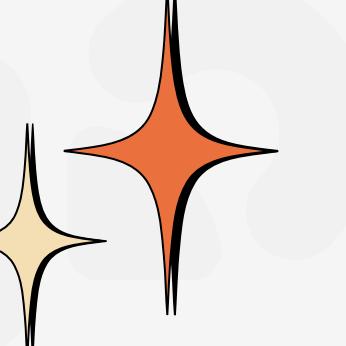




Project 1

Sales Forecasting: **Time Series Prediction using** **SARIMA and LSTM for** **Inventory Optimization**

Synergy Squad Team 





Meet Our Team



I Putu Ferry Wistika



**Lukas Yuliyanto
Gunawan**



Haikal Firdaus



**Muhammad
Egalingga Zainuri**



Adhi Kurniawan

linkedin.com/in/putuwistika

linkedin.com/in/lukas

linkedin.com/in/haikalfirdaus

linkedin.com/in/egalinggazainuri

linkedin.com/in/adhikurniawan

Outline

1

Background, Problem Statement & Objective

2

Understanding Dataset

3

**EDA
(Exploratory Data Analysis)**

4

Data Pre-Processing

5

Feature Engineering

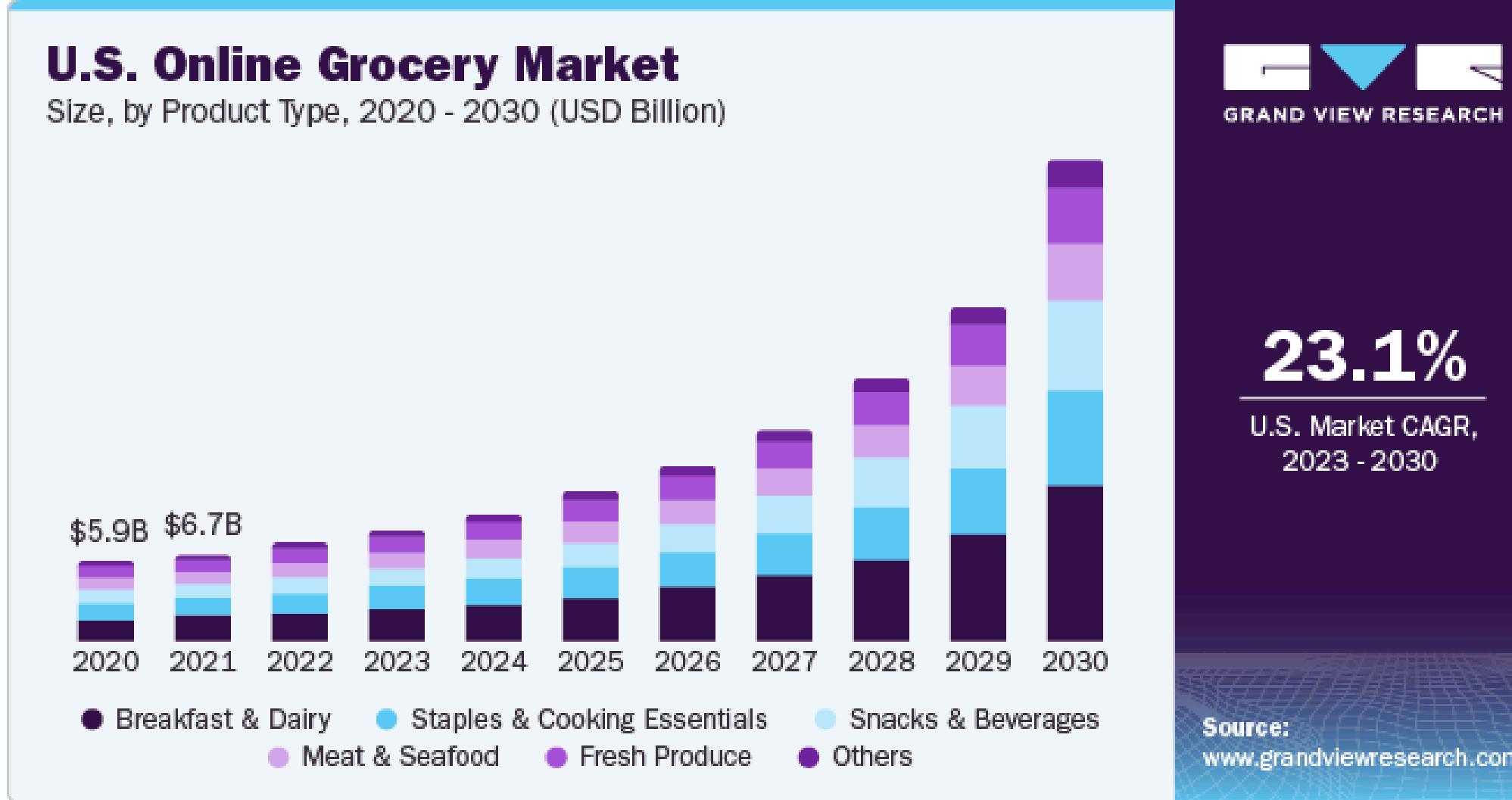
6

Modeling & Evaluation

7

Business Recommendation & Conclusion

Background & Problem Statement

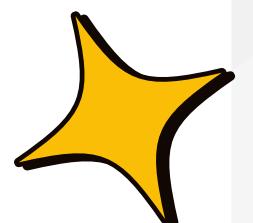


Tantangan Industri Retail Grocery di Era Digital

- Volatilitas Demand yang Tinggi
- Perubahan Perilaku Konsumen
- Kompleksitas Manajemen Inventory
- Customer Expectation

Dalam Konteks Global Market Grocery, Industri grocery mengalami transformasi luar biasa dengan pertumbuhan yang eksplosif

“Kesulitan signifikan dalam memprediksi penjualan produk GROCERY secara akurat, yang menyebabkan inefficiency inventory management dan kerugian finansial yang substansial”



Objectives & Scope

Tujuan utama dari proyek ini adalah untuk meningkatkan efisiensi operasional dan penjualan dengan membantu manajemen memperoleh wawasan yang dapat ditindaklanjuti dari data mereka. Melalui pemodelan time series yang dan teknik deep learning, kami berupaya memahami tren penjualan jangka panjang, mengidentifikasi pola yang memengaruhi kinerja penjualan, serta memberikan solusi untuk tantangan dalam pengelolaan inventaris.



Mengembangkan Model Prediksi yang Akurat dan Reliable
Mencapai MAPE $\leq 15\%$ untuk prediksi penjualan bulanan



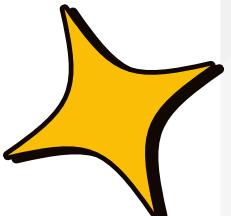
Operational Efficiency
Automate forecasting process and reduce manual effort



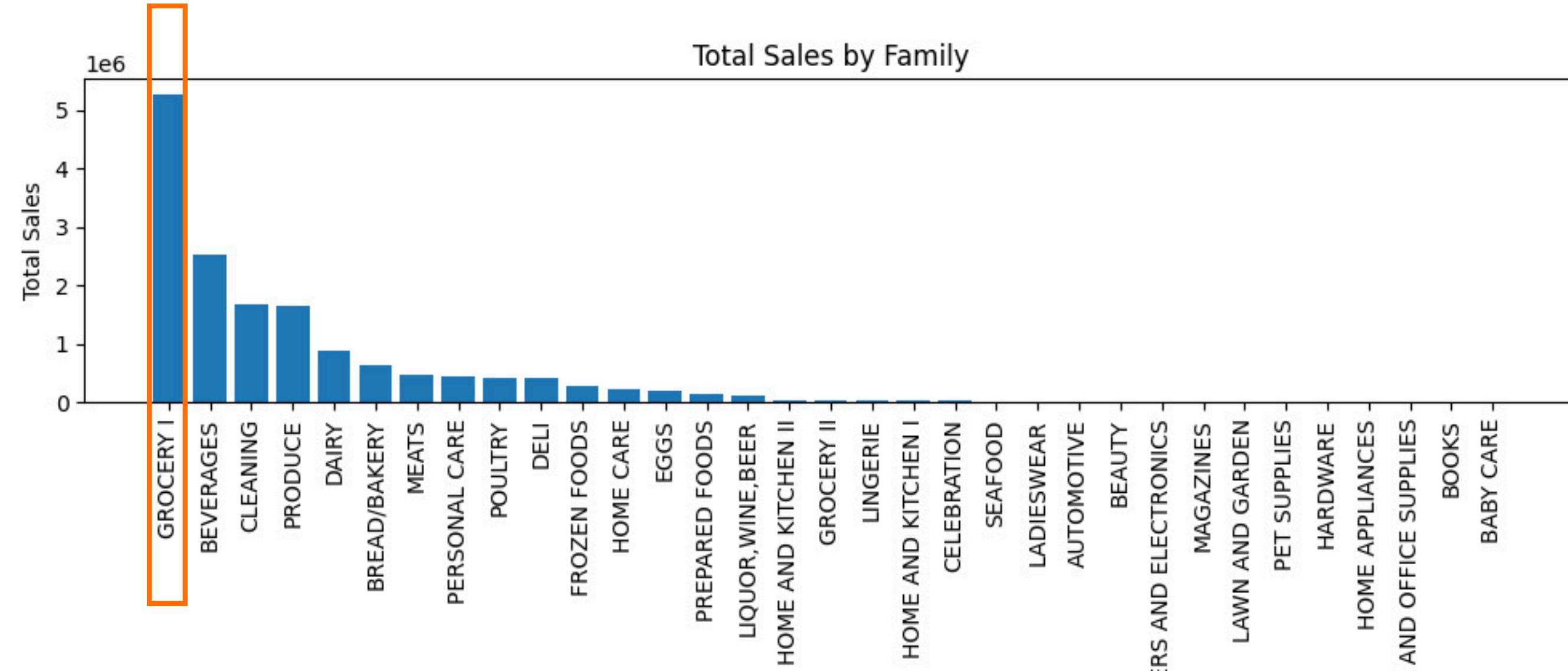
Business Value Creation
Inventory Optimization: Reduce stockout by 40% and excess inventory by 30%



Revenue Protection
Prevent lost sales dari availability issues



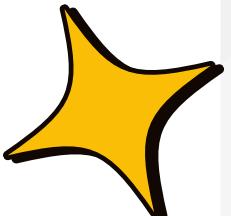
Objectives & Scope



Fokus Analisis: GROCERY I

Kategori produk utama dengan kontribusi 35,2% dari total pendapatan.

Menggunakan data historis >4 tahun untuk memproyeksikan penjualan masa depan dan mendukung pengambilan keputusan berbasis data melalui model prediktif berbasis time series.

