NYC Taxi Demand Analysis

Exploratory Data Analysis for Target Models

Dataset Information:
- Period: July 01, 2014 to January 31, 2015
- Total Records: 10,320
- Frequency: 30-minute intervals
- Total Days: 214
- Data Points per Day: 48

TARGET FORECASTING MODELS:

□ Naive Forecasting
- Baseline: Last value prediction
- Simple persistence model
- Benchmark for comparison

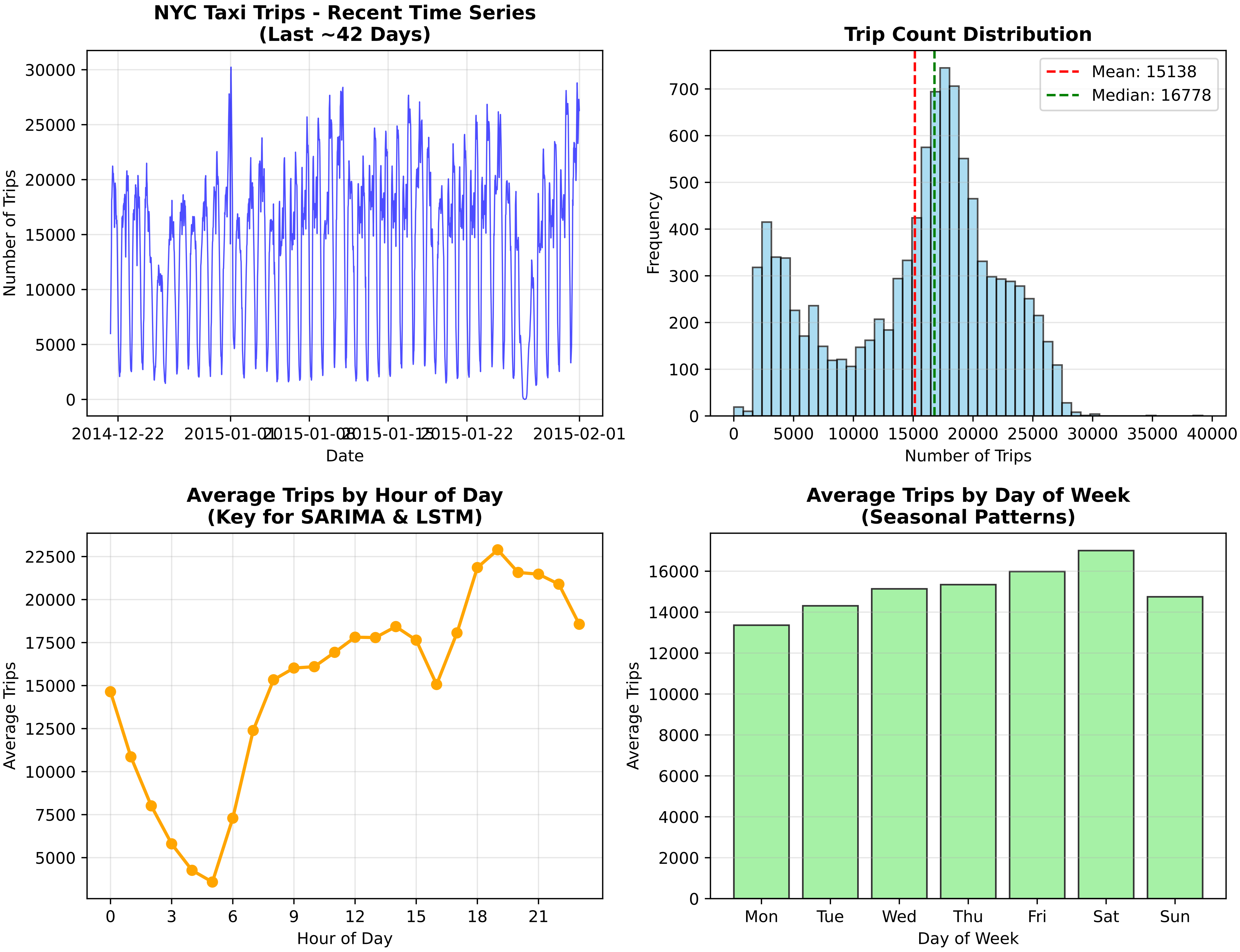
□ SARIMA (Seasonal ARIMA)
- Statistical time series model
- Handles trend and seasonality
- Requires stationary data

