TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI TP. HỒ CHÍ MINH

Viện công nghệ thông tin và Điện, điện tử

------o0o------

UTH

UNIVERSITY
OF TRANSPORT
HOCHIMINH CITY

BÁO CÁO ĐỒ ÁN
Khoa Học Dữ Liệu

Đề tài: Ứng dụng học máy trong phân cụm các chuyến xe khách tại theo mức giá và mức độ hài lòng

Giảng viên : Ts. Trần Thế Vinh
Mã học phần : 010112710701
Nhóm : 09
SL thành viên : 06

TP.Hồ Chí Minh 2025

Đóng góp của các thành viên			
STT	**Tên thành viên**	**MSSV**	**Các đóng góp**
1	Lê Văn An	056205001827	- Thiết kế và xây dựng pipeline crawling dữ liệu - Thực hiện làm sạch dữ liệu - Load dữ liệu vào cơ sở dữ liệu - Hỗ trợ trực quan hóa dữ liệu và thiết kế dashboard trên PowerBI - Phối hợp xây dựng mô hình phân cụm - Tạo và triển khai ứng dụng Streamlit
2	Phạm Gia Bảo	093205005065	- Viết báo cáo chương 1 - 4 - Hỗ trợ cào dữ liệu - Chỉnh bố cục bài báo cáo - Chỉnh lại nội dung chương 8
3	Dương Hưng	086205003910	- Thiết kế slide, tạo slide - Hỗ trợ cào dữ liệu - Hỗ trợ chỉnh sửa báo cáo
4	Lê Minh Duy Khang	060205001162	- Viết báo cáo chương 5 - 9 - Hỗ trợ cào dữ liệu
5	Trần Ngọc Lê Hưng	095205002614	- Hỗ trợ cào dữ liệu - Hỗ trợ làm sạch dữ liệu - Hỗ trợ chỉnh sửa, xây dựng mô hình phân cụm - Hỗ trợ thiết kế dashboard PowerBI
6	Trần Tô Khắc Huy	082205009536	- Hỗ trợ cào dữ liệu - Trực quan hóa dữ liệu và thiết kế dashboard trên PowerBI

MỤC LỤC

Link Repo: Link

Chương 1: Tổng Quan Đề Tài

1.1. Lý do chọn đề tài.

Sự phát triển nhanh chóng của các nền tảng đặt vé trực tuyến đã làm thay đổi đáng kể cách người dùng lựa chọn dịch vụ vận tải hành khách, như là thông tin về giá vé, chất lượng dịch vụ, đúng giờ và đặc biệt đánh giá từ khách hàng đã trở thành những yếu tố quan trọng trong quá trình đưa ra quyết định. Tuy nhiên, sự đa dạng của các nhà xe, sự biến động giá cả của ngày nay và mức độ nhu cầu theo từng dạng khách hàng đã làm cho dữ liệu ngày càng phức tạp và mơ hồ.

Trong bối cảnh đó, các kỹ thuật học máy không giám sát mang lại một hướng tiếp cận phù hợp, cho phép tìm hiểu các cấu trúc tiềm ẩn trong dữ liệu mà không cần nhãn. Việc phân cụm các chuyến xe dựa trên giá vé và đánh giá từ khách hàng có thể đánh giá dịch vụ của chuyến xe, từ nhóm giá rẻ đến chất lượng cao, hoặc những nhóm dịch vụ có mức giá không tương xứng với chất lượng thực tế. Những thông tin này không chỉ hữu ích cho người dùng trong việc lựa chọn dịch vụ phù hợp mà còn hỗ trợ doanh nghiệp vận tải tối ưu chiến lược giá, cải thiện chất lượng và nhận diện bất thường trong vận hành.

Ngoài ra, dữ liệu của thị trường vận tải hành khách hiện nay có đặc điểm phân tán, khó hệ thống hóa và thường được trình bày dưới dạng danh sách rời rạc trên nhiều nền tảng khác nhau. Việc thu thập, chuẩn hoá và phân tích dữ liệu theo phương pháp khoa học giúp hình thành một góc nhìn tổng thể hơn về thị trường, từ đó góp phần nâng cao tính minh bạch và khả năng so sánh giữa các dịch vụ.

Từ những lý do đã nêu trên, việc triển khai một nghiên cứu ứng dụng học máy không giám sát để phân cụm các chuyến xe khách dựa trên giá vé và đánh giá khách hàng được xem là cần thiết. Đề tài góp phần giải quyết khoảng trống trong phân tích dữ liệu vận tải, mang lại giá trị cả về mặt kỹ thuật lẫn thực tiễn, đồng thời mở ra cơ hội ứng dụng rộng hơn cho các hệ thống gợi ý, hỗ trợ ra quyết định và tối ưu hoá dịch vụ trong lĩnh vực giao thông vận tải.

1.2. Mục tiêu đề tài

Mục tiêu của đề tài là xây dựng một quy trình phân tích dữ liệu hoàn chỉnh nhằm khám phá cấu trúc tiềm ẩn trong tập dữ liệu các chuyến xe khách thông qua việc ứng dụng các kỹ thuật học máy không giám sát. Trọng tâm của nghiên cứu là thiết lập mô hình phân cụm có khả năng phân loại các chuyến xe thành những nhóm dịch vụ mang tính chất đặc trưng, dựa trên hai nhóm thông tin quan trọng gồm giá vé và mức đánh giá từ khách hàng.

Để đạt được mục tiêu tổng quát này, đề tài hướng đến việc thu thập và chuẩn hoá dữ liệu, đảm bảo tính nhất quán và khả năng sử dụng cho mô hình học máy. Quá trình phân tích khám phá dữ liệu được thực hiện nhằm nhận diện các mối quan hệ giữa giá vé, mức đánh giá và các

đặc tính phụ trợ khác, qua đó làm rõ sự biến thiên và các xu hướng trong hành vi định giá và chất lượng của các hãng xe.

Trên cơ sở đó, mô hình phân cụm được xây dựng nhằm tìm ra những nhóm chuyến xe có tính tương đồng cao, phản ánh những phân khúc dịch vụ khác nhau trong thị trường vận tải hành khách. Các cụm được tạo ra không chỉ mang ý nghĩa về mặt thống kê mà còn có giá trị diễn giải, giúp nhận diện các nhóm dịch vụ có giá thấp nhưng đánh giá tốt, các nhóm giá cao tương ứng chất lượng cao, hoặc những trường hợp mất cân đối giữa giá và chất lượng.

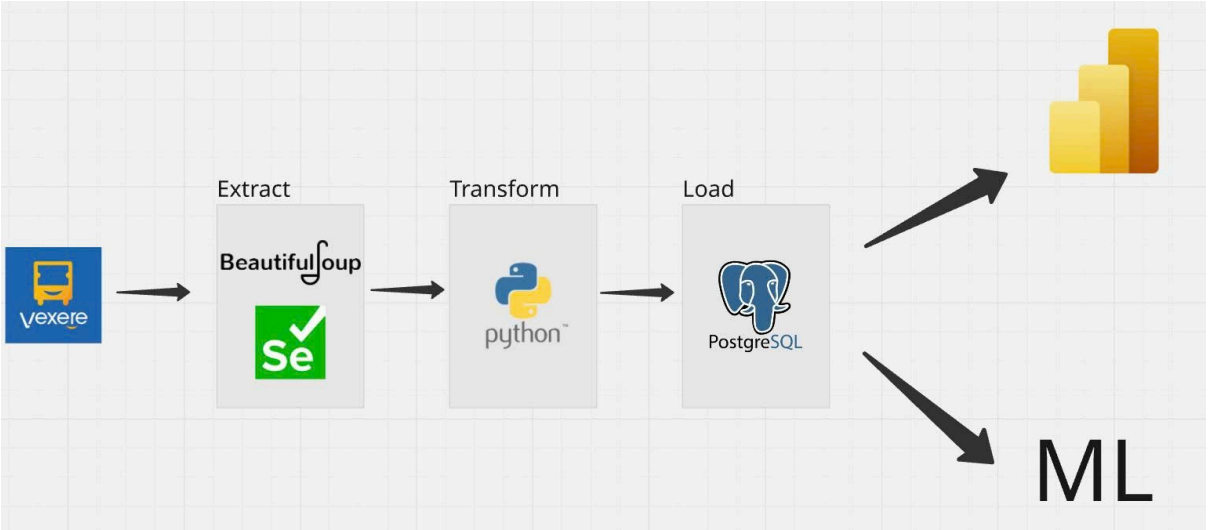
Cuối cùng, đề tài hướng đến việc đánh giá chất lượng mô hình bằng các tiêu chí đo lường phù hợp, đồng thời thảo luận về khả năng ứng dụng kết quả phân cụm trong hỗ trợ ra quyết định cho hành khách, doanh nghiệp vận tải và các hệ thống gợi ý trong tương lai.

1.3. Đối tượng và phạm vi nghiên cứu của đề tài

Đối tượng nghiên cứu của đề tài là tập dữ liệu mô tả các chuyến xe khách được cung cấp thông qua nền tảng đặt vé trực tuyến(vexere). Dữ liệu bao gồm các thuộc tính phản ánh đặc điểm dịch vụ, trong đó trọng tâm là giá vé và mức đánh giá từ khách hàng. Hai yếu tố này được xem là đại diện cho cấu trúc định giá và chất lượng cảm nhận của người sử dụng, đồng thời mang tính biến thiên lớn giữa các hãng xe và thời điểm đặt vé. Ngoài các biến cốt lõi, dữ liệu có thể bao gồm những thông tin bổ trợ khác như điểm đón, điểm trả, thời gian xuất phát, loại xe hoặc thời lượng chuyến đi, giúp mô hình phân cụm có thêm bối cảnh để mô tả sự khác biệt giữa các nhóm dịch vụ.

Phạm vi nghiên cứu tập trung vào việc thu thập, xử lý và phân tích dữ liệu dưới góc độ học máy không giám sát nhằm nhận diện các nhóm chuyến xe có đặc điểm tương đồng. Đề tài không xem xét yếu tố thời gian theo nghĩa xây dựng mô hình dự báo giá vé hoặc dự đoán mức đánh giá trong tương lai, mà chỉ khai thác mối quan hệ hiện hữu giữa các biến trong tập dữ liệu đã thu thập. Đồng thời, nghiên cứu giới hạn ở việc phân tích dữ liệu nền tảng nhà xe khách, mà không mở rộng sang toàn bộ thị trường vận tải hành khách.

Phạm vi kỹ thuật của đề tài tập trung vào các phương pháp tiền xử lý dữ liệu, phân tích khám phá dữ liệu và mô hình phân cụm, chủ yếu sử dụng các thuật toán phổ biến như K-Means, kết hợp với các phương pháp đánh giá cụm như Silhouette Score để xác định số lượng cụm hợp lý.



1.4. Phương pháp nghiên cứu

1.4.1. Phương pháp thu thập dữ liệu

Dữ liệu của bài nghiên cứu được thu thập từ trang web "vexere" với link:"Vexere.com"

1.4.2. Phương pháp phân tích và xử lý dữ liệu

Dữ liệu thô sau khi thu thập được đưa vào giai đoạn xử lý nhằm chuẩn hóa cấu trúc và đảm bảo tính toàn vẹn. Giai đoạn này bao gồm kiểm tra giá trị khuyết, loại bỏ bản ghi trùng lặp, chuẩn hóa kiểu dữ liệu và lọc bỏ các ngoại lệ không phù hợp.

Tiếp theo, dữ liệu đã xử lý được sử dụng cho phân tích khám phá (Exploratory Data Analysis – EDA). EDA được triển khai nhằm nhận diện các đặc điểm thống kê quan trọng, phân phối của các biến, mối tương quan giữa giá vé, đánh giá và các yếu tố vận hành.

Với dữ liệu đã chuẩn bị, mô hình học máy không giám sát được xây dựng chủ yếu dựa trên thuật toán K-Means, kết hợp thử nghiệm thêm các thuật toán phân cụm thay thế như DBSCAN hoặc Agglomerative để đánh giá độ ổn định.

Cuối cùng, dữ liệu và kết quả mô hình được lưu trữ trong hệ quản trị cơ sở dữ liệu PostgreSQL và được trực quan hóa bằng Microsoft Power BI để hỗ trợ việc diễn giải và trình bày các insight từ mô hình.