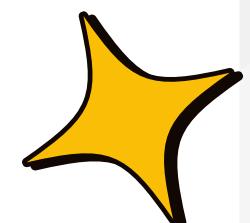


Data Collection & Preparation

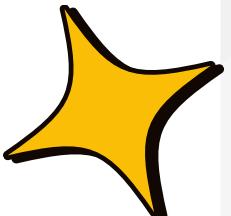
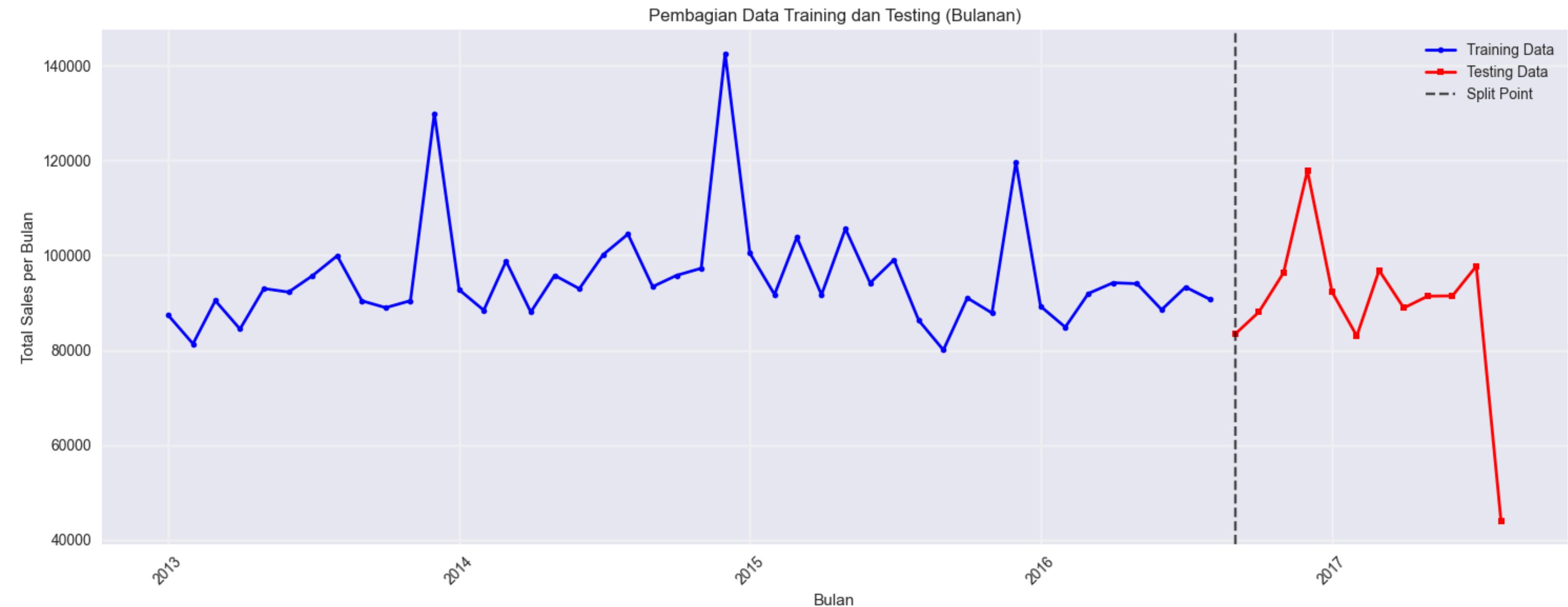
Sumber & Struktur Data

- Jumlah Data:** 55.572 observasi harian
- Rentang Tanggal:** 1 Jan 2013 – 15 Agustus 2017
- Frekuensi:** Harian, lalu diagregasi menjadi bulanan
- Lokasi:** Satu toko (Toko #5)

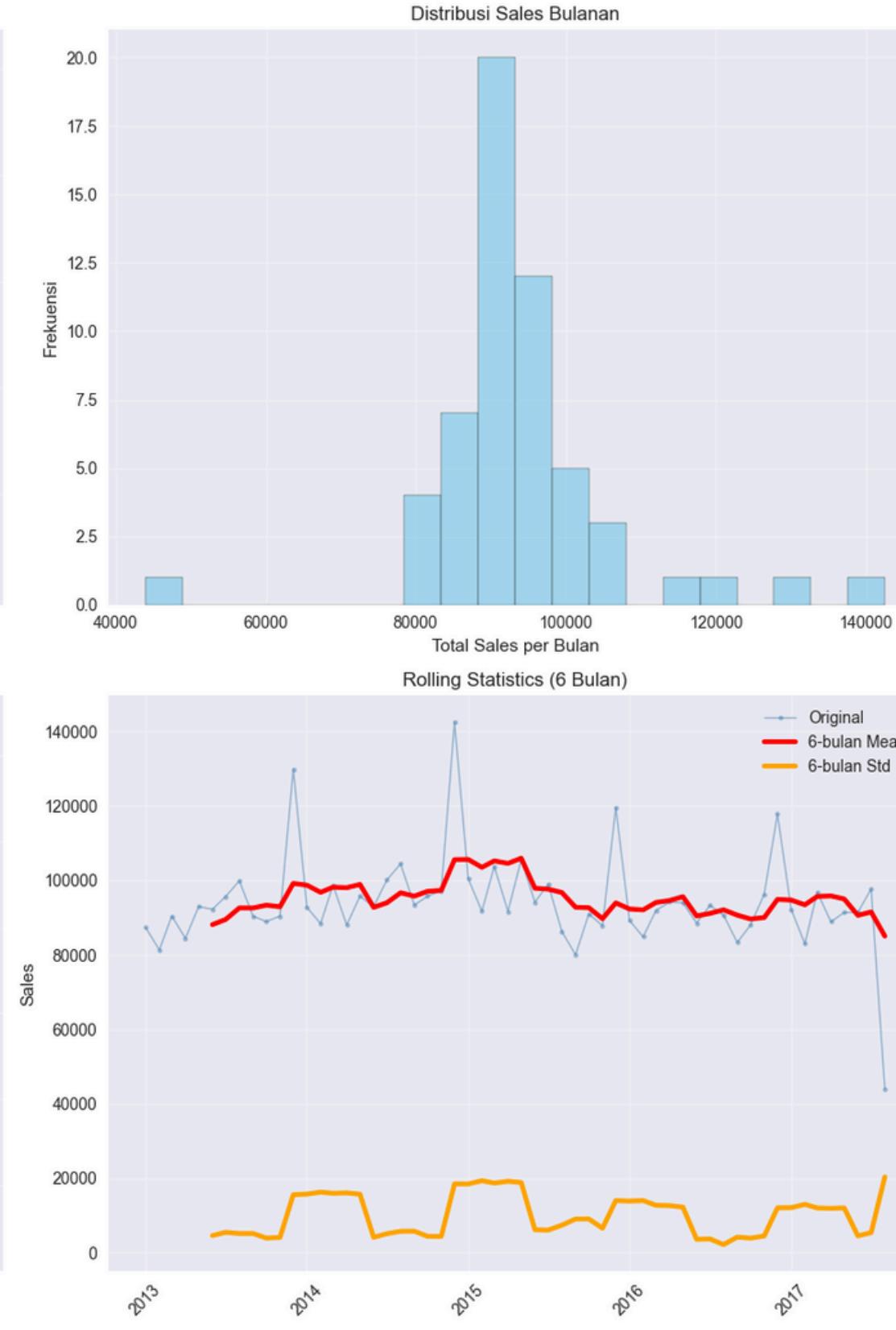
Variabel	Tipe	Deskripsi	Makna Bisnis
id	Integer	ID transaksi unik	Identifikasi sistem
date	Date	Tanggal transaksi	Tanggal penjualan
store_nbr	Integer	Nomor toko	Lokasi (5)
family	String	Kategori produk	Fokus: GROCERY I
sales	Float	Jumlah penjualan	Variabel target
onpromotion	Integer	Jumlah barang promo	Indikator dampak promosi
dcoilwtico	Float	Harga minyak dunia (WTI)	Indikator ekonomi eksternal



