

STATISTICAL MACHINE LEARNING

PyTorch, Selenium, Scikit-learn, BeautifulSoupoup

COMPREHENSIVE ML & DL APPROACHES
FOR HOTEL ROOM REVIEW SCORE
PREDICTION & INFLUENZA PROPORTION
FORECASTING: REGRESSION, CLUSTERING,
CLASSIFICATION, & TIME SERIES ANALYSIS

Khang Nguyen-Hoang

HUTECH UNIVERSITY

CONTENTS

1 INTRODUCTION

2 DATA PREPROCESSING

3 PROPOSED METHODS

- PROPOSED METHODS FOR REGRESSION TASK
- PROPOSED METHODS FOR CLASSIFICATION TASK
- PROPOSED METHODS FOR TIME SERIES FORECASTING TASK

4 EXPERIMENTAL RESULTS

INTRODUCTION: THE PROBLEM AND BOOKING.COM

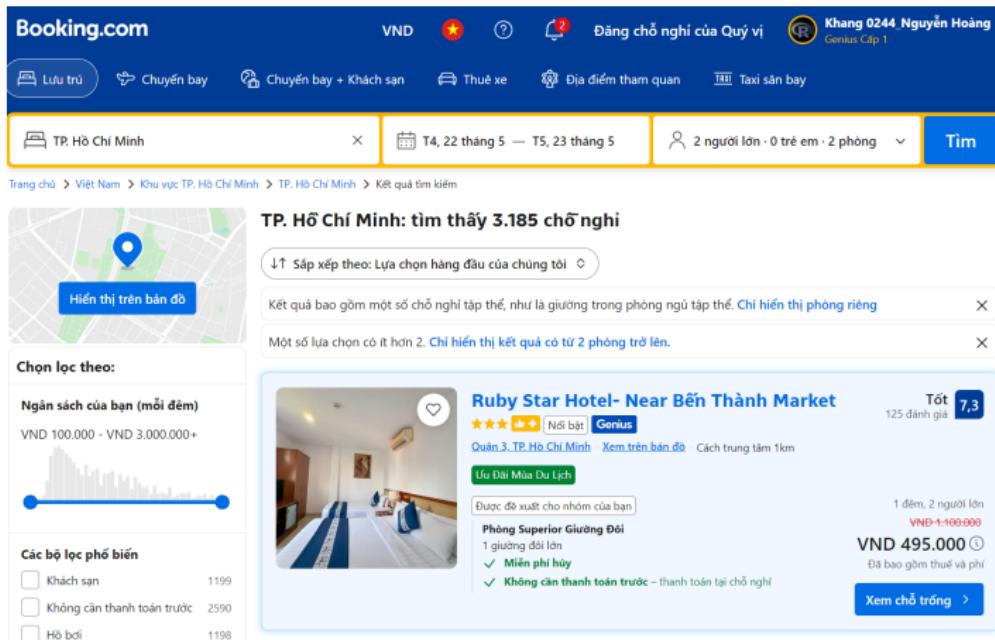


Figure 1: The interface of Booking.com

INTRODUCTION: THE PROBLEM AND BOOKING.COM

The screenshot shows a search interface for a hotel room. The search parameters are:

- Tìm chỗ nghỉ / điểm đến: TP. Hồ Chí Minh
- Ngày nhận phòng: Thứ 4, ngày 22 tháng 5 n...
- Ngày trả phòng: Thứ 5, ngày 23 tháng 5 n...
- Nghỉ 1 đêm: 2 người lớn · 0 trẻ em · 2 phòng
- Checkboxes selected: Nhà & căn hộ nguyên căn, Tôi đi công tác

A blue "Tìm" button is at the bottom.

The results page for "Language Exchange Hotel 3" includes:

- Rating: Genius
- Address: 212B C6/7 Nguyen Trai, District 1, Quận 1, TP. Hồ Chí Minh, Việt Nam
- Location map showing the hotel's proximity to Bitexco Financial Tower and Vạn Hạnh Mall.
- Room photos: A large main photo of a double bed with red curtains, and smaller thumbnail images of various rooms and interior details.
- Call-to-action buttons: "Đặt ngay" (Book now) and "Chúng Tôi Luôn Khớp Giá!" (We always match the price).

Figure 2: booking.com

INFLUENZA PROPORTION DATASET

1	date	start_date	end_date	REGION TYPE	REGION	YEAR	WEEK	% WEIGHTED ILI	OT
2	1997-09-29	1997-09-29	1997-10-05	National	X	1997	40	1.10148	1.21686
3	1997-10-06	1997-10-06	1997-10-12	National	X	1997	41	1.20007	1.28064
4	1997-10-13	1997-10-13	1997-10-19	National	X	1997	42	1.37876	1.23906
5	1997-10-20	1997-10-20	1997-10-26	National	X	1997	43	1.1992	1.14473
6	1997-10-27	1997-10-27	1997-11-02	National	X	1997	44	1.65618	1.26112
7	1997-11-03	1997-11-03	1997-11-09	National	X	1997	45	1.41326	1.28275
8	1997-11-10	1997-11-10	1997-11-16	National	X	1997	46	1.9868	1.44579
9	1997-11-17	1997-11-17	1997-11-23	National	X	1997	47	2.44749	1.64796
10	1997-11-24	1997-11-24	1997-11-30	National	X	1997	48	1.73901	1.67517
11	1997-12-01	1997-12-01	1997-12-07	National	X	1997	49	1.93919	1.61739
12	1997-12-08	1997-12-08	1997-12-14	National	X	1997	50	2.28166	1.76632
13	1997-12-15	1997-12-15	1997-12-21	National	X	1997	51	2.50761	2.10398
14	1997-12-22	1997-12-22	1997-12-28	National	X	1997	52	4.11353	2.43655