1.5. Ý nghĩa đề tài

Đề tài mang ý nghĩa ở cả hai phương diện: giá trị khoa học trong nghiên cứu học máy không giám sát và giá trị thực tiễn đối với lĩnh vực vận tải hành khách. Về mặt khoa học, việc áp dụng các thuật toán phân cụm vào dữ liệu dịch vụ xe khách cho phép kiểm chứng khả năng của mô hình trong việc nhận diện các cấu trúc tiềm ẩn mà không cần nhãn gán trước. Dữ liệu về giá vé, đánh giá người dùng và đặc điểm chuyến xe có bản chất đa chiều và biến động theo thời gian, do đó trở thành môi trường phù hợp để đánh giá mức độ ổn định của các

thuật toán phân cụm cũng như hiệu quả của các kỹ thuật tiền xử lý dữ liệu. Bên cạnh đó, quy trình từ thu thập, làm sạch, phân tích đến mô hình hóa giúp củng cố phương pháp luận trong các bài toán khai phá dữ liệu mang tính dịch vụ.

Ở góc độ ứng dụng, kết quả phân cụm mang lại khả năng phân loại tự động các nhóm chuyến xe có đặc điểm tương đồng về giá vé, chất lượng dịch vụ và trải nghiệm khách hàng. Điều này hỗ trợ các nhà vận hành trong việc định vị phân khúc thị trường, điều chỉnh chiến lược giá hoặc nâng cấp dịch vụ một cách chủ động. Đồng thời, người dùng cuối có thể được hưởng lợi khi có thêm công cụ giúp nhận diện các nhóm nhà xe theo chất lượng và mức giá, từ đó đưa ra quyết định phù hợp hơn với nhu cầu cá nhân. Việc định lượng hóa phản hồi người dùng thông qua phân cụm còn góp phần nâng cao tính minh bạch trong dịch vụ vận tải, tạo cơ sở dữ liệu hỗ trợ việc giám sát chất lượng và cải thiện trải nghiệm hành khách.

1.6. Cấu trúc đề tài

Gồm có 10 chương:

Chương 1: Tổng quan
Chương 2: Thu thập dữ liệu
Chương 3: Xử lý dữ liệu
Chương 4: Cơ sở lý thuyết
Chương 5:Lưu trữ dữ liệu database
Chương 6: Trực quan hóa dữ liệu
Chương 7: Phân tích khám phá dữ liệu
Chương 8: Mô hình hóa
Chương 9: Triển khai ứng dụng
Chương 10: Kết luận và hướng phát triển

Chương 2: Thu Thập Dữ Liệu

2.1. Khảo sát website và cấu trúc dữ liệu

Trước khi thu thập dữ liệu ta cần thực hiện khảo sát hệ thống nguồn dữ liệu để xác định các thực thể cần thu, các trường thông tin khả dụng và các hạn chế kỹ thuật. Trong nghiên cứu này, nguồn dữ liệu chính là nền tảng cung cấp thông tin chuyến xe công cộng mà hàm crawl_vexere trong module crawl. Danh sách các tuyến cần thu thập được lưu trong routes.json và được đọc bởi main.py. Kết quả khảo sát phải trả lời các câu hỏi then chốt: cấu trúc DOM (selector cho phần tử chuyến, giá, rating, thời lượng, loại ghế, điểm đi/đến), cách biểu diễn giá (ký hiệu tiền tệ, phân cách phần nghìn/thập phân), dạng rating (số thực, chuỗi dạng "4.5/5" hay nhãn chữ) và hành vi phân trang hay tải động của trang. Những thông tin này quyết định kỹ thuật thu thập (HTTP API vs. trình duyệt headless), cách parse chuỗi và các bước tiền xử lý cần thiết. Việc khảo sát cũng cần kiểm tra robots.txt và điều khoản sử dụng để đảm bảo tuân thủ pháp lý trước khi tiến hành crawling.

2.2. Thiết kế chiến lược data scraping (Selenium)

Thiết kế scraping được triển khai theo nguyên tắc mô-đun hoá: tách riêng phần logic trích xuất (crawl_vexere) và phần điều phối (main.py). main.py chịu trách nhiệm đọc routes.json, thiết lập tham số thời gian (DAYSOFF) và lặp qua từng cặp tuyến, trong khi crawl_vexere thực hiện tương tác với trang thông qua Selenium để xử lý DOM động và các hành vi tải bằng JavaScript. Lý do chọn Selenium là để xử lý các trang có rendering phía client và để mô phỏng hành vi người dùng khi cần (click, scroll, load lazy). Chiến lược cụ thể bao gồm chạy Chrome headless, khai báo implicit/explicit wait để chờ phần tử tải, áp dụng throttle giữa các request (main.py đặt time.sleep(8) giữa hai lần crawl liên tiếp) và dự trù cơ chế retry/backoff tại tầng crawl để xử lý lỗi kết nối hoặc timeout. Thiết kế còn bao gồm logging chi tiết trong module crawl để dễ truy nguyên lỗi và khả năng giới hạn phạm vi thu thập (max pages, days) nhằm kiểm soát khối lượng dữ liệu thu được.

2.3. Thu thập dữ liệu thô

Quy trình thu thập triển khai tuần tự theo từng tuyến: đối với mỗi cặp start_city và dest_city lấy từ routes.json, hàm crawl_vexere được gọi với tham số days tương ứng (do biến DAYSOFF quy định), trả về một DataFrame con chứa các bản ghi chuyến cho tuyến đó. main.py gom các DataFrame con này bằng pandas.concat thành một DataFrame toàn cục và lưu dưới dạng ./data/raw/{YYYY_MM_DD}_raw.csv, trong đó tên file sử dụng ngày mục tiêu để tạo snapshot theo thời gian. Lưu dữ liệu dưới dạng CSV ở bước thô có ưu điểm tương thích cao với nhiều công cụ phân tích và thuận tiện cho việc kiểm tra thủ công, đồng thời cho phép lưu giữ nguyên dạng trích xuất ban đầu để truy nguyên. Trong quá trình thu thập, cần chấp nhận rằng một số bản ghi có thể thiếu trường, có giá trị không chuẩn hoặc bị trùng do nhiều nguồn hiển thị cùng một chuyến; những vấn đề này sẽ được giải quyết ở bước xử lý.

2.4. Lưu trữ vào raw

Dữ liệu thô được tổ chức trong cây thư mục data/raw để phân tách rõ ràng giữa dữ liệu nguồn và dữ liệu đã xử lý. Việc ghi lại snapshot theo ngày cho phép lặp lại toàn bộ quy trình xử lý trên cùng một dữ liệu đầu vào, phục hồi khi cần và so sánh kết quả giữa các lần chạy. Trong môi trường thực nghiệm, cần bổ sung meta-data kèm theo file raw bao gồm: thời điểm thu thập (timestamp), tham số thu thập (DAYSOFF, danh sách tuyến), phiên bản mã nguồn (commit hash), và log tóm tắt số bản ghi thu được cùng các lỗi lớn nếu có. Việc lưu meta-data này hỗ trợ reproducibility và audit trail, đồng thời tạo điều kiện cho tái thực nghiệm khi cần kiểm chứng kết luận phân tích.

2.5. Các vấn đề phát sinh khi data scraping và cách xử lý

Trong thực tế, thao tác scraping thường gặp một số vấn đề chính: thay đổi cấu trúc HTML làm invalid selector, dữ liệu có nhiều dạng biểu diễn (price có ký hiệu, rating ở dạng chữ), giới hạn tần suất (rate limiting) hoặc chặn bot, lỗi kết nối, thời gian chờ,và dữ liệu bị load động dẫn đến mất bản ghi khi parse quá sớm. Cách xử lý bao gồm: (1) tách module crawl để dễ cập nhật selector và thêm unit-test cho các hàm parsing; (2) áp dụng cơ chế retry với backoff và kiểm tra điều kiện độ bền (stability) của phần tử DOM trước khi trích xuất; (3) giới hạn tốc độ truy vấn bằng throttle và random jitter, đồng thời tôn trọng robots.txt và chính sách của trang; (4) lưu log chi tiết và file raw chứa cả các bản ghi lỗi để có thể tái thu thập có chọn lọc; (5) nếu chịu tác động lớn từ chặn, xem xét sử dụng API chính thức hoặc hợp tác với nhà cung cấp dữ liệu. Những biện pháp này giảm thiểu rủi ro mất mát dữ liệu và đảm bảo rằng tập dữ liệu thô phản ánh trung thực nguồn.

Chương 3: Xử Lý Dữ Liệu

3.1. Đọc file raw

Bước đầu của pipeline xử lý là nạp file raw vào môi trường phân tích. main.py sử dụng pandas.read_csv để đọc file ./data/raw/{file_name}_raw.csv vào DataFrame. Mục tiêu của bước đọc không chỉ là chuyển dữ liệu sang dạng làm việc mà còn thu thập các thống kê ban đầu: số bản ghi, cấu trúc cột, kiểu dữ liệu sơ bộ và tỉ lệ giá trị thiếu. Việc đọc nên kèm theo việc kiểm tra encoding, xử lý các dòng corrupt và ghi log các hàng không parse được. Để đảm bảo reproducibility, cần ghi lại đường dẫn file raw và timestamp đọc dữ liệu trong meta-data.

3.2. Kiểm tra dữ liệu: missing value, duplicate, data types

Kiểm tra chất lượng dữ liệu gồm ba thành phần chính: phát hiện giá trị thiếu, phát hiện bản ghi trùng và xác thực kiểu dữ liệu. Phân tích missing cần báo cáo tỉ lệ missing theo cột và các mẫu liên quan (ví dụ những bản ghi thiếu rating tập trung ở nhà xe X hay tuyến Y), bởi vì nguyên nhân missing có ý nghĩa khác nhau và ảnh hưởng đến chiến lược xử lý. Phát hiện duplicate thực hiện trên tập các cột định danh như (departure_time, start_city, dest_city, operator) để loại bỏ lặp do crawl trùng; số bản ghi trước và sau khi drop_duplicates là chỉ số kiểm thử quan trọng. Kiểm tra kiểu dữ liệu nhằm đảm bảo các trường số (price, duration_minutes, rating) ở định dạng số học; các trường chuỗi được chuẩn hoá. Tài liệu kết quả kiểm tra cần trình bày số liệu tóm tắt: n_in (số bản ghi ban đầu), n_out (sau loại duplicate), tỉ lệ missing cho từng cột và danh sách các bản ghi bị lỗi parse.

3.3. Chuẩn hoá các trường

Chuẩn hoá định dạng là bước cốt lõi để đảm bảo tính nhất quán của các biến đầu vào cho phân tích. Với price, quá trình chuẩn hoá bao gồm loại bỏ ký tự tiền tệ, chuẩn hóa phân cách phần nghìn và decimal, và chuyển sang kiểu float; cần lưu ý quy tắc địa phương về dấu phẩy/dấu chấm. Với duration, cần chuyển các biểu diễn chuỗi như "3h 20m", "200 phút" thành một đơn vị chuẩn (phút) bằng hàm parse_duration. Với rating, nếu dữ liệu có dạng phân số "x/5" thì trích x; nếu là nhãn chữ thì áp dụng ánh xạ đến thang số đã định (ví dụ “Excellent” → 5.0). Các trường phân loại như seat_type nên được chuẩn hoá bằng chuẩn hóa văn bản (lowercase, trim) và gom các nhãn tương đương. Toàn bộ quy tắc chuẩn hoá cần được ghi lại trong tài liệu (processing spec) để đảm bảo tính minh bạch và khả năng tái tạo.

3.4 Tạo thêm các feature hữu ích

Việc này nhắm đến việc tạo các biến có ý nghĩa trong không gian phân cụm theo giá và mức độ hài lòng. Một biến mang tính thông tin cao là price_per_minute = price / duration_minutes, phản ánh mật độ giá trên thời lượng. Ngoài ra, có thể tạo các biến chỉ báo cho khoảng thời gian khởi hành (buổi sáng, chiều tối), loại ghế được mã hoá (one-hot hoặc

label encoding), và biến tương tác giữa nhà xe và giá trung bình tuyến để bắt tín hiệu khác biệt nhà cung cấp. Khi tạo feature, cần đảm bảo không tạo dữ liệu rò rỉ và giữ lại bản ghi nguồn với cờ giải thích cho từng biến mới. Mỗi biến mới cần được mô tả rõ ý nghĩa, phạm vi hợp lệ và cách xử lý giá trị thiếu.

