# Tùy Chỉnh Vải Treo Tường Decoals New York Poster NY Taxi NO.1 Dán Tường Đường Taxi Giấy Dán Tường Văn Phòng Bức Tranh Tường Chrismas Trang Trí|new york poster|wall stickernew york -K-MEAN CLUSTTERING WITH NEW YORK CITY TAXI TRIP DURATION

Data fields

* 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 the vehicle's 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

Tổng quan:

Dữ liệu gồm thông tin chuyến đi của 2 hãng xe giấu tên trong 6 tháng đầu năm 2016 tại New York: ta sẽ chia nhóm để kiểm soát từng nhóm >> tạo hàm phân loại để set các đơn vị tương tự trong tương lai vào các nhóm đã chia.

Các bước thực hiện:

1.Import, clean data

2.Biến đổi data

3.Sử dụng K-Mean Clustering chia nhóm data

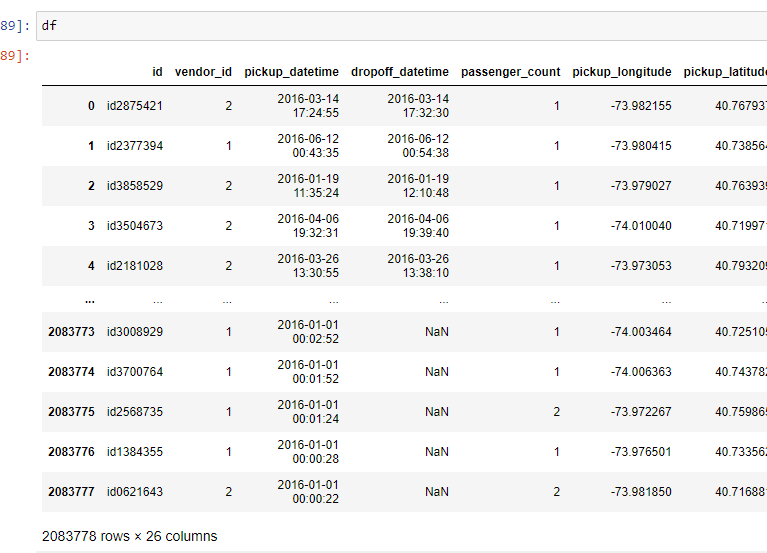
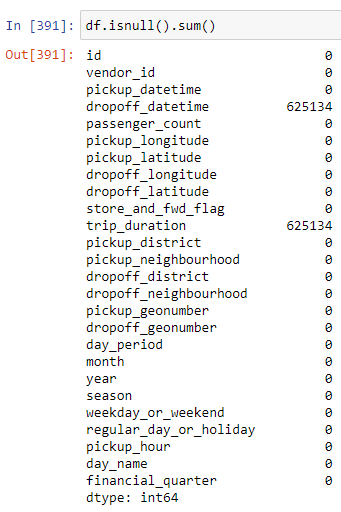
4.EDA, phân tích dữ liệu từng nhóm

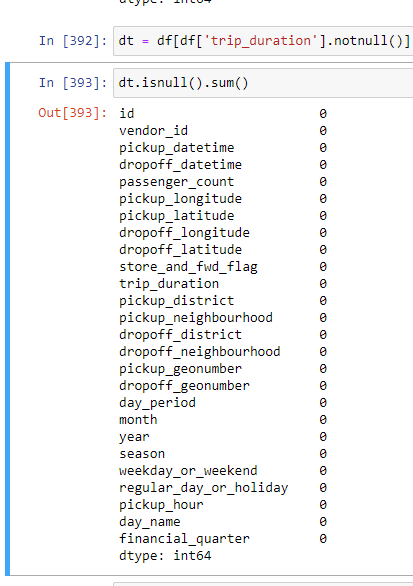
5.Lập hàm để tự động phân loại data sau này