THE RAW DATA (2153×12)

	link	hotel_name	review_score	review_count	price	address	image	facilities	subscore	star
	https://www.booking.com/hotel/vn/cactos-bar-am...	Cactos Bar & Homestay	NaN		VND 135.000	Quận 12	https://cf.bstatic.com/xdata/images/hotel/squa...	[2 hồ bơi, Phòng giá đình, Phòng không h...	(Nhân viên phục vụ: 10.0, Tiện nghi: 10.0,...	0
	https://www.booking.com/hotel/vn/saigon-city-c...	Saigon City Center Hostel	6,4	270.0	VND 170.000	Quận 1	https://cf.bstatic.com/xdata/images/hotel/squa...	[2 hồ bơi, Phòng giá đình, Lễ tân 24h, ...	(Nhân viên phục vụ: 7.3, Tiện nghi: 6.5,...	0
	https://www.booking.com/hotel/vn/sh-capsule-3...	Start House CAPSULE 5 min walk to the Airport	8,5	156.0	VND 180.000	Quận Tân Bình	https://cf.bstatic.com/xdata/images/hotel/squa...	[2 hồ bơi, Chỗ để xe miễn phí, Lễ tân...	(Nhân viên phục vụ: 9.6, Tiện nghi: 8.2,...	0
	https://www.booking.com/hotel/vn/kim-hostel-qu...	Kim Hostel	3,0	1.0	VND 200.000	Quận 1	https://cf.bstatic.com/xdata/images/hotel/squa...	[2 hồ bơi, Chỗ để xe miễn phí, Bữa sá...	(Nhân viên phục vụ: 7.5, Tiện nghi: 6.3,...	0
	https://www.booking.com/hotel/vn/kim-hostel-ho...	KIM HOSTEL	4,0	3.0	VND 200.000	Quận 1	https://cf.bstatic.com/xdata/images/hotel/squa...	Nan	(Nhân viên phục vụ: 5.8, Tiện nghi: 5.0,...	0

Figure 4: The raw data is crawled from Booking.com

THE PREPROCESSED DATA (1477×20)

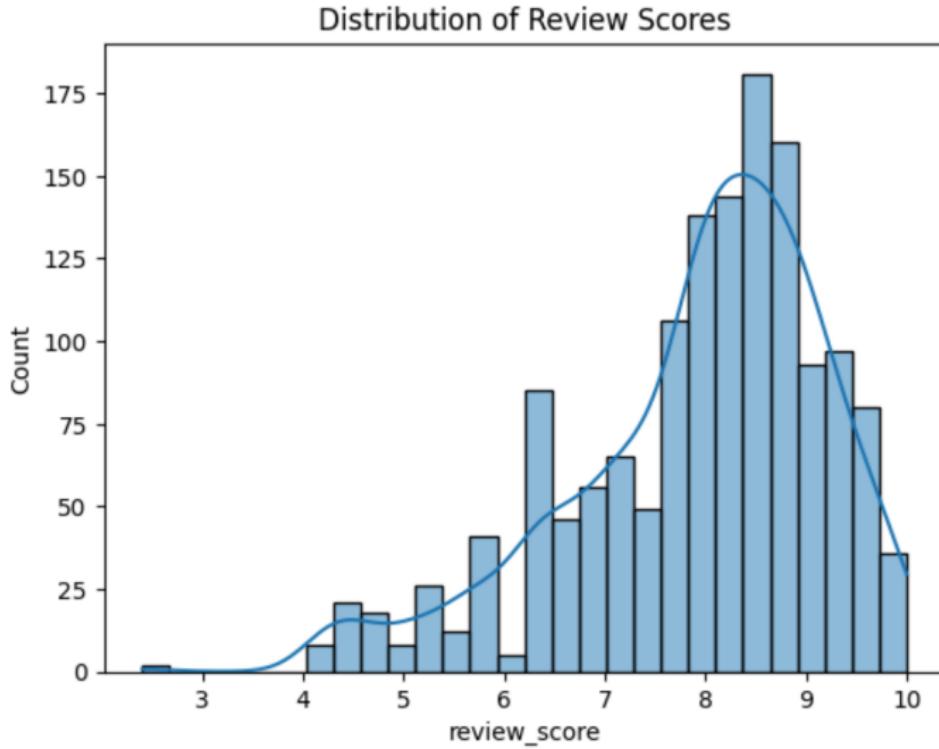
review_score	review_count	price	address	star	Pool	No_smoking_room	Families_room	Room_service	Free_parking	Breakfast	24h_front_desk	Airport_shuttle	Staff	Facilities
6.4	270	170000	Quận 1	0	1	0	1	1	0	0	0	1	1	1
8.5	156	180000	Quận Tân Bình	0	1	0	0	1	1	1	1	1	1	1
7.0	18	220000	Quận 1	0	1	0	1	1	1	1	0	0	1	1
8.0	274	230000	Quận 1	0	1	0	1	1	1	0	0	0	0	0
7.2	178	250000	Quận Bình Thạnh	0	1	1	1	0	1	1	1	1	1	0

```

review_score      1477 non-null   float64
review_count     1477 non-null   int64
price            1477 non-null   int64
address          1477 non-null   category
star             1477 non-null   int64
Pool              1477 non-null   category
No_smoking_room  1477 non-null   category
Families_room    1477 non-null   category
Room_service     1477 non-null   category
Free_parking     1477 non-null   category
Breakfast         1477 non-null   category
24h_front_desk   1477 non-null   category
Airport_shuttle  1477 non-null   category
Staff             1413 non-null   float64
Facilities        1413 non-null   float64

