شمایی کلی از دیتاست

	id	vendor_id	pickup_datetime	dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	trip_duration
836400	id3231567	2	2016-01-14 16:30:58	2016-01-14 16:42:10	2	-73.996132	40.743549	-74.006248	40.714748	N	672
1264281	id0842764	2	2016-02-09 16:19:57	2016-02-09 16:24:39	1	-73.967651	40.762859	-73.975113	40.750938	N	282
709782	id1500532	1	2016-02-04 17:30:26	2016-02-04 17:31:45	1	-73.950150	40.783993	-73.949257	40.785351	N	79
34961	id2704451	1	2016-02-21 18:42:06	2016-02-21 18:58:00	1	-73.991570	40.770222	-73.995056	40.749840	N	954
1093180	id2415554	1	2016-02-17 13:19:01	2016-02-17 13:46:25	3	-74.014061	40.713619	-74.006592	40.709755	N	1644
524299	id0714338	2	2016-05-31 23:21:01	2016-05-31 23:26:11	5	-73.974121	40.762959	-73.962410	40.759071	N	310
1227979	id2148140	2	2016-02-09 14:37:41	2016-02-09 15:03:48	1	-73.960838	40.765572	-73.999092	40.718941	N	1567
984753	id1665161	1	2016-03-19 10:32:05	2016-03-19 10:41:00	1	-73.985725	40.740936	-73.996666	40.732208	N	535
1413040	id1879384	2	2016-05-28 02:26:39	2016-05-28 02:35:54	1	-74.000641	40.735718	-73.978302	40.745621	N	555
23160	id2462720	2	2016-02-14 08:17:27	2016-02-14 08:25:40	1	-74.005623	40.725639	-73.984871	40.748112	N	493
759777	id3948018	2	2016-05-16 12:47:26	2016-05-16 13:41:09	1	-73.978821	40.763988	-73.865211	40.770668	N	3223
1271745	id0862679	2	2016-04-12 07:11:29	2016-04-12 07:17:57	1	-73.986626	40.734127	-73.976074	40.755611	N	388

اطلاعات ستون‌های دیتاست:

	Name	Count	Type
0	id	1458644 non-null	object
1	vendor_id	1458644 non-null	int64
2	pickup_datetime	1458644 non-null	object
3	dropoff_datetime	1458644 non-null	object
4	passenger_count	1458644 non-null	int64
5	pickup_longitude	1458644 non-null	float64
6	pickup_latitude	1458644 non-null	float64
7	dropoff_longitude	1458644 non-null	float64
8	dropoff_latitude	1458644 non-null	float64
9	store_and_fwd_flag	1458644 non-null	object
10	trip_duration	1458644 non-null	int64

همانطور که مشاهده میشود، دیتاست شامل 1458644 سطر میباشد که تمامی سطرها دارای 11 ویژگی منحصر به فرد میباشد که اطلاعات آنها موجود است.

توضیحات مربوط به هر ویژگی:

- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - date and time when the meter was engaged
- dropoff_datetime - date and time when the meter was disengaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged
- store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- trip_duration - duration of the trip in seconds

