XÂY DỰNG MÔ HÌNH DỰ ĐOÁN TIỀN TIP

Môn học: Machine Learning for Data Science

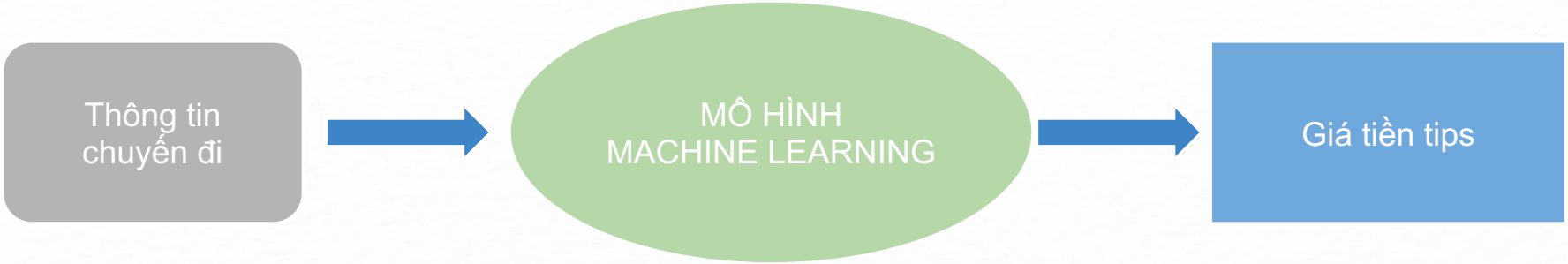
Học viên thực hiện:
Nguyễn Trọng Ân
Giáo viên hướng dẫn:
TS. Nguyễn Thanh Bình
TS. Huỳnh Thế Đăng

Nội dung báo cáo

1. Giới thiệu
2. Bộ dữ liệu
3. Xử lý dữ liệu
4. Lựa chọn đặc trưng và mô hình
5. Kết quả

1. Giới thiệu

BÀI TOÁN ĐẶT RA



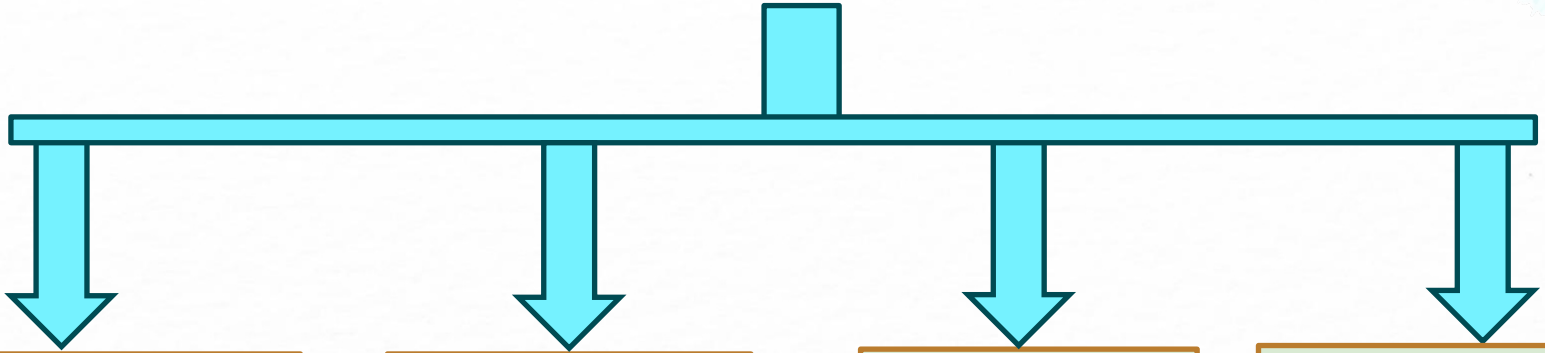
Hình 1: Tổng quan bài toán

1. Giới thiệu

- Nguồn bộ dữ liệu: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew
- Bộ dữ liệu gốc gồm hơn 200 triệu dòng dữ liệu mỗi dòng đại diện cho 1 chuyến đi.
- Bao gồm 23 thuộc tính.
- Bộ dữ liệu để thử nghiệm được trích xuất lấy 200.000 điểm dữ liệu: gồm dữ liệu của 3711 chiếc taxi và 47 hãng taxi