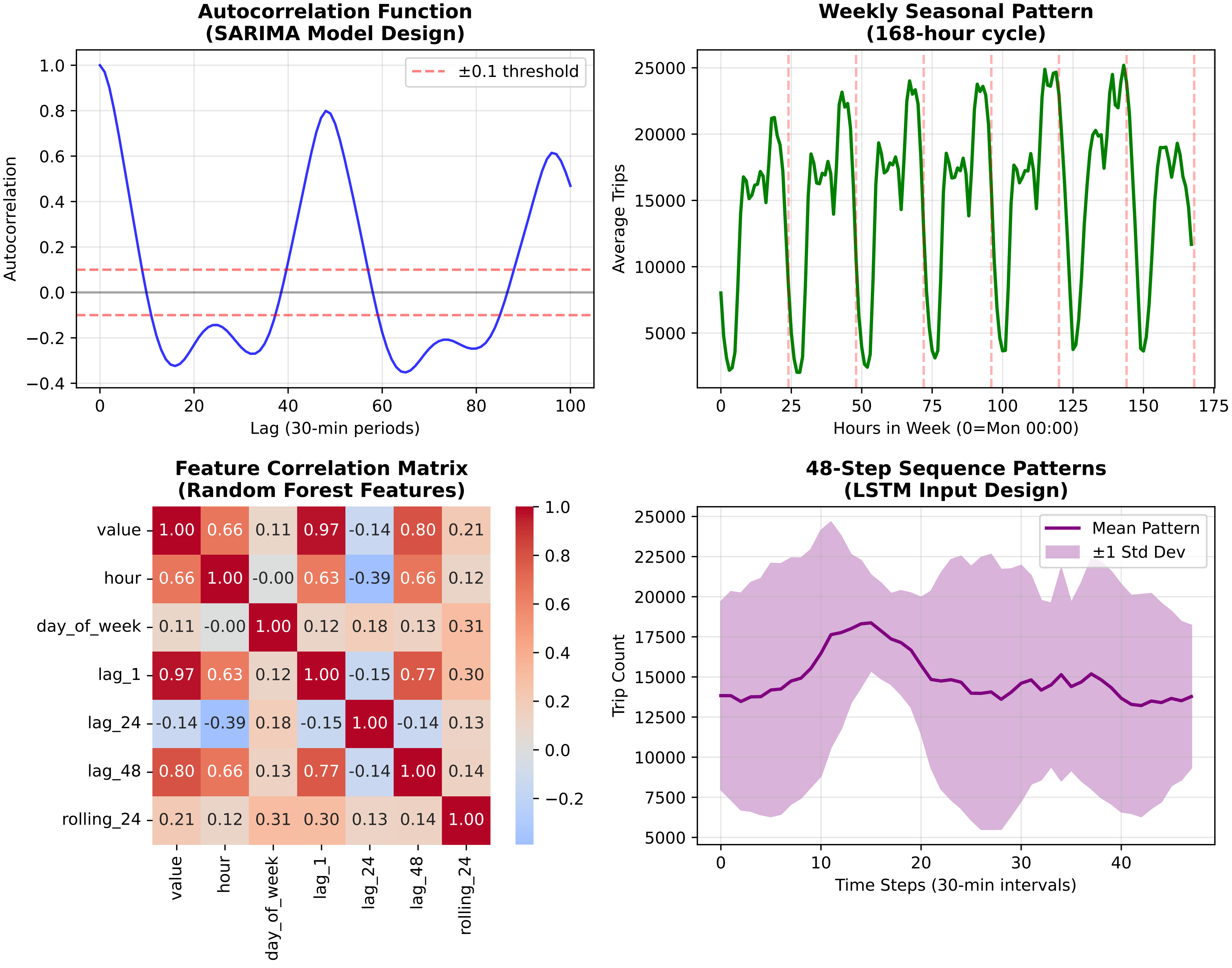
□ Random Forest
- Machine learning approach
- Feature engineering critical
- Handles non-linear patterns

