Creado por:

Isabel Maniega

Pandas y Modin

Objetivos:

- Aprender más sobre Modin
- Fomentar la proactividad del alumno

Tabla de contenidos:

Ordenador con el que se ha hecho el experimento

Características:

- Intel® Core™ i9-12900H CPU @ 2.40GHz × 4
- 32 GiB RAM DDR4
- 1T SSD m.2
- Sistema Operativo: Ubuntu
- Empleamos el archivo mencionado en el manual. (DATOS DEL 2018)

Pandas

In [1]:
```
# pip install pandas
```

In [2]:
```
%%time
import pandas as pd
```

```
CPU times: user 568 ms, sys: 1.3 s, total: 1.87 s
Wall time: 155 ms
```

In [3]:
```
%%time
df_pandas = pd.read_csv('yellow_tripdata.csv', low_memory=False)
df_pandas.head()
```

```
CPU times: user 7.54 s, sys: 1.47 s, total: 9.01 s
Wall time: 9 s
```

Out[3]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULc
0	0	1	2018-01-01 00:21:05	2018-01-01 00:24:23	1	0.5	1	N	
1	1	1	2018-01-01 00:44:55	2018-01-01 01:03:05	1	2.7	1	N	
2	2	1	2018-01-01 00:08:26	2018-01-01 00:14:21	2	0.8	1	N	
3	3	1	2018-01-01 00:20:22	2018-01-01 00:52:51	1	10.2	1	N	
4	4	1	2018-01-01 00:09:18	2018-01-01 00:27:06	2	2.5	1	N	

In [4]:
```
print(f"Number of Rows: {len(df_pandas.index)}")
print(f"Number of Columns: {len(df_pandas.axes[1])}")
```

```
Number of Rows: 8760687
Number of Columns: 20
```

In [5]:
```
%%time
df_pandas.describe()
```

```
CPU times: user 1.9 s, sys: 517 ms, total: 2.42 s
Wall time: 2.42 s
```

Out[5]:

	Unnamed: 0	VendorID	passenger_count	trip_distance	RatecodeID	PULocationID	DOLocationID	payment_type	fare_amour
count	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+0
mean	4.380343e+06	1.560978e+00	1.606807e+00	2.804022e+00	1.039545e+00	1.644579e+02	1.627270e+02	1.310613e+00	1.224443e+0
std	2.528993e+06	4.962678e-01	1.258420e+00	6.412050e+01	4.450619e-01	6.635990e+01	7.031145e+01	4.817808e-01	1.168321e+0
min	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	-4.500000e+0
25%	2.190172e+06	1.000000e+00	1.000000e+00	9.100000e-01	1.000000e+00	1.160000e+02	1.130000e+02	1.000000e+00	6.000000e+0
50%	4.380343e+06	2.000000e+00	1.000000e+00	1.550000e+00	1.000000e+00	1.620000e+02	1.620000e+02	1.000000e+00	9.000000e+0
75%	6.570514e+06	2.000000e+00	2.000000e+00	2.840000e+00	1.000000e+00	2.340000e+02	2.340000e+02	2.000000e+00	1.350000e+0
max	8.760686e+06	2.000000e+00	9.000000e+00	1.894838e+05	9.900000e+01	2.650000e+02	2.650000e+02	4.000000e+00	8.016000e+0

In [6]:
```
%%time
df_pandas.fare_amount.value_counts()
```

```
CPU times: user 51.1 ms, sys: 0 ns, total: 51.1 ms
Wall time: 49.8 ms
```

Out[6]:
```
6.00       473270
5.50       465207
6.50       461959
7.00       446414
5.00       433292
            ...
30.60           1
2409.00         1
168.88          1
201.50          1
33.96           1
Name: fare_amount, Length: 1714, dtype: int64
```

In [7]:
```
%%time
len(df_pandas), df_pandas.shape
```

```
CPU times: user 14 µs, sys: 0 ns, total: 14 µs
Wall time: 15.3 µs
```

Out[7]:
```
(8760687, (8760687, 20))
```

In [8]:
```
%%time
df_pandas.tail()
```

```
CPU times: user 56 µs, sys: 18 µs, total: 74 µs
Wall time: 75.8 µs
```

Out[8]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag
8760682	8760682	1	2018-01-31 23:21:35	2018-01-31 23:34:20	2	2.80	1	N
8760683	8760683	1	2018-01-31 23:35:51	2018-01-31 23:38:57	1	0.60	1	N
8760684	8760684	2	2018-01-31 23:28:00	2018-01-31 23:37:09	1	2.95	1	N
8760685	8760685	2	2018-01-31 23:24:40	2018-01-31 23:25:28	1	0.00	1	N
8760686	8760686	2	2018-01-31 23:28:16	2018-01-31 23:28:38	1	0.00	1	N

Modin

Modin y Ray

In [9]:
```
#pip install ray
```

In [10]:
```
# pip install modin
```

In [11]:
```
%%time
import ray
ray.init()
import modin
import modin.pandas as mpd
modin.config.Engine.put("ray")
```

```
2022-10-23 16:09:24,757 INFO worker.py:1518 -- Started a local Ray instance.
CPU times: user 208 ms, sys: 65.2 ms, total: 273 ms
Wall time: 1.43 s
```