1.Import, clean data

Link data = [New York City Taxi Trip Duration Extended | Kaggle](https://www.kaggle.com/datasets/neomatrix369/nyc-taxi-trip-duration-extended)

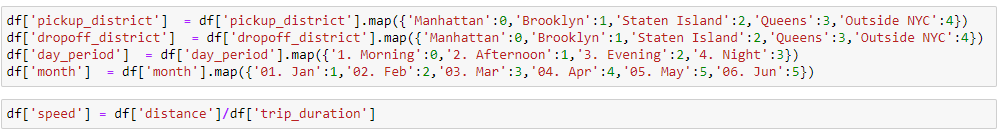




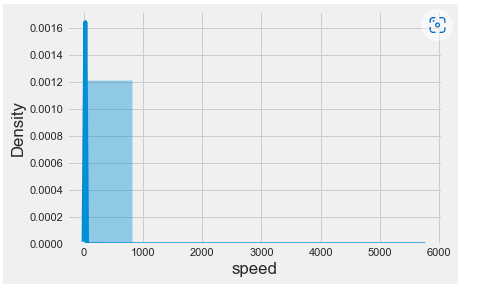
2.Biến đổi dữ liệu

Đổi thời gian từ giây sang giờ



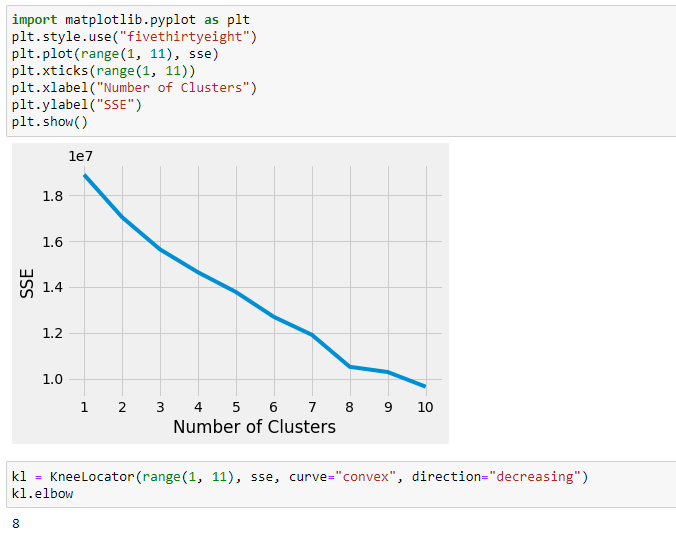


Thêm cột vận tốc để dễ học hơn.

Từ đây ta thấy được các dữ liệu đáng ngờ

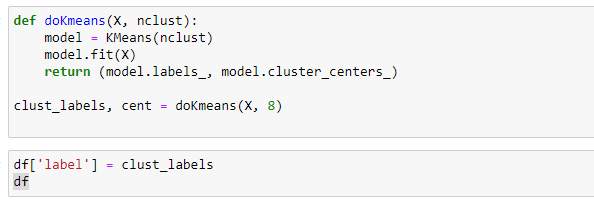
Nhưng mình không clean ngay, mình sẽ để máy tính làm việc đó, K-Mean có tương tác với các outliner.