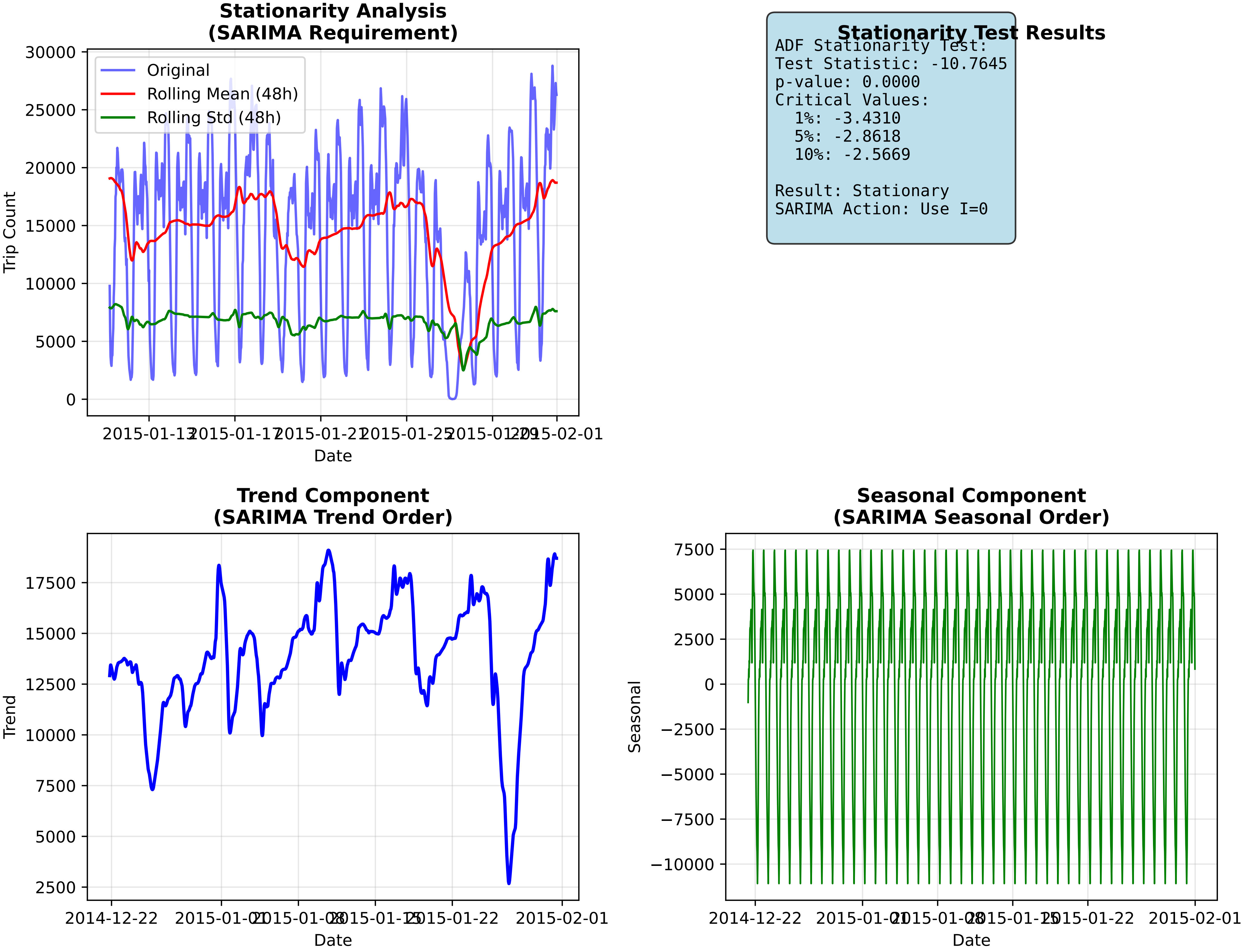
□ LSTM Neural Network
- Deep learning model
- Sequence-to-sequence learning
- Captures complex dependencies

