**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KỲ**

**MÔN HỌC: HỌC MÁY (MACHINE LEARNING)**

**DỰ ĐOÁN GIÁ CƯỚC TAXI Ở NEW YORK CITY**

**Mã lớp học phần: MALE431085\_23\_2\_04CLC**

**Học kỳ 2 – Năm học 2023-2024**

**Giảng viên hướng dẫn: ThS. Trần Quang Khải**

**Danh sách sinh viên thực hiện:**

|  |  |
| --- | --- |
| **MSSV** | **Họ tên** |
| 21110254 | Nguyễn Hoàng Phương Ngân |
| 21110313 | Nguyễn Dương Tiến Thông |

***Thành phố Hồ Chí Minh, tháng 05 năm 2024***

*Nhận xét của giảng viên*

*TP. Hồ Chí Minh, ngày … tháng… năm 2024*

*Giảng viên ký tên*

**MỤC LỤC**

[**PHẦN 1: MỞ ĐẦU 1**](#_Toc14341)

[**1. Lý do chọn đề tài 1**](#_Toc675)

[**2. Mục tiêu đề tài 1**](#_Toc12201)

[**3. Đối tượng và phạm vi nghiên cứu 1**](#_Toc2476)

[**PHẦN 2. NỘI DUNG 3**](#_Toc337)

[**CHƯƠNG 1: TÓM TẮT CƠ SỞ LÝ THUYẾT 3**](#_Toc20101)

[**1.1. Giới thiệu về Machine Learning 3**](#_Toc15093)

[**1.2. Giới thiệu về các thư viện thường được sử dụng trong Học Máy và môi trường lập trình 3**](#_Toc26672)

[**CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY 5**](#_Toc31946)

[**2.1. Các tập dữ liệu cần thu thập 5**](#_Toc20042)

[**CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY 7**](#_Toc16207)

[**3.1. Khởi tạo dữ liệu 7**](#_Toc16737)

[**3.1.1. Import các thư viện cần thiết 7**](#_Toc27326)

[**3.1.2. Xây dựng đường dẫn đến tập tin train.csv và test.csv 7**](#_Toc4892)

[**3.2. Xác thực và làm sạch dữ liệu (Data validation and cleaning) 8**](#_Toc21261)

[**3.3. Mô tả thống kê tập dữ liệu train.csv 13**](#_Toc13512)

[**3.4. Phân tích dữ liệu (Data analysis) 13**](#_Toc1248)

[**3.4.1. Phân tích phân phối giá cước taxi 13**](#_Toc26355)

[**3.4.2. Phân tích trung bình giá cước taxi theo thời gian 15**](#_Toc29464)

[**3.4.3. Phân tích vị trí đón và trả khách 18**](#_Toc31553)

[**3.4.4. Phân tích xu hướng giá cước taxi qua từng năm 19**](#_Toc20971)

[**CHƯƠNG 4: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN, KIỂM CHỨNG VÀ ĐÁNH GIÁ MÔ HÌNH CHO TẬP DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY 21**](#_Toc25002)

[**4.1. Tổng quan dữ liệu (Data Overview) 21**](#_Toc5031)

[**4.2. Trực quan hóa vị trí taxi đón khách sử dụng plotly express 23**](#_Toc17584)

[**4.3. Trực quan hóa các tuyến đường taxi sử dụng plotly graph objects 24**](#_Toc23688)

[**4.4. Tiền xử lý dữ liệu (Preprocessing Data) 25**](#_Toc21592)

[**4.4.1. Tính khoảng cách Haversine (Compute Haversine distance) 25**](#_Toc11853)

[**4.4.2. Tính khoảng cách radius (Compute Radius Distance) 26**](#_Toc7664)

[**4.4.3. Tính khoảng cách đến các địa điểm nổi bật ở New York (Add Airport Locate) 26**](#_Toc20850)

[**4.4.4. Tiền xử lý (Preprocess) 27**](#_Toc32477)

[**4.4. Phân tích đặc trưng (Feature Analytic) 28**](#_Toc28982)

[**4.5. Xây dựng Model (Build Model) 29**](#_Toc29452)

[**4.5.1. Train-Val split 29**](#_Toc18648)

[**4.5.2. Train Model 31**](#_Toc17339)

[**4.6. Mô hình Eval (Model Eval) 31**](#_Toc29614)

[**PHẦN KẾT LUẬN 33**](#_Toc434)

[**PHẦN TÀI LIỆU THAM KHẢO 34**](#_Toc16131)

**DANH MỤC HÌNH ẢNH**

[Hình 1 . Tập dữ liệu train.csv lưu trữ thông tin của các chuyến taxi 5](#_Toc2196)

[Hình 2 . Tập dữ liệu test.csv lưu trữ thông tin của các chuyến taxi 6](#_Toc12492)

[Hình 3 . Code import các thư viện cần thiết cho việc phân tích](#_Toc24905)

[và xây dự mô hình 7](#_Toc24905)

[Hình 4 . Code xây dựng đường dẫn đến tập tin train.csv và test.csv](#_Toc11487)

[trên máy tính 7](#_Toc11487)

[Hình 5 . Kết quả khi chạy code xây dựng đường dẫn 8](#_Toc9857)

[Hình 6 . Code xây dựng đường dẫn đến tập tin train.csv và test.csv trên Kaggle 8](#_Toc19902)

[Hình 7 . Code kiểm tra tập dữ liêu train.csv 9](#_Toc15446)

[Hình 8 . Một số kết quả khi kiểm tra tập dữ liệu train.csv 9](#_Toc1386)

[Hình 9 . Một số kết quả khi kiểm tra tập dữ liệu train.csv 10](#_Toc31490)

[Hình 10 . Kiểm nghiệm những dòng có giá trị pickup\_longtitude >0 11](#_Toc29961)

[Hình 11 . Code tạo các hàm để clean dữ liệu 12](#_Toc19697)

[Hình 12 . Code của việc mô tả thống kê tập dữ liệu train.csv 13](#_Toc7062)

[Hình 13 . Kết quả của việc mô tả thống kê tập dữ liệu train.csv 13](#_Toc29467)

[Hình 14 . Code tạo biểu đồ phân phối giá cước taxi 13](#_Toc23767)

[Hình 15 . Biểu đồ phân phối giá cước taxi 14](#_Toc25488)

[Hình 16 . Code tạo biểu đồ phân tích trung bình giá cước taxi](#_Toc29839)

[theo ngày trong tuần 15](#_Toc29839)

[Hình 17 . Biểu đồ trung bình giá cước taxi theo ngày trong tuần 15](#_Toc30292)

[Hình 18 . Code tạo biểu đồ phân tích trung bình giá cước taxi](#_Toc10742)

[theo giờ trong ngày 16](#_Toc10742)

[Hình 19 . Biểu đồ trung bình giá cước taxi theo giờ trong ngày 17](#_Toc21614)

[Hình 20 . Code vẽ biểu đồ phân tán phân tích vị trí đón khách 18](#_Toc22374)

[Hình 21 . Biểu đồ phân tích vị trí taxi đón khách và giá cước 18](#_Toc23868)

[Hình 22 . Code tạo biểu đồ đường phân tích xu hướng giá cước taxi](#_Toc6743)

[qua từng năm 19](#_Toc6743)

[Hình 23 . Biểu đồ trung bình giá cước taxi qua từng năm 19](#_Toc30217)

[Hình 24 . Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra](#_Toc16269)

[và clean data 21](#_Toc16269)

[Hình 25 . Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra](#_Toc30689)

[và clean data 22](#_Toc30689)

[Hình 26 . Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra](#_Toc16272)

[và clean data 23](#_Toc16272)

[Hình 27 . Code tạo bản đồ trực quan hóa các vị trí taxi đón khách](#_Toc12196)

[ở New York City 23](#_Toc12196)

[Hình 28 . Bản đồ các vị trí mà taxi đón khách ở New York City 24](#_Toc24413)

[Hình 29 . Code kết hợp các cột liên quan từ hai DataFrame train và test 24](#_Toc9306)

[Hình 30 . Code tạo bản đồ trực quan hóa các tuyến đường taxi](#_Toc18779)

[ở New York City 24](#_Toc18779)

[Hình 31 . Bản đồ các tuyến đường taxi ở New York City 25](#_Toc6086)

[Hình 32 . Code tạo hàm tính khoảng cách Haversine 25](#_Toc12944)

[Hình 33 . Code tạo hàm để tính toán hướng (bearing) 26](#_Toc30014)

[Hình 34 . Code tạo hàm tính khoảng cách đến các địa điểm nổi bật](#_Toc24597)

[ở New York City 27](#_Toc24597)

[Hình 35 . Code tạo các hàm tiền xử lý dữ liệu 28](#_Toc8890)

[Hình 36 . Code thực hiện phân tích và tiền xử lý dữ liệu 28](#_Toc8391)

[Hình 37 . Code trực quan hóa phân phối đặc trưng trong dữ liệu huấn luyện 29](#_Toc26666)

[Hình 38 . Kết quả trực quan hóa phân phối đặc trưng trong dữ liệu huấn luyện 30](#_Toc3904)

[Hình 39 . Code chuẩn bị dữ liệu huấn luyện và validation](#_Toc24845)

[cho mô hình LightGBM 30](#_Toc24845)

[Hình 40 . Code huấn luyện mô hình 31](#_Toc12294)

[Hình 41 .Đánh giá hiệu suất mô hình 32](#_Toc11327)

**DANH MỤC BẢNG**

[Bảng 1 . Bảng mô tả tập dữ liệu train.csv 5](#_Toc150)

[Bảng 2 . Bảng mô tả tập dữ liệu train.csv 6](#_Toc27695)

**BẢNG DANH MỤC CÁC TỪ VIẾT TẮT**

|  |  |
| --- | --- |
| **Ký hiệu chữ viết tắt** | **Cách viết đầy đủ** |
| NYC | New York City |
| JFK | John F. Kennedy International Airport |
| EWR | Newark Liberty International Airport |
| LGA | LaGuardia Airport |

**KẾ HOẠCH PHÂN CÔNG NHIỆM VỤ THỰC HIỆN ĐỀ TÀI**

**CUỐI KỲ MÔN LẬP TRÌNH PYTHON**

**HỌC KỲ I NĂM HỌC 2023-2024**

**1. Mã lớp môn học: MALE431085\_23\_2\_04CLC**

**2. Giảng viên hướng dẫn: ThS. Trần Quang Khải**

**3. Tên đề tài: Dự đoán giá cước taxi ở New York City**

**4. Bảng phân công nhiệm vụ:**

|  |  |
| --- | --- |
| **Sinh viên thực hiện** | **Nội dung thực hiện** |
| Nguyễn Hoàng Phương Ngân | * Code xử lý dữ liệu và phân tích dữ liệu. * Xây dựng báo cáo Phần Mở đầu và Phần Nội Dung (Chương 1, 2, 3). * Tổng hợp và hoàn thiện báo cáo. |
| Nguyễn Dương Tiến Thông | * Code xử lý dữ liệu và xây dựng, huấn luyện mô hình. * Xây dựng báo cáo Phần Kết Luận và Phần Nội Dung (Chương 4). |

**LỜI CẢM ƠN**

Để thực hiện được đề tài này, em xin gửi lời cảm ơn chân thành đến thầy Trần Quang Khải - giảng viên dạy học và hướng dẫn bộ môn Học Máy của lớp MALE431085\_23\_2\_04CLC vì đã chỉ dẫn và hỗ trợ chúng em hết mình trong việc học tập và trong quá trình thực hiện tiểu luận cuối kỳ.

Trong quá trình học tập, nghiên cứu và thực hiện đề tài chúng em đã nổ lực và cố gắng rất nhiều, nhưng do kinh nghiệm và hiểu biết còn hạn hẹp nên có thể phần báo cáo có thể còn nhiều thiếu sót. Chúng em rất mong nhận được sự động viên, hỗ trợ và góp ý từ thầy để nhóm em có thể hoàn thiện hơn không chỉ trong môn học này mà còn làm tốt hơn nữa trong những chặng đường sắp tới.

Một lần nữa, chúng em xin trân trọng cảm ơn sự quan tâm giúp đỡ của thầy trong suốt hành trình vừa qua.

# PHẦN 1: MỞ ĐẦU

## Lý do chọn đề tài

New York City là một trong những thành phố có hệ thống taxi phổ biến và nhộn nhịp nhất thế giới. Việc dự đoán chính xác giá cước taxi không chỉ hữu ích cho hành khách mà còn có lợi cho các tài xế và các công ty vận tải. Hành khách có thể ước lượng trước chi phí cho chuyến đi của mình, giúp họ lên kế hoạch tốt hơn về tài chính. Tài xế và công ty vận tải có thể sử dụng dự đoán này để tối ưu hóa lộ trình và tăng hiệu quả hoạt động.

Ngành công nghiệp taxi đóng vai trò quan trọng trong nền kinh tế đô thị, đặc biệt là các thành phố lớn như New York City. Với hàng ngàn chuyến đi mỗi ngày, thậm chí là hàng triệu chuyến, việc tối ưu hóa giá cước taxi có thể đem đến những lợi ích kinh tế đáng kể. Một mô hình dự đoán giá cước hiệu quả có thể giúp giảm chi phí cho hành khách, tăng thu nhập cho tài xế và cải thiện chất lượng dịch vụ.

Vậy câu hỏi đặt ra là làm thế nào để có thể tạo ra những mô hình dự đoán đó? Để giải đáp điều này, nhóm chúng em đã chọn thực hiện đề tài này dựa trên tất cả những kiến thức được lĩnh hội từ bộ môn Học Máy và các kiến thức về Phân tích dữ liệu để biểu diễn tập dữ liệu một cách trực quan và khái quát nhất.

## Mục tiêu đề tài

Mục tiêu cho đề tài “Dự đoán giá cước taxi ở New York City” là xây dựng một mô hình học máy để dự đoán chính xác lượng giá cước taxi dựa trên các yếu tố như thời gian, địa điểm và số lượng hành khách. Nhóm em sẽ xử lý và phân tích dữ liệu lớn từ hệ thống taxi của New York City để tạo ra một công cụ hữu ích cho các công ty taxi và tài xế, giúp tối ưu hóa lộ trình và chiến lược định giá.

## Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài này là:

* Tập dữ liệu taxi của New York City: Bao gồm thông tin về thời gian, địa điểm, số lượng hành khách, và giá cước của các chuyến đi taxi trong một khoảng thời gian nhất định.
* Hành khách và Tài xế Taxi: Những người sử dụng dịch vụ taxi ở New York City, cũng như những người cung cấp dịch vụ này.

Phạm vi nghiên cứu, chúng em sẽ tập trung vào việc:

* Phân tích dữ liệu: Tiến hành khai thác, xử lý dữ liệu lớn từ hệ thống taxi của New York City.
* Xây dựng mô hình dự đoán: Áp dụng các kỹ thuật học máy để xây dựng một mô hình dự đoán giá cước taxi chính xác và hiệu quả và đánh giá mô hình.

# PHẦN 2. NỘI DUNG

## CHƯƠNG 1: TÓM TẮT CƠ SỞ LÝ THUYẾT

### 1.1. Giới thiệu về Machine Learning

Machine Learning là khoa học (và nghệ thuật) lập trình máy tính để chúng có thể học từ dữ liệu.