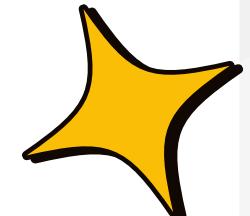
Data Collection & Preparation



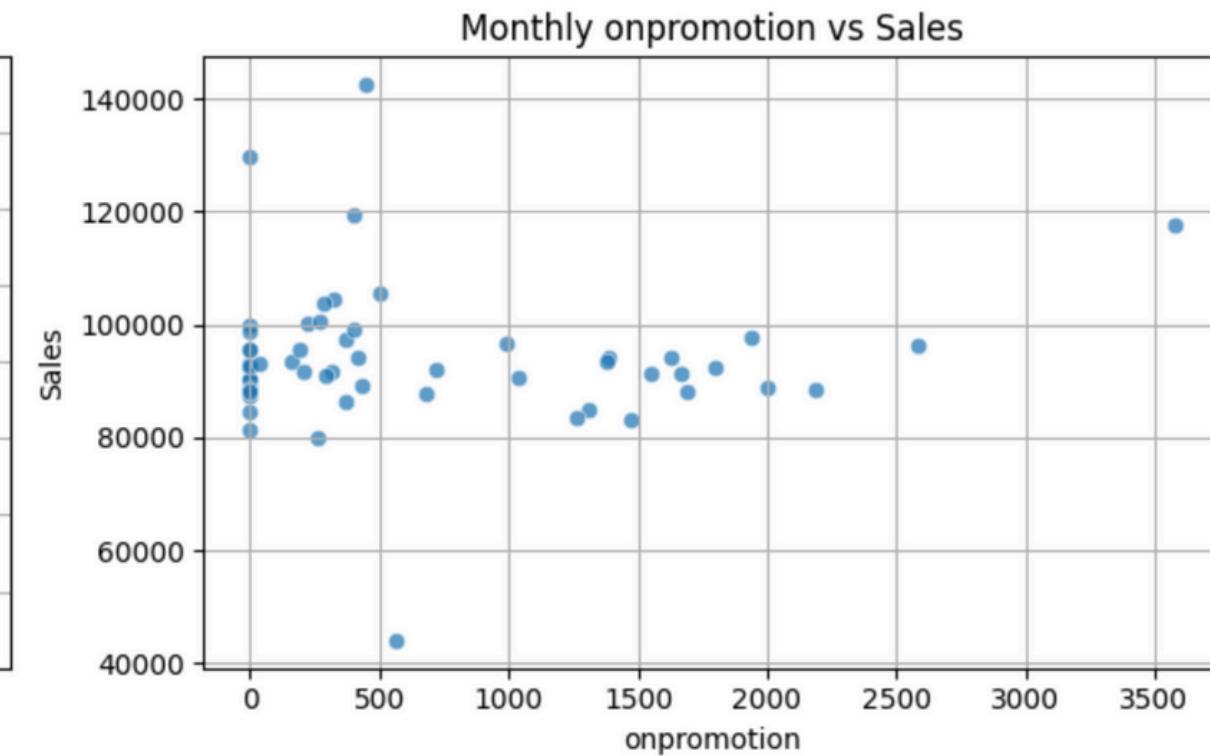
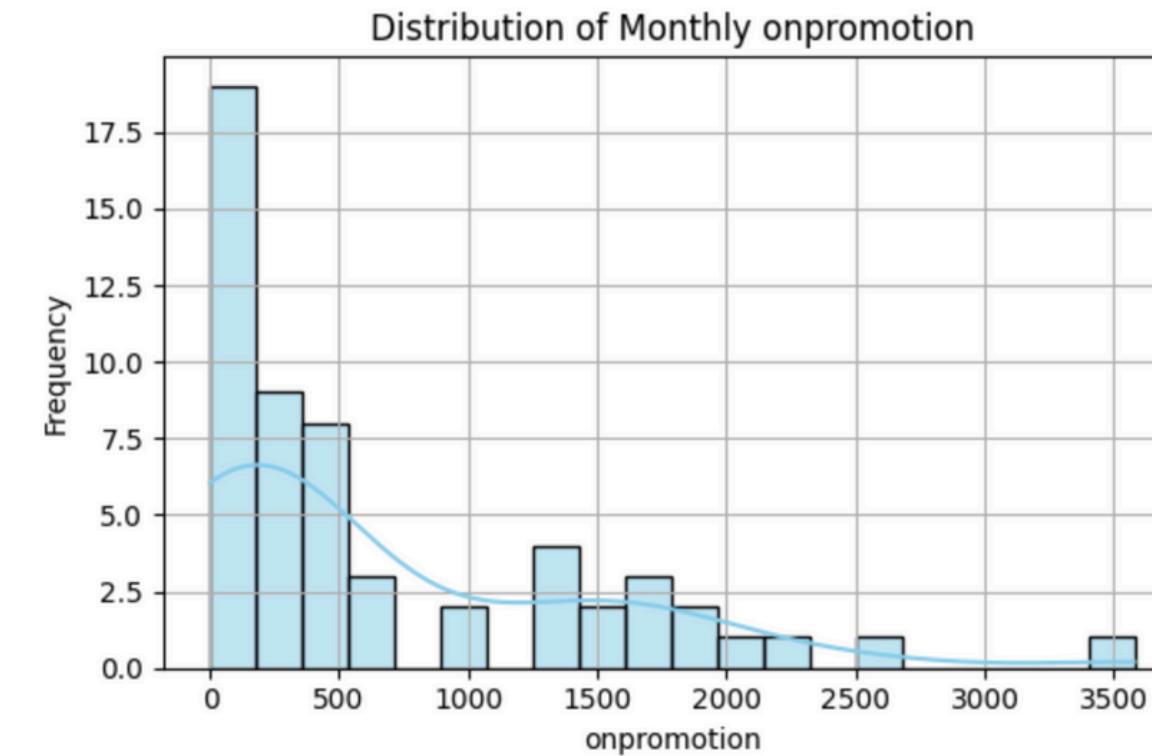
Exploratory Data Analysis



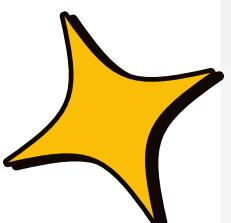
- **Tren Sales Bulanan menunjukkan fluktuasi signifikan dari tahun ke tahun, namun cenderung stabil setelah pertengahan periode, dengan satu penurunan ekstrem di akhir (kemungkinan anomali).**
- **Distribusi Sales berbentuk right-skewed, dengan mayoritas bulan memiliki penjualan antara 800.000–1.000.000 unit.**
- **Rata-Rata Bulanan memperlihatkan bahwa Desember memiliki rata-rata penjualan tertinggi, diikuti oleh Mei dan Februari, menunjukkan potensi efek musiman.**



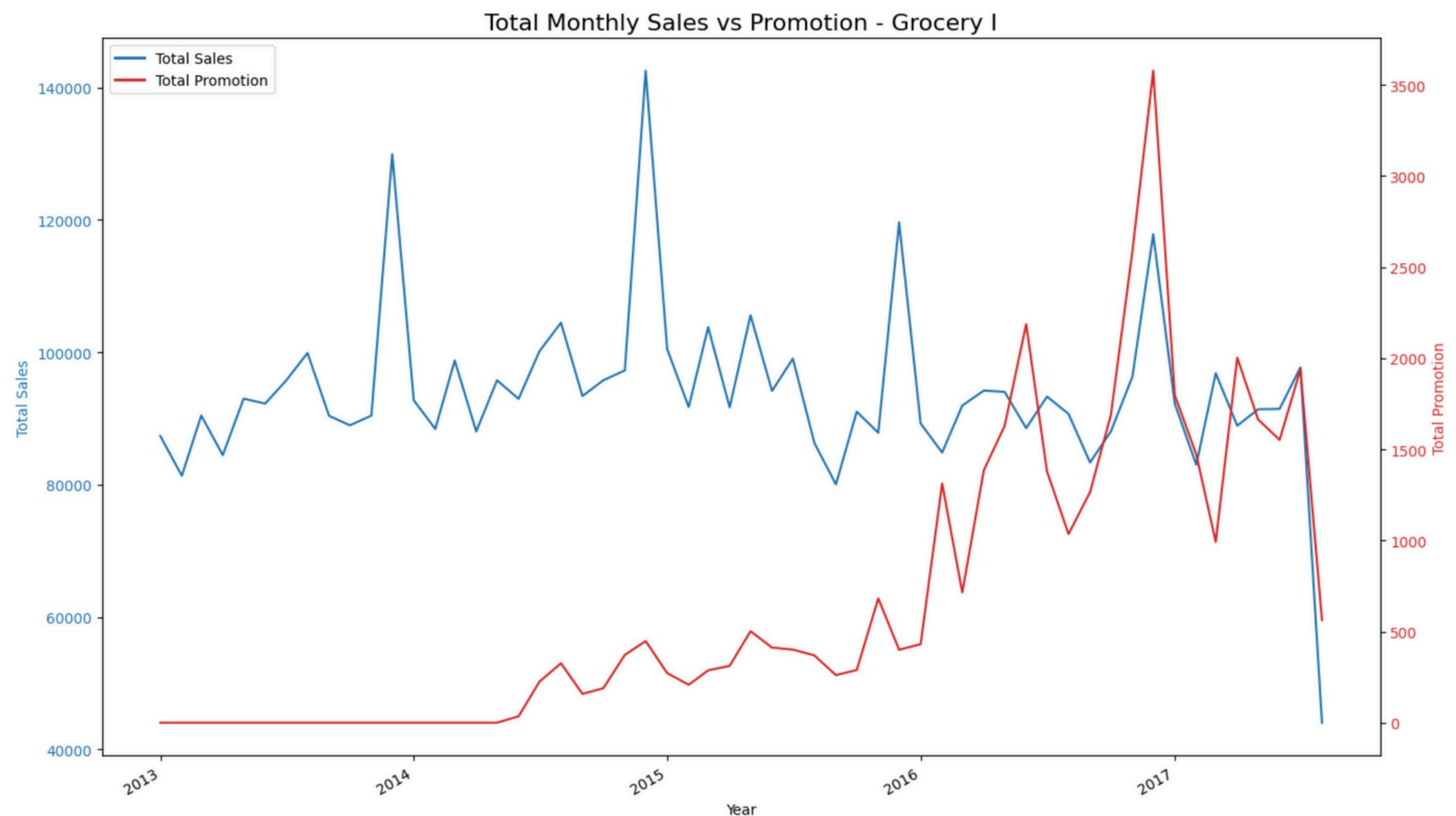
Exploratory Data Analysis



- Distribusi Onpromotion sangat right-skewed: sebagian besar bulan memiliki jumlah promosi di bawah 1000, menunjukkan aktivitas promosi yang tidak merata sepanjang waktu.
- Korelasi dengan Sales terlihat lemah: scatter plot tidak menunjukkan pola yang jelas antara jumlah promosi dan penjualan bulanan, mengindikasikan bahwa promosi belum tentu berdampak langsung terhadap peningkatan penjualan.



Exploratory Data Analysis



- Tren Penjualan Bulanan menunjukkan fluktuasi tajam dari tahun ke tahun, dengan puncak yang tinggi di momen tertentu
- Aktivitas Promosi mulai meningkat signifikan sejak 2015 dan mencapai puncaknya pada 2017.
- Meskipun ada kenaikan promosi, penjualan tidak selalu meningkat seiring—terlihat bahwa korelasi tidak konsisten, menandakan bahwa promosi belum tentu efektif mendorong penjualan.
- Potensi insight: promosi mungkin perlu diarahkan lebih strategis (misal: waktu, durasi, atau jenis produk).



ALUR PEMODELAN ANALISIS DERET WAKTU: ARIMA

Identifikasi Model

Stasioner

Mean dan variansi konstan

- Autocorrelation Function (ACF)
- Partial ACF (PACF)

Non-stasioner

Mean dan variansi tidak konstan

Estimasi Parameter

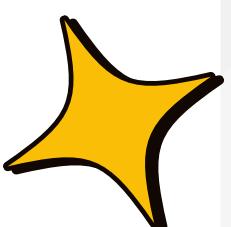
- Auto Regressive (AR)
- Moving Average (MA)
- Auto Regressive Moving Average (ARMA)

