

BÁO CÁO CUỐI KÌ  
NHÓM K

DATE  
07/28/2024



# BÁO CÁO CUỐI KÌ

**MÔN HỌC:** XỬ LÝ SỐ LIỆU THỐNG KÊ

**GIÁO VIÊN HƯỚNG DẪN:** TÔ ĐỨC KHÁNH

**NHÓM THỰC HIỆN:** NHÓM K

**ĐỀ TÀI: PROJECT 2 - THUÊ XE ĐẠP CÔNG CỘNG Ở SEOUL**



# THÀNH VIÊN NHÓM K

Họ và tên	Mã số sinh viên	Phân công công việc
Trần Ngọc Khánh Như	21280040	Tiền xử lý dữ liệu
Nguyễn Đăng Khôi	21280023	Trực quan hóa dữ liệu
Trần Minh Hiển	21280016	Xây dựng và đánh giá mô hình
Lâm Gia Phú	21280104	

# MỤC LỤC

- 01 GIỚI THIỆU DỮ LIỆU
- 02 BÀI TOÁN CẦN GIẢI QUYẾT
- 03 PHƯƠNG ÁN XỬ LÝ DỮ LIỆU
- 04 MÔ HÌNH ĐƯỢC SỬ DỤNG
- 05 KẾT QUẢ ĐẠT ĐƯỢC

# 1. GIỚI THIỆU DỮ LIỆU





# DỮ LIỆU

- Nguồn Dữ Liệu: SeoulBikeData.csv
- Thời Gian: 01/12/2017 - 30/11/2018

# MỤC TIÊU

- Đảm bảo cung cấp xe đạp ổn định, giảm thời gian chờ đợi.
- Dự đoán số lượng xe đạp cần thiết cho mỗi giờ.

- Ngày Tháng: Date (year-month-day)
- Số Lượng Xe Đạp Thuê: Rented Bike count
- Giờ Trong Ngày: Hour
- Nhiệt Độ: Temperature (Celsius)
- Độ Ẩm: Humidity (%)
- Tốc Độ Gió: Windspeed (m/s)
- Tầm Nhìn Visibility (10m)
- Nhiệt Độ Điểm Sương: Dew point temperature (Celsius)
- Bức Xạ Mặt Trời: Solar radiation (MJ/m<sup>2</sup>)
- Lượng Mưa: Rainfall (mm)
- Lượng Tuyết: Snowfall (cm)
- Mùa: Seasons (Winter, Spring, Summer, Autumn)
- Ngày Lễ: Holiday (Holiday/No holiday)
- Giờ Hoạt Động: Functional Day (NoFunc/ Fun)



CÁC BIẾN  
QUAN SÁT

## 2. BÀI TOÁN CẦN GIẢI QUYẾT



- Xác định xu hướng thuê xe theo thời gian (ngày/tháng/năm).
- Đánh giá các yếu tố ảnh hưởng đến nhu cầu thuê xe đạp.
- Hỗ trợ đề ra chiến lược kinh doanh phù hợp.



### 3. PHƯƠNG ÁN XỬ LÝ DỮ LIỆU



TẢI DỮ LIỆU VÀ  
TRÍCH XUẤT CÁC  
TÍNH NĂNG



LÀM SẠCH DỮ LIỆU



TRỰC QUAN HÓA  
DỮ LIỆU



RÚT GỌN VÀ BIẾN  
ĐỔI DỮ LIỆU

# TẢI DỮ LIỆU VÀ TRÍCH XUẤT CÁC TÍNH NĂNG

Date	Rented Bike Count	Hour	Temperature(°C)
Length:8760	Min. : 0.0	Min. : 0.00	Min. :-17.80
Class :character	1st Qu.: 191.0	1st Qu.: 5.75	1st Qu.: 3.50
Mode :character	Median : 504.5	Median :11.50	Median : 13.70
	Mean : 704.6	Mean :11.50	Mean : 12.88
	3rd Qu.:1065.2	3rd Qu.:17.25	3rd Qu.: 22.50
	Max. :3556.0	Max. :23.00	Max. : 39.40
Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)
Min. : 0.00	Min. :0.000	Min. : 27	Min. :-30.600
1st Qu.:42.00	1st Qu.:0.900	1st Qu.: 940	1st Qu.: -4.700
Median :57.00	Median :1.500	Median :1698	Median : 5.100
Mean :58.23	Mean :1.725	Mean :1437	Mean : 4.074
3rd Qu.:74.00	3rd Qu.:2.300	3rd Qu.:2000	3rd Qu.: 14.800
Max. :98.00	Max. :7.400	Max. :2000	Max. : 27.200
Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons
Min. :0.0000	Min. : 0.0000	Min. :0.00000	Length:8760
1st Qu.:0.0000	1st Qu.: 0.0000	1st Qu.:0.00000	Class :character
Median :0.0100	Median : 0.0000	Median :0.00000	Mode :character
Mean :0.5691	Mean : 0.1487	Mean :0.07507	
3rd Qu.:0.9300	3rd Qu.: 0.0000	3rd Qu.:0.00000	
Max. :3.5200	Max. :35.0000	Max. :8.80000	
Holiday	Functioning Day		
Length:8760	Length:8760		
Class :character	Class :character		
Mode :character	Mode :character		

# LÀM SẠCH DỮ LIỆU

<chr>	<chr>	<chr>	<chr>
Date	character	365	0
Rented Bike Count	integer	2166	0
Hour	integer	24	0
Temperature(°C)	numeric	546	0
Humidity(%)	integer	90	0
Wind speed (m/s)	numeric	65	0
Visibility (10m)	integer	1789	0
Dew point temperature(°C)	numeric	556	0
Solar Radiation (MJ/m2)	numeric	345	0
Rainfall(mm)	numeric	61	0
Snowfall (cm)	numeric	51	0
Seasons	character	4	0
Holiday	character	2	0
Functioning Day	character	2	0

