	CAUCA	CAFE	77,405.83	83,626.44
186	2015	CESAR	CAFE	25,948.50	22,240.81
187	2015	CHOCO	CAFE	137.47	158.20
188	2015	CUNDINAMARCA	CAFE	34,101.49	31,165.15
189	2015	HUILA	CAFE	130,452.40	145,168.10
190	2015	LA GUAJIRA	CAFE	5,631.53	4,317.50
191	2015	MAGDALENA	CAFE	17,996.31	16,691.31
192	2015	META	CAFE	2,922.21	3,206.35
193	2015	NARIÑO	CAFE	33,490.93	36,607.56
194	2015	NORTE DE SANTANDER	CAFE	22,940.64	20,267.64
195	2015	PUTUMAYO	CAFE	128.65	124.67
196	2015	QUINDIO	CAFE	21,491.21	24,694.56
197	2015	RISARALDA	CAFE	41,732.03	47,215.69
198	2015	SANTANDER	CAFE	42,679.11	47,304.16
199	2015	TOLIMA	CAFE	103,368.73	105,563.88
200	2015	VALLE DEL CAUCA	CAFE	54,938.79	57,583.56

29/9/2020

Untitled - Jupyter Notebook

```
In [63]: Estadistica_Anio2018=produccion_df.loc[produccion_df["Anio"]== 2018]
print (Estadistica_Anio2018)
# Indica los resultados estadísticos por departamento para el año 2018
```

Anio		Departamento	Producto	Area (ha)	Produccion (ton) \
244	2018	ANTIOQUIA	CAFE	98,038.15	141,898.91
245	2018	BOLIVAR	CAFE	1,182.13	734.91
246	2018	BOYACA	CAFE	9,653.45	7,780.34
247	2018	CALDAS	CAFE	50,762.22	68,670.96
248	2018	CAQUETA	CAFE	3,485.24	5,280.40
249	2018	CASANARE	CAFE	2,360.55	1,629.25
250	2018	CAUCA	CAFE	82,085.54	102,147.00
251	2018	CESAR	CAFE	23,915.45	14,943.62
252	2018	CHOCO	CAFE	140.33	181.42
253	2018	CUNDINAMARCA	CAFE	29,085.24	32,580.24
254	2018	HUILA	CAFE	122,002.46	136,161.86
255	2018	LA GUAJIRA	CAFE	4,810.97	2,990.91
256	2018	MAGDALENA	CAFE	17,414.32	10,826.24
257	2018	META	CAFE	2,761.01	3,877.62
258	2018	NARIÑO	CAFE	33,465.54	35,679.42
259	2018	NORTE DE SANTANDER	CAFE	20,873.04	23,471.69
260	2018	PUTUMAYO	CAFE	209.93	289.50
261	2018	QUINDIO	CAFE	16,374.73	17,739.03
262	2018	RISARALDA	CAFE	35,874.73	45,918.75
263	2018	SANTANDER	CAFE	42,269.07	55,918.71
264	2018	TOLIMA	CAFE	97,304.04	97,451.31
265	2018	VALLE DEL CAUCA	CAFE	48,305.31	49,667.88

	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
244	1.45	16.58	13.21
245	0.62	0.09	0.16
246	0.81	0.91	1.30
247	1.35	8.02	6.84
248	1.52	0.62	0.47
249	0.69	0.19	0.32
250	1.24	11.94	11.06
251	0.62	1.75	3.22
252	1.29	0.02	0.02
253	1.12	3.81	3.92
254	1.12	15.91	16.43
255	0.62	0.35	0.65
256	0.62	1.26	2.35
257	1.40	0.45	0.37
258	1.07	4.17	4.51
259	1.12	2.74	2.81
260	1.38	0.03	0.03
261	1.08	2.07	2.21
262	1.28	5.37	4.83
263	1.32	6.53	5.69
264	1.00	11.39	13.11
265	1.03	5.80	6.51

29/9/2020

Untitled - Jupyter Notebook

In [64]: `produccion_df[0:10]`
#Lista los primeros 10 elementos del dataframe

Out[64]:

	Anio	Departamento	Producto	Area (ha)	Produccion (ton)	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
0	2007	ANTIOQUIA	CAFE	112,343.60	120,500.80	1.07	14.54	14.66
1	2007	BOLIVAR	CAFE	502.00	446.00	0.89	0.05	0.07
2	2007	BOYACA	CAFE	11,374.50	9,683.10	0.85	1.17	1.48
3	2007	CALDAS	CAFE	78,393.65	92,815.00	1.18	11.20	10.23
4	2007	CAQUETA	CAFE	2,295.00	2,134.00	0.93	0.26	0.30
5	2007	CASANARE	CAFE	2,605.00	2,048.40	0.79	0.25	0.34
6	2007	CAUCA	CAFE	53,471.00	51,348.00	0.96	6.19	6.98
7	2007	CESAR	CAFE	23,172.00	13,278.50	0.57	1.60	3.02
8	2007	CHOCO	CAFE	290.00	205.90	0.71	0.02	0.04
9	2007	CUNDINAMARCA	CAFE	43,017.30	33,729.14	0.78	4.07	5.61

29/9/2020

Untitled - Jupyter Notebook

In [65]: `produccion_df[11:30]`
#Lista los elementos desde el 11 al 30 del dataframe

Out[65]:

	Anio	Departamento	Producto	Area (ha)	Produccion (ton)	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
11	2007	LA GUAJIRA	CAFE	4,785.00	2,958.70	0.62	0.36	0.62
12	2007	MAGDALENA	CAFE	17,506.00	14,005.00	0.80	1.69	2.28
13	2007	META	CAFE	2,048.00	1,617.20	0.79	0.20	0.27
14	2007	NARIÑO	CAFE	24,458.50	31,770.05	1.30	3.83	3.19
15	2007	NORTE DE SANTANDER	CAFE	30,171.84	13,593.24	0.45	1.64	3.94
16	2007	PUTUMAYO	CAFE	35.00	34.00	0.97	0.00	0.00
17	2007	QUINDIO	CAFE	19,904.00	25,426.00	1.28	3.07	2.60
18	2007	RISARALDA	CAFE	47,689.25	72,842.55	1.53	8.79	6.22
19	2007	SANTANDER	CAFE	34,406.67	29,469.52	0.86	3.56	4.49
20	2007	TOLIMA	CAFE	91,679.10	112,322.38	1.23	13.55	11.96
21	2007	VALLE DEL CAUCA	CAFE	76,667.80	69,618.24	0.91	8.40	10.00
22	2008	ANTIOQUIA	CAFE	114,694.00	113,505.20	0.99	13.70	15.13
23	2008	BOLIVAR	CAFE	572.00	711.00	1.24	0.09	0.08
24	2008	BOYACA	CAFE	10,778.50	9,547.30	0.89	1.15	1.42
25	2008	CALDAS	CAFE	74,897.00	86,884.00	1.16	10.49	9.88
26	2008	CAQUETA	CAFE	2,735.00	2,469.00	0.90	0.30	0.36
27	2008	CASANARE	CAFE	2,149.00	1,388.13	0.65	0.17	0.28
28	2008	CAUCA	CAFE	56,208.00	48,073.00	0.86	5.80	7.41
29	2008	CESAR	CAFE	23,198.00	13,841.45	0.60	1.67	3.06

```
In [71]: #Forma del DataFrame
print('Forma del DataFrame:')
print(produccion_df.shape)
print()
```

```
Forma del DataFrame:
(266, 8)
```

29/9/2020

Untitled - Jupyter Notebook

```
In [70]: #Correlación del DataFrame
print('Correlación del DataFrame')
print(produccion_df.corr())
```

```
Correlación del DataFrame:
          Anio  Rendimiento (ha/ton) \
Anio      1.000000             0.173474
Rendimiento (ha/ton)    0.173474             1.000000
Produccion Nacional (ton)  0.007957             0.385570
Area Nacional (ha)       0.008715             0.280677

                                         Produccion Nacional (ton)  Area Nacional (ha)
Anio                               0.007957             0.008715
Rendimiento (ha/ton)                 0.385570             0.280677
Produccion Nacional (ton)            1.000000             0.978409
Area Nacional (ha)                  0.978409             1.000000
```

```
In [72]: #Desviación estándar de cada columna del DataFrame
print('Desviación estándar de la columna del DataFrame:')
print(produccion_df.std())
```

```
Desviación estándar de la columna del DataFrame:
Anio              3.443484
Rendimiento (ha/ton)   0.267129
Produccion Nacional (ton)  4.950568
Area Nacional (ha)        4.565865
dtype: float64
```

29/9/2020

Untitled - Jupyter Notebook

```
In [73]: #Reemplaza los valores perdidos por La media  
print(produccion_df.fillna(produccion_df.mean ()))
```

	Anio	Departamento	Producto	Area (ha)	Produccion (ton)	\
0	2007	ANTIOQUIA	CAFE	112,343.60	120,500.80	
1	2007	BOLIVAR	CAFE	502.00	446.00	
2	2007	BOYACA	CAFE	11,374.50	9,683.10	
3	2007	CALDAS	CAFE	78,393.65	92,815.00	
4	2007	CAQUETA	CAFE	2,295.00	2,134.00	
..
261	2018	QUINDIO	CAFE	16,374.73	17,739.03	
262	2018	RISARALDA	CAFE	35,874.73	45,918.75	
263	2018	SANTANDER	CAFE	42,269.07	55,918.71	
264	2018	TOLIMA	CAFE	97,304.04	97,451.31	
265	2018	VALLE DEL CAUCA	CAFE	48,305.31	49,667.88	
		Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)		
0		1.07	14.54	14.66		
1		0.89	0.05	0.07		
2		0.85	1.17	1.48		
3		1.18	11.20	10.23		
4		0.93	0.26	0.30		
..			
261		1.08	2.07	2.21		
262		1.28	5.37	4.83		
263		1.32	6.53	5.69		
264		1.00	11.39	13.11		
265		1.03	5.80	6.51		

[266 rows x 8 columns]

```
In [74]: #Suma de datos nulos en el DataFrame  
print(produccion_df.isnull().sum())
```

Anio	0
Departamento	0
Producto	0
Area (ha)	0
Produccion (ton)	0
Rendimiento (ha/ton)	0
Produccion Nacional (ton)	0
Area Nacional (ha)	0
dtype: int64	

29/9/2020

Untitled - Jupyter Notebook

In [75]: `#Verificar si hay datos nulos en el DataFrame
print(produccion_df.isnull())`

```
Anio Departamento Producto Area (ha) Produccion (ton) \
0 False False False False False
1 False False False False False
2 False False False False False
3 False False False False False
4 False False False False False
.. ...
261 False False False False False
262 False False False False False
263 False False False False False
264 False False False False False
265 False False False False False

Rendimiento (ha/ton) Produccion Nacional (ton) Area Nacional (ha)
0 False False False
1 False False False
2 False False False
3 False False False
4 False False False
.. ...
261 False False False
262 False False False
263 False False False
264 False False False
265 False False False
```

[266 rows x 8 columns]

In [76]: `import matplotlib.pyplot as plt`

In [77]: `plt.close('Produccion Nacional (ton)')`

In [81]: `ts = pd.Series(np.random.randn(2000),
index=pd.date_range('3/3/2010', periods=2000))`

