Taller 1 Analítica Computacional para la toma de Decisiones

Sara Alejandra Gómez – 202022625

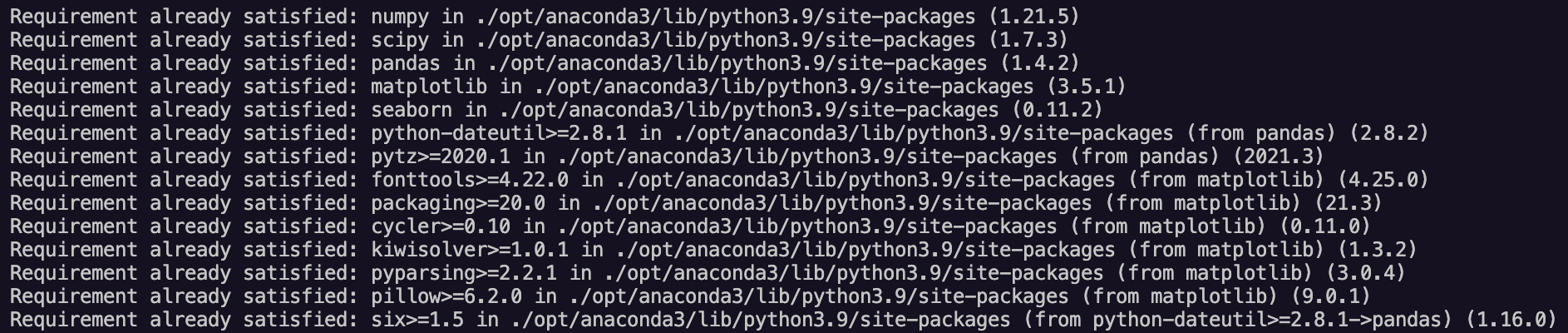
1. Instale su ambiente de desarrollo

1. /Users/saragomezgonzalez/Downloads/python-3.11.0-macos11.pkg

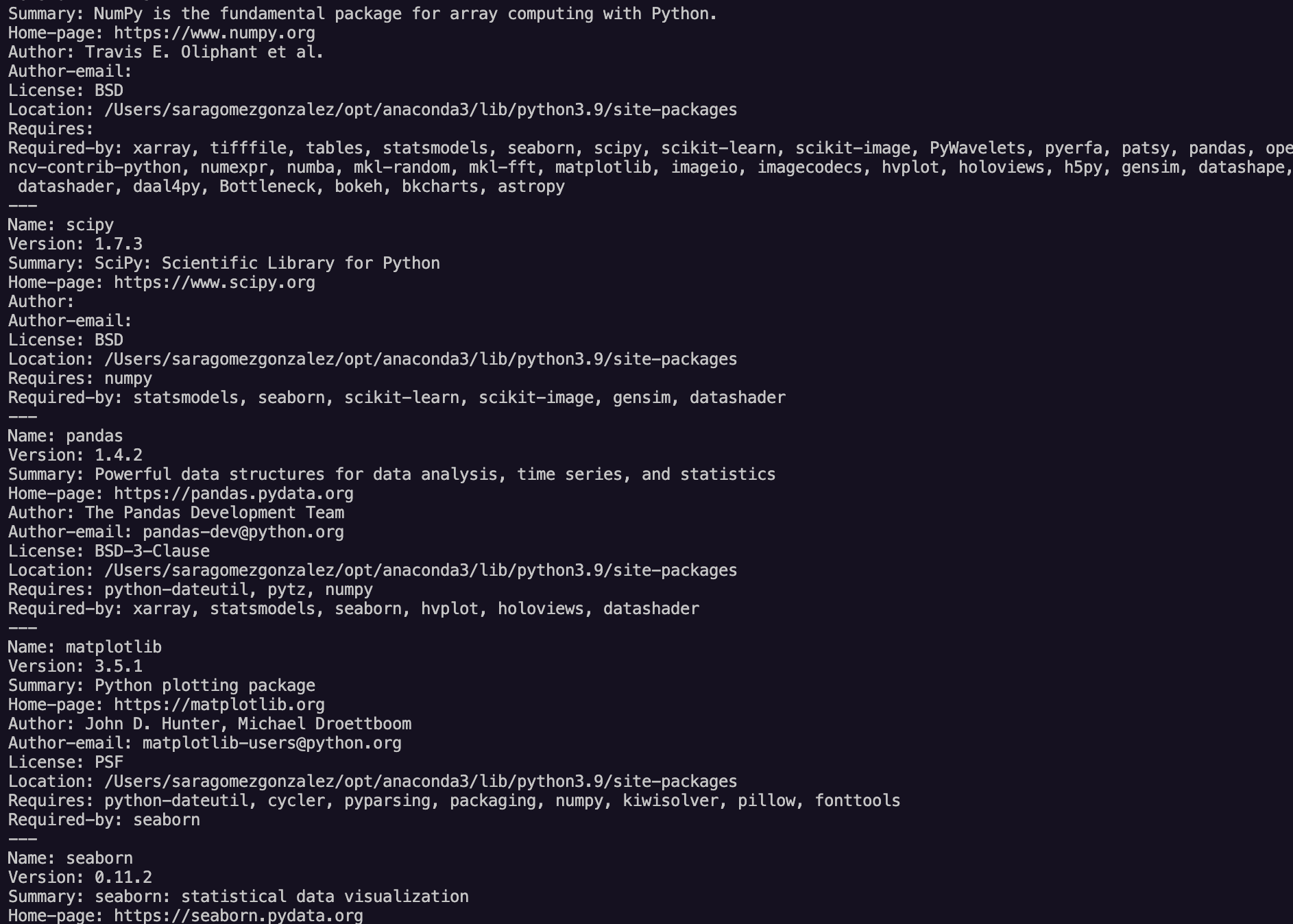
2. /Users/saragomezgonzalez/Downloads/Visual Studio Code.app

5. En la terminal instale los paquetes numpy, scipy, pandas, matplotlib y

seaborn conel comando (si está usando pypi)



6. Verifique la versión instalada de cada paquete con el comando

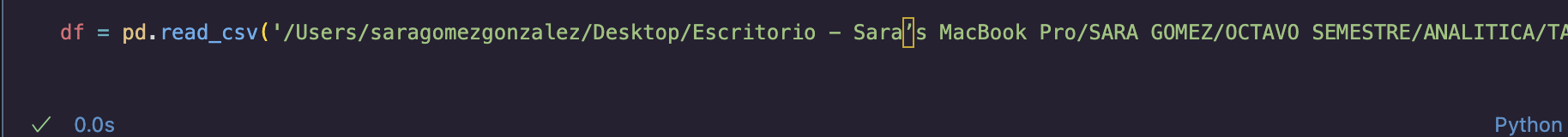


2. Exploración de datos en python

1. Adjunto a este taller encontrara un cuaderno de nombre bank-churners-init.ipynb y un archivo de valores separados por comas (CSV) BankChurn.csv.

2. Abra el cuaderno y revise en detalle las celdas.

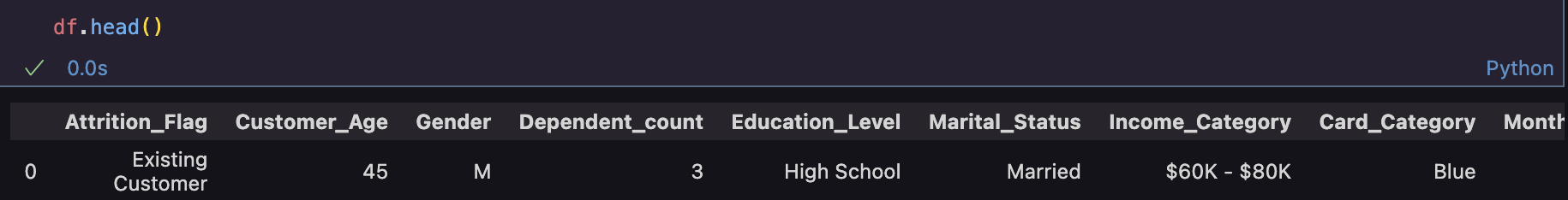
3. Ejecute cada una de las celdas (si es la primera vez que ejecuta un cuaderno en Visual Studio, le pedirá instalar la extensión para cuadernos de Jupyter).