3.5. Lọc nhiễu và xử lý ngoại lệ

Phát hiện ngoại lệ cần được thực hiện cho các biến liên tục. Phương pháp IQR (interquartile range) được áp dụng phổ biến: với IQR = Q3 − Q1, một quan sát được coi là ngoại lệ nếu giá trị nằm ngoài khoảng [Q1 − k·IQR, Q3 + k·IQR], với k thường là 1.5 hoặc giá trị điều chỉnh tùy mức độ muốn thận trọng. Công thức này cho phép đánh dấu ngoại lệ mà không giả định phân phối chuẩn. Ngoài ra, z-score cũng là lựa chọn nếu dữ liệu gần chuẩn phân phối; quan sát có |z| > t (ví dụ 3) có thể được coi là ngoại lệ. Sau khi phát hiện, ngoại lệ có thể được xử lý bằng cách gắn cờ, winsorization hoặc loại bỏ nếu xác định là lỗi nhập liệu. Tài liệu xử lý phải nêu rõ tiêu chí và số lượng bản ghi bị ảnh hưởng để người đọc đánh giá tác động lên kết quả phân tích.

3.6. Xuất data processed

Kết quả xử lý được lưu dưới dạng ./data/processed/{YYYY_MM_DD}_cleaned.csv như trong main.py; tùy nhu cầu hiệu năng, có thể xuất thêm định dạng columnar như Parquet để tối ưu lưu trữ và truy xuất. Trước khi xuất, cần thực hiện các kiểm thử cuối: đảm bảo ràng buộc nghiệp vụ (price ≥ 0, duration_minutes > 0, rating trong khoảng [0,5]), so sánh số bản ghi giữa raw và processed (với giải thích cho chênh lệch do loại bỏ lỗi), và lưu báo cáo tóm tắt chất lượng dữ liệu cùng file processed. Sau khi lưu, main.py đọc lại file processed và gọi DatabaseManager cùng hàm insert_trips_from_dataframe để nạp dữ liệu vào hệ quản trị quan hệ; trong báo cáo cần mô tả schema đích, các ràng buộc (NOT NULL, UNIQUE nếu có) và chỉ số kiểm tra nạp thành công/không thành công.

Chương 4: Cơ Sở Lý Thuyết

4.1. Giới thiệu Học máy không giám sát (Unsupervised Learning).

Trong thực tiễn khai phá dữ liệu, đặc biệt khi xử lý các tập thông tin thu thập qua web scraping, học máy không giám sát đóng vai trò quan trọng ở ba phương diện. Thứ nhất, phân cụm hỗ trợ nhận diện các nhóm mẫu có hành vi tương đồng, chẳng hạn phân nhóm tuyến xe hoặc mô hình giá vé có đặc trưng giống nhau để dễ dàng thiết kế chiến lược phân tích tiếp theo. Thứ hai, giảm chiều giúp mô hình hoá dữ liệu dạng bảng có nhiều thuộc tính, làm giảm tác động của nhiễu và cải thiện diễn giải bằng các không gian biểu diễn thấp chiều như PCA. Thứ ba, phát hiện bất thường hỗ trợ loại bỏ các mẫu lệch chuẩn phát sinh từ lỗi thu thập hoặc bất nhất hệ thống, đảm bảo chất lượng dữ liệu trước các bước mô hình hóa tiếp theo.

4.2. Khái niệm phân cụm (Clustering).

Ý nghĩa:

Phân cụm mang nhiều ý nghĩa trong nghiên cứu và ứng dụng thực tiễn. Thứ nhất, nó cung cấp phương tiện tổ chức dữ liệu phức tạp, từ đó nhận diện các mô hình, xu hướng hoặc hành vi ẩn trong tập dữ liệu. Thứ hai, phân cụm hỗ trợ việc giảm chiều dữ liệu và trích xuất các đặc trưng đại diện, qua đó cải thiện khả năng trực quan hóa và hiệu suất của các thuật toán học máy khác. Thứ ba, nó giúp phát hiện các ngoại lệ hoặc giá trị bất thường, từ đó đảm bảo tính toàn vẹn và độ tin cậy của dữ liệu trong các bước phân tích tiếp theo. Trong bối cảnh nghiên cứu chuyến xe khách, phân cụm giúp phân loại các chuyến đi theo mức giá, đánh giá từ khách hàng hoặc loại ghế, hỗ trợ việc ra quyết định quản lý và tối ưu hóa dịch vụ.

Ứng dụng:

Phân cụm có ứng dụng rộng rãi trong nhiều lĩnh vực. Trong thương mại và dịch vụ, nó được dùng để phân nhóm khách hàng theo hành vi mua sắm hoặc sở thích, từ đó xây dựng chiến lược marketing cá nhân hóa. Trong lĩnh vực giao thông và logistics, phân cụm giúp nhận diện các tuyến đường hoặc nhóm phương tiện có đặc điểm vận hành tương đồng, hỗ trợ phân bổ tài nguyên hiệu quả. Trong y học, phân cụm giúp phát hiện các loại bệnh lý hoặc nhóm bệnh nhân dựa trên đặc trưng sinh học, dữ liệu lâm sàng hoặc xét nghiệm. Trong đề tài về phân cụm các chuyến xe khách, ứng dụng cụ thể bao gồm phân loại tuyến xe, đánh giá mức độ phổ biến của các loại ghế, nhóm các chuyến đi theo mức giá và rating, từ đó tạo cơ sở dữ liệu cho phân tích hành vi khách hàng, gợi ý dịch vụ và tối ưu hóa vận hành hệ thống.

4.3. Thuật toán K-Means.

K-Means là thuật toán phân cụm dựa trên trung tâm (center-based clustering), thuộc học máy không giám sát. Tư tưởng cốt lõi của thuật toán là chia tập dữ liệu $X=\{x1,x2,\ldots,xn\}$ thành k cụm sao cho các đối tượng trong cùng một cụm có độ tương đồng cao, còn các đối

tượng thuộc các cụm khác nhau có sự khác biệt lớn. Khoảng cách thường được đo bằng khoảng cách Euclid hoặc các phép đo khác phù hợp với bản chất dữ liệu. Thuật toán thực hiện lặp đi lặp lại hai bước cơ bản: gán mỗi điểm dữ liệu vào cụm gần nhất và cập nhật tâm cụm dựa trên các điểm thuộc cụm đó, cho đến khi hội tụ.

4.4. Đánh giá mô hình phân cụm.

Silhouette Score:

Silhouette Score là một chỉ số nội tại phổ biến dùng để đánh giá mức độ phù hợp của mỗi điểm dữ liệu với cụm mà nó thuộc về. Với mỗi điểm xi định nghĩa:

a(i): khoảng cách trung bình từ xi đến tất cả các điểm khác trong cùng cụm.
b(i): khoảng cách trung bình từ xi đến các điểm trong cụm gần nhất không chứa xi.
Silhouette Score của xi được tính bằng:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Giá trị s(i) nằm trong khoảng [−1,1] Giá trị gần 1 cho thấy điểm được phân cụm tốt, gần 0 biểu thị điểm nằm ở ranh giới giữa các cụm, còn giá trị âm chỉ ra phân cụm không hợp lý. Silhouette Score trung bình trên toàn bộ tập dữ liệu cung cấp một thước đo tổng quan về chất lượng phân cụm. Trong đề tài về phân cụm chuyến xe khách, chỉ số này giúp đánh giá mức độ phân biệt giữa các nhóm chuyến đi theo giá vé, rating, loại ghế và số ghế trống.

Elbow Method:

Elbow Method là một phương pháp trực quan để lựa chọn số cụm kkk tối ưu trong thuật toán K-Means. Nguyên lý của phương pháp dựa trên việc tính toán tổng bình phương sai số (SSE) cho nhiều giá trị k khác nhau:

$$SSE(k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

Khi k tăng, SSE giảm dần, nhưng tốc độ giảm sẽ chậm lại sau một số cụm nhất định. Điểm mà tại đó tốc độ giảm đột ngột "gập khúc" (elbow) được chọn làm giá trị k tối ưu, bởi nó cân bằng giữa việc giảm SSE và tránh số lượng cụm quá nhiều gây phức tạp hoặc overfitting.

Trong thực tiễn, khi phân cụm các chuyến xe khách, Elbow Method giúp xác định số nhóm chuyến đi hợp lý để mô hình phân loại các tuyến xe hoặc phân khúc khách hàng mà không tạo ra quá nhiều cụm nhỏ không mang ý nghĩa thực tiễn.

4.5. Các kỹ thuật tiền xử lý dữ liệu cho phân cụm.

Scaling:

Scaling hay chuẩn hóa dữ liệu nhằm đưa các biến về cùng một thang đo, tránh tình trạng biến có giá trị lớn chi phối khoảng cách và làm lệch kết quả phân cụm. Một số phương pháp phổ biến bao gồm:

StandardScaler: Chuẩn hóa dữ liệu về phân phối chuẩn với trung bình bằng 0 và độ lệch chuẩn bằng 1. Phương pháp này phù hợp với các biến có phân phối gần Gaussian, giúp duy trì tính biến thiên tương đối giữa các giá trị.

$$x' = \frac{x - \mu}{\sigma}$$

RobustScaler: Chuẩn hóa dựa trên trung vị và khoảng tứ phân vị, giúp giảm tác động của các giá trị ngoại lệ. Phương pháp này đặc biệt hiệu quả khi dữ liệu có nhiều outlier, ví dụ như giá vé bất thường hoặc số ghế trống không hợp lý.

$$x' = \frac{x - \text{median}(x)}{IQR(x)}$$

Encoding:

Đối với các biến phân loại, chẳng hạn như loại ghế (ngồi, giường nằm, giường đôi), cần chuyển đổi sang dạng số để thuật toán phân cụm có thể xử lý. OneHotEncoder là phương pháp phổ biến, tạo ra các biến nhị phân thể hiện sự xuất hiện của từng giá trị. Ví dụ, nếu biến loại ghế có ba loại, OneHotEncoder tạo ba cột nhị phân, mỗi cột đánh dấu một loại ghế. Phương pháp này duy trì tính phân biệt giữa các giá trị phân loại mà không tạo ra thứ tự giả định giữa các nhãn.

Xử lý outlier:

Ngoài việc chuẩn hóa, việc xử lý các giá trị ngoại lệ (outlier) cũng là bước thiết yếu. Các thuật toán phân cụm dựa trên khoảng cách như K-Means nhạy cảm với outlier, vì các điểm bất thường có thể kéo centroid lệch và làm biến dạng hình dạng cụm. Các kỹ thuật xử lý bao gồm:

Loại bỏ các giá trị bất thường vượt quá ngưỡng định trước (ví dụ: giá vé quá cao hoặc số ghế trống âm).

Giới hạn giá trị về biên an toàn (capping) hoặc thay thế bằng giá trị trung bình/trung vị.

Sử dụng các phương pháp scaling chống outlier như RobustScaler để giảm ảnh hưởng của các giá trị ngoại lai.

4.6. Các khái niệm nền tảng khác.

Feature Engineering:

Feature engineering là quá trình tạo ra các đặc trưng mới từ dữ liệu gốc nhằm nâng cao khả năng mô hình hóa và phát hiện cấu trúc ẩn. Đối với dữ liệu chuyến xe khách, các feature mới có thể bao gồm: tỷ lệ số ghế trống trên tổng số ghế, thời lượng chuyến đi chuẩn hóa theo khoảng cách, hoặc nhóm giá vé theo phân vị. Việc tạo ra các đặc trưng hữu ích giúp tăng tính phân biệt giữa các điểm dữ liệu, từ đó làm rõ cấu trúc cụm và cải thiện chất lượng phân cụm. Feature engineering không chỉ giúp mô hình học máy nhận biết các mẫu phức tạp mà còn hỗ trợ trực quan hóa dữ liệu một cách hiệu quả.