Dưới đây là những định nghĩa khái quát về Machine Learning:

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed. (*Học Máy là lĩnh vực nghiên cứu cho phép máy tính học hỏi mà không cần được lập trình một cách rõ ràng.*)

—Arthur Samuel, 1959

A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. (*Một chương trình máy tính được cho là học từ kinh nghiệm E đối với một số nhiệm vụ T và một số thước đo hiệu suất P, nếu hiệu suất của nó trên T, được đo bằng P, cải thiện theo kinh nghiệm E.*)

—Tom Mitchell, 1997

### 1.2. Giới thiệu về các thư viện thường được sử dụng trong Học Máy và môi trường lập trình

Các thư viện phổ biến được sử dụng trong Machine Learning bao gồm:

1. NumPy (Numerical Python)

NumPy cung cấp một đối tượng đa chiều (ndarray) hiệu quả, giúp thực hiện các phép toán số học nhanh chóng. Dùng trong các phép toán ma trận và các thao tác tính toán khoa học

2. Pandas

Pandas cung cấp các cấu trúc dữ liệu linh hoạt DataFarme giúp làm việc với dữ liệu dạng bảng dễ dàng. Đọc được nhiều định dạng file như CSV, SQL, …, làm sạch dữ liệu, thực hiện các thao tác cơ bản trên dữ liệu.

3. Matplotlib

Matplotlib tạo các biểu đồ và đồ thị để hiển thị dữ liệu. Hỗ trợ vẽ nhiều loại biểu đồ như đường, cột, hình trò và nhiều loại biểu đồ khác.

5. Scikit-learn

Scikit-learn là một thư viện học máy phổ biến và mạnh mẽ, hỗ trợ nhiều loại thuật toán cũng như các công cụ để trực quan hóa Machine Learning, tiền xử lý, điều chỉnh, lựa chọn và đánh giá mô hình.

6. TensorFlow

TensorFlow: Là một thư viện mã nguồn mở của Google được sử dụng rộng rãi cho việc xây dựng và huấn luyện mô hình Machine Learning và Deep Learning.

7. Keras

Keras: Là một thư viện mã nguồn mở cao cấp được sử dụng để xây dựng các mô hình Deep Learning một cách dễ dàng và linh hoạt.

8. PyTorch

PyTorch: Là một thư viện Machine Learning mã nguồn mở được phát triển bởi Facebook, cung cấp một cách tiếp cận linh hoạt và hiệu quả cho việc xây dựng mô hình Deep Learning.

Môi trường lập trình phổ biến cho Machine Learning bao gồm Jupyter Notebook, Google Colab, và các IDE như PyCharm, Visual Studio Code. Đây là môi trường linh hoạt và hỗ trợ tốt cho việc phát triển và thử nghiệm các mô hình Machine Learning.

## CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY

### 2.1. Các tập dữ liệu cần thu thập

Ở đề tài này, nhóm em đã thu thập những tập dữ liệu có chứa những thông tin chi tiết về các chuyến taxi ở New York City:

* File train.csv:
  + Được sử dụng để huấn luyện mô hình.
  + Bao gồm các đặc trưng đầu vào (input features) và giá cước thực tế (target fare\_amount) các chuyến đi taxi.
  + Thường có số lượng dòng (samples) lớn hơn để mô hình có đủ dữ liệu để học.

|  |  |
| --- | --- |
| **Thông tin** | **Mô tả** |
| Tên tập dữ liệu | train.csv |
| Kích thước | 8 cột và khoảng 55 triệu dòng |
| Loại dữ liệu | Số nguyên, số thực, thời gian, chuỗi ký tự |
| *Hình 1**. Tập dữ liệu train.csv lưu trữ thông tin của các chuyến taxi*  *ở New York City* | |

*Bảng 1**. Bảng mô tả tập dữ liệu train.csv*

* File test.csv:
  + Được sử dụng để kiểm định và đánh giá hiệu suất của mô hình.
  + Bao gồm các đặc trưng đầu vào (input features) cho các chuyến đi taxi, nhưng không có giá cước (fare\_amount) đi kèm.
  + Thường có số lượng dòng (samples) nhỏ hơn so với tập dữ liệu huấn luyện để tiết kiệm thời gian tính toán và đánh giá.

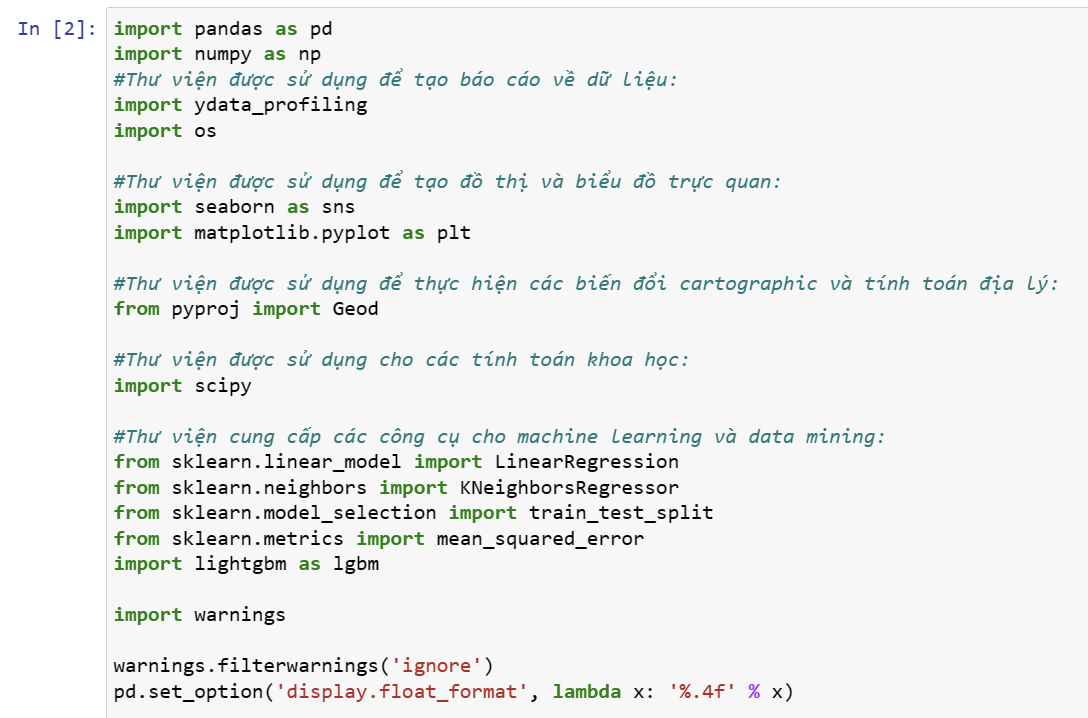
|  |  |
| --- | --- |
| **Thông tin** | **Mô tả** |
| Tên tập dữ liệu | test.csv |
| Kích thước | 7 cột và 9915 dòng |
| Loại dữ liệu | Số nguyên, số thực, thời gian, chuỗi ký tự |
| *Hình 2**. Tập dữ liệu test.csv lưu trữ thông tin của các chuyến taxi*  *ở New York City* | |

*Bảng 2**. Bảng mô tả tập dữ liệu train.csv*

## CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY

### 3.1. Khởi tạo dữ liệu

#### 3.1.1. Import các thư viện cần thiết



*Hình 3**. Code import các thư viện cần thiết cho việc phân tích và xây dự mô hình*

*dự đoán giá cước taxi ở New York City.*

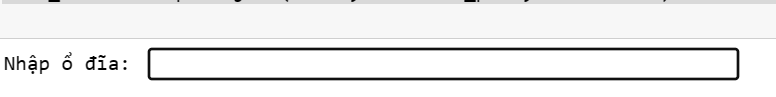
#### 3.1.2. Xây dựng đường dẫn đến tập tin train.csv và test.csv

Tùy vào thư mục chứa các tập tin này ở đường dẫn nào thì điều hướng sang đường dẫn đó.

* Cách 1: Lưu tập tin trên máy tính cá nhân:



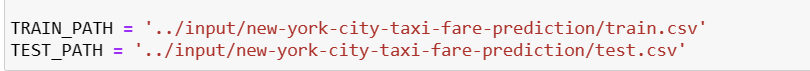
*Hình 4**. Code xây dựng đường dẫn đến tập tin train.csv và test.csv trên máy tính*

Kết quả khi chạy code:  




*Hình 5**. Kết quả khi chạy code xây dựng đường dẫn*

* Cách 2: Lưu tập tin trên Kaggle

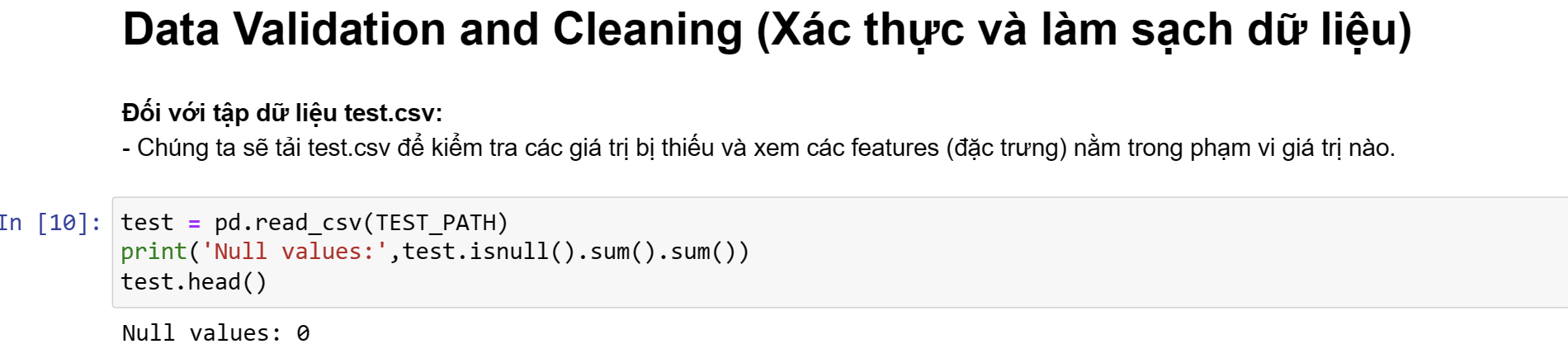


*Hình 6**. Code xây dựng đường dẫn đến tập tin train.csv và test.csv trên Kaggle*

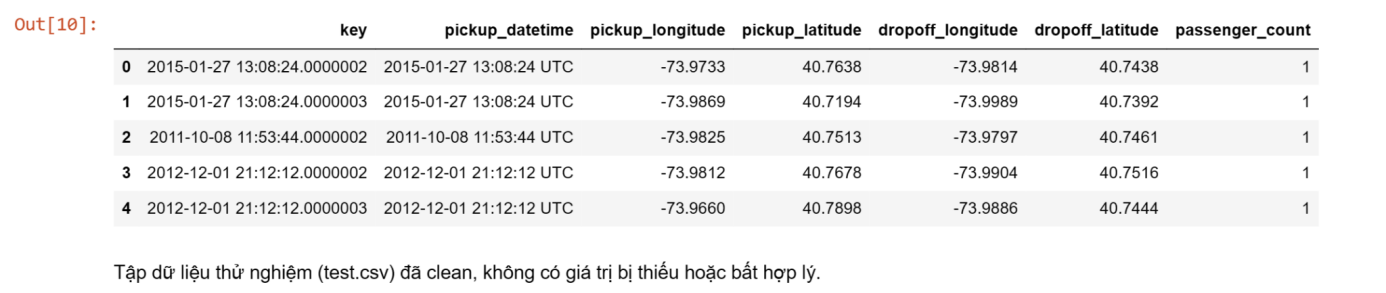
### 3.2. Xác thực và làm sạch dữ liệu (Data validation and cleaning)

Ở bước này, chúng ta sẽ kiểm tra xem tập dữ liệu có giá trị null hay có những giá trị nào bị lỗi không và xử lý chúng.

* **Kiểm tra tập dữ liệu test.csv:**



*Hình: Code kiểm tra tập dữ liệu test.csv*



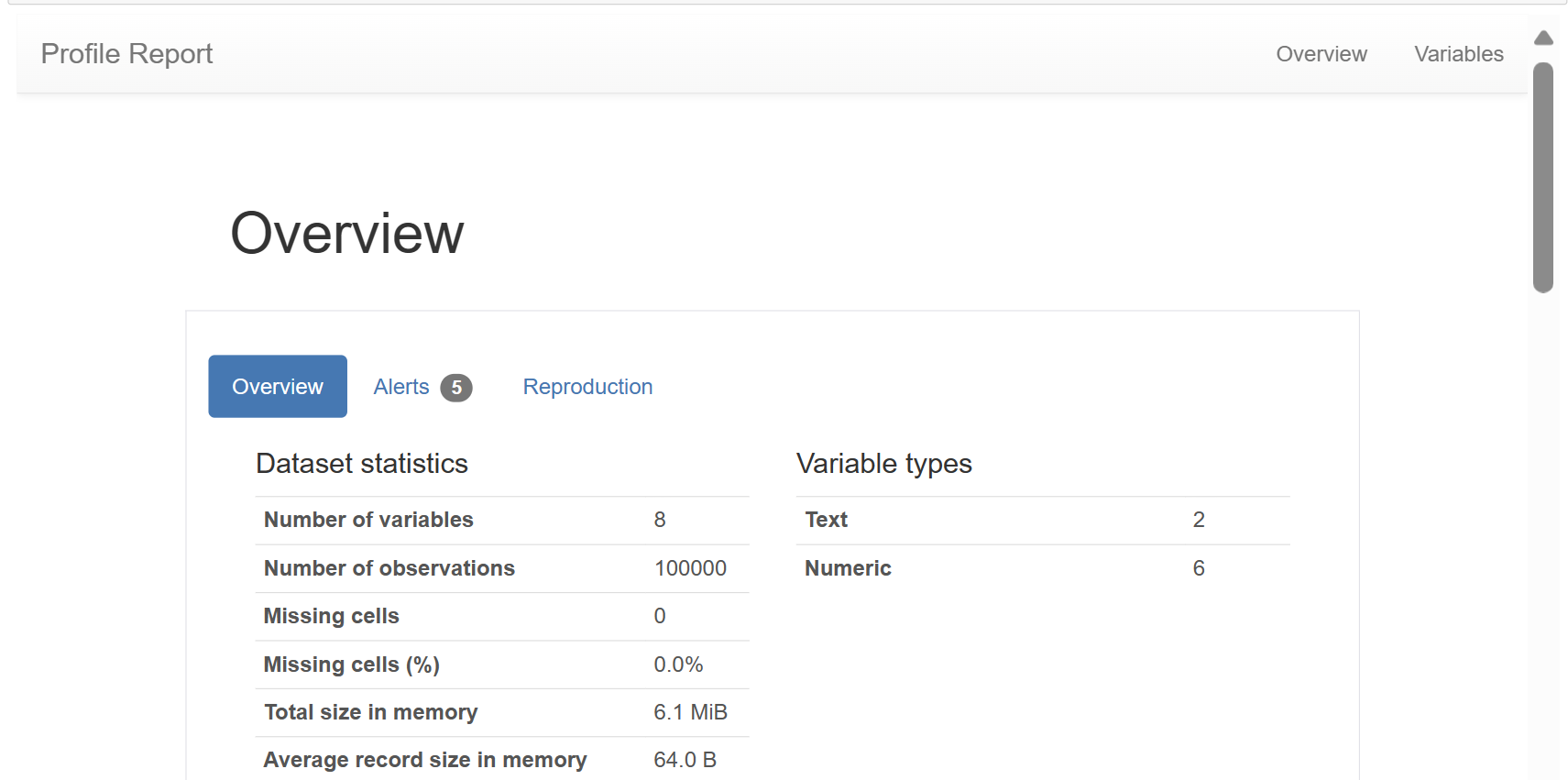
*Hình: Kết quả kiểm tra tập dữ liệu test.csv*

* **Kiểm tra tập dữ liệu train.csv:**

Do tập dữ liệu này có số lượng hàng khá lớn nên khi thực hiện các thao tác sẽ vô cùng chậm, vì vậy chúng ta sẽ tạo ra 1 Profile Report để kiểm tra tập dữ liệu train.csv nhanh hơn và một cách khái quát hơn.