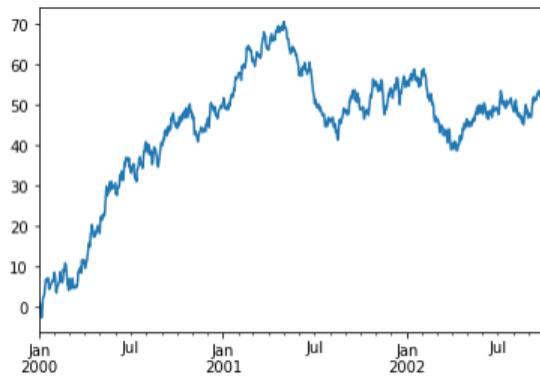
In [82]: `ts = ts.cumsum()`

29/9/2020

Untitled - Jupyter Notebook

In [80]: `ts.plot()`

Out[80]: <matplotlib.axes._subplots.AxesSubplot at 0x28ad5e1d788>



Length

Max length	10
Median length	9
Mean length	8.409774436
Min length	4

Produccion (ton)

Categorical

HIGH

CARDINALITY

UNIFORM

Distinct	262
Distinct (%)	98.5%
Missing	0
Missing (%)	0.0%
Memory size	2.1 KiB

Value	Count	Frequency (%)
510.00	2	0.8%
26.70	2	0.8%
0.00	2	0.8%
---	-	- - -

Overview

Dataset statistics

Number of variables	8
Number of observations	266
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	16.8 KiB
Average record size in memory	64.5 B

Variable types

NUM	4
CAT	4

Warnings

Producto has constant value "266"	Constant
Area (ha) has a high cardinality: 261 distinct values	High cardinality
Produccion (ton) has a high cardinality: 262 distinct values	High cardinality
Area Nacional (ha) is highly correlated with Produccion	High correlation
...	

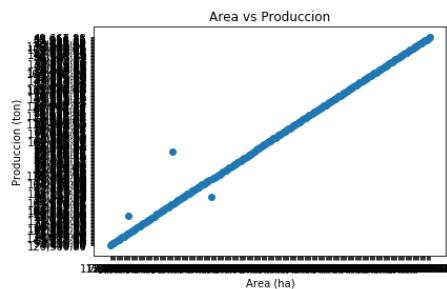
```
In [36]: import numpy as np # libreria para calculos
import matplotlib.pyplot as plt
#%matplotlib inline
import seaborn as sns #Esta libreria permite construir graficos muy particulares
#si se requiere se puede Definir un indice para listar la informacion del datafra
# por ejemplo - -->produccion_df=produccion_df.set_index('Departamento')
# pero en este ejemplo no se emplea el indice
produccion_df.head()
```

```
Out[36]:
```

	Anio	Departamento	Producto	Area (ha)	Produccion (ton)	Rendimiento (ton/ha)	Produccion Nacional (ton)	Area Nacional (ha)
0	2007	ANTIOQUIA	CAFE	112,343.60	120,500.80	1.07	14.54	14.66
1	2007	BOLIVAR	CAFE	502.00	446.00	0.89	0.05	0.07
2	2007	BOYACA	CAFE	11,374.50	9,683.10	0.85	1.17	1.48
3	2007	CALDAS	CAFE	78,393.65	92,815.00	1.18	11.20	10.23
4	2007	CAQUETA	CAFE	2,295.00	2,134.00	0.93	0.26	0.30

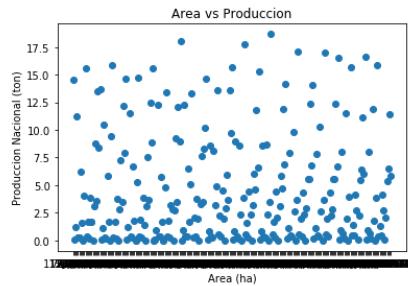
```
In [37]: # Gráfico del comportamiento del Area versus Produccion.
plt.scatter(producción_df['Area (ha)'],producción_df['Producción (ton)'])
plt.title('Área vs Producción')
plt.xlabel('Área (ha)')
plt.ylabel("Producción (ton)")
```

```
Out[37]: Text(0, 0.5, 'Producción (ton)')
```



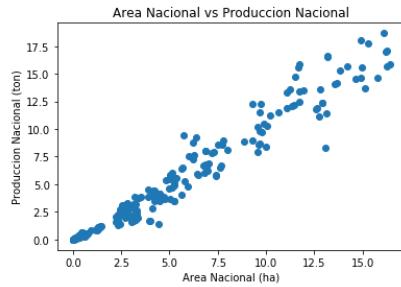
```
In [46]: # Gráfico del comportamiento del Área versus Producción Nacional
plt.scatter(produccion_df['Área (ha)'],produccion_df['Producción Nacional (ton)'])
plt.title('Área vs Producción')
plt.xlabel('Área (ha)')
plt.ylabel("Producción Nacional (ton)")
```

```
Out[46]: Text(0, 0.5, 'Producción Nacional (ton)')
```



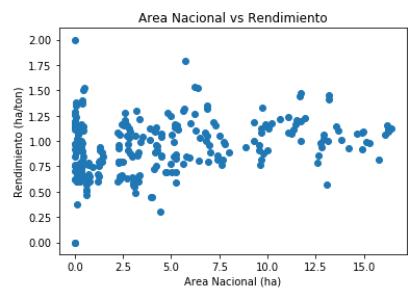
```
In [45]: # Gráfico del comportamiento del Área Nacional versus Producción Nacional
plt.scatter(produccion_df['Área Nacional (ha)'],produccion_df['Producción Nacional (ton)'])
plt.title('Área Nacional vs Producción Nacional')
plt.xlabel('Área Nacional (ha)')
plt.ylabel("Producción Nacional (ton)")
```

```
Out[45]: Text(0, 0.5, 'Producción Nacional (ton)')
```