Dimensionality Reduction (PCA)

Khi dữ liệu có nhiều thuộc tính (high-dimensional), việc phân cụm trực tiếp có thể gặp khó khăn do hiệu ứng "curse of dimensionality", khiến khoảng cách giữa các điểm trở nên đồng nhất và làm giảm khả năng phân tách cụm. Principal Component Analysis (PCA) là một kỹ thuật giảm chiều phổ biến, nhằm ánh xạ dữ liệu sang không gian có số chiều thấp hơn đồng thời giữ lại phần lớn phương sai. Công thức tổng quát:

$$Z = XW$$

trong đó X là ma trận dữ liệu chuẩn hóa, W là ma trận vector riêng tương ứng với các thành phần chính, và Z là dữ liệu chiếu xuống không gian mới. PCA giúp loại bỏ nhiễu và các feature ít thông tin, làm giảm tính toán và cải thiện hiệu quả của các thuật toán phân cụm như K-Means.

Robust Scaling:

Trong các phương pháp chuẩn hóa dữ liệu, RobustScaler là lựa chọn phù hợp khi dữ liệu chứa nhiều ngoại lệ — một đặc điểm thường gặp trong dữ liệu giá vé, số ghế trống hoặc rating từ người dùng. Khác với StandardScaler dựa trên trung bình và độ lệch chuẩn, RobustScaler chuẩn hóa dữ liệu dựa trên median và khoảng tứ phân vị (IQR), giúp giảm ảnh hưởng của các giá trị bất thường. Công thức tổng quát:

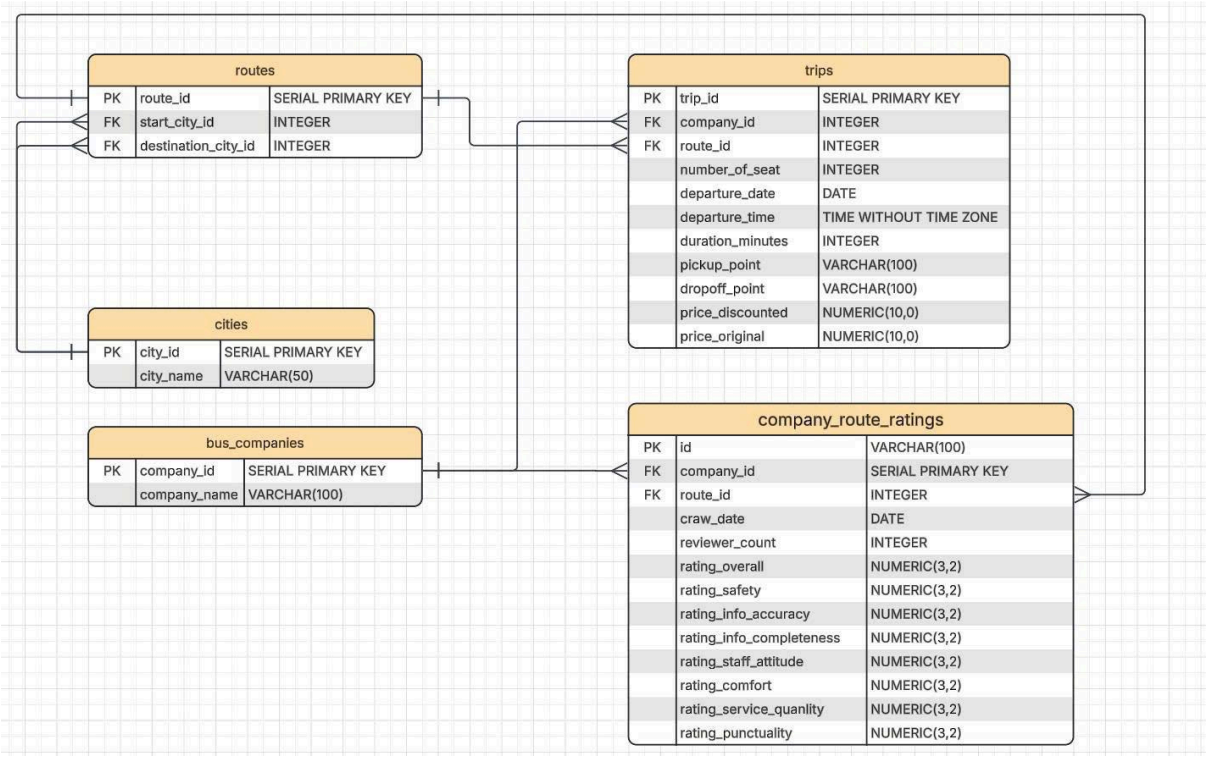
$$x' = \frac{x - \text{median}(x)}{IQR(x)}$$

Sử dụng RobustScaler đảm bảo rằng các giá trị cực đoan không gây sai lệch mạnh đến khoảng cách giữa các điểm dữ liệu, từ đó cải thiện đáng kể độ ổn định của các thuật toán phân cụm dựa trên metric như K-Means. Bản chất chống outlier của phương pháp này phù hợp với đặc thù của bộ dữ liệu trong nghiên cứu, nơi giá vé và lượt đánh giá thường có độ biến thiên lớn và xuất hiện những giá trị lệch chuẩn.

Chương 5: Lưu Trữ Dữ Liệu Database

5.1. Thiết kế cấu trúc bảng

Trong giai đoạn hậu xử lý dữ liệu, việc xây dựng một hệ thống lưu trữ có tổ chức đóng vai trò quan trọng nhằm đảm bảo khả năng truy vấn, phân tích và mở rộng trong tương lai. Cấu trúc bảng được thiết kế dựa trên nguyên tắc chuẩn hóa dữ liệu, đảm bảo mỗi thực thể được mô tả đầy đủ và không trùng lặp thông tin. Bảng dữ liệu chính mô tả các chuyến xe khách bao gồm các trường thông tin về mã chuyến xe, nhà xe, điểm xuất phát, điểm đến, giá vé, mức độ hài lòng của hành khách và các đặc trưng được tạo ra trong quá trình tiền xử lý để phục vụ mô hình phân cụm. Các thuộc tính dạng số được chuẩn hóa về cùng một thang đo nhằm đảm bảo tính nhất quán khi sử dụng trong phân tích, trong khi các trường dạng văn bản được mã hóa hoặc rút trích đặc trưng tùy theo yêu cầu của thuật toán. Toàn bộ cấu trúc bảng hướng đến mục tiêu vừa hỗ trợ phục vụ thuật toán máy học, vừa đảm bảo khả năng lưu trữ ổn định cho hoạt động báo cáo và truy vấn sau này.



5.2. Tạo kết nối database (PostgreSQL / MySQL)

Hệ thống quản lý cơ sở dữ liệu được triển khai dựa trên một trong hai nền tảng phổ biến là PostgreSQL hoặc MySQL, tùy thuộc vào môi trường triển khai và yêu cầu phần mềm. Kết nối từ môi trường xử lý dữ liệu sang cơ sở dữ liệu được thiết lập thông qua các thư viện phổ biến như SQLAlchemy hoặc psycopg2 đối với PostgreSQL, hay mysql-connector đối với MySQL. Việc cấu hình kết nối bao gồm khai báo thông tin máy chủ, tên cơ sở dữ liệu, tài khoản truy cập và các tham số bảo mật bổ sung nếu hệ thống yêu cầu. Sau khi kết nối được thiết lập thành công, môi trường phân tích có khả năng thực hiện truy vấn, ghi dữ liệu và

kiểm thử trực tiếp trên hệ thống lưu trữ, đảm bảo tính liền mạch giữa các giai đoạn trong pipeline xử lý dữ liệu.

5.3. Quy trình load data processed → DB

Dữ liệu sau khi được tiền xử lý và sinh các đặc trưng cần thiết cho mô hình phân cụm sẽ được đưa vào cơ sở dữ liệu thông qua quy trình nạp dữ liệu được xây dựng thống nhất. Quy trình này bao gồm bước chuyển đổi định dạng dữ liệu từ dataframe sang dạng bảng tương thích với SQL, tiếp theo là kiểm tra ràng buộc kiểu dữ liệu để đảm bảo dữ liệu không vi phạm các chuẩn mà hệ quản trị đã định nghĩa. Sau đó dữ liệu được ghi vào bảng đích thông qua thao tác ghi hàng loạt nhằm tối ưu hiệu năng xử lý. Đối với các hệ thống yêu cầu lịch sử dữ liệu hoặc cập nhật theo phiên, quy trình nạp dữ liệu cũng đồng thời kiểm tra trùng lặp và thực hiện cập nhật có điều kiện, bảo đảm rằng thông tin lưu trữ luôn phản ánh chính xác trạng thái mới nhất. Việc tổ chức quy trình một cách tự động giúp giảm thiểu lỗi thao tác thủ công và tăng tính ổn định trong suốt vòng đời dự án.

5.4. Kiểm thử dữ liệu sau khi load

Sau khi quá trình nạp dữ liệu hoàn tất, việc kiểm thử dữ liệu trong cơ sở dữ liệu là bước cuối cùng nhằm xác nhận tính toàn vẹn và mức độ chính xác của dữ liệu lưu trữ. Hoạt động kiểm thử bao gồm việc kiểm tra số lượng bản ghi trước và sau khi nạp để đảm bảo không có mất mát dữ liệu, đồng thời thực hiện các truy vấn kiểm tra đối với các thuộc tính quan trọng như giá vé, mức độ hài lòng và các đặc trưng đầu vào của mô hình phân cụm. Các quy tắc kiểm thử cũng bao gồm việc đánh giá sự phù hợp của kiểu dữ liệu, tính hợp lệ của giá trị ràng buộc và phát hiện các giá trị ngoại lai có thể phát sinh trong quá trình nạp. Khi dữ liệu vượt qua toàn bộ các bước kiểm thử, cơ sở dữ liệu được xem là sẵn sàng phục vụ cho quá trình phân tích và trực quan hóa tiếp theo, đồng thời trở thành nền tảng ổn định cho việc tích hợp với các ứng dụng khác trong hệ thống.

Chương 6: Trực Quan Hóa Dữ Liệu (Power BI Dashboard)

6.1. Import dữ liệu từ DB / CSV

Việc trực quan hóa dữ liệu trong Power BI bắt đầu bằng quá trình nhập dữ liệu từ cơ sở dữ liệu hoặc tệp CSV được tạo ra từ giai đoạn tiền xử lý. Khi sử dụng dữ liệu từ hệ quản trị cơ sở dữ liệu như PostgreSQL hoặc MySQL, Power BI thiết lập kết nối thông qua các trình điều khiển chuyên dụng, cho phép truy vấn trực tiếp dữ liệu đã được chuẩn hóa và lưu trữ ở chương trước. Trong trường hợp dữ liệu được cung cấp dưới dạng tệp CSV, Power BI thực hiện nhận diện loại dữ liệu, tự động gán kiểu trường và hỗ trợ người dùng bổ sung các thao tác chuyển đổi đơn giản trong Power Query Editor. Toàn bộ quá trình nhập liệu được thiết kế nhằm bảo đảm rằng dữ liệu được đưa vào hệ thống trực quan hóa ở trạng thái sạch, thống nhất và sẵn sàng cho các thao tác phân tích tiếp theo.

6.2. Xây dựng các visual:

Sau khi dữ liệu được nạp thành công, các biểu đồ trực quan được xây dựng nhằm mô tả các đặc trưng quan trọng của tập dữ liệu và hỗ trợ phân tích mô hình phân cụm. Các biểu đồ phân phối giá vé, mức độ hài lòng và số lượng chuyến xe theo tuyến giúp hình dung tổng quan đặc điểm dữ liệu. Đối với kết quả phân cụm, các biểu đồ scatter plot hai chiều và ba chiều đóng vai trò then chốt trong việc thể hiện cấu trúc nhóm, cho phép quan sát rõ ràng sự phân chia của các cụm theo giá vé và mức độ hài lòng, đồng thời hỗ trợ đánh giá tính hợp lý của kết quả mô hình. Các visual dạng bảng cho phép người dùng theo dõi chi tiết từng chuyến xe cùng nhóm cụm tương ứng, trong khi các biểu đồ tương tác giúp hệ thống trở nên trực quan và linh hoạt hơn, đặc biệt trong việc khám phá dữ liệu theo từng chiều thông tin cụ thể. Power BI cung cấp khả năng gắn bộ lọc động, cho phép người dùng tùy chỉnh không gian quan sát dữ liệu và tập trung vào các tuyến, nhà xe hoặc phân khúc cụ thể.