4. En su reporte explique qué se hace en cada una de las celdas (brevemente, una línea por celda).

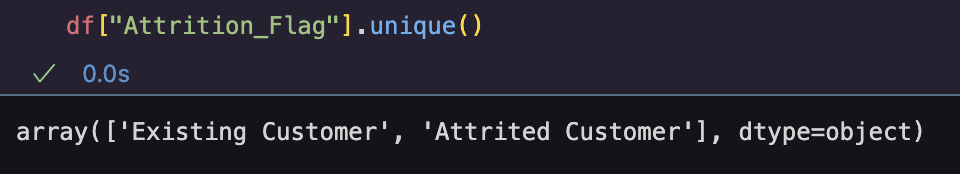
En la línea, se hace click en la barra izquierda, y se copia la ruta exacta del archivo CSV.



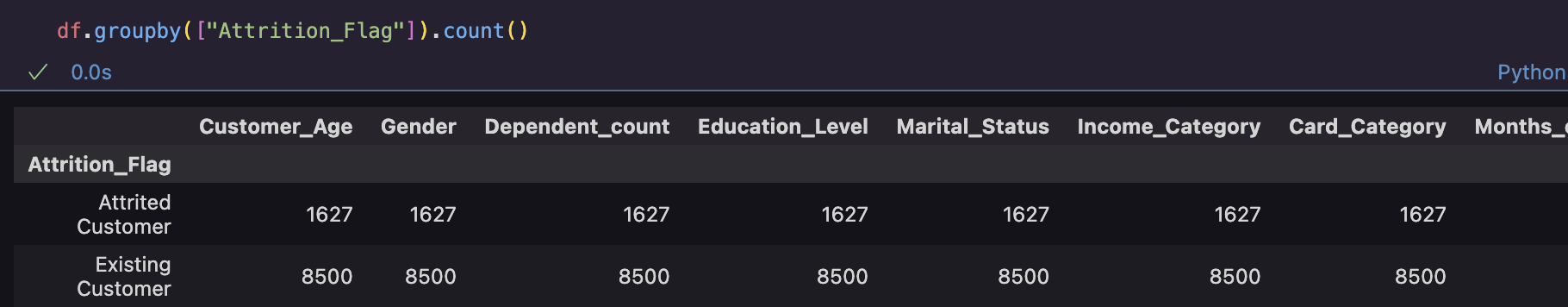
Se cuentan la cantidad de columnas y filas del data frame.



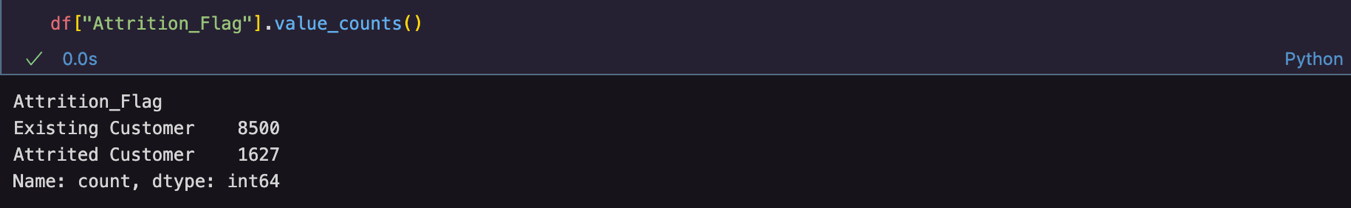
En la línea, se muestran los primeros 5 datos del data frame para una mayor claridad de la información.



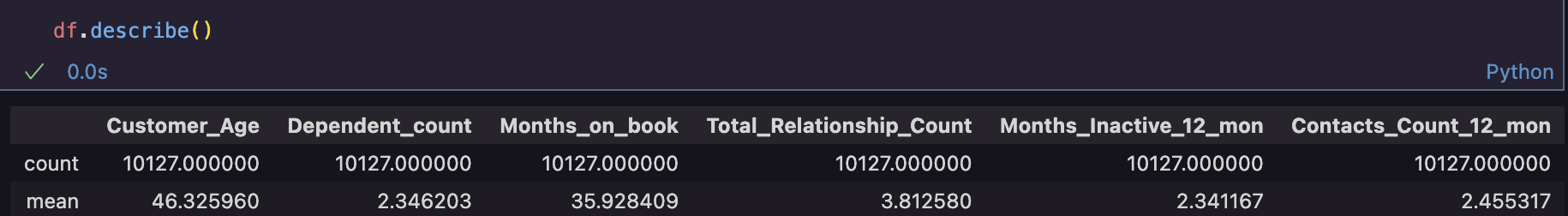
Esta función encuentra los valores únicos de la variable Attrition Flag, en forma de array.



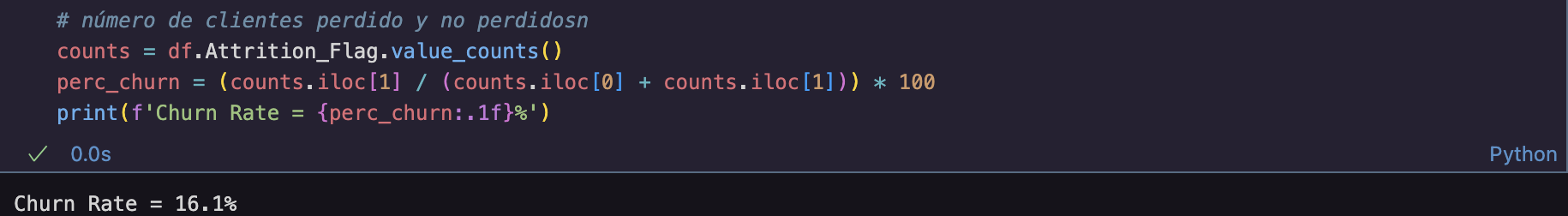
Esta función es muy interesante porque el groupby funciona como un filtro, para que agrupe los datos con base en un atributo en este caso es el Attrition Flag, y cuenta la cantidad de variables que hay para cada atributo.



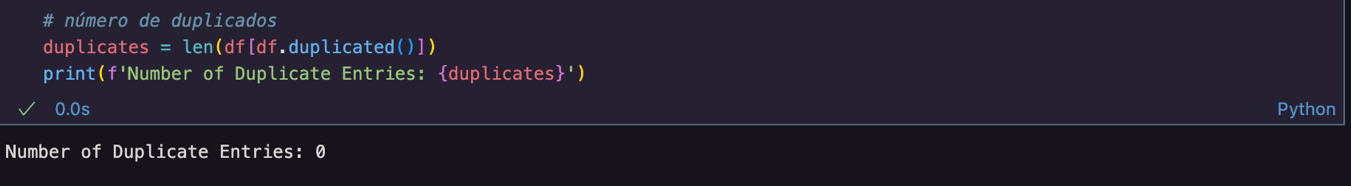
La función cuenta la cantidad de veces que esta un valor único por el atributo de la variable Attrition Flag.