3.Sử dụng K-Mean Clustering chia nhóm data

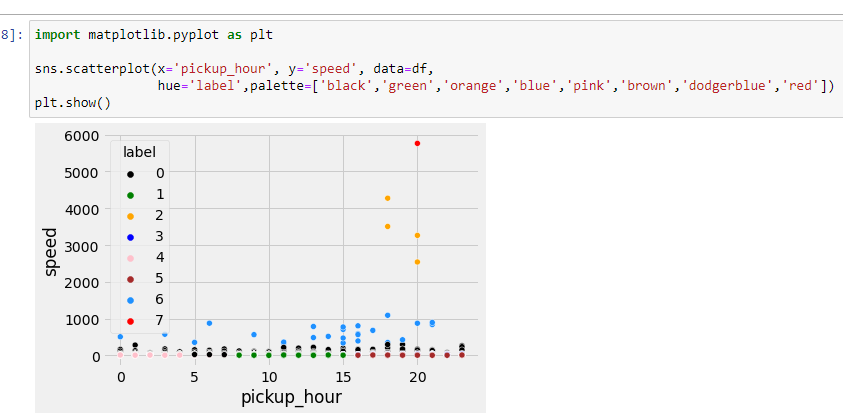


Từ biểu đồ và hàm ta thấy khá rõ giá trị k =8

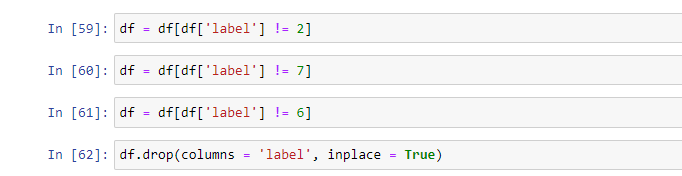
Sau đó sử dụng hàm để dán nhãn các nhóm cho data

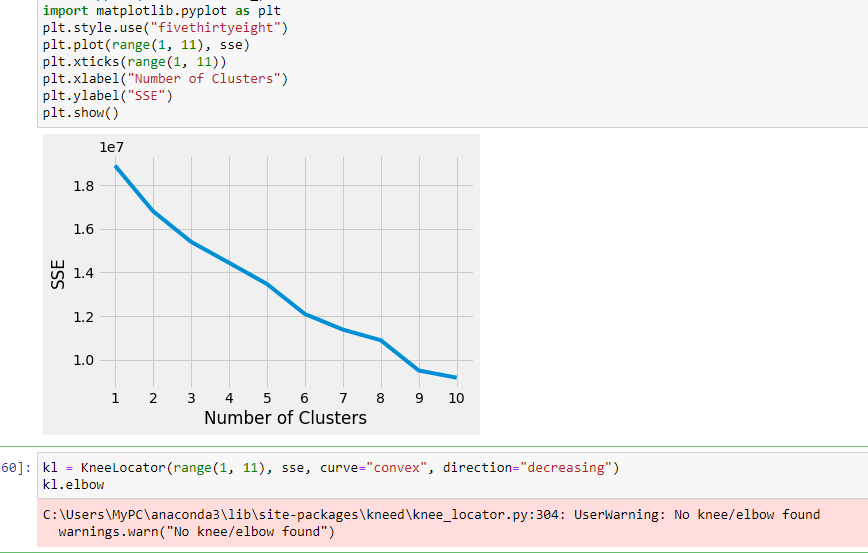


Soi ra các outliner từ biểu đồ

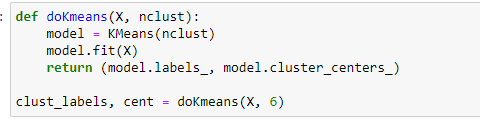


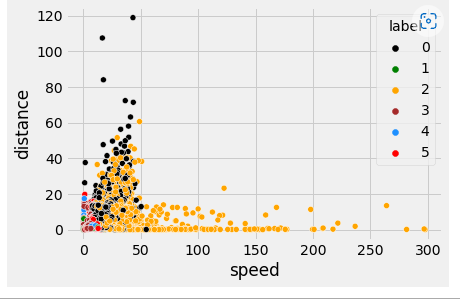
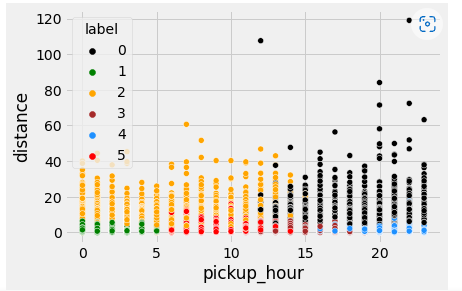
>>> xử lý outliner và chia nhóm lại