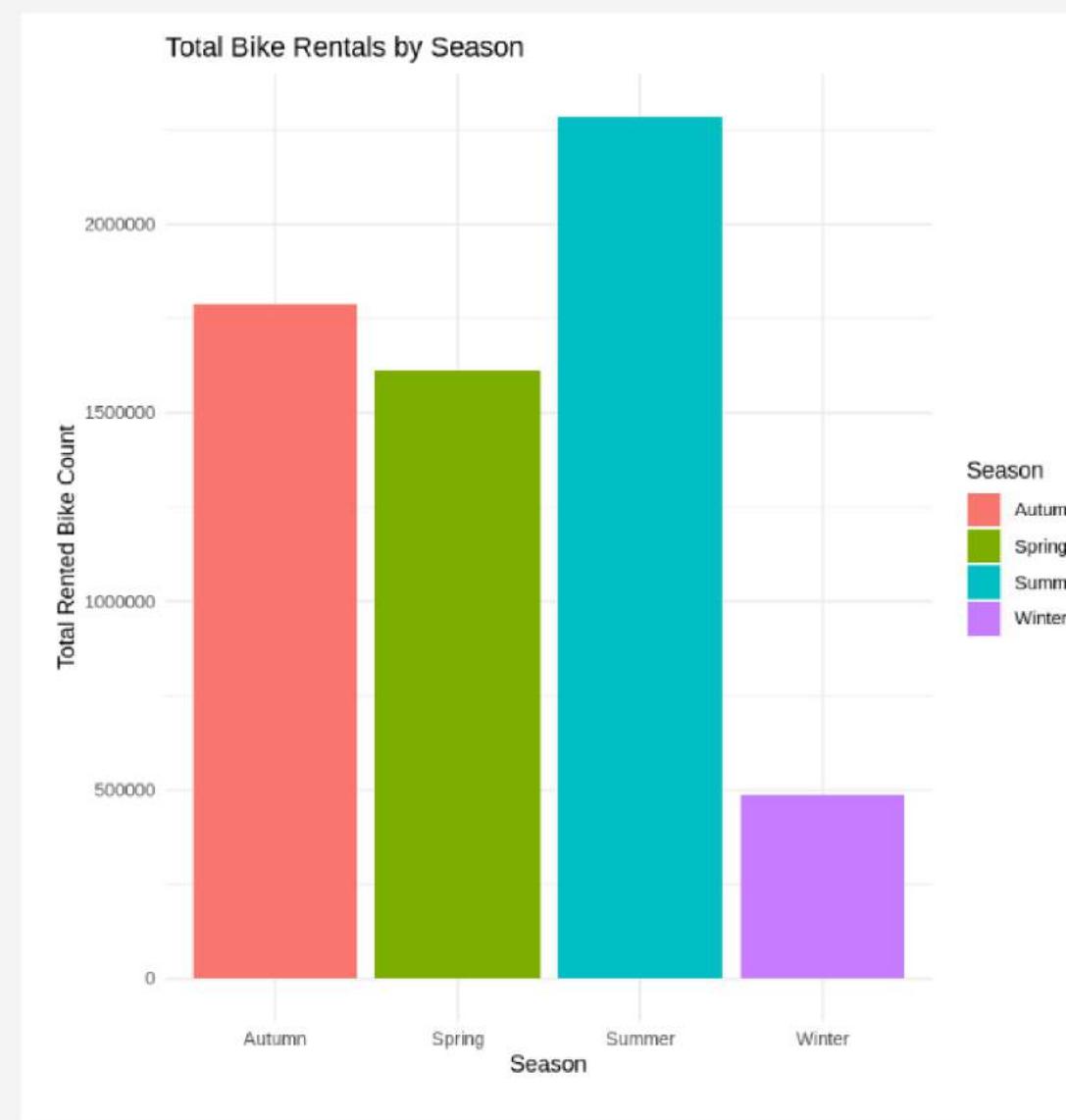
BÁO CÁO CUỐI KÌ  
NHÓM K

# TRỰC QUAN HÓA DỮ LIỆU

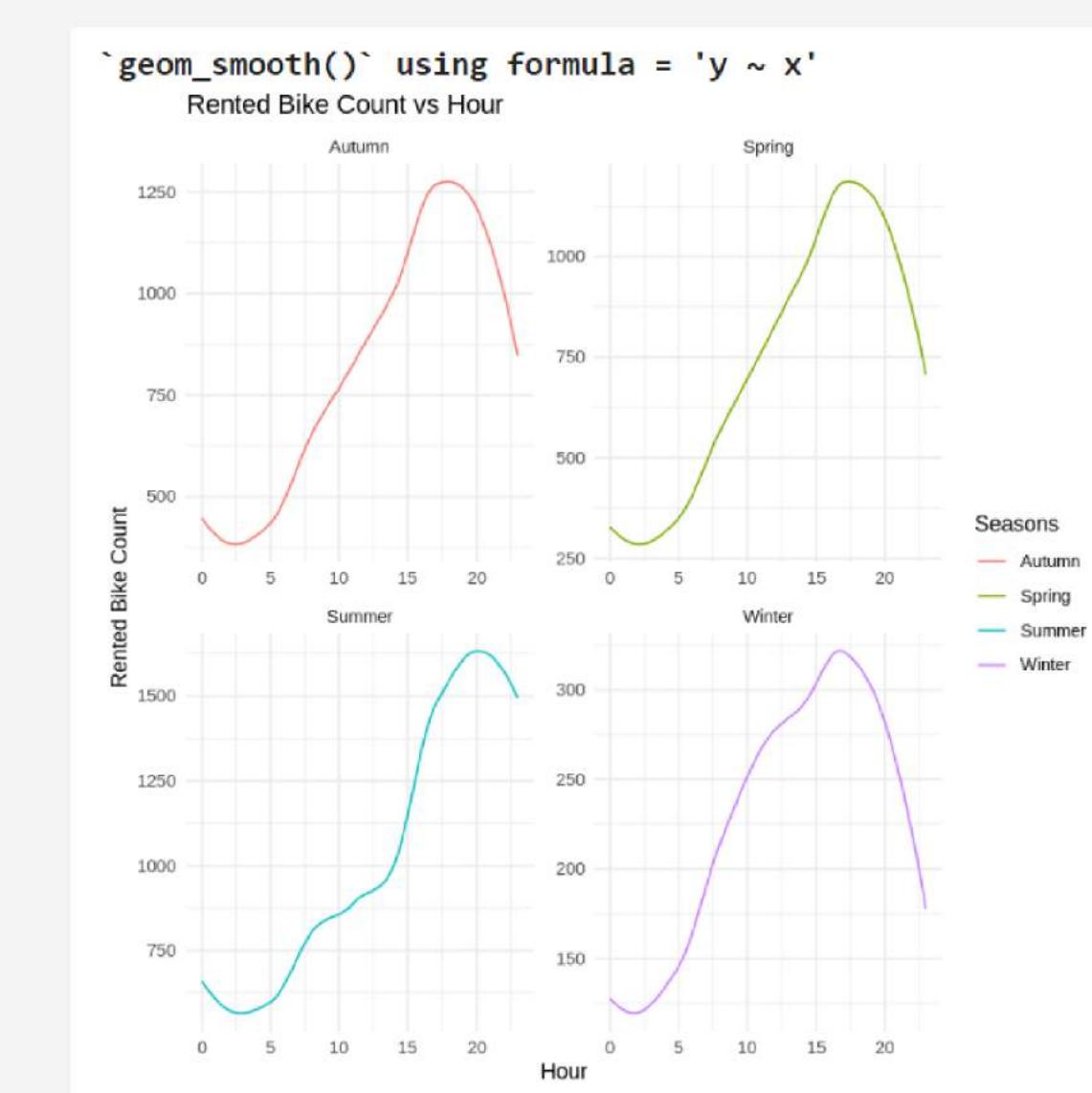


DATE  
07/28/2024

# BIỂU ĐỒ SỐ LƯỢNG XE ĐẠP THUÊ THEO MÙA

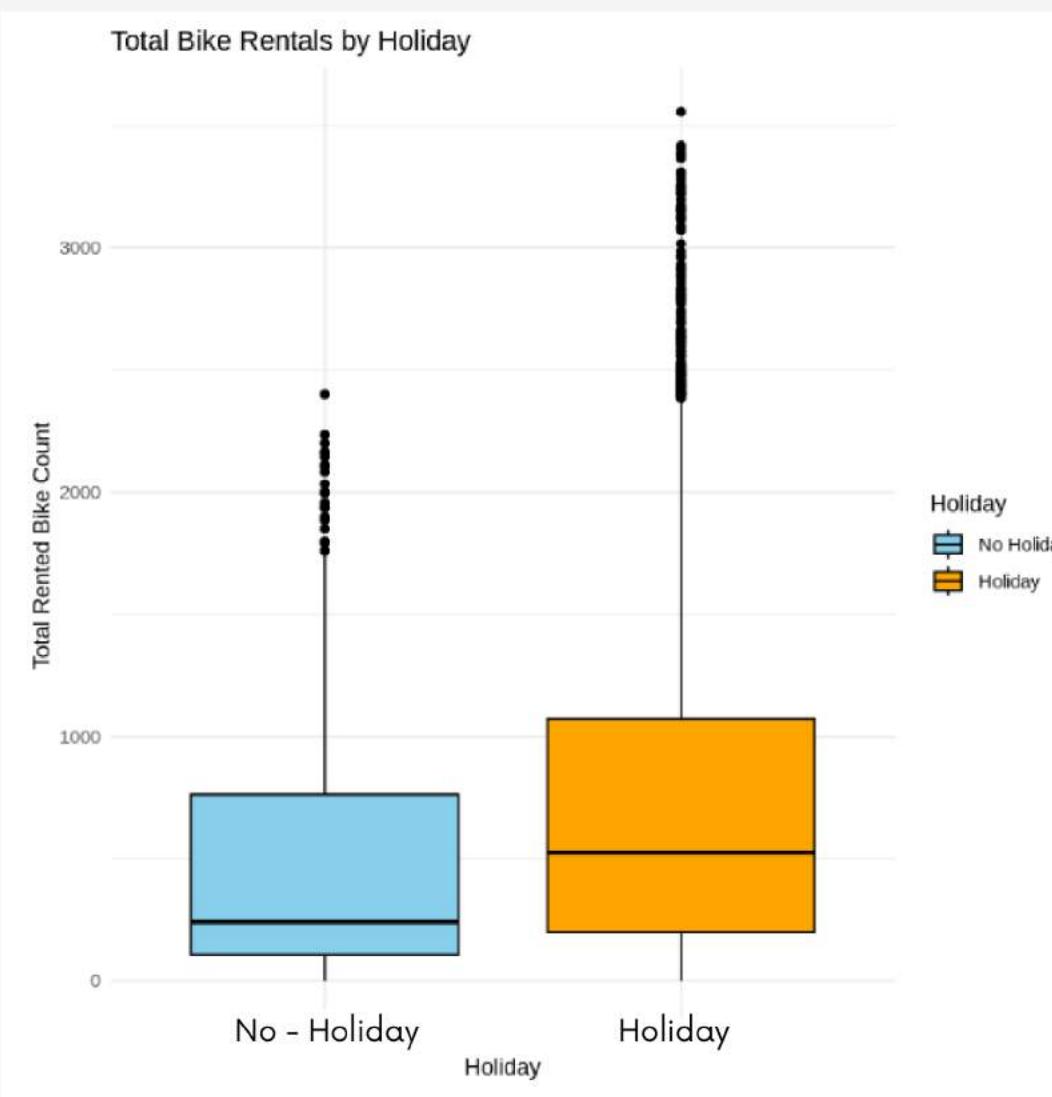


# BIỂU ĐỒ SỐ LƯỢNG XE ĐẠP THUÊ THEO GIỜ CHO TỪNG MÙA



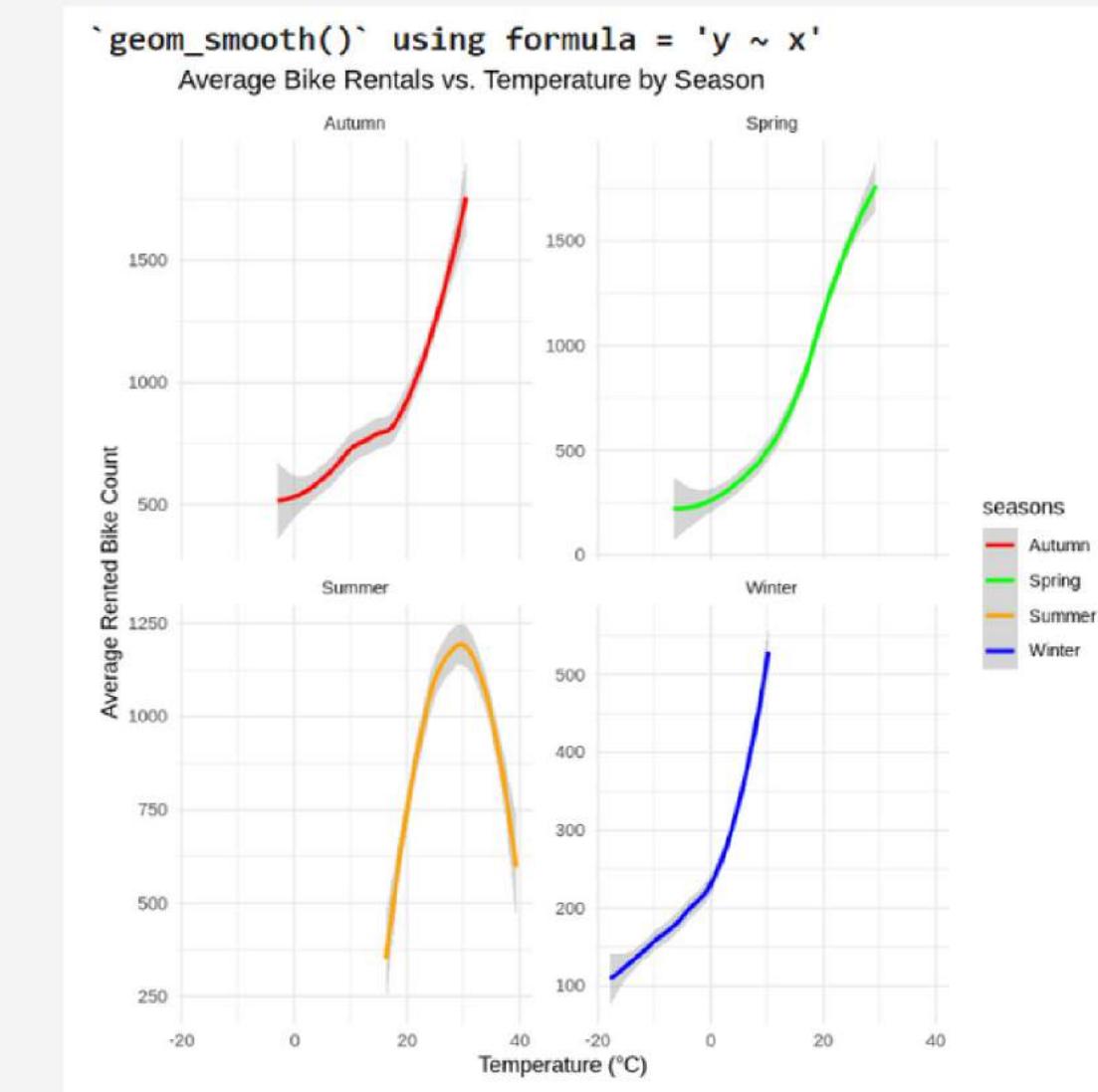
- Sử dụng cao nhất:** Buổi tối mùa hè khoảng 17:00-18:00.
- Sử dụng thấp nhất:** Buổi sáng sớm trong tất cả các mùa.
- Thời gian cao điểm nhất:** 17:00 là thời gian thuê xe đạp cao điểm nhất trong tất cả các mùa.