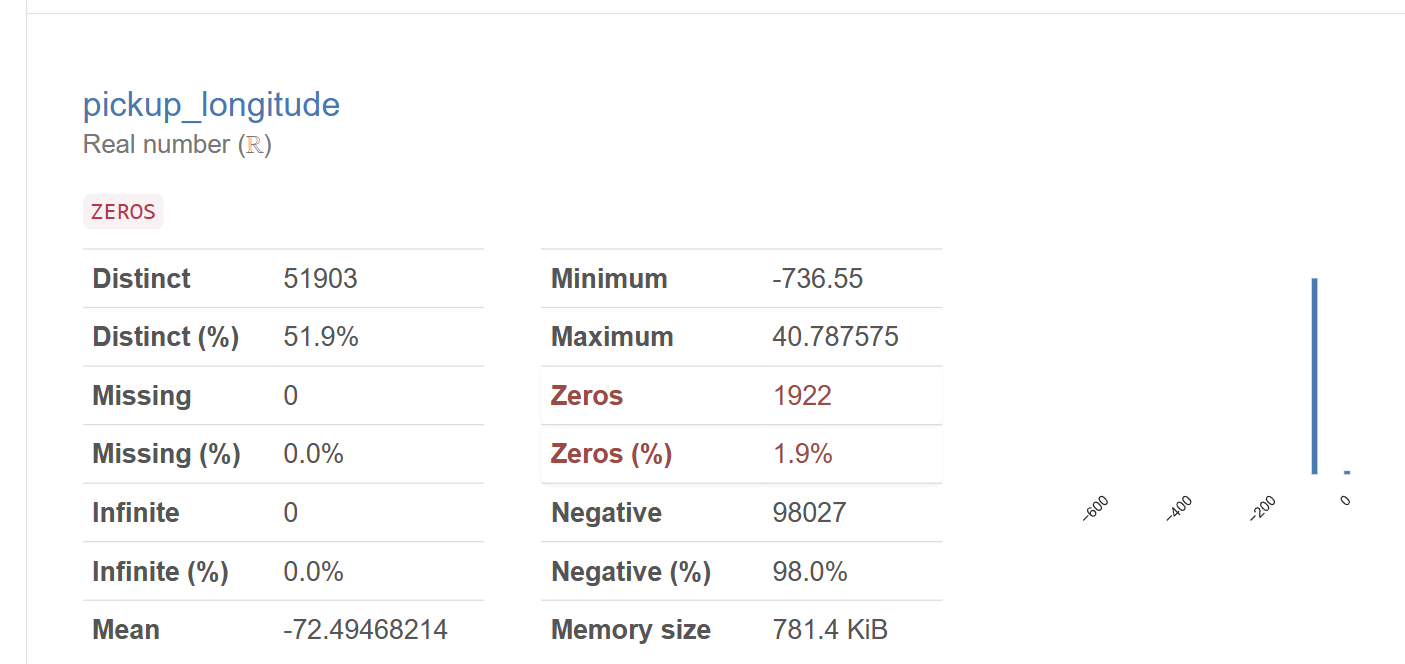
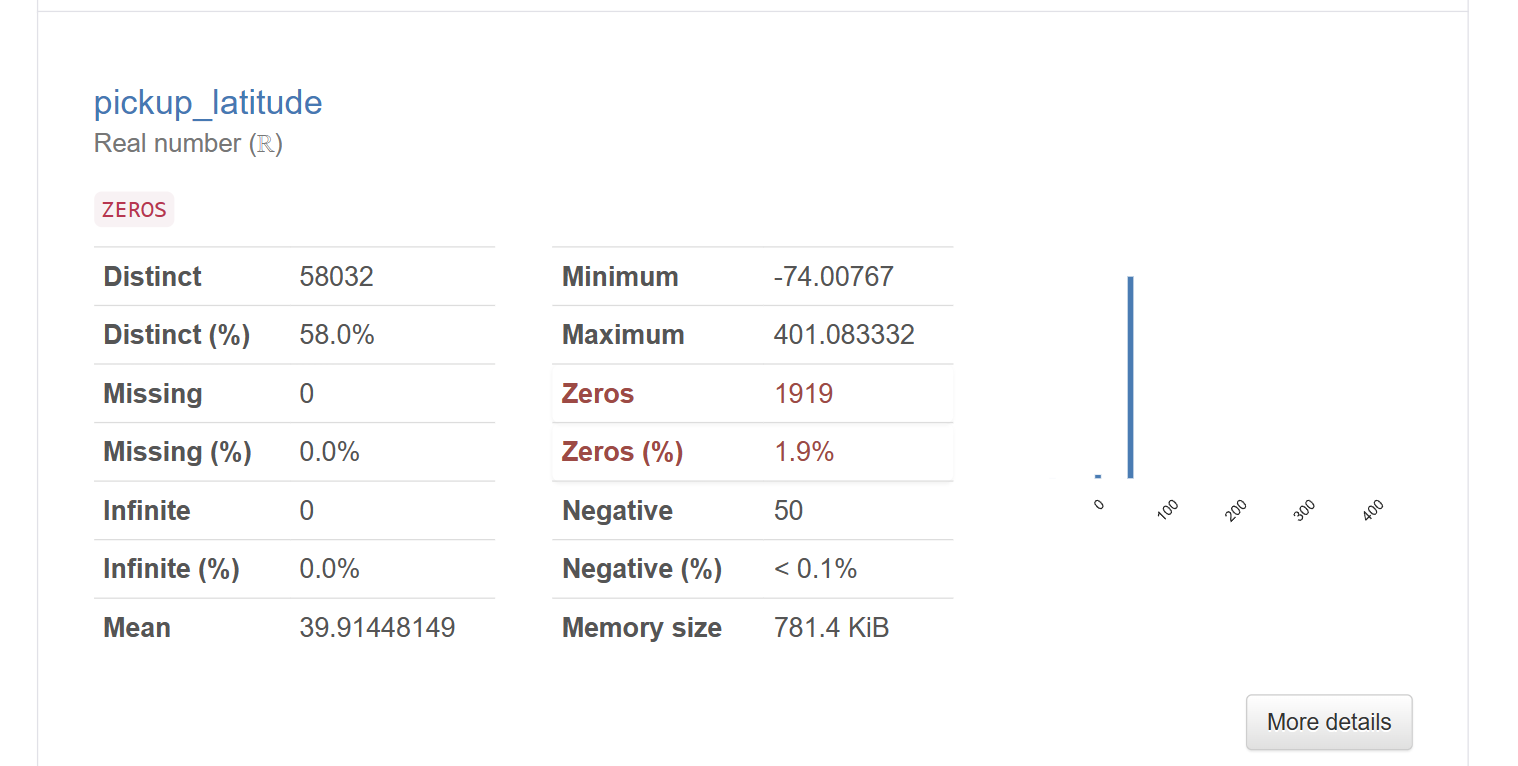


*Hình 7**. Code kiểm tra tập dữ liêu train.csv*





*Hình 8**. Một số kết quả khi kiểm tra tập dữ liệu train.csv*



*Hình 9**. Một số kết quả khi kiểm tra tập dữ liệu train.csv*

Dựa vào Profile Report, ta sẽ rút ra được những thông tin sau:

* Cột key được sử dụng làm chỉ mục và có thể bị xóa khi làm việc với dữ liệu.
* Cột fare\_amount chứa một số giá trị âm, chúng ta sẽ xóa các hàng này và đặt giới hạn trên hợp lý (plausible upper limit).
* Cột passenger\_count chứa một số chuyến đi không có hành khách, chúng ta sẽ xóa các hàng này và đặt giới hạn trên hợp lý.
* Ở các cột như pickup\_longitude, pickup\_latitude, dropoff\_longitude và dropoff\_latitude có một số giá trị tọa độ kinh độ (longitude) và vĩ độ (latitude) của các vị trí ở New York City có vẻ không hợp lý. Cụ thể, vĩ độ của New York City thường nằm trong khoảng 40 và kinh độ thường nằm trong khoảng -73. Tuy nhiên, trong tập dữ liệu, có một số hàng có giá trị vĩ độ tối thiểu là -74 và kinh độ tối đa là 40, điều này cho thấy có thể các giá trị này đã bị hoán đổi cho nhau trong một số hàng.
* Giải thích điều này:
  + Vĩ độ (latitude) thường là một số dương và nằm trong khoảng từ 40 đến 41 cho khu vực New York City.
  + Kinh độ (longitude) thường là một số âm và nằm trong khoảng từ -75 đến -73 cho khu vực New York City.

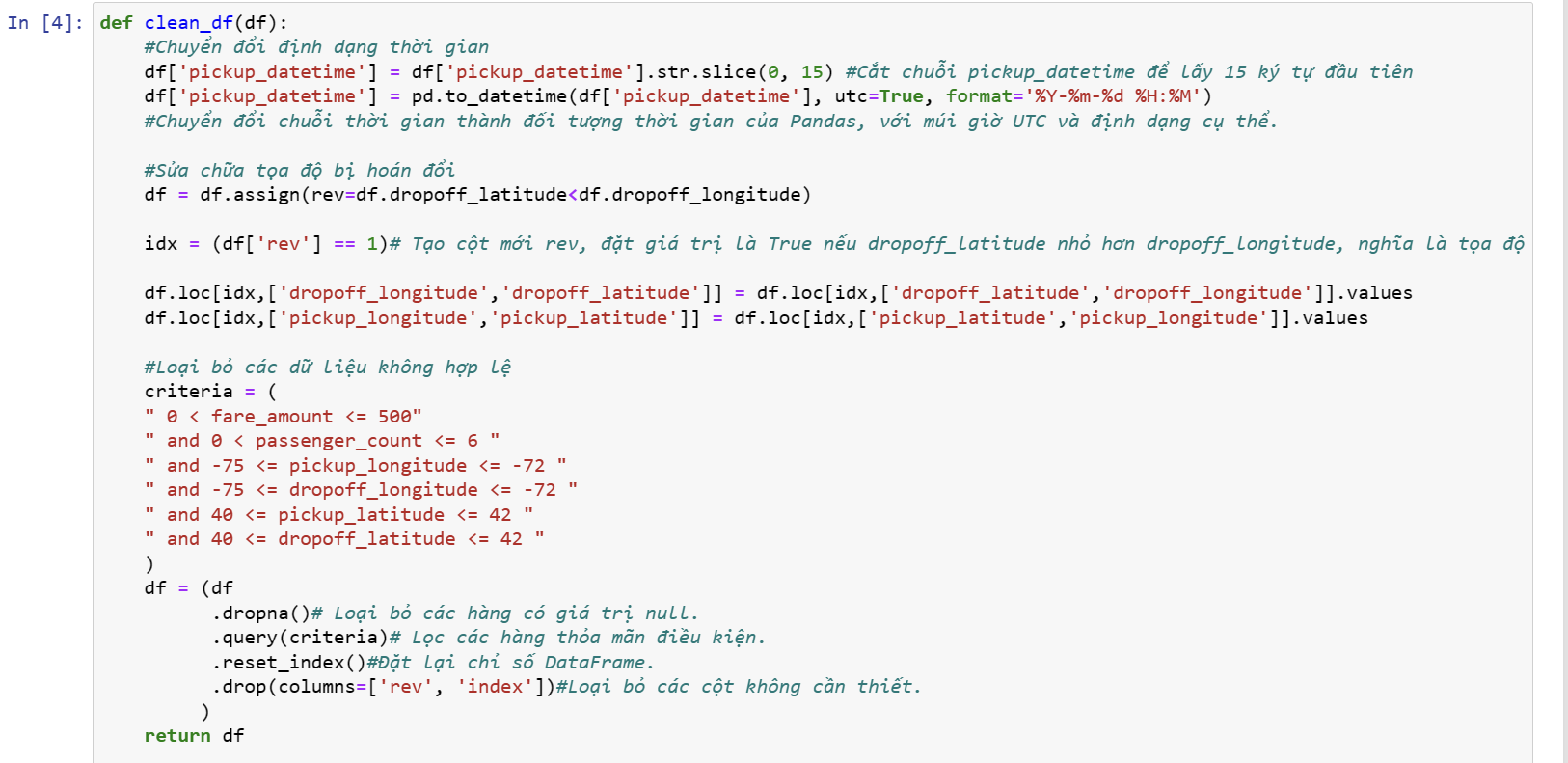
Chúng ta sẽ thử lấy 5 dòng có pickup\_longitude>0 trong tập dữ liệu train.csv để kiểm chứng điều vừa rồi:

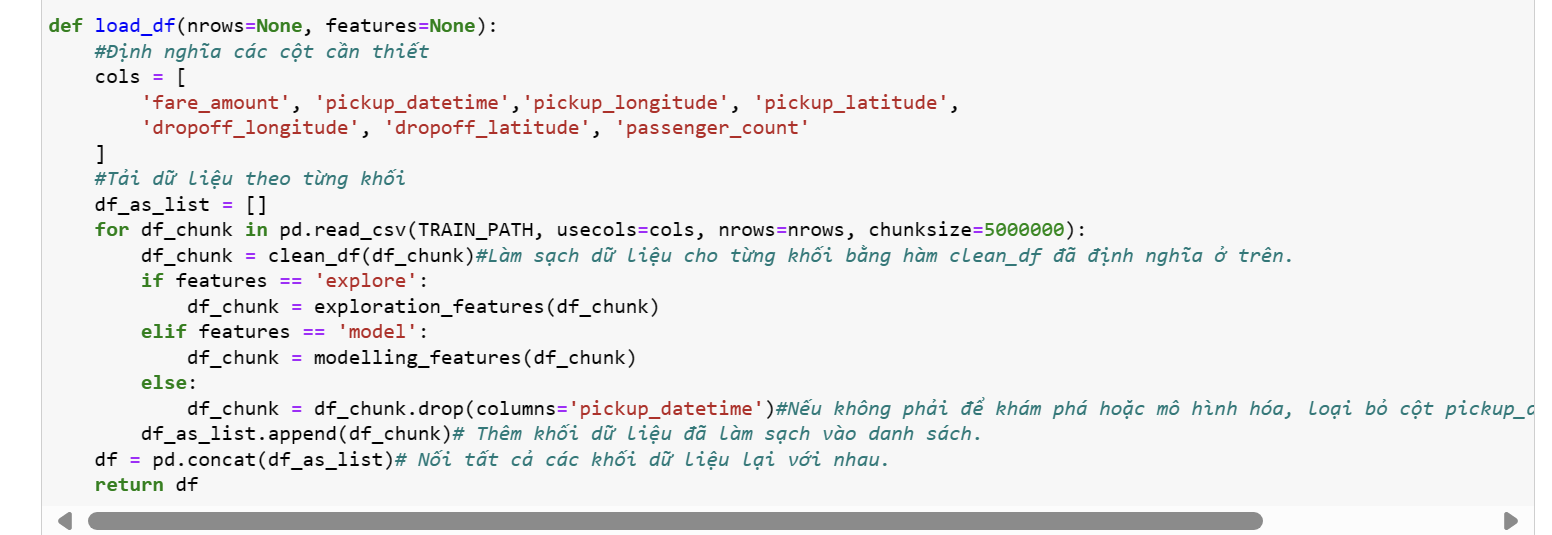


*Hình 10**. Kiểm nghiệm những dòng có giá trị pickup\_longtitude >0*

Dựa vào kết trên, chúng ta đã có thể xác nhận mọi nghi ngờ về việc một số hàng có giá trị kinh độ/vĩ độ đảo ngược là hoàn toàn chính xác. Chúng ta sẽ chuyển các giá trị này về hướng chính xác, sau đó xóa các hàng không chứa giá trị kinh độ/vĩ độ trong New York City.

