Universidad Nacional
ARTURO JAURETCHE

Probabilidad y Estadística.

2do cuatrimestre 2023.

Alumnos:
Montaño, Eduardo.
Ferreyra, Matías
Bazan, Geronimo

Docente: *Guillermo Filia.*

En el siguiente informe se realizará la explicación del ***Notebook*** realizado para el ***"Tp_Obligatorio"*** de la materia ***"Probabilidad y Estadística"*** .

Enunciado del trabajo:

1. ***Encontrar un problema de interés.***
2. ***Buscar una base de datos con información pertinente.***
3. ***Cargar la base de datos en un notebook como la del ejemplo suministrado (ver link abajo).***
4. ***Explorar la base de datos.***
5. ***Encontrar las variables que describan la información en la que ustedes están interesados.***
6. ***Caracterizar las distribuciones de esas variables con medidas de tendencia central y de forma. (Medias, medianas, frecuencias, varianzas, cuartiles, curtosis, etc.). Mirar en la bibliografía los capítulos dedicados a Estadística Descriptiva.***
7. ***Representar esa distribución de varias maneras (histogramas, gráficos de barras, gráficos de cajas, etc).***
8. ***Sacar conclusiones respecto de la información obtenida.***
9. ***Escribir un informe contando el análisis realizado y los resultados encontrados.***

Índice:

1_ Introducción.

Lo primero que se realizó fue buscar una base de datos de nuestro interés para realizar un estudio de la misma realizando diversos gráficos.

datos proporcionados por el gobierno nacional Argentino mediante su sitio Web ***"Datos.gob.ar"***.

Los datos consultados son datos proporcionados por el gobierno Nacional Argentino mediante su sitio web "Datos.gob.ar" , en el cual podemos encontrar una gran cantidad de bases de diversas áreas disponibles de manera pública para su uso.

Para el desarrollo de este informe se eligió el área de "Energía" utilizando en concreto la siguiente base de datos ***"Productos Elaborados a partir del Petróleo"***.

2_ Base de datos(Tabla).

Podemos ver el tamaño de la base de datos de la siguiente manera tener en cuenta la cantidad de datos disponibles.

```
print("El tamaño de la primera base de datos es:");
print(Data1.shape);
```

```
El tamaño de la primera base de datos es:
(284934, 11)
```

Observamos una parte de las tablas para ver cómo están dispuestos los datos.

anio	mes	idempresa	empresa	idrefineria	refineria	idconce...	concepto	cantida...	cantida...	observa...
2009	1	SOU	Petrolera...	ANTA	Antartida	18	Aerokero...	200	200	
2009	1	SOU	Petrolera...	ANTA	Antartida	17	Aeronaftas	0	0	
2009	1	SOU	Petrolera...	ANTA	Antartida	13	Alconaft...	0	0	
2009	1	SOU	Petrolera...	ANTA	Antartida	30	Bases L...	0	0	
2009	1	SOU	Petrolera...	ANTA	Antartida	1	Gas de ...	0	0	
2009	1	SOU	Petrolera...	ANTA	Antartida	39	Otros Pr...	0	0	
2009	1	NAO	NEW AM...	PHN	Plaza Hu...	28	Diesel Oil	32.56	27.35	
2009	1	BRUE	BRUER...	PLVC	Planta Vi...	37	Otros Pr...	0	0	
2009	1	CARB	CARBO...	CAMS	Campan...	21	Gasoil G...	0	0	
2009	1	ENAR	ENARSA...	CAME	Campan...	24	Gasoil Bi...	0	0	
2009	1	ENAR	ENARSA...	CAME	Campan...	25	Gasoil Bi...	0	0	
2009	1	CARB	CARBO...	CAMS	Campan...	22	Gasoil G...	0	0	
2009	1	ENAR	ENARSA...	CAME	Campan...	26	Gasoil Bi...	0	0	
2009	1	ENAR	ENARSA...	CAME	Campan...	21	Gasoil G...	0	0	
2009	1	BRUE	BRUER...	PLVC	Planta Vi...	38	Otros Pr...	0	0	
2009	1	CARB	CARBO...	CAMS	Campan...	23	Gasoil G...	0	0	
2009	1	ENAR	ENARSA...	CAME	Campan...	22	Gasoil G...	0	0	
2009	1	ENAR	ENARSA...	CAME	Campan...	23	Gasoil G...	0	0	

