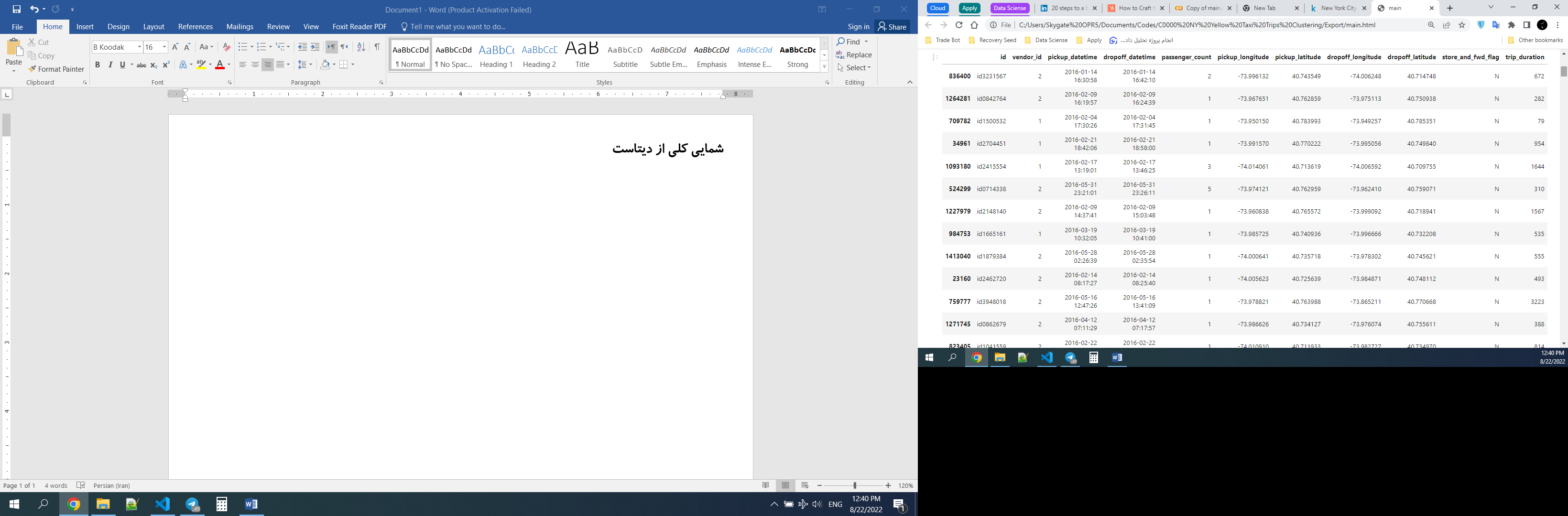
شمایی کلی از دیتاست



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

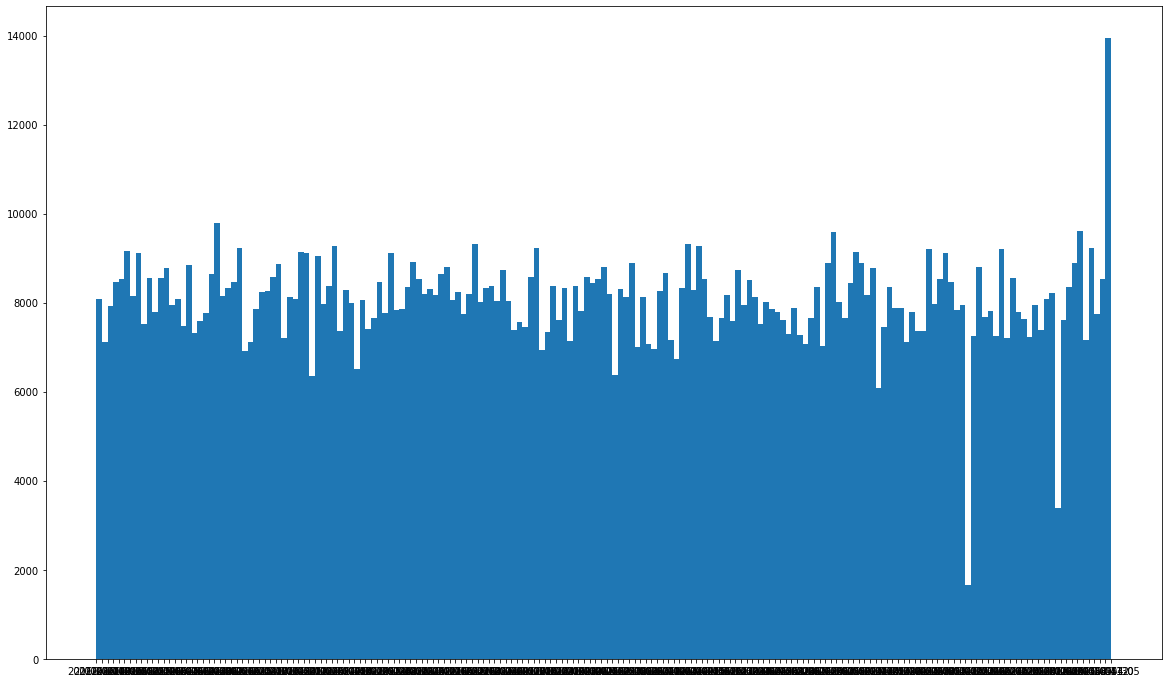
بررسی ویژگی­های:

1) **pickup\_datetime**

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

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

توزیع داده­ها طبق **pickup\_datetime:**

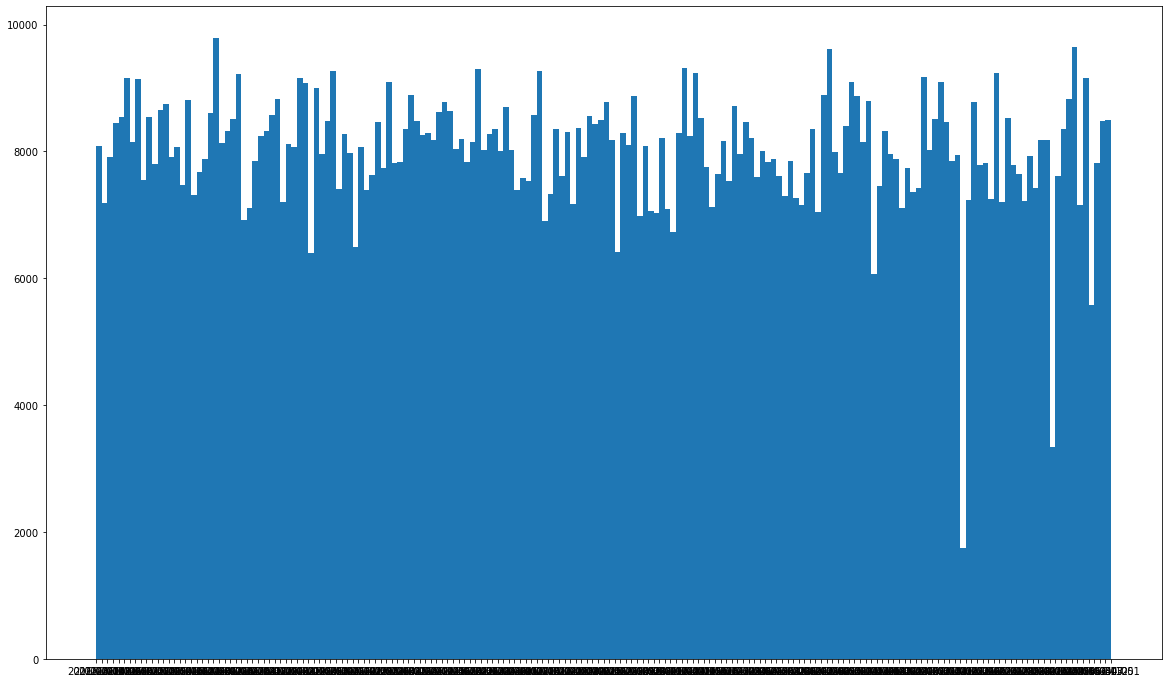


2) **dropoff\_datetime**

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

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

توزیع داده­ها طبق **dropoff\_datetime:**

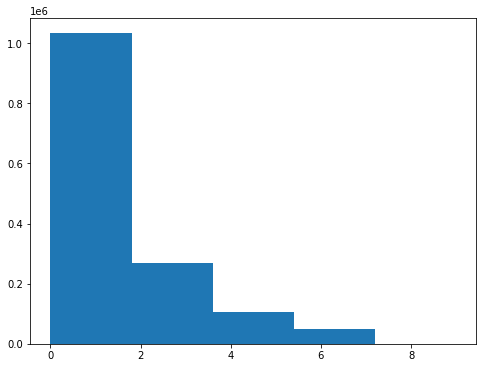