```

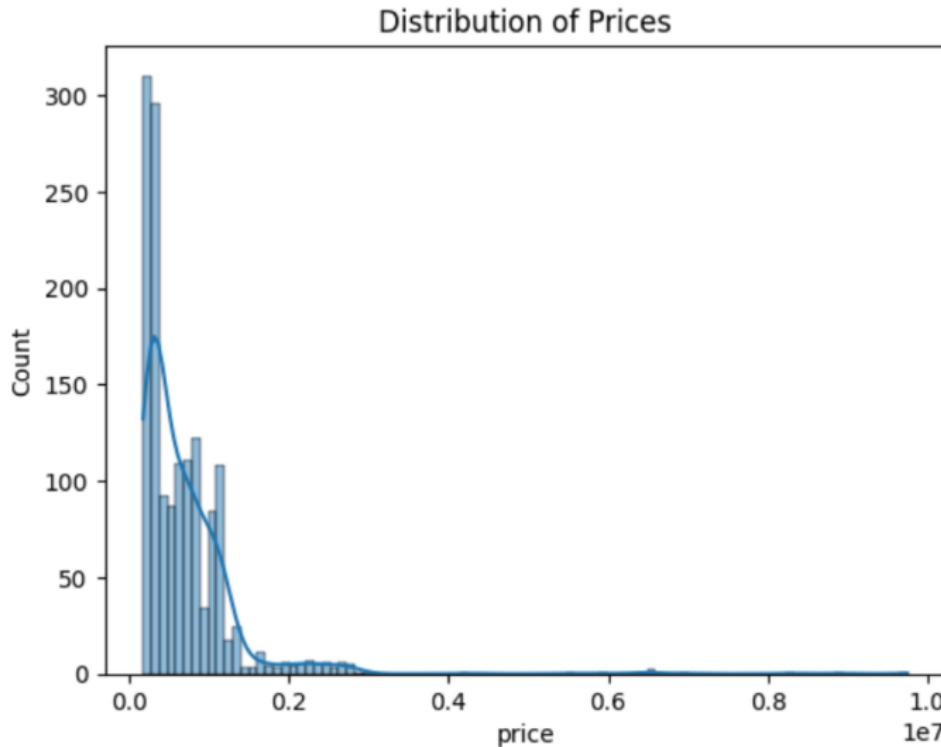
VISUALIZE THE DISTRIBUTION OF NUM-COLS



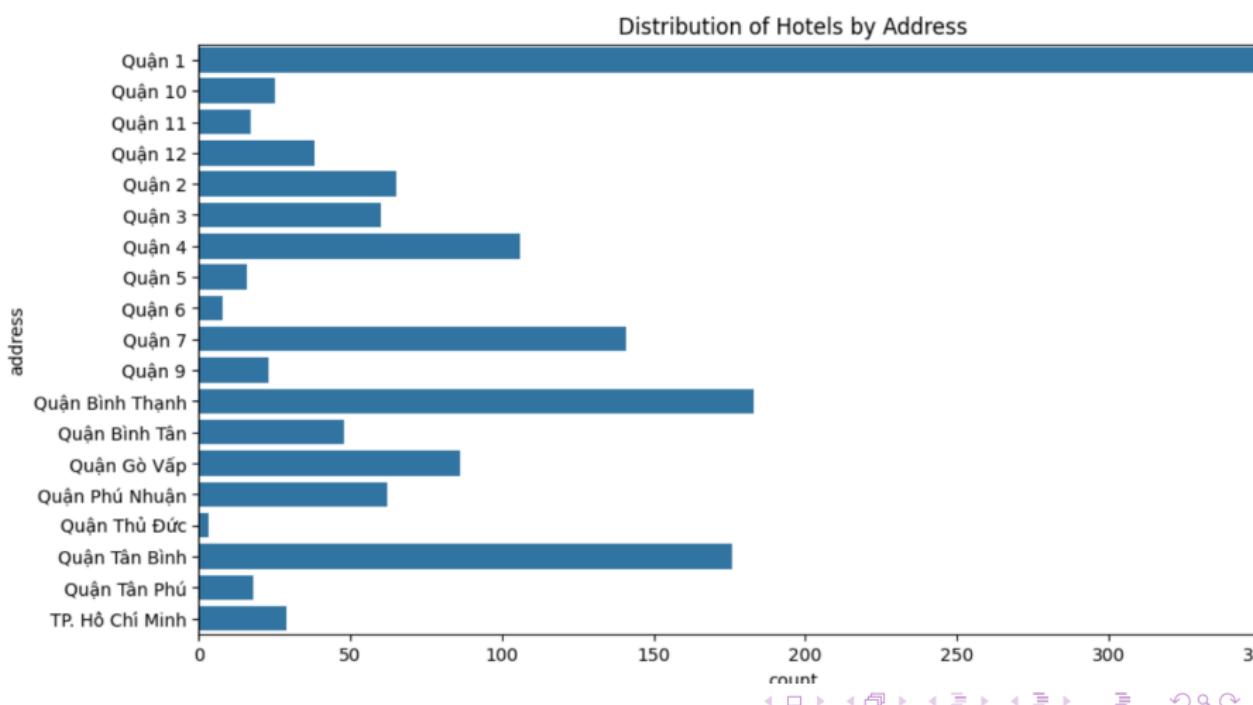
OVERVIEW ABOUT THE DATA

	review_score	review_count	price	star	Staff	\			
count	1477.000000	1477.000000	1.477000e+03	1477.000000	1413.000000				
mean	7.916858	143.483412	6.881338e+05	2.033852	8.620878				
std	1.291086	175.637038	7.101059e+05	1.615119	0.991670				
min	2.400000	5.000000	1.700000e+05	0.000000	3.100000				
25%	7.200000	22.000000	2.800000e+05	0.000000	8.100000				
50%	8.100000	69.000000	5.644300e+05	2.000000	8.800000				
75%	8.800000	199.000000	8.644300e+05	3.000000	9.300000				
max	10.000000	987.000000	9.750000e+06	5.000000	10.000000				
Facilities Cleanliness Comfort Affordable Location \									
count	1413.000000	1413.000000	1413.000000	1413.000000	1413.000000				
mean	7.979193	8.126115	8.172187	8.31005	8.577849				
std	1.256493	1.245276	1.166657	1.07331	0.934686				
min	3.100000	3.100000	3.800000	3.800000	3.800000				
25%	7.200000	7.300000	7.300000	7.60000	8.000000				
50%	8.200000	8.300000	8.300000	8.50000	8.800000				
75%	8.900000	9.100000	9.100000	9.10000	9.300000				
max	10.000000	10.000000	10.000000	10.00000	10.000000				
Wifi									
count	1012.000000								
mean	8.421344								
std	1.393830								
min	2.500000								
25%	7.500000								
50%	8.800000								
75%	9.700000								
max	10.000000								
address Pool No_smoking_room Families_room Room_service Free_parking Breakfast 24h_front_desk Airport_shuttle									
count	1477	1477	1477	1477	1477	1477	1477	1477	1477
unique	19	2	2	2	2	2	2	2	2
top	Quận 1	1	0	1	1	1	0	1	0
freq	373	1221	796	1146	851	1062	1200	952	996