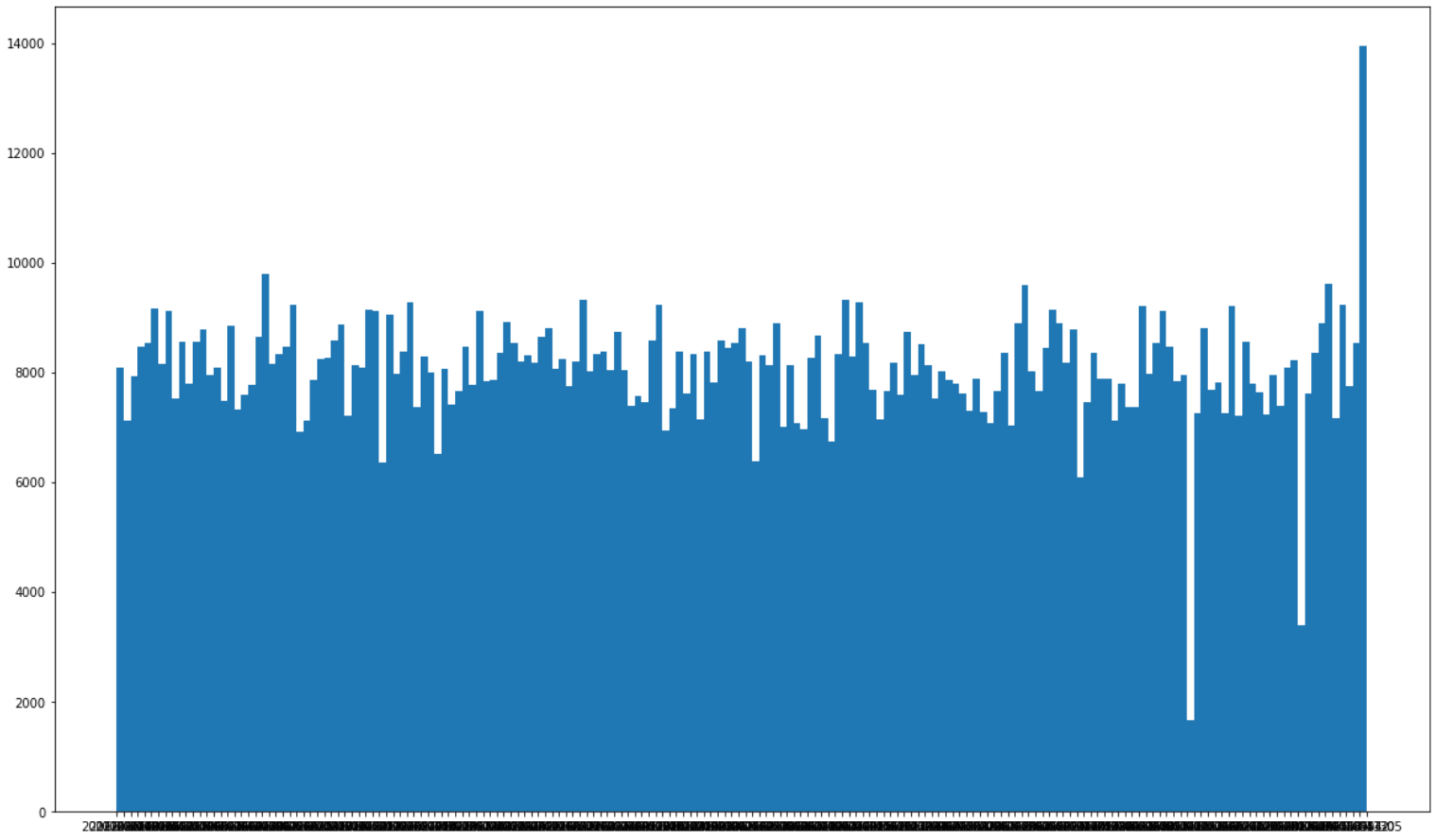
بررسی ویژگی‌های:

1) **pickup_datetime**

Min: 2016-01-01 00:00:17

Max: 2016-06-30 23:59:39

توزیع داده‌ها طبق pickup_datetime:

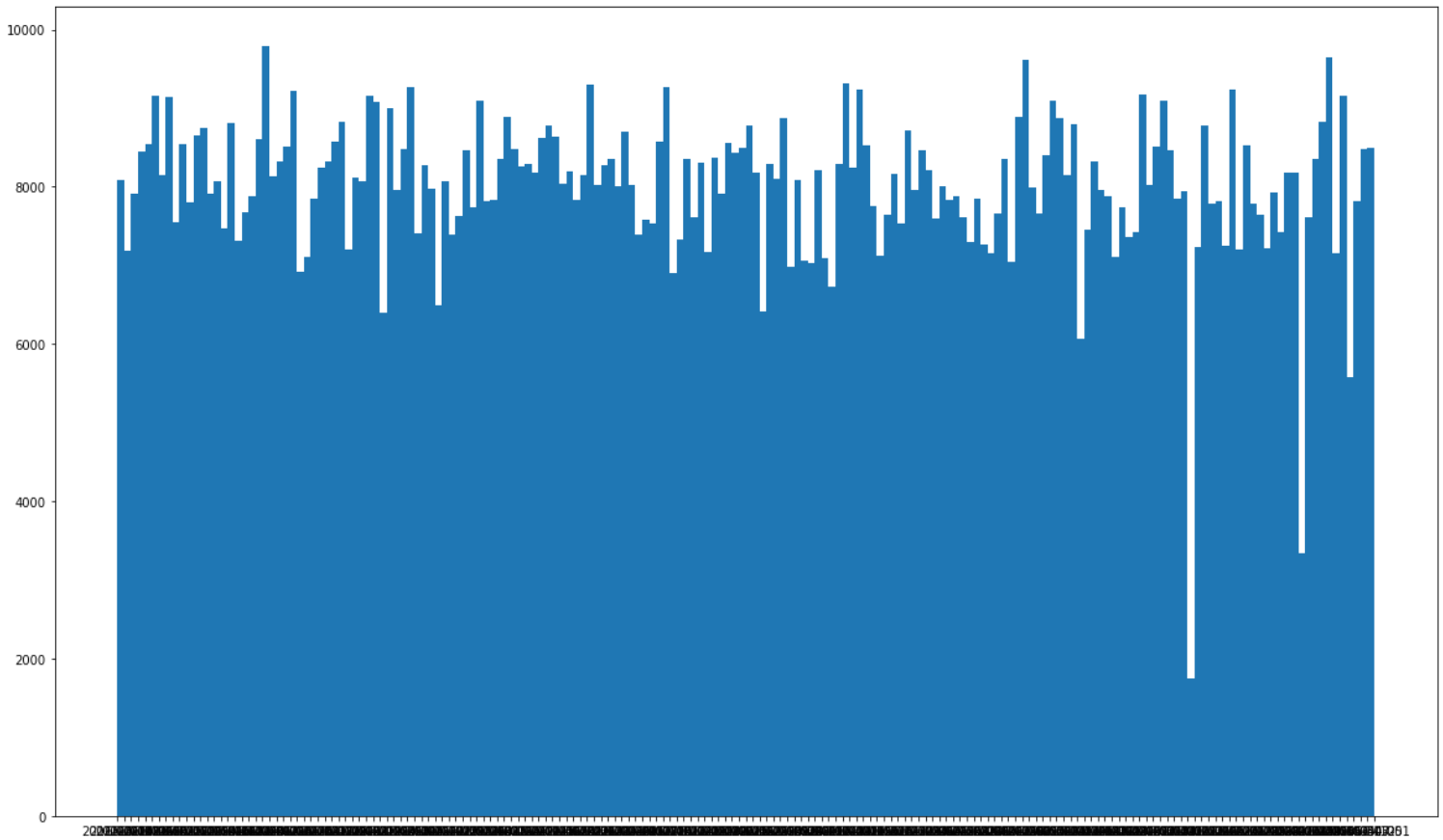


2) **dropoff_datetime**

Min: 2016-01-01 00:03:31

Max: 2016-07-01 23:02:03

توزیع داده‌ها طبق dropoff_datetime:

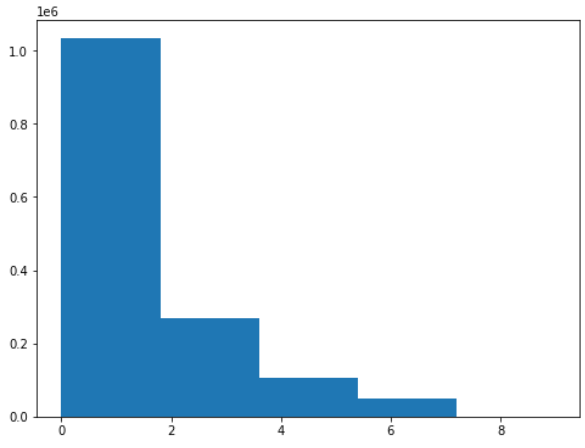


3) **passenger_count**

Min: 0

Max: 9

توزیع داده‌ها طبق passenger_count:



4) **pickup_longitude**

Min: -121.933342

Max: -61.335529

5) **pickup_latitude**

Min: 34.359695

Max: 51.881084

6) **dropoff_longitude**

Min: -121.933304

Max: -61.335529

7) **dropoff_latitude**

Min: 32.181141

Max: 43.921028

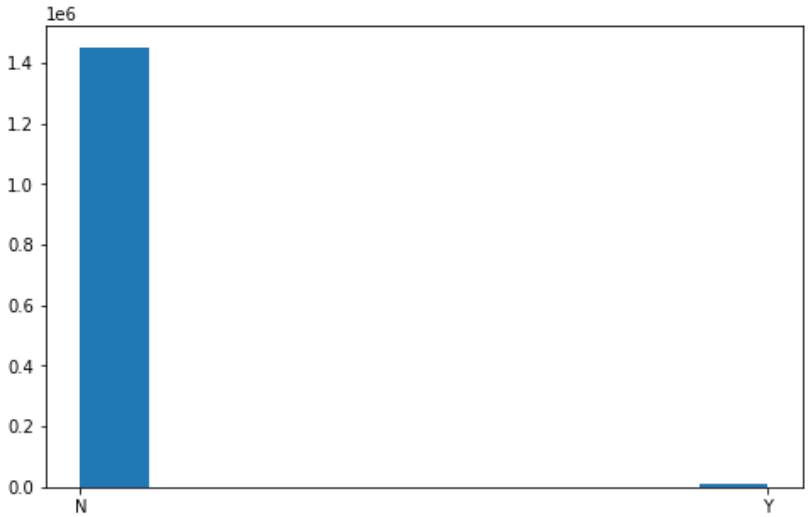
8) **store_and_fwd_flag**

Values: N, Y

Count N: 1450599

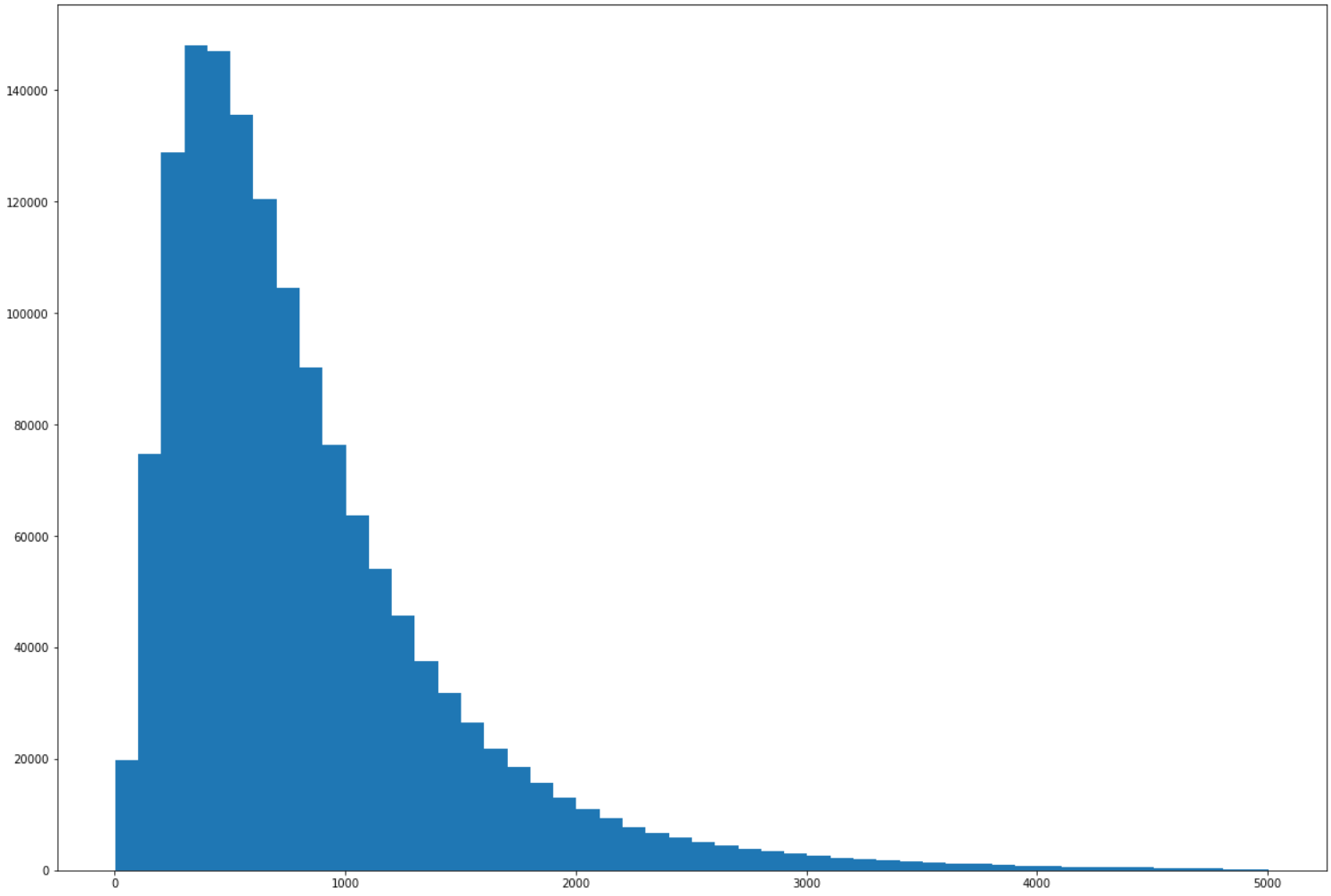
Count Y: 8045

توزیع داده‌ها طبق store_and_fwd_flag:



9) **trip_duration**

Mean	**959.4923**
Std	**5237.432**
Min	**1**
25%	**397**
50%	**662**
75%	**1075**
Max	**3526282**



جداسازی و محاسبه اطلاعات لازم برای بررسی زمانی-مکانی:

برای اطلاعات زمانی ستون trip_duration کافی است و این ستون را در دیتاست جدید که برای ساخت مدل لازم داریم، قرار میدهیم.

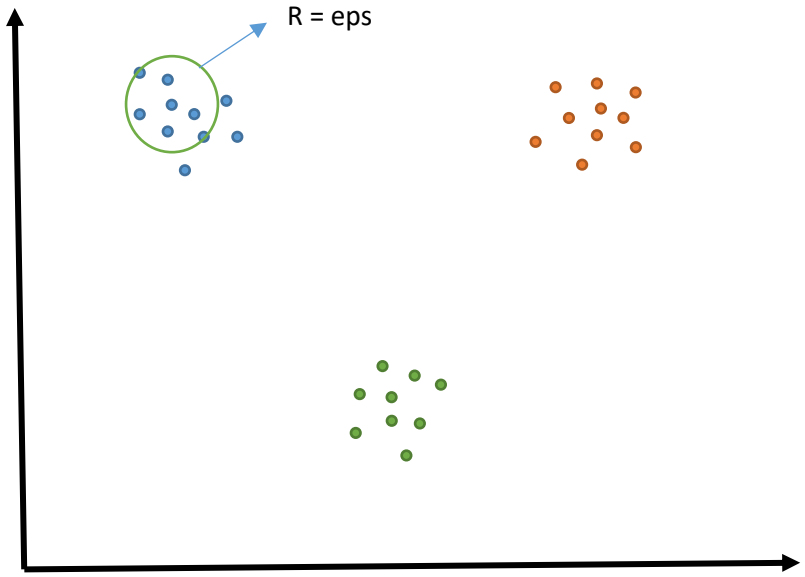
اطلاعات مکانی هم که به صورت طول و عرض جغرافیایی محل سوار شدن مسافر و پیاده شدن آن در دسترس است. با استفاده فرمول زیر میتوانیم فاصله اقلیدسی این دو نقطه را بدست آوریم:

$$d = \sqrt{\left(\text{dropoff}_{\text{longitude}} - \text{pickup}_{\text{longitude}}\right)^2 + \left(\text{dropoff}_{\text{latitude}} - \text{pickup}_{\text{latittude}}\right)^2}$$

حال با داشتن trip_duration و distance میتوان گروه بندی را انجام داد.

برای گروهبندی از الگوریتم DBScan استفاده میکنیم. این الگوریتم برای ایجاد مدل از روی اطلاعات نیاز به دو پارامتر از پیش تعیین شده دارد: eps, min_samples که eps به معنی شعاع بررسی میباشد و min_samples به معنای حداقل اعضای درون دایره رسم شده میباشد.

نحوه عملکرد الگوریتم DBScan به صورت زیر میباشد:



مطابق شکل بالا، روی هر کدام از اعضای مجموعه داده، دایره ای به شعاع eps زده میشود و اگر تعداد عضوهای درون دایره به تعداد min_samples میرسید، آنها را به عنوان یک گروه تعیین میکنند. سپس روی اعضای این گروه اینکار تکرار میشود تا تمامی داده‌های نزدیک در گروه های مناسب قرار بگیرند.

روی اطلاعات بدست آمده از دیتاست، بهترین پارامترها به شرح زیر میباشد:

'algorithm': 'auto',

'eps': 1.95,

'leaf_size': 30,

'metric': 'euclidean',

'metric_params': None,

'min_samples': 300,

'n_jobs': None,

' p': None

با این پارامترها، مدل نمره 2750 **را از الگوریتم** Calinski_Harabasz **دریافت میکند.**