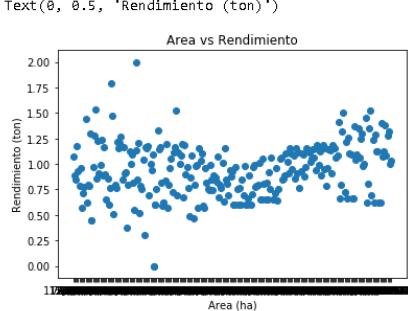
```
In [47]: # Gráfico del comportamiento del Área Nacional versus Rendimiento
plt.scatter(produccion_df['Área Nacional (ha)'],produccion_df['Rendimiento (ha/ton)'])
plt.title('Área Nacional vs Rendimiento')
plt.xlabel('Área Nacional (ha)')
plt.ylabel("Rendimiento (ha/ton)")
```

```
Out[47]: Text(0, 0.5, 'Rendimiento (ha/ton)')
```



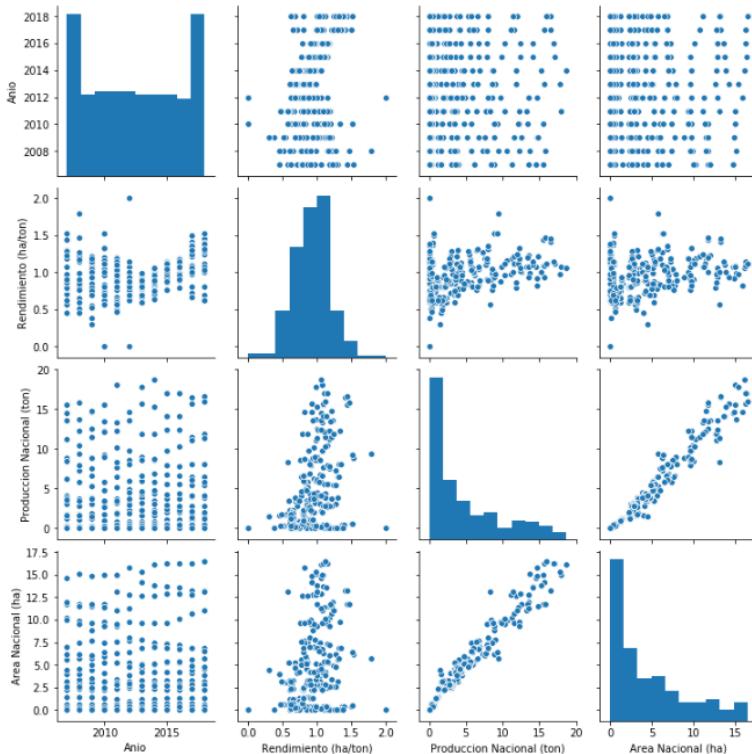
```
In [44]: # Gráfico del comportamiento del Área versus Rendimiento
plt.scatter(produccion_df['Área (ha)'],producción_df['Rendimiento (ha/ton)'])
plt.title('Área vs Rendimiento')
plt.xlabel('Área (ha)')
plt.ylabel("Rendimiento (ton)")

Out[44]: Text(0, 0.5, 'Rendimiento (ton)')
```



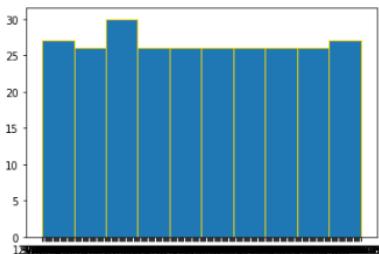
```
In [48]: import seaborn as sns
sns.pairplot(producción_df)

Out[48]: <seaborn.axisgrid.PairGrid at 0x10846410>
```



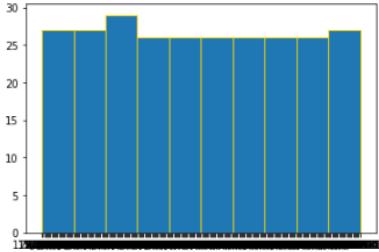
```
In [50]: # Histograma de la Producción de Café  
plt.hist(produccion_df['Producción (ton)'], edgecolor='gold', linewidth=1)
```

```
Out[50]: (array([27., 26., 30., 26., 26., 26., 26., 26., 26., 27.]),  
 array([ 0., 26.1, 52.2, 78.3, 104.4, 130.5, 156.6, 182.7, 208.8,  
 234.9, 261.]),  
<a list of 10 Patch objects>)
```



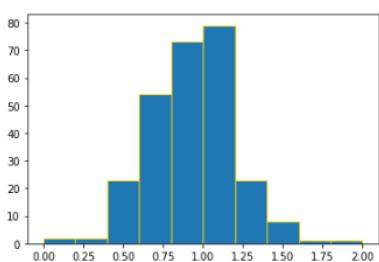
```
In [52]: # Histograma del Área sembrada  
plt.hist(produccion_df['Área (ha)'], edgecolor='gold', linewidth=1)
```

```
Out[52]: (array([27., 27., 29., 26., 26., 26., 26., 26., 27.]),  
 array([ 0., 26., 52., 78., 104., 130., 156., 182., 208., 234., 260.]),  
<a list of 10 Patch objects>)
```

