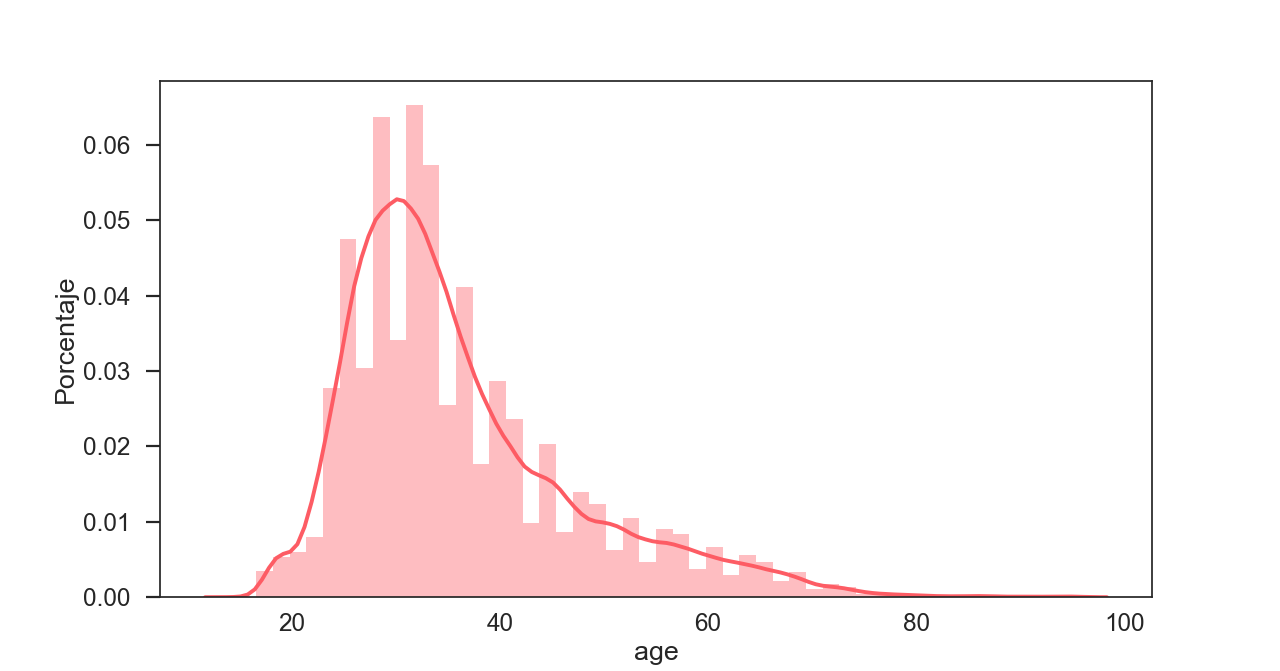
# **Etapa 2: Análisis Estadístico**

Después de haber llevado a cabo la exploración de los datos de las reservaciones de airbnb, se llegó a la conclusión, de que habían varias columnas que podrían ser interesantes de analizar, ya que se mostraba una relación entre estas y nuestro objetivo. La variable crítica en la que se centra nuestro análisis es la fecha del primer booking realizado por los usuarios. De esta variable se encontraron 3 aspectos que podrían ayudarnos a lograr el objetivo. Estas son: reservaciones por edad, reservación por mes y tiempo de la sesión para realizar una reservación.

## **Análisis estadístico descriptivo**

A continuación se presentará el análisis realizado con respecto a la variable de primera reservación del usuario (*first booking*) en tres diferentes aspectos: reservaciones por edad, reservación por mes y tiempo de la sesión para realizar una reservación.

