MARKET BASKET ANALYSIS

PRESENTED TO

Aruna Mam

Ananya Mam

PRESENTED BY

Amit Kumar Makkad

Mihir Karandikar

Problem Definition

Our problem is to analyze customer purchase behavior, recommend products based on their purchase history, and forecast sales for an e-commerce company.

Proposed Solution

Data Understanding

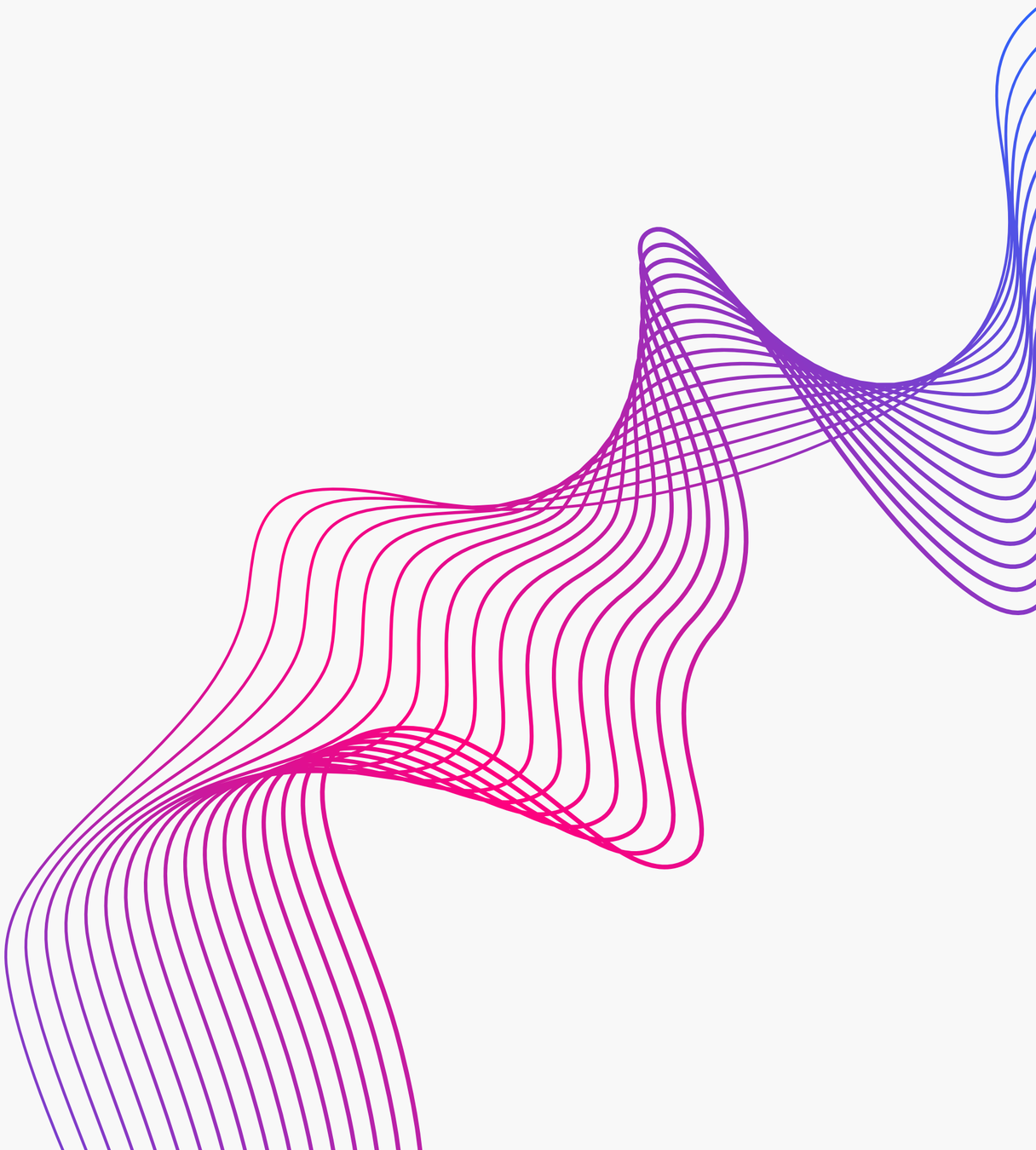
It involves preprocessing of data also

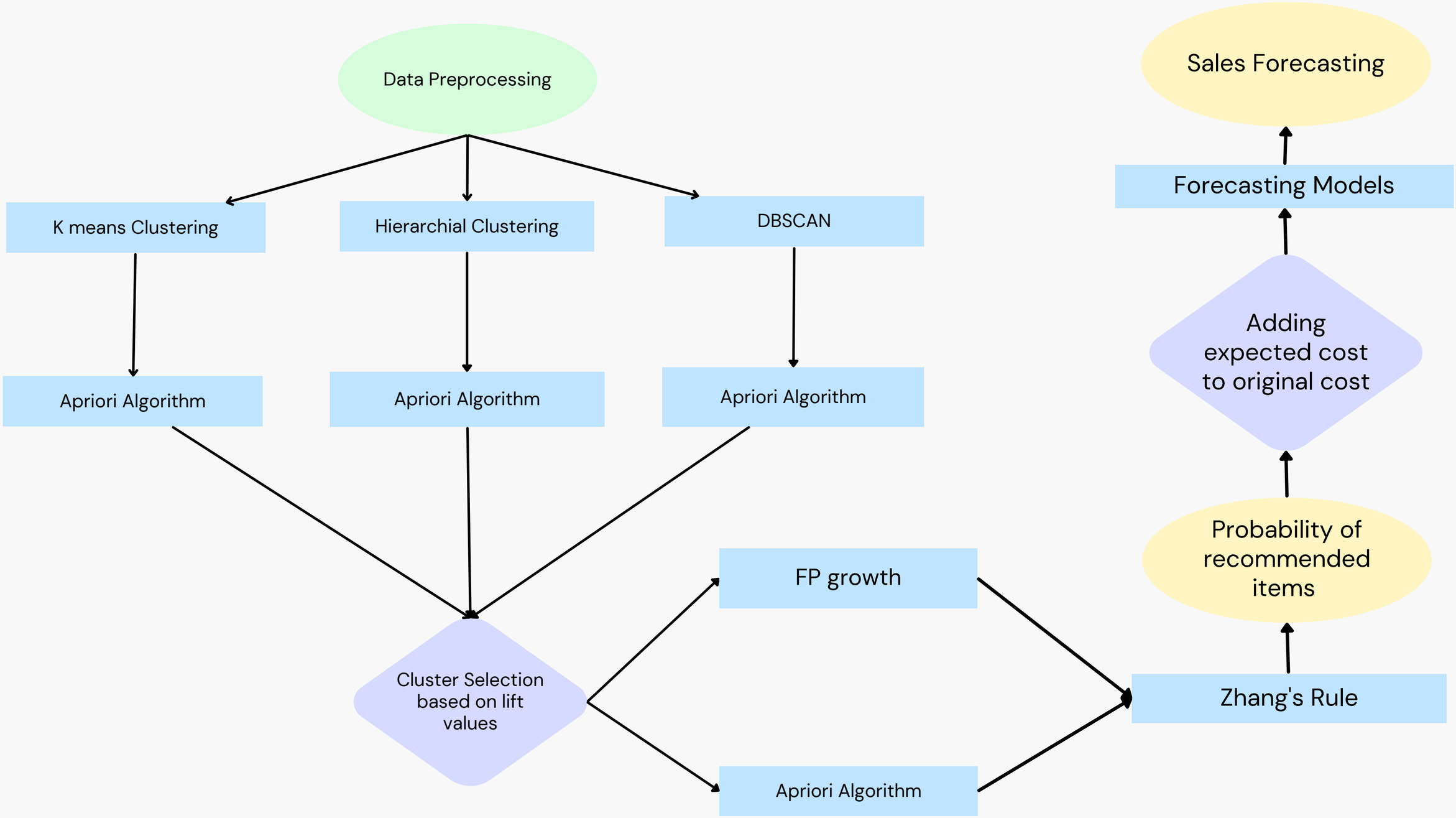
Customer purchase behaviour

First, we make customer segments using clustering
Second, we find association rules between items
Third, we find probability of buying recommended product

Sales forecasting

For a basket, we add the expected cost of buying recommended product
Then we used forcasting model to find expected sales




Data Preprocessing
K means Clustering
Hierarchial Clustering
DBSCAN
Apriori Algorithm
Apriori Algorithm
Apriori Algorithm
Cluster Selection based on lift values
FP growth
Apriori Algorithm
Zhang's Rule
Probability of recommended items
Adding expected cost to original cost
Forecasting Models
Sales Forecasting

