SUPPLY CHAIN ANALYSIS

GROUP - 06

ABHISHEK UDANSHIV
ARYAN ALPESH AGARWAL
CHAEYA LEE
HANEELA REDDY AVUTHU
MANAS GARG
SMRUTI JAGDISH JADHAV

AGENDA


Introduction
Dataset Explanation
Research Questions
Recommenda-tions
Conclusion
Future Steps

INTRODUCTION

Background & Objective:

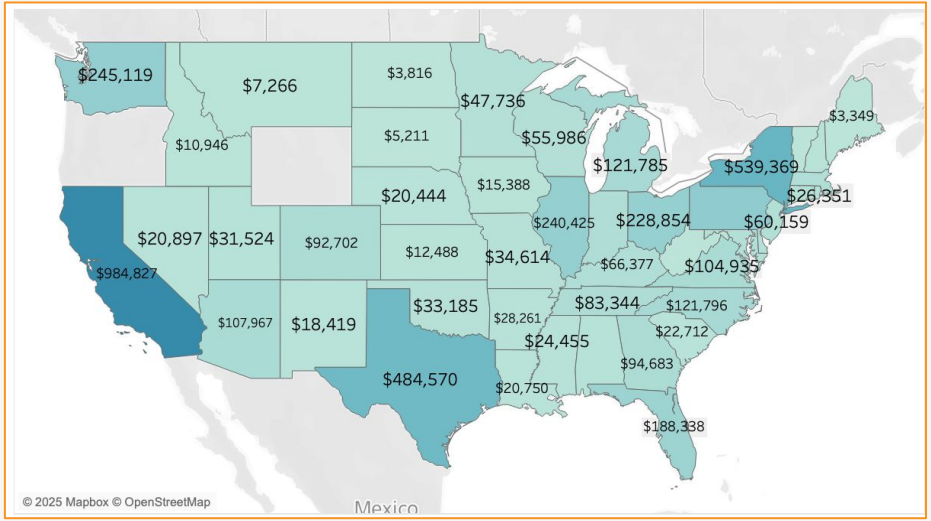
- A large transactional dataset from retail business
- Customer behavior
- Fraudulent activities
- Sales trends
- Statistical analysis and machine learning techniques.

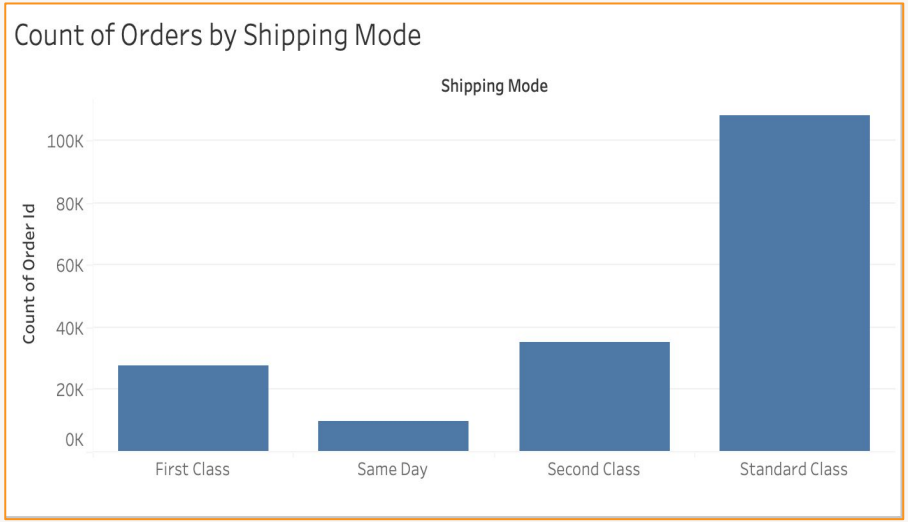
Expected Outcome:

- Targeted marketing for each segment
- Fraud prevention strategies
- Accurate sales forecasting
- Support data-driven decision-making in operations, finance, and customer relations.

DATASET EXPLANATION

- The Dataset is obtained from Mendeley Data
- Company Name : DataCo
- It consists of approximately 180,000 rows and 53 columns



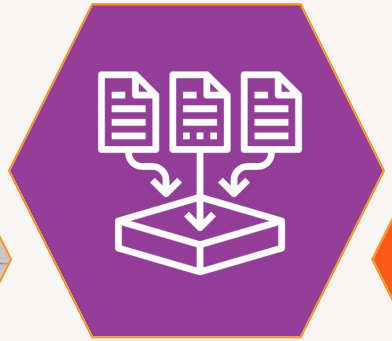


RESEARCH QUESTION #1

How can we segment customers based on their purchasing behavior,shipping preferences, and order history?

DATA PREPROCESSING

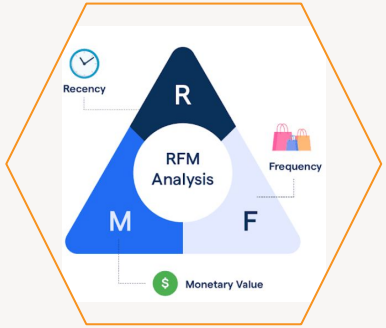
Column Standardization



Data Subsetting

Normalizing with Minmax()

Data Type Conversion



Recency, Frequency, Monetary Data Splitting

DATASET EXPLANATION

RFM Feature Summary:

- **Recency:** Avg = 220 days, Range = 0–1125
- **Frequency:** Avg = 3 orders, Max = 15
- **Monetary:** Avg = $1600, Max = $9436

	customer_id	recency	frequency	monetary
0	1	792	1	472.450012
1	2	136	4	1618.660042
2	3	229	5	3189.200037
3	4	380	4	1480.709993
4	5	457	3	1101.919998
...	...	...	...	...
20647	20753	0	1	161.869995
20648	20754	0	1	172.660004
20649	20755	0	1	314.640015
20650	20756	0	1	10.910000
20651	20757	0	1	34.980000