En primera instancia estudiamos la primera reservación en base a la edad del usuario:



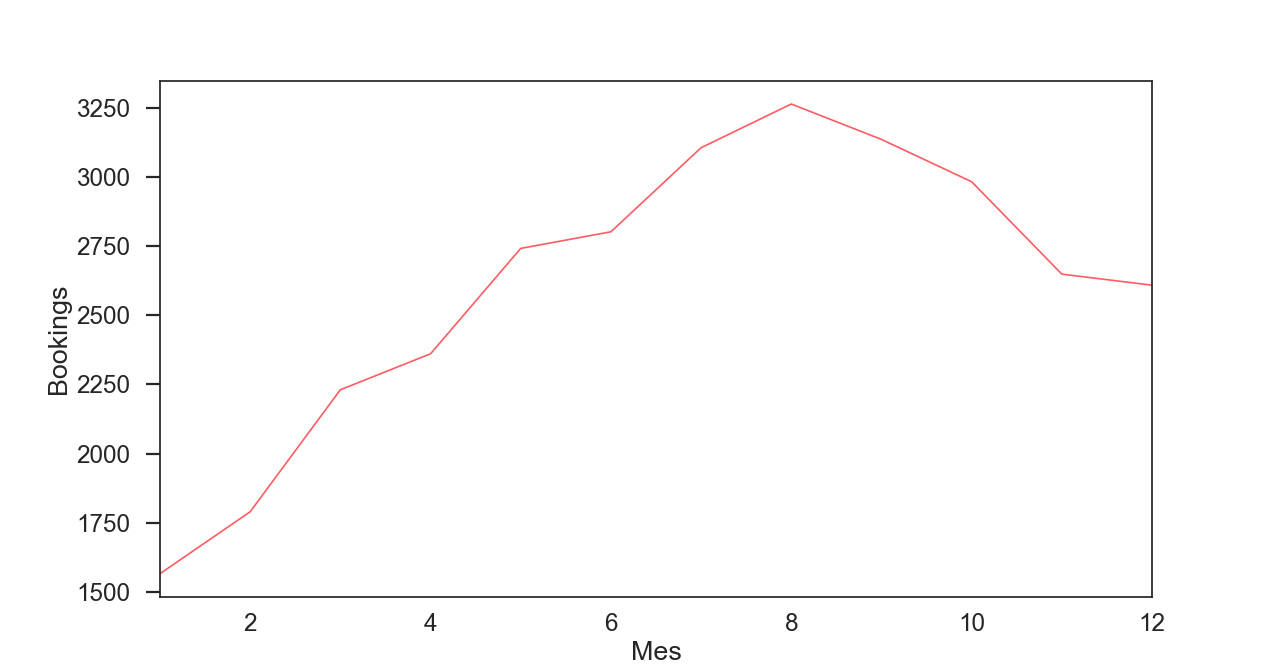
En la gráfica anterior se presenta un histograma donde se presentan los porcentajes correspondientes a la edad de los usuarios que realizaron su primer reservación. Realizando el análisis de los datos correspondientes a este histograma, se obtuvieron los siguientes valores:

**\*Age descriptive statistics\***

**rows: 22008 min and max: (15.0, 95.0) mean: 35.119 variance: 129.84 skewness: 1.2615 kurtosis: 1.3957**

A partir de estos resultados notamos que la distribución correspondiente a lo anterior tiene un parecido significante a una distribución de tipo lognormal.

Después analizamos la relación entre el número de primeras reservaciones realizadas y una determinada época del año. Para facilitar el estudio de los datos, decidimos agruparlos en meses.