Uji Diagnostik

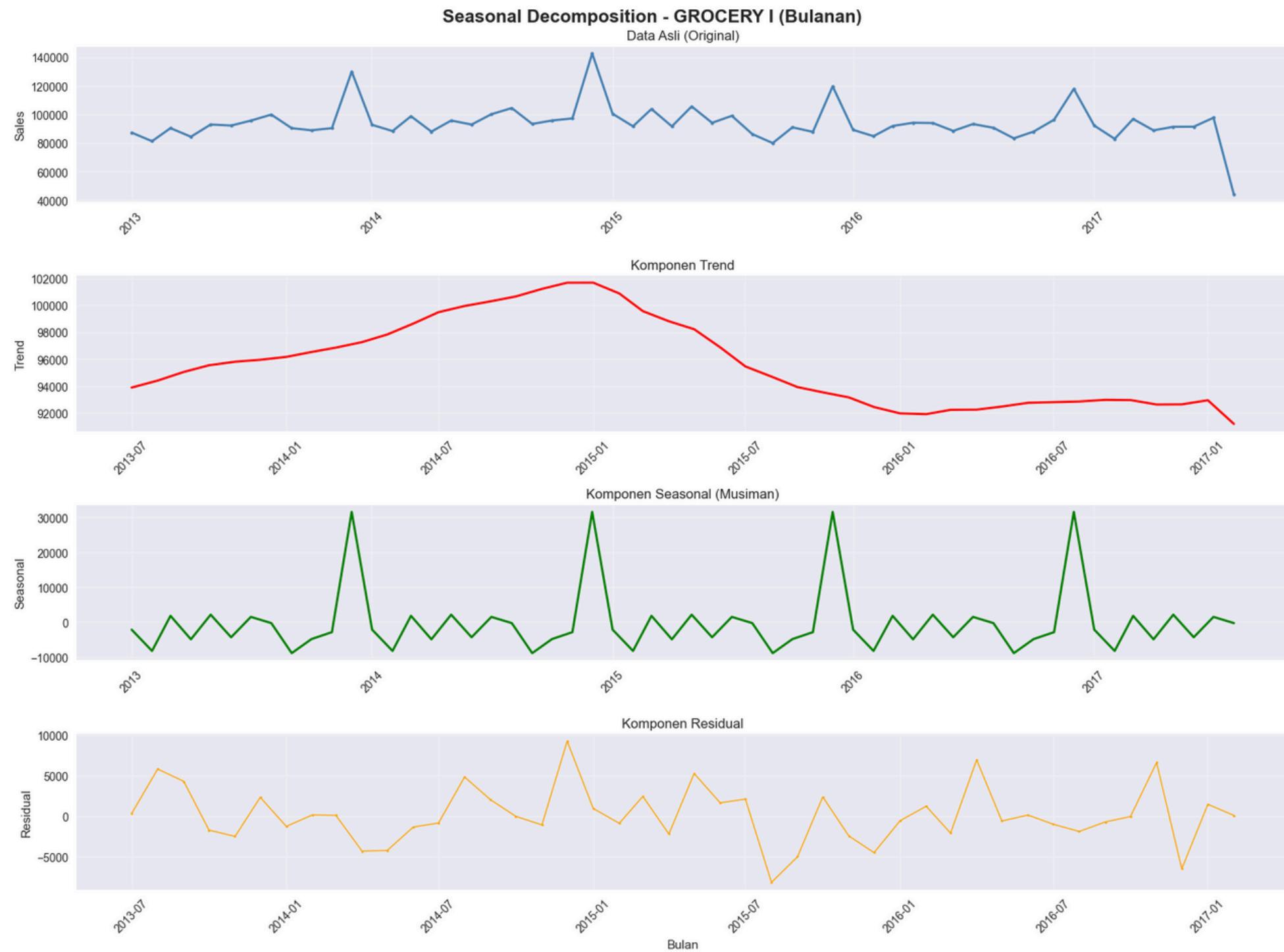
Uji White Noise :

1. Normalitas Residual
2. Independensi Residual

Uji Ljung-Box (uji residu)



Uji Kestasioneran

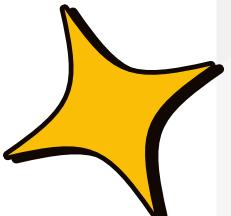


Kestasioneran data dapat langsung dilihat dari plot obsevasi terhadap waktu, dengan ciri-ciri:

- Tidak terdapat unsur musiman dan trend.
- Memiliki sifat rataan dan variansi yang konstan.
- Kovariansi antar data satu sama lain konstan (tidak bergantung pada waktu t).

Pada plot ini tidak memenuhi ciri-ciri stasioner.
Di yakinkan dengan uji ADF

✓ UJI ADF untuk Data Bulanan Sales GROCERY I:
 - ADF Statistic: 0.3592
 - p-value: 0.9799
 - Critical Values:
 1%: -3.5886
 5%: -2.9299
 10%: -2.6032
△ HASIL: Data Bulanan Sales GROCERY I adalah NON-STASIONER (p-value > 5%)



ARIMA Model Development

Identifikasi Model

Ciri-ciri model **stasioner tidak terpenuhi**, maka disebut model non stasioner.

Menstasionerkan model dengan cara **mendiferensikan** data tersebut.

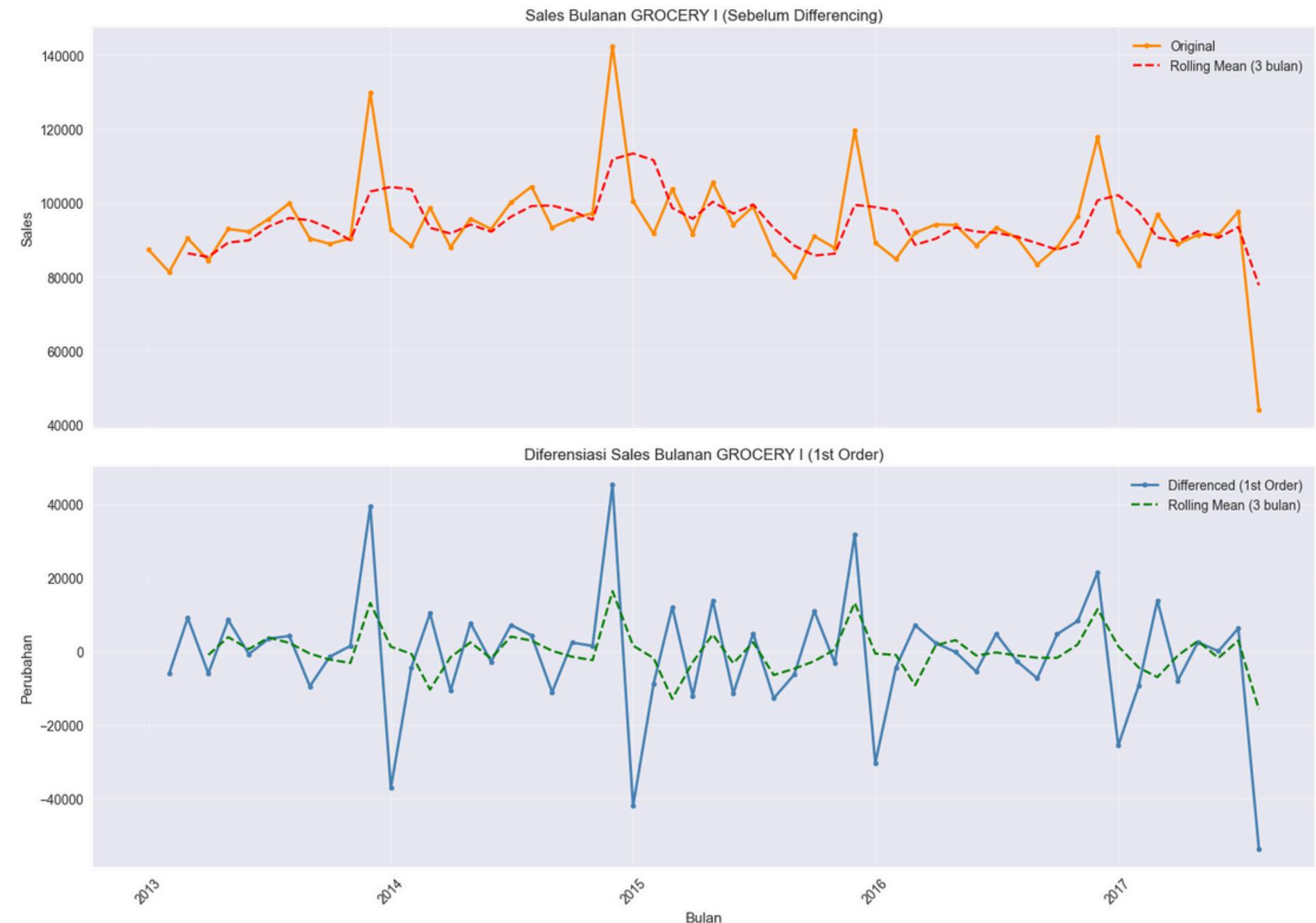
Misalkan $\{X_t\}$ mengikuti suatu proses deret waktu,
maka proses diferensiasi dapat dilakukan dengan

$$Y_t = X_t - X_{t-1}$$

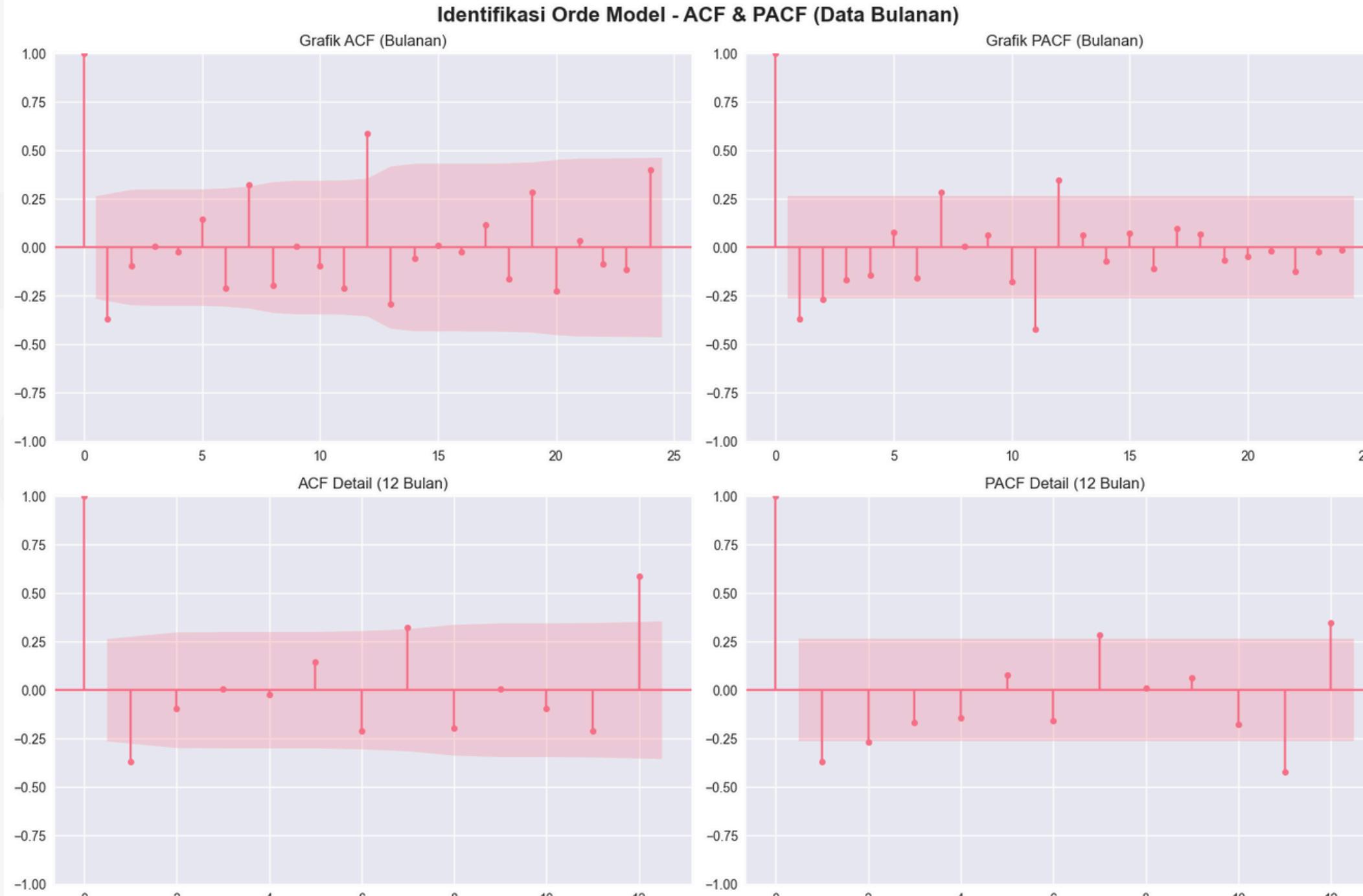
```
[ ] result = adfuller(monthly_groceries['sales'].diff().dropna())
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])

→ ADF Statistic: -5.038874
p-value: 0.000019
```