# BIỂU ĐỒ SỐ LƯỢNG XE ĐẠP ĐƯỢC THUÊ THEO CÁC NGÀY LỄ VÀ NGÀY THƯỜNG



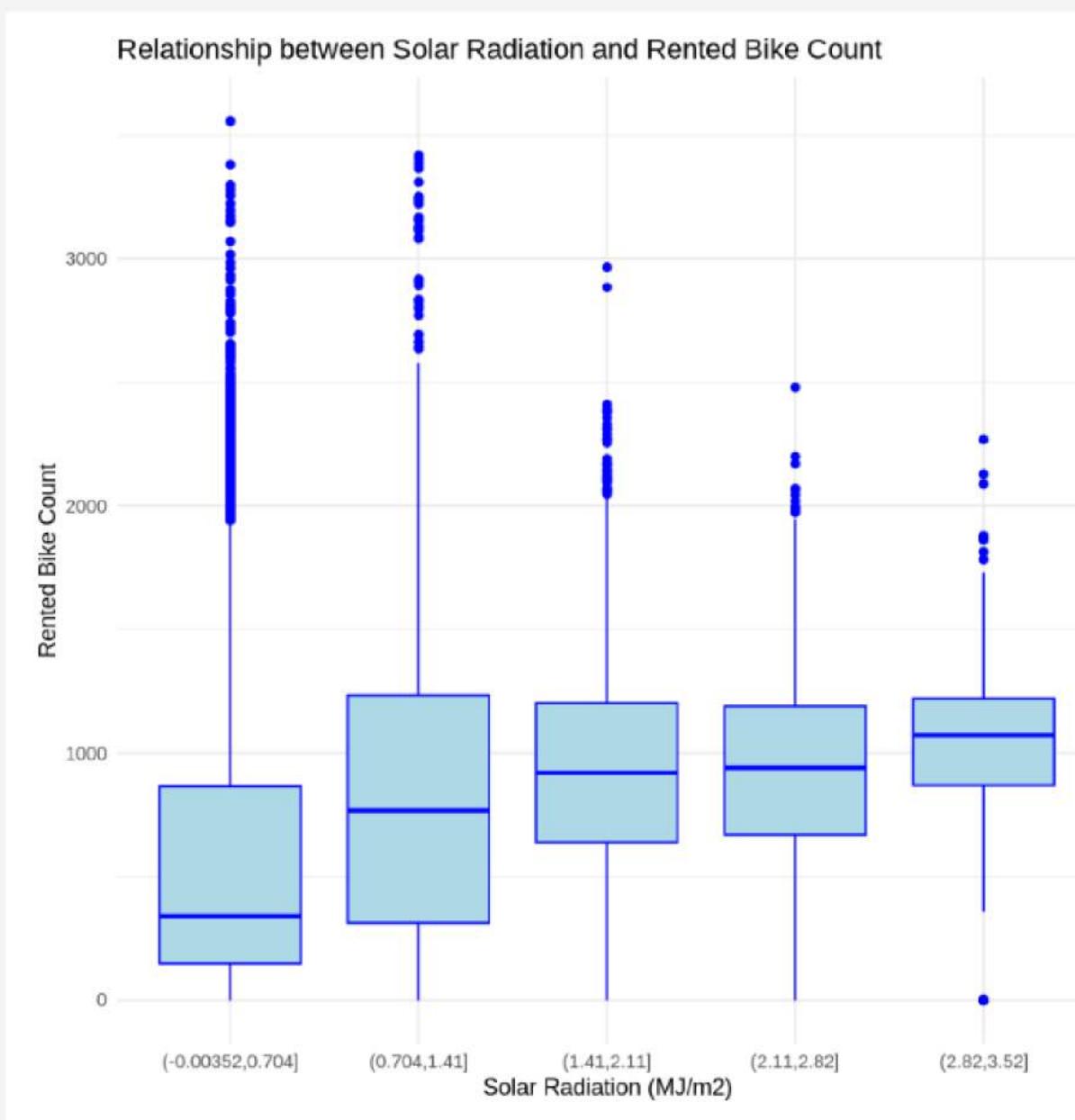
- **Ngày lễ** có số lượng thuê xe đạp trung bình nhiều hơn nhưng biến động hơn
- **Ngày không phải ngày lễ** có số lượng thuê xe đạp ít hơn nhưng ổn định hơn.

# BIỂU ĐỒ SỐ LƯỢNG XE ĐẠP ĐƯỢC THUÊ TRUNG BÌNH CHO MÔI NHIỆT ĐỘ VÀ MÙA



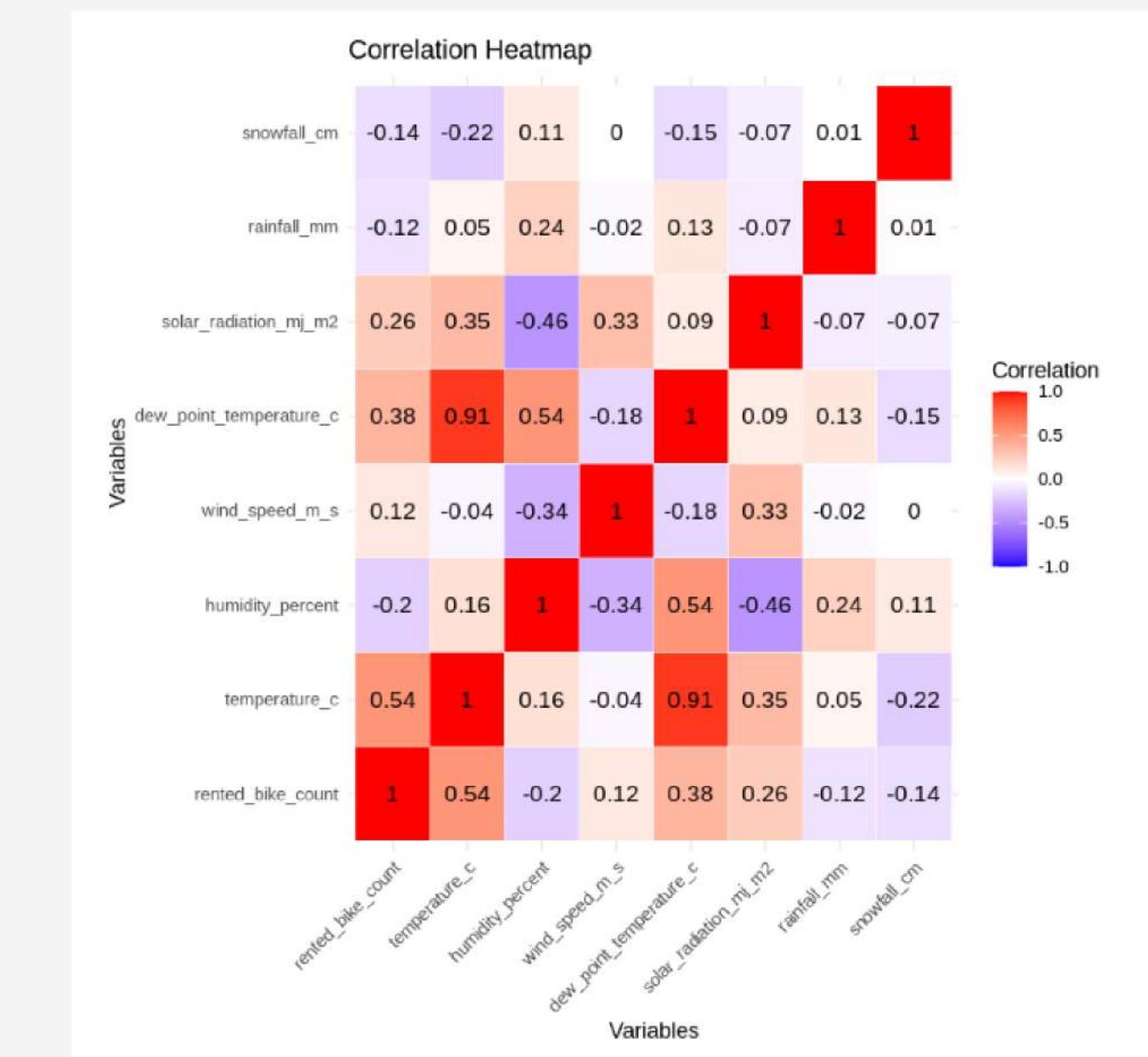
- **Nhiệt độ và lượng thuê**: Nhiệt độ từ 24.5°C đến 30.1°C thì lượng thuê cao. Dưới 17°C, lượng thuê giảm.