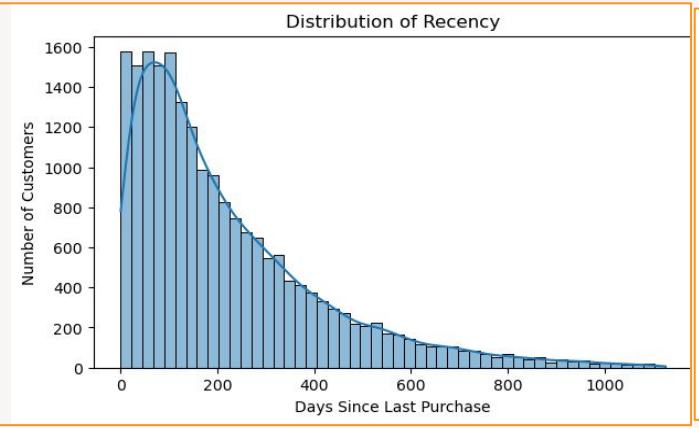
	recency	frequency	monetary
count	20652.000000	20652.000000	20652.000000
mean	220.089144	3.183808	1600.542436
std	199.395441	2.430699	1508.417956
min	0.000000	1.000000	8.470000
25%	75.000000	1.000000	254.940002
50%	159.000000	3.000000	1294.504997
75%	307.000000	5.000000	2621.140007
max	1125.000000	15.000000	9436.610088

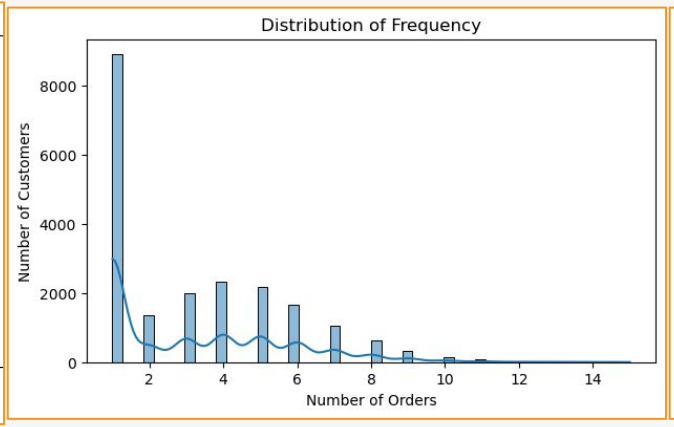
RFM ANALYSIS

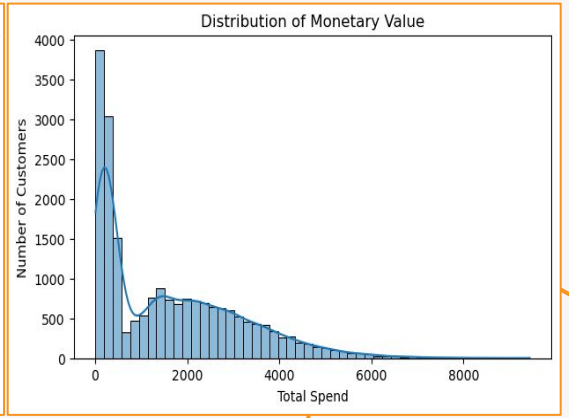
Recency: Most customers purchased recently; activity declines sharply over time.

Frequency: Majority of users placed only 1–3 orders, indicating low repeat behavior.

Monetary: Spending is skewed — a few customers account for high revenue.