2. Bộ dữ liệu

Chicago Taxi Trips



Thông tin chuyến đi

Trip ID
Taxi ID
Pickup Census Tract
Dropoff Census Tract
Pickup Community Area
Dropoff Community Area
Payment Type

Thời gian

Trip Start Timestamp
Trip End Timestamp
Trip Seconds
Trip Miles

Chi phí

Fare
Tolls
Tips
Extras
Trip Total

Tọa độ

Pickup Centroid Latitude
Pickup Centroid Longitude
Pickup Centroid Location
Dropoff Centroid Latitude
Dropoff Centroid Longitude
Dropoff Centroid Location

3. Xử lý dữ liệu

Bảng 1: Ví dụ xử lý dữ liệu kiểu Datetime.

Trip Start Timestamp	Trip End Timestamp
01/01/2020 01:00:00 AM	01/02/2020 02:30:00 PM
01/01/2020 02:30:00 AM	01/02/2020 12:15:00 PM
01/01/2020 03:30:00 AM	01/01/2020 08:30:00 PM

start time	start daytime	end time	end daytime
01:00:00	morning	14:30:00	night
02:30:00	morning	12:15:00	night
03:30:00	morning	20:30:00	night

3. Xử lý dữ liệu

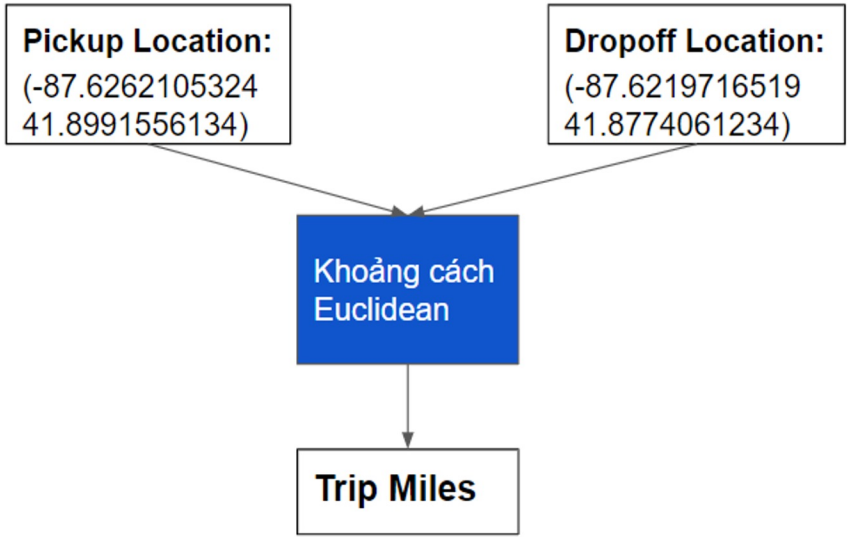
Bảng 2: Ví dụ xử lý thuộc tính 'Trip Seconds':