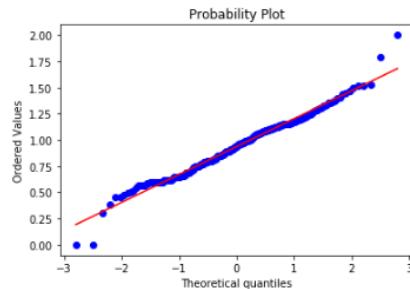


```
In [53]: # Histograma del rendimiento del Café  
plt.hist(produccion_df['Rendimiento (ha/ton)'], edgecolor='gold', linewidth=1)
```

```
Out[53]: (array([ 2., 2., 23., 54., 73., 79., 23., 8., 1., 1.]),  
 array([ 0., 0.2, 0.4, 0.6, 0.8, 1., 1.2, 1.4, 1.6, 1.8, 2.]),  
<a list of 10 Patch objects>)
```



```
In [58]: #librerias para construir estos tipos de graficos
import pylab
import scipy.stats as stats #librerias para construir estos tipos de graficos
stats.probplot(produccion_df['Rendimiento (ha/ton)'], dist= 'norm', plot = pylab)
pylab.show()
```



```
In [59]: # importar la libreria shapiro para realizar el TEST DE SHAPIRO WILK,
# el test de Shapiro Wilk CONFIRMA EFECTIVAMENTE la correlacion entre las variables
from scipy.stats import shapiro
estadistico,p_value = shapiro(produccion_df['Rendimiento (ha/ton)'])
print('Estadistica=%f, El Valor de: p_value=%f' % (estadistico,p_value))
# Si el valor entregado en la variable P_VALUE es MENOR a 0.05,
# indica que si existe una distribucion normal y correlacion
# entre las variables
```

Estadística=0.983, El Valor de: p_value=0.003

```
In [60]: # valores correlacion de Spearman
import numpy as np
produccion_correlacion_spearman = produccion_df.corr(method='spearman')
produccion_correlacion_spearman
# los valores del COEFICIENTE DE SPEARMAN cercanos a cero o inferiores a (+)(-0.4)
# indica que las variables no tienen correlacion
# los valores del COEFICIENTE DE SPEARMAN mayores a (+)(-0.4)
```

Out[60]:

	Anio	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
Anio	1.000000	0.180205	0.037725	0.023246
Rendimiento (ha/ton)	0.180205	1.000000	0.366952	0.264041
Produccion Nacional (ton)	0.037725	0.366952	1.000000	0.986380
Area Nacional (ha)	0.023246	0.264041	0.986380	1.000000

```
In [61]: # valores correlacion de Pearson
import numpy as np
produccion_correlacion_pearson = produccion_df.corr(method='pearson')
produccion_correlacion_pearson
```

Out[61]:

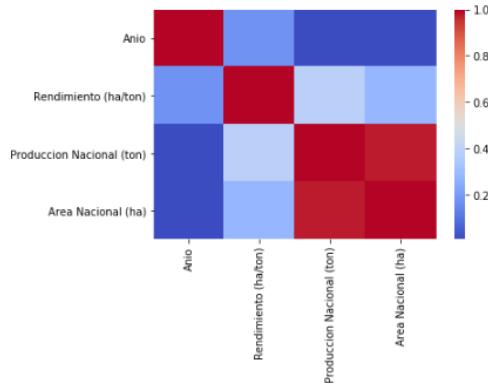
	Anio	Rendimiento (ha/ton)	Produccion Nacional (ton)	Area Nacional (ha)
Anio	1.000000	0.173474	0.007957	0.008715
Rendimiento (ha/ton)	0.173474	1.000000	0.385570	0.280677
Produccion Nacional (ton)	0.007957	0.385570	1.000000	0.978409
Area Nacional (ha)	0.008715	0.280677	0.978409	1.000000

```
In [62]: # valores correlacion de Kendall
import numpy as np
produccion_correlacion_kendall = produccion_df.corr(method='kendall')
produccion_correlacion_kendall
```

```
Out[62]:
      Anio  Rendimiento (ha/ton)  Produccion Nacional (ton)  Area Nacional (ha)
Anio  1.000000          0.140836          0.026879        0.016567
Rendimiento (ha/ton)  0.140836          1.000000          0.265165        0.186979
Produccion Nacional (ton)  0.026879          0.265165          1.000000        0.909233
Area Nacional (ha)  0.016567          0.186979          0.909233        1.000000
```

```
In [64]: # Generacion de mapa de calor para observar fácilmente las variables correlacionadas
# las rojas son correlaciones fuertes positivas y las azules correlaciones negativas
import seaborn as sns # esta librería permite crear gráficos estadísticos
sns.heatmap(produccion_correlacion_pearson,
            xticklabels=produccion_correlacion_pearson.columns,
            yticklabels=produccion_correlacion_pearson.columns,
            cmap='coolwarm'
)
```

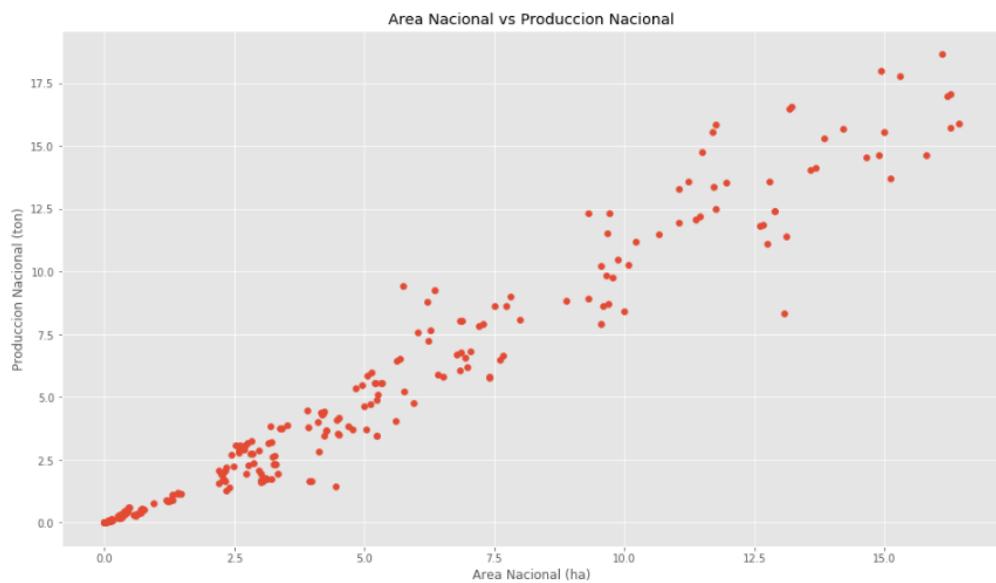
```
Out[64]: <matplotlib.axes._subplots.AxesSubplot at 0x13040f30>
```



```
In [65]: # Imports necesarios
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
%matplotlib inline
from mpl_toolkits.mplot3d import Axes3D
from matplotlib import cm
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score
```

```
In [67]: # Gráfico de dispersión del comportamiento del Área Nacional versus Producción Nacional
plt.scatter(produccion_df['Área Nacional (ha)'],produccion_df['Producción Nacional (ton)'])
plt.title('Área Nacional vs Producción Nacional')
plt.xlabel('Área Nacional (ha)')
plt.ylabel("Producción Nacional (ton)")

Out[67]: Text(0, 0.5, 'Producción Nacional (ton)')
```