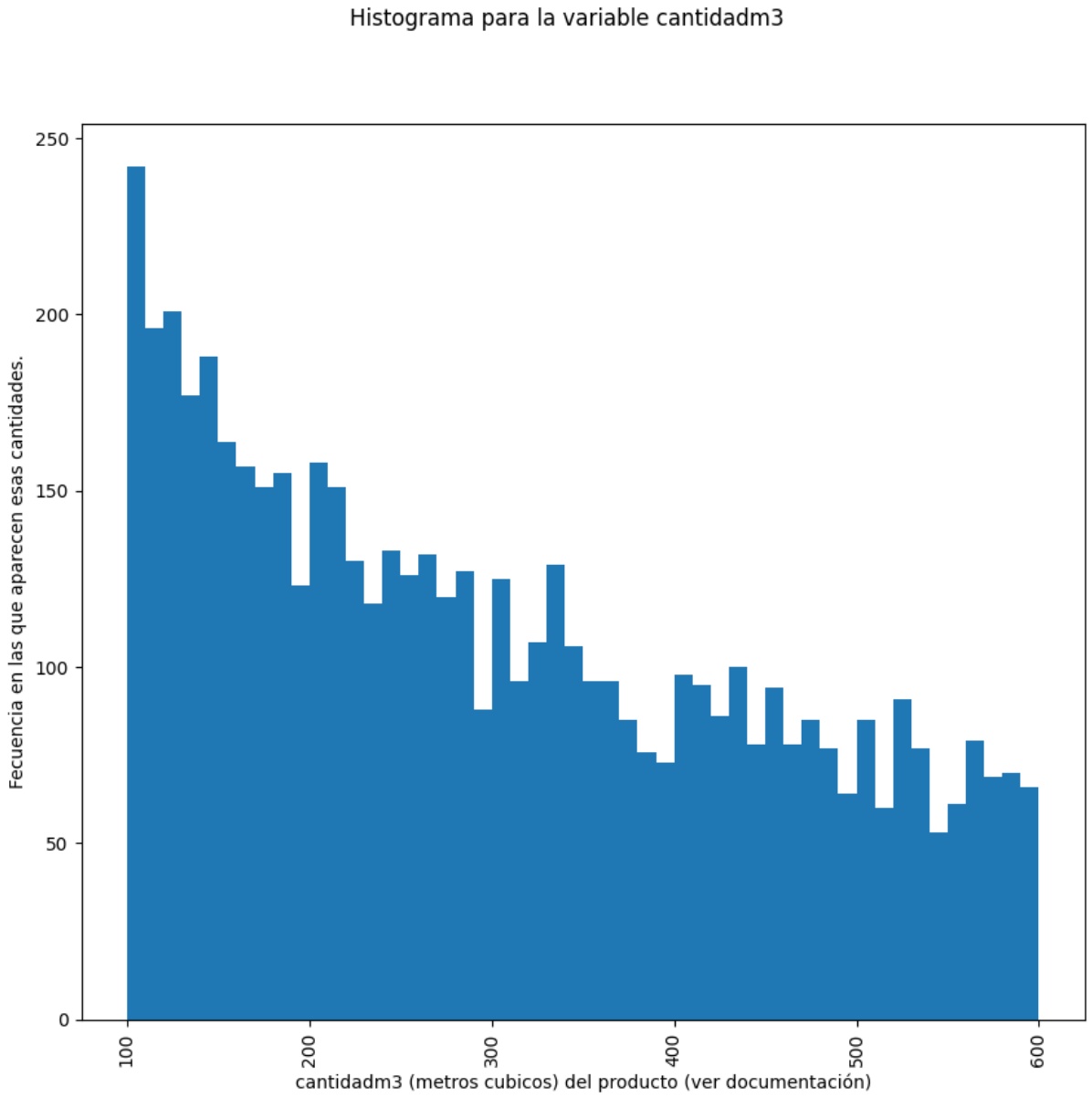
3_ Gráficos.

Lo primero antes de poder realizar los gráficos sería definir variables que analizaremos.

```
cantm3DistCero = Data1[Data1.cantidadm3 != 0]; #  Quito a toda la base de datos los valores donde la cantidadm3 es 0
cantidad = cantm3DistCero['cantidadm3'] #Seleccionamos la columna "cantidadm3" para realizar los graficos.
aux = cantidad;                         #Definimos una variable auxiliar donde capturaremos la columnas.
name = cantidad.name                    #Capturamos el nombre de la propia columna.
```

Una vez realizado esto podemos proceder a realizar el gráfico de la columna en concreto seleccionada (en este caso cantidadm3, Cantidad en Metros cúbicos).