# BIỂU ĐỒ SỐ LƯỢNG XE ĐẠP THEO BỨC XẠ MẶT TRỜI



- Theo bức xạ mặt trời :** số lượng thuê xe trung bình tăng vào thời tiết đẹp, nắng tốt.

# BIỂU ĐỒ TƯƠNG QUAN



- Biểu đồ tương quan :** biến temperarture\_c có độ tương quan cao nhất với 0.54 với biến target

# CHIẾN LƯỢC KINH DOANH CHO DỊCH VỤ CHO THUÊ XE ĐẠP QUA TRỰC QUAN HÓA

- Tối Ưu Hóa Nguồn Cung: Điều chỉnh cung cấp theo mùa và thời tiết.
- Chiến Lược Vận Hành Hiệu Quả: Tối ưu giờ hoạt động, bảo trì xe vào giờ ít thuê.
- Chiến Dịch Tiếp Thị và Khuyến Mãi: Tiếp thị địa phương, khuyến mãi vào ngày thường và trong giờ hành chính.
- Cải Thiện Dịch Vụ: Thu thập phản hồi, mở rộng trạm xe.
- Quản Lý Dựa Trên Dữ Liệu: Giám sát thời gian thực, đánh giá và điều chỉnh chiến lược.



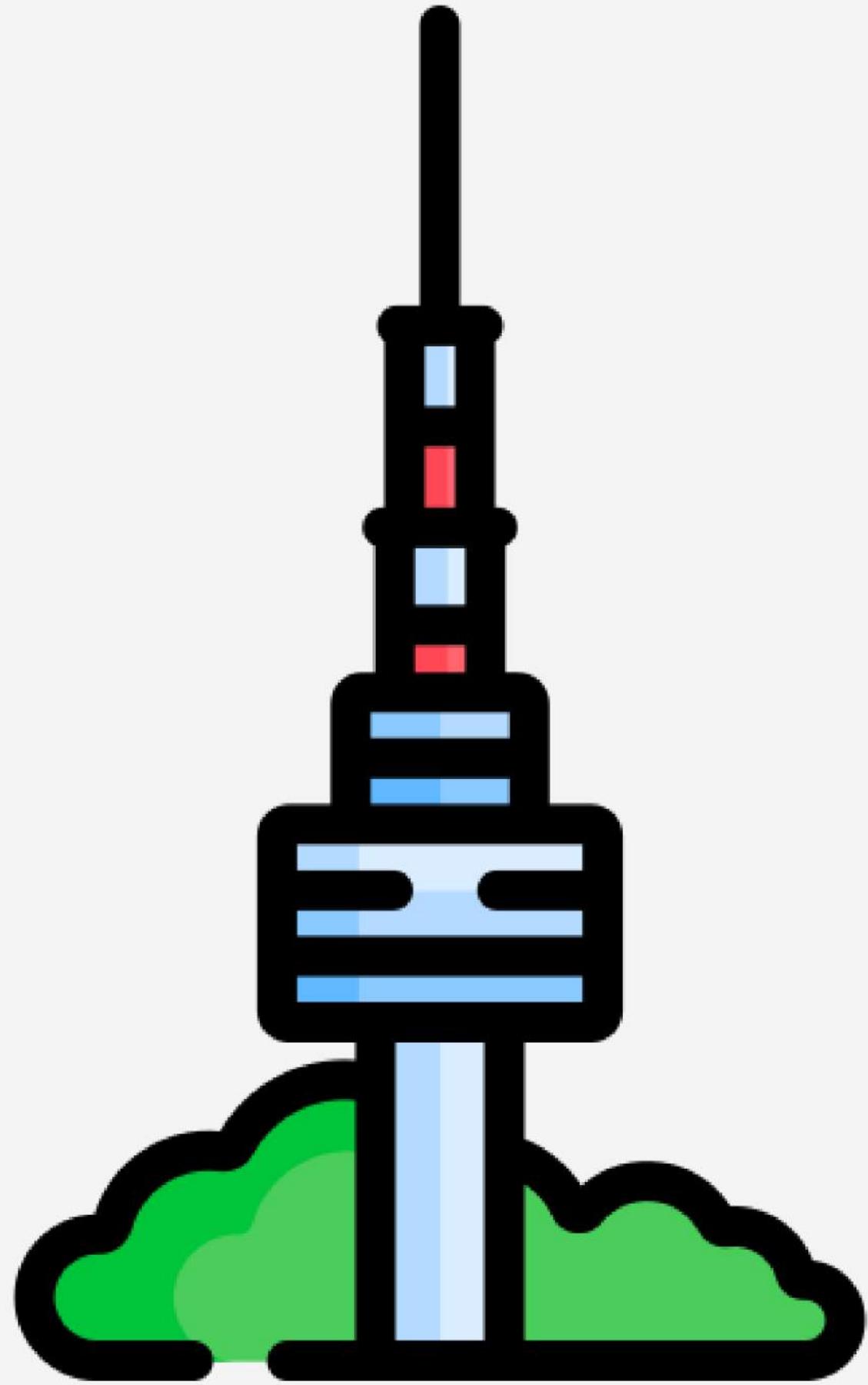
# RÚT GỌN VÀ BIẾN ĐỔI DỮ LIỆU

---

Phương pháp scale dữ liệu được sử dụng để chuẩn hóa dữ liệu cho train\_data:

- Hàm scale(train\_data): tính toán giá trị trung bình và độ lệch chuẩn của từng biến trong train\_data, sau đó chuẩn hóa từng biến theo công thức trên.
  - Kết quả là một ma trận với các giá trị đã được chuẩn hóa, có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1 cho mỗi biến.
- 





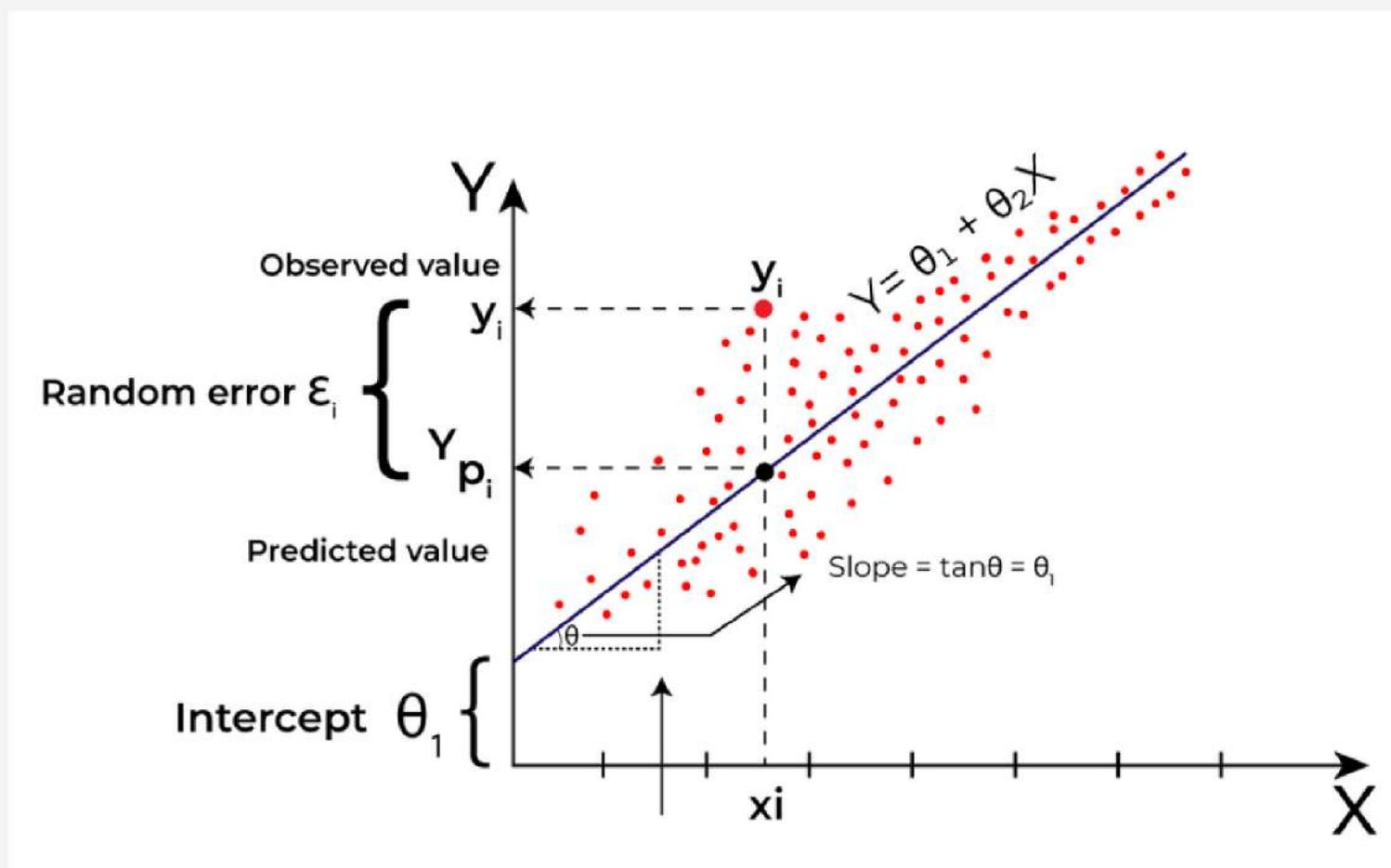
## **4. MÔ HÌNH ĐƯỢC SỬ DỤNG**

# MÔ HÌNH HỒI QUY

Mô hình hồi quy **Poisson** dựa trên trung bình:

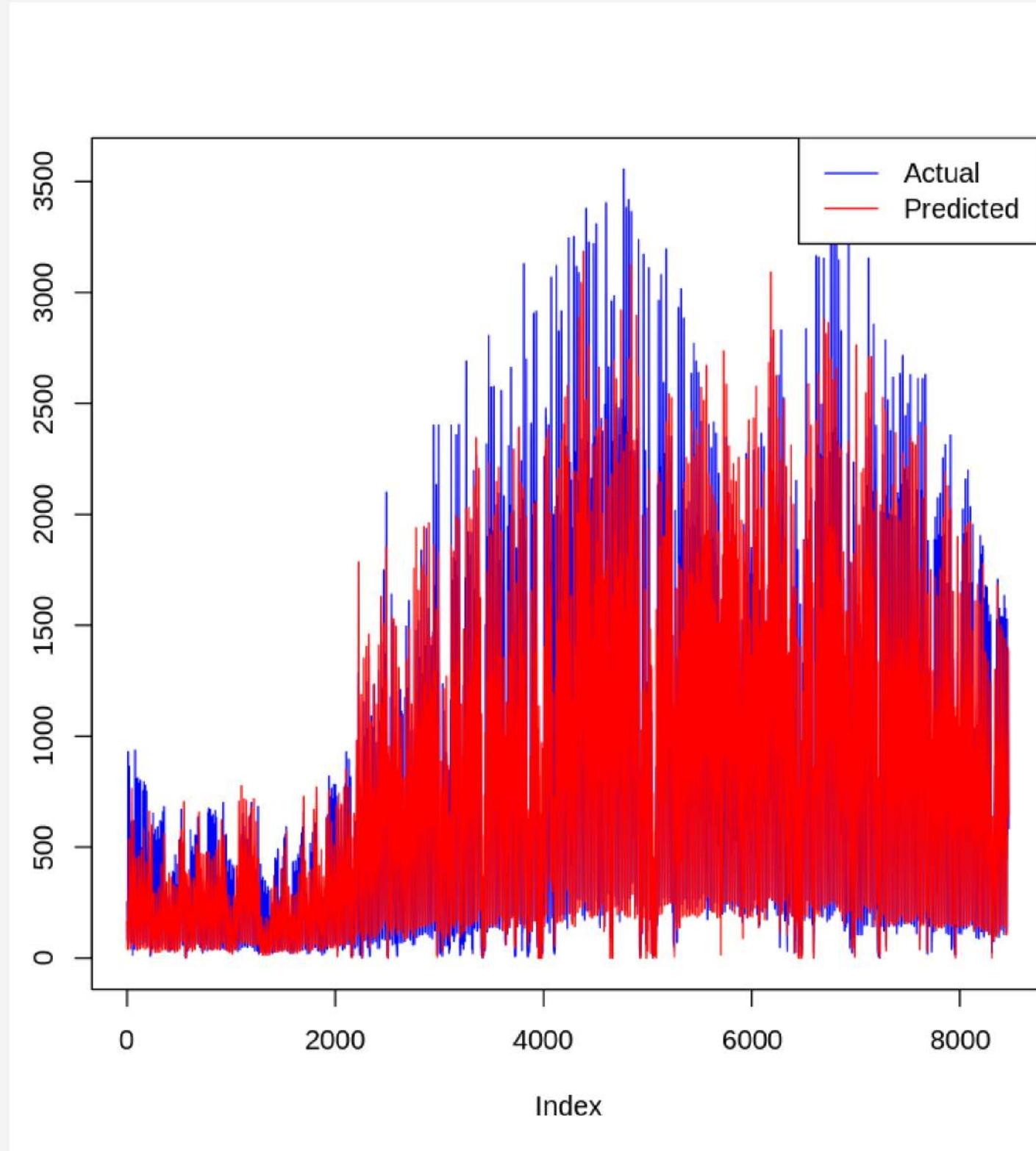
$$\lambda(x) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)$$

Rented.Bike.count ~ Temperature + Humidity + Windspeed + Visibility + Dew.point.temperature + Solar.radiation + Rainfall + Snowfall + factor(Hour) + factor(Seasons) + factor(Holiday) + factor(Functional.Day)



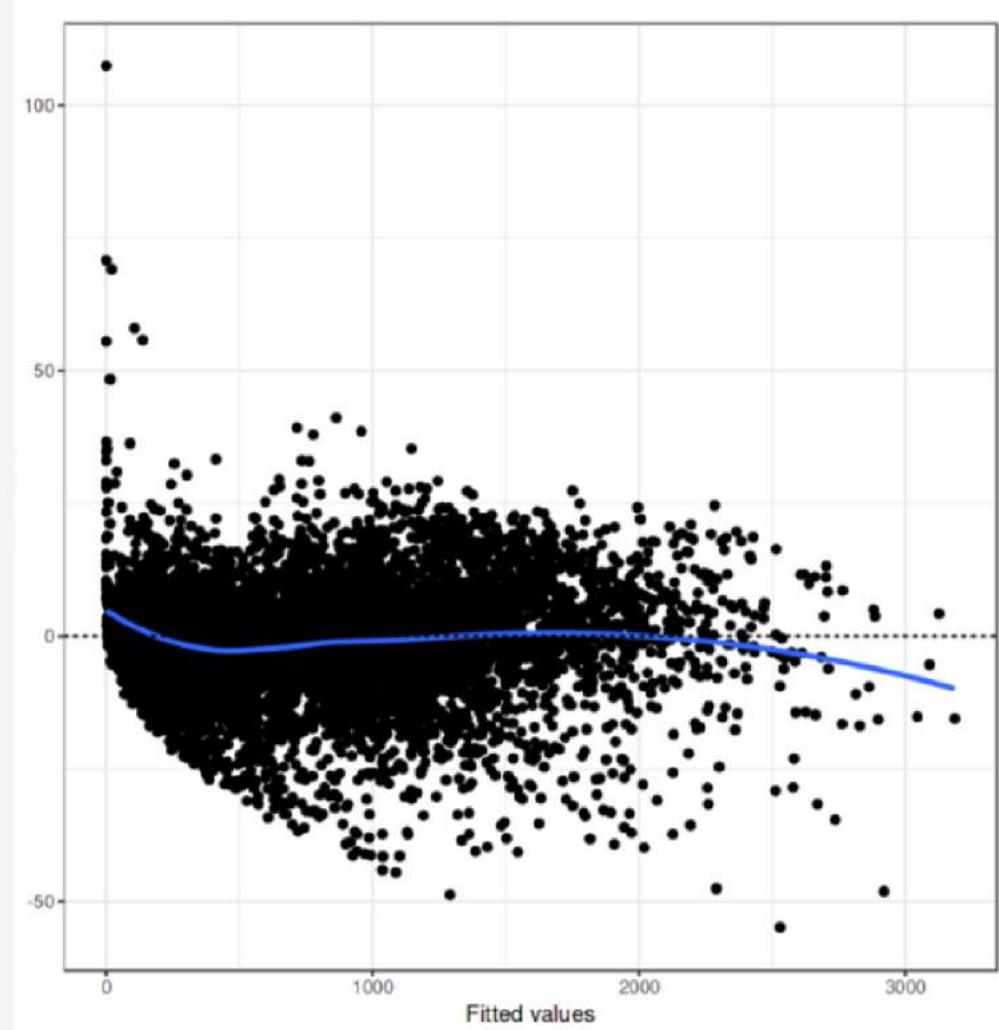
# CHUẨN ĐOÁN MÔ HÌNH

- Mean Absolute Error (MAE): 214.7058445
- Root Mean Squared Error (RMSE): 322.508723
- R-squared: 0.747890502
- Mean Absolute Percentage Error (MAPE):  
79.21346001