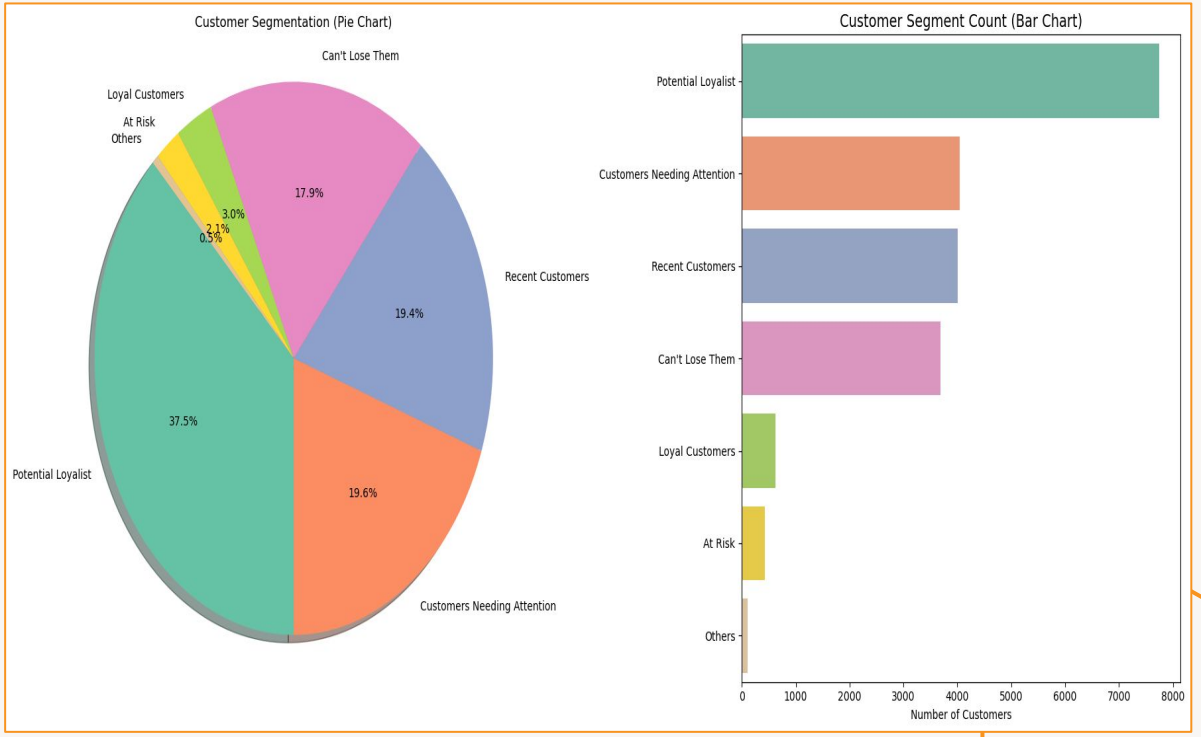
SEGMENTATION

Approach:

- Customers scored on **RFM** (Recency, Frequency, Monetary) metrics using quantile-based binning.
- Combined RFM scores used to classify customers into 8 named segments:
    - *Champions*, *Loyal*, *Potential*, *Recent*, *Needing Attention*, *At Risk*, *Lost*, *Others*

Most Common Segments:

- Potential Loyalist (7,745)
- Customers Needing Attention (4,041)
- Recent Customers (4,012)



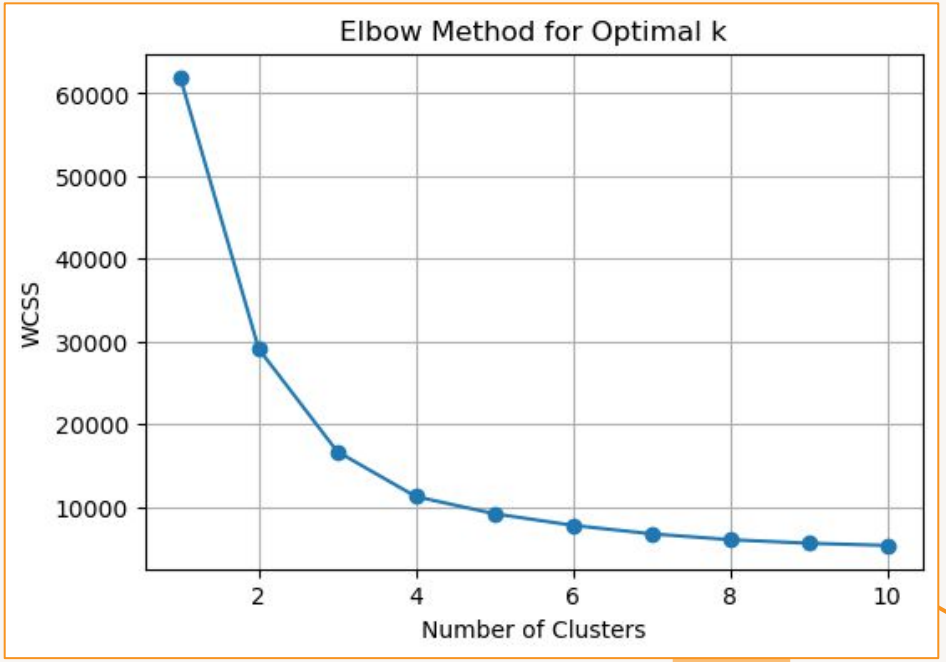
MODELING: OPTIMAL NUMBER OF CLUSTERS

Elbow Method Summary:

- Evaluated K values from 1 to 10 using **WCSS (Within-Cluster Sum of Squares)**
- Sharp drop observed until **k = 4**, after which the curve flattens
- Selected **4 clusters** as the optimal trade-off between performance and complexity

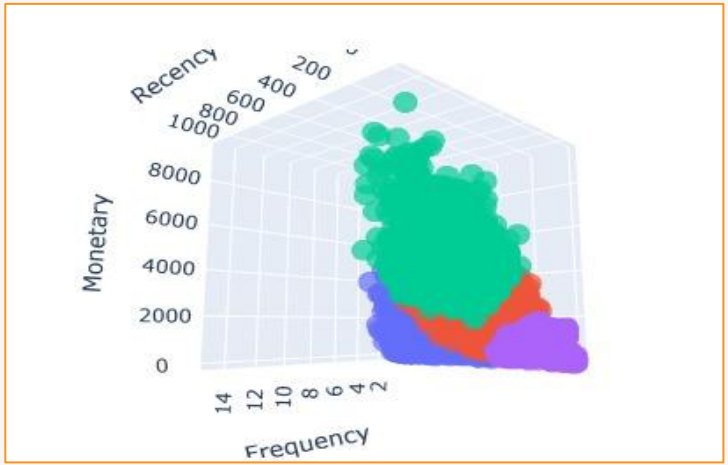
Visual Insight:

- Elbow curve clearly shows diminishing returns after 4 clusters



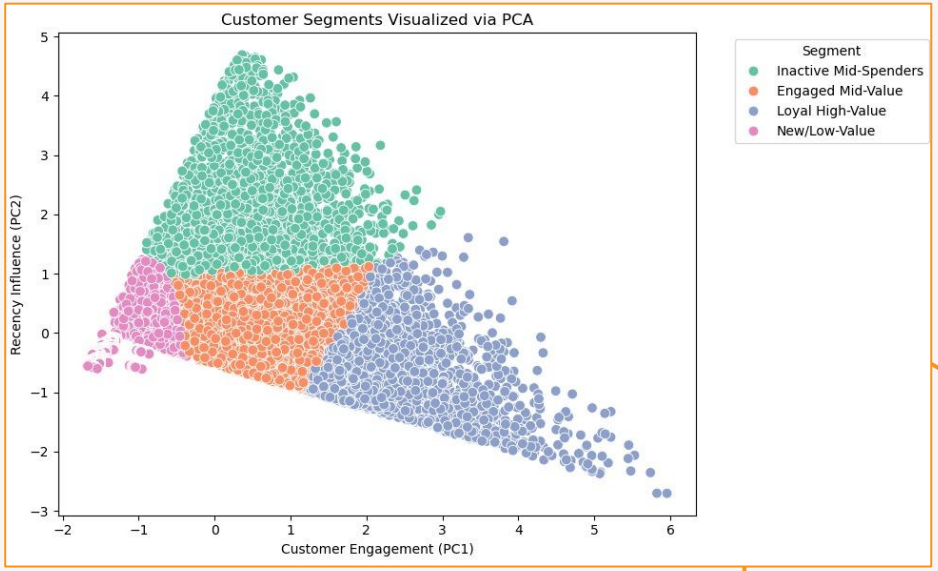
CUSTOMER SEGMENT INSIGHTS

Segment	Size	Frequency	Monetary ($)	Recency
Loyal High-Value	3,330	7.3	4,160.70	239.5
New/Low-Value	8,878	1.1	301.1	69.6
Inactive Mid-Spenders	2,373	2.7	1,365.80	642
Engaged Mid-Value	6,071	4.3	2,188.30	264.6



Visual Summary:

- PCA plot shows clean separation across 4 customer profiles.
- Clusters vary in size, spending potential, and engagement behavior.

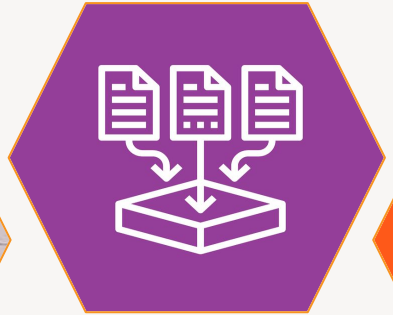


RESEARCH QUESTION #2

Can we predict future sales based on past transaction data, seasonality, and regional sales trends?