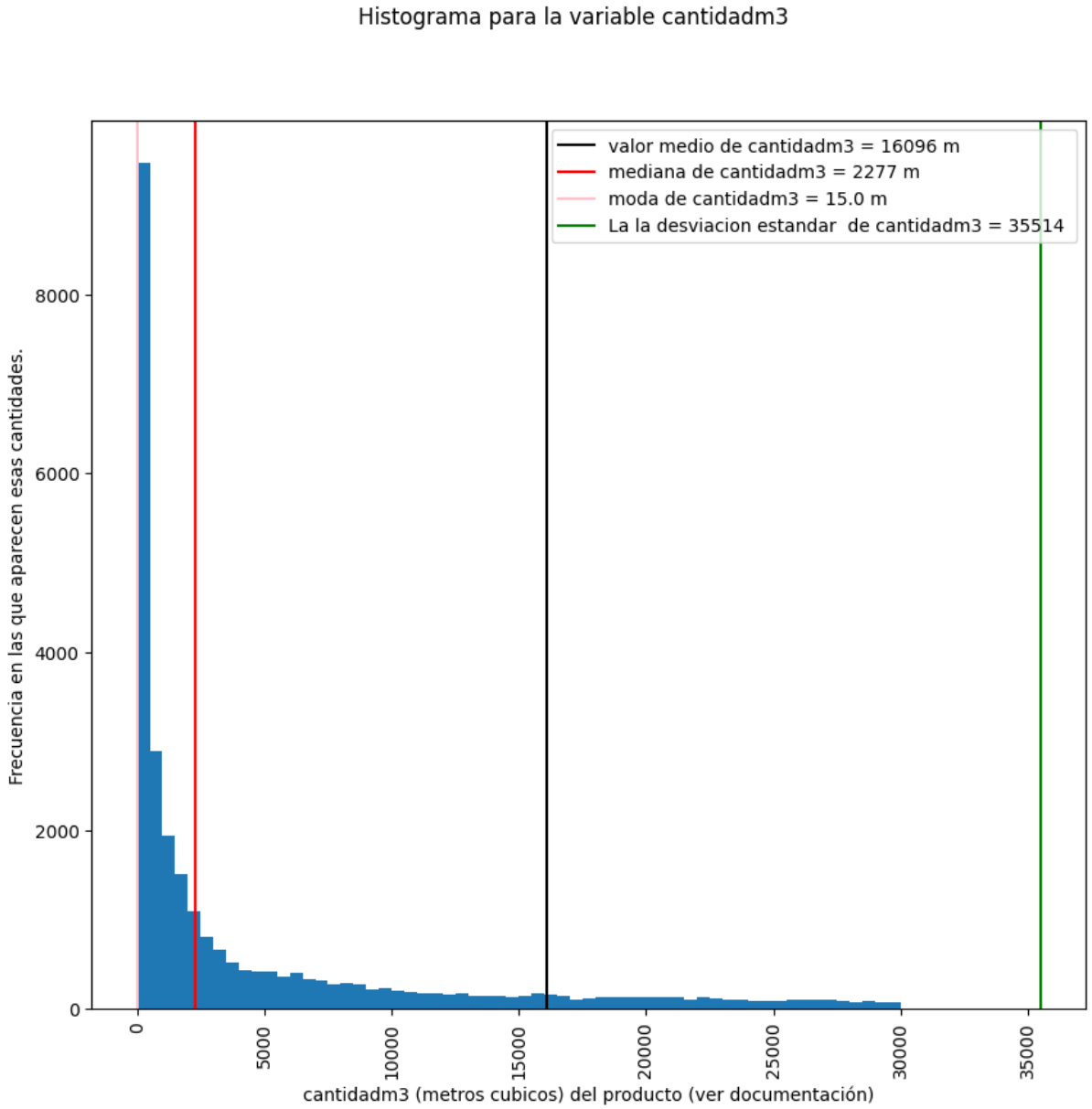
3.1_ Fig1: Análisis de la columna de "cantidadm3".



De este histograma podemos afirmar que ***"no"*** tenemos una distribución normal, aun así podemos calcular algunos datos como: ***la media, mediana y moda***.

En este histograma además ,vemos un sesgo positivo del histograma. En donde la media tendrá un valor mayor que la mediana y la mediana tendrá un valor mayor que la moda.

3.2_ Fig2: <u>Media, Mediana y Moda</u>.