6.3. Insight rút ra từ dashboard

Việc tổng hợp thông tin từ dashboard mang lại nhiều insight quan trọng liên quan đến hoạt động vận tải hành khách. Sự phân bố của giá vé và mức độ hài lòng cho thấy sự tồn tại của những nhóm hành khách có hành vi và kỳ vọng khác nhau, từ đó giúp doanh nghiệp nhận diện được các phân khúc khách hàng đặc trưng. Các nhóm phân cụm thể hiện rõ sự khác biệt về mức giá và chất lượng dịch vụ, cho phép nhà xe đánh giá mức độ cạnh tranh giữa các tuyến cũng như hiệu quả định giá hiện tại. Một số cụm đặc biệt cho thấy sự mất cân đối giữa giá vé và mức độ hài lòng, từ đó gợi ý những khu vực cần cải thiện chất lượng dịch vụ hoặc điều chỉnh chiến lược giá để tăng mức độ hài lòng của hành khách. Bên cạnh đó, visual về phân bố theo tuyến và thời gian giúp xác định các khung giờ hoặc tuyến đường có nhu cầu cao, tạo điều kiện cho việc tối ưu tần suất vận hành. Những insight này trở thành nền tảng quan trọng giúp doanh nghiệp đưa ra quyết định dựa trên dữ liệu, đồng thời củng cố giá trị ứng dụng của mô hình phân cụm trong thực tiễn.


TOTAL TRIPS
Total Number ...
28.6...
Total Bus...
344
Average Ticket...
308.6...
4.28
Average Overall Ra...
Filter by Departure Date
11/12/2025
11/23/2025
Ngày giờ khởi hành
Trip Volume Trend by Day
Trip Volume...
50M
0M
Trip Count by Departure City
Trip Count
10K
0K
Sài Gòn
Hà Nội
Đà Nẵng
Departure City
Trip Count by Destination City
Destination City
Nha Trang - Khán...
Ninh Bình
Quảng Ninh
Thừa Thiên-Huế
0M
20M
40M
Trip Count
Route Price vs Rating Scatter
Avg Price, Avg Price by Route, Sum of trip_id, Avg Price, First route_display, Avg Rating and Total Trips by route_display
route_display
Đà Nẵng ...
Avg Pric...
0.5M
0.0M
Route Performance Summary
route_display
Giá trung bình
Xếp hạng trung bình theo tuyến đườ
Sài Gòn → Phú Yên
361,821.159 ₫
Sài Gòn → Ninh Thuận
297,677.258 ₫
Sài Gòn → Nha Trang - Khánh Hòa
336,516.713 ₫
Sài Gòn → Gia Lai
367,109.673 ₫
Total
308,612.230 ₫

Chương 7: Phân Tích Khám Phá Dữ Liệu (EDA)

7.1. Kiểm tra phân phối các biến

Quá trình phân tích khám phá dữ liệu được khởi đầu bằng việc kiểm tra phân phối của các biến quan trọng nhằm nhận diện đặc điểm tổng quan của tập dữ liệu. Các biến định lượng như giá vé, mức độ hài lòng và thời lượng chuyến đi thường được mô tả thông qua các biểu đồ histogram hoặc density plot, giúp làm rõ xu hướng trung tâm, mức độ phân tán và cấu trúc lệch của dữ liệu. Những quan sát này cung cấp thông tin quan trọng về sự tồn tại của các nhóm giá trị cực trị hoặc các phân phối không chuẩn, từ đó hỗ trợ việc lựa chọn kỹ thuật chuẩn hóa phù hợp cho giai đoạn xây dựng mô hình. Đối với các biến định tính như loại ghế hoặc phân loại tuyến, biểu đồ tần suất giúp mô tả mức độ phổ biến của từng nhóm và nhận diện sự mất cân đối trong phân bố dữ liệu.

7.2. Tương quan giữa giá vé – rating – loại ghế – thời lượng

Sau khi nắm được phân phối của từng biến, quá trình EDA tập trung vào việc phân tích mối quan hệ giữa các biến có khả năng ảnh hưởng lẫn nhau. Mối tương quan giữa giá vé và mức độ hài lòng được xem xét nhằm đánh giá liệu giá cao có thực sự đi kèm với chất lượng dịch vụ tốt hơn hay không. Phân tích tương quan tiếp tục được mở rộng đối với loại ghế và thời lượng chuyến đi, qua đó cho phép nhận diện các thành tố cấu thành chiến lược định giá của nhà xe. Các visual như scatter plot, boxplot và heatmap tương quan giúp làm rõ xu hướng tuyến tính hoặc phi tuyến giữa các biến; đồng thời chỉ ra các đặc điểm bất thường như sự tập trung của một số loại ghế trong các mức giá cố định hoặc mức độ hài lòng khác biệt rõ rệt giữa các tuyến có thời lượng dài. Những quan sát này đóng vai trò định hướng việc hình thành các đặc trưng đầu vào phù hợp cho mô hình phân cụm.

7.3. Phát hiện outlier

Outlier có thể xuất hiện từ lỗi nhập liệu, sự cố kỹ thuật hoặc các hành vi bất thường trong quá trình thu thập dữ liệu. Việc phát hiện các giá trị bất thường được thực hiện thông qua nhiều phương pháp khác nhau như quan sát phân phối, sử dụng boxplot, hoặc áp dụng các thước đo định lượng như z-score và IQR. Nhóm dữ liệu giá vé thường xuất hiện những giá trị cao bất thường khi có dịch vụ cao cấp hoặc các tuyến đặc biệt, trong khi mức độ hài lòng có thể xuất hiện điểm thấp đột ngột phản ánh sự cố dịch vụ. Đối với thời lượng hành trình, các chuyến kéo dài vượt mức thông thường có thể là tín hiệu của sự sai lệch trong xử lý dữ liệu hoặc điều kiện vận tải ngoại lệ. Outlier cần được phân tích thận trọng vì chúng có thể ảnh hưởng mạnh đến kết quả phân cụm; do đó, tùy vào bản chất giá trị, chúng có thể được loại bỏ, điều chỉnh hoặc giữ lại nếu mang ý nghĩa thực tiễn.

7.4. Chuẩn bị dữ liệu cho Modeling

Trước khi dữ liệu được đưa vào mô hình phân cụm, quá trình chuẩn bị bao gồm xử lý thiếu dữ liệu, chuẩn hóa thang đo và mã hóa các biến phân loại. Các giá trị thiếu trong dữ liệu

được xử lý thông qua các phương pháp nội suy hoặc loại bỏ có kiểm soát, tùy thuộc vào mức độ ảnh hưởng. Việc chuẩn hóa dữ liệu theo các phương pháp như MinMaxScaler hoặc StandardScaler giúp các biến số có thang đo đồng nhất, bảo đảm thuật toán phân cụm không bị chi phối bởi các biến có đơn vị lớn. Biến loại ghế, nếu mang tính phân loại, sẽ được chuyển đổi sang dạng số thông qua các phương pháp mã hóa thích hợp. Giai đoạn này cũng bao gồm việc tạo ra các đặc trưng bổ sung nếu cần thiết để làm rõ hơn cấu trúc dữ liệu và tăng khả năng phân tách giữa các nhóm.

7.5. Kỹ thuật đặc trưng (Feature Engineering)

Trước khi tiến hành phân cụm, nhóm nghiên cứu nhận thấy dữ liệu thô ban đầu chưa đủ để phản ánh toàn diện các khía cạnh của một chuyến xe. Do đó, dựa trên bước Khám phá dữ liệu (EDA) trước đó, chúng tôi đã tiến hành kỹ thuật đặc trưng (Feature Engineering) để xây dựng thêm các biến phát sinh quan trọng nhằm làm rõ hơn cấu trúc dữ liệu.

Các đặc trưng mới được thiết lập bao gồm:

- **Đặc trưng về Giá cả:**

real_price: Giá vé thực tế mà khách hàng phải trả sau khi đã trừ đi các khoản giảm giá (nếu có), phản ánh chính xác chi phí của người dùng.

price_per_minute: Giá vé chia cho thời gian di chuyển, dùng để đo lường chi phí trên đơn vị thời gian.

price_per_seat: Giá vé trên mỗi ghế, giúp chuẩn hóa giá giữa các loại xe có sức chứa khác nhau.

- **Đặc trưng về Chất lượng và Uy tín:**

service_score (Điểm chất lượng dịch vụ): Biến tổng hợp từ các đánh giá thành phần như: thái độ nhân viên, tiện nghi trên xe và chất lượng dịch vụ chung.

trust_score (Điểm độ tin cậy): Biến tổng hợp từ các yếu tố: mức độ an toàn, tính đúng giờ và độ chính xác của thông tin.

wilson_score: Điểm đánh giá tổng hợp đã được chuẩn hóa thống kê (Wilson Lower Bound), giúp cân bằng giữa điểm số trung bình và số lượng lượt đánh giá, tránh thiên vị các nhà xe có ít review.

- **Đặc trưng kết hợp:**

price_rating_ratio_stable (hoặc Fairness Index): Tỷ số giữa giá vé và chất lượng, đại diện cho mức độ "đáng tiền" của chuyến xe.

7.6. Lựa chọn features để phân cụm

Việc lựa chọn đặc trưng có ảnh hưởng quyết định đến chất lượng mô hình phân cụm (Clustering). Các biến được chọn cần thể hiện đầy đủ sự khác biệt giữa các chuyến xe và phản ánh được sự đa dạng trong trải nghiệm của hành khách.

Dựa trên danh sách các đặc trưng đã chuẩn bị ở mục 7.5, nhóm quyết định **chỉ giữ lại các feature liên quan trực tiếp đến hai khía cạnh cốt lõi: GIÁ CẢ và CHẤT LƯỢNG**. Cụ thể:

- Các đặc trưng định lượng như **giá vé (`real_price`)**, **mức độ hài lòng (`wilson_score`)** được giữ lại làm trọng tâm vì chúng mang tính định lượng rõ ràng và có khả năng phân tách cao.
- Các biến tổng hợp như **`trust_score`** và **`service_score`** được ưu tiên sử dụng thay vì các điểm đánh giá thành phần rời rạc để giảm nhiễu và giảm số chiều dữ liệu.
- Yếu tố loại ghế hay thời lượng hành trình, dù quan trọng, nhưng có thể đóng góp giá trị tốt hơn khi được mã hóa hợp lý hoặc sử dụng làm biến phụ trợ để giải thích kết quả sau phân cụm.

Dưới đây là chi tiết về phương pháp tính toán và công thức của các chỉ số được sử dụng:

7.6.1. Chỉ số đánh giá độ tin cậy (Wilson Score)

Để khắc phục nhược điểm của điểm đánh giá trung bình (Rating) – vốn thường bị sai lệch khi số lượng đánh giá quá ít (ví dụ: 5.0 sao nhưng chỉ có 1 lượt đánh giá), nhóm sử dụng công thức **Wilson Lower Bound Score**. Đây là phương pháp thống kê giúp ước lượng cận dưới của khoảng tin cậy cho điểm đánh giá thực tế.

Công thức toán học:

$$W = \frac{\hat{p} + \frac{z^2}{2n} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Trong đó:

- W: Điểm Wilson Score (giá trị từ 0 đến 1).
- $\hat{p}$: Tỷ lệ đánh giá tích cực quan sát được. Được tính bằng: $\text{Rating trung bình}/5$.
- n: Tổng số lượng đánh giá (Review Count) của nhà xe.
- z: Phân vị chuẩn tắc của mức độ tin cậy. Trong đồ án này, nhóm chọn $z = 1.96$, tương ứng với độ tin cậy 95%.

7.6.2. Các biến tổng hợp (Aggregated Features)

Để giảm số chiều dữ liệu và tập trung vào các khía cạnh cốt lõi của trải nghiệm khách hàng, các tiêu chí đánh giá chi tiết được gom nhóm thành hai chỉ số chính bằng phương pháp trung bình cộng.