E Commerce Data

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12-01-2010 08:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12-01-2010 08:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12-01-2010 08:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12-01-2010 08:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12-01-2010 08:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12-01-2010 08:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12-01-2010 08:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	12-01-2010 08:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	12-01-2010 08:28	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12-01-2010 08:34	1.69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12-01-2010 08:34	2.1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12-01-2010 08:34	2.1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12-01-2010 08:34	3.75	13047	United Kingdom
536367	22310	IVORY KNITTED MUG COSY	6	12-01-2010 08:34	1.65	13047	United Kingdom

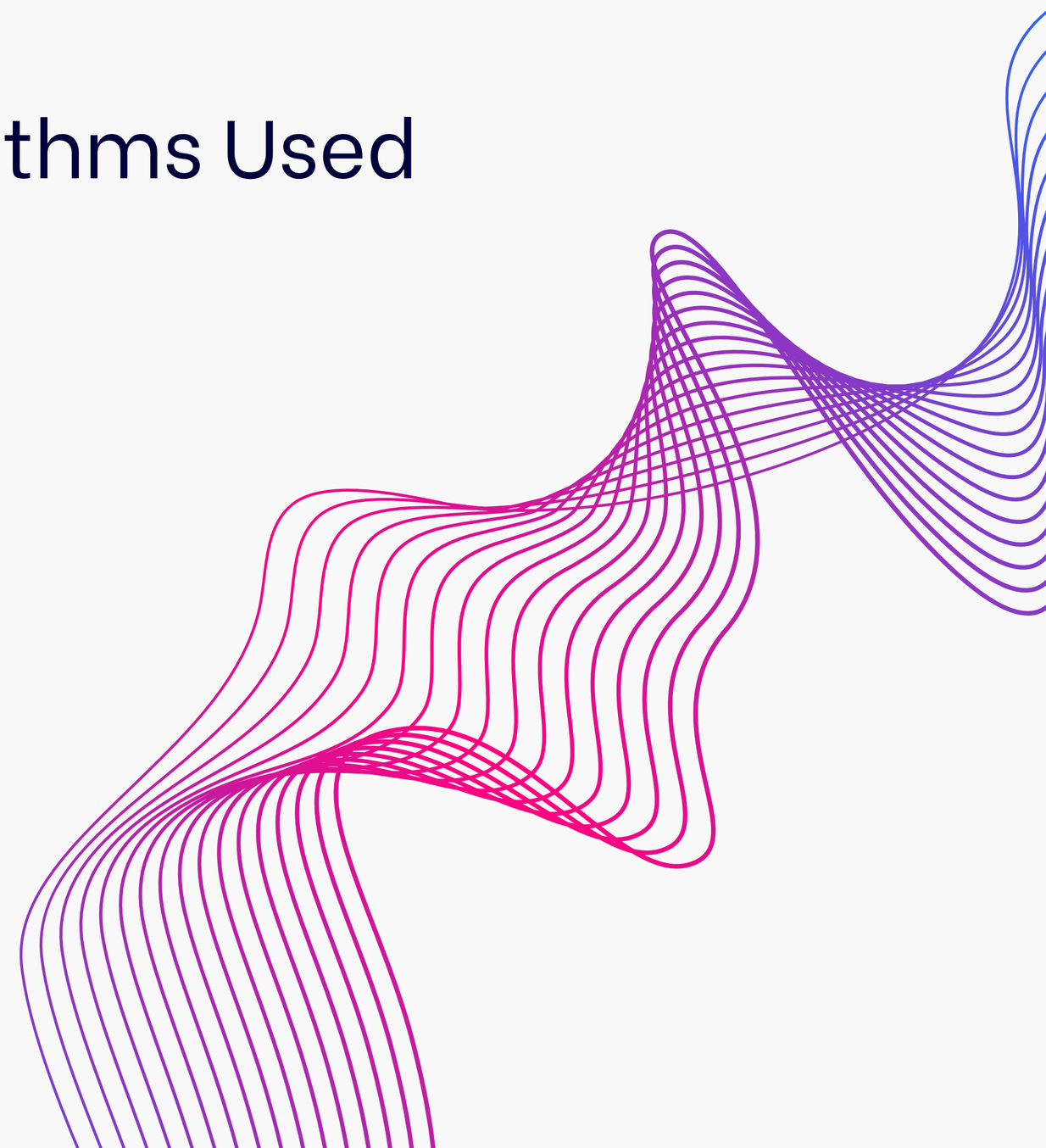
Source: kaggle

Different Clustering Algorithms Used

K means Clustering

DB Scan

Hierarchical Clustering



K Means Clustering

1	K-Means clustering is an unsupervised algorithm that clusters the input data points into K different clusters groups based on direct or indirect patterns present in the data.
2	We start with K random cluster centroids, and in every iteration, we assign the data points to the closest centroid and then recalculate the cluster centroids.
3	The algorithm runs until the centroids are unchanged across multiple iterations or a certain number of iterations has been reached.
4	The algorithm’s performance depends on the initial selection of cluster centers, and the number of clusters K must be given beforehand.