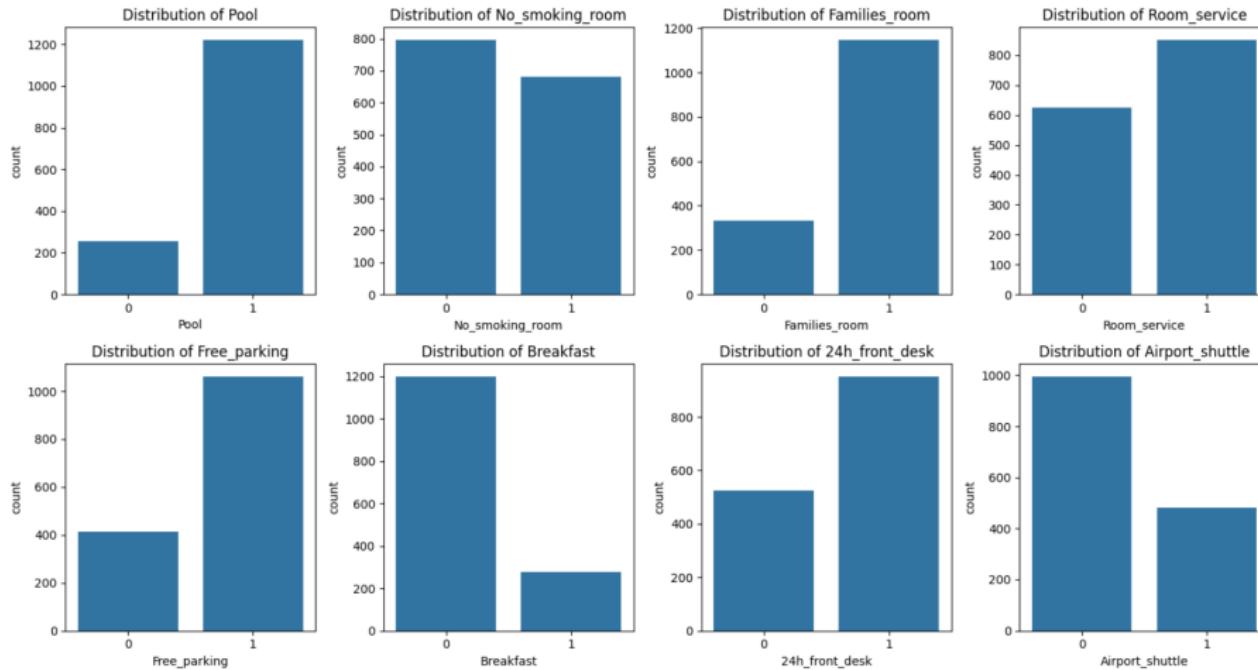
VISUALIZE THE DISTRIBUTION OF NUM-COLS



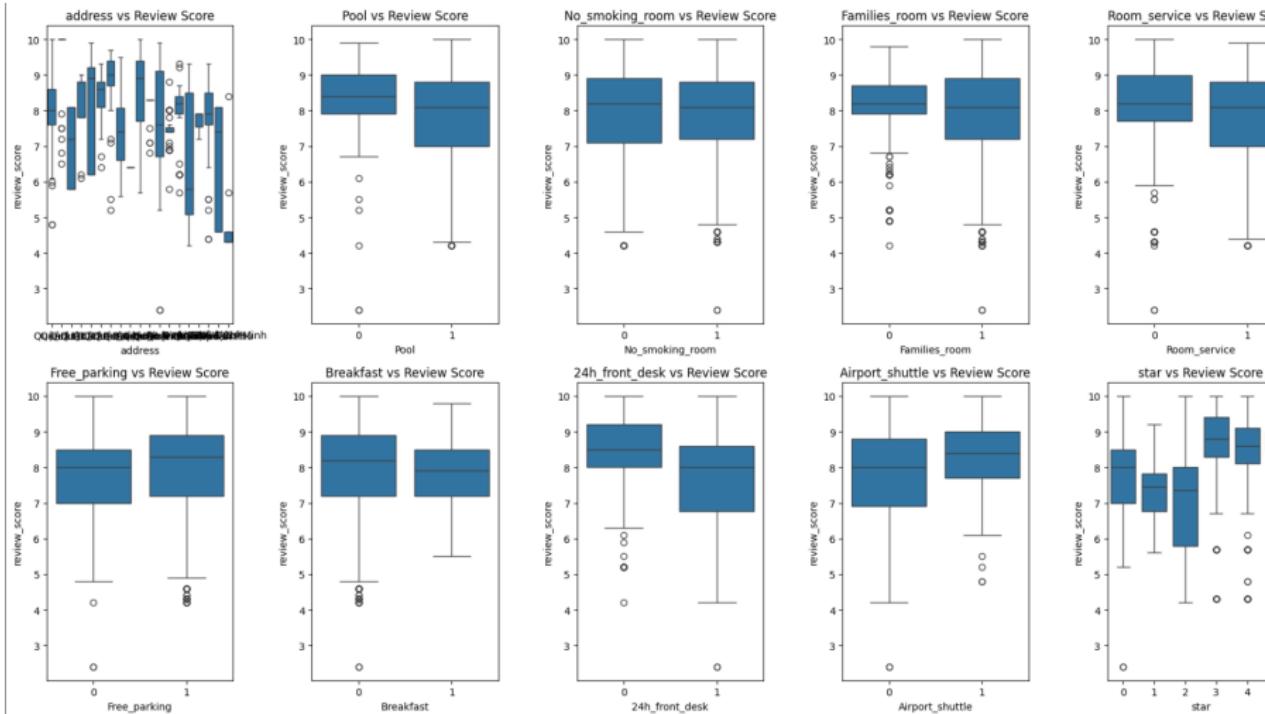
EXPLORE CATEGORICAL VARIABLES



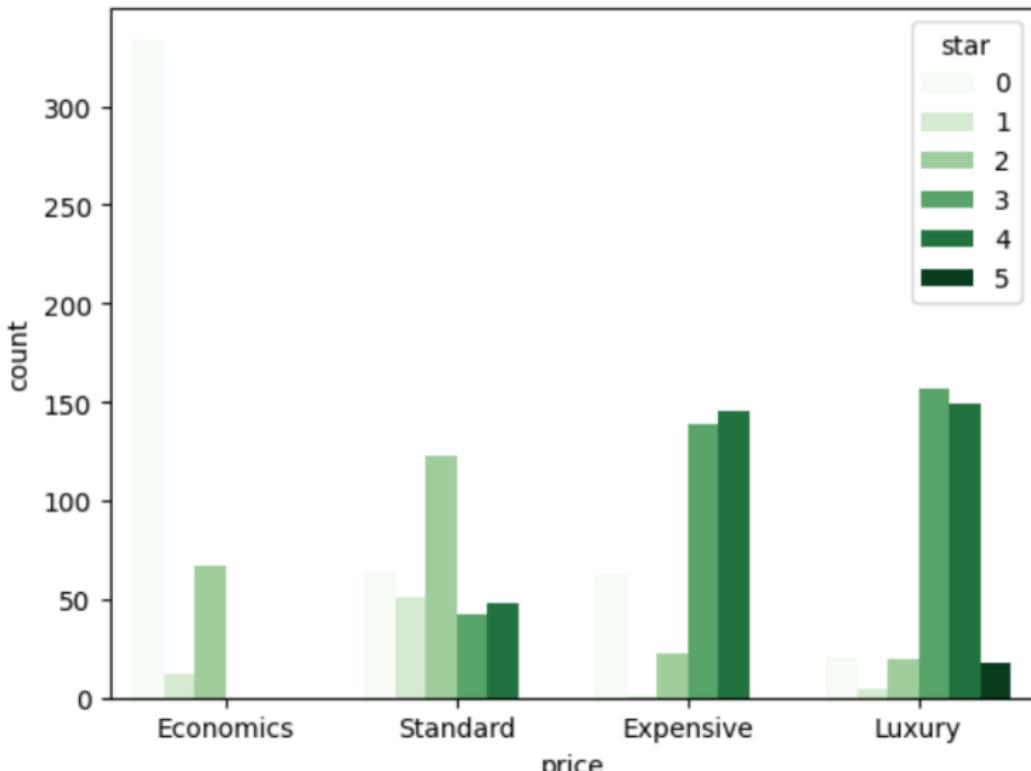
EXPLORE CATEGORICAL VARIABLES



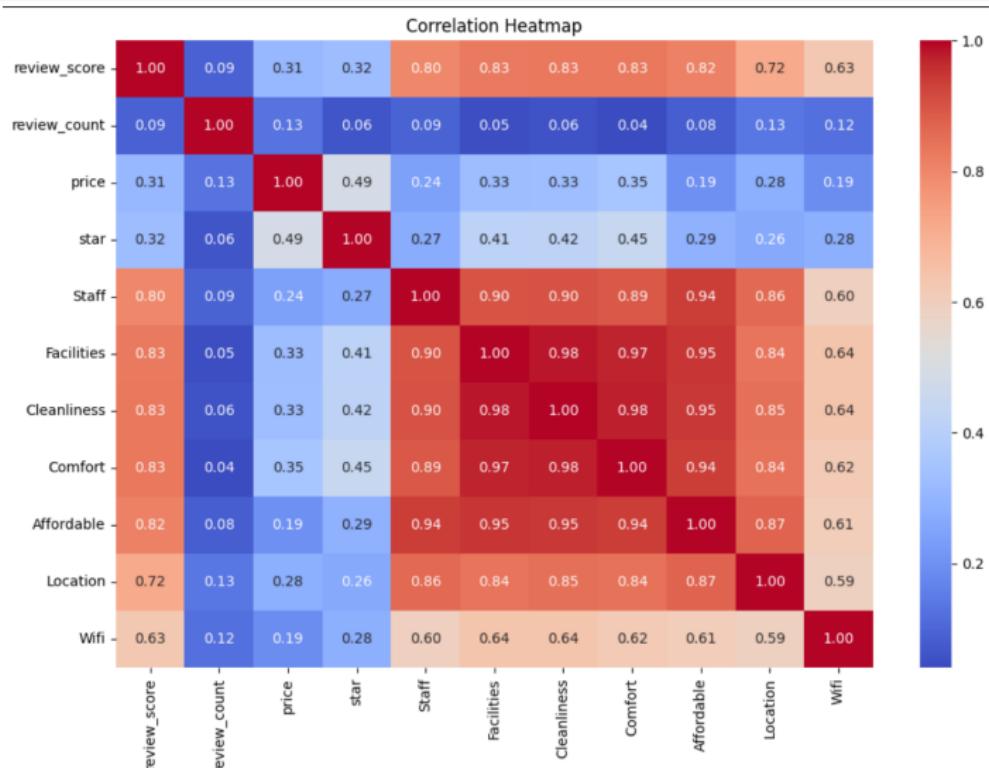
EXPLORE CATEGORICAL VARIABLES



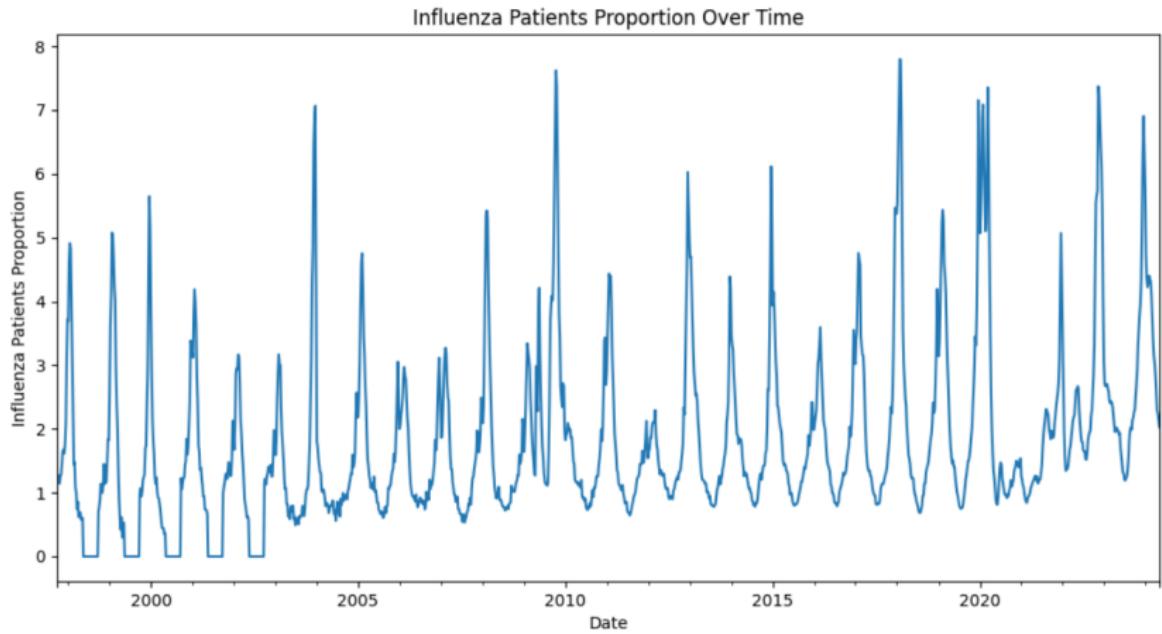
EXPLORE CATEGORICAL VARIABLES



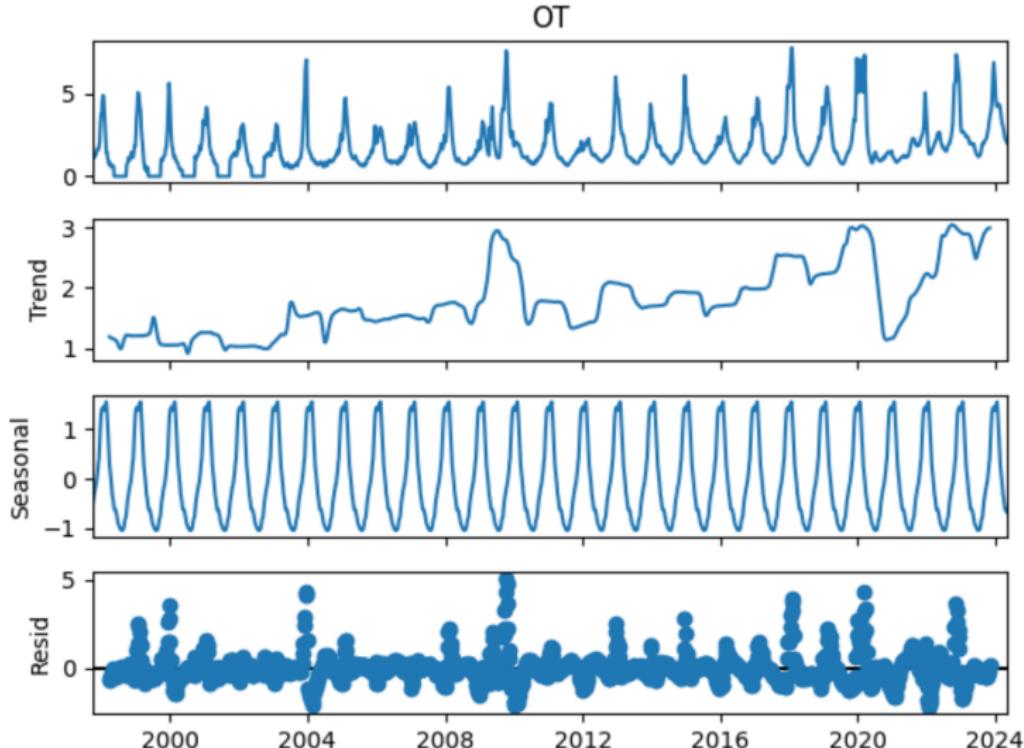
CORRELATION HEATMAP FOR NUMERIC-COLS



INFLUENZA PROPORTION OVER TIME PLOT



DECOMPOSITION



Stationarity Test and Model Metrics

Stationarity Test on Original Data

- **ADF Statistic:** -8.877553258384136, **p-value:** 1.3451218329750162e-14
- **KPSS Statistic:** 1.253054251563682, **p-value:** 0.01

Stationarity Test After First-Order Differencing

- **ADF Statistic:** -10.117359464526357, **p-value:** 9.645353311405768e-18
- **KPSS Statistic:** 0.004671689966883386, **p-value:** 0.1