Sau khi kiểm tra xem các tập dữ liệu có đang gặp vấn dề gì không, bước kế đến chúng ta sẽ tiến hành “làm sạch dữ liệu”.



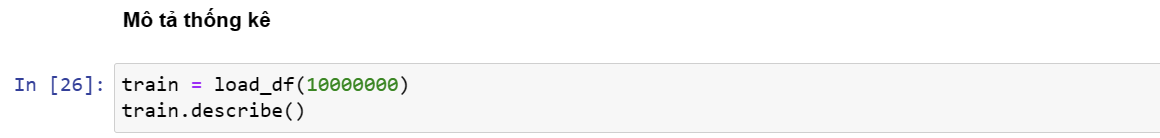


*Hình 11**. Code tạo các hàm để clean dữ liệu*

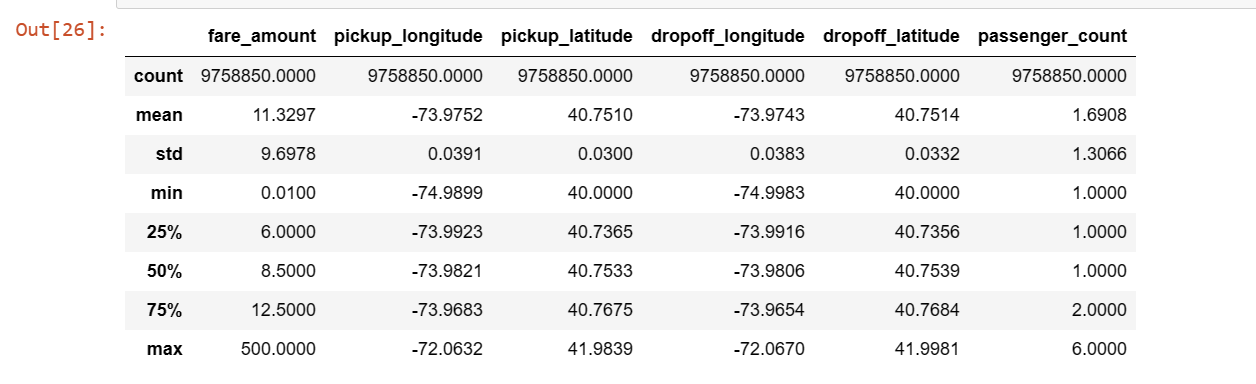
Tóm lại:

* clean\_df: Hàm này làm sạch dữ liệu bằng cách chuyển đổi định dạng thời gian, sửa lỗi hoán đổi tọa độ, và loại bỏ các giá trị không hợp lệ.
* load\_df: Hàm này tải dữ liệu từ file CSV theo từng khối, làm sạch dữ liệu từng khối và nối các khối dữ liệu lại với nhau, sau đó trả về DataFrame hoàn chỉnh.

### 3.3. Mô tả thống kê tập dữ liệu train.csv



*Hình 12**. Code của việc mô tả thống kê tập dữ liệu train.csv*



*Hình 13**. Kết quả của việc mô tả thống kê tập dữ liệu train.csv*

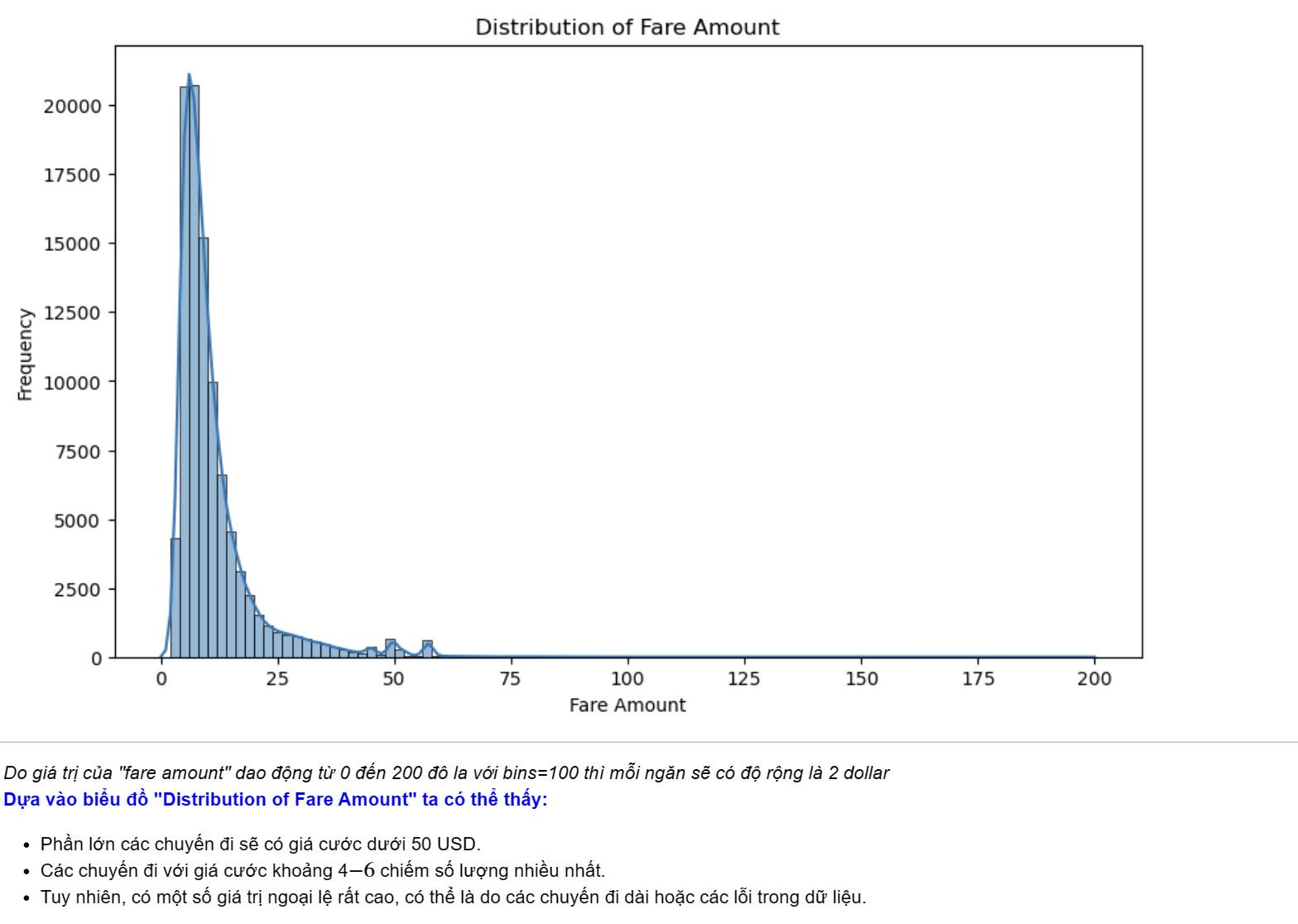
### 3.4. Phân tích dữ liệu (Data analysis)

#### 3.4.1. Phân tích phân phối giá cước taxi

Tại phần này, chúng ta sẽ phân tích phân phối giá cước taxi để kiểm tra xem các chuyến đi ở mức giá bao nhiêu là có số lượng nhiều nhất.



*Hình 14**. Code tạo biểu đồ phân phối giá cước taxi*



*Hình 15**. Biểu đồ phân phối giá cước taxi*

Do giá trị của “fare\_amount” dao động từ 0 đến 200 đô la với bins = 100 thì mỗi ngăn sẽ có độ rộng là 2 đô la.

Dựa vào biểu đồ "Distribution of Fare Amount" ta có thể thấy:

* Phần lớn các chuyến đi sẽ có giá cước dưới 50$.
* Các chuyến đi với giá cước khoảng 4$−6$ chiếm số lượng nhiều nhất.
* Tuy nhiên, có một số giá trị ngoại lệ rất cao, có thể là do các chuyến đi dài hoặc các lỗi trong dữ liệu.

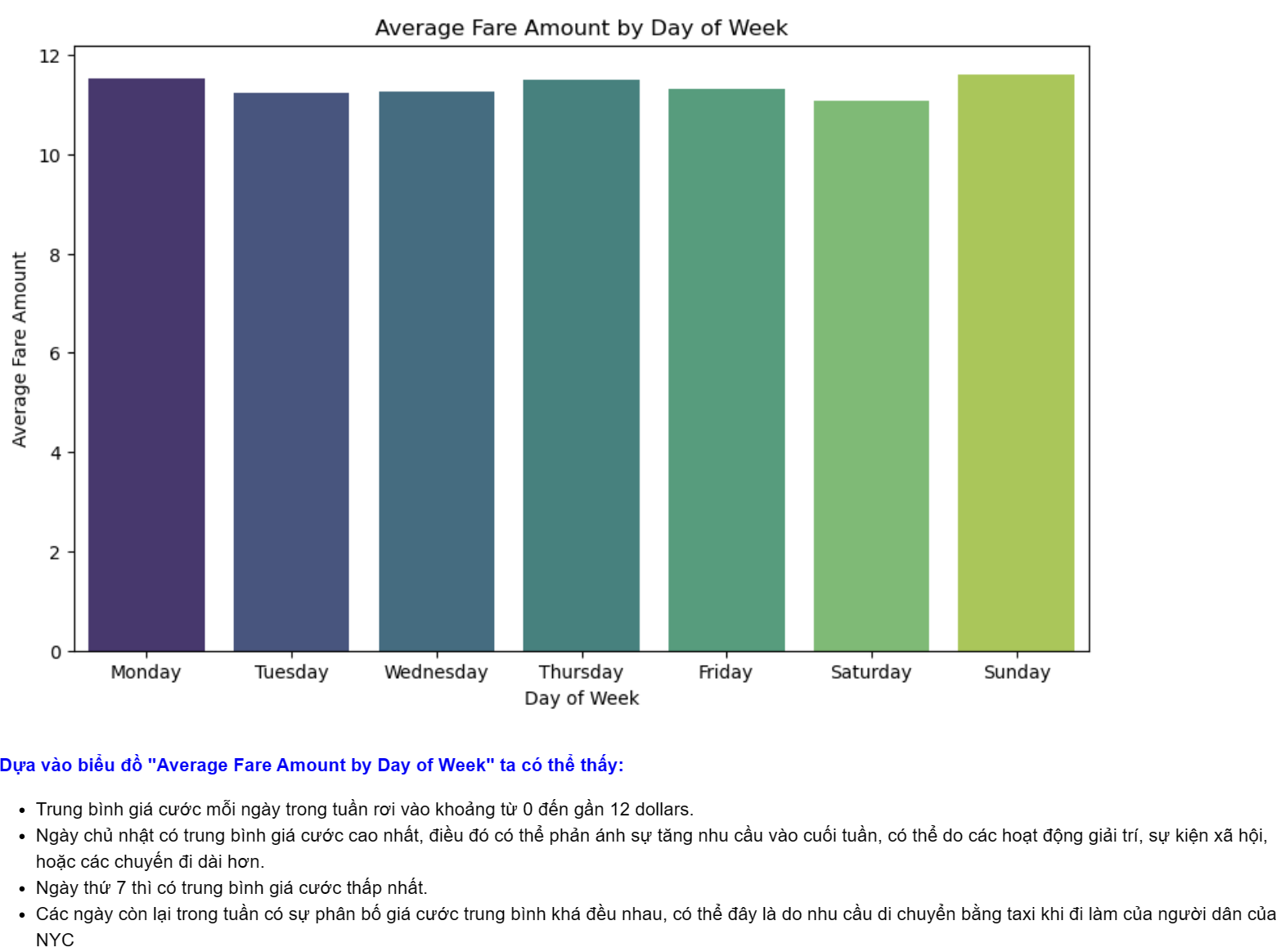
#### 3.4.2. Phân tích trung bình giá cước taxi theo thời gian

Tại đây, chúng ta sẽ phân tích sự phân bố trung bình giá cước taxi theo ngày trong tuần, giờ trong ngày để tìm ra các xu hướng theo thời gian.

##### 3.4.2.1. Phân tích trung bình giá cước taxi theo ngày trong tuần



Hình 16. *Code tạo biểu đồ phân tích trung bình giá cước taxi theo ngày trong tuần*

**

*Hình 17**. Biểu đồ trung bình giá cước taxi theo ngày trong tuần*

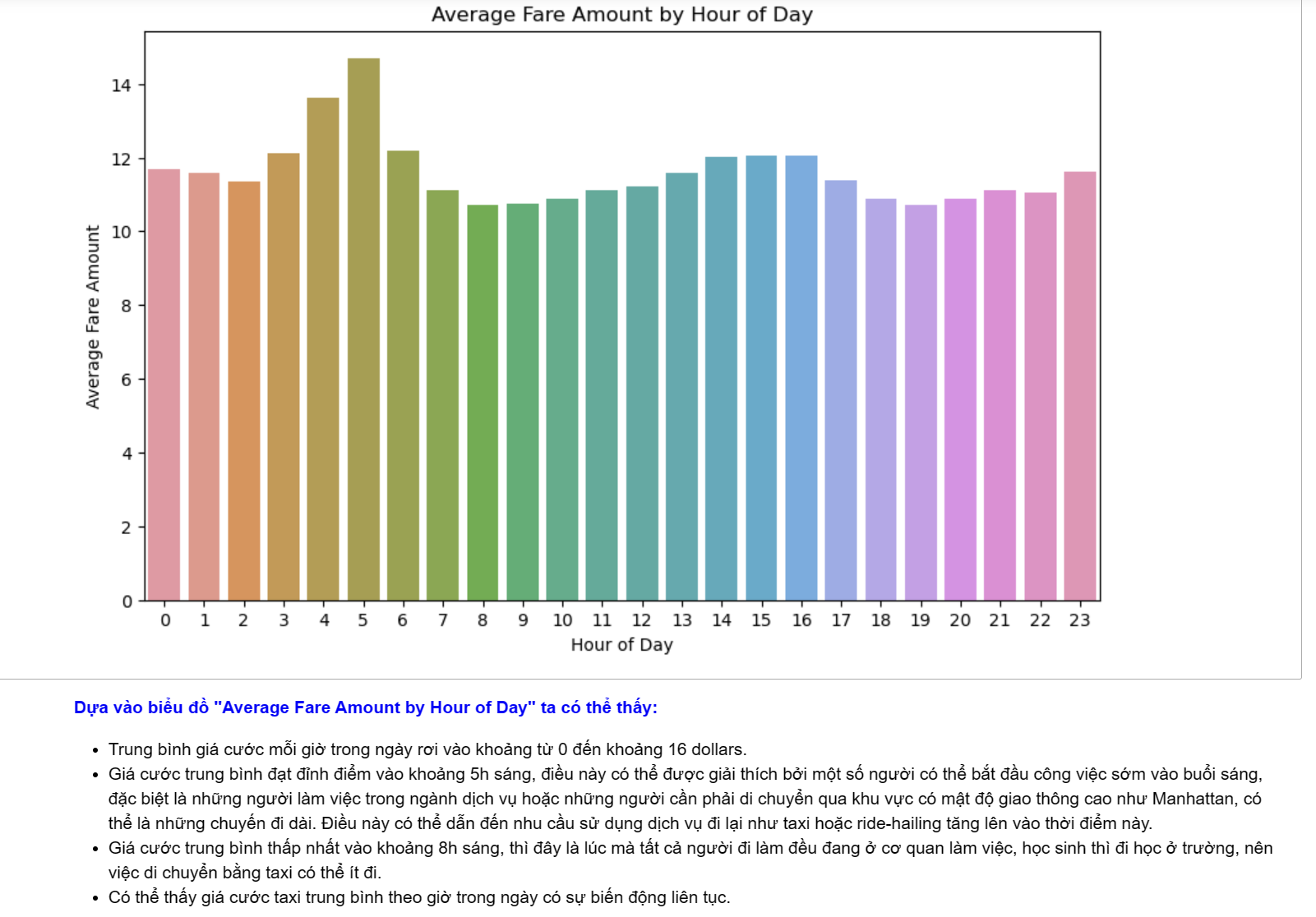
Dựa vào biểu đồ "Average Fare Amount by Day of Week" ta có thể thấy:

* Trung bình giá cước mỗi ngày trong tuần rơi vào khoảng từ 0 đến gần 12 dollars.
* Ngày chủ nhật có trung bình giá cước cao nhất, điều đó có thể phản ánh sự tăng nhu cầu vào cuối tuần, có thể do các hoạt động giải trí, sự kiện xã hội, hoặc các chuyến đi dài hơn.
* Ngày thứ 7 thì có trung bình giá cước thấp nhất.
* Các ngày còn lại trong tuần có sự phân bố giá cước trung bình khá đều nhau, có thể đây là do nhu cầu di chuyển bằng taxi khi đi làm của người dân ở New York City.