Điểm Dịch vụ (Service Score)

Đo lường mức độ hài lòng về trải nghiệm vật lý và cảm xúc của hành khách.

$$\text{Service Score} = \frac{\text{Thái độ nhân viên} + \text{Chất lượng xe} + \text{Tiện nghi}}{3}$$

Điểm Tin cậy (Trust Score)

Đo lường mức độ uy tín trong vận hành và cam kết của nhà xe.

$$\text{Trust Score} = \frac{\text{An toàn} + \text{Đúng giờ} + \text{Thông tin chính xác}}{3}$$

7.6.3. Chỉ số Công bằng (Fairness Index)

Đây là chỉ số phái sinh được nhóm xây dựng nhằm đo lường hiệu suất giá trị (Value for Money) – tức là chất lượng khách hàng nhận được so với số tiền họ bỏ ra.

$$\text{Fairness Index} = \frac{\text{Wilson Score}}{\sqrt{\text{Real Price}}}$$

Trong đó:

- **Wilson Score**: Điểm chất lượng đã được chuẩn hóa (tính ở mục 1).
- **Real Price**: Giá vé thực tế (sau khi trừ khuyến mãi).
- *Lưu ý:* Giá vé được lấy căn bậc hai ($\sqrt{...}$) để thu hẹp biên độ dao động giá, giúp chỉ số không bị quá nhỏ và cân bằng trọng số giữa Giá và Chất lượng.

Chương 8: Mô Hình Hóa (Modeling - Unsupervised Learning)

8.1. Lựa chọn thuật toán

Trong bài toán phân cụm các chuyến xe khách dựa trên giá vé và mức độ hài lòng, việc lựa chọn thuật toán phân cụm phù hợp có vai trò quyết định đến chất lượng và khả năng diễn giải của mô hình do đó ta sẽ lựa chọn thuật toán K Means vì là thuật toán được sử dụng phổ biến nhờ tính đơn giản và khả năng mở rộng tốt đối với dữ liệu có kích thước lớn.

8.2. Chuẩn hóa dữ liệu (RobustScaler, MinMaxScaler…)

Do các biến đầu vào như giá vé, thời lượng hành trình và mức độ hài lòng có thang đo không đồng nhất, quá trình chuẩn hóa dữ liệu là bước bắt buộc nhằm đảm bảo mô hình phân cụm hoạt động ổn định. RobustScaler được sử dụng trong trường hợp dữ liệu chứa nhiều outlier, khi đó thuật toán sẽ chuẩn hóa dựa trên median và interquartile range nhằm giảm ảnh hưởng của các giá trị cực đoan. Trong khi đó, MinMaxScaler đưa dữ liệu về cùng khoảng [0, 1], giúp các thuật toán nhạy cảm với khoảng cách như KMeans đạt hiệu quả tốt hơn và vì dữ liệu có một số biến có ngoại lệ (outlier) và phân phối lệch nên trong bài nghiên cứu này ta sẽ sử dụng RobustScaler để chuẩn hóa dữ liệu

```
scaler = RobustScaler()
X_scaled = scaler.fit_transform(df_cluster[features])
X_scaled[:5]
```

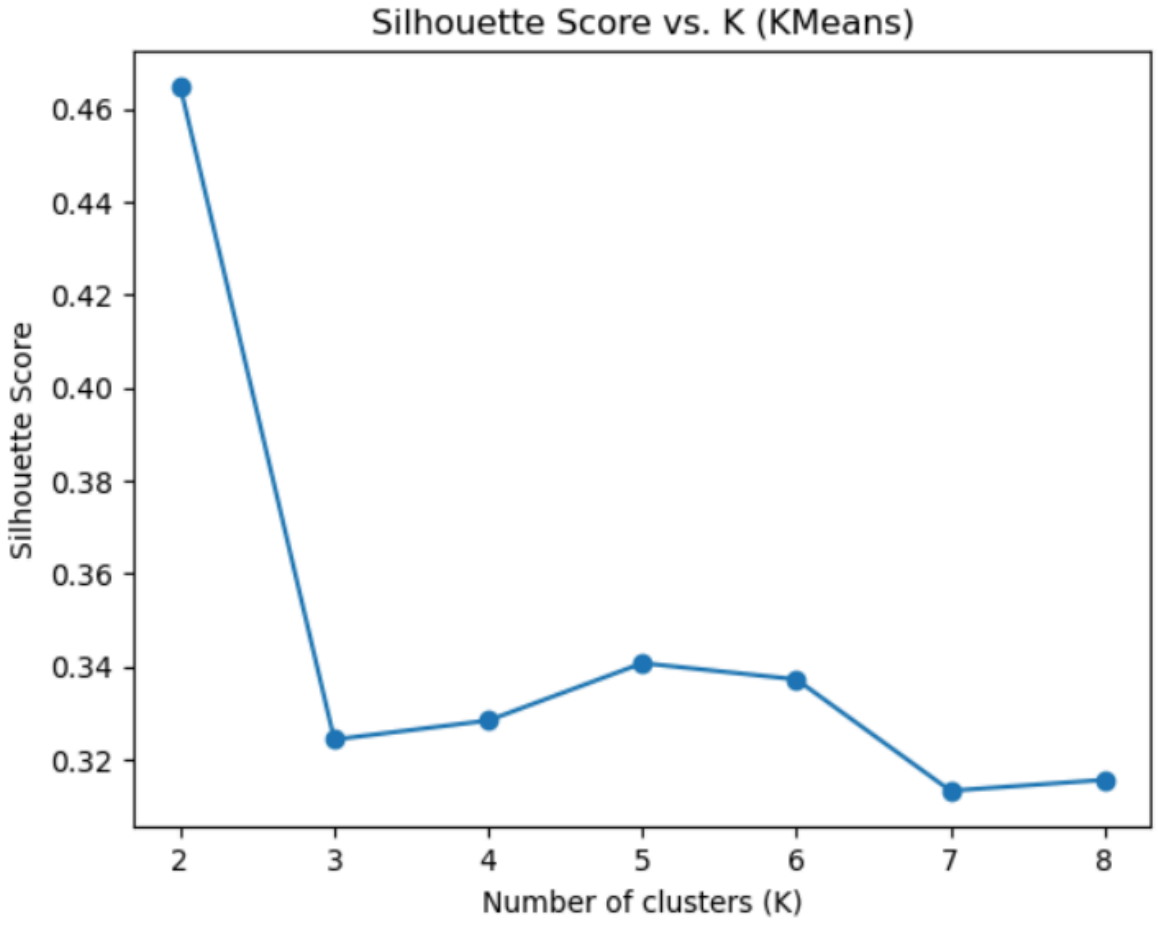
```
array([[ 0.50653262, -0.06266591,  0.39887517,  0.4       ,  0.31578947],
       [ 0.18387972, -0.08831917,  0.24376506,  0.33333333,  0.36842105],
       [ 0.27553245,  0.48235091, -0.11529946, -0.06666667,  0.        ],
       [ 0.33162087,  0.23193753,  0.08475004,  0.        ,  0.15789474],
       [ 0.45895135, -0.06266591,  0.37314172,  0.13333333,  0.10526316]])
```

8.3. Thử nghiệm nhiều K và đánh giá silhouette score

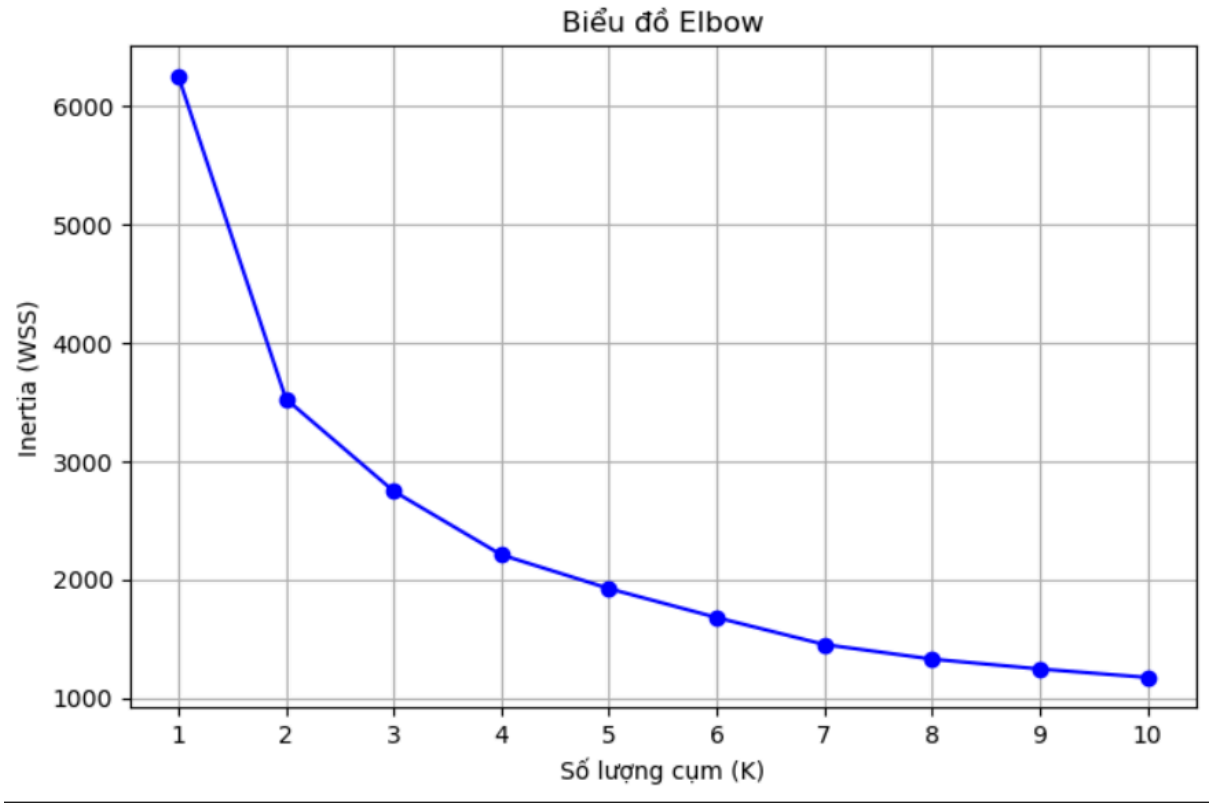
Đối với các mô hình yêu cầu lựa chọn số lượng cụm như KMeans hoặc Agglomerative, việc xác định giá trị K tối ưu được thực hiện thông qua quá trình thử nghiệm lặp lại với nhiều giá trị khác nhau. Silhouette score đóng vai trò là chỉ số đánh giá chính, phản ánh mức độ tách biệt giữa các cụm và sự gắn kết của các quan sát trong từng nhóm. Một giá trị silhouette cao cho thấy ranh giới giữa các cụm rõ ràng, trong khi giá trị thấp hoặc âm phản ánh sự chồng chéo giữa các cụm và cấu trúc không ổn định. Ngoài ra, có thể kết hợp với phương pháp Elbow để quan sát độ giảm của tổng bình phương sai trong cụm, tạo nền tảng bổ sung cho quá trình lựa chọn K. Kết quả thử nghiệm cho thấy những giá trị K tối ưu tập trung vào các trường hợp phản ánh rõ nét sự phân tầng về giá vé và mức độ hài lòng của hành khách.

```
K=2, Silhouette Score=0.4648
K=3, Silhouette Score=0.3242
K=4, Silhouette Score=0.3283
K=5, Silhouette Score=0.3407
K=6, Silhouette Score=0.3372
K=7, Silhouette Score=0.3132
K=8, Silhouette Score=0.3155

Số cụm tốt nhất theo Silhouette: 2
```



Vẽ Silhouette theo K để trực quan hóa



Từ K=1 xuống K=2: Độ dốc giảm cực mạnh (Inertia giảm từ >6000 xuống ~3500).

Từ K=2 xuống K=3: Độ dốc vẫn còn khá lớn (Inertia giảm từ ~3500 xuống ~2800).

Từ K=3 trở đi: Đường biểu đồ bắt đầu thoải dần và đi ngang.

Kết luận từ Elbow: Điểm gập khúc (khuỷu tay) nằm ở khoảng K=3.