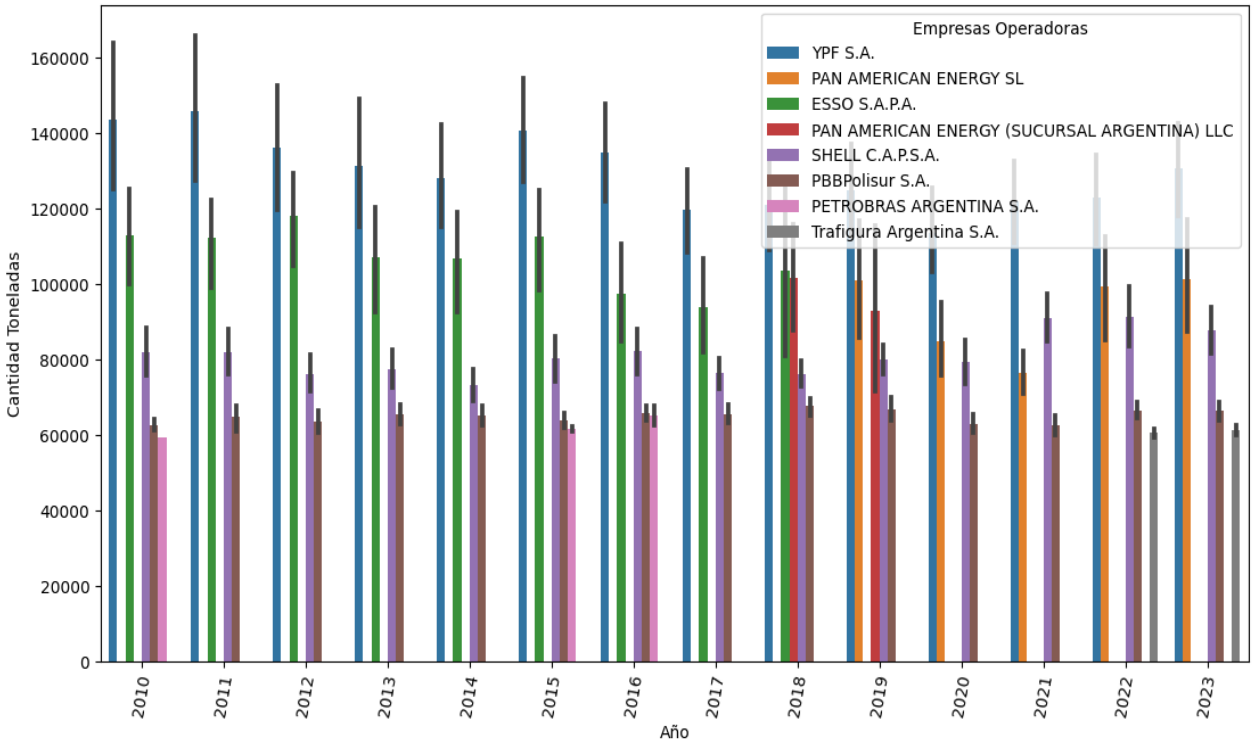
La Media: Es el valor que representa el centro o la tendencia central de un conjunto de datos.

La Mediana: Es el valor que se encuentra en el centro de los datos analizados.

La Moda: Es el valor que más se repite en los datos analizados.

*En **Fig2: <u>Media, Mediana y Moda</u>.** Se muestran los valores*
promedio de ***16.096m³***, una mediana de ***2.277m³*** y una moda de ***15.0m³***. Tal como se mencionó anteriormente los valores son ***promedio>mediana>moda***.