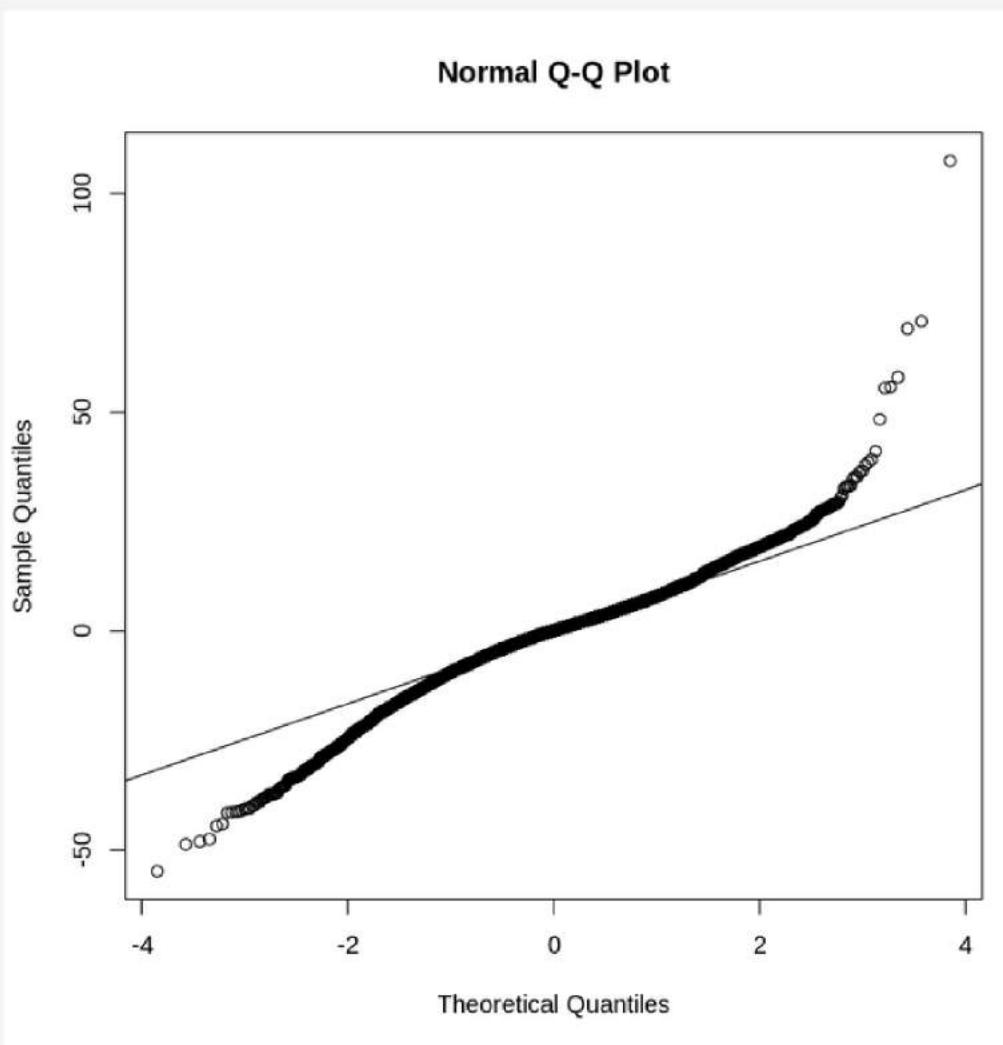


# CHUẨN ĐOÁN MÔ HÌNH

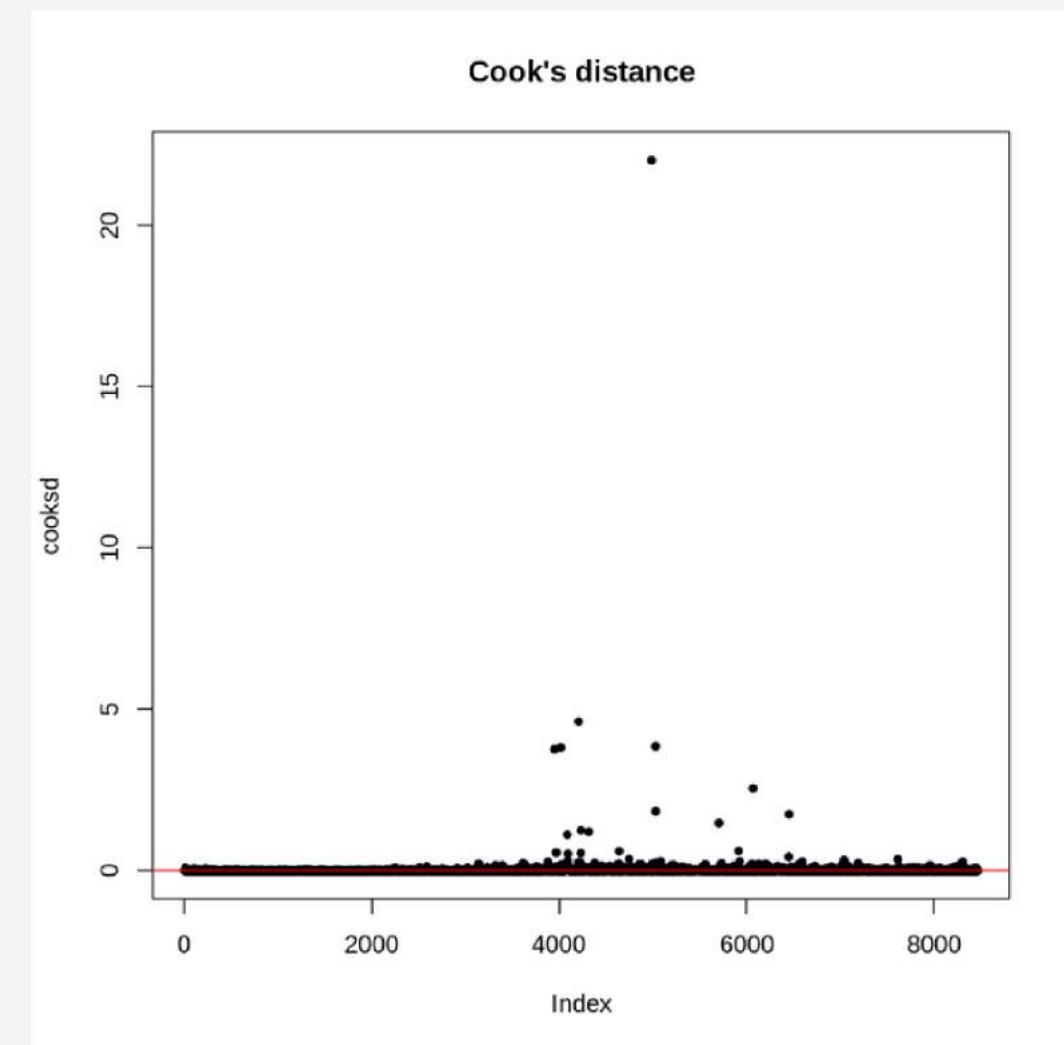
Biểu đồ phần dư và  
giá trị dự đoán



Biểu đồ Q-Q chuẩn



Biểu đồ Cook's Distance

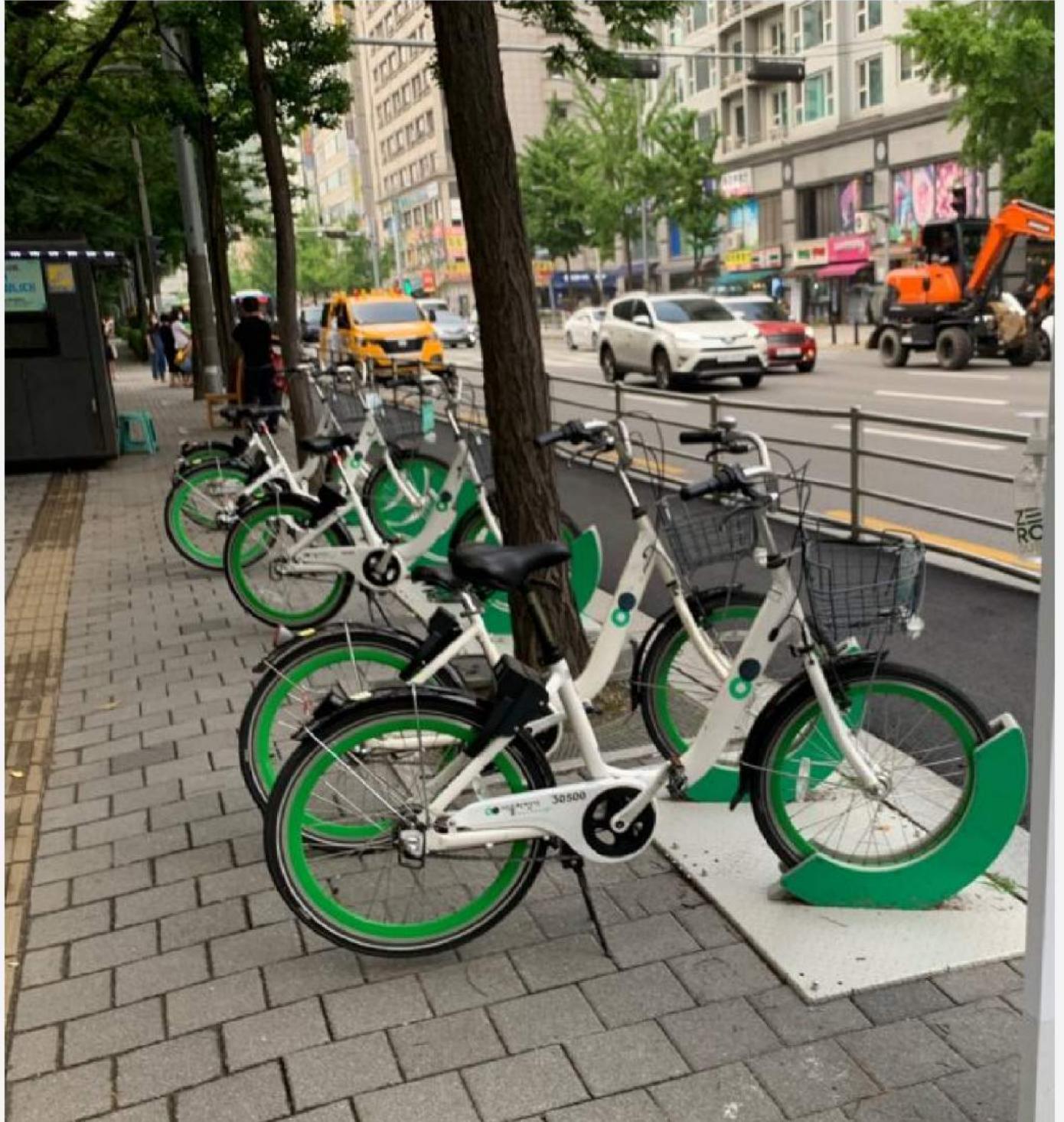


# MỞ RỘNG MÔ HÌNH

Mô hình tuyến tính tổng quát (**Generalized Linear Model - GLMs**):