Lectura de csv utilizando Modin-Ray

In [12]:
```
%%time
df_modin_ray = mpd.read_csv('yellow_tripdata.csv')
df_modin_ray.head()
```

```
UserWarning: When using a pre-initialized Ray cluster, please ensure that the runtime env sets environment variable __MODIN_AUTOIMPORT_PANDAS__ to 1
CPU times: user 293 ms, sys: 299 ms, total: 592 ms
Wall time: 3.36 s
```

Out[12]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULc
0	0	1	2018-01-01 00:21:05	2018-01-01 00:24:23	1	0.5	1	N	
1	1	1	2018-01-01 00:44:55	2018-01-01 01:03:05	1	2.7	1	N	
2	2	1	2018-01-01 00:08:26	2018-01-01 00:14:21	2	0.8	1	N	
3	3	1	2018-01-01 00:20:22	2018-01-01 00:52:51	1	10.2	1	N	
4	4	1	2018-01-01 00:09:18	2018-01-01 00:27:06	2	2.5	1	N	

In [13]:
```
%%time
df_modin_ray.describe()
```

```
CPU times: user 33.9 ms, sys: 4.61 ms, total: 38.6 ms
Wall time: 29 ms
```

Out[13]:

	Unnamed: 0	VendorID	passenger_count	trip_distance	RatecodeID	PULocationID	DOLocationID	payment_type	fare_amour
count	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+0
mean	4.380343e+06	1.560978e+00	1.606807e+00	2.804022e+00	1.039545e+00	1.644579e+02	1.627270e+02	1.310613e+00	1.224443e+0
std	2.528993e+06	4.962678e-01	1.258420e+00	6.412050e+01	4.450619e-01	6.635990e+01	7.031145e+01	4.817808e-01	1.168321e+0
min	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	-4.500000e+0
25%	2.190172e+06	1.000000e+00	1.000000e+00	9.100000e-01	1.000000e+00	1.160000e+02	1.130000e+02	1.000000e+00	6.000000e+0
50%	4.380343e+06	2.000000e+00	1.000000e+00	1.550000e+00	1.000000e+00	1.620000e+02	1.620000e+02	1.000000e+00	9.000000e+0
75%	6.570514e+06	2.000000e+00	2.000000e+00	2.840000e+00	1.000000e+00	2.340000e+02	2.340000e+02	2.000000e+00	1.350000e+0
max	8.760686e+06	2.000000e+00	9.000000e+00	1.894838e+05	9.900000e+01	2.650000e+02	2.650000e+02	4.000000e+00	8.016000e+0

In [14]:
```
%%time
df_modin_ray.fare_amount.value_counts()
```

```
CPU times: user 117 ms, sys: 36.3 ms, total: 153 ms
Wall time: 353 ms
UserWarning: sort_values is not currently supported by PandasOnRay, defaulting to pandas implementation.
Please refer to https://modin.readthedocs.io/en/stable/supported_apis/defaulting_to_pandas.html for explanation.
```

Out[14]:
```
6.00       473270
5.50       465207
6.50       461959
7.00       446414
5.00       433292
            ...
60.06           1
60.30           1
60.53           1
60.55           1
8016.00         1
Name: fare_amount, Length: 1714, dtype: int64
```

In [15]:
```
%%time
len(df_modin_ray), df_modin_ray.shape
```

```
CPU times: user 61 µs, sys: 21 µs, total: 82 µs
Wall time: 87.7 µs
```

Out[15]:
```
(8760687, (8760687, 20))
```

In [16]:
```
%%time
df_modin_ray.tail()
```

```
CPU times: user 1.16 ms, sys: 393 µs, total: 1.55 ms
Wall time: 1.23 ms
```

Out[16]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag
8760682	8760682	1	2018-01-31 23:21:35	2018-01-31 23:34:20	2	2.80	1	N
8760683	8760683	1	2018-01-31 23:35:51	2018-01-31 23:38:57	1	0.60	1	N
8760684	8760684	2	2018-01-31 23:28:00	2018-01-31 23:37:09	1	2.95	1	N
8760685	8760685	2	2018-01-31 23:24:40	2018-01-31 23:25:28	1	0.00	1	N
8760686	8760686	2	2018-01-31 23:28:16	2018-01-31 23:28:38	1	0.00	1	N

Modin y Dask

In [17]:
```
# pip install dask[distributed]
```

In [18]:
```
%%time
import dask
from dask.distributed import Client
client = Client()
import modin
import modin.pandas as mpd
modin.config.Engine.put("dask")
```

```
CPU times: user 152 ms, sys: 75.2 ms, total: 227 ms
Wall time: 622 ms
```

Lectura de csv utilizando Modin-Dask

In [19]:
```
%%time
df_modin_dask = mpd.read_csv('yellow_tripdata.csv')
df_modin_dask.head()
```

```
CPU times: user 611 ms, sys: 445 ms, total: 1.06 s
Wall time: 3.94 s
```