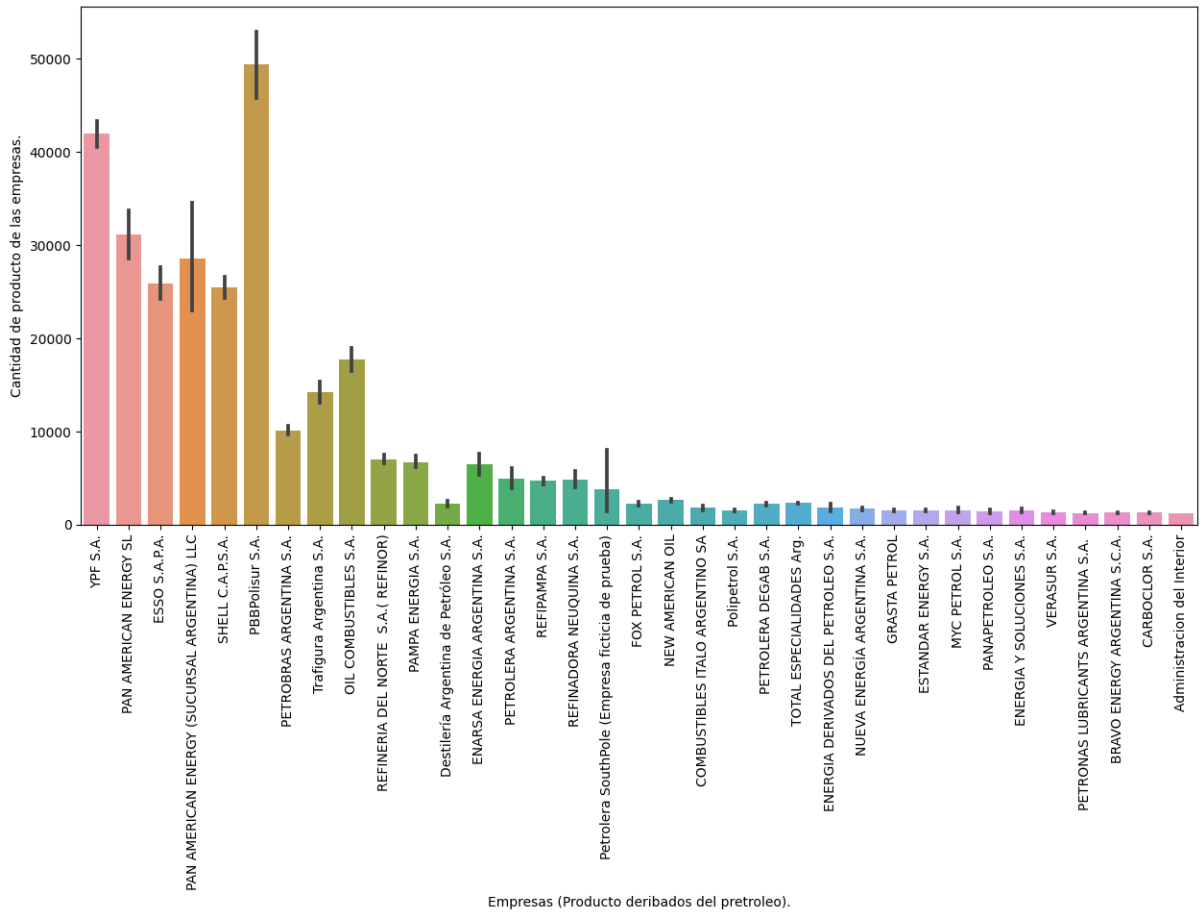
3.3_ Fig3: <u>comparación año, empresas operadoras y cantidades</u>.



En este gráfico ordenamos de manera decreciente la columna ***"<u>cantidadtns</u>"*** para luego comparar la cantidad de producto derivado del petróleo producido en ***"<u>toneladas</u>"*** por cada año *(desde 2010 hasta la actualidad)*, de este gráfico podemos analizar qué empresa tuvo una mayor cantidad de producto en un año en concreto y ver cómo fue aumentando o disminuyendo su capacidad.

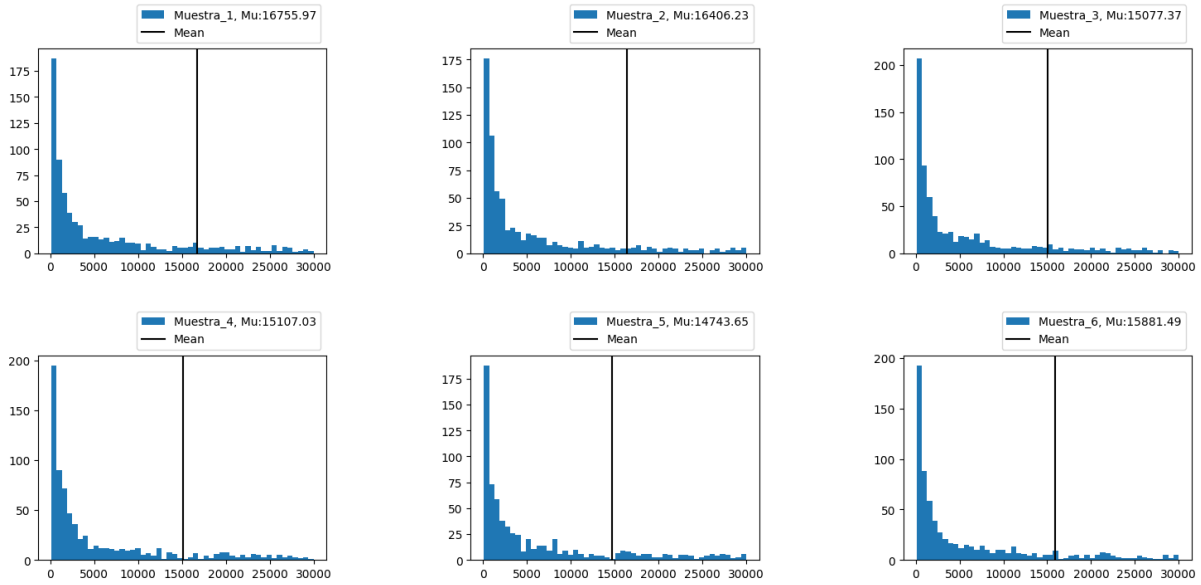
Tambien se puedem observar las media de cada empresa por año con las barras negras que sobre salen de cada barra de color.