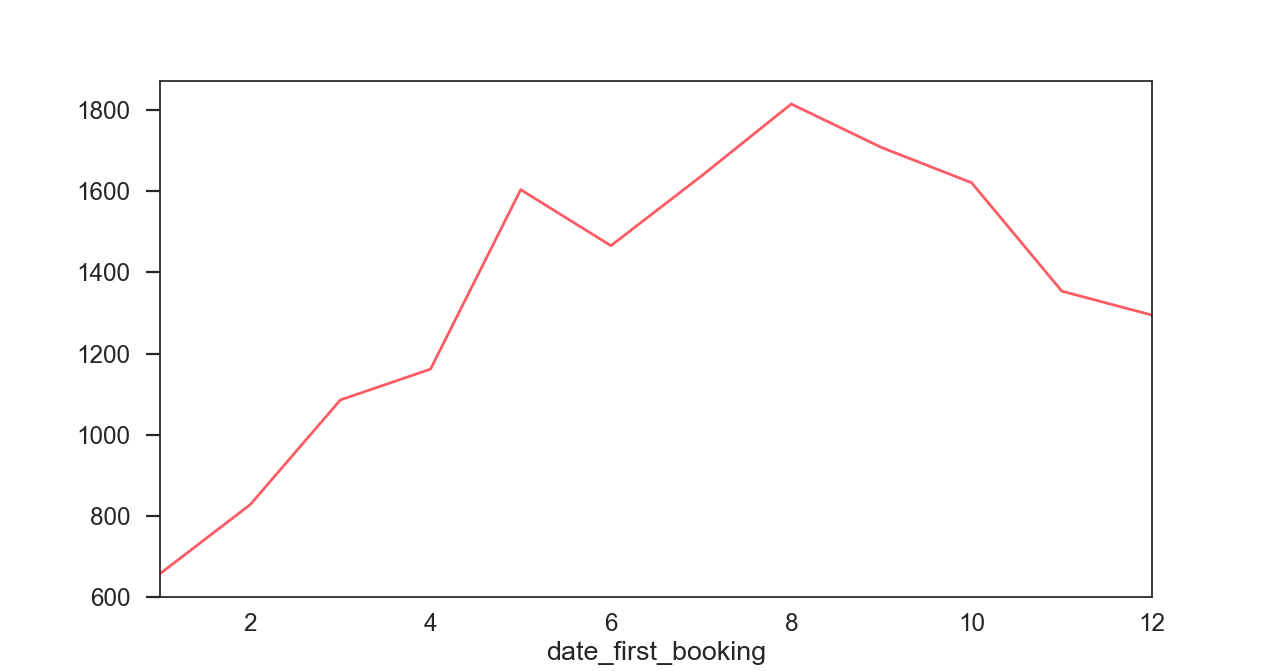


En la gráfica anterior se presentan las primeras reservaciones de los usuarios por mes, correspondiente al año 2013. Con base en este grupo de datos se llevó a cabo un análisis descriptivo de los datos y se obtuvieron los siguientes valores:

**\*Months 2013 descriptive statistics\***

**rows: 12 min and max: (1566, 3264) mean: 2603.2 variance: 2.8298e+05 skewness: -0.68595 kurtosis: -0.55355**

A manera de comparación, se llevó a cabo el análisis de los datos correspondientes al año 2012.



En la gráfica anterior, se presentan las primeras reservaciones por mes de los usuarios, correspondiente al año 2012. A partir del análisis descriptivo de los datos utilizados para este histograma, se obtuvieron los siguientes valores:

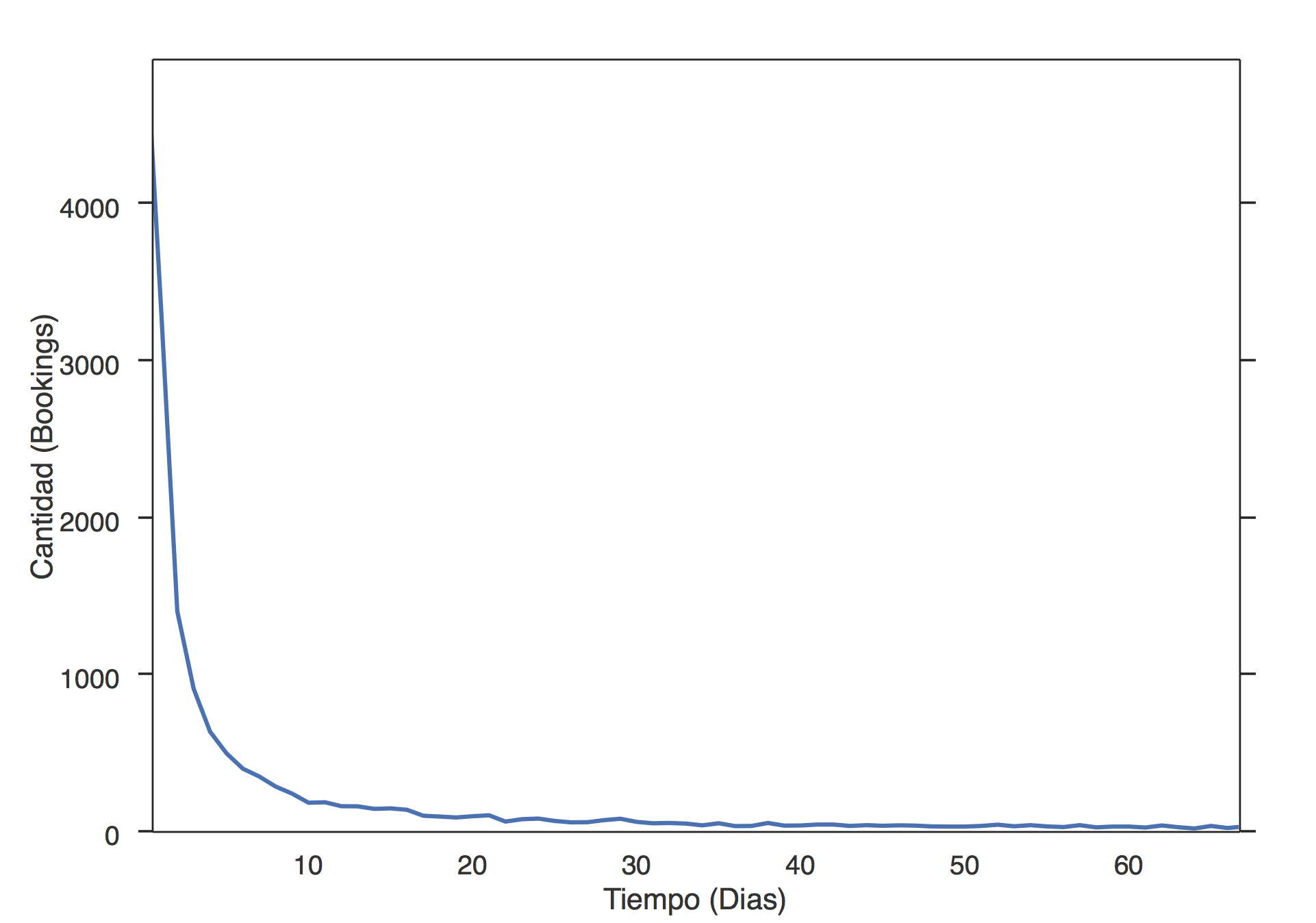
**\*Months 2012 descriptive statistics\***

**rows: 12 min and max: (658, 1815) mean: 1352.8 variance: 1.3068e+05 skewness: -0.60925 kurtosis: -0.71929**

Dados los resultados anteriores, podemos observar que la distribución manejada por esta relación tiene un parecido a una distribución de tipo normal o gamma.

Por último, estudiamos el vínculo entre tiempo empleado dentro de una sesión y la cantidad de primeras reservaciones realizadas.

Gráfica sesiones



En la gráfica anterior se presenta el tiempo que le tomó al usuario realizar una primera reservación desde que se creó la cuenta. Tras analizar los datos que se utilizaron en la gráfica, se obtuvieron los siguientes valores:

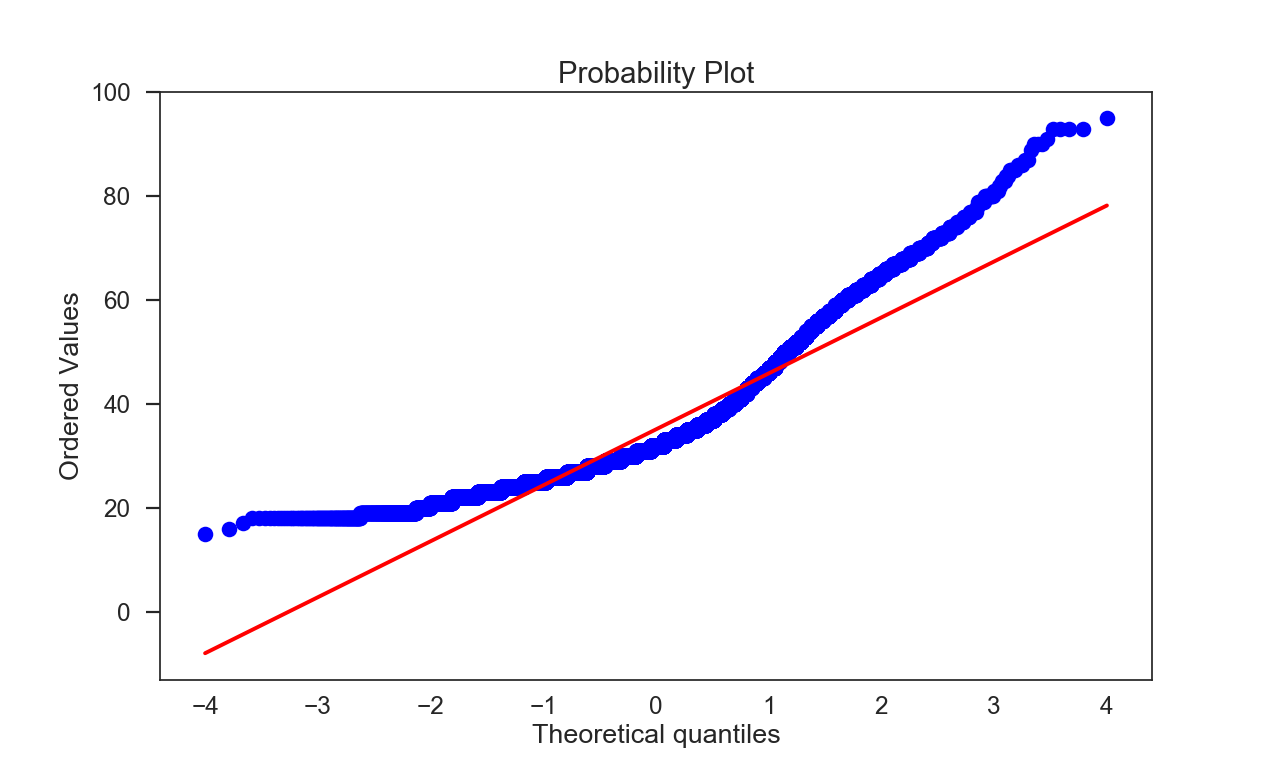
**\*Session duration descriptive statistics\***