DATA PREPROCESSING

Null Values Removed



Aggregated daily sales

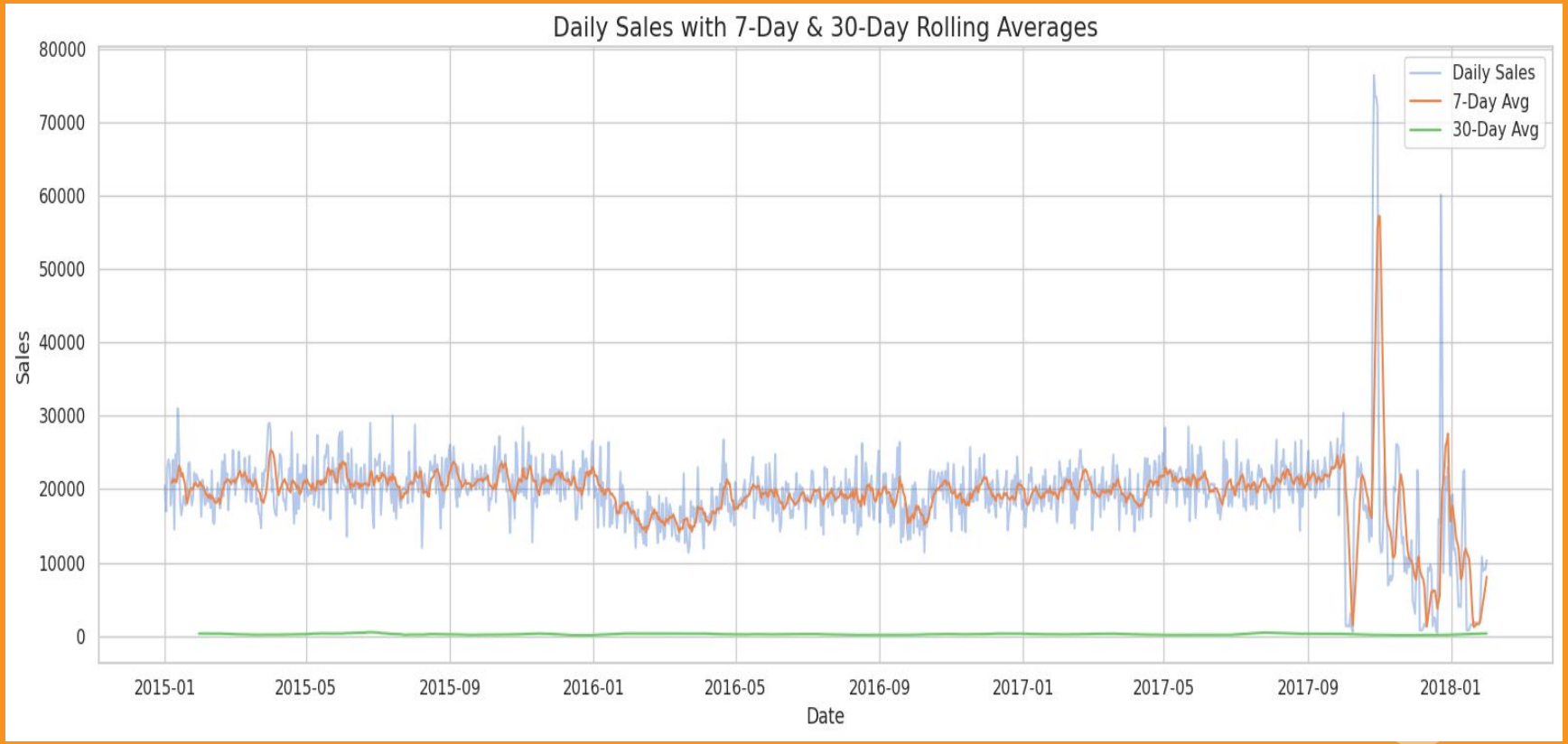
Reindexed missing dates

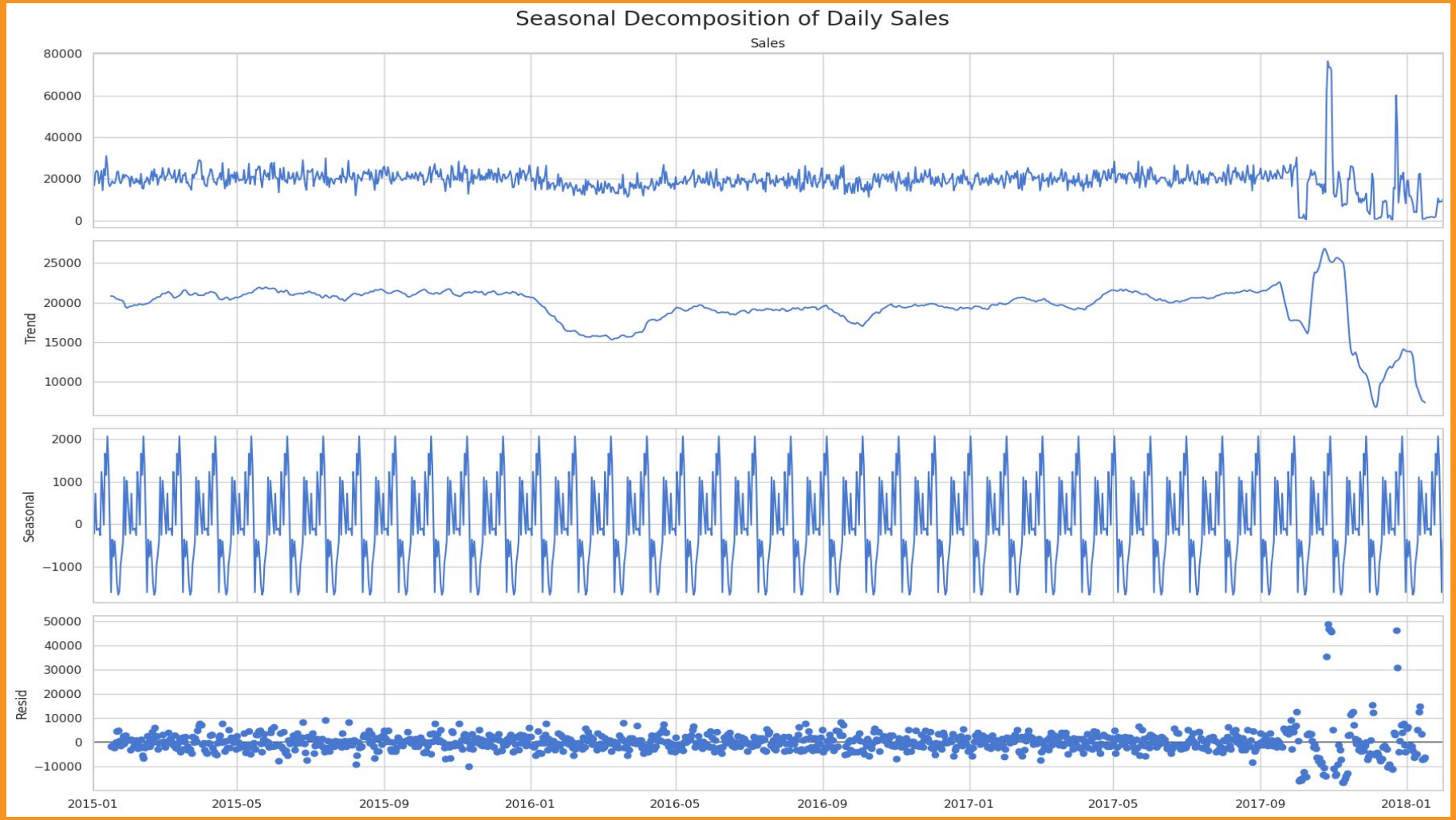
Converted order date to datetime

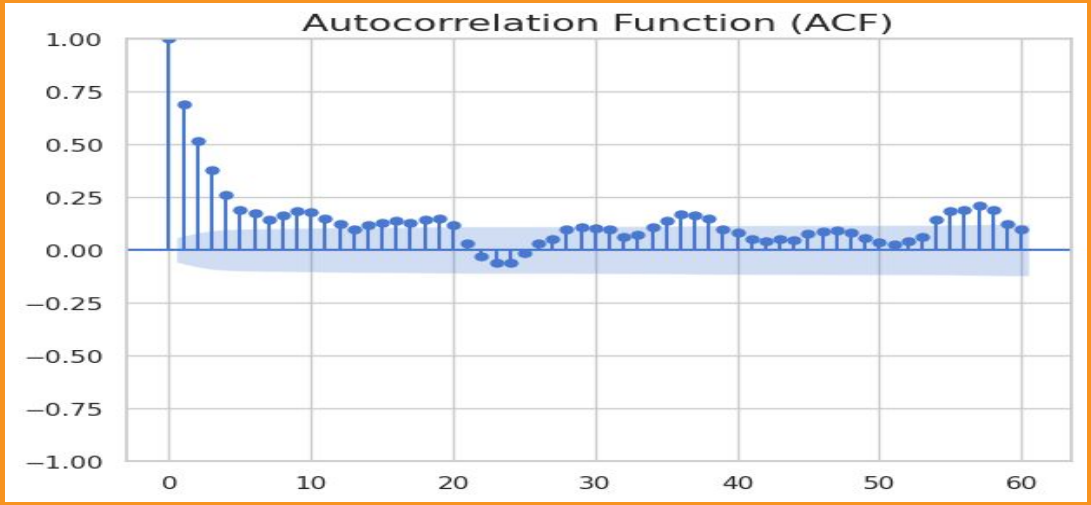
Splitting Data (Time Dependent)

