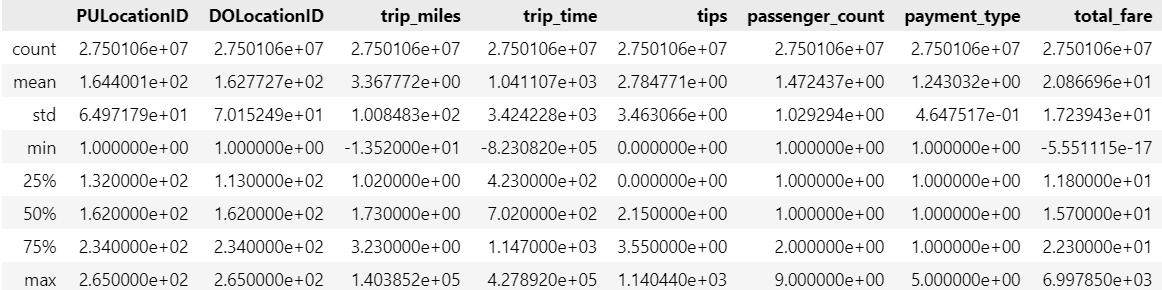
**Análisis de outliers**

En primera instancia, para optimizar el procesamiento, convertimos todas las variables numéricas a 32 bit (Originalmente estaban en 64 bit).

Para evitar distorsiones en la utilización de los datos en nuestras predicciones y cálculos, analizaremos algunas columnas significativas de nuestra tabla de taxis amarillos.

Inicialmente utilizamos el comando *describe()* que arroja el siguiente resultado:



Podemos observar valores extremos llamativos (máximos y mínimos) en las columnas *trip\_miles*, *trip\_time* y *total\_fare*.

A continuación, efectuaremos algunos análisis que nos ayudarán a establecer los criterios necesarios para filtrarlos.

La tarifa mínima a cobrar cuando tomamos un taxi amarillo en la ciudad de Nueva York es de 3 dólares, por lo cual eliminaremos todos aquellos registros con tarifas menores.

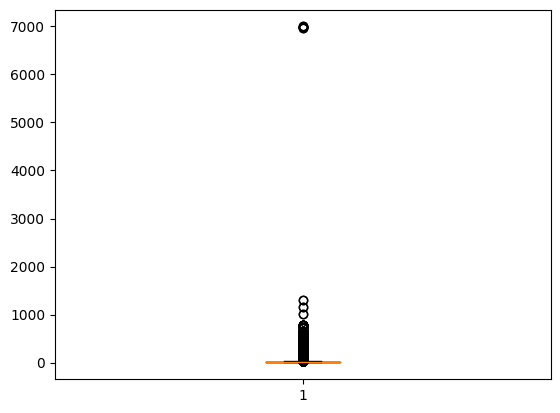
A partir de allí debemos sumar 70 centavos por cada quinto de milla recorrida (*https://www.nyc.gov/site/tlc/passengers/taxi-fare.page*). Por lo cual filtraremos todos aquellos valores que tengan registros en la columna *trip\_miles* menores a 0,2 millas.

Los registros con viajes en disputa (*payment\_type* = 4) en su gran mayoría contienen valores espejados de la tarifa cobrada. Es decir, un viaje donde se facturó un monto X y posteriormente otro donde se facturó el monto -X. En ambos casos el monto final a cobrar fue de $0. Eliminaremos también estos registros.

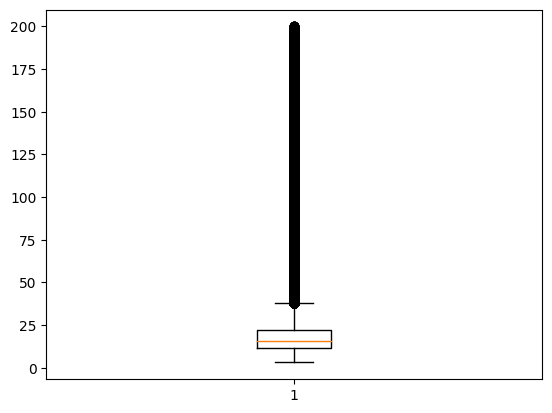
Adicionalmente eliminaremos aquellos registros con duración de viaje negativos.