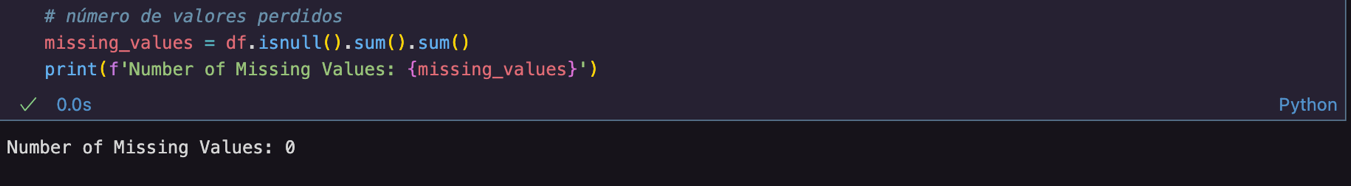
La función describe, realiza una descripción estadística para cada una de las variables numéricas del data frame.



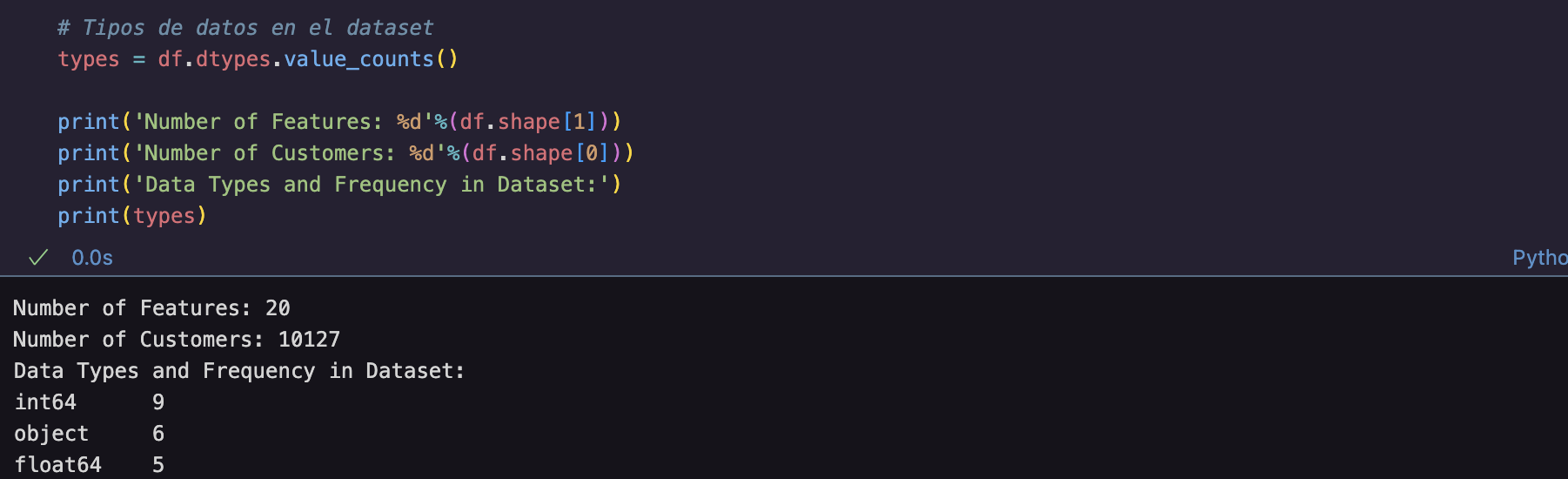
La función, calcula el número de clientes perdidos, como un promedio, entonces calcula el número de clientes perdidos sobre el total de clientes y lo convierte en porcentaje.



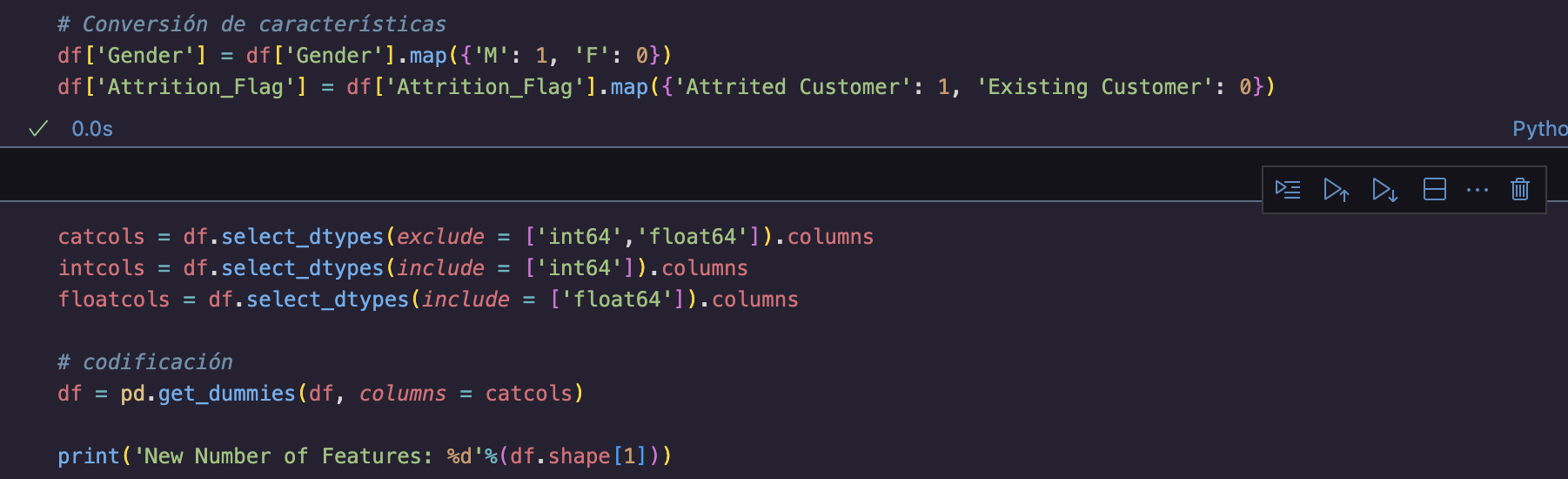
La función duplicated, devuelve valores de True y False, y vuelve a llamar el data frame para que solo muestre los valores que están duplicados. Y luego len calcula el número de elementos el data frame.



La función isnull primero convierte los valores a True y False donde True es indicativo de que es un valor nulo. Luego, el primer sum suma todos los valores de la serie y el otro sum suma el valor total de la suma de todas las serie.



La línea, lo que hace es primero acceder al data frame y mediante dtypes, arroja el tipo de valor que tienen los datos, luego, los cuenta y los mete en una serie con su respectivo tipo y la frecuencia en el data set. Ademas las funciones de df.shape cuentan la cantidad de elementos en la posición mencionada.

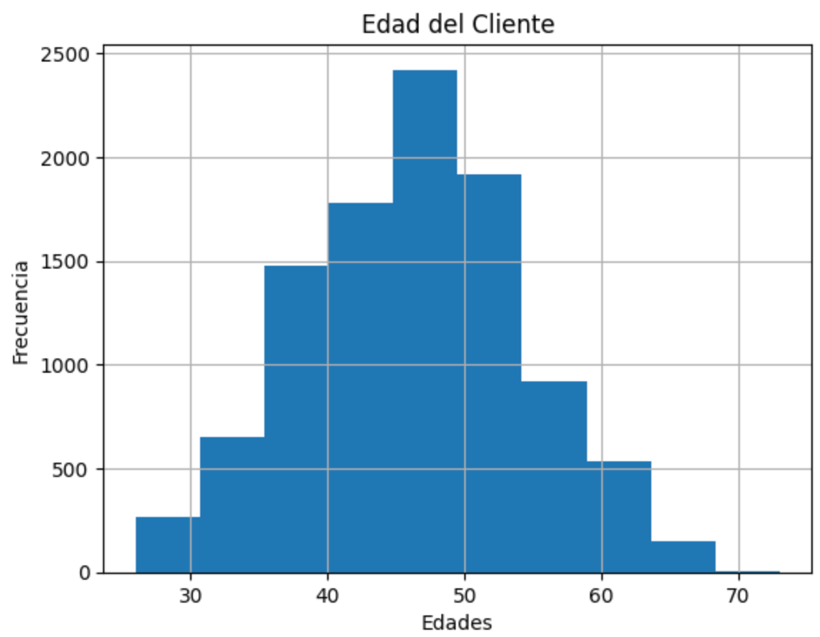


En el primer fragmento del código, se realiza una transformación de las características mediante la función map, convirtiendo los géneros en 1 y 0 así como los elementos del Attrition Flag.