LINEAR REGRESSION

Loss:

$$\mathcal{L}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y) \sim \hat{p}_{data}} [\|\mathbf{x} \cdot \mathbf{w} - y\|^2] = \mathbb{E}_{(\mathbf{x},y) \sim \hat{p}_{data}} [\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2]$$

Normal Equation: $\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{X}^\top \mathbf{y}$

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\text{RMSE} = \sqrt{\mathbb{E}_{\mathbf{y} \sim \mathbb{P}(Y|\mathbf{X})} [(\mathbf{y} - \hat{\mathbf{y}})^2]}$$

Ridge Regression: $\mathcal{L}(\mathbf{w}) = \mathbb{E}_{\mathbf{y} \sim \mathbb{P}(Y|\mathbf{X})} [(\mathbf{y} - \mathbf{X}\mathbf{w})^2] + \lambda \|\mathbf{w}\|^2$

$$\text{VIF}_j = \frac{1}{1-R_j^2}$$

VIF ANALYSIS AFTER REMOVING MULTICOLLINEARITY

Feature	VIF
Free_parking	3.6386
24h_front_desk	3.3036
star	3.2864
Room_service	2.6647
No_smoking_room	2.0670
Airport_shuttle	1.9207
Breakfast	1.4637
price	1.4609
Facilities	1.4362
review_count	1.1065

REGRESSION MODEL INTERPRETATION

$$\text{SSR} = \sum_{i=1}^n (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^2$$

$$\text{SST} = \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})^2$$

$$\text{SSE} = \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2$$

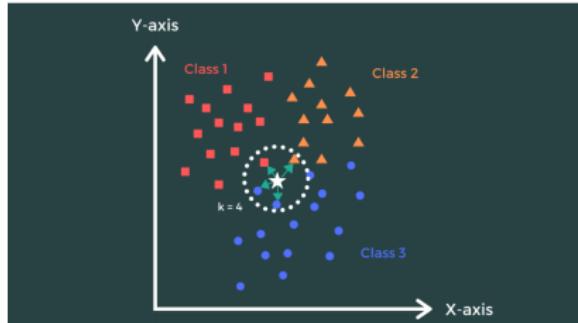
$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$\text{Adjusted } R^2 = 1 - \left(\frac{\text{SSE}/(n - p - 1)}{\text{SST}/(n - 1)} \right) = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1} \right)$$

TRANSFORM REVIEW _ SCORE AND KNN

$$f(\text{score}) = \begin{cases} 0 & \text{if } < 4 \\ 1 & \text{if } 4 \leq \text{score} < 7 \\ 2 & \text{if } 7 \leq \text{score} < 9 \\ 3 & \text{if } \geq 9 \end{cases}$$

$$\hat{y} = \arg \max_{c \in \mathcal{C}} \sum_{i \in N_k(\mathbf{x})} \mathbb{I}(y_i = c)$$



