New York City Taxi Fare Prediction.

Giới thiệu

Trong dự án này, tôi được cung cấp bộ dữ liệu đào tạo lưu trữ các chuyến đi Taxi hàng ngày ở New York, Mỹ từ năm 2009 đến năm 2015. Bộ dữ liệu này bao gồm 55 triệu bản ghi tương đương với 55 triệu chuyến đi được ghi lại.

Mục tiêu của dự án này là dự đoán được giá vé của mỗi chuyến Taxi dựa trên các thông tin về địa điểm đón, trả khách, ngày giờ đón và số lượng khách đi.

Dữ liệu

Dữ liệu được cung cấp bởi Kaggle (https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data) hoàn toàn miễn phí. Trong dự án này, tôi chỉ lấy 1 triệu bản trong số 55 triệu bản ghi để thực hiện phân tích.

Cấu trúc các cột trong dữ liệu:

Feature Name	Feature Description	Feature Data Type
Key	This is the unique identifier. This is combination of pickup datetime and an unique identifier	string
pickup_datetime	Date time when trip started	timestamp
pickup_longitude	longitude coordinate of where trip started	float
pickup_latitude	latitude coordinate of where the trip started	float
dropoff_longitude	longitude coordinate of where trip ended	float
dropoff_latitude	latitude coordinate of where the trip ended	float
passenger_count	number of passengers in taxi ride	integer
fare_amount	cost of the taxi ride in dollars. This is the value to be predicted. It is not present in the test data	float

Một số bản ghi đại diện cho dữ liệu:

	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	2009-06-15 17:26:21.0000001	4.5	2009-06-15 17:26:21+00:00	-73.844311	40.721319	-73.841610	40.712278	1
1	2010-01-05 16:52:16.0000002	16.9	2010-01-05 16:52:16+00:00	-74.016048	40.711303	-73.979268	40.782004	1
2	2011-08-18 00:35:00.00000049	5.7	2011-08-18 00:35:00+00:00	-73.982738	40.761270	-73.991242	40.750562	2
3	2012-04-21 04:30:42.0000001	7.7	2012-04-21 04:30:42+00:00	-73.987130	40.733143	-73.991567	40.758092	1
4	2010-03-09 07:51:00.000000135	5.3	2010-03-09 07:51:00+00:00	-73.968095	40.768008	-73.956655	40.783762	1

Câu hỏi

Quá trình vận hành một doanh nghiệp taxi, chúng tôi lo ngại rằng việc tính sai tiền vé cho mỗi chuyến đi sẽ làm ảnh hưởng đến mức độ hài lòng và hình ảnh doanh nghiệp trong khách hàng. Những sai số đó có thể đến từ sự hự hỏng của máy móc hoặc sự gian lận của tài xế. Để phát hiện ra những gian lận như vậy, chúng tôi cần giải thích được sự phụ thuộc của giá tiền vào các yếu tố khách trong chuyến đi.

Giả thuyết và công nghệ lựa chọn

- Quảng đường di chuyển càng dài thì giá vé sẽ cao hơn.
- Thời gian di chuyển vào giờ cao điểm giá vé sẽ cao hơn.
- Ngày cuối tuần giá vé sẽ cao hơn.
- Giá vé năm sau sẽ cao hơn giá vé năm trước đó.
- Chuyến đi có càng nhiều khách thì giá vé sẽ càng cao.
- Giá vé sẽ khách nhau tùy theo các quận khách nhau trong thành phố.
- Đón trả khách tại sân bay giá vé sẽ cao hơn.

➔ Technique: Sử dụng các mô hình Regression (linear, decision tree)

Làm sạch và phân tích dự liệu

1. Loại bỏ tất cả các bản ghi chứa thông tin *null*

```
key                  0
fare_amount          0
pickup_datetime      0
pickup_longitude     0
pickup_latitude      0
dropoff_longitude    10
dropoff_latitude     10
passenger_count      0
dtype: int64
```