En la segunda parte del código, en el caso de catcols, se añade a las columnas aquellos elementos cuyo atributo sea diferente a un entero y float. Seguidamente, convierte los elementos del data frame en variables binarias y se vuelve a obtener el número de features con las nuevas características.

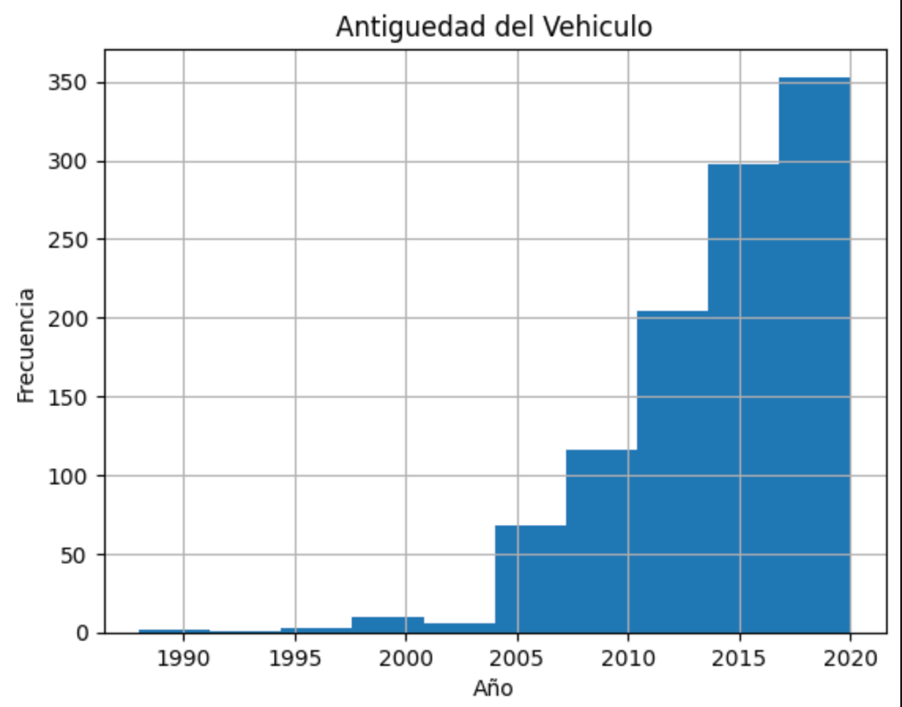
5. En la documentación de Pandas o Matplotlib busque cómo graficar un histograma.

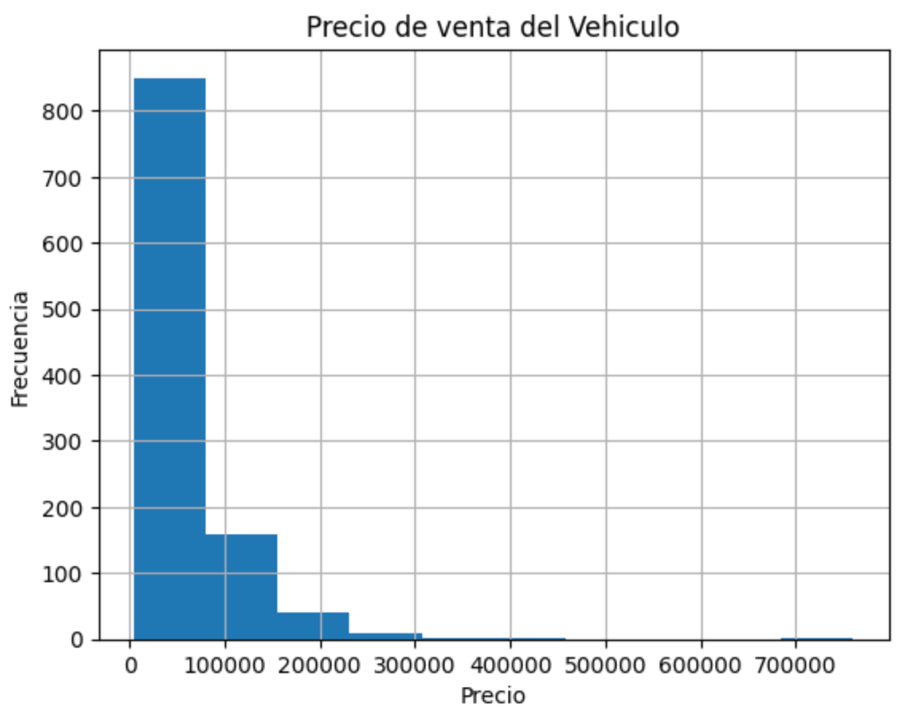
Realice un histograma de una variable de interés e inclúyalo en su reporte.



Histograma 1: Distribución de edad de los clientes

6. Repita el ejercicio anterior usando los datos BikePrices.csv que encontrara en Bloque Neón, en un nuevo cuaderno. Incluya el cuaderno como parte de su entrega (archivo de soporte).





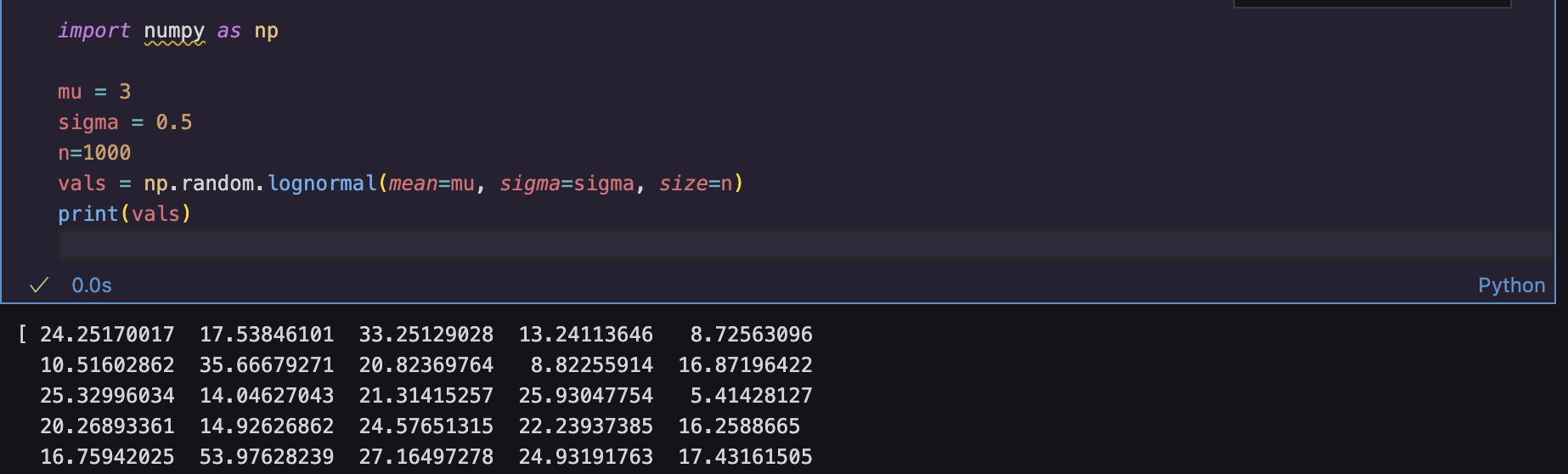
3. Números aleatorios y bondad de ajuste en python