EDA FOCUS AREAS:

✓ Temporal patterns for SARIMA optimization
✓ Feature engineering for Random Forest
✓ Sequence patterns for LSTM design
✓ Data quality and preprocessing needs
✓ Seasonal decomposition analysis
✓ Stationarity testing for SARIMA

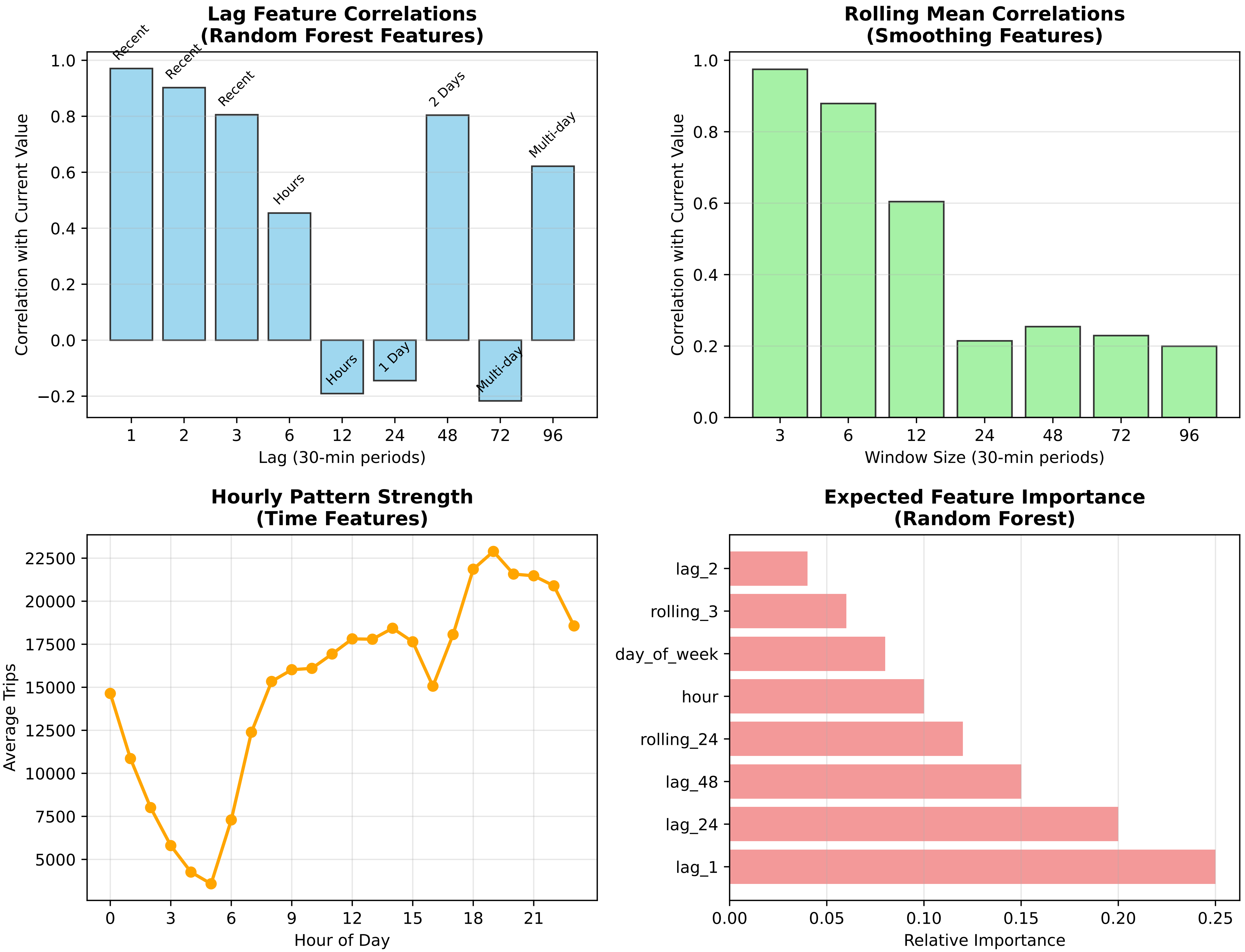

NYC Taxi Trips - Recent Time Series
(Last ~42 Days)
Number of Trips
0
5000
10000
15000
20000
25000
30000
2014-12-22
2015-01-01
2015-01-08
2015-01-15
2015-01-22
2015-02-01
Date
Trip Count Distribution
Mean: 15138
Median: 16778
Frequency
0
100
200
300
400
500
600
700
0
5000
10000
15000
20000
25000
30000
35000
40000
Number of Trips
Average Trips by Hour of Day
(Key for SARIMA & LSTM)
Average Trips
5000
7500
10000
12500
15000
17500
20000
22500
0
3
6
9
12
15
18
21
Hour of Day
Average Trips by Day of Week
(Seasonal Patterns)
Average Trips
0
2000
4000
6000
8000
10000
12000
14000
16000
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Day of Week