2. Loại bỏ các bản ghi có thông tin ngoại lệ

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	999990.000000	999990.000000	999990.000000	999990.000000	999990.000000	999990.000000
mean	11.347953	-72.526699	39.929040	-72.527860	39.919954	1.684941
std	9.821790	12.057778	7.626087	11.324494	8.201418	1.323907
min	-44.900000	-3377.680935	-3116.285383	-3383.296608	-3114.338567	0.000000
25%	6.000000	-73.992060	40.734965	-73.991385	40.734046	1.000000
50%	8.500000	-73.981792	40.752695	-73.980135	40.753166	1.000000
75%	12.500000	-73.967094	40.767154	-73.963654	40.768129	2.000000
max	500.000000	2522.271325	2621.628430	45.581619	1651.553433	208.000000

Nhìn vào mô tả của dữ liệu ở hình trên, tôi thấy rằng có những bản ghi mà giá trị của cột **fare_amount** là âm. Điều này là không đúng với thực tế, nên tôi đã loại bỏ toàn bộ các giá trị này.

Đồng thời tôi cũng quan sát thấy rằng, có một số bản ghì có số lượng khách quá lớn:

fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
3.3	2009-07-30 11:54:00+00:00	0.0	0.0	0.0	0.0	208

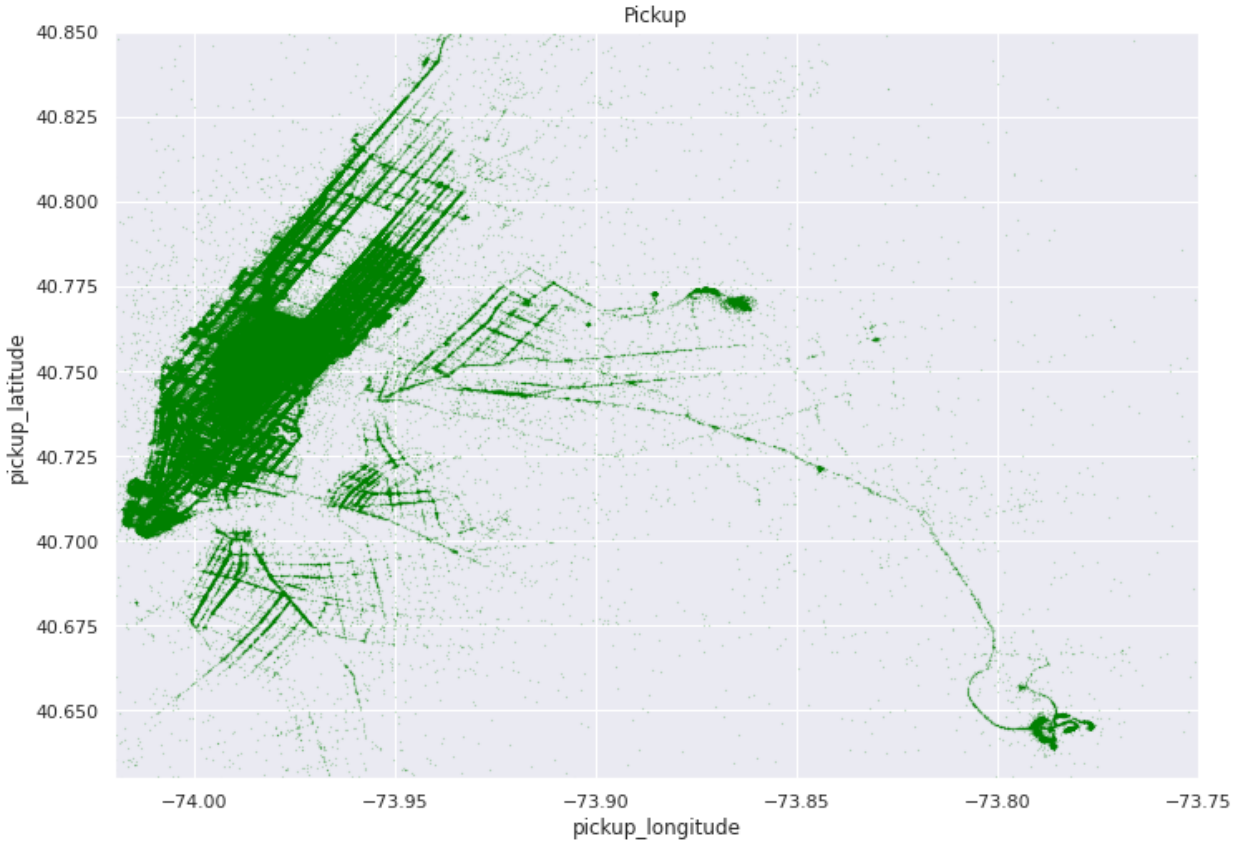
Trong thực tế, không có chuyến xe Taxi nào có thể chở 208 khách cùng một lúc, nên tôi đã loại bỏ những bản ghi mà giá trị của cột **passeger_count** là lớn hơn 8.

