# PHÂN TÍCH BỘ DỮ LIỆU YELLOW TAXI TRIP NĂM 2021

## 1. Bộ dữ liệu

### 1.1. Nguồn dữ liệu

<https://data.cityofnewyork.us/Transportation/2021-Yellow-Taxi-Trip-Data/m6nq-qud6>

### 1.2. Tổng quan về bộ dữ liệu

- Số lượng dữ liệu: 5,967,103

- Số cột: 18

- %NaN: 5.44%

- Ý nghĩa: Các dữ liệu được tạo từ các chuyến đi do Nhà cung cấp dịch vụ công nghệ taxi (TSP) thực hiện. Mỗi hàng đại diện cho một chuyến đi bằng taxi màu vàng vào năm 2021. Dữ liệu bao gồm các cột ghi ngày/giờ đón và trả khách, địa điểm đón và trả khách, khoảng cách chuyến đi, giá vé được chia thành từng khoản, loại giá, hình thức thanh toán và số lượng hành khách do tài xế báo cáo.

**Bảng 1.1:** Tổng quan về bộ dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **%NaN** | **Kiểu dữ liệu** | **Ý nghĩa** |
| VendorID | 5.43845 | Int64 | Mã nhà cung cấp taxi |
| tpep\_pickup\_datetime | 0.00007 | Datetime | Thời gian taxi bắt đầu chạy |
| tpep\_dropoff\_datetime | 0.00007 | Datetime | Thời gian taxi kết thúc chuyến đi |
| passenger\_count | 5.43854 | Float64 | Số lượng hành khách trên xe |
| trip\_distance | 0.00008 | Float64 | Khoảng cách của chuyến đi |
| RatecodeID | 5.43854 | Float64 | Mã chương trình giá để xác định phương pháp tính cước cho chuyến đi |
| store\_and\_fwd\_flag | 5.43854 | Object | bản ghi chuyến đi có được lưu trữ trong bộ nhớ của phương tiện trước khi gửi đến nhà cung cấp (công ty taxi) hay không |
| PULocationID | 0.00008 | Float64 | Khu vực đồng hồ taxi được đặt vào, đại diện cho vị trí lấy khách |
| DOLocationID | 0.00008 | Float64 | Khu vực đồng hồ taxi ngừng hoạt động, đại diện cho vị trí trả khách |
| payment\_type | 5.43854 | Float64 | Mã xác định cách hành khách thanh toán cho chuyến đi |
| fare\_amount | 0.00008 | Float64 | Cước thời gian và khoảng cách tính bằng đồng hồ taxi |
| extra | 0.00008 | Float64 | Các khoản phụ và phụ phí khác được thêm vào cước, bao gồm các khoản phí giờ cao điểm và qua đêm |
| mta\_tax | 0.00008 | Float64 | Thuế MTA $0.5 tự động kích hoạt dựa trên tỷ lệ tính theo đồng hồ taxi |
| tip\_amount | 0.00008 | Float64 | Số tiền tip được thêm vào cước |
| tolls\_amount | 0.00008 | Float64 | Tổng số tiền phải trả cho tất cả các trạm thu phí trong chuyến đi |
| improvement\_surcharge | 0.00008 | Float64 | Phụ phí cải tiến $0.3 áp dụng cho chuyến đi khi khởi hành |
| total\_amount | 0.00008 | Float64 | Tổng số tiền tính cho hành khách cho chuyến đi, bao gồm tất cả các cước, phụ phí và tiền phải trả |
| congestion\_surcharge | 0.00008 | Float64 | Cột này không được mô tả đầy đủ trong thông tin được cung cấp, nhưng có thể đại diện cho bất kỳ khoản phụ phí nào liên quan đến giá cả khi kẹt xe, nếu có áp dụng trong tập dữ liệu |

## Tiền xử lý dữ liệu

### Môi trường sử dụng

* Python: 3.11.2
* Pandas: 2.0.1
* Matplotlib: 3.7.2
* Searborn: 0.12.2

### Kiểm tra kiểu dữ liệu

Kiểm tra kiểu dữ liệu của từng cột sau khi đọc file csv. Kết quả thu được hầu hết các cột đều khớp với kiểu dữ liệu, ngoại trừ:

**Bảng 2.1:** Thông tin cột và kiểu dữ liệu

|  |  |  |
| --- | --- | --- |
| **Tên cột** | **Kiểu dữ liệu ban đầu** | **Kiểu dữ liệu cần chuyển đổi** |
| VendorID | Object | Int64 |
| tpep\_pickup\_datetime | Object | Datetime64 |
| Tpep\_dropoff\_datetime | Object | Datetime64 |