3.4_ Fig4: Cantidad total hidrocarburos y las empresas (en Metros cubicos).



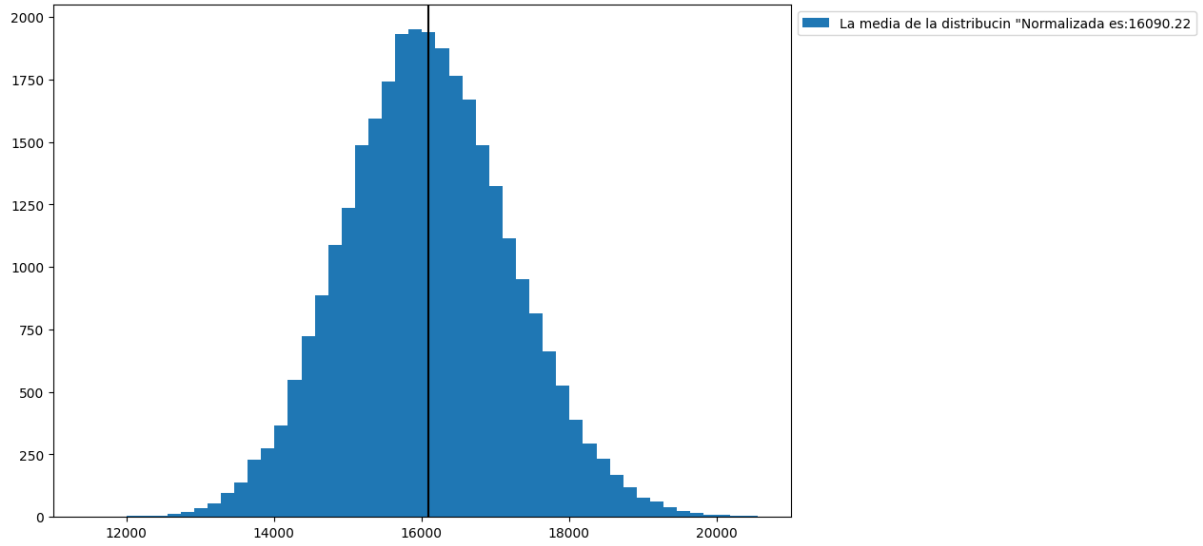
En este grafico se replica lo mismo que en la ***"Fig3"*** pero con ***"cantidadm3"***, el cual nos muestra la cantidad de producto derivado del petróleo producido por diferentes empresas.

3.5_ Fig5: Tipos de muestras.