**rows: 22008 min and max: (0, 365) mean: 50.872 variance: 8452.6 skewness: 1.9752 kurtosis: 2.7391**

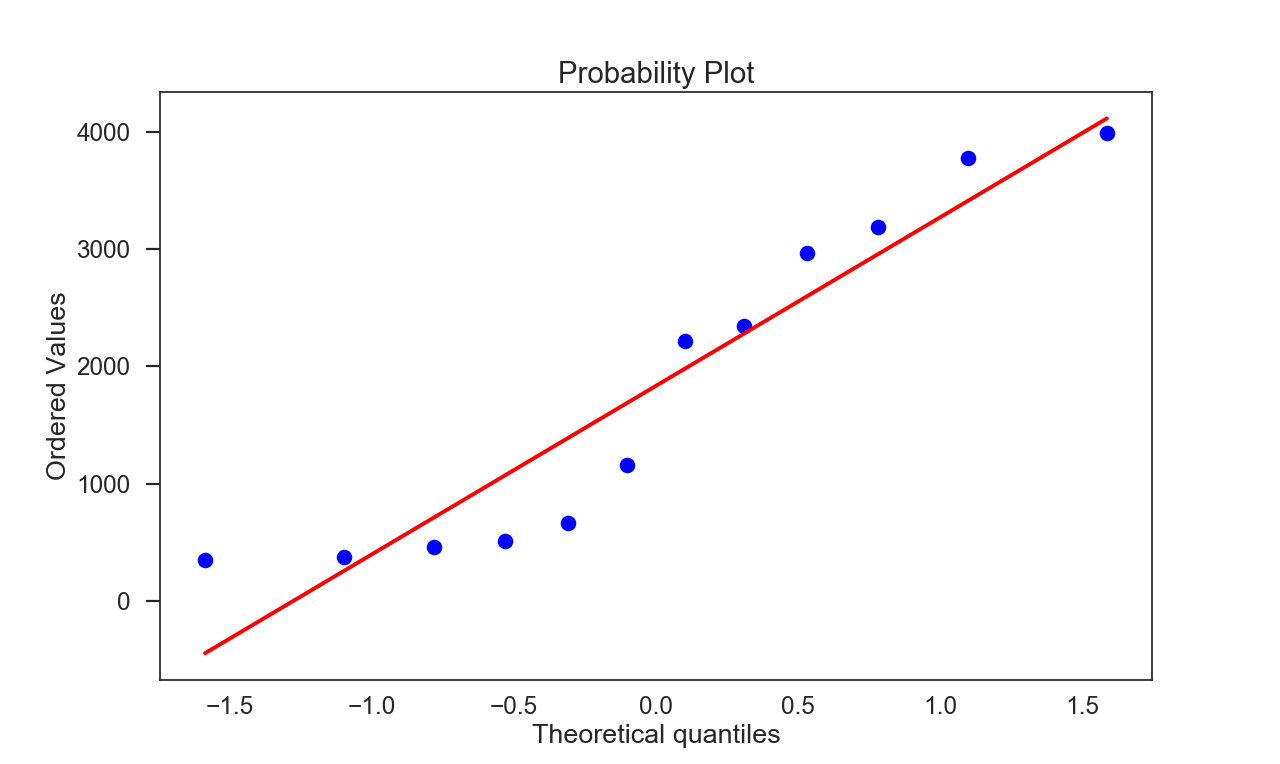
A partir del análisis descriptivo de los datos, se puede observar que la distribución de los datos anteriores tiene un parecido a una distribución de tipo normal o gamma.

## **Determinación de la normalidad de los datos**

Utilizando Python en conjunto con la librería Scipy, generamos el gráfico Cuantil-Cuantil (*Q-Q Plot*) para cada una de nuestras variables. Los C-C se crearon a partir de una distribución normal, por lo que nos fueron útiles en determinar si era necesario ajustar las distribuciones de nuestros datos.