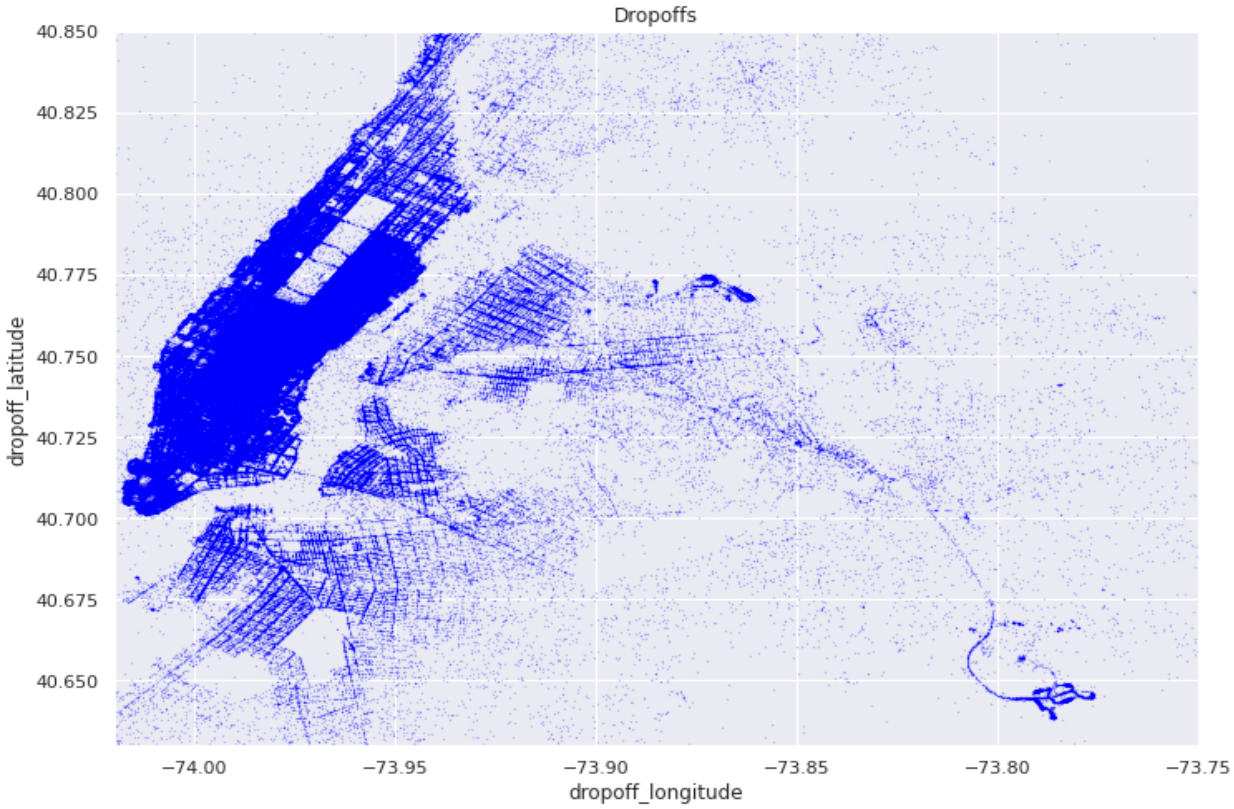
Tiếp theo, tôi thấy rằng có những vị trí mà dữ liệu ghi lại là không thuộc thành phố New York hoặc thâm chí nó là một vị trí không hợp lệ trên bản đồ. Tôi đã loại bỏ những bản ghi này bằng cách giờ hạn lại khung tọa độ hợp lý cho dữ liệu là: vĩ độ từ 40.573143 đến 41.709555 và kinh độ từ -74.263242 đến -72.986532.