1. Adjunto a este taller encontrara un cuaderno de nombre data-gen.ipynb.

2. Abra el cuaderno y revise en detalle las celdas.

3. Ejecute cada una de las celdas.

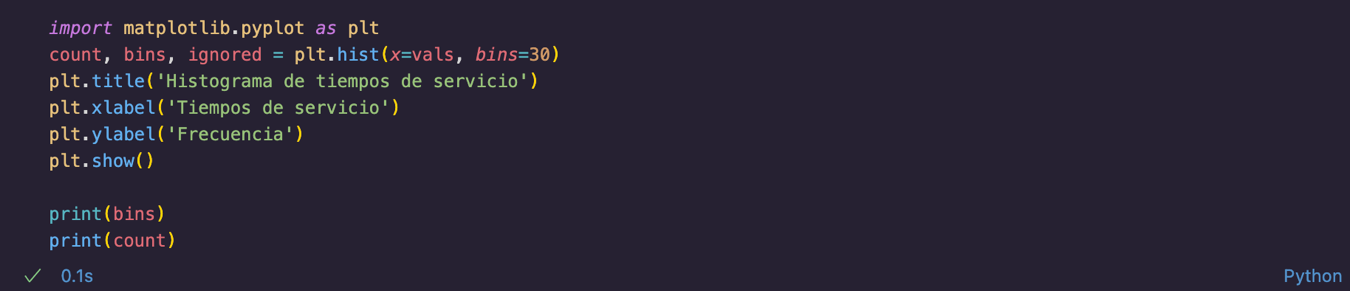
4. En su reporte explique qué se hace en cada una de las celdas (brevemente, una línea por celda).

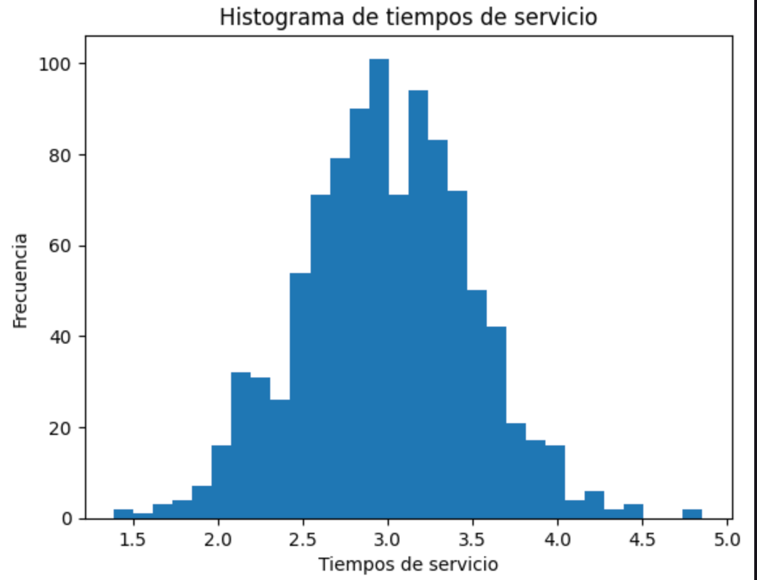


El código genera 1000 números aleatorios que siguen una distribución normal estándar con media 3 y desviación estándar 0.5.

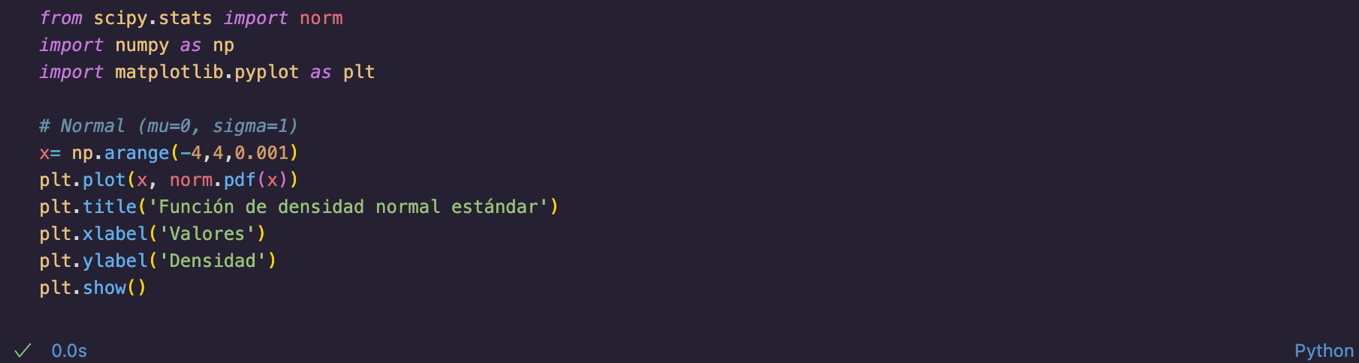


La función crea un data frame a partir de los valores de vals generados aleatoriamente, luego con la función describe, crea las estadísticas descriptivas sobre el conjunto de números.

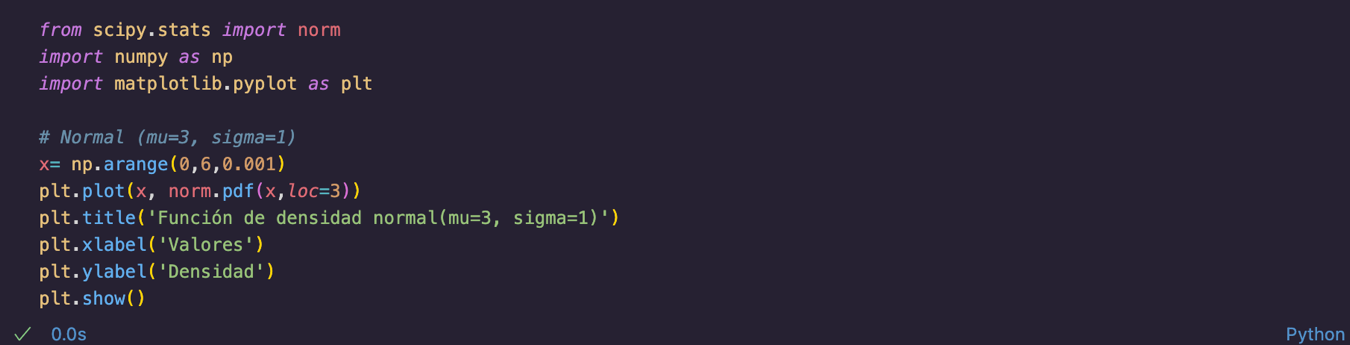


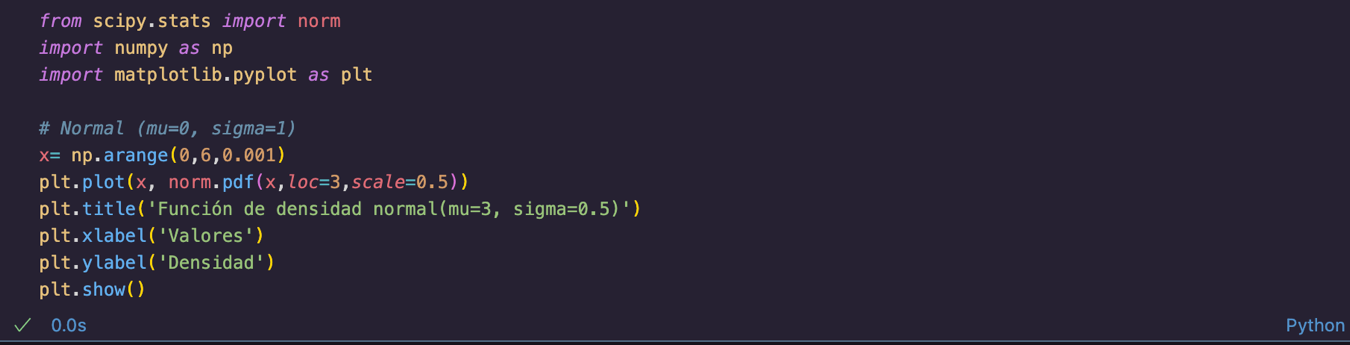


El anterior código genera un histograma donde los valores siguen la distribución normal y contiene 30 bins, los cuales son los intervalos. Por su parte, el count imprime el número de elementos en cada uno de los intervalos.