##### 3.4.2.2. Phân tích trung bình giá cước taxi theo giờ trong ngày



*Hình 18**. Code tạo biểu đồ phân tích trung bình giá cước taxi theo giờ trong ngày*



*Hình 19**. Biểu đồ trung bình giá cước taxi theo giờ trong ngày*

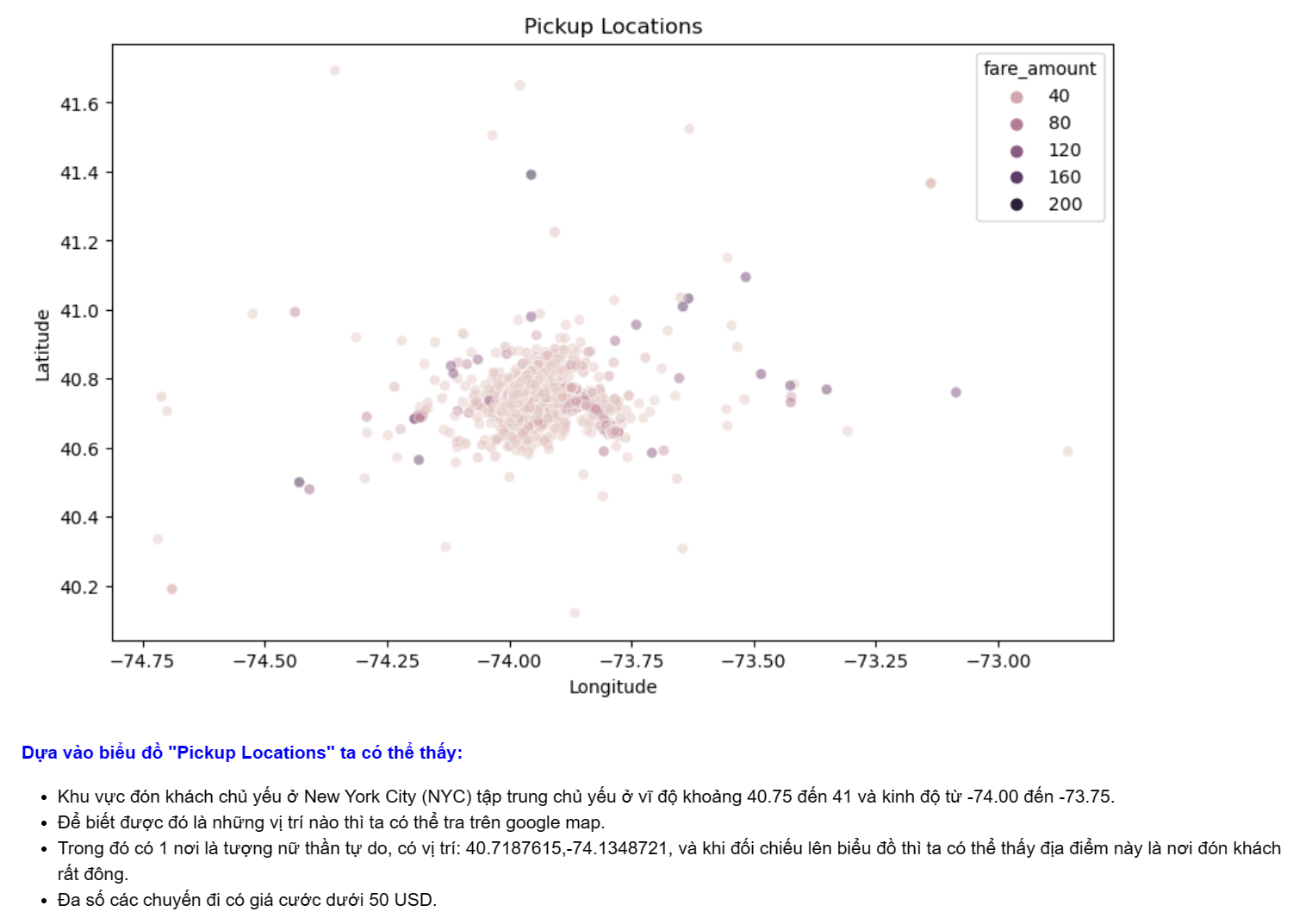
Dựa vào biểu đồ "Average Fare Amount by Hour of Day" ta có thể thấy:

* Trung bình giá cước mỗi giờ trong ngày rơi vào khoảng từ 0 đến khoảng 16 dollars.
* Giá cước trung bình đạt đỉnh điểm vào khoảng 5h sáng, điều này có thể được giải thích bởi một số người có thể bắt đầu công việc sớm vào buổi sáng, đặc biệt là những người làm việc trong ngành dịch vụ hoặc những người cần phải di chuyển qua khu vực có mật độ giao thông cao như Manhattan, có thể là những chuyến đi dài. Điều này có thể dẫn đến nhu cầu sử dụng dịch vụ đi lại như taxi hoặc ride-hailing tăng lên vào thời điểm này.
* Giá cước trung bình thấp nhất vào khoảng 8h sáng, thì đây là lúc mà tất cả người đi làm đều đang ở cơ quan làm việc, học sinh thì đi học ở trường, nên việc di chuyển bằng taxi có thể ít đi.
* Có thể thấy giá cước taxi trung bình theo giờ trong ngày có sự biến động liên tục.

3.4.3. Phân tích vị trí đón và trả khách  
 Trong phần này chúng ta sẽ tìm ra đâu là vị trí mà các chuyến taxi đón và trả khách nhiều nhất.



*Hình 20**. Code vẽ biểu đồ phân tán phân tích vị trí đón khách*

*Hình 21**. Biểu đồ phân tích vị trí taxi đón khách và giá cước*

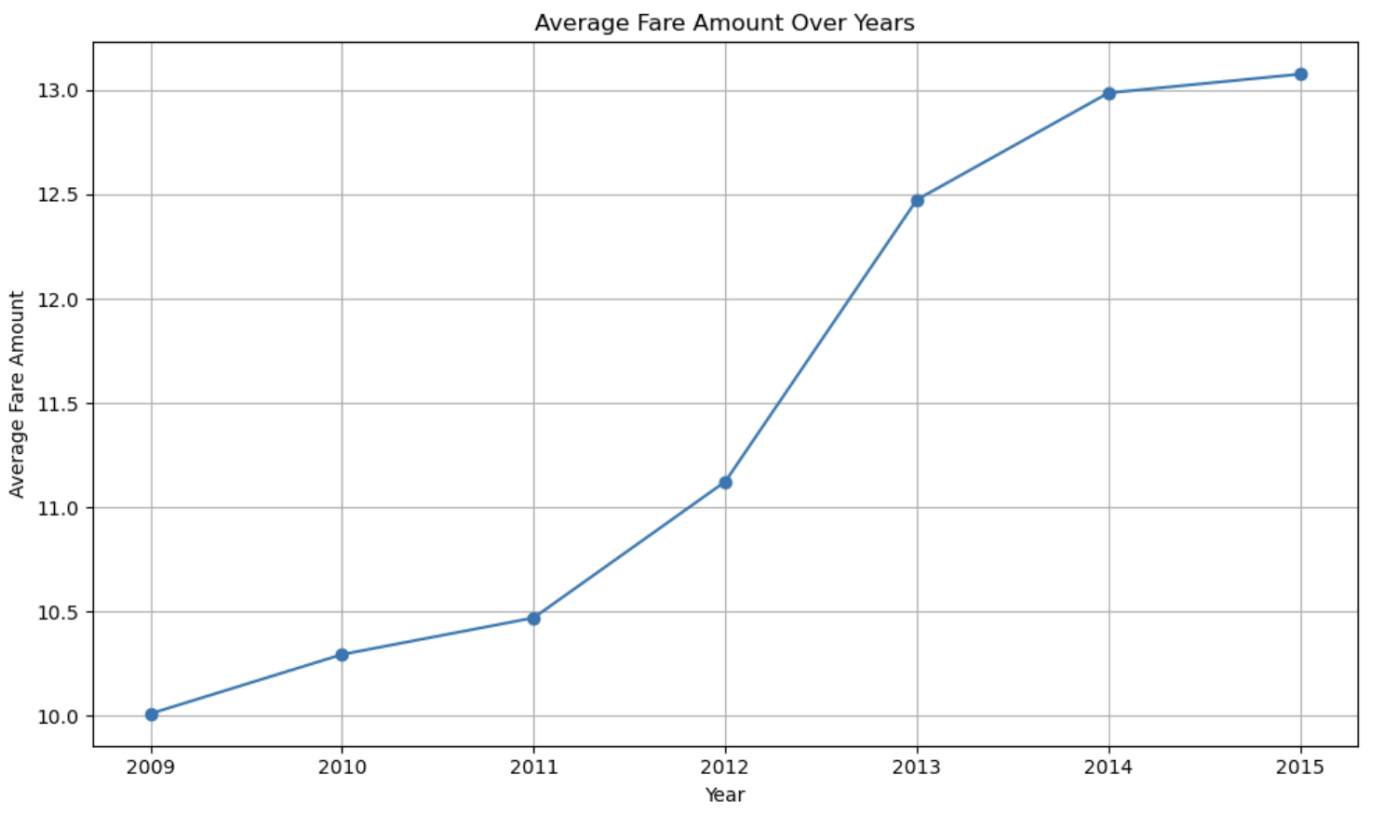
Dựa vào biểu đồ "Pickup Locations" ta có thể thấy:

* Khu vực đón khách chủ yếu ở New York City (NYC) tập trung chủ yếu ở vĩ độ khoảng 40.75 đến 41 và kinh độ từ -74.00 đến -73.75.
* Để biết được đó là những vị trí nào thì ta có thể tra trên Google Map.
* Trong đó có 1 nơi là tượng nữ thần tự do, có vị trí: 40.7187615,-74.1348721, và khi đối chiếu lên biểu đồ thì ta có thể thấy địa điểm này là nơi đón khách rất đông.
* Đa số các chuyến đi có giá cước dưới 50$.

#### 3.4.4. Phân tích xu hướng giá cước taxi qua từng năm



Hình 22. Code tạo biểu đồ đường phân tích xu hướng giá cước taxi qua từng năm



*Hình 23**. Biểu đồ trung bình giá cước taxi qua từng năm*

**Từ những nội dung phân tích bên trên, chúng ta rút ra được những thông tin sau:**

- Phần lớn các chuyến đi nằm trong mức giá cước dưới 50$.

- Giá cước cao nhất vào khoảng 5 giờ sáng, chủ yếu do các chuyến đi dài rời khỏi thành phố hoặc do người dân đi làm từ sớm.

- Trung bình giá cước taxi có mức cao nhất vào ngày chủ nhật.

- Sân bay và trung tâm thành phố, tượng nữ thần tự do là những điểm nóng mà taxi hay đến.

- Trung bình giá cước taxi có xu hướng tăng qua các năm từ 2019-2015 (tăng từ 10.0$ đến trên 13.0$).

**Kết luận từ việc phân tích tập dữ liệu:**

- Mức giá cước từ 50$ trở xuống là khá hợp lí đối với người dân ở New York City, bên cạnh đó điều này còn cho thấy rằng dịch vụ taxi là một lựa chọn phổ biến và hợp lý cho nhu cầu đi lại của người dân và du khách.

**Dự đoán tương lai:**

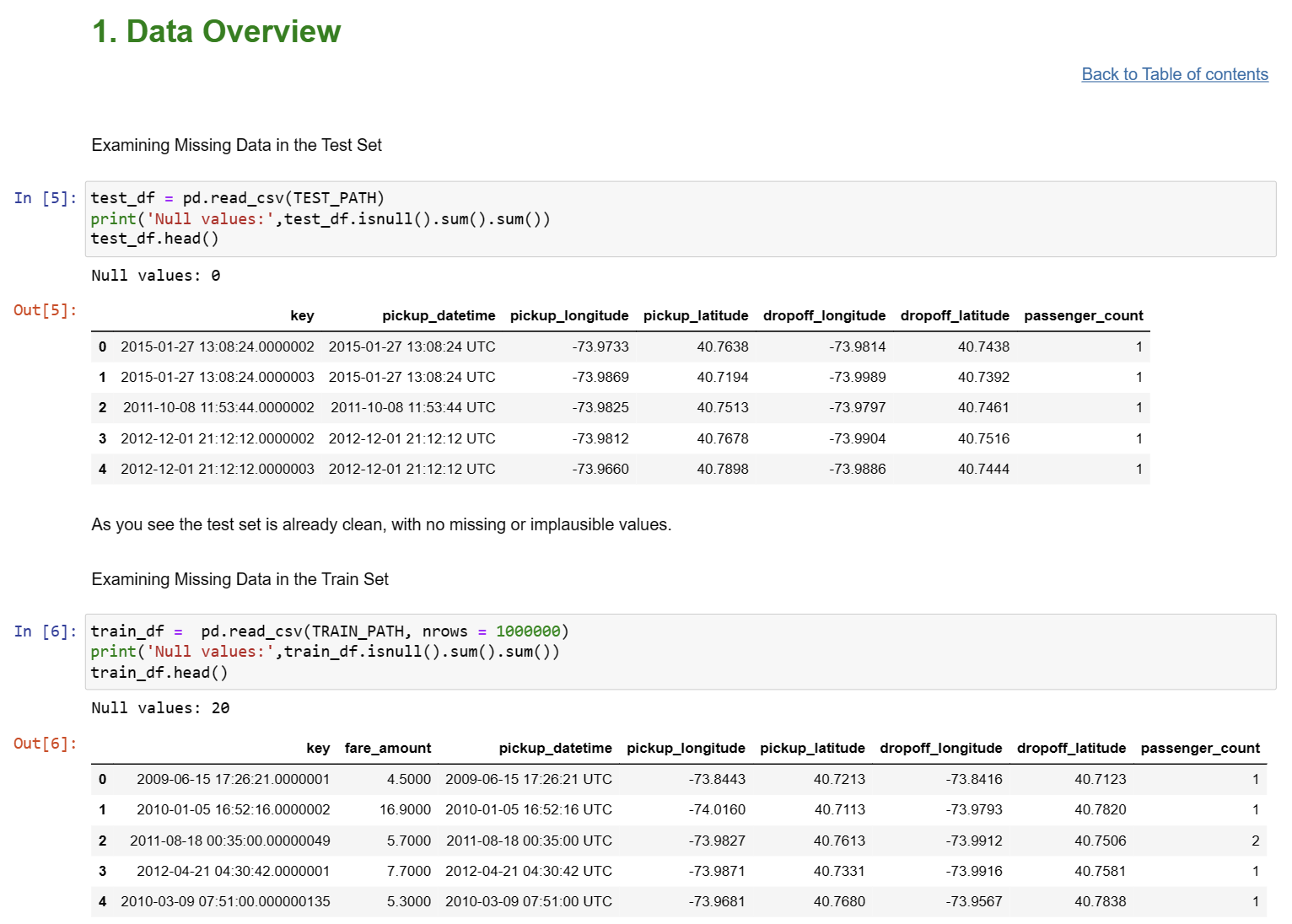
- Tăng giá cước vào khoảng 5 giờ sáng và vào cuối tuần: Với sự tăng trưởng kinh tế và tăng cường hoạt động của thành phố, dự đoán giá cước vào những thời điểm này sẽ tiếp tục tăng lên do nhu cầu đi lại tăng cao từ người dân đi làm từ sớm và từ các hoạt động giải trí, du lịch vào cuối tuần.