Sau đó, tôi thực hiện tính độ dài quảng đường di chuyển của tất cả các chuyến đi bằng cách sử dụng geodesic (khảng cách trắc địa trên địa cầu). Tôi phát hiện ra một số bản ghi có khoảng cách bằng 0, điều này là do các điểm đón và trả khách là cũng một vị trí. Có thể đây không phải là những bản ghi sai, chuyến đi này có thế đến một điểm nào đó và sau đó được yêu cầu về lại chỗ cũ. Nhưng trong phạm vi của mình, tôi không thể xử lý được những bản ghi này nên tôi đã quyết định loại bỏ hoàn toàn chúng ra khỏi dữ liệu.

3. Xác đinh ảnh hưởng của các vị trí địa lý đối với giá vé.

Chúng tôi, biểu diễn tất cả các điểm vị trí đón và trả khách lên biểu đồ tọa độ:


Pickup
40.850
40.825
40.800
40.775
40.750
40.725
40.700
40.675
40.650
pickup_latitude
−74.00
−73.95
−73.90
−73.85
−73.80
−73.75
pickup_longitude


Dropoffs
40.850
40.825
40.800
40.775
40.750
40.725
40.700
40.675
40.650
dropoff_latitude
−74.00
−73.95
−73.90
−73.85
−73.80
−73.75
dropoff_longitude

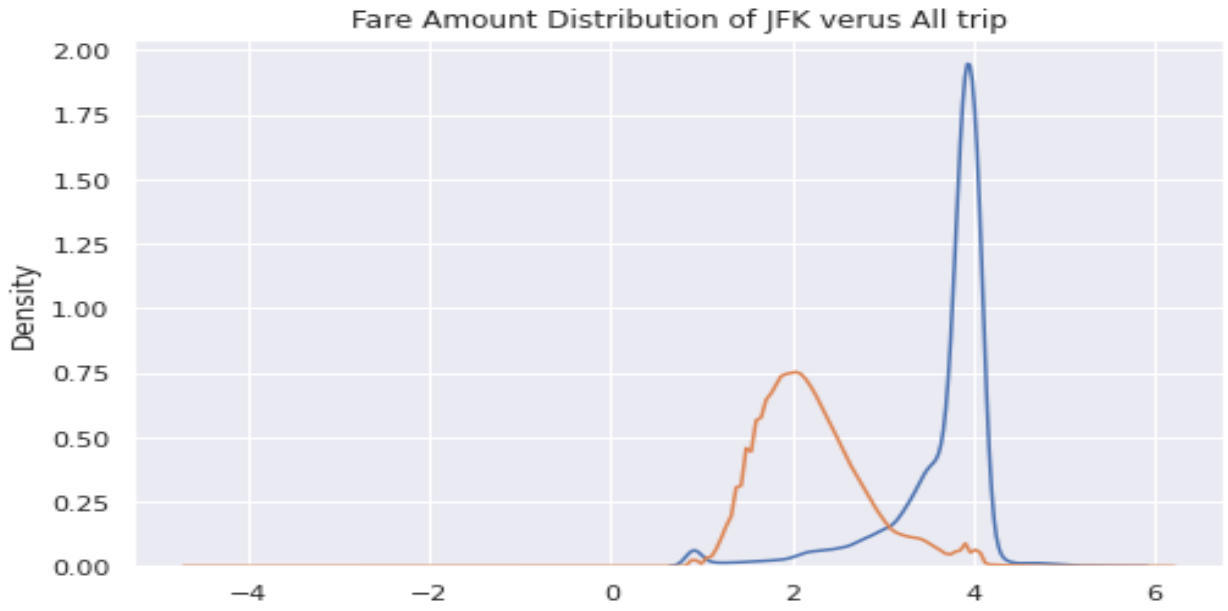
Sau khi quan sát, chúng tôi thấy rằng các ví trí đón và trả khách có mật độ phân bổ tương tự nhau. Chúng đều tập trung chủ yếu về một vùng có màu đậm, tôi xác định được những điểm này trên bản đồ có vị trí trùng với các sân bay:

LaGuardia Airport(40.7730135746,-73.8702298524).

John F. Kennedy InternationalAirport(40.641766,-73.780968).

Câu hỏi đặt ra ở đây là liệu rằng các vị trí đón và trả khách ở sân bay có ảnh hưởng đến giá vé Taxi hay không?