A screenshot of a computer

Description automatically generated

**Hình 2.1:** Thông tin kiểu dữ liệu cho các cột

\*Đối với cột tpep\_dropoff\_datetime:

Xuất hiện 1 giá trị bị nhiễu nên tiến hành xóa giá trị nhiễu rồi chuyển đổi kiểu dữ liệu

A computer screen shot of text

Description automatically generated

**Hình 2.2:** Xóa điểm dữ liệu không phù hợp

### Xử lý giá trị NaN

Đối với bộ dữ liệu, tiến hành xử lý các cột có số lượng NaN theo 2 cách: drop đối với những cột có số lượng NaN không đáng kể (%NaN <0.1%) và visualize để xem xét cách xử lý NaN đối với cột có %NaN > 5%.

A screen shot of a computer

Description automatically generated

**Hình 2.3:** Thông tin các điểm dữ liệu cuối cùng của bộ dữ liệu

A screenshot of a computer

Description automatically generated

**Hình 2.4:** Hình thể hiện % các giá trị NaN của các cột

Từ hình 2.3 và hình 2.4, chia quá trình xử lý các giá trị NaN làm hai phần:

* + 1. Các cột có số lượng NaN < 0.1%

Vì các cột có số lượng NaN <=5 tập trung ở các record cuối nên tiến hành xóa các record này:

A computer screen shot of a code

Description automatically generated

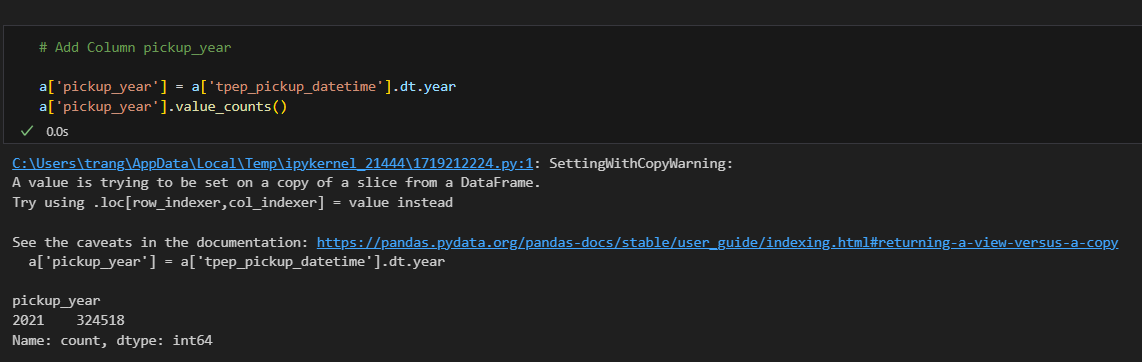
* + 1. Các cột có số lượng NaN > 5%

Tạo dataframe a chỉ chứa giá trị NaN của cột VendorID và nhận thấy rằng 5 cột đều có NaN ở cùng index:

A screenshot of a computer screen

Description automatically generated

Đồng thời, các giá trị NaN đều được ghi nhận vào thời gian năm 2021:



Trong khi đó, năm phân bổ của bộ dữ liệu như sau:

A computer screen shot of a program

Description automatically generated

Vì vậy, tiến hành filter bộ dữ liệu theo năm 2021:

A screen shot of a computer

Description automatically generated

## Visualization

### 3.1. Biểu đồ histogram data trước và sau khi xử lý

\*Trước khi xử lý:

A screenshot of a graph

Description automatically generated

**Hình 3.1:** Biểu đồ thể hiện các cột dữ liệu bị NaN > 5% trước khi xử lý

\*Phương pháp xử lý:

- Đối với 4 cột “passenger\_count”, “RatecodeID”, “store\_and\_fwd\_flag”, “payment\_type”, fillna bằng giá trị xuất hiện nhiều nhất

A screenshot of a computer program

Description automatically generated

- Đối với cột “VendorID”:

Do count giá trị “VendorID” = “1” chiếm gần 50% “VendorID” = “2” nên tiến hành so sánh để chọn giữa phương pháp bfill và ffill.

A screenshot of a computer program

Description automatically generated

Sau khi so sánh, nhận thấy dùng phương pháp ffill sẽ làm ít ảnh hưởng tới phân phối của bộ dữ liệu nhất nên tiến hành ffill và chuyển đổi kiểu dữ liệu của cột “VendorID” thành int64.