DB Scan

1	DBSCAN primarily uses two metrics, the neighborhood size and density threshold.
2	Based on these, it clusters data points with many neighbors by first constructing a neighborhood around each data point based on the neighborhood size.
3	It has several advantages over the K-Means algorithm, such that it can locate clusters of any shape and does not need the number of clusters to be predetermined.
4	Also, it outperforms K-Means in situations where clusters have different forms and densities.

Hierarchical Clustering

1	An unsupervised machine learning approach, hierarchical clustering, combines or divides clusters based on a similarity score to group comparable data points together.
2	The algorithm produces a dendrogram that displays the order and distance of merges or splits, and there are two forms of hierarchical clustering: agglomerative and divisive.
3	It has many applications but can be computationally expensive and sensitive to the chosen connection criterion and similarity metric.

Parameters to understand strenght between associations

Support is the fraction of the total number of transactions in which the itemset occurs

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

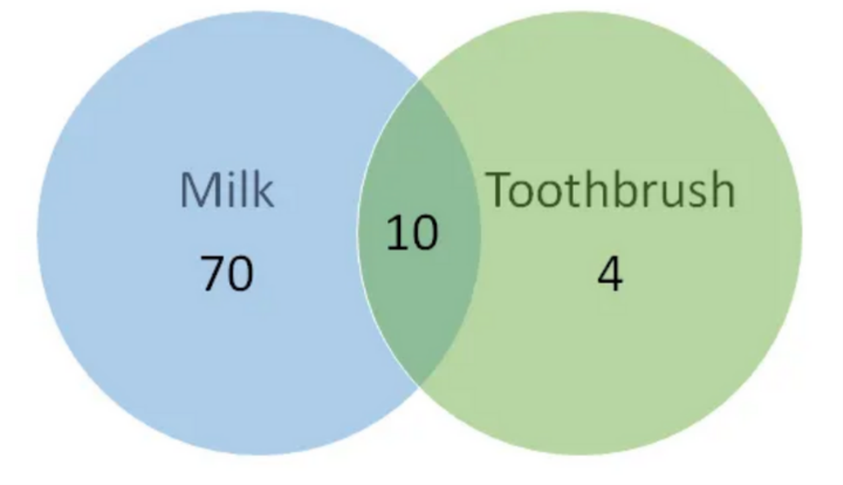
Confidence is the conditional probability of occurrence of consequent given the antecedent.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

If the frequency of consequent is very high then no matter what antecedent is, the confidence value is very high.

Lift is the rise in probability of having {Y} on the cart with the knowledge of {X} being present over the probability of having {Y} on the cart without any knowledge about presence of {X}. Mathematically,

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$



Confidence for {Toothbrush} → {Milk} will be 10/(10+4) = 0.7 as it is between 0 to 1, 0.7 means good association

Lift for {Toothbrush} → {Milk} will be 0.7/0.8=0.87 < 1 which means no association

Zhang's metric is a metric to measure the quality of association rules, that can be used to identify not only frequent itemsets but can identify different types of associations like complementarity, exclusivity, and causality between the items in the itemset.

For example, A positive Zhang value, indicates that item A and item B are complementary, meaning that if a customer purchases one of these items, they are more likely to purchase the other item. While negative zhang value denotes that A and b are exclusive.

$$Zhang(A,B) = (support(A,B) / support(B)) - support(A)$$

support(A) = 0.6
support(B) = 0.4
support(A, B) = 0.3

lift(A -> B) = (0.3 / 1) / (0.6 * 0.4) = 1.875
lift(B -> A) = (0.3 / 1) / (0.4 * 0.6) = 1.875
zhang(A -> B) = (0.3 / 0.6) - 0.4 = 0.1
zhang(B -> A) = (0.3 / 0.4) - 0.6 = -0.2

Association Rule Mining - Apriori Algorithm

Apriori algorithm is used to discover associations and patterns in transactional data.