start time	end time
01:00:00	14:30:00
02:30:00	12:15:00

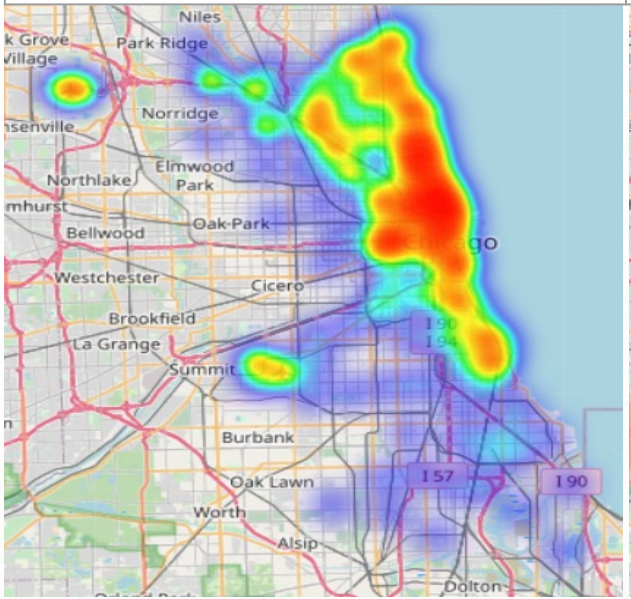
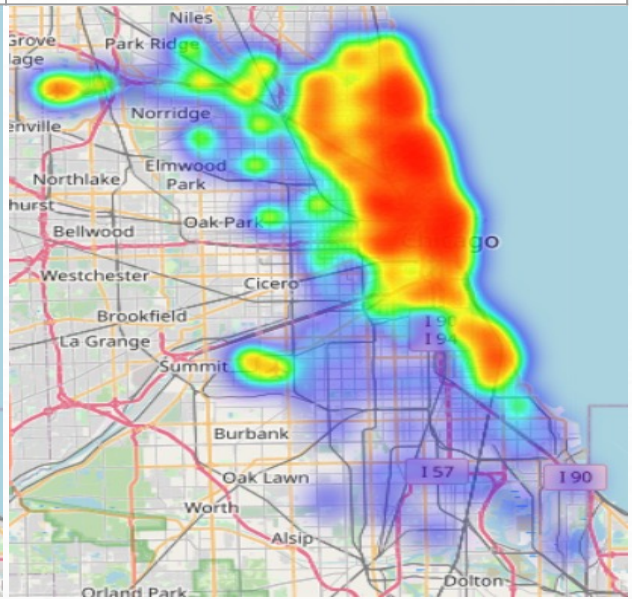
Trip Seconds
48600.0
35100.0

3. Xử lý dữ liệu

Trip Miles có thể sử dụng tọa độ từ các cột Pickup/ Dropoff Location để tính khoảng cách Euclidean rồi chuyển về đơn vị mile.