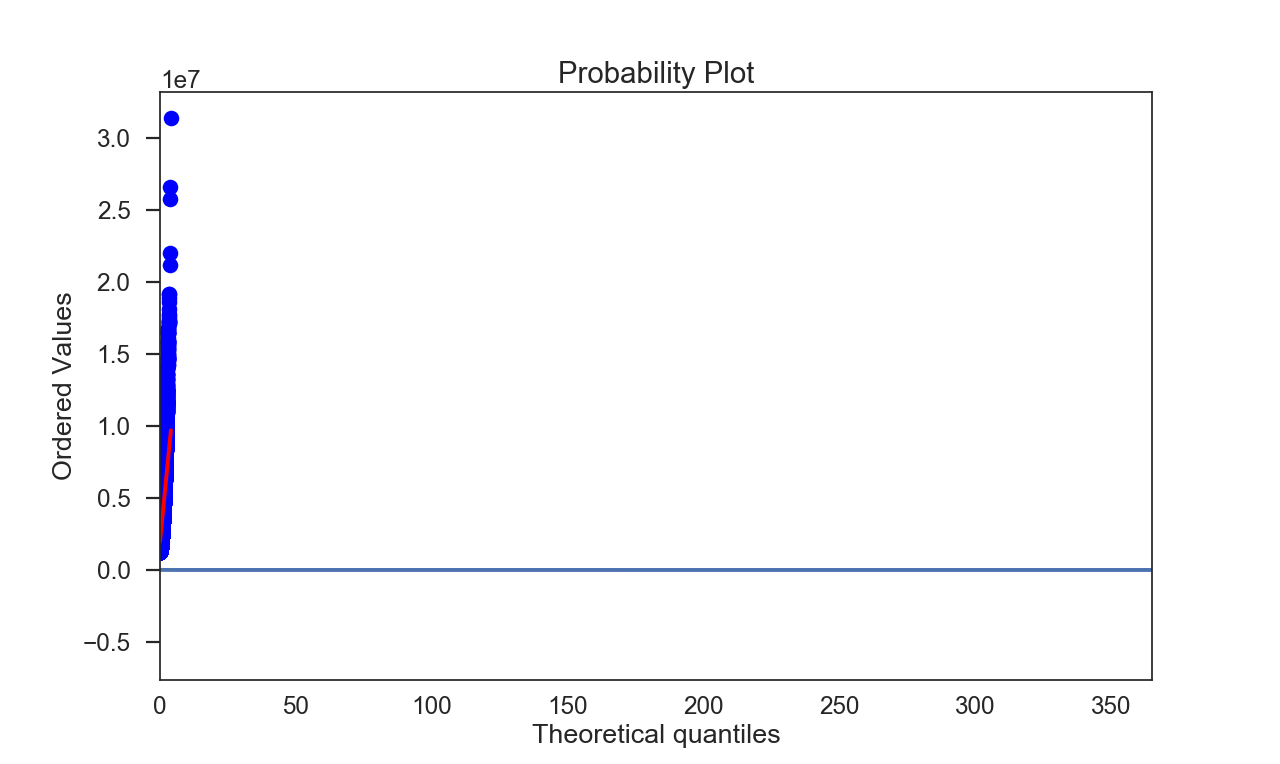


El gráfico C-C anterior corresponde a las primeras reservaciones realizadas de acuerdo a las edades de los usuarios. Podemos observar como la curva de nuestra distribución definitivamente no se comporta de forma similar a la de la normal. Por lo tanto, será necesario ajustarla.



La segunda gráfico muestra lo referente a las primeras reservaciones agendadas según la época del año (mes). Al igual que en el caso anterior, es evidente que la distribución de los datos difiere de la normal.

En el último gráfico de *sessions* podemos ver que la línea empata con una distribución normal. Esta conclusión la podemos obtener de acuerdo a la librería stats de python, que de acuerdo a un resumen, dicha probabilidad fue la mejor y por mucho. A continuación se muestran los resultados:



## 

## **Ajuste de distribuciones de probabilidad**

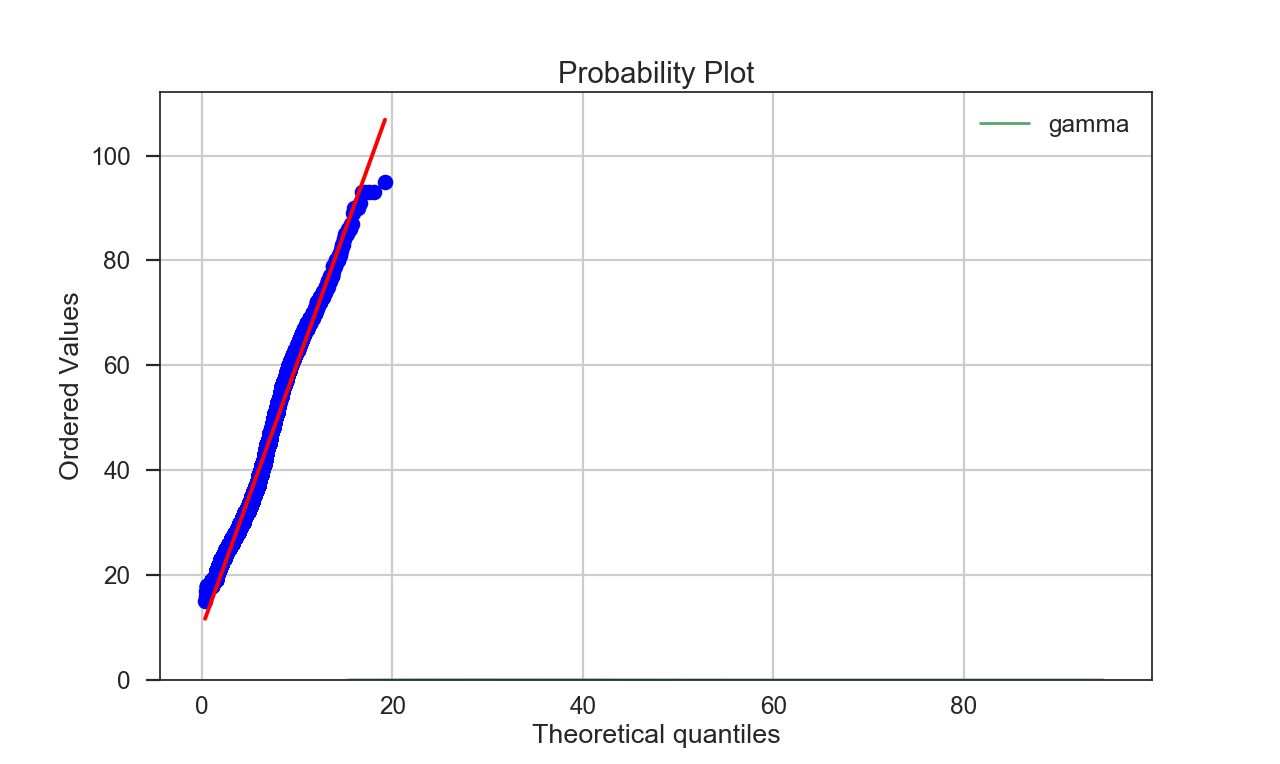
Dados los resultados anteriores, nos vimos en la necesidad realizar ajustes de distribución de probabilidad a nuestras variables. Decidimos nuevamente utilizar Python con este fin, en conjunto con la librería externa Fitter. Los resultados fueron los siguientes:

**\*Age best fit\***

**Fitted gamma distribution with error=0.0125265827934)**

**Fitted expon distribution with error=0.031248102592)  
lognorm was not fitted. no parameters available  
pareto was not fitted. no parameters available  
weibull\_min was not fitted. no parameters available**

Para el caso de las primeras reservaciones según la edad del usuario, encontramos que una distribución Gamma con k = 5 era el mejor ajuste dentro de las anteriores:



**\*Month best fit\***

**Fitted gamma distribution with error=5.60705344497e-05)  
Fitted expon distribution with error=0.000296074842948)  
Fitted norm distribution with error=6.83035217354e-05)  
Fitted weibull\_min distribution with error=5.11487942271e-05)  
Fitted pareto distribution with error=5.63857265574e-05)**

En segunda instancia, analizamos la variable de época del año (mes) y encontramos que la distribución que más se aproximaba a la suya era la Weibull con una k = 1.6.

En la última instancia (sessions) se obtuvieron los siguientes resultados:

SKIPPED gamma distribution (taking more than 30 seconds)

SKIPPED expon distribution (taking more than 30 seconds)

**Fitted norm distribution with error=3.441098494e-13)**

SKIPPED weibull\_min distribution (taking more than 30 seconds)

SKIPPED pareto distribution (taking more than 30 seconds)

expon was not fitted. no parameters available

gamma was not fitted. no parameters available

pareto was not fitted. no parameters available

weibull\_min was not fitted. no parameters available

Después de haberlos visto, pudimos concluir que el mejor *fit* es un ajuste normal debido a que valor tiende a 0.

# 

# 