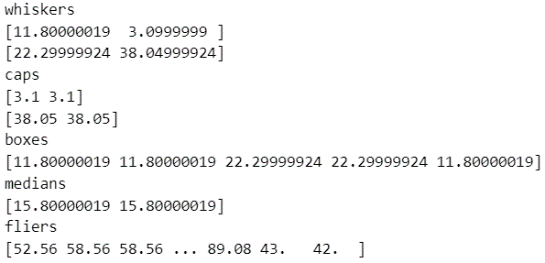
**Columna *total\_fare***

Diagrama de caja



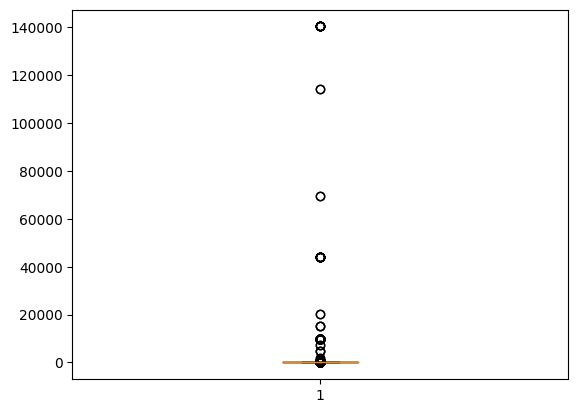
Para analizar mejor qué sucede en el gráfico, nos limitamos a ver los valores inferiores a 200 dólares.



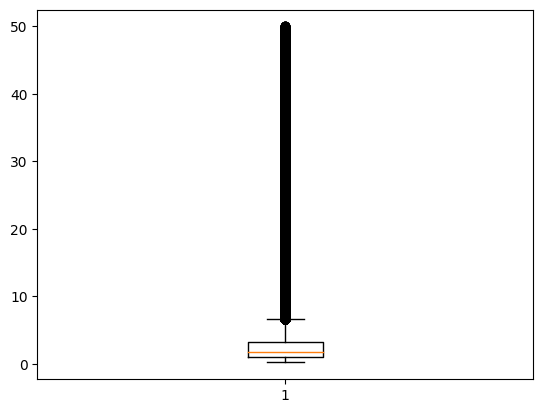


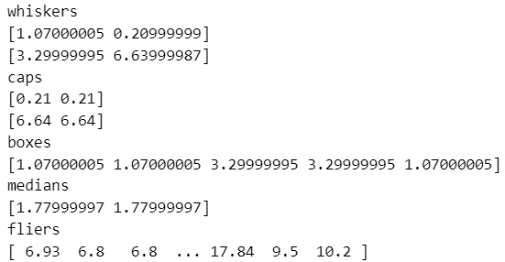
**Columna *trip\_miles***

Diagrama de caja



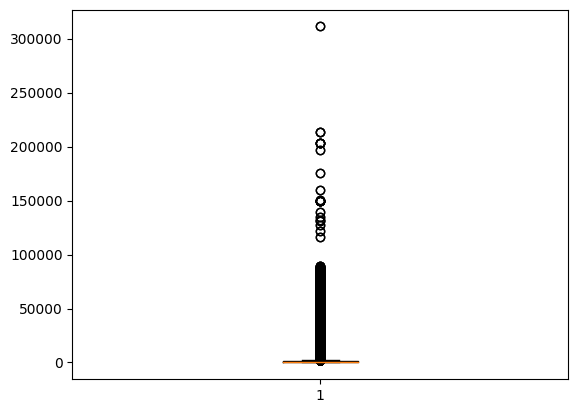
Para analizar mejor qué sucede en el gráfico, nos limitamos a ver los valores inferiores a 50 millas.



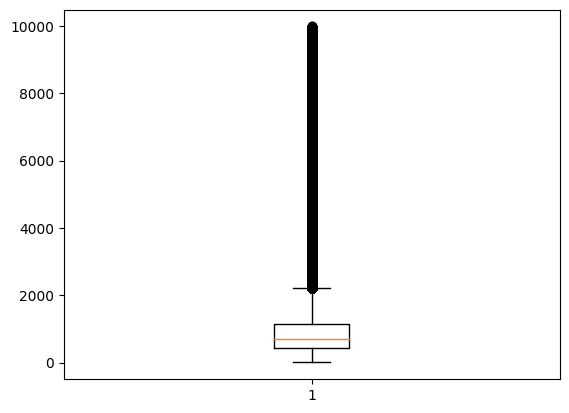
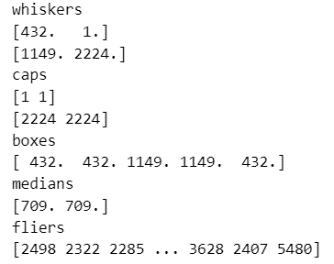


**Columna *trip\_time***

Diagrama de caja



Para analizar mejor qué sucede en el gráfico, nos limitamos a ver los valores inferiores a 10.000 segundos.



Finalmente, luego de haber analizado los diagramas de caja precedentes, fijamos los límites para aplicar los filtros correspondientes a nuestros registros.

* 200 > 'total\_fare' >= 3 (dólares norteamericanos)
* 50 > 'trip\_miles' >= 0.2 (millas)
* 5000 > 'trip\_time' >= 100 (segundos)