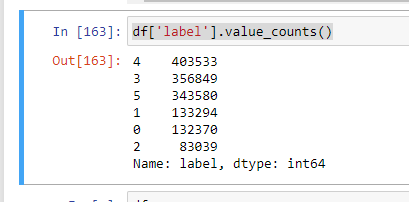


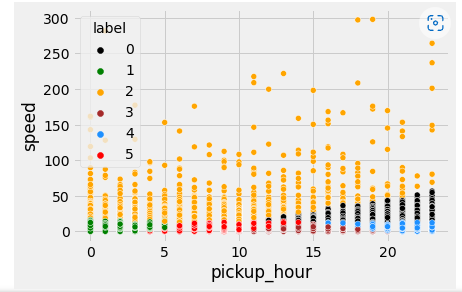
Lúc này ta cần tự tin chọn 1 số k cho mình.

Ở đây mình chọn k = 6









Vàng: Siêu tốc độ, cự li trung bình và xa, xuất hiện đủ loại khung giờ (83k lượt)

Xanh lá: đón khách cự li ngắn ở khung giờ rạng sáng (133k lượt)

Đỏ: Đón khách di chuyển giờ hành chính, cự li ngắn (132k lượt)

Nâu: Đón khách di chuyển giờ tan tầm, cự li ngắn (357k lượt)

Xanh dương: Đón khách di chuyển ban đêm, cự li ngắn, tốc độ chậm (403k)

Đen: Đón khách di chuyển ban đêm, cự li trung bình đến xa, tốc độ vừa phải (343k)

>>> nhu cầu taxi ở NYC tăng cao đột biến vào buổi chiều đến tối – gấp 3 lần so với nửa ngày còn lại

>>> Từ đây, nếu chúng ta có sản phẩm cần quảng cáo, ta có thể lựa chọn kênh radio vào buổi chiều tối cho các chiến dịch của mình > với khả nẳng tiếp cận khách hàng áp đảo.

>>> khoảng cách xa thì bình thường tài xế sẽ đua nhanh, thể hiện rõ ở nhóm Vàng và Đen

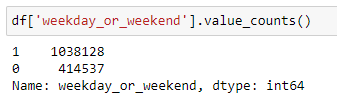
>>> chúng ta có thể tăng giá các cuốc xe ở khoảng cách gần, do thời gian hoàn thành sẽ khá lâu, vận tốc chậm >> tối ưu thu nhập cho tài xế

>>> lưu lại thông tin của các tài xế lái những chuyến xe ở nhóm vàng, cho máy học đường đi của nhóm vàng để tối ưu tốc độ chỉ đường của map, và cũng để kiểm soát – cảnh báo nguy cơ ở các tuyến đường có công an, hoặc tai nạn, do nhóm này có ngưỡng tốc độ max khá cao.



>>> tình hình đường xá ở NYC có vẻ khá đông đúc khi nhóm Vàng chiếm tỷ trọng rất thấp so với phần còn lại >>>







>>> Người dân có xu hướng nghỉ ngơi, ít di chuyển bằng taxi vào cuối tuần, cho thấy taxi đã là phương tiện di chuyển chủ yếu để đi làm việc của khá nhiều người.

Từ đây, nếu chúng ta có sản phẩm cần quảng cáo, ta có thể lựa chọn kênh radio vào buổi chiều tối cho các chiến dịch của mình > với khả nẳng tiếp cận khách hàng áp đảo.



Cuối cùng ta thêm hàm LogisticRegression để sử dụng cho các data về sau này được thuận tiện hơn và cũng dễ theo dõi, không phải mất quá nhiều thời gian xử lý, từ đó lại tiếp tục ra các phương án cụ thể với từng sự biến động của data.



**>.-.THE END.-.<**