3) **passenger\_count**

**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**

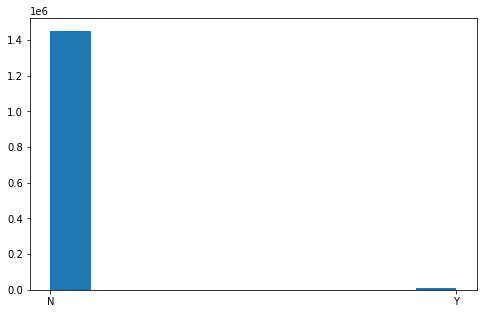
8) **store\_and\_fwd\_flag**

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**

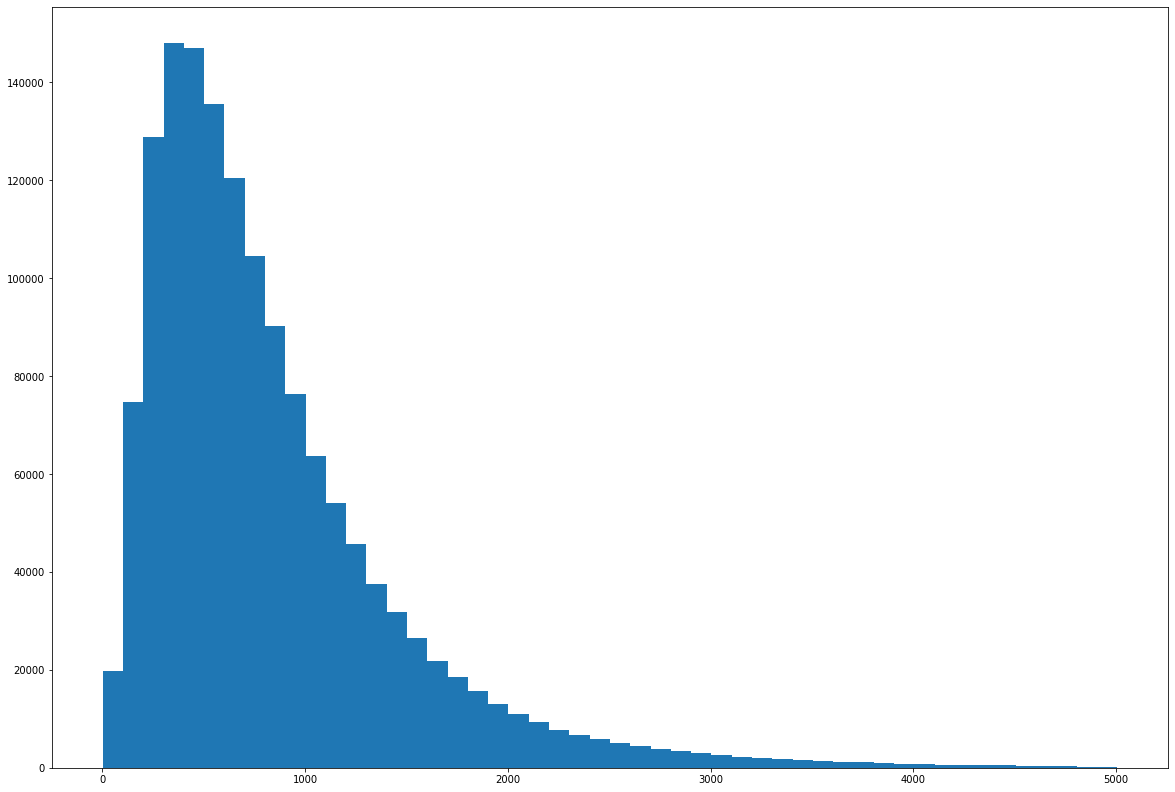
**Min 1**

**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:

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:

R = eps

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.