- Sân bay và trung tâm thành phố vẫn là điểm nóng của taxi: Với sự phát triển của ngành du lịch và tăng cường hoạt động kinh doanh, dự đoán rằng sân bay và trung tâm thành phố sẽ tiếp tục là những điểm tập trung lớn của người dân và du khách, là nơi mà taxi thường đến và đi.

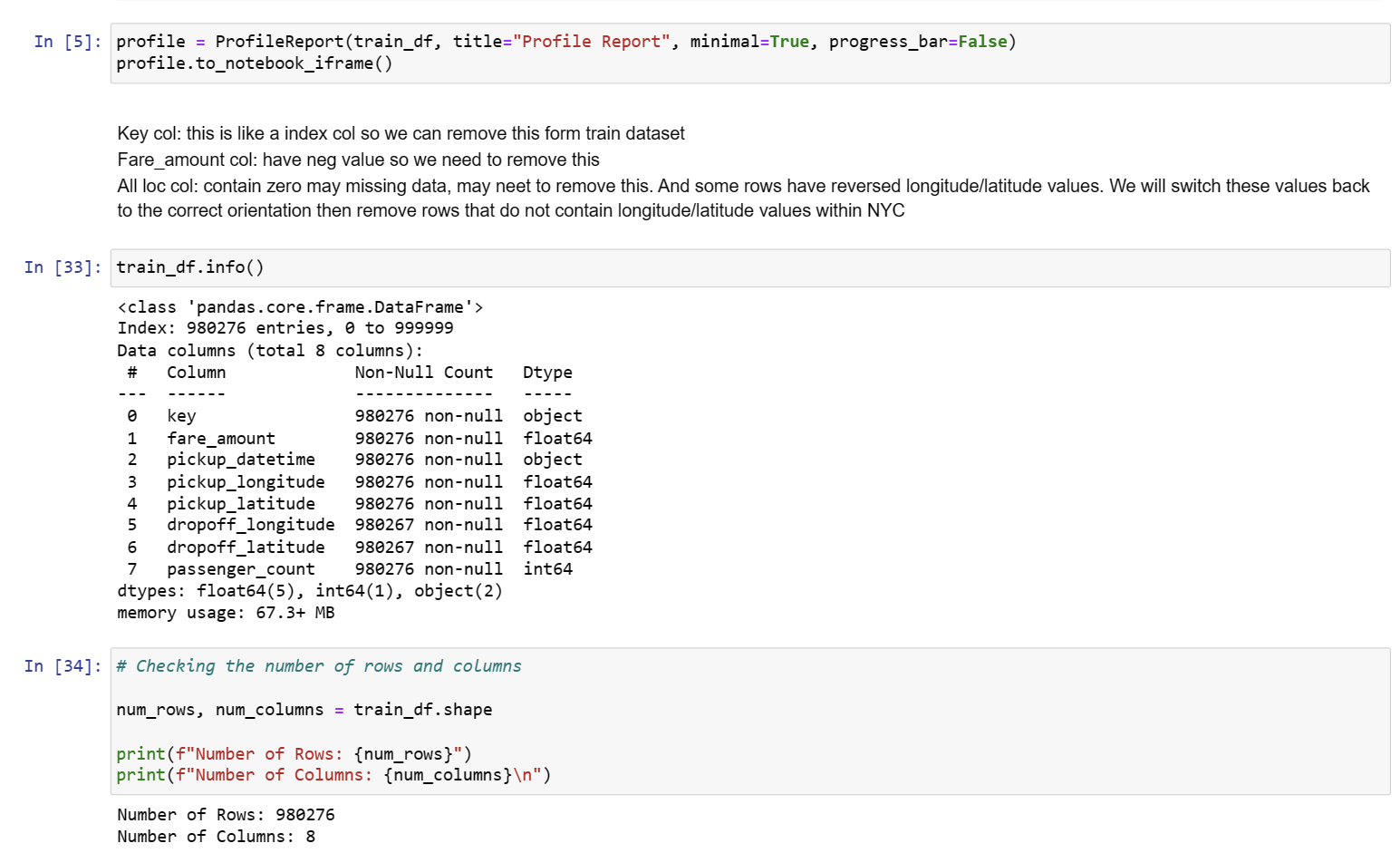
## CHƯƠNG 4: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN, KIỂM CHỨNG VÀ ĐÁNH GIÁ MÔ HÌNH CHO TẬP DỮ LIỆU GIÁ CƯỚC TAXI Ở NEW YORK CITY

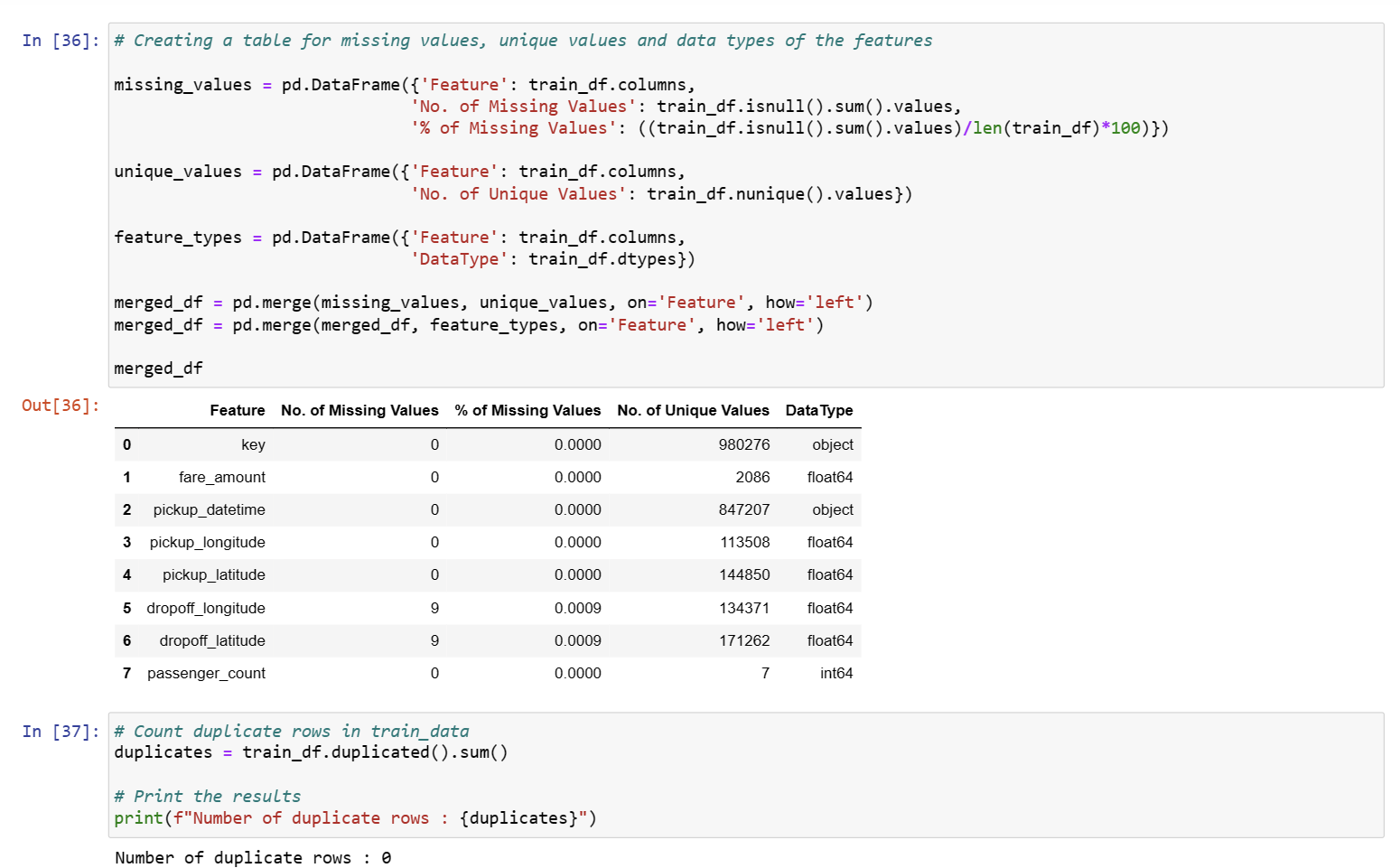
### 4.1. Tổng quan dữ liệu (Data Overview)

Kiểm nghiệm lại các tập dữ liệu train.csv, test.csv và clean data để thực hiện xây dựng mô hình dự đoán.

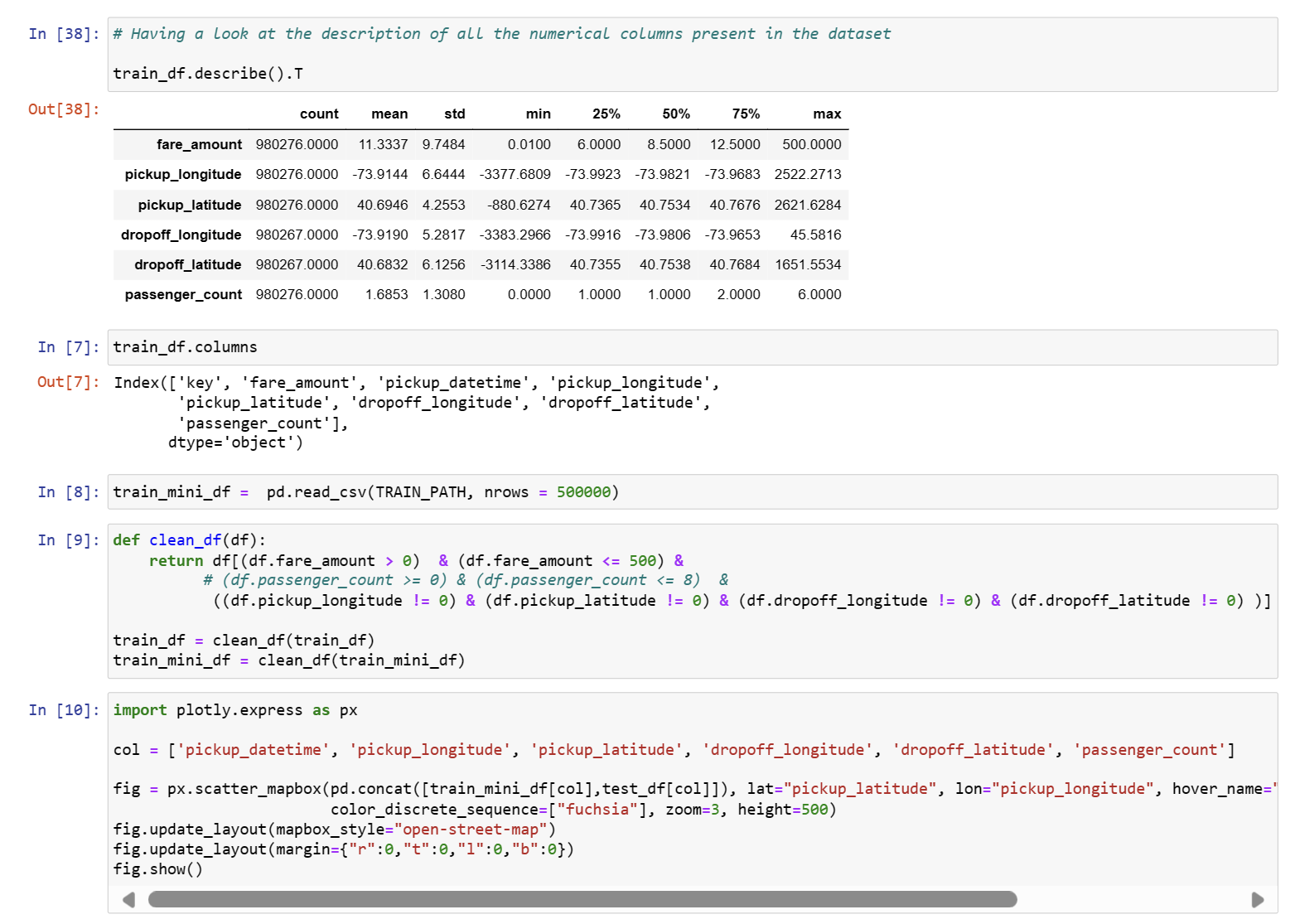


*Hình 24**. Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra và clean data*





*Hình 25**. Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra và clean data*



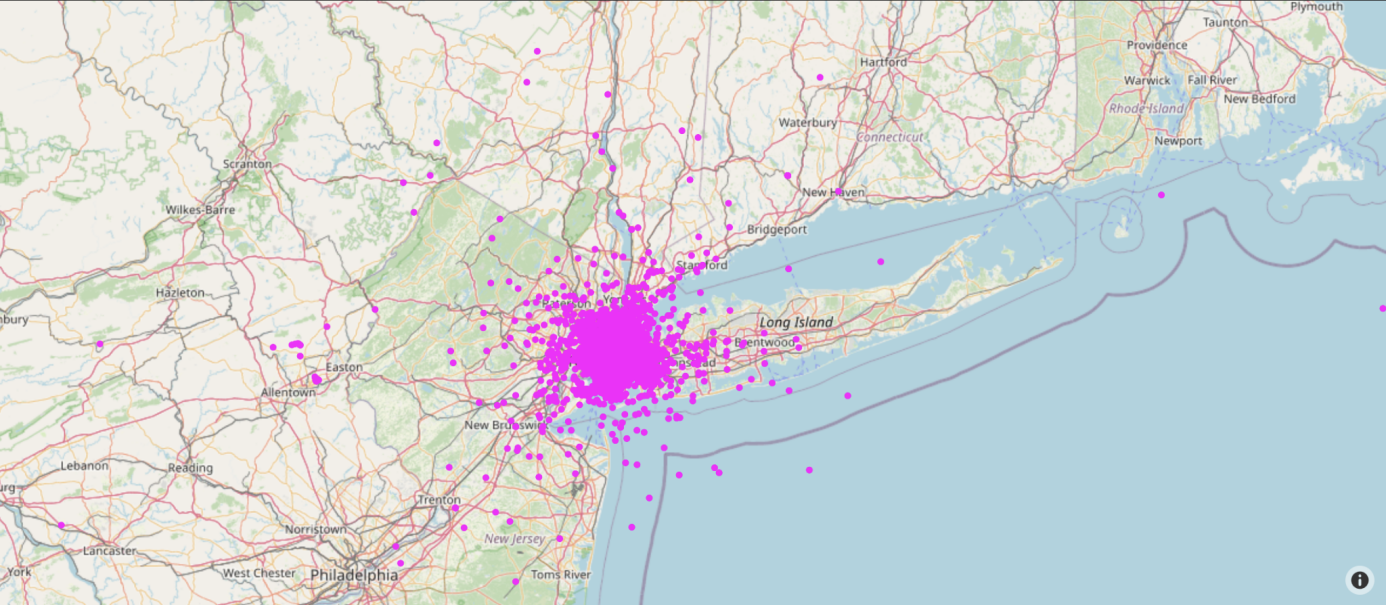
*Hình 26**. Code kiểm nghiệm các tập dữ liệu huấn luyện, kiểm tra và clean data*

### 4.2. Trực quan hóa vị trí taxi đón khách sử dụng plotly express

Trong phần này, chúng ta sẽ hiển thị các điểm đón khách của các chuyến xe taxi trên bản đồ sử dụng vĩ độ và kinh độ từ dữ liệu kết hợp của các tập dữ liệu huấn luyện và kiểm tra.