8.4. Diễn giải cụm và đặc điểm từng cluster

Sau khi mô hình phân cụm được lựa chọn và huấn luyện, giai đoạn diễn giải cụm giúp hiểu rõ hơn ý nghĩa thực tiễn của các nhóm được tạo ra. Mỗi cụm được phân tích dựa trên giá trị trung bình, phương sai và phân phối của các thuộc tính quan trọng, từ đó rút ra đặc điểm hành khách hoặc loại chuyến xe tiêu biểu cho từng nhóm. Chẳng hạn, một cụm có thể đại diện cho các chuyến xe giá thấp nhưng mức độ hài lòng trung bình, trong khi một cụm khác thể hiện nhóm dịch vụ cao cấp với giá vé và đánh giá trải nghiệm cùng tăng. Việc diễn giải các cụm không chỉ hỗ trợ đánh giá tính hợp lý của mô hình, mà còn mang lại insight giá trị cho doanh nghiệp trong việc định hướng chiến lược giá, cải thiện chất lượng dịch vụ hoặc tối ưu hóa mô hình vận hành. Các cụm có sự giao thoa cần được phân tích kỹ lưỡng nhằm đánh giá khả năng mô hình bị ảnh hưởng bởi nhiễu hoặc thiếu đặc trưng quan trọng.

cluster	wilson_score	log_price	fairness_index	trust_score	service_score
0	0.85	12.13	0.0	4.61	4.51
1	0.60	12.92	0.0	3.97	3.77
2	0.87	12.80	0.0	4.66	4.55

Số lượng chuyến xe trong từng cụm:

cluster	count
0	478
1	450
2	873

Dựa trên các chỉ số trung bình về Giá (Log Price) và Chất lượng (Trust Score, Service Score, Wilson Score), mô hình phân cụm với K=3 đã phân tách dữ liệu thành 3 phân khúc thị trường rất rõ ràng.

Cụm 0: Phân khúc "Ngon - Bổ - Rẻ"

Đặc điểm:

Giá vé: Thấp nhất trong 3 nhóm (`log_price` ~ 12.13).

Chất lượng: Rất tốt, ngang ngửa với nhóm cao cấp (`trust_score`: 4.61, `service_score`: 4.51).

Độ tin cậy: Điểm Wilson cao (0.85), cho thấy lượng đánh giá tích cực ổn định.

Ý nghĩa: Đây là nhóm đại diện cho sự "Tối ưu hóa giá trị". Các nhà xe trong nhóm này đã thành công trong việc cung cấp dịch vụ chất lượng cao với chi phí vận hành cạnh tranh. Đây là nhóm chủ lực để thu hút và giữ chân khách hàng trung thành.

Cụm 1: Phân khúc "Giá ảo - Chất lượng thấp"

Đặc điểm:

Giá vé: Cao nhất (`log_price` ~ 12.92), đắt hơn cả nhóm Cao cấp.

Chất lượng: Tệ nhất (`trust_score`: 3.97, `service_score`: 3.77).

Độ tin cậy: Thấp nhất (0.60).

Ý nghĩa: Nhóm chuyến xe có hiệu suất kém, giá không tương xứng với chất lượng (có thể do độc quyền tuyến, tăng giá mùa cao điểm nhưng phục vụ kém).

Cụm 2: Phân khúc "Cao cấp - Đáng tiền"

Đặc điểm:

Giá vé: Cao (`log_price` ~ 12.80).

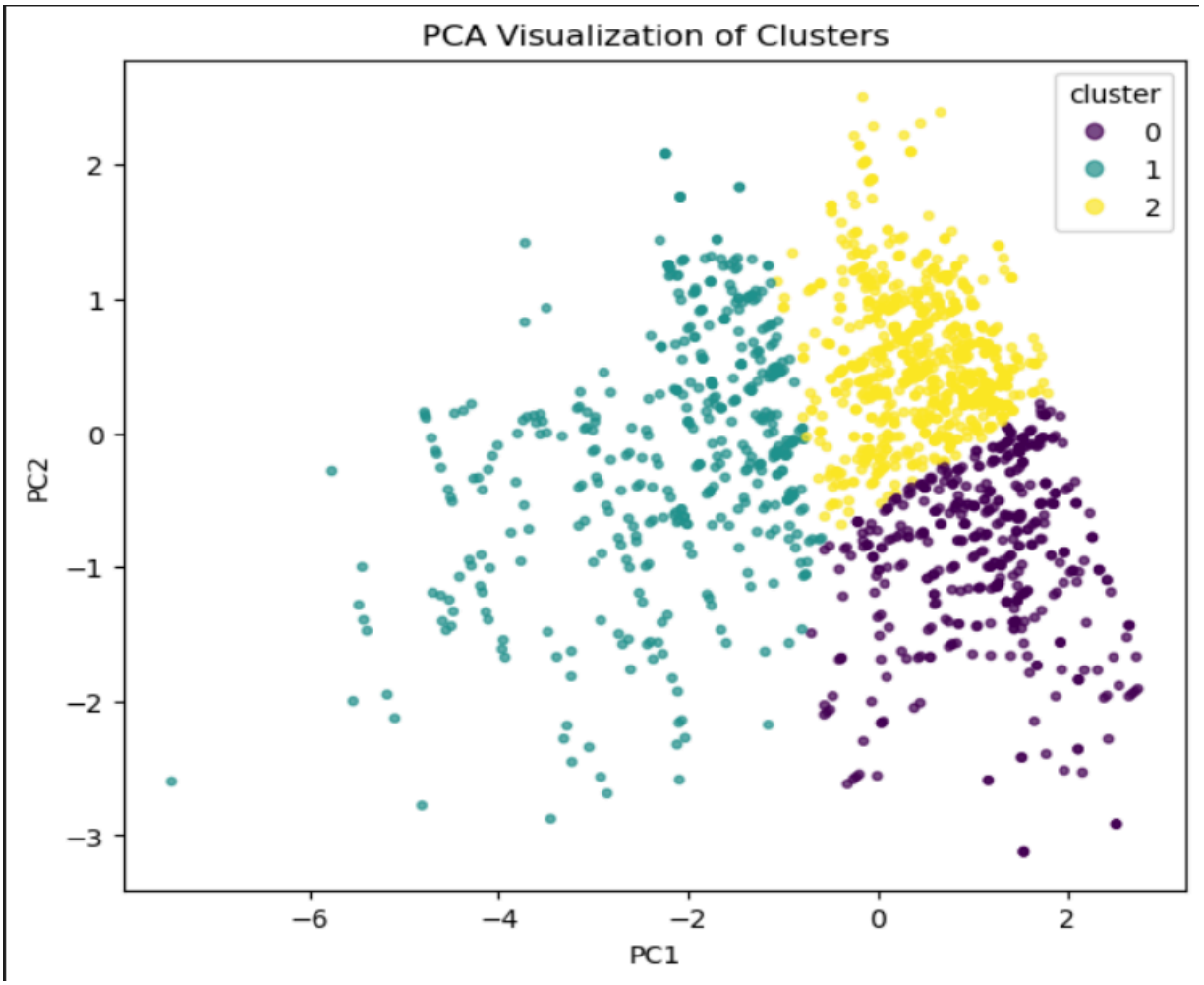
Chất lượng: Tốt nhất thị trường (`trust_score`: 4.66, `service_score`: 4.55).

Số lượng: Chiếm số lượng lớn nhất trong dữ liệu (873 chuyến).

Ý nghĩa: Nhóm "tiền nào của nấy". Khách hàng trả giá cao để nhận lại dịch vụ 5 sao.

8.5. Trực quan hoá các cụm (2D / 3D scatter plot)

Để giúp người đọc hiểu rõ hơn cấu trúc các cụm, việc trực quan hóa kết quả mô hình là bước thiết yếu. Biểu đồ scatter hai chiều thường tập trung vào hai biến chính như giá vé và mức độ hài lòng, với màu sắc đại diện cho từng cụm. Điều này cung cấp cái nhìn trực quan về cách dữ liệu được tách biệt theo từng nhóm và mức độ rõ ràng của ranh giới giữa các cụm. Trong trường hợp mô hình sử dụng nhiều đặc trưng hơn, việc trực quan hóa ba chiều giúp mô phỏng cấu trúc không gian phức tạp hơn, phản ánh sự ảnh hưởng của các biến bổ sung như thời lượng hành trình hoặc loại ghế. Các biểu đồ này cũng giúp đánh giá trực quan mức độ chồng lấp giữa các cụm, phát hiện các điểm nhiễu hoặc quan sát có đặc điểm quá khác biệt so với nhóm. Toàn bộ quá trình trực quan hóa đóng vai trò hỗ trợ diễn giải, đồng thời củng cố tính minh bạch và thuyết phục của mô hình phân cụm trong báo cáo.



Trực quan hóa bằng PCA

Chương 9: Triển khai ứng dụng (Deployment)

9.1. Mục tiêu và phạm vi triển khai

Mục tiêu của giai đoạn triển khai là đưa mô hình phân cụm các chuyến xe khách theo giá vé và chất lượng dịch vụ vào một ứng dụng web tương tác, giúp người dùng không cần môi trường lập trình vẫn có thể sử dụng mô hình. Ứng dụng được xây dựng bằng Streamlit, cho phép tải dữ liệu đã xử lý từ thư mục *data/processed*, huấn luyện mô hình KMeans với số cụm cố định, trực quan hóa kết quả phân cụm bằng PCA hai chiều và dự đoán cụm cho các chuyến xe mới dựa trên thông tin giá vé và các điểm đánh giá dịch vụ

	company_name	departure_time	pickup_point	arrival_time	dropoff_point	price_original	price_discounted	departure_date	start_point	destination	rating_safety	rating_info_accuracy	rating_info_cor
0	Hoàng Thuỷ	18:45:00	Bến xe	05:45	Bến xe	350000	300000	2025-11-12	Sài Gòn	Gia Lai	4.8	4.7	
1	Sinh Diên Hồng	17:30:00	Bến xe	04:15	Bến xe	370000	296000	2025-11-12	Sài Gòn	Gia Lai	4.7	4.7	
2	Phong Phú	20:00:00	Văn phòng	07:45	Văn phòng	530000	399000	2025-11-12	Sài Gòn	Gia Lai	4.5	4.5	
3	Kính Diên Hồng	18:40:00	Bến xe	05:40	Bến xe	350000	0	2025-11-12	Sài Gòn	Gia Lai	4.7	4.6	
4	Đức Đạt	20:10:00	Bến xe	06:35	Bến xe	350000	300000	2025-11-12	Sài Gòn	Gia Lai	4.6	4.5	

9.2. Môi trường và cấu trúc ứng dụng

Ứng dụng được triển khai trong môi trường Python 3 với các thư viện chính gồm pandas, numpy, scikit-learn (RobustScaler, KMeans, PCA), matplotlib và Streamlit. Toàn bộ logic được hiện thực trong tệp app.py, trong đó chương trình tự động quét các tệp CSV trong thư mục data/processed, đọc và hợp nhất dữ liệu, chuẩn hóa kiểu dữ liệu số, sau đó áp dụng một hàm *feature_engineering* dùng chung cho cả giai đoạn huấn luyện và dự đoán. Hàm này xây dựng các đặc trưng như giá thực tế *real_price*, *log_price*, chỉ số dịch vụ *service_score*, chỉ số tin cậy *trust_score*, điểm Wilson *wilson_score* và chỉ số công bằng *fairness_index*, trước khi chọn ra tập đặc trưng cuối cùng để đưa vào mô hình phân cụm. Về cấu trúc, app.py đóng vai trò điểm vào của ứng dụng, thư mục data/processed là nguồn dữ liệu huấn luyện, các bước xử lý và mô hình được thực hiện trực tiếp trong tiến trình *Streamlit* mà không cần thêm một dịch vụ cơ sở dữ liệu hay API trung gian ở giai đoạn này. Tại mục này có thể chèn Hình X.2 là sơ đồ kiến trúc logic của hệ thống, mô tả luồng: người dùng truy cập giao diện web → ứng dụng Streamlit đọc dữ liệu và trích xuất đặc trưng → mô hình KMeans phân cụm → ứng dụng trả về kết quả và trực quan hóa.