TBA: DECISION TREE & RANDOM FOREST

$$\text{Shannon Entropy} = \mathbb{E}_{x \sim \mathbb{P}(X)} [-\log \mathbb{P}(x)] = - \sum_{i=1}^n \mathbb{P}(x_i) \log \mathbb{P}(x_i)$$

$$\text{Information Gain}(X) = \text{Entropy}(Y) - \sum_{i=1}^k \frac{|Y_i|}{|Y|} \text{Entropy}(Y_i)$$

$$\mathbb{E}_{\mathbb{P}(\text{sample})} [\text{Error}_{\text{Bagging}}] = \frac{1}{M} \sum_{m=1}^M \text{Error}_m$$

MULTI-CLASS S.V.M

- Score matrix: $\mathbf{W}^\top \mathbf{X} + \mathbf{b} \triangleq \mathbf{S} \in \mathbb{R}^{C \times N}$
- Score matrix (sử dụng *bias trick*): $\mathbf{S} = \mathbf{W}^\top \mathbf{X}$
- Loss tại quan sát thứ i :

$$\mathcal{L}^{(i)}(\mathbf{W}; \mathbf{x}^{(i)}, y^{(i)}) = \sum_{j=1, j \neq y^{(i)}}^C \max(0, s_j^{(i)} - (s_{y^{(i)}}^{(i)} - \nu))$$

- Loss function:

$$\mathcal{J}(\mathbf{W}; \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1, j \neq y^{(i)}}^C \max(0, \nu - s_j^{(i)} - s_{y^{(i)}}^{(i)})$$

- Loss function (có sử dụng *weight decay*):

$$\tilde{\mathcal{J}}(\mathbf{W}; \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1, j \neq y^{(i)}}^C \max(0, \nu - s_j^{(i)} - s_{y^{(i)}}^{(i)}) + \frac{\alpha}{2} \|\mathbf{W}\|_{\mathcal{F}}^2$$

LOGISTIC REGRESSION

$$\text{Cross Entropy} = \mathbb{E}_{y \sim \mathbb{P}(Y)} [-\log \mathbb{P}(\hat{y}|y)] = -\mathbb{E}_{\mathbf{y} \sim \mathbb{P}(Y)} \left[\sum_{i=1}^n \mathbf{y}_i \log \hat{\mathbf{y}}_i \right]$$

$$\mathcal{L}(\mathbf{w}) = -\mathbb{E}_{\mathbf{y} \sim \mathbb{P}(Y|\mathbf{x})} [\mathbf{y} \log \hat{\mathbf{y}} + (1 - \mathbf{y}) \log(1 - \hat{\mathbf{y}})] + \lambda \|\mathbf{w}\|^2$$

$$\text{Sigmoid: } \sigma(\mathbf{z}) = \frac{1}{1 + e^{-\mathbf{z}}}$$

$$\text{Softmax: } \sigma(\mathbf{z})_i = \frac{e^{\mathbf{z}_i}}{\sum_j e^{\mathbf{z}_j}}$$

$$\text{Log-Odds: } \log \frac{\mathbb{P}(\mathbf{y} = 1|\mathbf{x})}{\mathbb{P}(\mathbf{y} = 0|\mathbf{x})} = \mathbf{w}^\top \mathbf{x}$$

$$\text{AIC} = 2k - 2 \ln(\mathbb{L})$$

$$\text{BIC} = k \ln(n) - 2 \ln(\mathbb{L})$$

ARIMA

$$y_t^{(d)} = \Delta^d y_t, y'_t = y_t - y_{t-1}$$

$$y_t^{(d)} = c + \sum_{i=1}^p \phi_i y_{t-i}^{(d)} + \epsilon_t + \sum_{j=1}^q \theta_j \epsilon_{t-j}$$

$$\rho_k = \frac{\text{Cov}(y_t, y_{t-k})}{\sqrt{\text{Var}(y_t) \cdot \text{Var}(y_{t-k})}}$$

$$\rho_k = \frac{\sum_{t=k+1}^N (y_t - \mu)(y_{t-k} - \mu)}{\sum_{t=1}^N (y_t - \mu)^2}$$

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_k y_{t-k} + \epsilon_t$$

Frame Title

Regression Coefficients:

Feature	Coefficient
price	0.0826
review_count	0.0836
Facilities	1.0158
star	-0.0280
No_smoking_room	-0.2862
Room_service	0.2798
Free_parking	0.0023
Breakfast	-0.1007
24h_front_desk	-0.3212
Airport_shuttle	0.2126
Intercept	8.0834

Final Results

Validation Metrics:

- R^2 : 0.6899
- Adjusted R^2 : 0.6672

Test Metrics:

- RMSE: 0.4727
- R^2 : 0.8554
- Adjusted R^2 : 0.8448

RESULTS FOR REGRESSION BY RIDGE LR

Average Results of 5-fold Cross Validation

Average Metrics across 5-fold Cross Validation:

Metric	Value
RMSE	0.6931
R^2	0.7070
Adjusted R^2	0.6885

Model Information:

- Model: Ridge Regression
- Alpha (regularization strength): 1.0

Regression Coefficients for Ridge Regression

Regression Coefficients:

Feature	Coefficient
price	0.0634
review_count	0.0811
Comfort	0.2649
Cleanliness	0.4702
Facilities	0.3051
star	-0.0409
Pool	-0.0641
No_smoking_room	-0.3051
Families_room	0.0581
Room_service	0.2730
Free_parking	0.0320
Breakfast	-0.0685
24h_front_desk	-0.3179

Final Results

Validation Metrics:

- RMSE: 0.6973
- R^2 : 0.6891
- Adjusted R^2 : 0.6564

Test Metrics:

- RMSE: 0.4525
- R^2 : 0.8675
- Adjusted R^2 : 0.8536

RESULTS FOR REGRESSION BY COMBINED MODEL

Loss Metrics

Epoch	Train Loss	Val Loss	Test Loss
1	0.6822	0.3867	0.4392
2	0.5542	0.0902	0.2441
3	0.4349	0.0648	0.1939