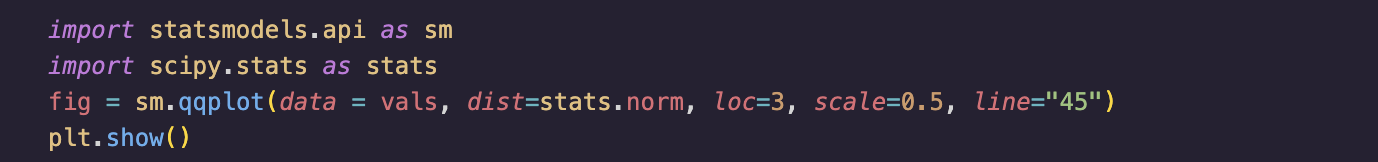


El código, en primer lugar, crea un rango de números desde el -4 hasta el 4 con un incremento de 0.001 para que sea la amplitud del gráfico. Seguidamente, se obtiene la función de densidad normal con la función norm.pdf.

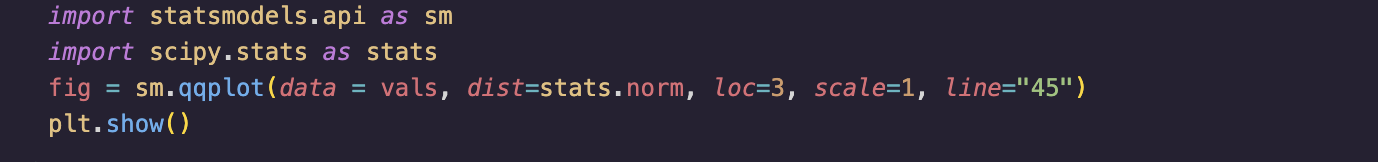


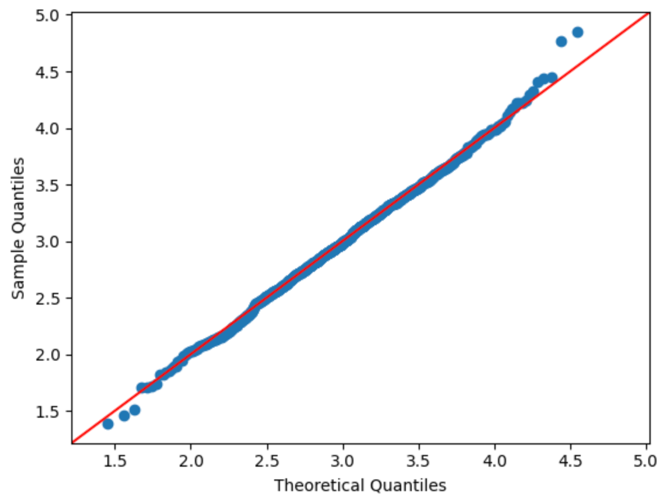


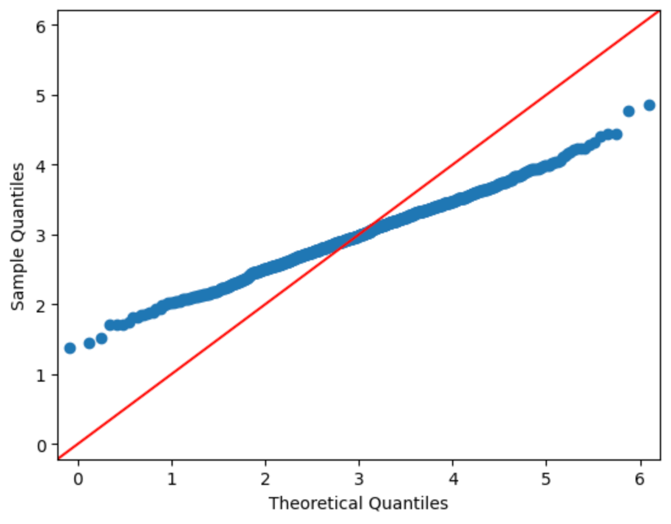
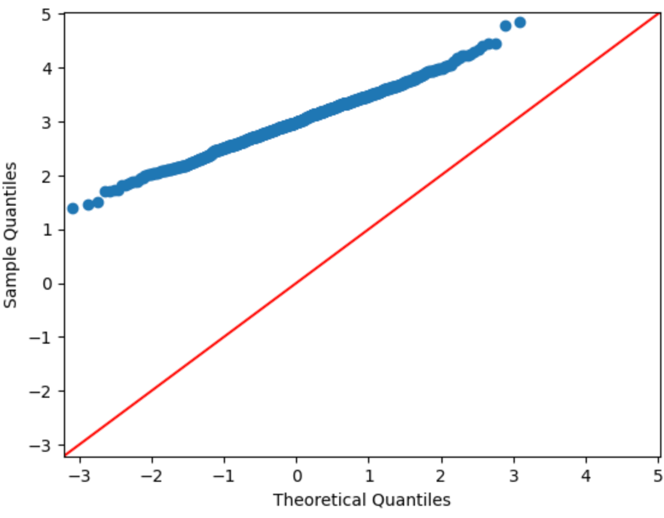
En los dos siguientes bloques de código, se cambian los parámetros de la función normal para obtener dos nuevos graficos, este cambio se ve en la parte del rango que se utiliza que cambia para ser de 0 a 6 y luego, en la norm.pdf, se cambian los parámetros de loc y scale.



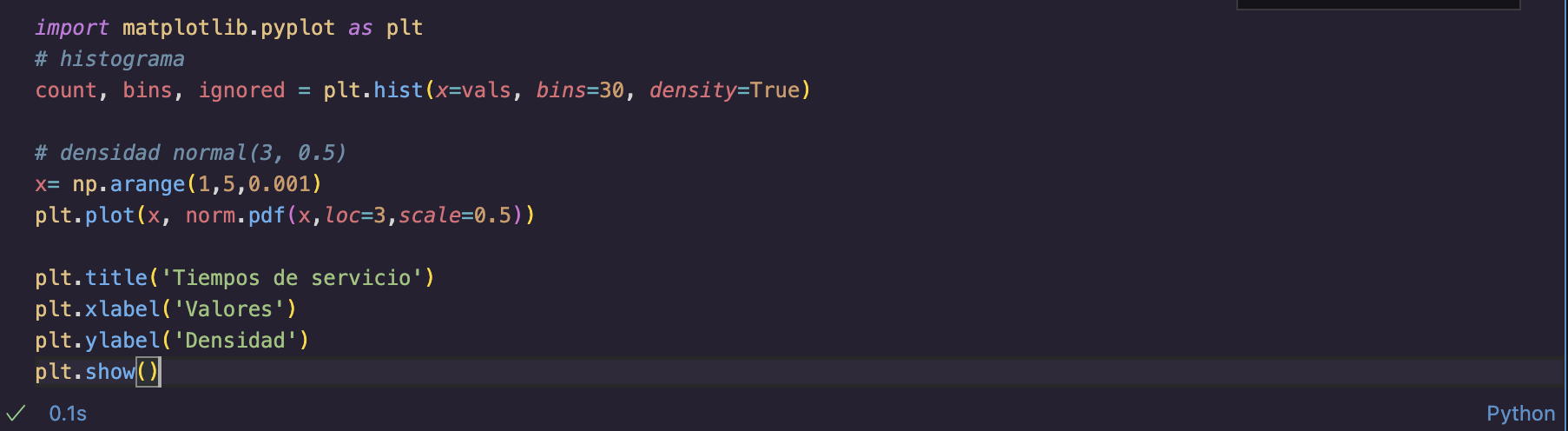
Este código, se utiliza para calcular el qq plot correspondiente a los datos, y en el parámetro de dist es necesario poner el tipo de distribución que se quiere utilizar en este caso es una normal estándar con media 3 y desviación 0.5 y el parámetro de line es la línea que se utiliza de referencia para la alineación de los datos .