9.3. Quy trình triển khai ứng dụng

Quy trình triển khai ứng dụng bắt đầu từ việc chuẩn bị môi trường Python, cài đặt các thư viện phụ thuộc được liệt kê trong tệp requirements.txt, sau đó đảm bảo thư mục *data/processed* đã chứa tập dữ liệu chuyến xe đã qua bước tiền xử lý. Ứng dụng được khởi chạy bằng lệnh streamlit run app.py trên máy cục bộ hoặc máy chủ, lúc này Streamlit tự động mở một phiên làm việc trên trình duyệt và sẵn sàng cho người dùng tương tác. Khi ứng dụng hoạt động, ở lần tải đầu tiên, hệ thống sẽ đọc toàn bộ dữ liệu huấn luyện từ các tệp CSV, chuyển đổi các cột số về đúng kiểu, áp dụng pipeline tạo đặc trưng, chuẩn hóa dữ liệu bằng RobustScaler và huấn luyện mô hình KMeans với số cụm K = 3. Kết quả phân cụm trên dữ liệu huấn luyện được trực quan hóa bằng PCA hai chiều để thể hiện tương quan tương đối giữa các cụm. Sau đó người dùng có thể nhập thông tin của một chuyến xe mới gồm giá gốc, giá khuyến mãi, các điểm đánh giá thành phần và số lượng người đánh giá trực tiếp trên giao diện. Ứng dụng sử dụng lại chính pipeline tạo đặc trưng và bộ chuẩn hóa đã được huấn luyện từ dữ liệu gốc để suy ra cụm dự đoán cho chuyến xe mới, đồng thời hiển thị bảng thông tin chi tiết cùng nhãn cụm và phần diễn giải ý nghĩa kinh doanh của từng cụm (chẳng hạn như cụm giá hợp lý với dịch vụ ổn định, cụm giá cao nhưng trải nghiệm chưa tương xứng, hay cụm dịch vụ chất lượng cao với trải nghiệm toàn diện).

🔮 Kết quả dự đoán cụm cho dữ liệu mới

	price_original	price_discounted	rating_overall	reviewer_count	real_price	log_price	wilson_score
0	400.000	350.000	4.3	80	350.000	12.7657	0.7674

📘 Giải thích ý nghĩa các cụm xuất hiện trong dự đoán

🎯 Cluster 2 – Dịch vụ chất lượng cao – Trải nghiệm trọn vẹn

📌 **Cụm 2 – Dịch vụ chất lượng cao – Trải nghiệm trọn vẹn**

- Giá vé thuộc nhóm cao, đi kèm chất lượng phục vụ tốt
- Điểm hài lòng ổn định và mức độ tin cậy vượt trội
- Wilson Score cao → phản ánh sự đồng thuận lớn từ người dùng

👉 Cụm này đại diện cho **dịch vụ cao cấp**, phù hợp hành khách chú trọng trải nghiệm, sự an toàn và tính chuyên nghiệp trong suốt hành trình.

Các dòng thuộc cụm 2: [0]

Chương 10: Kết Luận Và Hướng Phát Triển

10.1. Những gì mô hình đã làm được

Kết quả nghiên cứu cho thấy mô hình phân cụm với số cụm tối ưu K=3 đã thành công trong việc phân tách dữ liệu thành ba phân khúc thị trường có đặc điểm khác biệt rõ ràng. Dựa trên các chỉ số trung bình về giá vé (log price) và chất lượng dịch vụ (trust score, service score, Wilson score), mô hình đã hình thành một cấu trúc phân khúc phản ánh chính xác sự đa dạng của thị trường vận tải hành khách. Cụm thứ nhất đại diện cho phân khúc "Ngon - Bổ - Rẻ", nơi các chuyến xe có mức giá thấp nhưng chất lượng dịch vụ lại duy trì ở mức cao, đồng thời độ tin cậy được đảm bảo bởi lượng đánh giá tích cực ổn định. Đây là minh chứng cho khả năng tối ưu hóa giá trị, khi doanh nghiệp có thể cung cấp dịch vụ chất lượng với chi phí cạnh tranh, từ đó tạo ra lợi thế trong việc thu hút và duy trì khách hàng trung thành.

Ngược lại, cụm thứ hai phản ánh phân khúc "Giá ảo - Chất lượng thấp", nơi giá vé được đẩy lên mức cao nhất nhưng chất lượng dịch vụ lại ở mức thấp nhất, đồng thời độ tin cậy cũng suy giảm đáng kể. Kết quả này cho thấy sự bất cân đối giữa giá và chất lượng, có thể bắt nguồn từ tình trạng độc quyền tuyến hoặc chiến lược tăng giá trong mùa cao điểm mà không đi kèm cải thiện dịch vụ. Đây là nhóm có hiệu suất kém, và việc nhận diện được phân khúc này mang lại giá trị thực tiễn trong việc cảnh báo doanh nghiệp về nguy cơ mất khách hàng và giảm uy tín thương hiệu.

Cuối cùng, cụm thứ ba đại diện cho phân khúc "Cao cấp - Đáng tiền", nơi giá vé cao đi kèm với chất lượng dịch vụ tốt nhất thị trường. Đây là nhóm chiếm tỷ trọng lớn nhất trong dữ liệu, phản ánh xu hướng "tiền nào của nấy" khi khách hàng sẵn sàng trả chi phí cao để nhận lại trải nghiệm dịch vụ năm sao. Kết quả này cho thấy sự tồn tại của một phân khúc cao cấp ổn định, đóng vai trò quan trọng trong việc duy trì hình ảnh và uy tín của ngành vận tải hành khách.

Nhìn chung, mô hình đã chứng minh khả năng phân loại dữ liệu theo đặc trưng giá và chất lượng, đồng thời cung cấp những insight có giá trị thực tiễn cho cả hành khách và doanh nghiệp. Hành khách có thể dựa vào kết quả phân cụm để đưa ra quyết định lựa chọn dịch vụ phù hợp, trong khi doanh nghiệp có thể sử dụng thông tin này để tối ưu chiến lược giá, nâng cao chất lượng dịch vụ và nhận diện bất thường trong vận hành.

10.2. Những hạn chế của dữ liệu & mô hình

Mặc dù đạt được những kết quả khả quan, nghiên cứu vẫn tồn tại một số hạn chế cần được thừa nhận. Trước hết, dữ liệu được thu thập chỉ từ một nền tảng duy nhất là vexere, do đó chưa phản ánh đầy đủ toàn bộ thị trường vận tải hành khách. Việc giới hạn nguồn dữ liệu có thể dẫn đến sai lệch trong việc khái quát hóa kết quả, đặc biệt khi các nền tảng khác có thể sở hữu đặc điểm giá vé và chất lượng dịch vụ khác biệt.

Thứ hai, yếu tố thời gian chưa được đưa vào phân tích. Giá vé và đánh giá dịch vụ thường biến động theo mùa vụ, ngày lễ hoặc giờ cao điểm, nhưng mô hình hiện tại mới chỉ dừng lại ở việc phân tích snapshot dữ liệu tại một thời điểm. Điều này hạn chế khả năng nhận diện xu hướng động và làm giảm tính ứng dụng trong dự báo.

Ngoài ra, dữ liệu đánh giá khách hàng vốn mang tính chủ quan, có thể bị thiên lệch hoặc thiếu tính nhất quán giữa các tuyến và nhà xe. Điều này ảnh hưởng đến độ tin cậy của các chỉ số chất lượng, và do đó có thể tác động đến kết quả phân cụm.

Cuối cùng, mô hình phân cụm mới dừng lại ở thuật toán cơ bản như K-Means. Mặc dù thuật toán này đã cho kết quả khả quan, song chưa khai thác hết tiềm năng của các phương pháp nâng cao như Gaussian Mixture Model, Spectral Clustering hoặc các kỹ thuật học sâu. Khả năng tổng quát hóa của mô hình vì thế còn hạn chế, đặc biệt khi dữ liệu chưa đa dạng và chưa kết hợp thêm các yếu tố nhân khẩu học hoặc hành vi đặt vé.

10.3. Hướng mở rộng trong tương lai

Để khắc phục những hạn chế và nâng cao giá trị ứng dụng, nghiên cứu có thể được mở rộng theo nhiều hướng. Trước hết, việc mở rộng nguồn dữ liệu từ nhiều nền tảng đặt vé khác nhau sẽ giúp tăng tính đại diện và độ tin cậy của kết quả phân tích. Kết hợp dữ liệu trực tiếp từ các doanh nghiệp vận tải cũng có thể bổ sung thêm thông tin về vận hành, chi phí và chiến lược giá, từ đó làm phong phú tập dữ liệu đầu vào.

Tiếp theo, việc đưa yếu tố thời gian vào phân tích sẽ cho phép nhận diện xu hướng biến động giá vé và chất lượng dịch vụ theo mùa vụ, ngày lễ hoặc giờ cao điểm. Điều này không chỉ nâng cao khả năng mô tả mà còn mở ra cơ hội xây dựng mô hình dự báo, hỗ trợ doanh nghiệp trong việc hoạch định chiến lược giá động.

Ngoài ra, việc thử nghiệm các thuật toán phân cụm nâng cao như Gaussian Mixture Model, Spectral Clustering hoặc Deep Clustering sẽ giúp cải thiện độ chính xác và khả năng diễn giải của mô hình. Các phương pháp này có thể khai thác tốt hơn cấu trúc tiềm ẩn trong dữ liệu, đặc biệt khi dữ liệu có tính phi tuyến hoặc phân phối phức tạp.

Một hướng phát triển quan trọng khác là tích hợp kết quả phân cụm vào hệ thống gợi ý, nhằm hỗ trợ hành khách lựa chọn chuyến xe tối ưu theo nhu cầu cá nhân. Kết hợp thêm dữ liệu nhân khẩu học và lịch sử đặt vé sẽ cho phép phân tích hành vi khách hàng, từ đó xây dựng các mô hình gợi ý cá nhân hóa có giá trị thực tiễn cao.

Cuối cùng, việc triển khai kết quả nghiên cứu vào thực tế thông qua dashboard cho doanh nghiệp vận tải sẽ mang lại lợi ích trực tiếp trong quản lý và vận hành. Các doanh nghiệp có thể sử dụng thông tin phân cụm để tối ưu chiến lược giá, nâng cao chất lượng dịch vụ và giám sát hiệu quả hoạt động, từ đó cải thiện trải nghiệm hành khách và nâng cao năng lực cạnh tranh trên thị trường.

Tài Liệu Tham Khảo

Trong quá trình thực hiện đề tài, nhóm đã tham khảo nhiều nguồn tài liệu khác nhau, bao gồm các công trình nghiên cứu học thuật, giáo trình chuyên ngành, cũng như các tài liệu trực tuyến và công cụ phần mềm. Các nguồn này không chỉ cung cấp nền tảng lý thuyết về học máy và phân cụm dữ liệu, mà còn hỗ trợ nhóm trong việc triển khai kỹ thuật thu thập, xử lý và trực quan hóa dữ liệu. Danh mục tài liệu tham khảo được trình bày dưới đây nhằm đảm bảo tính minh bạch và khoa học của báo cáo.

Tài liệu học thuật và sách chuyên ngành

1. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
3. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
4. Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.
5. Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining*. Pearson.

Nguồn trực tuyến và công cụ phần mềm

6. Scikit-learn Documentation. (2025). *Clustering algorithms*. Retrieved from https://scikit-learn.org
7. Pandas Documentation. (2025). *Data analysis and manipulation in Python*. Retrieved from https://pandas.pydata.org
8. PostgreSQL Documentation. (2025). *Relational database management system*. Retrieved from https://www.postgresql.org
9. Microsoft Power BI Documentation. (2025). *Data visualization and business intelligence*. Retrieved from https://powerbi.microsoft.com
10. Selenium Documentation. (2025). *Web automation and scraping*. Retrieved from https://www.selenium.dev

Nguồn dữ liệu thực nghiệm

11. Vexere. (2025). *Nền tảng đặt vé xe khách trực tuyến*. Retrieved from https://vexere.com

Nhận xét về phần tham khảo

Phần tài liệu tham khảo này đã được sắp xếp hợp lý, bao gồm cả nền tảng lý thuyết, công cụ triển khai và nguồn dữ liệu thực tế. Việc kết hợp các nguồn học thuật với tài liệu trực tuyến giúp báo cáo vừa có tính khoa học, vừa gắn với ứng dụng thực tiễn.