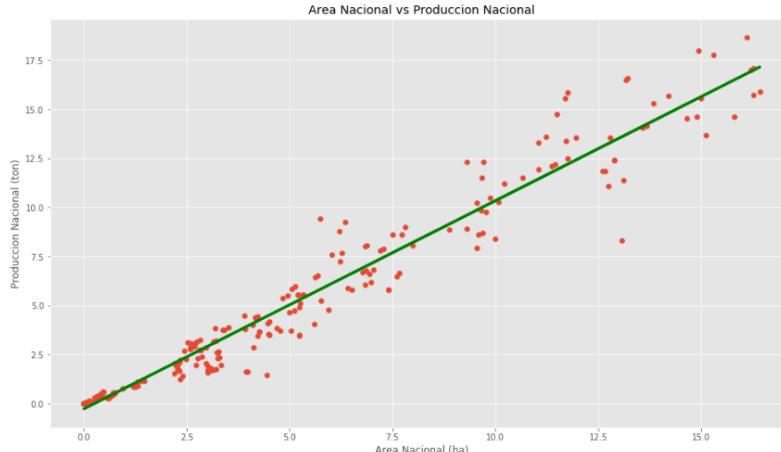
```
In [68]: dataX = produccion_df[['Área Nacional (ha)']]
X_train = np.array(dataX)
y_train = produccion_df['Producción Nacional (ton)'].values
```

```
In [72]: # Creamos la función objeto para determinar la Regresión Lineal Y= mX+b
regr = linear_model.LinearRegression()
# Entrenamos nuestro modelo de regresión lineal, con la siguiente función
regr.fit(X_train, y_train)
# Hacemos las predicciones según el modelo de regresión lineal
y_pred = regr.predict(X_train)
# Ahora imprimimos los resultados obtenidos
# Vemos el valor de la pendiente, o sea la variable m, el coeficiente de la variable
print("Valor de la pendiente (m) o Coefficients:===== > ", regr.coef_)
# Ahora el valor de la constante b0, es decir el valor donde la recta corta el eje
print("Valor de la constante o Independent term: =====> ", regr.intercept_)
```

```
Valor de la pendiente (m) o Coefficients:===== > [ 1.06084584]
Valor de la constante o Independent term: =====> -0.2743751434833568
```

```
In [86]: plt.scatter(produccion_df['Área Nacional (ha)'],producción_df['Producción Nacional (ton)'])
plt.title('Área Nacional vs Producción Nacional')
plt.xlabel('Área Nacional (ha)')
plt.ylabel("Producción Nacional (ton)")

# A continuación se grafica en color azul, la función lineal obtenida a partir del
plt.plot(X_train[:,0], y_pred, color='green', linewidth=3)
plt.show()
```



```
In [92]: # Ahora vamos a predecir utilizando La función obtenida, La producción nacional ( 
# Queremos predecir cuántos toneladas de producción nacional de Café vamos a obtener
# según nuestro modelo, hacemos:
produccion_obtenida = regr.predict([[2]])
print('Estimación de la Producción Nacional del Café en toneladas==>%.3f' %produ
```

Estimación de la Producción Nacional del Café en toneladas==>1.847

```
In [93]: # Ahora vamos a predecir utilizando La función obtenida, La producción nacional ( 
# Queremos predecir cuántos toneladas de producción nacional de Café vamos a obtener
# según nuestro modelo, hacemos:
produccion_obtenida = regr.predict([[2.5]])
print('Estimación de la Producción Nacional del Café en toneladas==>%.3f' %produ
```

Estimación de la Producción Nacional del Café en toneladas==>2.378

```
In [94]: # Ahora vamos a predecir utilizando La función obtenida, La producción nacional ( 
# Queremos predecir cuántos toneladas de producción nacional de Café vamos a obtener
# según nuestro modelo, hacemos:
produccion_obtenida = regr.predict([[8]])
print('Estimación de la Producción Nacional del Café en toneladas==>%.3f' %produ
```

Estimación de la Producción Nacional del Café en toneladas==>8.212

```
In [95]: # Ahora vamos a predecir utilizando La función obtenida, La producción nacional ( 
# Queremos predecir cuántos toneladas de producción nacional de Café vamos a obtener
# según nuestro modelo, hacemos:
produccion_obtenida = regr.predict([[11]])
print('Estimación de la Producción Nacional del Café en toneladas==>%.3f' %produ
```

Estimación de la Producción Nacional del Café en toneladas==>11.395

```
In [96]: # Ahora vamos a predecir utilizando La función obtenida, La producción nacional ( 
# Queremos predecir cuántos toneladas de producción nacional de Café vamos a obtener
# según nuestro modelo, hacemos:
produccion_obtenida = regr.predict([[15]])
print('Estimación de la Producción Nacional del Café en toneladas==>%.3f' %produ
```