Uji ADF : Data sudah Stasioner dengan p-value < 5%



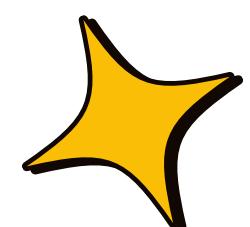
Identifikasi Orde



- Plot **ACF (Autocorrelation Function)** menunjukkan signifikansi pada lag ke-1, 7, dan 12, yang mengindikasikan adanya seasonal pattern bulanan dan mingguan.
- Plot **PACF (Partial Autocorrelation Function)** juga menunjukkan puncak signifikan pada lag yang sama, mendukung komponen autoregressive (AR) pada titik tersebut.

Dari analisis ACF dan PACF, diperoleh beberapa kombinasi orde ARIMA/SARIMA yang potensial

- (1, 1, 1), (1, 1, 7), (1, 1, 12)
- (7, 1, 1), (7, 1, 7), (7, 1, 12)
- (11, 1, 1), (11, 1, 7), (11, 1, 12)



ARIMA Model Development

Pemilihan Model Deret Waktu

Untuk mengevaluasi performa model, dilakukan pencarian otomatis parameter terbaik menggunakan Auto ARIMA.

Hasil terbaik:

ARIMA(1, 1, 1)

AIC: 950.437

Namun, model ini belum mempertimbangkan komponen musiman, padahal:

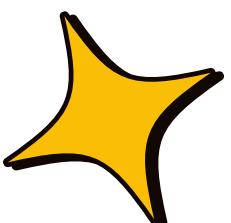
- Pola penjualan menunjukkan siklus musiman tahunan (bulan ke-12)
- Hal ini teridentifikasi dari plot ACF & PACF, serta tren data historis

Oleh Sebab itu Kita Tuning Model menggunakan SARIMA

Best ARIMA model is ARIMA(1, 1, 1) with AIC: 950.437

Summary of the Best ARIMA Model:

SARIMAX Results						
Dep. Variable:	sales	No. Observations:	44			
Model:	ARIMA(1, 1, 1)	Log Likelihood	-472.218			
Date:	Tue, 27 May 2025	AIC	950.437			
Time:	16:21:08	BIC	955.720			
Sample:	0 - 44	HQIC	952.385			
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.1155	0.213	0.543	0.587	-0.302	0.533
ma.L1	-0.5639	0.221	-2.554	0.011	-0.997	-0.131
sigma2	1.711e+08	4.84e-10	3.54e+17	0.000	1.71e+08	1.71e+08
Ljung-Box (L1) (Q):			0.93	Jarque-Bera (JB):		14.40
Prob(Q):			0.33	Prob(JB):		0.00
Heteroskedasticity (H):			0.49	Skew:		1.14
Prob(H) (two-sided):			0.19	Kurtosis:		4.68



ARIMA Model Development

Pemilihan Model Deret Waktu

Untuk mengakomodasi pola musiman, digunakan model SARIMA (Seasonal ARIMA)

Model yang digunakan:

python

```
ARIMA(train_df['sales'], order=(1,1,1), seasonal_order=(1,1,1,12))
```

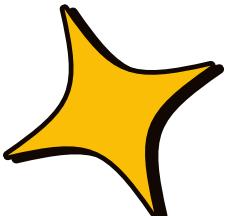
- Artinya:

- (1,1,1) → Komponen non-musiman (AR, I, MA)
- (1,1,1,12) → Komponen musiman dengan periode 12 bulan

Model	AIC	Catatan
ARIMA(1,1,1)	950.437	Tanpa komponen musiman
SARIMA(1,1,1,12)	644.363 	Memasukkan pola musiman bulanan

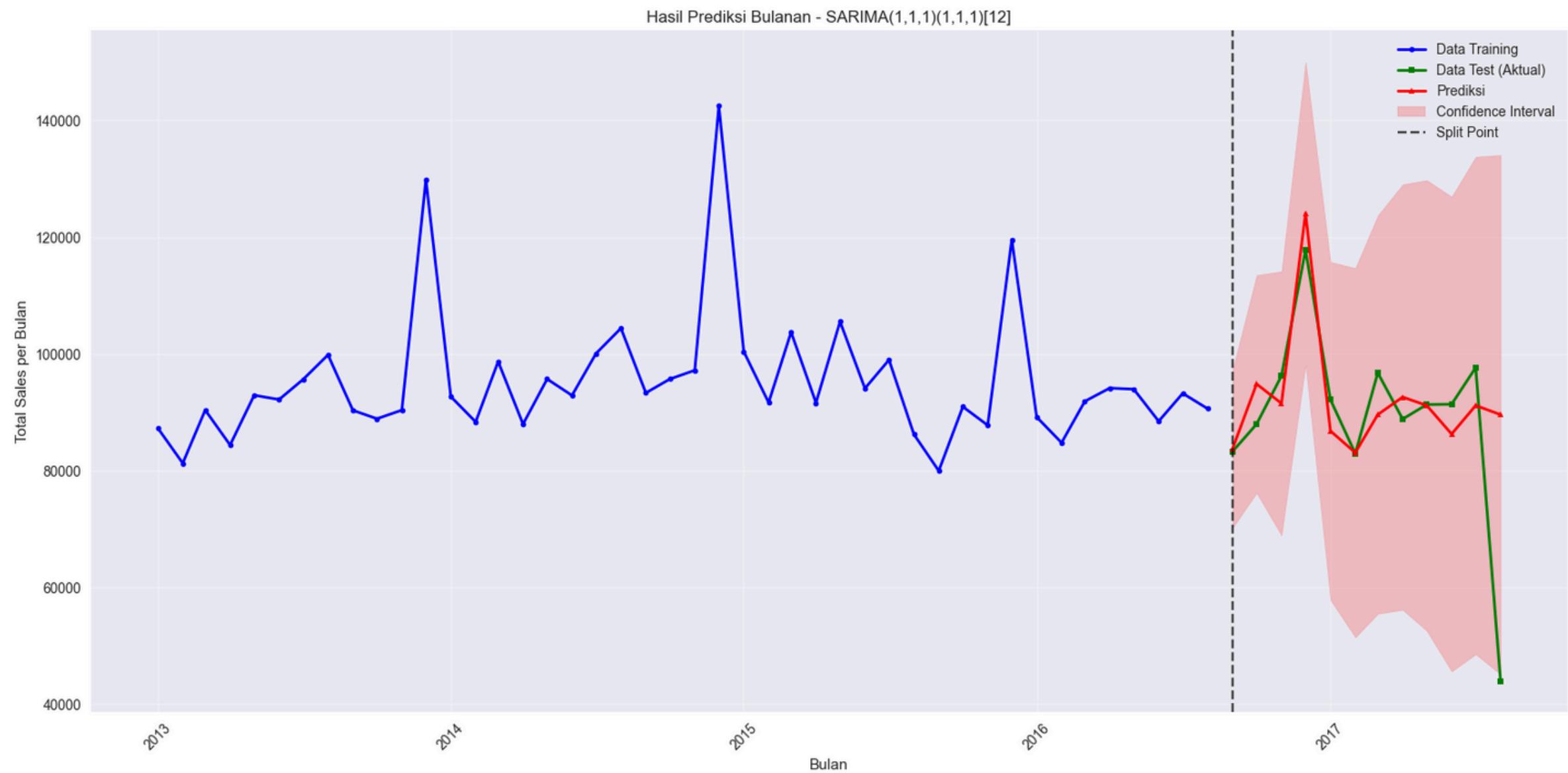
SARIMAX Results						
Dep. Variable:	sales	No. Observations:	44			
Model:	ARIMA(1, 1, 1)x(1, 1, 1, 12)	Log Likelihood	-317.181			
Date:	Tue, 27 May 2025	AIC	644.363			
Time:	16:21:08	BIC	651.533			
Sample:	0 - 44	HQIC	646.700			
Covariance Type: opg						
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.1054	3.339	0.032	0.975	-6.438	6.649
ma.L1	-0.1553	3.386	-0.046	0.963	-6.791	6.480
ar.S.L12	-0.9091	0.353	-2.576	0.010	-1.601	-0.217
ma.S.L12	0.9730	0.569	1.709	0.087	-0.143	2.089
sigma2	4.68e+07	5.08e-08	9.21e+14	0.000	4.68e+07	4.68e+07
Ljung-Box (L1) (Q): 3.15 Jarque-Bera (JB): 1.12						
Prob(Q):	0.08	Prob(JB):	0.57			
Heteroskedasticity (H):	10.95	Skew:	-0.42			
Prob(H) (two-sided):	0.00	Kurtosis:	3.39			

Semakin rendah nilai AIC, semakin baik performa model dalam menyeimbangkan akurasi dan kompleksitas.



ARIMA Model Development

Prediksi Dengan SARIMA Terbaik



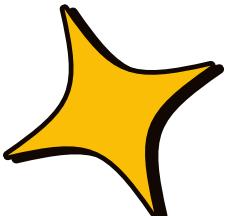
Prakiraan dilakukan menggunakan model SARIMA terbaik yang telah diperoleh:

SARIMA(1,1,1)(1,1,1)₁₂

dan akan dilakukan prakiraan pada s waktu yang akan datang, maka

$$\phi(B)(1 - B)^d(1 - B^s)^D Y_{t+h} = \theta(B)\Theta(B^s)\varepsilon_{t+h}$$

dengan s merupakan lag waktu untuk prakiraan di waktu yang akan datang.