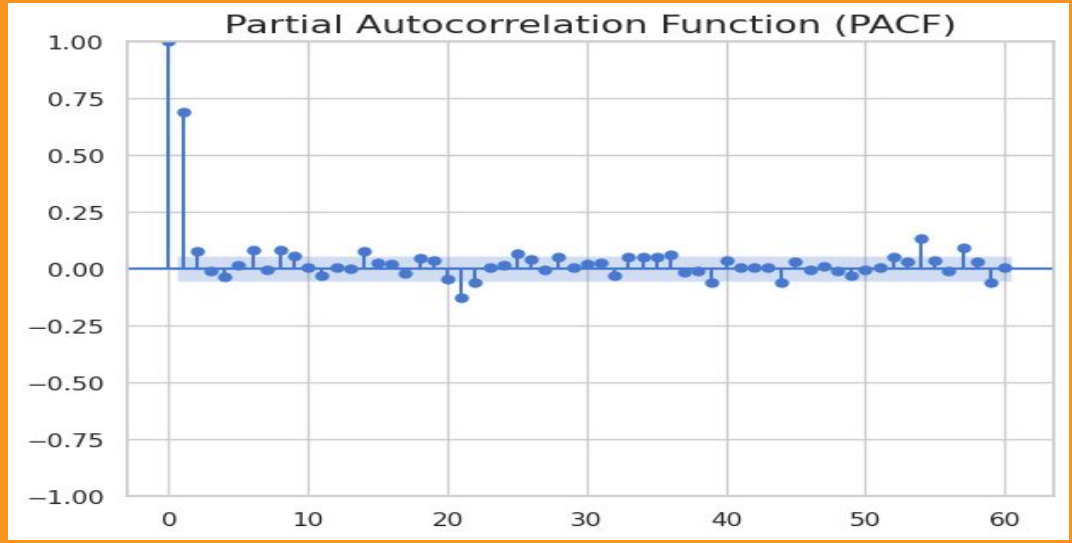
MODEL BUILDING

- **Used models:** Naive Forecasting, Moving Average, Simple Exponential Smoothing, ARIMA, Univariate LSTM, Multivariate LSTM, Stacked LSTM, Ensemble Forecasting.
- **Evaluated with metrics:** RMSE, MAE.
- **Hyperparameter tuning using** ARIMA orders (p, d, q), LSTM architecture (layers, units, dropout).
- **Validation approach:** Train/validation/test splits, early stopping, residual diagnostics.


Daily Sales with 7-Day & 30-Day Rolling Averages
Daily Sales
7-Day Avg
30-Day Avg
Sales
80000
70000
60000
50000
40000
30000
20000
10000
0
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
Date


Seasonal Decomposition of Daily Sales
Sales
80000
60000
40000
20000
0
Trend
25000
20000
15000
10000
Seasonal
2000
1000
0
−1000
Resid
50000
40000
30000
20000
10000
0
−10000
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01


Autocorrelation Function (ACF)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
60


Partial Autocorrelation Function (PACF)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50
60

BASELINE MODELS

1. Naïve Forecast

- Predicts next day's sales as today's value
- RMSE: **7,534.70**, MAE: **3,978.51**
- Surprisingly effective due to stable short-term sales before 2017

2. Moving Average (7-Day)

- Uses mean of past 7 days for prediction
- RMSE: **11,891.98**
- Failed to capture sudden drops and spikes; oversmooths

3. Simple Exponential Smoothing (SES)

- Models short-term level without trend or seasonality
- RMSE: **12,035.98**
- Slightly worse than moving average due to lack of adaptability

Key Takeaway:

- These baseline models performed reasonably during stable periods

ARIMA - AutoRegressive Integrated Moving Average

- Input: 45-day sliding window of past sales
- Data scaled using MinMaxScaler
- Architecture: LSTM → Dropout → Dense
- Trained on 80% data with early stopping
- **Model Performance:**
- RMSE: **12,049.21**
- MAE: **7,680.37**

UNIVARIATE LSTM

- Scaled sales values using MinMaxScaler
- Created 45-day sliding windows for input sequences
- Built LSTM → Dropout → Dense model
- Trained using 80:20 split and tracked loss/validation loss
- **Model Performance:**
- RMSE: **11,201.60**
- MAE: **6,895.11**

MULTIVARIATE LSTM

- Added time-based (day, month) and lag-based features
- Trained LSTM on multivariate sequences
- No improvement in prediction accuracy by modeling feature dependencies
- **Model Performance:**
- RMSE: **11,829.26**
- MAE: **7,252.62**

STACKED LSTM

- Engineered features: Lag1, Lag2, 3-day MA, 7-day MA, Day, Month, Year
- Used stacked architecture: LSTM (64) → LSTM (32) → Dense
- Train/val/test split: 70/10/20
- Trained for 30 epochs with dropout for regularization
- Early stopping stabilized training and restored the best-performing weight.
- RMSE: 11,523 and MAE: 7,179