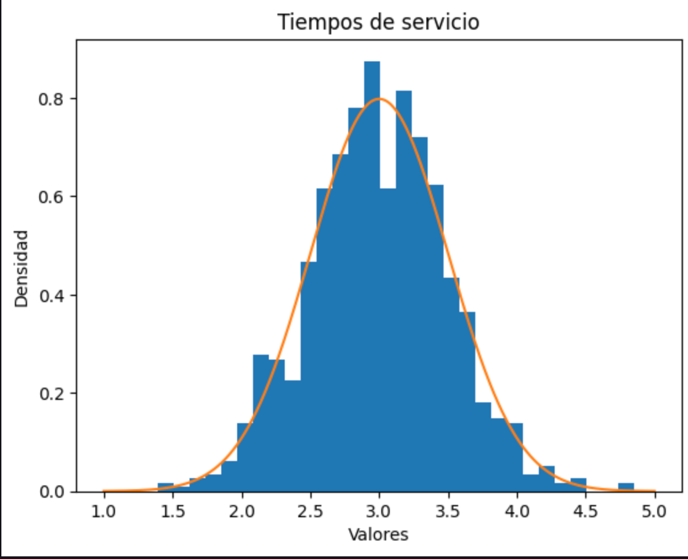




El primer gráfico, sigue los parámetros definidos inicialmente y la alineación con la línea de 45 grados, lo que indica que cumple la distribución normal. Por otro lado, el segundo grafico tiene un cambio en la media a 0 y desviación estándar de 1, lo que indica que no sigue una distribución normal y está muy distanciado de los puntos de referencia. El tercer grafico tiene una media de 3 y una desviación estándar de 1, podemos ver que hay una alineación en los puntos medios, sin embargo los extremos están fuera de la línea por lo que el cambio de los parámetros afecta la distribución.



El código, crea un histograma con una línea que muestra la densidad de la función, y se hace un plot de la función de densidad normal con media de 3 y desviación de 0.5.

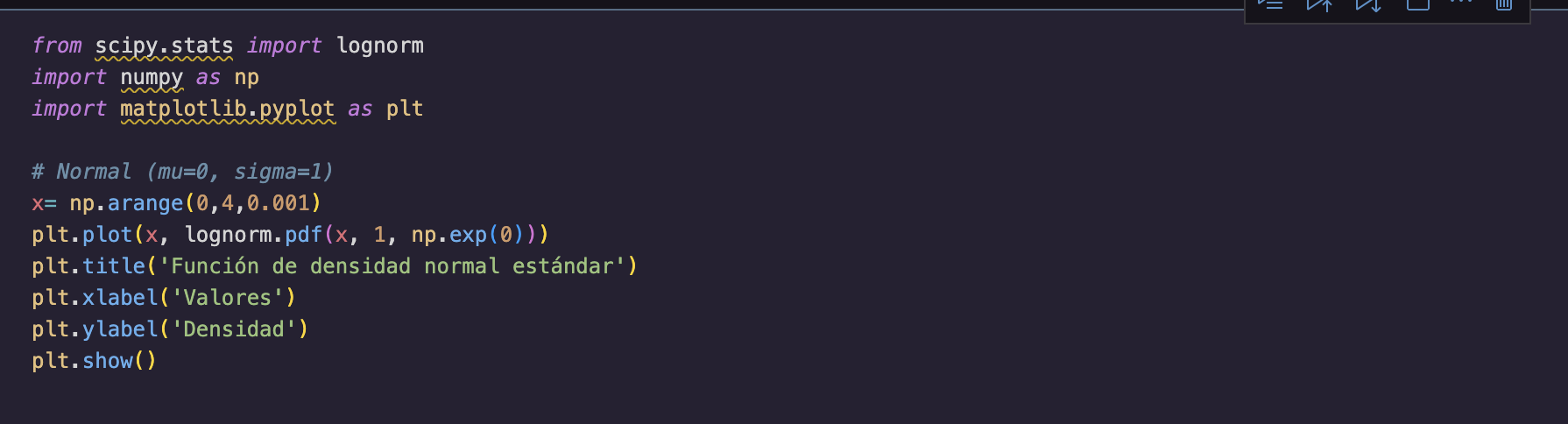


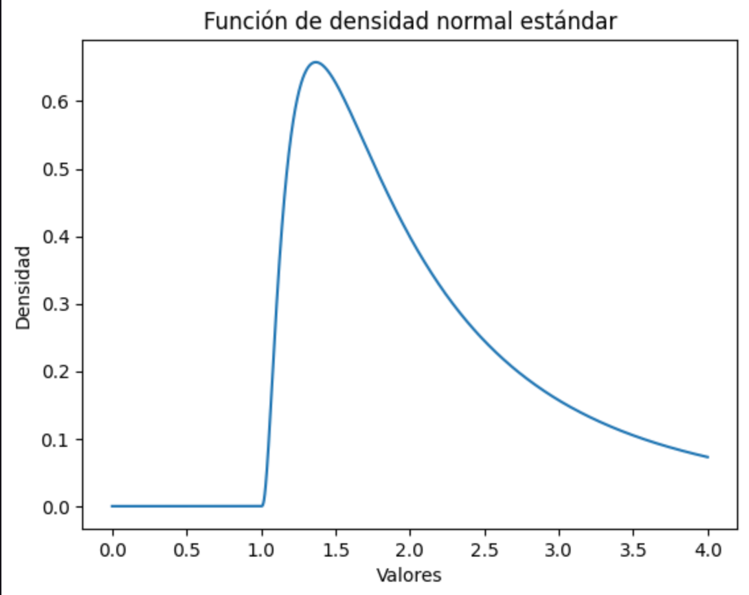
El resultado muestra que los datos siguen una distribución de probabilidad normal por la forma de campana de la línea de densidad.

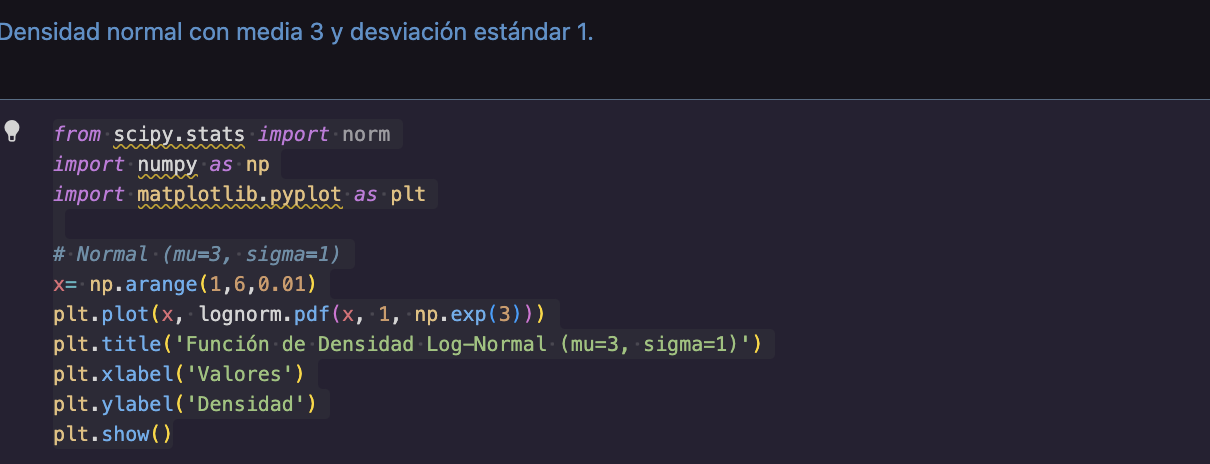
5. Repita el ejercicio del cuaderno modificando la distribución de probabilidad. Seleccione una de las disponibles (de la Beta a la Zipf) en la página <https://numpy.org/doc/stable/reference/random/generated/numpy.random.beta.html>.