ARIMA Model Development

Akurasi Model

Metric	Nilai	Interpretasi
MAE	4.232,01	Rata-rata kesalahan absolut dalam satuan penjualan
MAPE	4,42% 	Tingkat kesalahan relatif rendah (di bawah 10%) — model sangat akurat
MASE	0,424	Lebih baik dari model naif ($MASE < 1$)
RMSE	4.972,83	Ukuran kesalahan yang mempertimbangkan variansi — cocok untuk membandingkan model

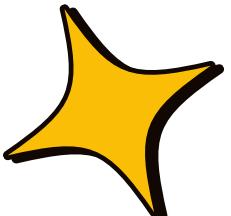
Model menghasilkan forecast yang sangat akurat dan stabil, layak digunakan untuk pengambilan keputusan bisnis.

Uji Diagnostik Ljung-Box

Komponen	Hasil	Interpretasi
Q-Statistic	2,8382	Nilai statistik uji
p-value	0,9850 	$>> 0,05 \rightarrow$ residu acak
Hasil	 Tidak ada autokorelasi pada residu	

Uji Ljung-Box:

- Residu model tidak menunjukkan pola tertentu
- Model sudah sesuai (**well-fitted**) dan residu bersifat white noise



ALUR PEMODELAN ANALISIS DERET WAKTU: LSTM

Data Preparation & Feature Engineering

Uji Stasioneritas

- ADF Test (Augmented Dickey-Fuller)
- Visual Plot Analysis

Transformasi Data

- Differencing (jika non-stasioner)
- MinMax Scaling (-1, 1)

Lag Features Creation

- Time Series → Supervised Learning
- Lag_1, Lag_2, ..., Lag_n
- Feature Importance Analysis

Model Architecture & Training

LSTM Architecture

- Sequential Model
- LSTM Layer (neurons, dropout)
- Dense Output Layer

Hyperparameters

- Learning Rate
- Batch Size
- Epochs & Early Stopping

Training Process

- Train-Validation Split (temporal)
- Loss: MSE, Optimizer: Adam
- Monitor Training Loss

Evaluation & Validation

Performance Metrics

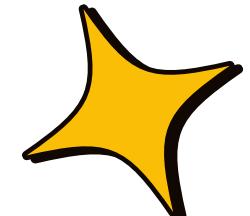
- MAE (Mean Absolute Error)
- RMSE (Root Mean Square Error)
- MAPE (Mean Absolute Percentage Error)

Residual Analysis

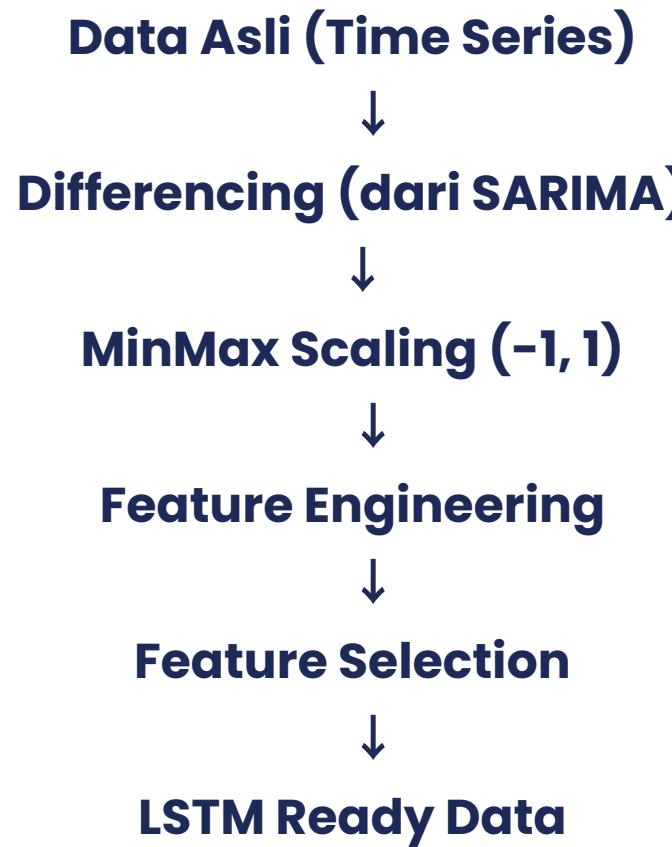
- Residual Plot Analysis
- Autocorrelation of Residuals
- Distribution Normality Test

Validation Techniques

- Time Series Cross-Validation
- Walk-Forward Validation
- Out-of-Sample Testing



TRANSFORMASI DATA



Key Points:

- **Differencing: Sudah dilakukan pada tahap SARIMA**
- **Scaling: MinMax normalisasi (-1, 1) untuk LSTM**
- **Engineering: Lag features + Seasonal + Rolling statistics**
- **Selection: R-squared analysis untuk kombinasi terbaik**

...

TARGET VARIABLE (y) - 1 feature:

1. sales_diff ← Differenced sales (stationary)

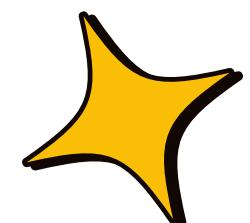
INPUT FEATURES (X) - 7 features:

1. month_num	← Month number (1=Jan, 12=Dec)
2. quarter	← Quarter (1=Q1, 2=Q2, 3=Q3, 4=Q4)
3. rolling_mean_3	← 3-month rolling average
4. rolling_mean_6	← 6-month rolling average
5. lag_1	← Sales_diff 1 month(s) ago
6. lag_2	← Sales_diff 2 month(s) ago
7. lag_3	← Sales_diff 3 month(s) ago

BEST FEATURE COMBINATION (dari Feature Importance Analysis):

Combination 4 menggunakan 5 features dengan $R^2 = 1.0000$:

1. lag_1
2. lag_2
3. lag_3
4. month_num
5. rolling_mean_3



Model Arsitektur & Training

LSTM Architecture

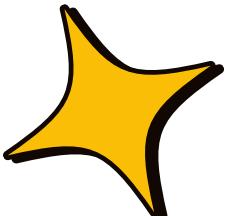
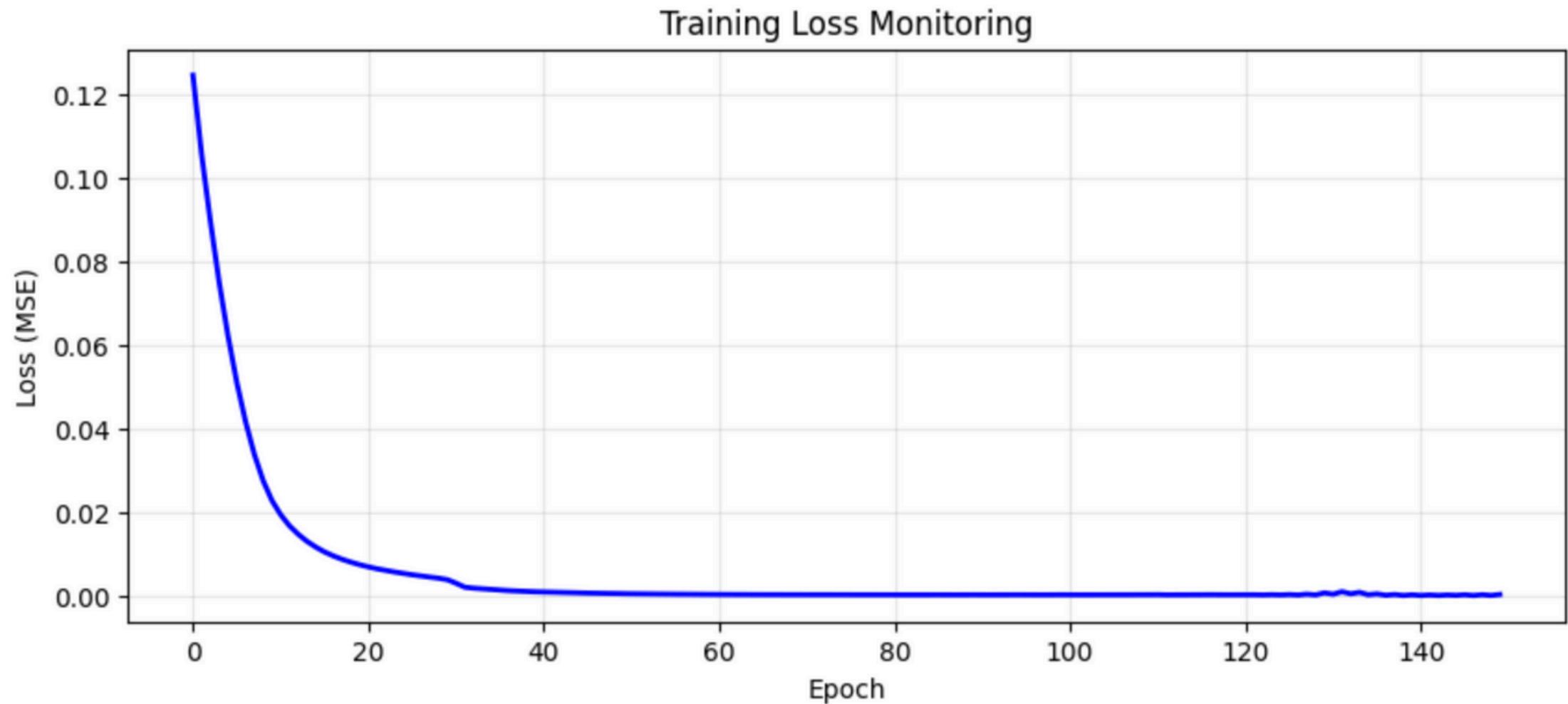
Model dibangun menggunakan pendekatan Sequential dengan fokus pada pemrosesan data berurutan. Arsitektur ini dirancang untuk mempelajari pola temporal dari data time-series.