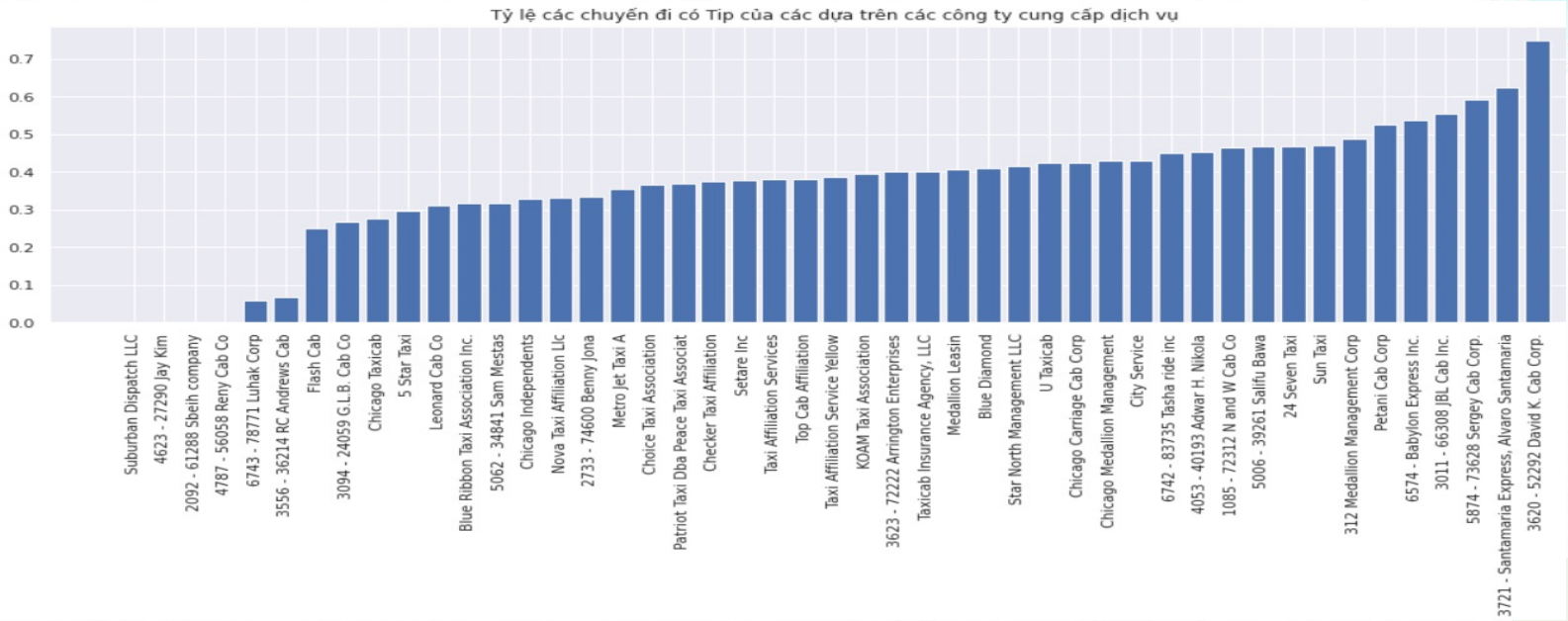


4. Lựa chọn đặc trưng và mô hình

Nơi bắt đầu của những chuyến đi có Tip	Nơi kết thúc của những chuyến đi có Tip
	

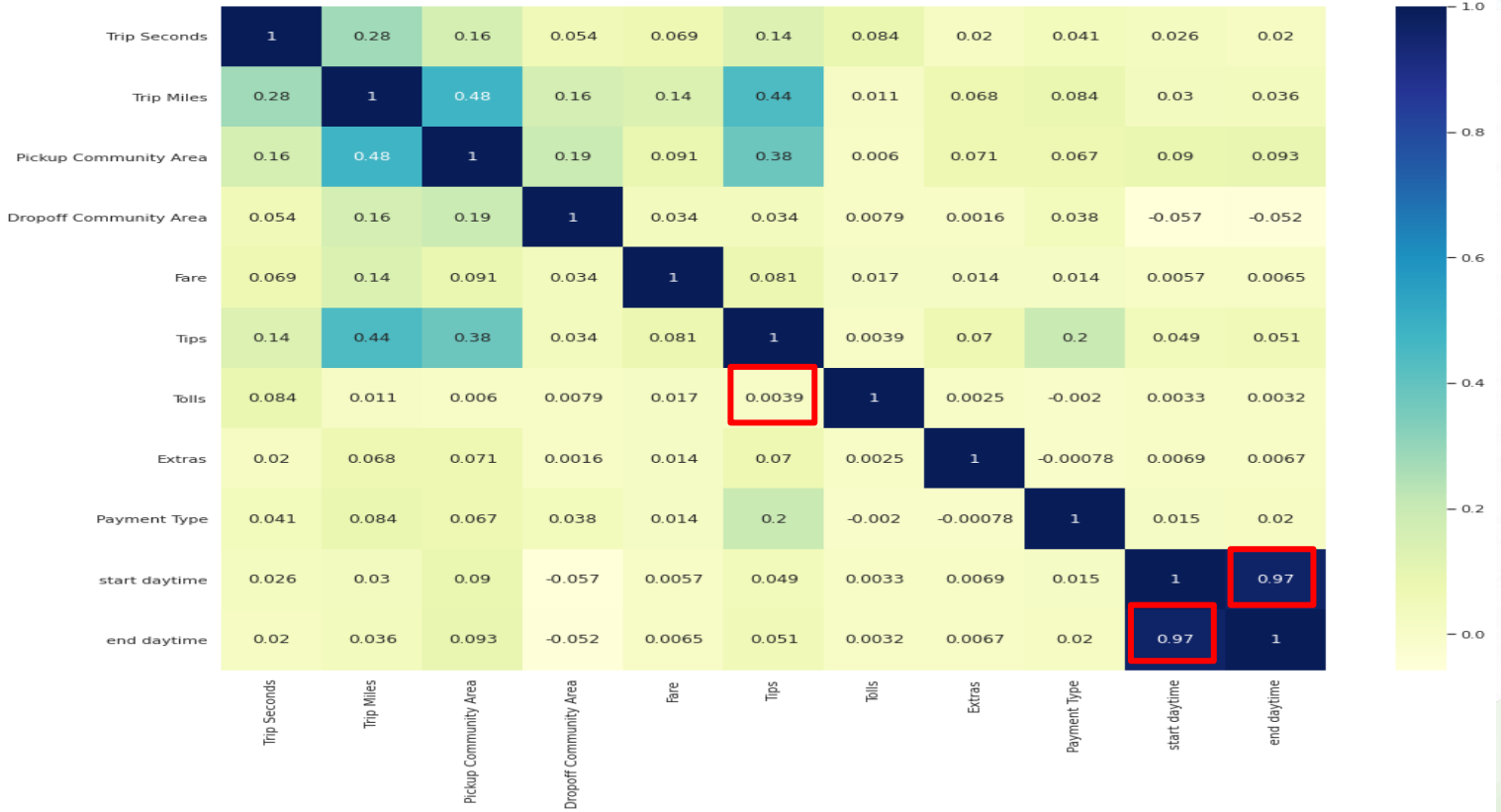
Hình 2: Heatmap trực quan các chuyến đi có Tip