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

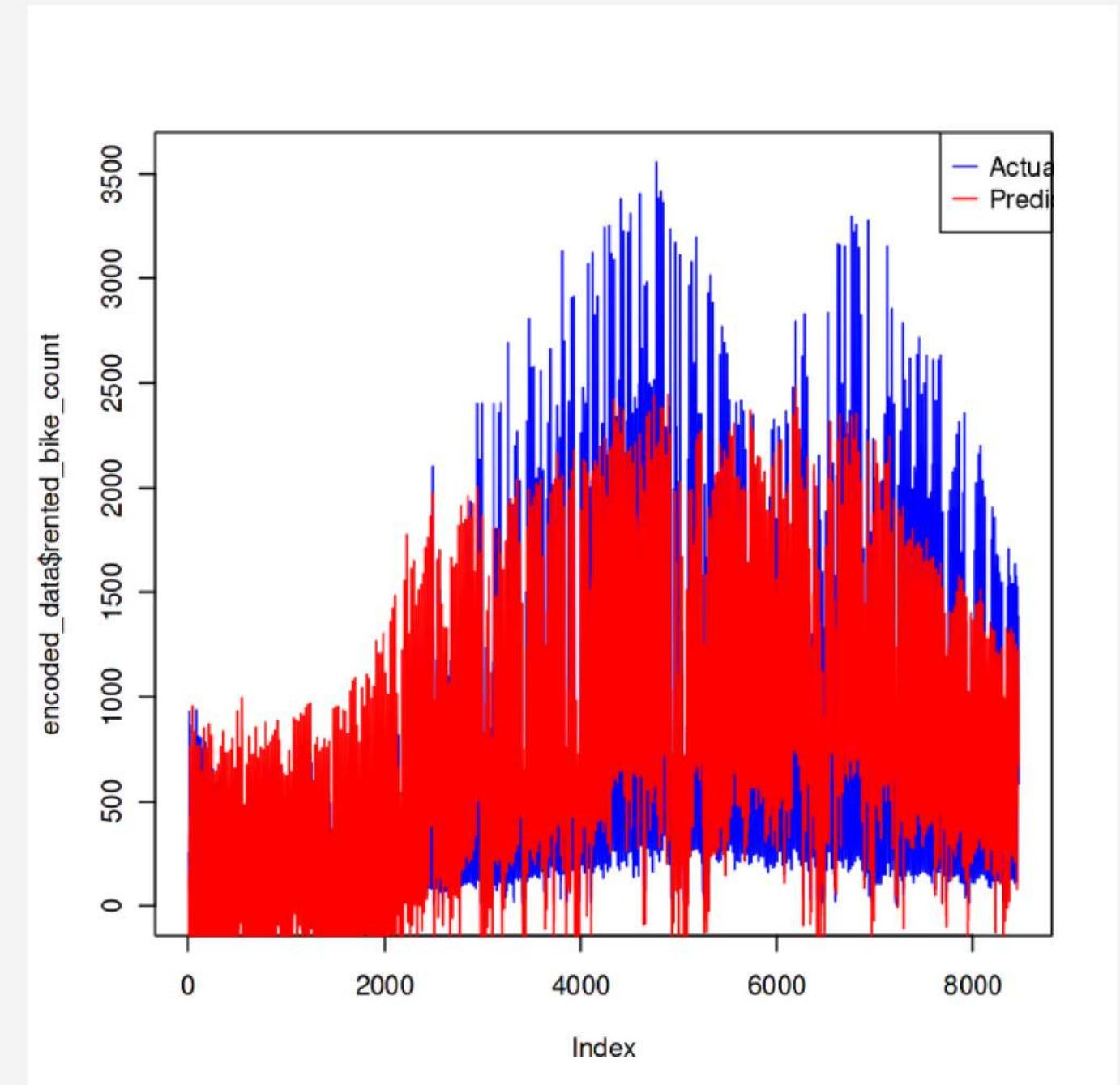
- Phân phối của biến phản hồi: Họ phân phối hàm mũ (exponential family) như chuẩn, nhị thức, Poisson, Gamma.
- Hàm liên kết (Link function): Liên kết kỳ vọng của biến phản hồi với tổ hợp tuyến tính của các biến dự đoán.
- Tổ hợp tuyến tính (Linear predictor): Tổ hợp tuyến tính của các biến dự đoán.



# MỞ RỘNG MÔ HÌNH

Mô hình Tổng quát Bổ sung (**Generalized Additive Models - GAM**) là một phần mở rộng của mô hình tuyến tính tổng quát (GLM) cho phép sự linh hoạt hơn trong việc mô hình hóa các mối quan hệ phi tuyến giữa các biến dự đoán và biến phản hồi.

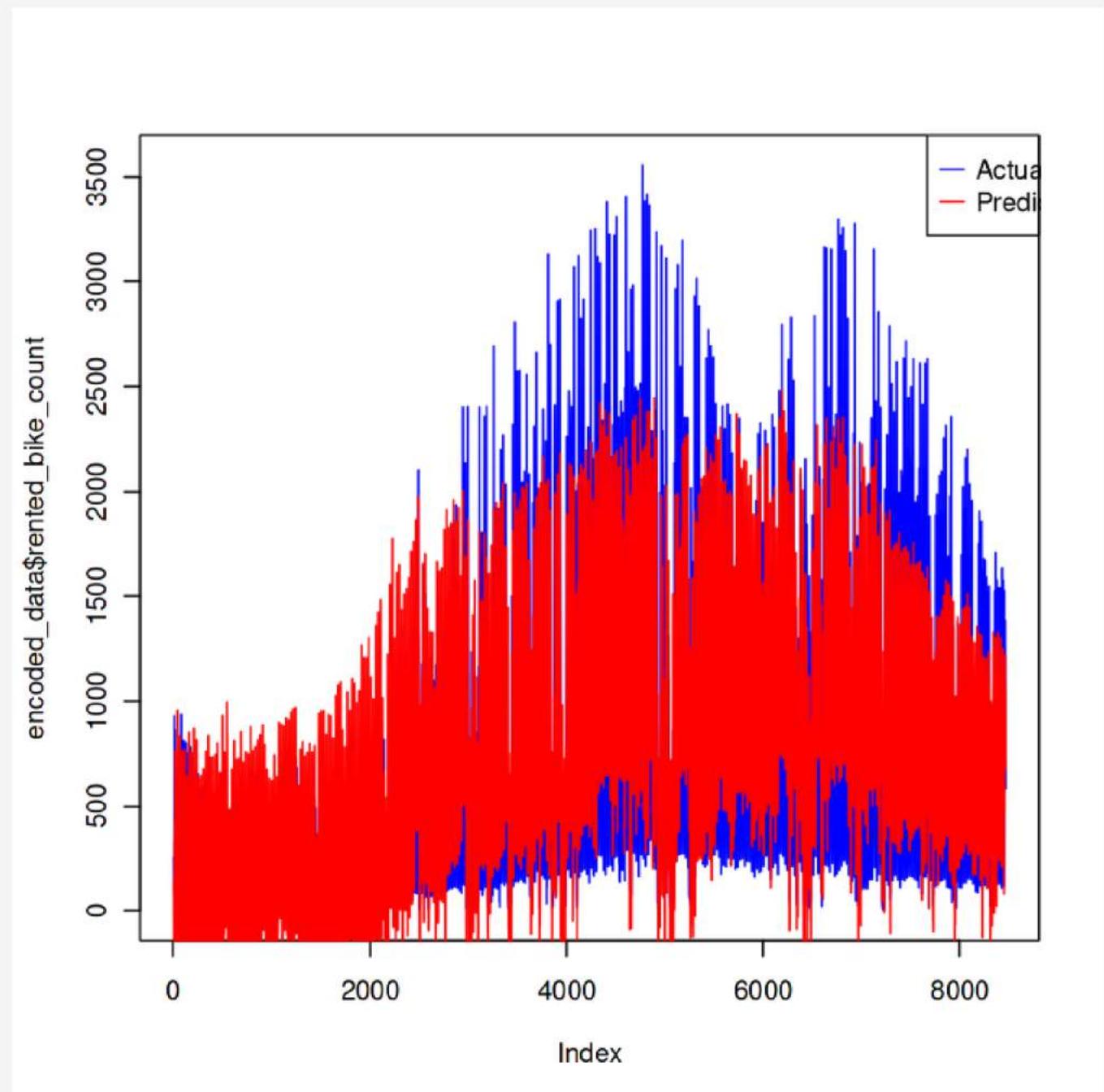
$$g(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \dots + f_p(x_{ip})$$



# MỞ RỘNG MÔ HÌNH

Kết quả thu được

- Mean Absolute Error (MAE): 230.0882713
- Root Mean Squared Error (RMSE): 309.0201
- R-squared: 0.7685
- Mean Absolute Percentage Error (MAPE): 136.27713