It uses a level-wise approach, starting with single itemsets and progressively building more significant itemsets by combining smaller frequent itemsets.

Then we select subsets of frequent itemsets and calculate their support and confidence.

Then we can find items bought together frequently based on passing the minimum threshold value of support, confidence, and lift.

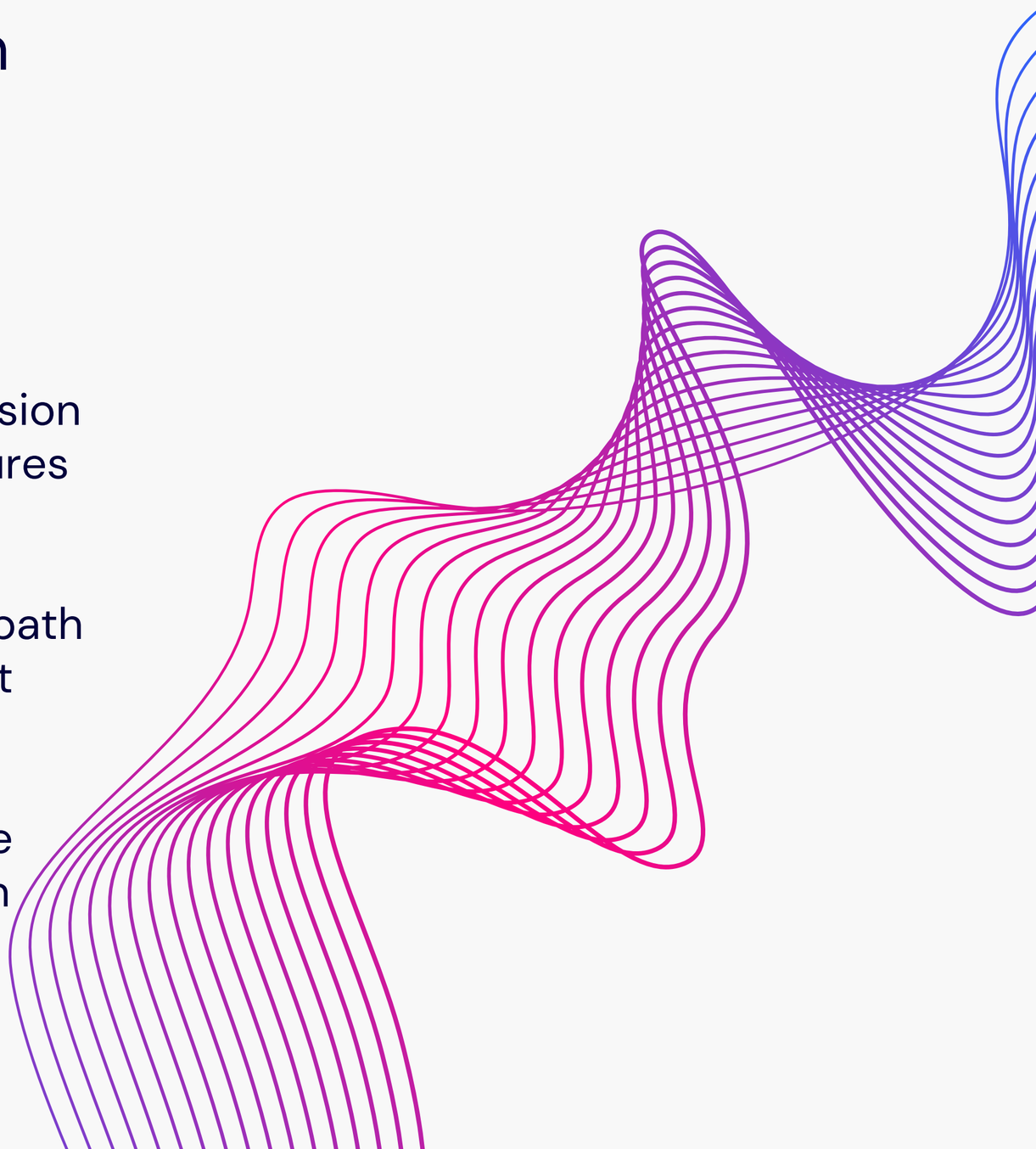
Association Rule Mining - FP growth

A data mining method called FP-growth identifies recurring trends in transactional data.

It functions by building an FP-tree, a compact version of the transaction database that effectively captures all frequent itemsets.