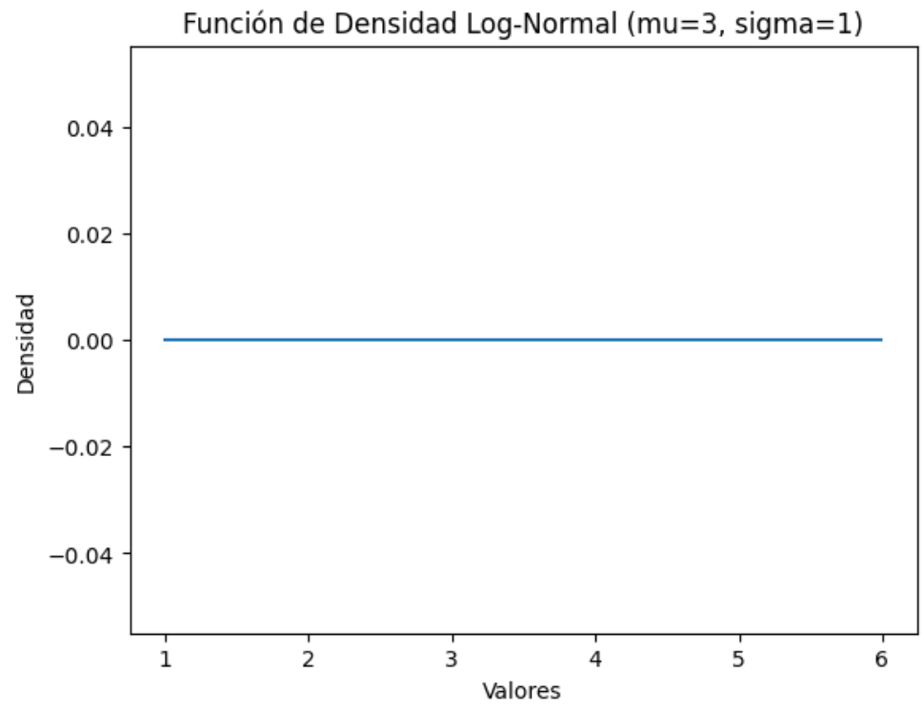
Documente sus resultados en su reporte.

Se utiliza la función de probabilidad de lognormal.

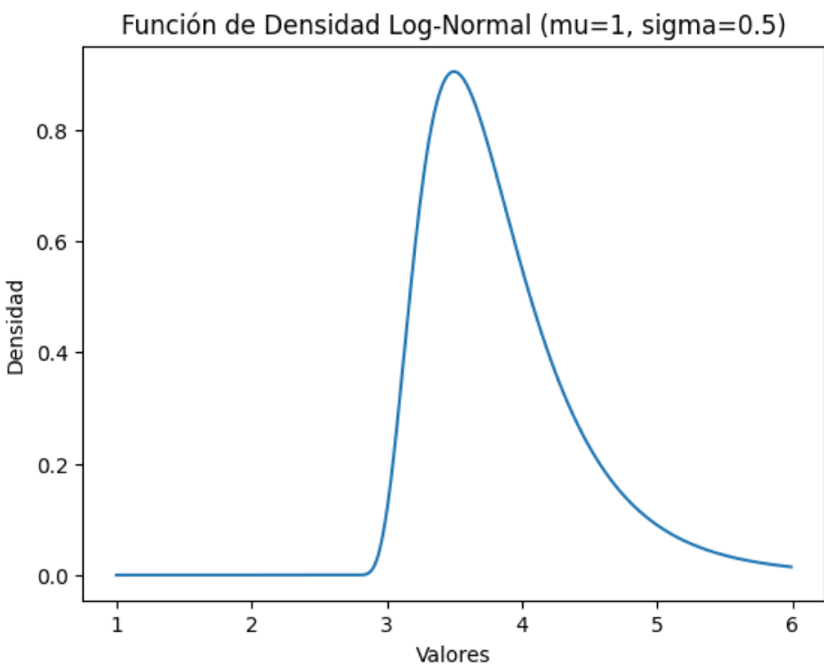




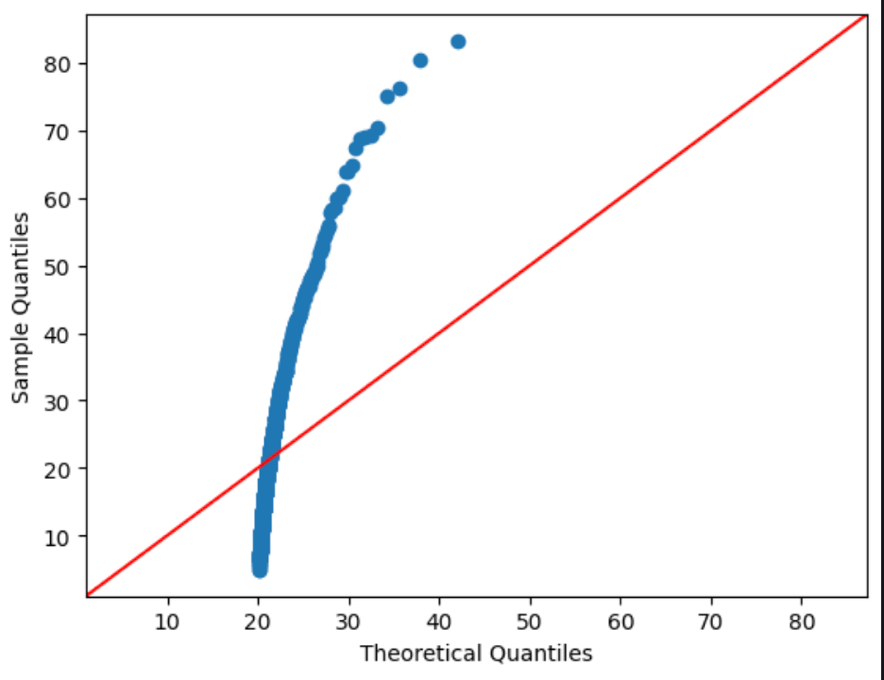
Seguidamente hacemos el cambio de los parámetros para la media:  




Se entiende que para valores de la media mayores a 1 el grafico se convierte en una línea recta, en cambio sí utilizamos una media de 1 y desviación de 0.5, obtenemos un gráfico muy similar nuevamente al primer grafico mostrado.



Al realizar el qq-plot, podemos analizar que no se están usando los parámetros correctos para la función log-normal, por lo tanto hay una desviación significativa de los extremos y el centro de la función sobre la línea de referencia.



Por ultimo realizamos el histograma:



Podemos concluir que el modelo de distribución log-normal se ajusta adecuadamente a los datos de los tiempos de servicio, sin evidenciar valores atípicos significativos. La asimetría positiva de la distribución se explica por la presencia de algunos tiempos de servicio prolongados.