Để trả lời cho câu hỏi này, tôi đã mô tả mật độ giá vé của tất cả các chuyến đi so với mật độ giá vé của những bản ghi có vị trí đón và trả khách ở sân bay:



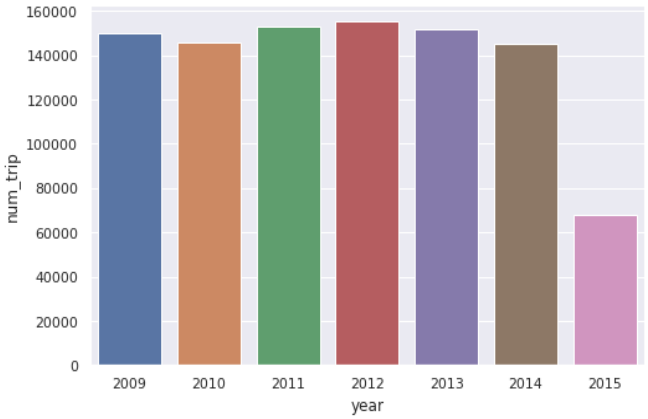
Tôi thấy rằng những chuyến đi đón trả ở sân bay, có mật độ giá vé cao cao hơn rất nhiều.

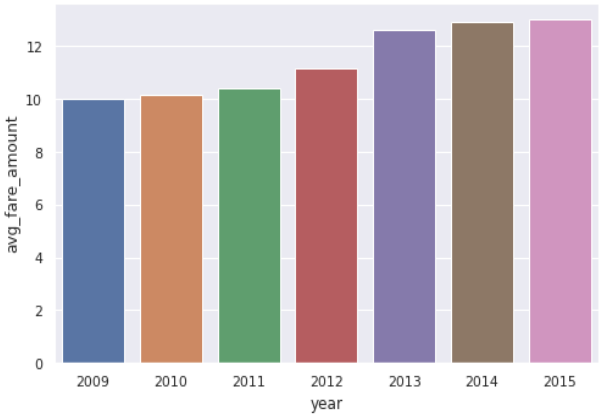
4. Xác định ảnh hưởng của thời gian chuyến đi đến giá vé

Bước đầu tiên, tôi chuyển đổi các giá trị thời gian đón khách thành year, month, day, day_of_week, hour.

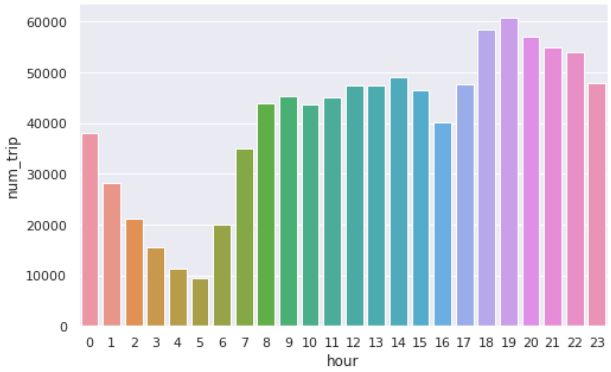
```
df_train['year']=pd.to_datetime(df_train['pickup_datetime']).dt.year
df_train['month']=pd.to_datetime(df_train['pickup_datetime']).dt.mont
h
df_train['day']=pd.to_datetime(df_train['pickup_datetime']).dt.day
df_train['day_of_week']=pd.to_datetime(df_train['pickup_datetime']).d
t.dayofweek
df_train['hour']=pd.to_datetime(df_train['pickup_datetime']).dt.hour
```

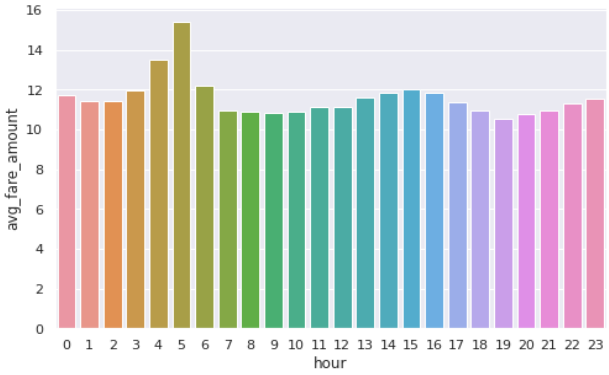
Sau đó, tôi thấy rằng giá vé trung bình tăng lên theo mỗi năm.