\*Sau khi xử lý:

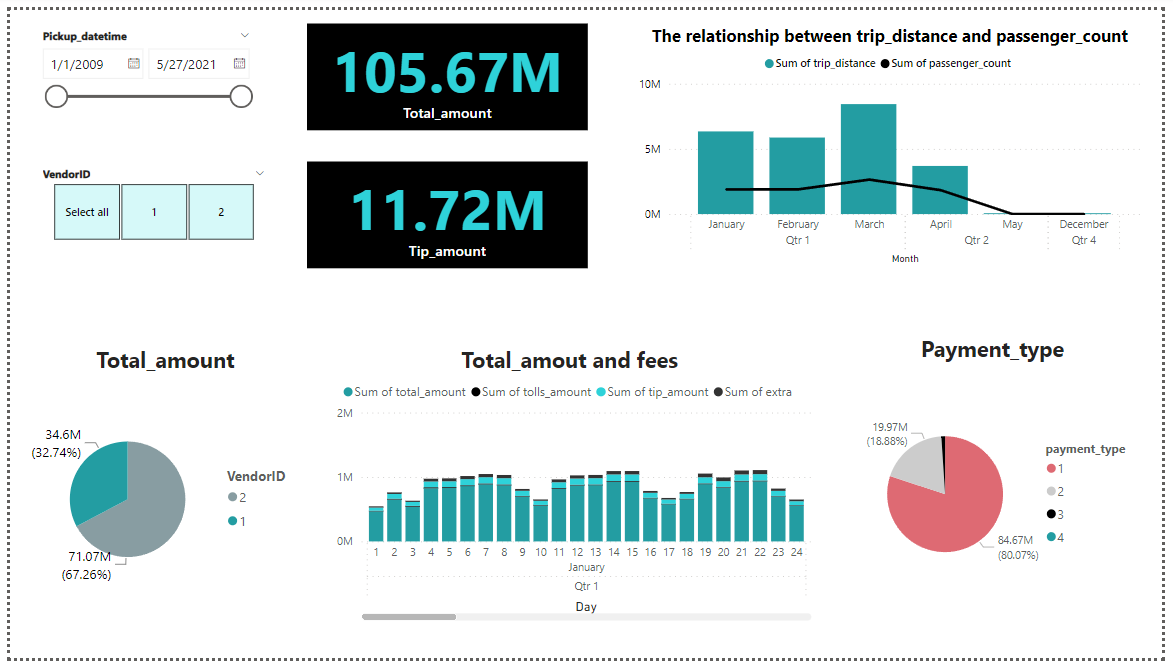
A collage of red and white graphs

Description automatically generated

**Hình 3.2:** Biểu đồ thể hiện các cột dữ liệu bị NaN > 5% sau khi xử lý

*Nhận xét: Phân phối của dữ liệu không có sự thay đổi đáng kể sau khi xử lý giá trị NaN.*

### 4.2. Power Bi



**Hình 4.1.** Biểu đồ tổng quan

A screenshot of a data presentation

Description automatically generated

**Hình 4.2.** Biểu đồ (tại VendorID = 1)

A screenshot of a computer

Description automatically generated

**Hình 4.3.** Biểu đồ (tại VendorID = 2)

\*Nhận xét chung:

- Dựa vào ba biểu đồ trên, có thể thấy rằng tổng số lượng tiền thu được (total\_amount) thông qua các chuyến đi taxi của VendorID 2 cao hơn VendorID1, cụ thể lần lượt chiến 67.26% và 32.74%.

- Ngoài ra, số lượng tiền tip của VendorID 2 cũng cao hơn đáng kể so với VendorID 1, cụ thể lần lượt bằng 8.08M và 3.63M.

- Đối với Vendor 2, số lượng hành khách và khoảng cách của chuyến đi tỷ lệ thuận với nhau, điều này được giải thích bởi thực tế khi đi xe taxi, quãng đường càng xa hành khách có xu hướng đi với số lượng đông hơn. Tuy nhiên, đối với VendorID 1, số lượng hành khách biến động bất thường hơn hẳn so với VendorID 2.

- Tổng số tiền phải trả cho tất cả các trạm thu phí trong chuyến đi của VendorID 1 chiếm phần lớn trong tổng tolls\_amount, hơn hẳn so với VendorID 2, đây cũng là một trong những yếu tố làm total\_amount của VendorID 1 thu lại ít hơn so với VendorID 2.

- Về phương thức thanh toán, đối với VendorID 1, khách hàng có xu hướng thanh toán bằng cả 4 phương thức, trong đó chủ yếu vẫn là phương thức 1 – thanh toán bằng tiền mặt. Đối với VendorID 2, khách hàng thanh toán chỉ qua 3 phương thức (phương thức 1,2,3). Tuy nhiên, nhìn chung, phương thức 1 vẫn là phương thức phổ biến và được khách hàng sử dụng nhiều nhất.