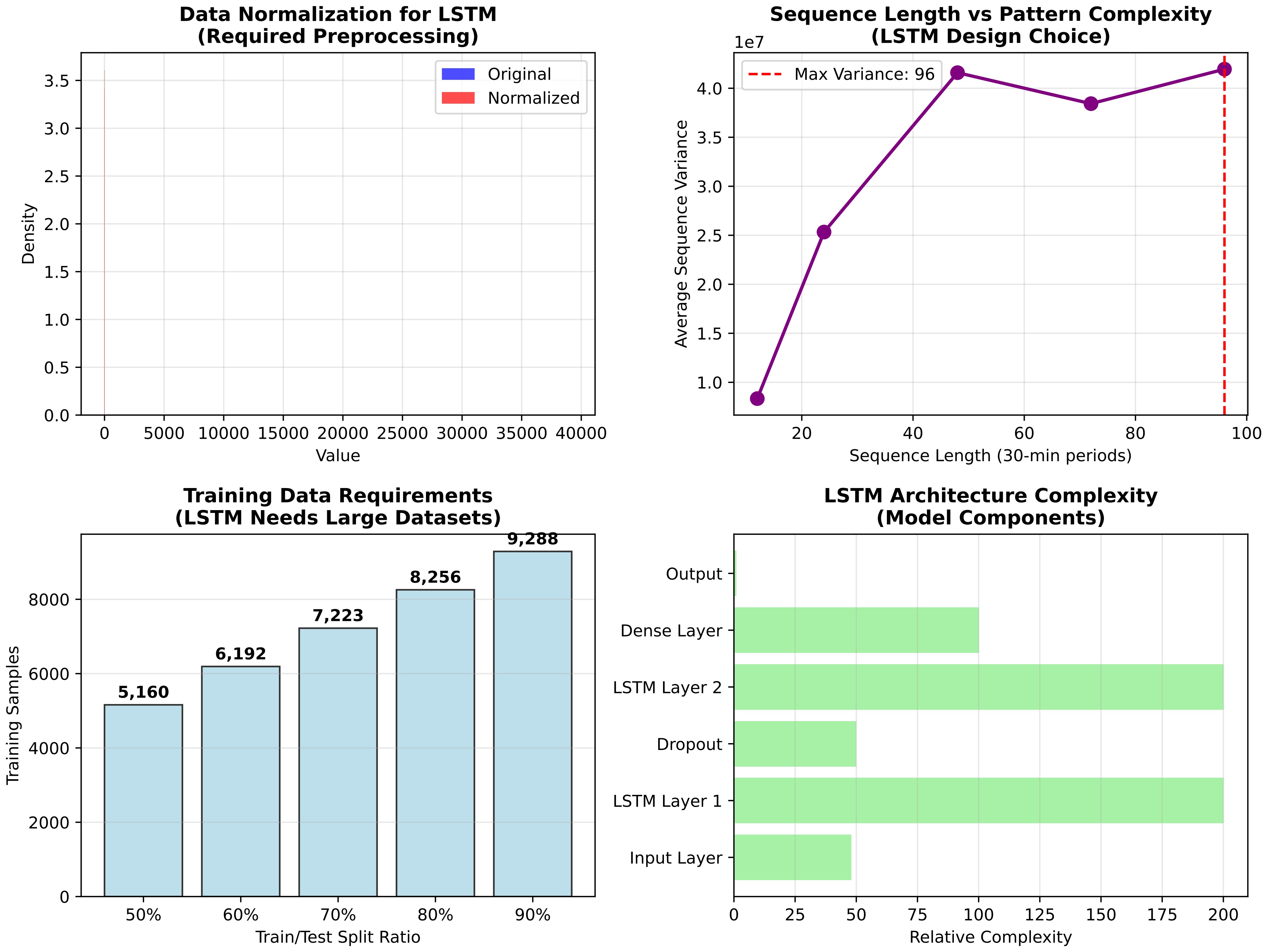

Autocorrelation Function
(SARIMA Model Design)
±0.1 threshold
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
20
40
60
80
100
Lag (30-min periods)
Weekly Seasonal Pattern
(168-hour cycle)
Average Trips
25000
20000
15000
10000
5000
0
25
50
75
100
125
150
175
Hours in Week (0=Mon 00:00)
Feature Correlation Matrix
(Random Forest Features)
value
hour
day_of_week
lag_1
lag_24
lag_48
rolling_24
1.00 0.66 0.11 0.97 -0.14 0.80 0.21
0.66 1.00 -0.00 0.63 -0.39 0.66 0.12
0.11 -0.00 1.00 0.12 0.18 0.13 0.31
0.97 0.63 0.12 1.00 -0.15 0.77 0.30
-0.14 -0.39 0.18 -0.15 1.00 -0.14 0.13
0.80 0.66 0.13 0.77 -0.14 1.00 0.14
0.21 0.12 0.31 0.30 0.13 0.14 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
48-Step Sequence Patterns
(LSTM Input Design)
Mean Pattern
±1 Std Dev
Trip Count
25000
22500
20000
17500
15000
12500
10000
7500
5000
0
10
20
30
40
Time Steps (30-min intervals)



```
ADF Stationarity Test:
Test Statistic: -10.7645
p-value: 0.0000
Critical Values:
  1%: -3.4310
  5%: -2.8618
  10%: -2.5669

Result: Stationary
SARIMA Action: Use I=0
```


Lag Feature Correlations
(Random Forest Features)
Correlation with Current Value
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Recent
Recent
Recent
Hours
Hours
1 Day
2 Days
Multi-day
Multi-day
1
2
3
6
12
24
48
72
96
Lag (30-min periods)
Rolling Mean Correlations
(Smoothing Features)
Correlation with Current Value
1.0
0.8
0.6
0.4
0.2
0.0
3
6
12
24
48
72
96
Window Size (30-min periods)
Hourly Pattern Strength
(Time Features)
Average Trips
22500
20000
17500
15000
12500
10000
7500
5000
0
3
6
9
12
15
18
21
Hour of Day
Expected Feature Importance
(Random Forest)
lag_2
rolling_3
day_of_week
hour
rolling_24
lag_48
lag_24
lag_1
0.00
0.05
0.10
0.15
0.20
0.25
Relative Importance