Out[19]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULc
0	0	1	2018-01-01 00:21:05	2018-01-01 00:24:23	1	0.5	1	N	
1	1	1	2018-01-01 00:44:55	2018-01-01 01:03:05	1	2.7	1	N	
2	2	1	2018-01-01 00:08:26	2018-01-01 00:14:21	2	0.8	1	N	
3	3	1	2018-01-01 00:20:22	2018-01-01 00:52:51	1	10.2	1	N	
4	4	1	2018-01-01 00:09:18	2018-01-01 00:27:06	2	2.5	1	N	

In [20]:
```
%%time
df_modin_dask.describe()
```

```
CPU times: user 209 ms, sys: 14.1 ms, total: 224 ms
Wall time: 221 ms
2022-10-23 16:10:21,294 - distributed.worker_memory - WARNING - Worker is at 81% memory usage. Pausing worker. Process memory: 5.06 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:22,630 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.62 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:22,631 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.62 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:22,728 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.60 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:22,826 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.56 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:22,926 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.69 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,025 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,127 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.57 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,227 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.69 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,326 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.61 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,426 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,526 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.61 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,628 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,727 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.61 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,827 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:23,927 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.69 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,025 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,154 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.60 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,226 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.56 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,325 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.57 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,425 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.57 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,527 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.56 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,627 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.56 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,726 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,826 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.67 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:24,925 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,026 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.67 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,127 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,226 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.67 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,327 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,425 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.57 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,527 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.70 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,627 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.57 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,727 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.63 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:25,826 - distributed.worker_memory - WARNING - Unmanaged memory use is high. This may indicate a memory leak or the memory may not be released to the OS; see https://distributed.dask.org/en/latest/worker-memory.html#memory-not-released-back-to-the-os for more information. -- Unmanaged memory: 5.76 GiB -- Worker memory limit: 6.21 GiB
2022-10-23 16:10:26,147 - distributed.worker_memory - WARNING - Worker is at 1% memory usage. Resuming worker. Process memory: 117.24 MiB -- Worker memory limit: 6.21 GiB
```

Out[20]:

	Unnamed: 0	VendorID	passenger_count	trip_distance	RatecodeID	PULocationID	DOLocationID	payment_type	fare_amour
count	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+0
mean	4.380343e+06	1.560978e+00	1.606807e+00	2.804022e+00	1.039545e+00	1.644579e+02	1.627270e+02	1.310613e+00	1.224443e+0
std	2.528993e+06	4.962678e-01	1.258420e+00	6.412050e+01	4.450619e-01	6.635990e+01	7.031145e+01	4.817808e-01	1.168321e+0
min	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	-4.500000e+0
25%	2.190172e+06	1.000000e+00	1.000000e+00	9.100000e-01	1.000000e+00	1.160000e+02	1.130000e+02	1.000000e+00	6.000000e+0
50%	4.380343e+06	2.000000e+00	1.000000e+00	1.550000e+00	1.000000e+00	1.620000e+02	1.620000e+02	1.000000e+00	9.000000e+0
75%	6.570514e+06	2.000000e+00	2.000000e+00	2.840000e+00	1.000000e+00	2.340000e+02	2.340000e+02	2.000000e+00	1.350000e+0
max	8.760686e+06	2.000000e+00	9.000000e+00	1.894838e+05	9.900000e+01	2.650000e+02	2.650000e+02	4.000000e+00	8.016000e+0

In [21]:
```
%%time
df_modin_dask.fare_amount.value_counts()
```

```
UserWarning: sort_values is not currently supported by PandasOnDask, defaulting to pandas implementation.
CPU times: user 830 ms, sys: 65.4 ms, total: 896 ms
Wall time: 2.98 s
```

Out[21]:
```
6.00       473270
5.50       465207
6.50       461959
7.00       446414
5.00       433292
            ...
60.06           1
60.30           1
60.53           1
60.55           1
8016.00         1
Name: fare_amount, Length: 1714, dtype: int64
```

In [22]:
```
%%time
len(df_modin_dask), df_modin_dask.shape
```

```
CPU times: user 53 µs, sys: 18 µs, total: 71 µs
Wall time: 76.8 µs
```

Out[22]:
```
(8760687, (8760687, 20))
```

In [23]:
```
%%time
df_modin_dask.tail()
```

```
CPU times: user 2.75 ms, sys: 909 µs, total: 3.66 ms
Wall time: 1.71 ms
```

Out[23]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag
8760682	8760682	1	2018-01-31 23:21:35	2018-01-31 23:34:20	2	2.80	1	N
8760683	8760683	1	2018-01-31 23:35:51	2018-01-31 23:38:57	1	0.60	1	N
8760684	8760684	2	2018-01-31 23:28:00	2018-01-31 23:37:09	1	2.95	1	N
8760685	8760685	2	2018-01-31 23:24:40	2018-01-31 23:25:28	1	0.00	1	N
8760686	8760686	2	2018-01-31 23:28:16	2018-01-31 23:28:38	1	0.00	1	N

Creado por:

Isabel Maniega