ENSEMBLE

= LSTM + Naïve Ensemble Forecast

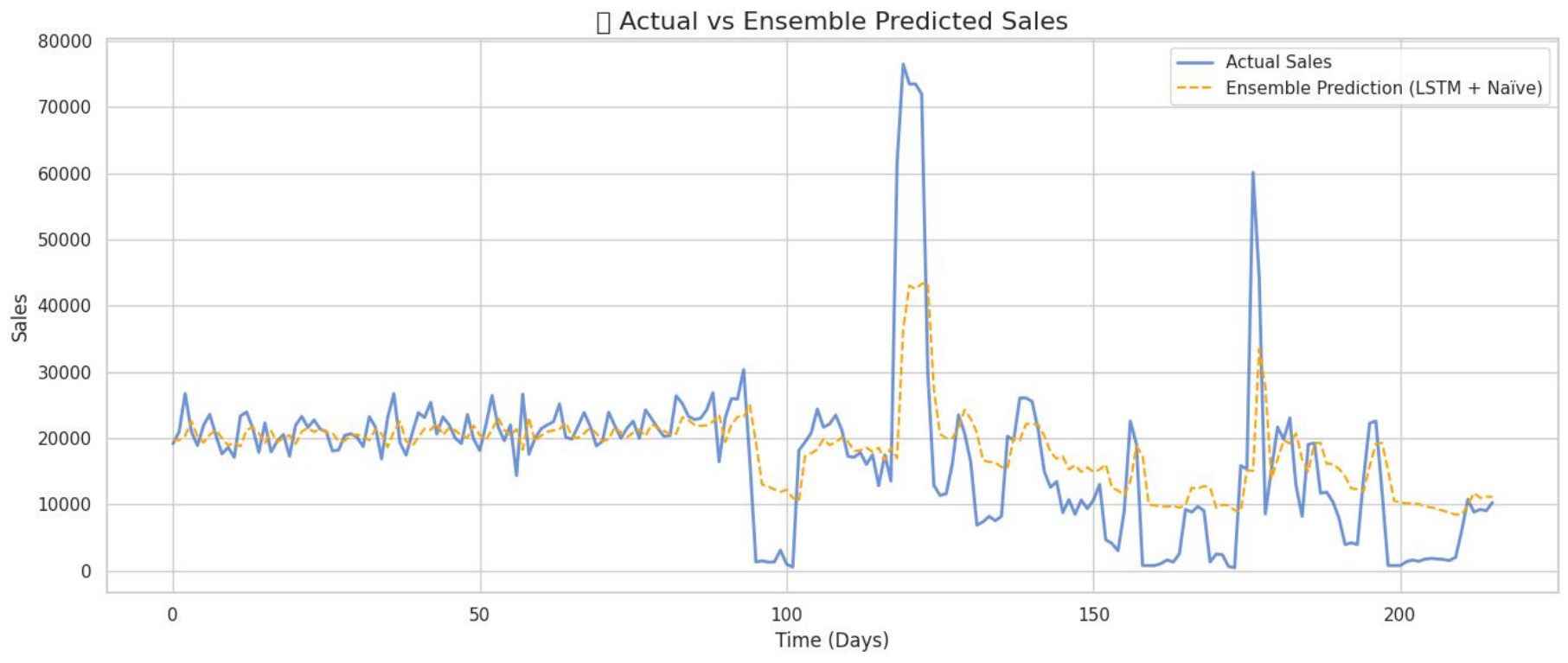
- LSTM's ability to capture temporal patterns + Naïve model’s short-term accuracy

🧮 Formula:

```
y_ensemble = α × y_pred_lstm + (1 - α) × y_pred_naive
```

- α = 0.6
- Gave the LSTM model 60% of the weight and the naïve model 40%.
- Reduced RMSE to **8,513** and MAE to **5,410**

ENSEMBLE LSTM



RESULTS & EVALUATION

- **Model performance heavily impacted by data imbalance**
  a. Sudden, extreme sales spike in mid-2017 caused inflated RMSE across all models
  b. Models predicted well during normal periods but failed to capture post-spike volatility

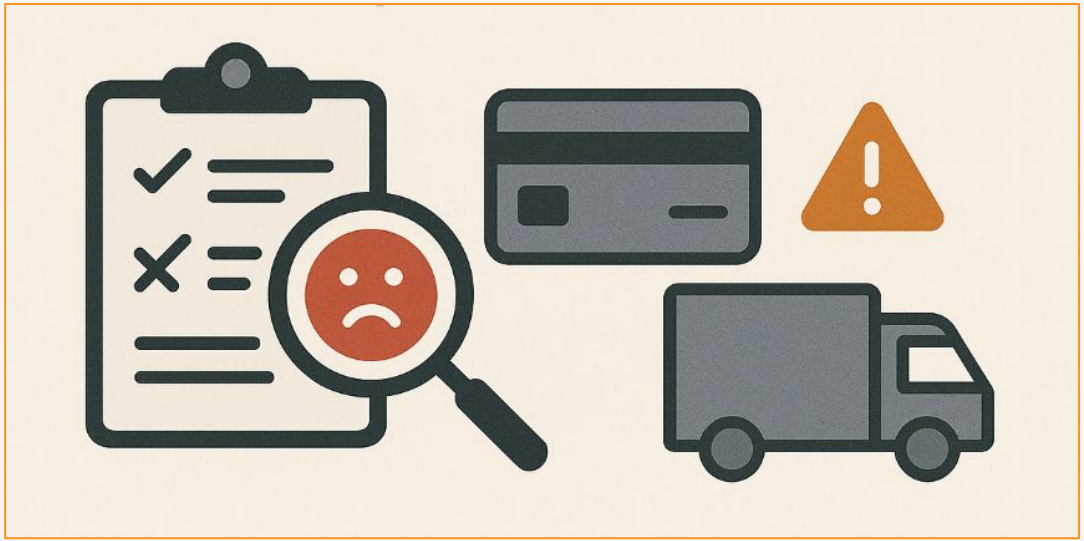
- **RMSE not a fair sole metric** in this case
  a. High RMSE ≠ poor model, but rather reflects **extreme variance** and **anomalies**
  b. MAE offers better stability and interpretability in this case

- **Best Model**:
  a. **Ensemble (LSTM + Naïve)**
  b. RMSE: 8,513.25, MAE: 5,409.59
  c. Balanced between overfitting and underfitting

- **Key Insight**:
  Forecasting performance can only be as good as the stability and representativeness of the data.