Estimación de la Producción Nacional del Café en toneladas==>15.638

```
In [97]: # Ahora vamos a predecir utilizando La función obtenida, La producción nacional ( 
# Queremos predecir cuántos toneladas de producción nacional de Café vamos a obtener
# según nuestro modelo, hacemos:
produccion_obtenida = regr.predict([[35]])
print('Estimación de la Producción Nacional del Café en toneladas==>%.3f' %produ
```

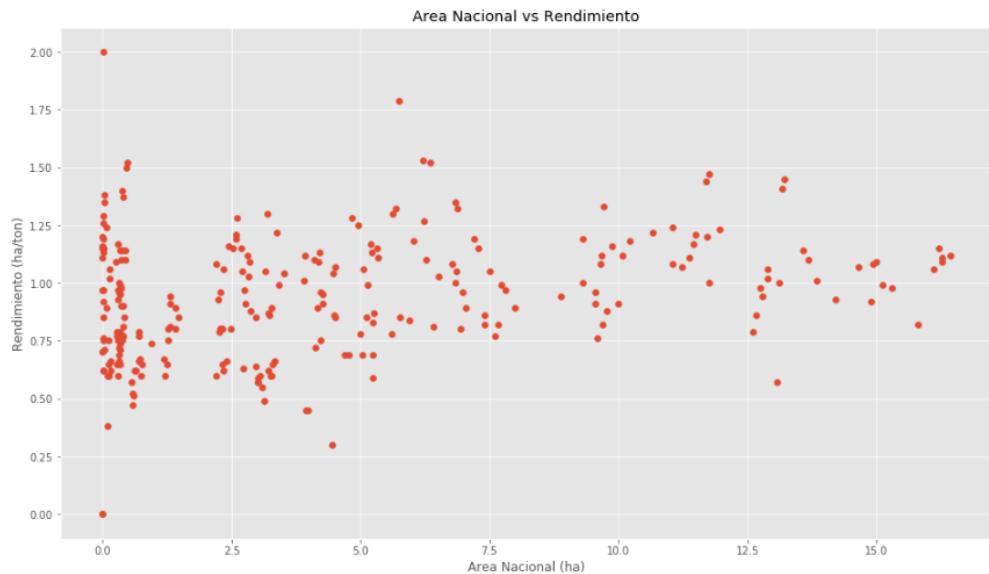
Estimación de la Producción Nacional del Café en toneladas==>36.855

```
In [98]: # AHORA, VAMOS A CONSTRUIR NUEVOS MODELOS PREDICTIVOS UTILIZANDO EL METODO DE RE
```

```
In [105]: # Así, de esta manera, ya conociendo el proceso de analítica determinística, podemos
# Iniciamos el proceso para determinar el modelo de regresión Lineal
# Asignamos a nuestra variable de entrada X (En este caso corresponde al Área Nacional)
# Asignamos a la variable dependiente Y (En este caso corresponde a Rendimiento)
dataX = producción_df[['Área Nacional (ha)']]
X_train = np.array(dataX)
y_train = producción_df['Rendimiento (ha/ton)'].values
```

```
In [79]: plt.scatter(producción_df['Área Nacional (ha)'], producción_df['Rendimiento (ha/ton)'])
plt.title('Área Nacional vs Rendimiento')
plt.xlabel('Área Nacional (ha)')
plt.ylabel('Rendimiento (ha/ton)')
```

```
Out[79]: Text(0, 0.5, 'Rendimiento (ha/ton)')
```



```
In [107]: # Creamos la función objeto para determinar La Regresión Lineal Y= mX+b
regr = linear_model.LinearRegression()

# Entrenamos nuestro modelo de regresión Lineal, con la siguiente función
regr.fit(X_train, y_train)

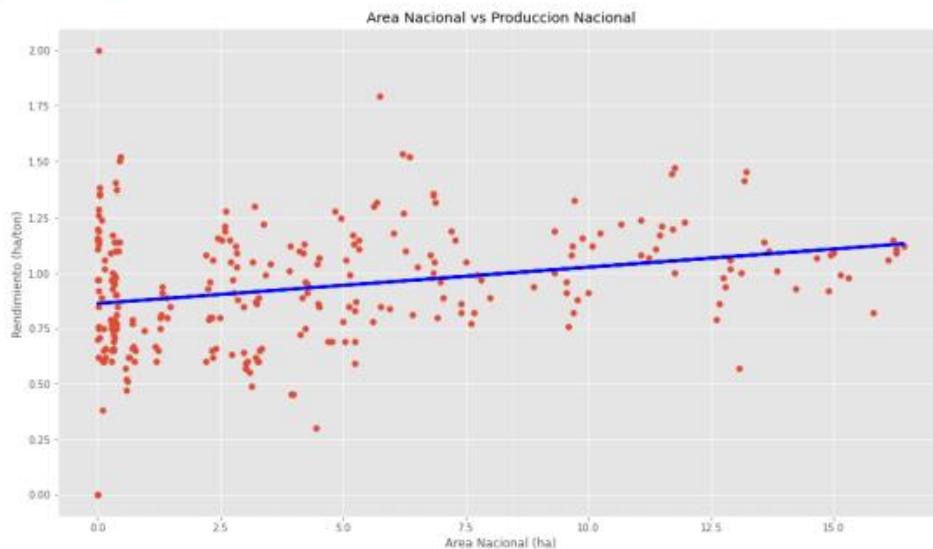
# Hacemos las predicciones según el modelo de regresión Lineal
y_pred = regr.predict(X_train)

# Ahora imprimimos los resultados obtenidos
# Vemos el valor de la pendiente, o sea la variable m, el coeficiente de la variable
print('Valor de la tangente (m) o Coefficients:=====》 ', regr.coef_)
# Ahora el valor de la constante b0, es decir el valor donde la recta corta el eje
print('Valor de la constante o Independent term: =====》 ', regr.intercept_)
# Se imprime el Error Cuadrado Medio
print("Error cuadrado medio o Mean squared error:=====》 %.2f " % mean_squared_error(y_train, y_pred))
# Puntaje o valor de Varianza. El mejor puntaje es un 1.0
print('valor de la varianza o Variance score:=====》 %.2f' % r2_score(y_train, y_pred))
```

```
Valor de la tangente (m) o Coefficients:=====》 [0.01642121]
Valor de la constante o Independent term: =====》 0.8623491800082033
Error cuadrado medio o Mean squared error:=====》 0.07
valor de la varianza o Variance score:=====》 0.08
```

```
In [108]: # Gráfico de dispersión del comportamiento del Área Nacional versus Producción Nacional
plt.scatter(produccion_df['Área Nacional (ha)'], produccion_df['Rendimiento (ha/ton)'])
plt.title('Área Nacional vs Producción Nacional')
plt.xlabel('Área Nacional (ha)')
plt.ylabel("Rendimiento (ha/ton)")

# A continuación se grafica en color azul, La función Lineal obtenida a partir del modelo
plt.plot(X_train[:,0], y_pred, color='blue', linewidth=3)
plt.show()
```



```
In [110]: # Ahora vamos a predecir utilizando la función obtenida, el RENDIMIENTO (ha/ton).
# Queremos predecir el rendimiento de la producción nacional de Café vamos a obtenerlo
# según nuestro modelo, hacemos:
producción_obtenida = regr.predict([[2]])
print('Estimación del rendimiento del Café en (hectareas/toneladas)====>%.3f' %producción)
```