- Sequential Model
- LSTM Layer: 16 neurons
- Dense Layer: 4 neurons (ReLU)
- Output Layer: 1 neuron (untuk regresi)

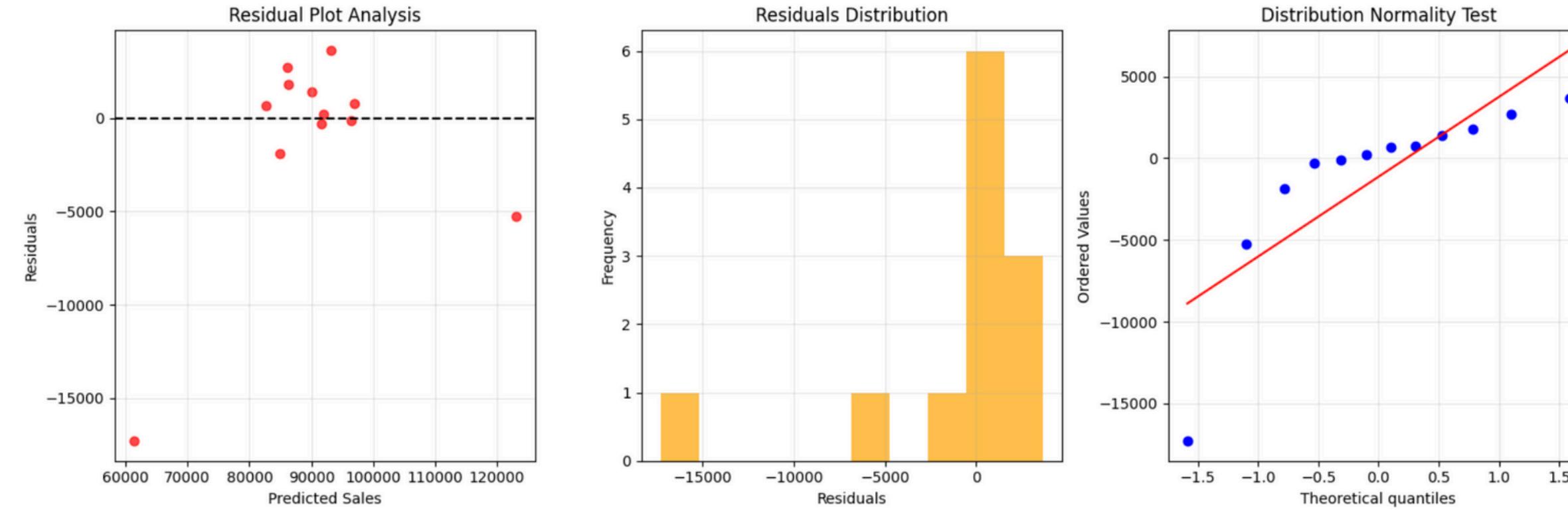
Hyperparameters

Pemilihan hyperparameter dilakukan secara manual berdasarkan eksperimen awal untuk memastikan konvergensi model secara stabil.

- Learning Rate: 0.001
- Batch Size: 1
- Epochs: 150
- Loss Function: Mean Squared Error (MSE)
- Optimizer: Adam



Model Evaluation



- **Residual Analysis:** Sebagian besar residual mendekati garis nol, menunjukkan bahwa model cukup baik dalam menangkap pola utama dari data.
- **Kekuatan Model:** Model mampu menghasilkan prediksi yang relatif stabil pada sebagian besar titik data, dengan kesalahan yang kecil dan terkontrol di area pusat distribusi.

Distribusi Residual:
Terdapat indikasi ketidaksimetrian dan outlier, terutama di nilai ekstrem, yang menunjukkan bahwa model belum optimal dalam menangani seluruh variasi data.

Normality Test (Q-Q Plot):
Beberapa deviasi dari garis normal menunjukkan bahwa residual tidak sepenuhnya normal, namun untuk kebutuhan prediksi, hal ini masih dapat ditoleransi.



LSTM Model Development



Model Validation

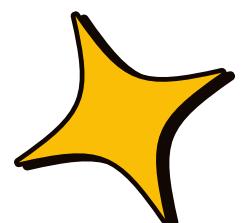
Model Configuration

Komponen	Detail
Arsitektur	LSTM (Sequential)
Hidden Layer	1 LSTM Layer (16 neuron) + Dense Layer (4 neuron, ReLU)
Output Layer	1 neuron
Optimizer	Adam
Loss Function	Mean Squared Error (MSE)
Learning Rate	0.001
Batch Size	1
Epochs	150

Performance Metrics

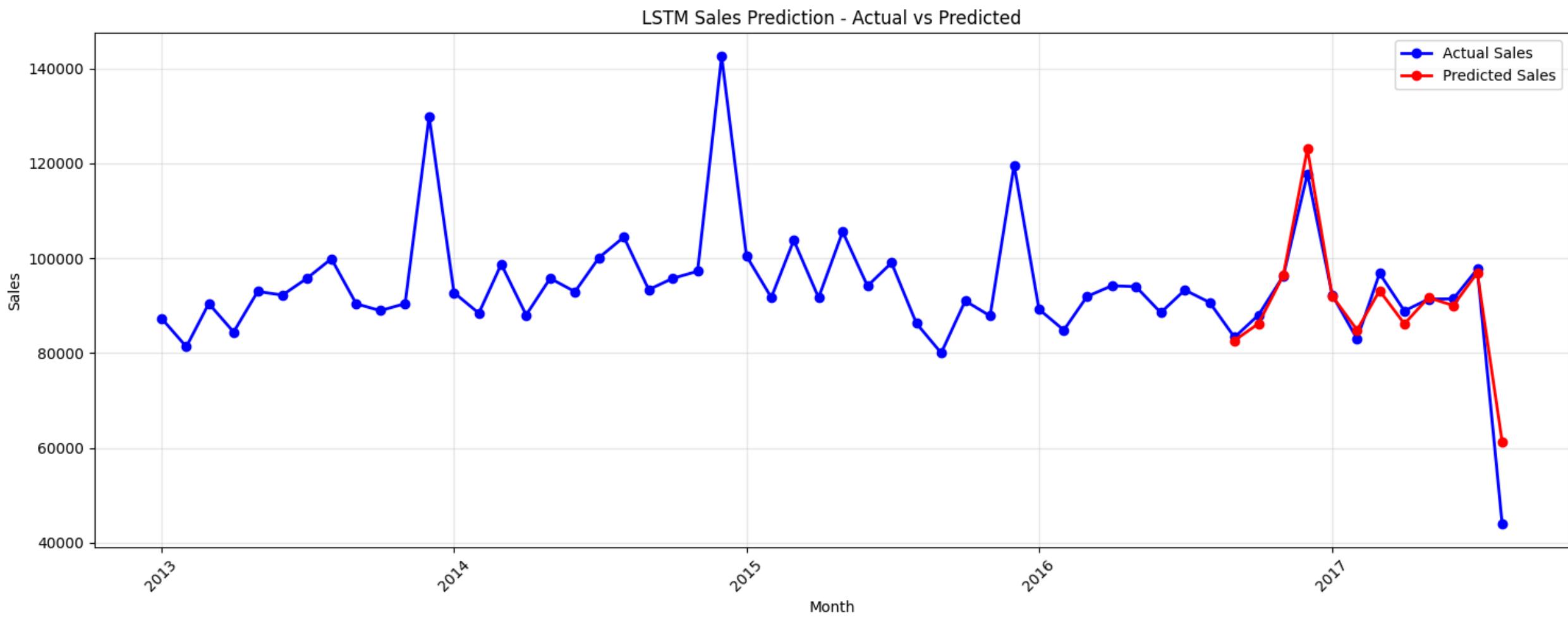
Metode Evaluasi	Nilai	Interpretasi Singkat
MAE	3008.33	Rata-rata kesalahan absolut prediksi
RMSE	5457.25	Lebih sensitif terhadap outlier
MAPE	4.89%	Error relatif terhadap nilai aktual

Model menunjukkan performa yang baik dengan MAPE di bawah 5%, yang berarti prediksi cukup akurat secara relatif. MAE dan RMSE masih dalam batas wajar, walaupun RMSE sedikit tinggi karena adanya outlier pada data.



LSTM Model Development

Prediksi



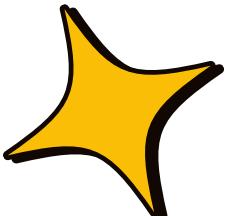
--- HASIL PREDIKSI ---

Prediction Results:

	month	sales	pred_value
44	2016-09-01	83334.0	82652.0
45	2016-10-01	88021.0	86214.0
46	2016-11-01	96295.0	96424.0
47	2016-12-01	117826.0	123061.0
48	2017-01-01	92220.0	92017.0
49	2017-02-01	82980.0	84862.0
50	2017-03-01	96812.0	93163.0
51	2017-04-01	88886.0	86166.0
52	2017-05-01	91373.0	91689.0
53	2017-06-01	91430.0	90023.0
54	2017-07-01	97700.0	96929.0
55	2017-08-01	44023.0	61322.0

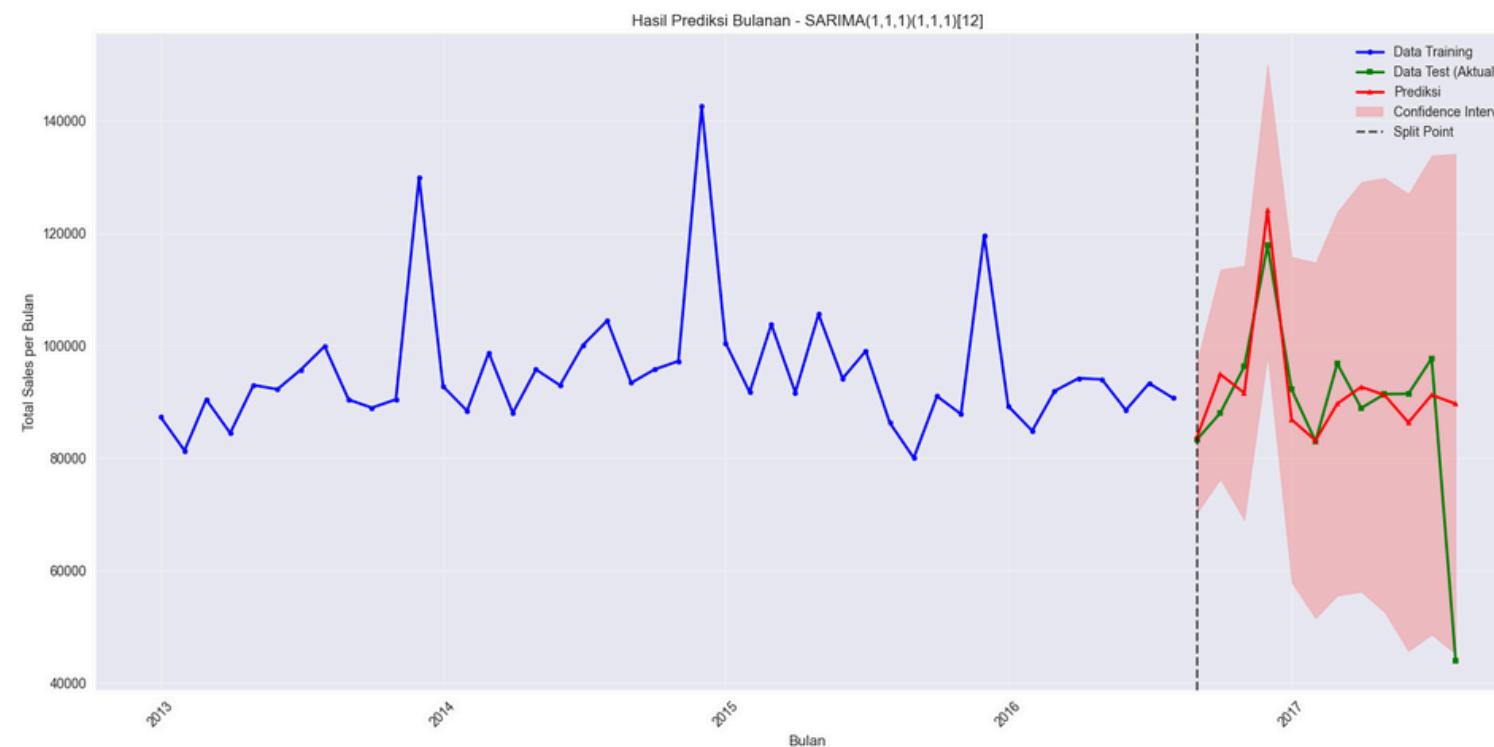
- Sebagian besar prediksi cukup dekat dengan nilai aktual.
- Model overestimate di Agustus 2017 → kemungkinan karena outlier atau tren mendadak menurun.
- Prediksi cukup akurat pada bulan-bulan tinggi (Des 2016, Jan 2017)

- Model berhasil menangkap pola musiman dan tren secara umum.
- Dapat digunakan untuk forecast jangka pendek, namun perlu peningkatan akurasi pada bulan-bulan fluktuatif.