4. Lựa chọn đặc trưng và mô hình



Hình 3: Biểu đồ thể hiện tỉ lệ chuyến đi có tips theo hãng taxi 'Company'.

4. Lựa chọn đặc trưng và mô hình



Hình 4: Bảng tương quan giữa các thuộc tính.

4. Lựa chọn đặc trưng và mô hình

Sau toàn bộ các bước thăm dò và tiền xử lý bộ dữ liệu còn lại :

- 8 biến độc lập: ‘Trip Seconds’, ‘Trip Miles’, ‘Pickup Community Area’, ‘Dropoff Community Area’, ‘Fare’, ‘Extras’, ‘Payment Type’, ‘start daytime’.
- 1 biến mục tiêu: ‘Tips’.

- Các đặc trưng quan trọng sẽ được tổ hợp lại với nhau nhằm tìm ra bộ đặc trưng tốt nhất.
- Các mô hình thực nghiệm sẽ là mô hình Linear Regression, Multiple Linear Regression và Polynomial Linear Regression (bậc 2,3,4,5).
- Tỷ lệ tập train:test là 8:2

4. Lựa chọn đặc trưng và mô hình

- Simple linear regression sẽ dùng một biết độc lập để đưa ra dự đoán.

$$y = b_0 + b_1 x$$

- Multiple linear regression sẽ dùng nhiều biết độc lập để đưa ra dự đoán.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

- Polynominal Regresstion là một trường hợp đặc biệt của mô hình Multiple linear regression với số bậc cao hơn 1.

$$y = b_0 + b_1 x_1 + b_2 (x_1)^2 + b_3 (x_1)^3 \ldots$$

5. Kết quả

Bảng 3: Năm mô hình có kết quả cao nhất của mô hình hồi quy tuyến tính đơn biến.

	Model	Feature_details	RMSE	R^2_train	R^2_test	4_Fold_Validation	5_Fold_Validation	Note
0	Multiple Linear Regression	['TM', 'F', 'PCA', 'DCA', 'PT']	2.124521	0.254302	0.291223	0.252505	0.254774	Test_Size = 0.2, Number_Feature = 5
1	Multiple Linear Regression	['TM', 'PCA', 'DCA', 'PT']	2.124147	0.254015	0.291472	0.252777	0.254750	Test_Size = 0.2, Number_Feature = 4
2	Multiple Linear Regression	['TM', 'F', 'PCA', 'DCA', 'PT', 'SD']	2.124230	0.254459	0.291417	0.252389	0.254745	Test_Size = 0.2, Number_Feature = 6
3	Multiple Linear Regression	['TM', 'PCA', 'DCA', 'PT', 'SD']	2.123862	0.254172	0.291662	0.252664	0.254720	Test_Size = 0.2, Number_Feature = 5
4	Multiple Linear Regression	['TS', 'TM', 'F', 'PCA', 'DCA', 'PT']	2.124852	0.254369	0.291002	0.252178	0.254228	Test_Size = 0.2, Number_Feature = 6

Bảng 4: Năm mô hình có kết quả cao nhất của mô hình hồi quy tuyến tính đa biến.