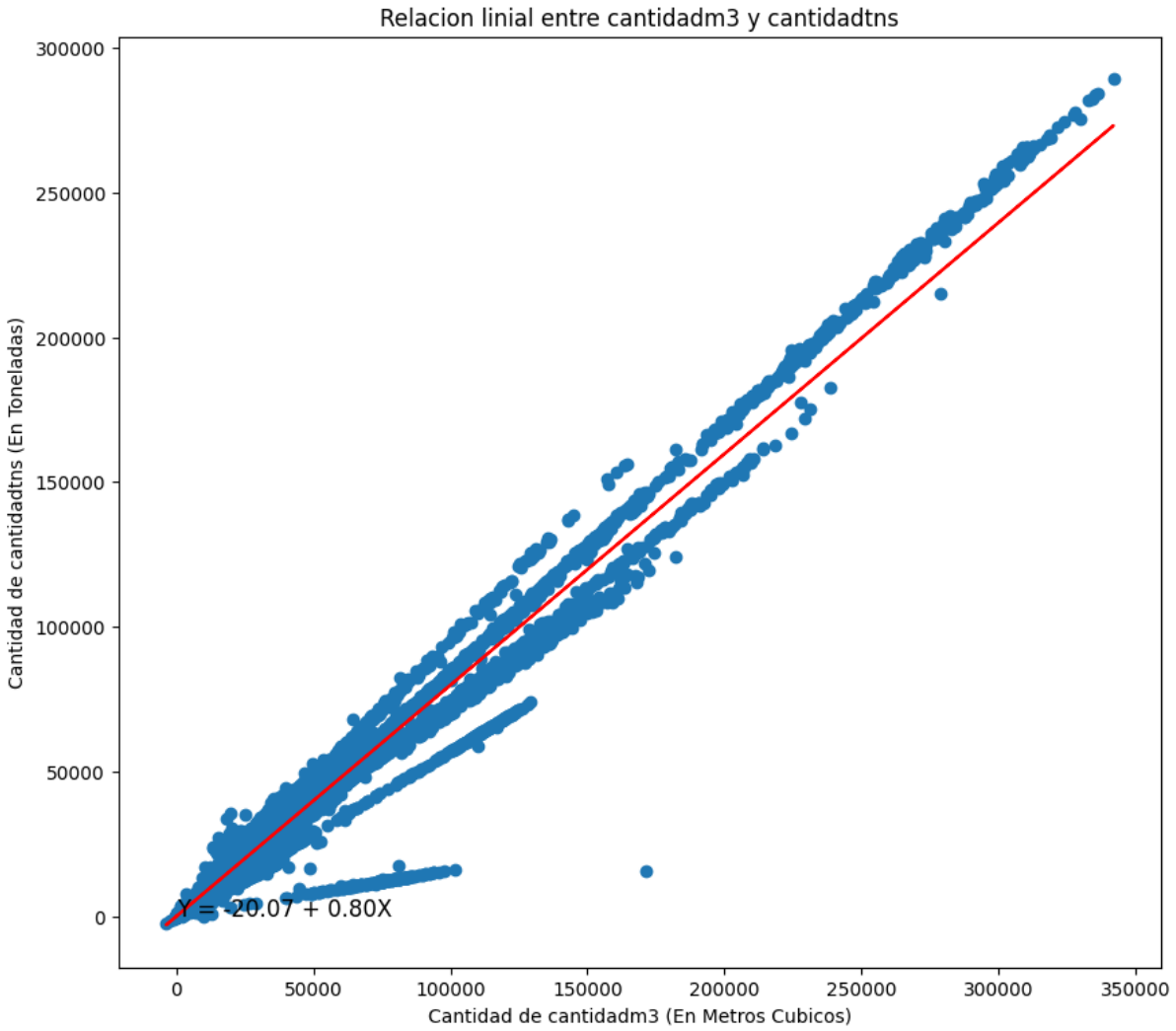
En este gráfico tomamos 6 muestras con datos aleatorios de la tabla de ***"cantidadm3"*** de las cuales le sacamos sus ***"medias"*** .

3.6_ Fig6: Distribución Normal.



Luego lo que realizamos 30.000 (treinta mil muestras) tomando 1000 (mil) elementos en cada una es decir nuestro ***"n"*** por lo que podemos podemos decir que la ***"Esperanza es normal"***.

3.7_ Fig7: Relación lineal entre metros cúbicos y toneladas.



En este gráfico se puede ver una relación lineal entre ***"cantidadm3"*** y ***"cantidadtns"***, con una pendiente positiva lo que quiere decir que por cada unidad del eje ***X (cantidadm3)***, se espera un aumento de unidades en el eje ***Y(cantidadtns)***.

3.7.1_ Covarianza , Coeficiente de correlación.

Podemos realizar el cálculo de la covarianza con la siguiente fórmula.

$$Cov(x,y) = E(x,y) - E(x)E(y)$$

Así como el coeficiente de correlación con la siguiente fórmula.