Training and Validation RMSE by Epoch

RMSE Metrics

Epoch	Train RMSE	Val RMSE	Test RMSE
1	0.8913	0.6249	0.6733
2	0.7778	0.2932	0.5001
3	0.6836	0.2366	0.4490

SUMMARIZE RESULTS FOR REGRESSION

Metric	Linear Regression (Test)	Ridge Regression (Test)
RMSE	0.4727	0.4525
R^2	0.8554	0.8675
Adjusted R^2	0.8448	0.8536

Test RMSE Comparison

Model	Test RMSE
Linear Regression	0.4727
Ridge Regression	0.4525
Combined Model (Epoch 3)	0.4490

RESULTS FOR CLASSIFICATION BY SOFTMAX REGRESSION

Model Configuration and Parameters

Model Configuration

Parameter	Value
Solver	lbfgs
Max iterations	100
Convergence achieved	True
Number of iterations	41

Intercept Terms

Class	Intercept
Class 1	0.3917
Class 2	2.3919
Class 3	-2.7835

Feature Coefficients by Class

Feature Coefficients

Feature	Class 1	Class 2	Class 3
price	-1.1394	0.4520	0.6875
review_count	-0.4015	0.1896	0.2119
Comfort	-1.6322	-0.9319	2.5641
Cleanliness	-1.5122	0.4023	1.1099
Facilities	-0.2497	-0.6817	0.9314
star	0.5740	-0.1361	-0.4379
Pool	0.2249	0.0312	-0.2560
No_smoking_room	0.3253	0.2681	-0.5933
Families_room	-0.0577	0.0165	0.0412
Room_service	-0.3503	0.1926	0.1578
Free_parking	-0.3306	0.1443	0.1863
Breakfast	0.3599	0.0816	-0.4415
24h_front_desk	1.1567	0.1092	-1.2659

Cross-validation Results and Test Evaluation

Cross-validation Results

Metric	Value (\pm Std Dev)
Accuracy	0.8441 (\pm 0.0410)
Precision	0.8671 (\pm 0.0252)
Recall	0.8441 (\pm 0.0410)
F1-Score	0.8479 (\pm 0.0381)
AUC-ROC	0.9405 (\pm 0.0173)
AUC-PR	0.9243 (\pm 0.0186)

Test Evaluation Results

Metric	Value
Accuracy	0.8475
Precision	0.8620
F1-Score	0.8503
AUC-ROC	0.9272

RESULTS FOR CLASSIFICATION BY STACKING MODEL

Ensemble Model Configuration and Cross-validation Results

Class Distribution in Training Set (after Merging)

Class	Samples
Class 1	196
Class 2	613
Class 3	224

5-fold Cross-validation Results

Metric	Value (\pm Std Dev)
Accuracy	0.9729 (\pm 0.0095)
Precision	0.9731 (\pm 0.0094)
Recall	0.9729 (\pm 0.0095)
F1-Score	0.9729 (\pm 0.0094)
AUC-ROC	0.9958 (\pm 0.0024)
AUC-PR	0.9944 (+0.0031)

Base Model Configurations and Meta Model Coefficients

SVM Model Configuration

Parameter	Value
Kernel	rbf
C	1.0
Support Vectors	[53, 267, 60]

KNN Model Configuration

Parameter	Value
n_neighbors	5
weights	distance
algorithm	auto

Meta Model Coefficients in Ensemble

Meta Model Coefficients (Softmax Regression)

Meta Feature	Class 1	Class 2	Class 3
svm_class_0	0.7830	-0.2981	-0.4848
svm_class_1	-0.2485	0.6698	-0.4213
svm_class_2	-0.5082	-0.4196	0.9278
knn_class_0	1.2760	-0.7048	-0.5712
knn_class_1	-0.6771	1.3067	-0.6297
knn_class_2	-0.5727	-0.6499	1.2225
dt_class_0	1.2760	-0.7048	-0.5712
dt_class_1	-0.6771	1.3067	-0.6297
dt_class_2	-0.5727	-0.6499	1.2225
rf_class_0	1.1465	-0.6073	-0.5392
rf_class_1	-0.5766	1.1161	-0.5395
rf_class_2	-0.5437	-0.5566	1.1003

Test Set Evaluation Results for Ensemble Model

Test Set Evaluation Results

Metric	Value
Accuracy	0.9552
Precision	0.9555
Recall	0.9552
F1-Score	0.9550
AUC-ROC	0.9903
AUC-PR	0.9873

SUMMARIZE RESULTS FOR CLASSIFICATION TASK

Performance Comparison on Test Set

Metric	Logistic Regression	Ensemble Model
Accuracy	0.8475	0.9552
Precision	0.8620	0.9555
Recall	0.8475	0.9552
F1-Score	0.8503	0.9550
AUC-ROC	0.9272	0.9903
AUC-PR	0.9009	0.9873

RESULTS FOR CLUSTERING BY K-MEANS

Clustering Performance: KMeans and DBSCAN

KMeans Fold Metrics

Fold	Silhouette Score	Inertia
Fold 1	0.1955	6384.05
Fold 2	0.1978	6271.58
Fold 3	0.1759	6235.52
Fold 4	0.2141	6492.29
Fold 5	0.1842	6303.61

KMeans Summary

Metric	Value
Best Silhouette Score	0.2141
Test Silhouette Score	0.2039
Test Inertia	6492.29
Best Parameters	$n_clusters = 10$
Timestamp	20241102_080352

DBSCAN Performance Metrics

DBSCAN Fold Metrics

Fold	Silhouette Score
Fold 1	0.2201
Fold 2	0.0482
Fold 3	0.0531
Fold 4	0.3911
Fold 5	0.3049

DBSCAN Summary

Metric	Value
Best Silhouette Score	0.3911
Test Silhouette Score	0.3582
Best Parameters	$eps = 0.8, min_samples = 3$
Timestamp	20241102_080352

RESULTS FOR CLUSTERING BY DBSCAN

SUMMARIZE RESULTS FOR CLUSTERING TASK

Comparison on Test Set

Metric	KMeans	DBSCAN
Silhouette Score	0.2039	0.3582
Inertia	6492.29	N/A
Best Parameters	$n_clusters = 10$	$eps = 0.8, min_samples = 3$

RESULTS FOR TIME SERIES FORECASTING

RMSE: 0.3141