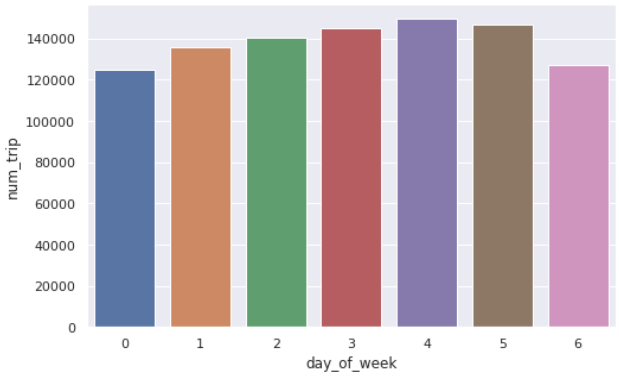


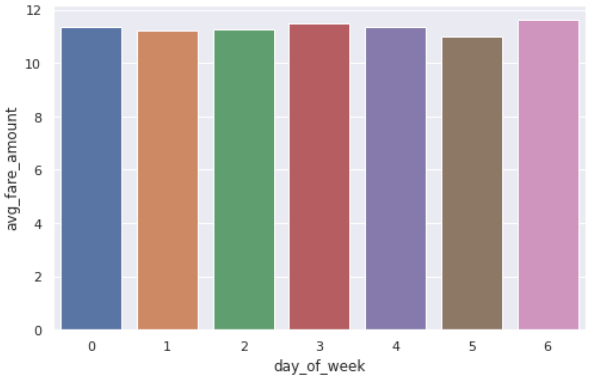
Theo các giờ trong ngày, những chuyến đi tập trung vào khoảng từ 17-21h. Nhưng giá vé trung bình lại cao nhất cho các chuyến đi vào khoảng 5h, mặc dù số lượng chuyến đi là rất thấp vào thời gian này.





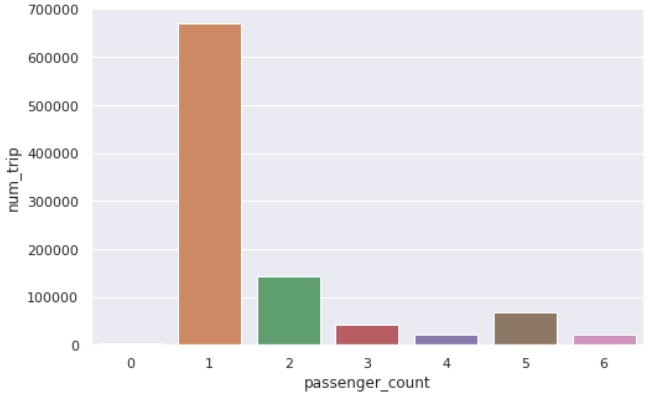
Các ngày trong tuần, giá vé trung bình gần như không quá chênh lệch.