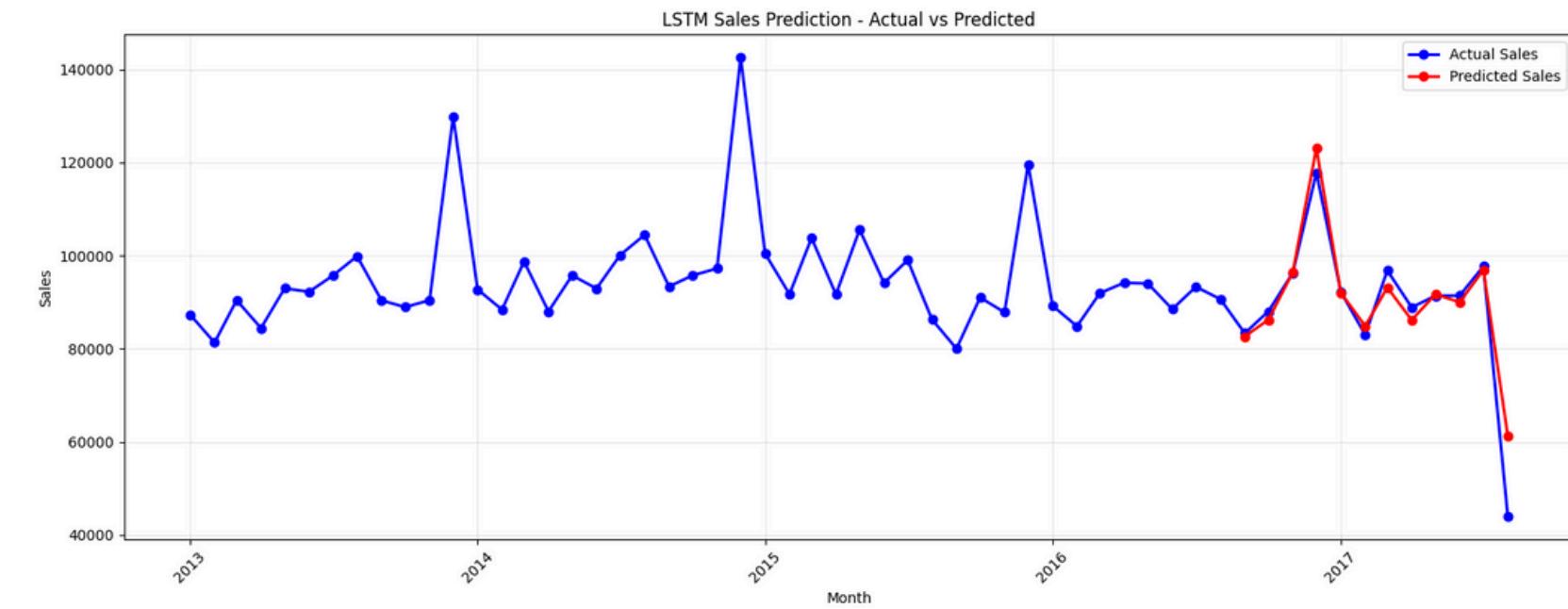


Perbandingan Model

SARIMA

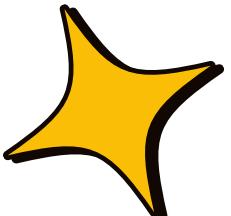


LSTM



Metric	LSTM	SARIMA
MAE	3008.33	4232
RMSE	5457.25	4972
MAPE	4.89%	4.42% 

- **SARIMA lebih stabil dan akurat secara persentase, cocok untuk data dengan pola musiman yang jelas.**
- **LSTM lebih unggul dalam mengurangi error absolut dan dapat menangkap pola non-linear yang kompleks.**



Real-world Application

Tujuan utama dari proyek ini adalah untuk meningkatkan efisiensi operasional dan penjualan dengan membantu manajemen memperoleh wawasan yang dapat ditindaklanjuti dari data mereka. Melalui pemodelan time series yang dan teknik deep learning, kami berupaya memahami tren penjualan jangka panjang, mengidentifikasi pola yang memengaruhi kinerja penjualan, serta memberikan solusi untuk tantangan dalam pengelolaan inventaris.



Mengembangkan Model Prediksi yang Akurat dan Reliable
Mencapai MAPE $\leq 15\%$ untuk prediksi penjualan bulanan



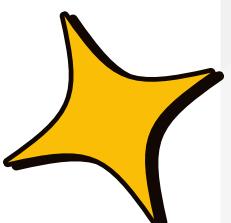
Operational Efficiency
Automate forecasting process and reduce manual effort



Business Value Creation
Inventory Optimization: Reduce stockout by 40% and excess inventory by 30%



Revenue Protection
Prevent lost sales dari availability issues



Future Improvement

Model Optimization & Tuning

Melakukan hyperparameter tuning (misalnya jumlah neuron, batch size, learning rate) untuk meningkatkan akurasi prediksi.



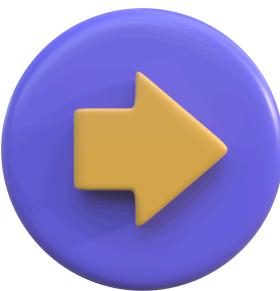
Visualisasi Interaktif & Dashboard

Mengembangkan dashboard dengan Power BI atau Tableau agar tim non-teknis dapat memantau prediksi penjualan secara real-time.



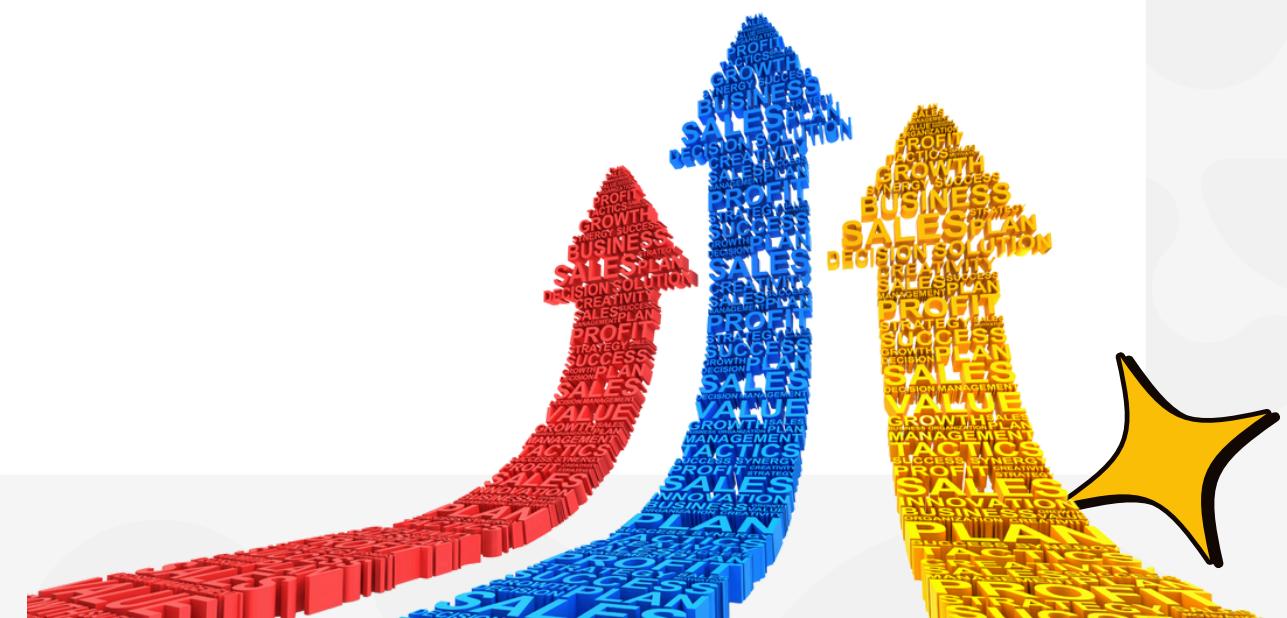
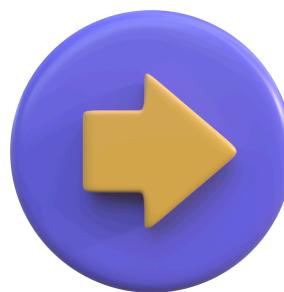
Penambahan Variabel Eksternal

Mengintegrasikan data cuaca, hari libur nasional, dan promosi sebagai variabel tambahan yang bisa memengaruhi penjualan.



Integrasi dengan Sistem Operasional

Menghubungkan hasil prediksi langsung ke sistem ERP atau SCM, sehingga bisa digunakan untuk otomatisasi pemesanan stok dan perencanaan distribusi.



Simpulan

Penelitian ini berhasil mengembangkan dan membandingkan dua pendekatan forecasting untuk optimasi inventaris Grocery I: **model statistik klasik (SARIMA)** dan **deep learning (LSTM)**. Kedua model menunjukkan performa yang sangat baik dengan **MAPE di bawah 5%**, menandakan tingkat akurasi prediksi yang excellent untuk kebutuhan bisnis.

Key Insights:

- **LSTM: Unggul dalam prediksi absolut (MAE lebih rendah 29%)**
- **SARIMA: Lebih stabil dan akurat secara persentase (MAPE & RMSE)**
- **Kedua model: Mencapai kategori "EXCELLENT" (MAPE < 5%)**
- **Actionable Insights: Clear understanding tentang pola demand Grocery I**
- **Inventory Optimization: Framework untuk mengurangi stockout dan overstock**
- **Scalable Solution: Architecture yang dapat diperluas ke kategori lain**





Project 1



Terima Kasih

Synergy Squad Team