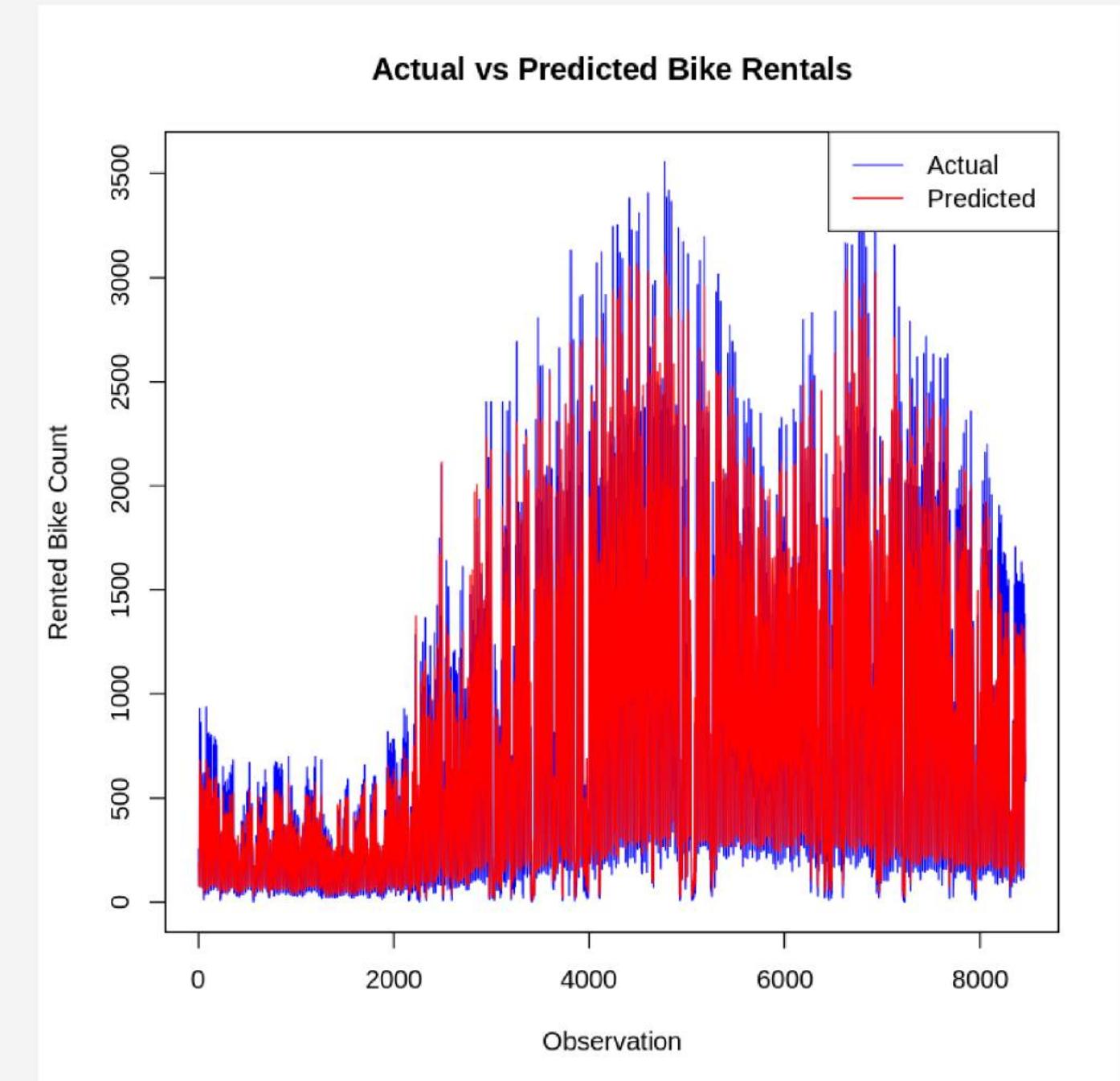


# TÌM HIỂU THÊM

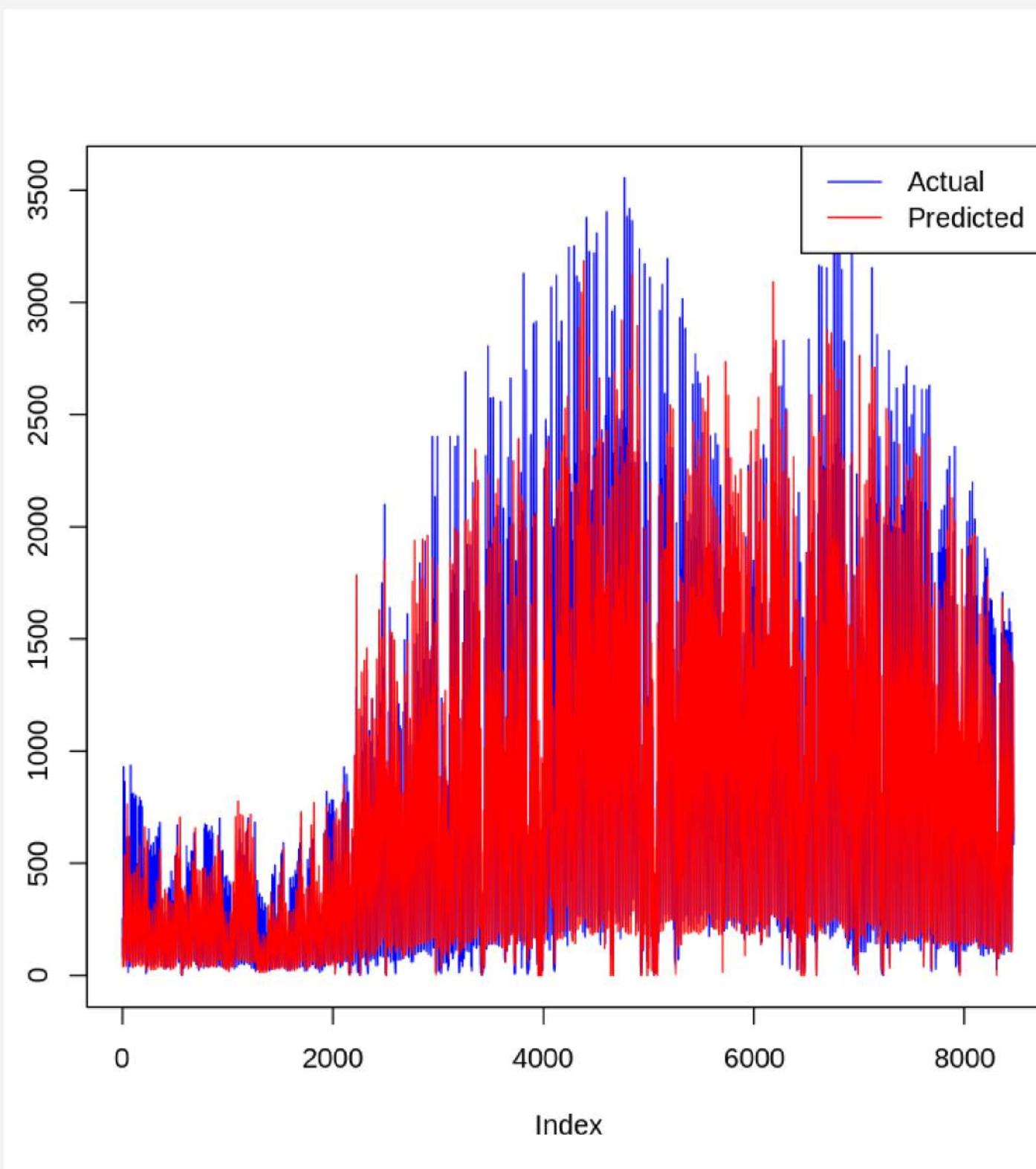
**Random Forest** là một tập hợp của nhiều cây quyết định (ensemble method), nơi mỗi cây trong rừng được huấn luyện trên một tập con của dữ liệu gốc, và dự đoán của mô hình được tính bằng cách lấy trung bình (cho hồi quy) hoặc bằng cách lấy phiếu bầu đa số (cho phân loại).

Kết quả mô hình:

- Mean Absolute Error (MAE): 74.3358212580072
- Root Mean Squared Error (RMSE):  
114.990945402978
- R-squared: 0.967949591947861
- Mean Absolute Percentage Error (MAPE):  
26.81230657925



# TÌM HIỂU THÊM



Cải tiến mô hình hồi quy poisson:

- Mean Absolute Error (MAE): 196.659802
- Root Mean Squared Error (RMSE):  
295.9415103
- R-squared: 0.787715617
- Mean Absolute Percentage Error (MAPE):  
78.3413927

# NHẬN XÉT MÔ HÌNH

Random Forest ( $R^2 = 0.97$ ) là tốt nhất.

Lựa chọn khác:

- Poisson Regression - Cải tiến ( $R^2 = 0.79$ )
- GAM ( $R^2 = 0.7685$ ).

Mô hình	R-squared
Possion Regression	0.7478
Weighted Regression	0.66
Polynomial Regression	0.66
Bsplines	0.69
GAM	0.7685
Random Forest (Tìm hiểu thêm)	0.97
Possion Regression (Cải tiến)	0.79



## 5. KẾT QUẢ ĐẠT ĐƯỢC

# KẾT QUẢ DỰ ĐOÁN:

Sử dụng mô hình tốt nhất để dự đoán số lượng xe đạp cần thuê mỗi giờ dựa trên các yếu tố khí hậu, thời gian trong ngày, mùa, ngày lễ, và ngày chức năng.



# CHIẾN LƯỢC

- Dựa trên thời gian trong ngày: Đảm bảo số lượng xe đạp đủ trong các giờ cao điểm.
- Dựa trên mùa: Điều chỉnh số lượng xe đạp theo mùa, ví dụ, tăng số lượng xe đạp vào mùa xuân và mùa hè.



- Dựa trên ngày lễ: Đảm bảo số lượng xe đạp đủ vào các ngày lễ và ngày cuối tuần.
- Dựa trên dự báo thời tiết: Điều chỉnh số lượng xe đạp dựa trên dự báo thời tiết như nhiệt độ, độ ẩm, và lượng mưa.



**CẢM ƠN THẦY VÀ CÁC BẠN  
ĐÃ LẮNG NGHE**