Estimación del rendimiento del Café en (hectareas/toneladas)====>0.895

```
In [111]: # Ahora vamos a predecir utilizando la función obtenida, el RENDIMIENTO (ha/ton).
# Queremos predecir el rendimiento de la producción nacional de Café vamos a obtenerlo
# según nuestro modelo, hacemos:
producción_obtenida = regr.predict([[6]])
print('Estimación del rendimiento del Café en (hectareas/toneladas)====>%.3f' %producción)
```

Estimación del rendimiento del Café en (hectareas/toneladas)====>0.961

```
In [112]: # Ahora vamos a predecir utilizando la función obtenida, el RENDIMIENTO (ha/ton).
# Queremos predecir el rendimiento de la producción nacional de Café vamos a obtenerlo
# según nuestro modelo, hacemos:
producción_obtenida = regr.predict([[11]])
print('Estimación del rendimiento del Café en (hectareas/toneladas)====>%.3f' %producción)
```

Estimación del rendimiento del Café en (hectareas/toneladas)====>1.043

```
In [113]: # Ahora vamos a predecir utilizando la función obtenida, el RENDIMIENTO (ha/ton).
# Queremos predecir el rendimiento de la producción nacional de Café vamos a obtenerlo
# según nuestro modelo, hacemos:
producción_obtenida = regr.predict([[26]])
print('Estimación del rendimiento del Café en (hectareas/toneladas)====>%.3f' %producción)
```

Estimación del rendimiento del Café en (hectareas/toneladas)====>1.289

CONCLUSIÓN

El café tiene una gran importancia en la economía, esta afecta significativamente la crematística del país tanto en el proceso de recolección como en el de exportación.

Este tiene una gran variedad, pero en este informe solo se habla de la producción de Colombia, hay que recalcar que nuestro país colombiano ocupa el 3 puesto de producción de café mundialmente, los primeros puestos se les conceden a países como Brasil y Vietnam, obviamente estos países tienen datos con tendencias diferentes pueden ser mayores o menores.

Aprendimos que la calidad del café colombiano es muy buena, por eso es muy apreciado en todo el mundo. Ha generado muchos empleos. También aprendemos que a lo largo de los años su producción ha sido muy alta gracias a las grandes empresas de exportación de café.

BIBLIOGRAFIA

- <http://www.fao.org/faostat/es/#data/QC/visualize>
- <https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=3>
- <https://www.realacademiadelcafe.com/>
- <https://ebookcentral-proquest-com.bdigital.sena.edu.co/lib/senavirtualsp/detail.action?docID=5756204>
- <https://www-alfaomegacloud-com.bdigital.sena.edu.co/reader/ciencia-de-datos?location=13>
- [https://ebookcentral-proquest-com.bdigital.sena.edu.co/lib/senavirtualsp/detail.action?docID=5214094&query=rst
udio](https://ebookcentral-proquest-com.bdigital.sena.edu.co/lib/senavirtualsp/detail.action?docID=5214094&query=rstudio)
- https://wwwalfaomegacloud-com.bdigital.sena.edu.co/auth/ip?intended_url=https://www-lfaomegacloudcom.
- https://wwwalfaomegacloud-com.bdigital.sena.edu.co/auth/ip?intended_url=https://www-lfaomegacloudcom.
- <http://www.fcharte.com/libros/ExploraVisualizaConR-Fcharte.pdf>
- [https://ebookcentral-proquest-com.bdigital.sena.edu.co/lib/senavirtualsp/detail.action?docID=5308389&query=BI
G+DATA+CON+R](https://ebookcentral-proquest-com.bdigital.sena.edu.co/lib/senavirtualsp/detail.action?docID=5308389&query=BIG+DATA+CON+R)
-

GLOSARIO

Análisis	Enfoque	Prescriptiva
Big Data	Estadística	Proceso
Calidad de los Datos	Etapa	Productividad
Ciencia de Datos	ETL	Python
Conocimiento	Información	Rendimiento
CRISP-DM	Informática	Rentabilidad
CRM	Inteligencia de Negocios	RStudio
Cualitativo	Investigación	Síntesis
Cuantitativo	Jupyter Notebook	Sistema Operativo
Dato	KDD	Smart Data
Deep Data	Minería de Datos	Toma de decisiones
Descriptiva	Modelo	Validación
Diagnóstico	Negocio	Variable
Eficacia	Predictiva	Eficiencia