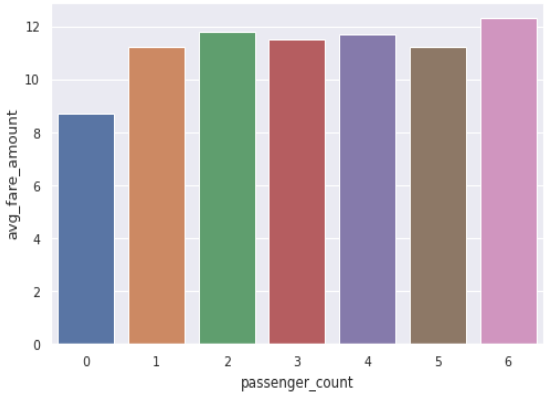




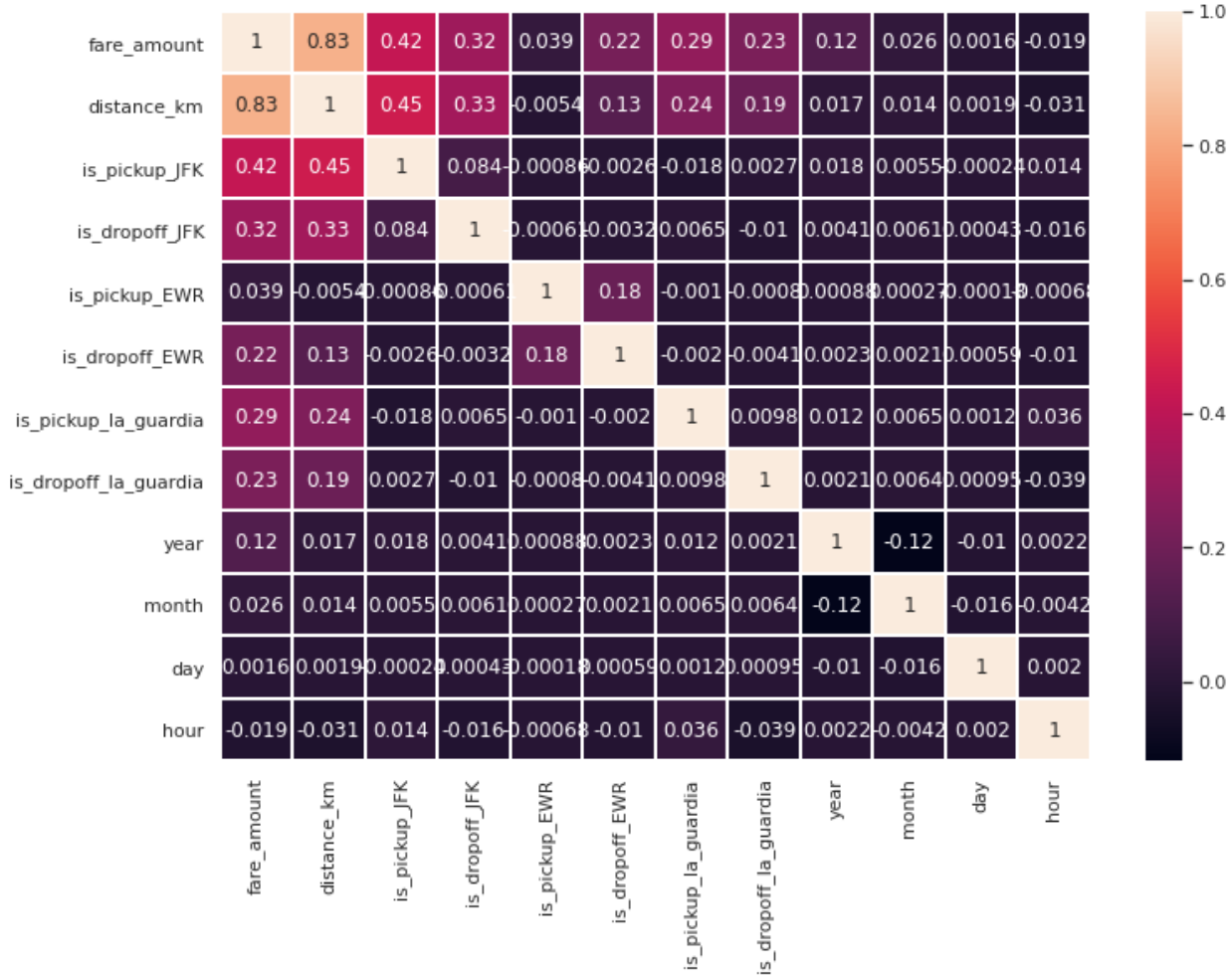
5. Xác định ảnh hưởng của số lượng khách đối với giá vé

Tôi thấy rằng, số lượng khách chủ yếu là 1 người và giá vé trung bình là không chênh lêch nhiều khi số lượng khách thay đổi. Tuy nhiên, có một số bản ghi có số lượng khách là 0, tôi sẽ loại bỏ chúng khỏi tập dữ liệu của mình.





Tìm model tối ưu



Đối với những cột có hệ số tương quan thấp, tôi sẽ loại bỏ chúng ra khỏi việc xây dựng model Regression. Tôi lựa chọn các biến độc lập là distance_km, is_pickup_JFK, is_dropoff_JFK, is_pickup_EWR, is_dropoff_EWR, is_pickup_la_guardia, is_dropoff_la_guardia, year, hour.

1. Using Statistical Model
2. LinearRegression Model
3. Random Forest Model