$$Pxy=Cov(x,y)/V(x)V(y)$$

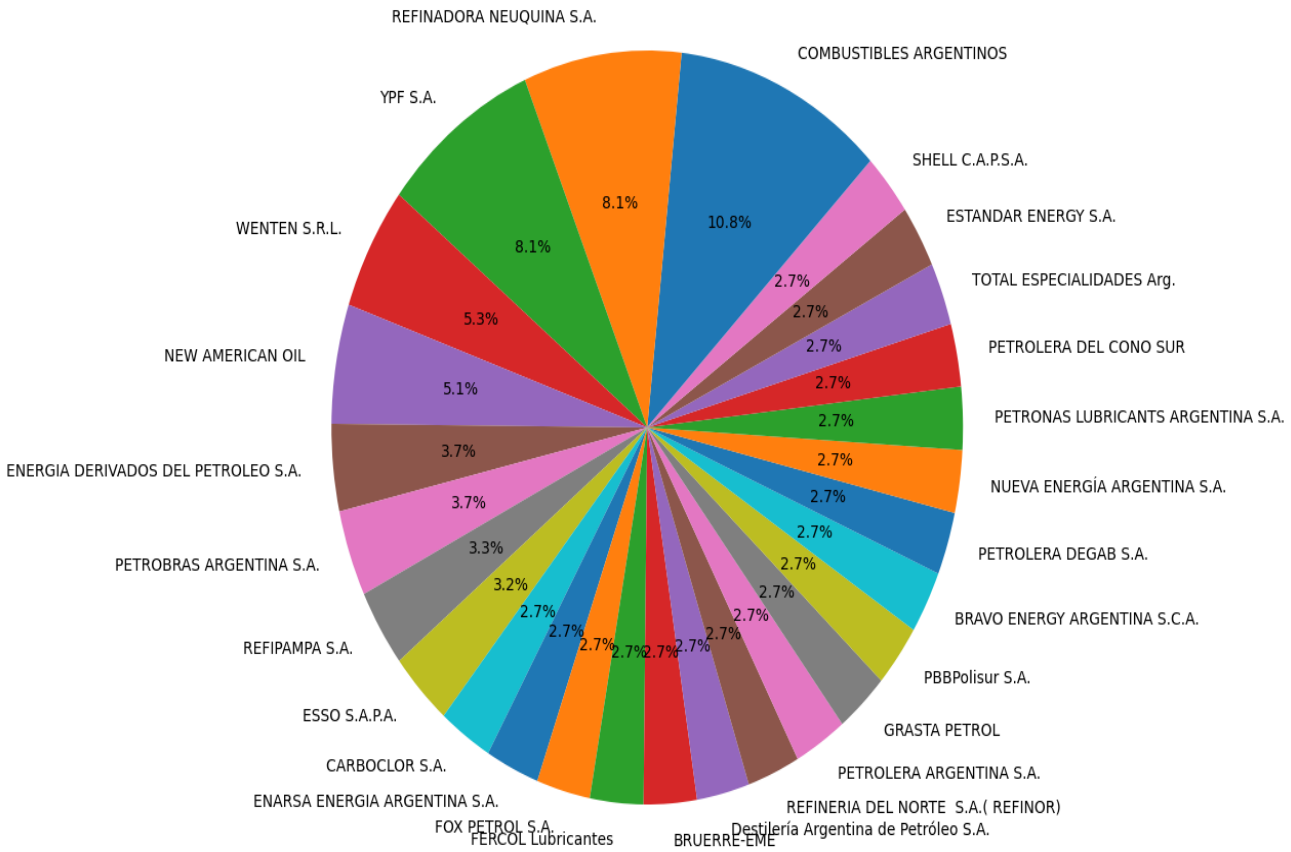
para facilitar los cálculos usamos la librería ***"pandas"***

```
Pendiente (slope): 0.7987834260038912
Termino independiente (intercept): -20.07441440361913
Valor p (Grado de relacion): 0.0
La covarianza es: 137628014.5108512
Este es el cooeficiente de correlación:0.9884403576102463
```

En la covarianza tenemos un valor positivo, lo cual indica en la figura que ambas variables tienden a aumentar de manera conjunta**.**

En cuanto al coeficiente de correlación se puede afirmar que tiende a ***"1"*** es decir que tiene una ***"relación lineal positiva total"***.

2.8_ Fig8: Relación entre empresas y porcentajes.



Tipo de distribución porcentual, gráfico pastel, en este caso nos muestra la distribución del porcentaje de cada empresa del total de hidrocarburos, derivados del petróleo. Se puede observar rápidamente las empresas con mayor producción.

Nota: *para facilitar la visualización mostramos las empresas que cuentan con un porcentaje mayor a* ***2%****.*

4_ Conclusiones.

Hemos llegado a la conclusión que la esperanza que se puede observar en la ***"Fig2"*** es equivalente a la esperanza que nos dio como resultado en la ***"Fig6".*** Para cada muestra decidimos tomar una cantidad de elementos mayor 30 (como dice la teoría), es decir aproximadamente 16096. En cuanto a nuestra varianza la calculamos como:

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Dado que nuestra varianza es 35514 (en la ***"Fig2"***) los cálculos quedan expresados de la siguiente manera:

$$\sigma_{\bar{x}} = 35514 / \sqrt{1000} = 1123.05$$

Cómo detallamos en la ***"Fig7"*** se da una ***relación lineal positiva total***. Además, los datos que se proporcionan en la ***"Fig8"*** se pueden utilizar para calcular una probabilidad condicional.

5_Bibliografía:

NoteBook:
https://colab.research.google.com/drive/1vL46Sp1GLHS7oezMOZali8xq1pwp7tX6#scrollTo=CH_ty10IrVHN

[1] https://github.com/thepycoach/python-course-for-excel-users

[2] https://youtube.com/watch?v=XEG4eh5l_qU

[3] https://www.youtube.com/watch?v=QPaHUBphDgo

[4] https://www.youtube.com/watch?v=zAIWnwqHGok