Data Normalization for LSTM (Required Preprocessing)

Sequence Length vs Pattern Complexity (LSTM Design Choice)

Training Data Requirements (LSTM Needs Large Datasets)

LSTM Architecture Complexity (Model Components)

Model-Specific Data Insights & Recommendations

```
DATA CHARACTERISTICS SUMMARY:

Total Data Points: 10,320
Average Trips: 15138 ± 6939
Range: 8 to 39197
Coefficient of Variation: 0.458
Missing Data Rate: 0.000%
Seasonal Strength: 0.378

MODEL-SPECIFIC RECOMMENDATIONS:

□ NAIVE FORECASTING:
Strengths:
• Excellent baseline with minimal computation
• Robust to data quality issues
• Fast execution for real-time applications

Considerations:
• Will perform poorly in volatile periods
• Coefficient of variation 0.458 suggests moderate volatility
• Best used as benchmark for other models

Data Requirements: □ Minimal - just last observation

□ SARIMA MODELING:
Strengths:
• Strong seasonal patterns detected (strength: 0.378)
• Suitable for 48-period seasonal cycle (24-hour days)
• Statistical foundation with interpretable parameters

Considerations:
• May need differencing for stationarity
• Requires parameter tuning (p,d,q)(P,D,Q,s)
• Sensitive to outliers and structural breaks

Data Requirements: □ Good - 10,320 points sufficient
Recommended Configuration: SARIMA(1,1,1)(1,1,1,48)

□ RANDOM FOREST:
Strengths:
• Can handle non-linear patterns
• Robust to outliers and missing data
• Feature importance interpretability
• Good with engineered features

Considerations:
• Requires extensive feature engineering
• May overfit with too many features
• Computationally more intensive

Data Requirements: □ Excellent - 10,320 points ideal
Feature Strategy:
• Lag features: 1, 2, 3, 24, 48 periods
• Rolling means: 3, 12, 24 period windows
• Time features: hour, day_of_week, month
• Seasonal features: sin/cos transformations

□ LSTM NEURAL NETWORK:
Strengths:
• Captures complex temporal dependencies
• Excellent for sequence-to-sequence learning
• Can model non-linear relationships
• Handles multiple input features naturally

Considerations:
• Requires significant computational resources
• Needs careful hyperparameter tuning
• Data normalization critical
• Risk of overfitting with small datasets

Data Requirements: □ Good - 10,320 points adequate
Architecture Recommendations:
• Sequence Length: 48 periods (24 hours)
• Hidden Units: 50-100 per layer
• Layers: 2 LSTM layers with dropout
• Batch Size: 32-64
• Epochs: 50-100 with early stopping

PREPROCESSING RECOMMENDATIONS:

For All Models:
• Data quality: 100.0% complete □
• Outlier detection and handling
• Consistent time intervals validation

For SARIMA:
• Stationarity testing and differencing
• Seasonal decomposition analysis
• Parameter selection via AIC/BIC

For Random Forest:
• Feature scaling (optional)
• Lag feature creation
• Rolling statistics computation
• Categorical encoding for time features

For LSTM:
• MinMax normalization to [0,1] range □ Critical
• Sequence windowing (48 time steps)
• Train/validation/test split: 70/15/15
• Early stopping to prevent overfitting

EXPECTED PERFORMANCE RANKING:
Based on data characteristics and model capabilities:

1. LSTM (Best) - Complex patterns, sufficient data
2. Random Forest - Good with features
3. SARIMA - Strong seasonality
4. Naive (Baseline) - Simple persistence

SUCCESS FACTORS:
• Strong daily/weekly seasonality favors SARIMA and LSTM
• Large dataset size supports complex models
• Moderate volatility suggests all models viable
• Clear temporal patterns support sequence models
```