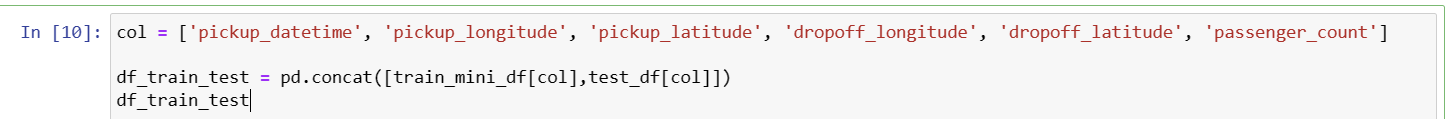
*Hình 27**. Code tạo bản đồ trực quan hóa các vị trí taxi đón khách ở New York City*



*Hình 28**. Bản đồ các vị trí mà taxi đón khách ở New York City*

### 4.3. Trực quan hóa các tuyến đường taxi sử dụng plotly graph objects

Trong phần này, chúng ta sẽ hiển thị các tuyến đường của các chuyến xe taxi trên bản đồ, sử dụng vĩ độ và kinh độ từ dữ liệu kết hợp của các tập dữ liệu huấn luyện và kiểm tra. Mỗi tuyến đường được biểu diễn bằng một đường thẳng nối giữa vị trí đón (pickup) và trả (dropoff) khách.



Hình 29. *Code kết hợp các cột liên quan từ hai DataFrame train và test*

*thành một DataFrame duy nhất*



*Hình 30**. Code tạo bản đồ trực quan hóa các tuyến đường taxi ở New York City*



*Hình 31**. Bản đồ các tuyến đường taxi ở New York City*

### 4.4. Tiền xử lý dữ liệu (Preprocessing Data)

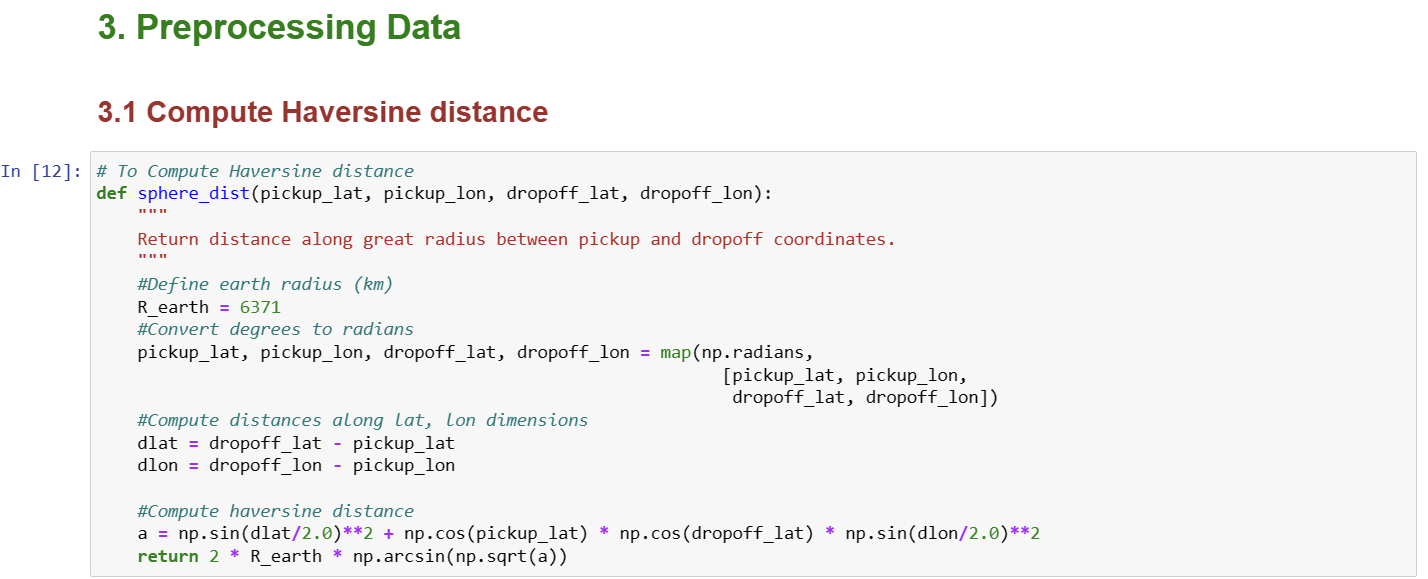
Chúng ta cần tạo ra các thông tin có khả năng và mối liên quan mật thiết nhất, có sự tác động cao đối với giá cước.

Trong trường hợp này khoản cách, vị trí phổ biến là các thông tin có mối tương quan cao và có thể đo dựa trên tập dữ liệu.

Cũng có nhũng thông tin khác có mối tương quan cao như quảng đường di chuyển, tốc độ, thời gian(cao điểm), các yếu tố kỹ thuật, điều kiện môi trường,…

#### 4.4.1. Tính khoảng cách Haversine (Compute Haversine distance)

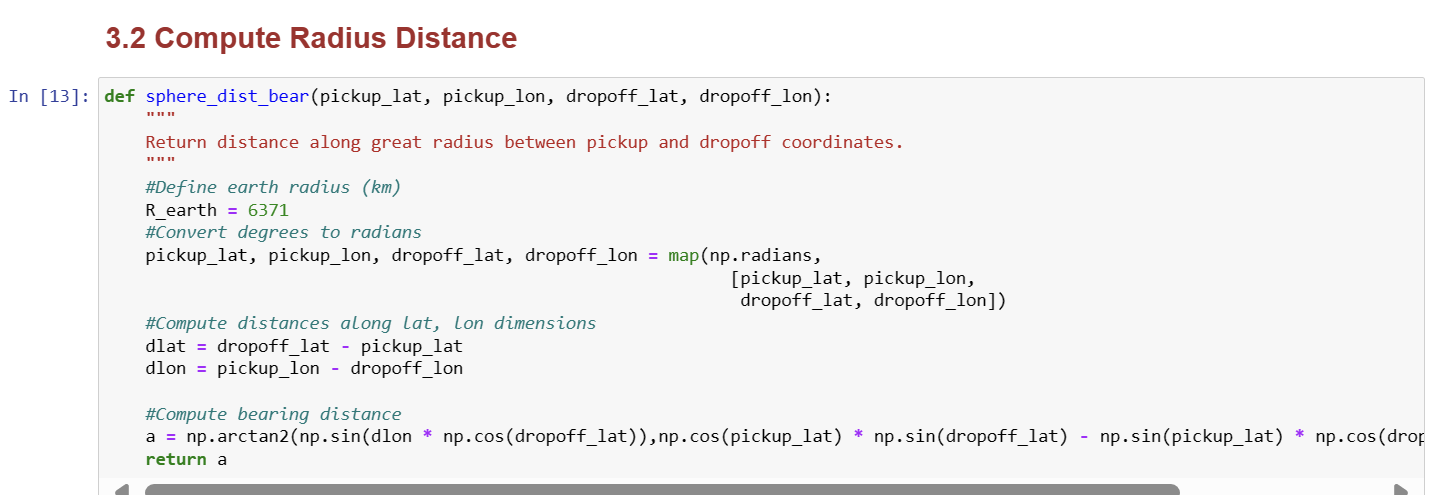
Tạo hàm để tính khoảng cách Haversine giữa hai điểm trên bề mặt Trái Đất, sử dụng tọa độ vĩ độ và kinh độ của điểm đón và điểm trả khách.



*Hình 32**. Code tạo hàm tính khoảng cách Haversine*

#### 4.4.2. Tính khoảng cách radius (Compute Radius Distance)

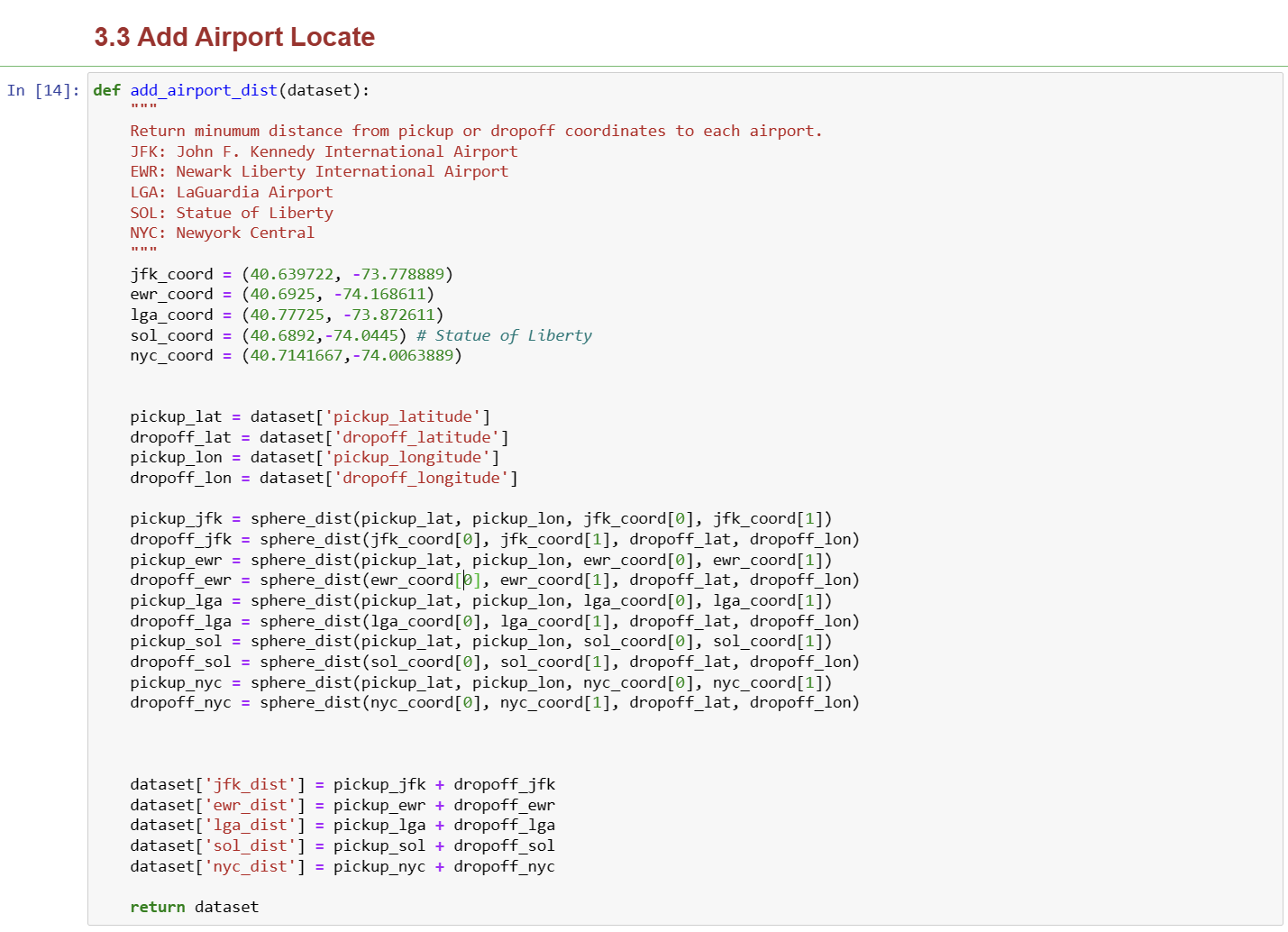
Tạo hàm để tính toán khoảng cách theo hướng (bearing) giữa hai điểm, sử dụng tọa độ vĩ độ và kinh độ của điểm đón và điểm trả khách.



*Hình 33**. Code tạo hàm để tính toán hướng (bearing)*

#### 4.4.3. Tính khoảng cách đến các địa điểm nổi bật ở New York (Add Airport Locate)

Tạo hàm để tính toán khoảng cách từ điểm đón và trả khách của mỗi chuyến taxi đến năm địa điểm quan trọng (JFK, EWR, LGA *(mã của ba sân bay quốc tế chính ở New York)*, Tượng Nữ thần Tự do, và trung tâm New York) và thêm các khoảng cách này vào DataFrame dưới dạng các cột mới. Mỗi cột mới đại diện cho tổng khoảng cách từ điểm đón và trả khách đến một địa điểm cụ thể.