	Model	Feature_details	RMSE	R^2_train	R^2_test	4_Fold_Validation	5_Fold_Validation	Note
0	PolynomialFeatures	['F', 'PT']	1.182926	0.744581	0.780263	0.732633	0.733684	Test_Size = 0.2, Degree = 4, Number_Feature = 2
1	PolynomialFeatures	['F', 'PCA', 'PT']	1.235830	0.726472	0.760169	0.728137	0.729938	Test_Size = 0.2, Degree = 3, Number_Feature = 3
2	PolynomialFeatures	['F', 'PCA', 'PT']	1.170155	0.746315	0.784982	0.723116	0.728944	Test_Size = 0.2, Degree = 4, Number_Feature = 3
3	PolynomialFeatures	['F', 'DCA', 'PT']	1.252279	0.722129	0.753742	0.724216	0.725472	Test_Size = 0.2, Degree = 3, Number_Feature = 3
4	PolynomialFeatures	['F', 'PT']	1.250390	0.721849	0.754485	0.724082	0.725329	Test_Size = 0.2, Degree = 3, Number_Feature = 2
5	PolynomialFeatures	['TS', 'F', 'PCA', 'PT']	1.233431	0.727044	0.761099	0.721382	0.724598	Test_Size = 0.2, Degree = 3, Number_Feature = 4
6	PolynomialFeatures	['TS', 'F', 'PT']	1.245558	0.722823	0.756378	0.712413	0.714251	Test_Size = 0.2, Degree = 3, Number_Feature = 3
8	PolynomialFeatures	['TS', 'F', 'DCA', 'PT']	1.247897	0.723195	0.755462	0.703290	0.704504	Test_Size = 0.2, Degree = 3, Number_Feature = 4
9	PolynomialFeatures	['F', 'PCA', 'DCA', 'PT']	1.279389	0.727338	0.742964	0.697954	0.700462	Test_Size = 0.2, Degree = 3, Number_Feature = 4
10	PolynomialFeatures	['TS', 'F', 'PCA', 'DCA', 'PT']	1.266162	0.727860	0.748252	0.694238	0.699459	Test_Size = 0.2, Degree = 3, Number_Feature = 5

Cám ơn mọi người
đã lắng nghe!

Column Name	Miêu tả tên cột	Data type
Trip ID	id của chuyến đi	Text
Taxi ID	id của xe taxi	Text
Trip Start Timestamp	thời gian chuyến xe khởi hành, được làm tròn thành 15 phút gần nhất	Date & Time
Trip End Timestamp	thời gian chuyến xe kết thúc, được làm tròn thành 15 phút gần nhất	Date & Time
Trip Seconds	Độ dài của chuyến đi tính bằng giây	Number
Trip Miles	Khoảng cách di chuyển tính bằng dặm	Number
Pickup Census Tract	Census tract nơi bắt đầu chuyến đi. Để bảo mật thông tin một số chuyến đi sẽ không có thuộc tính này. Những địa điểm nằm ngoài Chicago thường không có giá trị cho thuộc tính này	Text

Column Name	Miêu tả tên cột	Data type
Dropoff Census Tract	Census tract nơi kết thúc chuyến đi. Để bảo mật thông tin một số chuyến đi sẽ không có thuộc tính này. Những địa điểm nằm ngoài Chicago thường không có giá trị cho thuộc tính này	Text
Pickup Community Area	Khu vực dân cư nơi chuyến đi bắt đầu. Nếu địa điểm nằm ngoài Chicago thì trường này sẽ trống	Number
Dropoff Community Area	Khu vực dân cư nơi chuyến đi kết thúc. Nếu địa điểm nằm ngoài Chicago thì trường này sẽ trống	Number
Fare	Giá chuyến đi	Number
Tips	Tiền tip khách hàng đưa cho driver. Tip bằng tiền mặt sẽ không được lưu lại	Number
Tolls	Phí đường bộ	Number
Extras	Chi phí phát sinh thêm	Number
Trip Total	Tổng chi phí	Number

Column Name	Miêu tả tên cột	Data type
Payment Type	Phương thức thanh toán	Text
Company	Hãng taxi	Text
Pickup Centroid Latitude	Vĩ độ của trung tâm nơi đón khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm nằm ngoài Chicago thường sẽ không có trường này	Number
Pickup Centroid Longitude	Kinh độ của trung tâm nơi đón khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm nằm ngoài Chicago thường sẽ không có trường này	Number
Pickup Centroid Location	Địa điểm của trung tâm nơi đón khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm nằm ngoài Chicago thường sẽ không có trường này	Point