RESEARCH QUESTION #3

Can we identify fraudulent transactions based on order patterns, payment types, and late delivery risks?



DATA PREPROCESSING

Null Values Removed

Splitting Data (Stratified Split)

Encoding Categorical Variables (Ordinal Encoding)

Handled Imbalanced Data Using SMOTE

MODEL BUILDING

- **Used models:** Random Forest, XGBoost, LightGBM, CatBoost, ADABoost, Gradient Boost

- **Evaluated with metrics:**
  - **Recall:** to measure how well the model detects actual frauds
  - **F1 Score:** which balances precision and recall for fraud detection
  - **Accuracy (Generalisation):** to assess overall performance and generalization to unseen data

NEURAL NETWORK MODELING INSIGHTS

Artificial Neural Networks (ANNs) using various **hyperparameter tuning strategies**, including:

- Adding **batch normalization** and **dropout layers**
- Adjusting **% of Dropout Layers for regularisation / preventing our model from overfitting**
- Testing different **optimizers** and **activation functions**

CONCLUSION:

- Despite extensive tuning, the model achieved a maximum of 97% accuracy & recall
- Underperformed Baseline Model.

MODEL COMPARISON

Model	Accuracy	ROC-AUC	Precision	Recall	F1
Random Forest	0.9849	0.9873	0.9898	0.9948	0.9923
CatBoost	0.9875	0.9895	0.991	**0.9962**	0.9936
ANN	0.9531	0.9751	1	0.96	0.98

RESULTS & EVALUATION

Recommendation:

- **Best Overall:**
  CatBoost — it has the highest balance of precision, recall, and F1-score, with outstanding accuracy.

- **Best Business-Friendly Tradeoff:**
  Random Forest — if you want strong performance across all metrics with simplicity and interpretability.

RECOMMENDATIONS & FUTURE STEPS

Automate Segmentation and Targeting	Integrate Real-Time Fraud Monitoring	Refine Forecasting Models	Optimize Shipping Strategies	Invest in Data Infrastructure
Use the segmentation model results to develop dynamic marketing campaigns (e.g., loyalty rewards for high-value customers, incentives for low-frequency buyers).	Deploy the fraud detection pipeline in real-time systems to flag suspicious transactions before they are processed, especially focusing on high-risk shipping types.	Incorporate external variables like promotions, holidays, and competitor pricing into the sales prediction models to improve forecast accuracy.	Address the high late delivery rate (especially in Standard Class, which accounts for ~60% of delays) by negotiating with logistics providers or promoting alternative shipping options.	Build a centralized analytics platform with automated ETL pipelines, model retraining schedules, and a dashboard for continuous monitoring of key KPIs.

Conclusion

Customer Segmentation

By leveraging clustering techniques such as K-Means, we successfully segmented customers into **four distinct groups** based on purchasing behavior, shipping preferences, and order history.

Cluster 0: Loyal High-Value

Cluster **1: New/Low-Value**

Cluster 2: inactive Mid-Spenders,

Cluster 3: Engaged Mid-Value

kmeans_cluster	recency	frequency	monetary	num_customers
0	239.5	7.3	4160.7	3330
1	69.6	1.1	301.1	8878
2	642.0	2.7	1365.8	2373
3	264.6	4.3	2188.3	6071

These analysis revealed clear patterns among different customer types:

- customers Cluster 1 New/low value, Despite low individual value, this is the **largest segment**, representing a significant opportunity for growth if engaged properly.

These insights enable businesses to **personalize marketing strategies**, **optimize shipping options**.

Conclusion

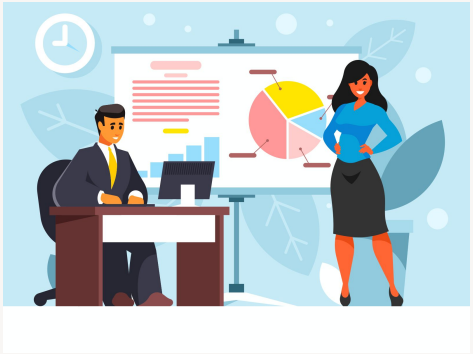
Fraud Detection

To detect fraudulent activities, we applied machine learning techniques that analyzed customer **payment methods**, **order frequency**, and **delivery behavior**. The models identified outliers and suspicious patterns that deviated from normal purchasing behavior, such as sudden high-value orders or unusual shipping addresses.

By flagging these anomalies early, businesses can take **proactive measures to investigate and prevent fraud**, reducing potential revenue loss and maintaining customer trust. The system provides a scalable, data-driven framework that can continuously learn and adapt to new fraud patterns.

Conclusion

Sales Forecasting



Using time series analysis we forecasted future sales based on **historical data**, capturing **seasonal trends**, **quarterly fluctuations**, and **regional demand variations**. The predictive insights help businesses **align inventory**, **plan marketing efforts**, and **optimize supply chain operations**.

Accurate forecasting supports better decision-making by anticipating demand surges or slowdowns, thereby improving **resource allocation**, **minimizing waste**, and **maximizing revenue potential**.

THANK YOU!



QUESTIONS?