**

*Hình 34**. Code tạo hàm tính khoảng cách đến các địa điểm nổi bật ở New York City*

4.4.4. Tiền xử lý (Preprocess)

Tạo các hàm tiền xử lý dữ liệu:

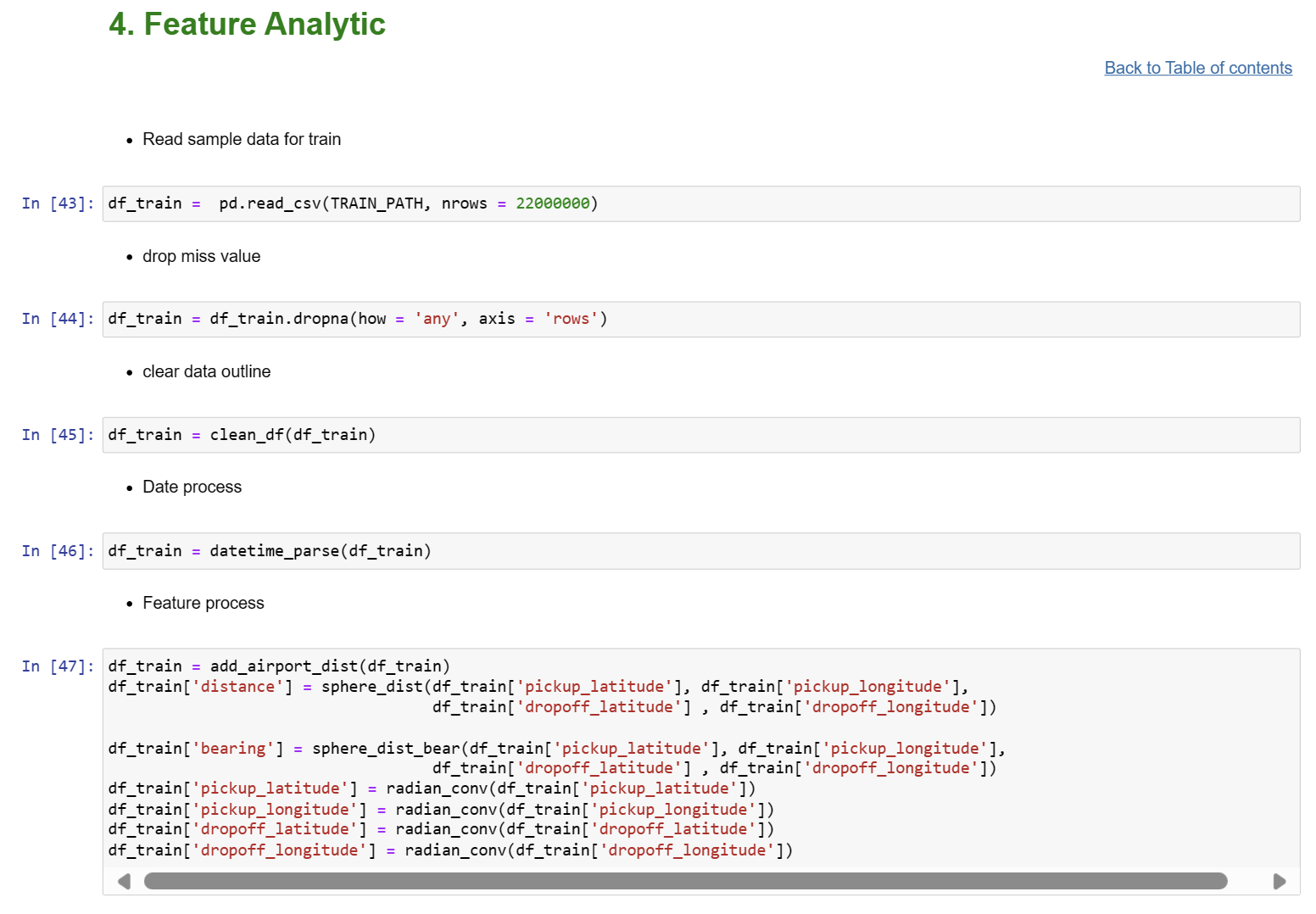
* Hàm radian\_conv: Chuyển đổi độ sang radian.
* Hàm datetime\_parse: Chuyển đổi cột thời gian sang định dạng datetime và trích xuất thông tin thời gian.
* Hàm clean\_df: Làm sạch dữ liệu bằng cách loại bỏ các bản ghi không hợp lệ hoặc nằm ngoài phạm vi địa lý được xác định.



*Hình 35**. Code tạo các hàm tiền xử lý dữ liệu*

### 4.4. Phân tích đặc trưng (Feature Analytic)

Phân tích và tiền xử lý dữ liệu để chuẩn bị cho việc tạo ra các đặc trưng (features) để sử dụng trong mô hình dự đoán giá cước taxi.



*Hình 36**. Code thực hiện phân tích và tiền xử lý dữ liệu*

### 4.5. Xây dựng Model (Build Model)

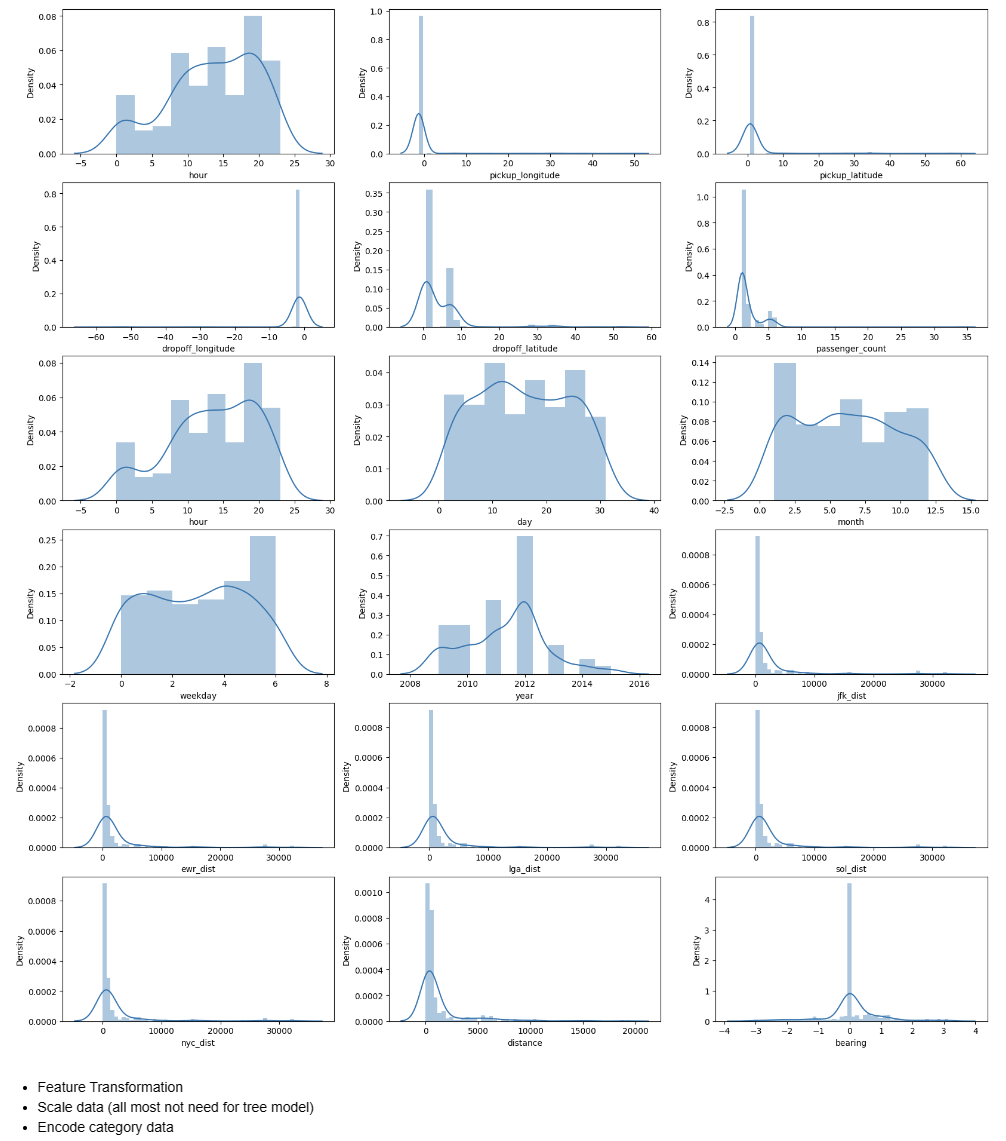
#### 4.5.1. Train-Val split

Trước khi xây dựng mô hình dự đoán giá cước taxi, chúng ta cần chia dữ liệu thành tập huấn luyện và tập validation để đánh giá hiệu suất của mô hình. Tiếp theo, chúng ta sẽ vẽ biểu đồ phân phối của các đặc trưng trong tập huấn luyện để hiểu rõ hơn về phân bố của chúng. Điều này sẽ giúp chúng ta đưa ra quyết định thông minh hơn khi xây dựng và điều chỉnh mô hình dự đoán giá cước taxi.

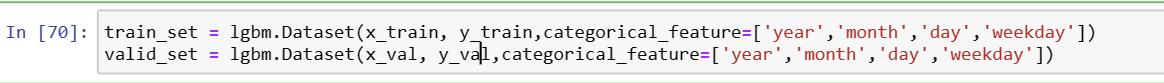




*Hình 37**. Code trực quan hóa phân phối đặc trưng trong dữ liệu huấn luyện*



*Hình 38**. Kết quả trực quan hóa phân phối đặc trưng trong dữ liệu huấn luyện*

**

*Hình 39**. Code chuẩn bị dữ liệu huấn luyện và validation cho mô hình LightGBM*

#### 4.5.2. Train Model

Trong phần này chúng ta sẽ thực hiện việc huấn luyện mô hình sử dụng thuật toán Gradient Boosting Decision Tree (GBDT) thông qua thư viện LightGBM.



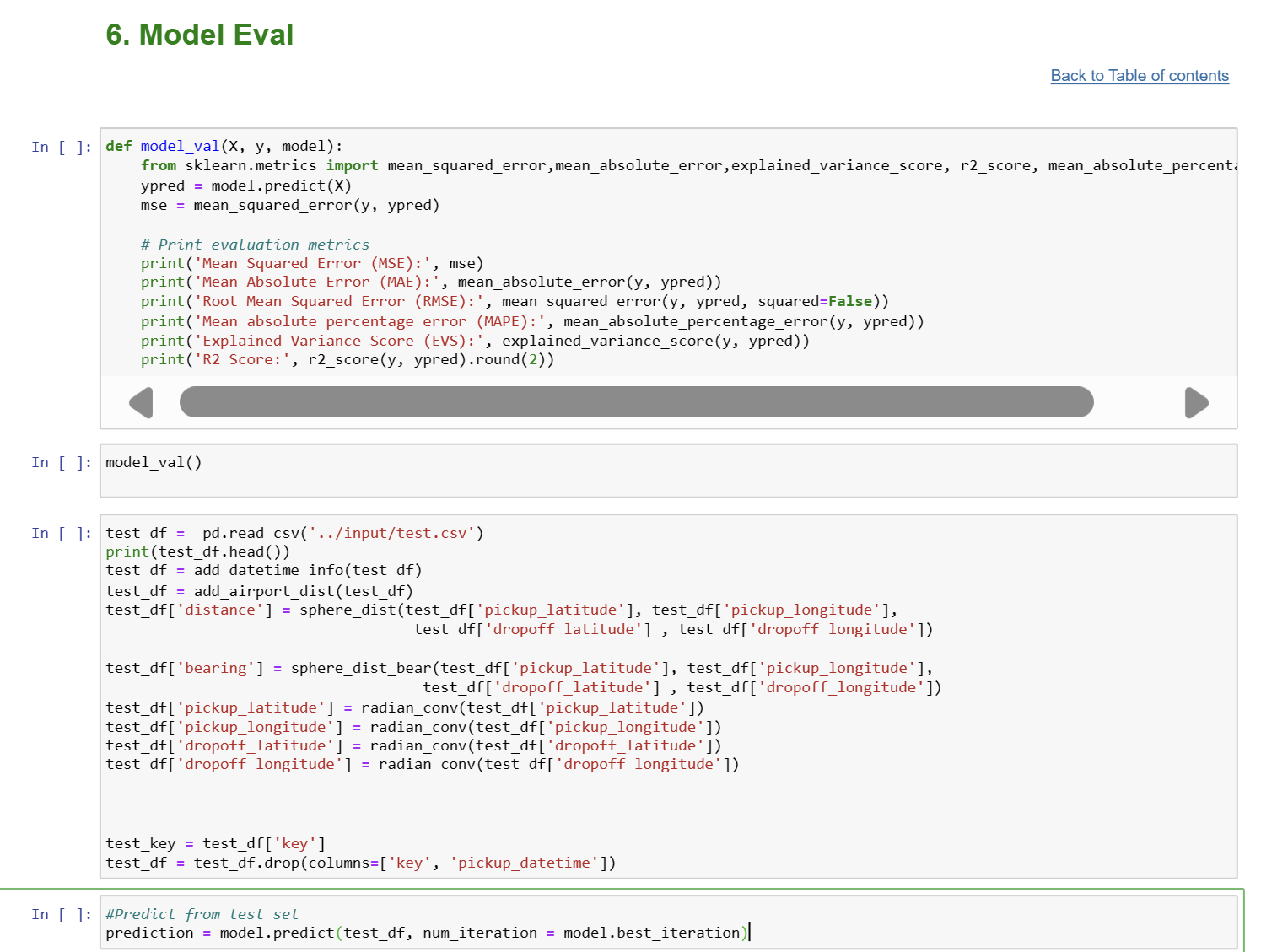
A screenshot of a computer program

Description automatically generated

*Hình 40**. Code huấn luyện mô hình*

### 4.6. Mô hình Eval (Model Eval)

Ở phần này chúng ta sẽ tiến hành đánh giá hiệu suất của mô hình dự đoán giá cước taxi. Bằng cách sử dụng các độ đo như Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Explained Variance Score (EVS), và R2 Score. Sau đó chúng ta sẽ dự đoán giá cước taxi cho tập dữ liệu kiểm tra.



*Kêt quả đánh giá:*

*A screenshot of a computer

Description automatically generated*

*A screenshot of a computer code

Description automatically generated*

*Hình 41**. đánh giá hiệu suất mô hình*

# PHẦN KẾT LUẬN

* + - 1. **Kết quả đạt được**
* Về kiến thức

+ Đựa trên tập dữ liệu:

Kiến thức về phần tích dữ liệu lớn, đến rất lớn. Tập trung hiểu được cốt lõi các giá trị thành phần chính quan trọng của dữ liệu, cách chúng tác động tới nhau.

Bằng các bước tìm hiểu trên tập dữ liệu lớn giúp chúng em: Bước đầu nắm bắt được những lỗi trong dữ liệu và trong quá trình thu thập. Kế đến các bước xử lí bằng sample, các tập con nhỏ, có tính khái quát cao, và các tập lớn hơn một cách hơp lý.

Tốc độ tính toán trên tập dữ liệu lớn khá chậm, nên cần có các phương pháp xử lý giúp cải thiện tốc độ. Mang lại hiệu qua bao phủ cao.

+ Trên quá trình train model:

Hiểu dữ liệu là một bước vô cùng quan trọn, từ đó chúng em có khả năng đưa ra được các phân, lực chọn các đặt trương hợp lý, có mới tương quan cao.

Hiểu được việc xử lý dữ liệu ảnh hưởng đến kết quả model như thế nào.

Biết được các tối ưu các siêu tham số để tạo được tập model tốt nhất.

* Về sản phẩm

+ Sản phẩm có khả năng dự đoán tốt giá cả hợp lý, độ chính xác để đề xuất mức giá dự đoán trước cao.

+ Tính ứng dụng model cho các mô hình báo giá taxi tốt, có thể làm mức giá quy chuẩn cho các doanh nghiêp taxi

* + - 1. **Ưu và nhược điểm**
* Ưu điểm: Mô hình dự đoán nhanh, mạnh, tính chính xác cao có khả năng ứng dụng triển khai trên các sản phẩm di động.
* Nhược điểm: Độ bao phủ thấp (tập trung ở new york), chưa có sự linh hoạt do thiếu các đặt trưng về môi trường, chi phí, điều kiện vận hành.
  + - 1. **Hướng phát triển**
* Fine turning trên nhiều tập dataset của nhiều khu vực hơn.
* Thêm các yếu tố môi trường cho sản phẩm.

# PHẦN TÀI LIỆU THAM KHẢO

[1]. Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow,* Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2019

[2]. Kaggle (2018), *New York City Taxi Fare Prediction*, truy cập ngày 10/5/2024, đường dẫn: <https://www.kaggle.com/competitions/new-york-city-taxi-fare-prediction/data>