It then builds conditional FP trees for each prefix path and iteratively mines the tree to produce frequent itemsets.

This makes it more scalable and effective than the Apriori approach, particularly for big datasets with many elements.



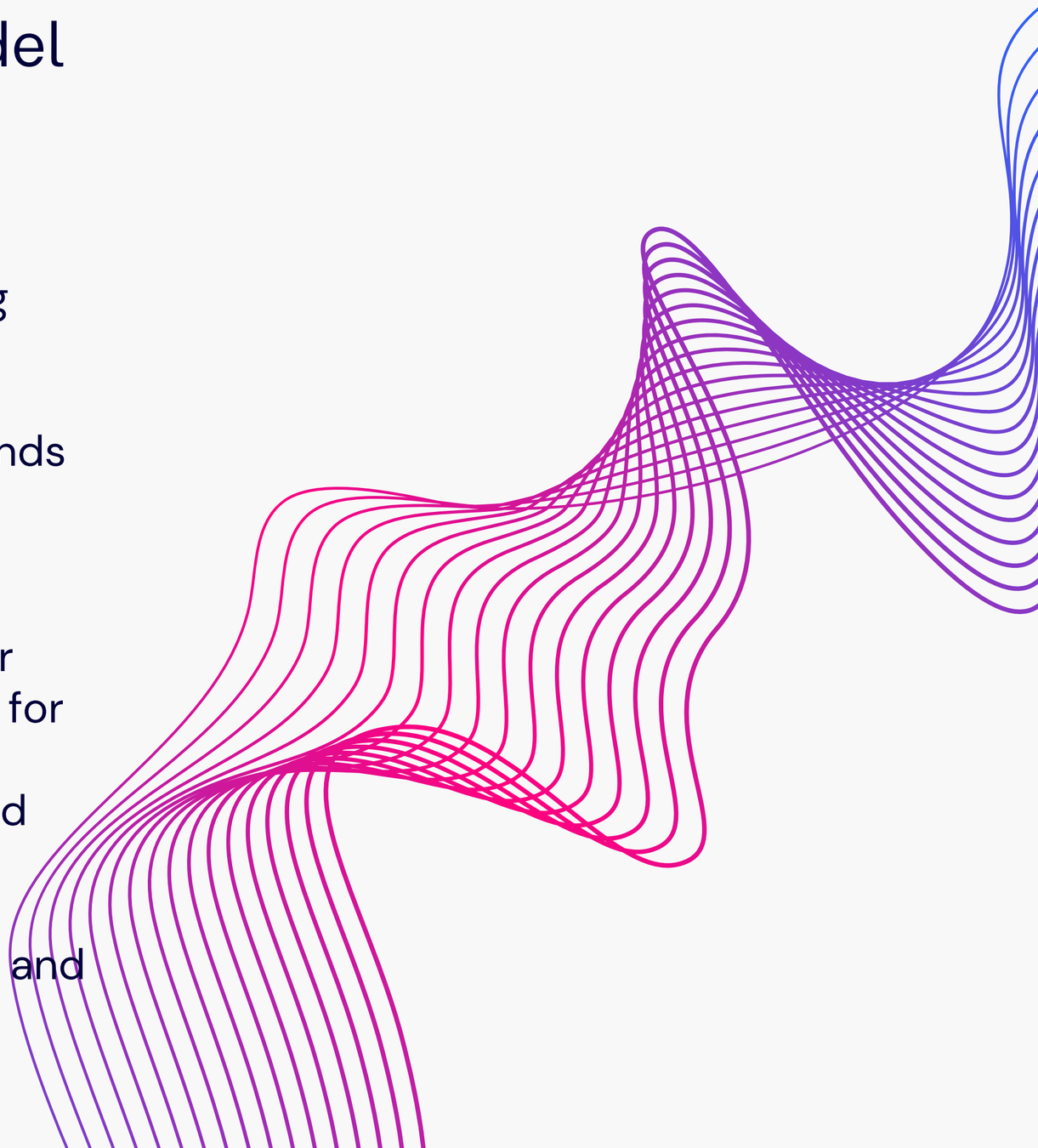
Time Series Forcasting - ARIMA Model

ARIMA (Autoregressive Integrated Moving Average) model is used to forecast future sales using existing sales data

It uses three techniques to detect patterns and trends in the data: moving average, differencing, and autoregression.

The moving average component models the error or residual term, the differencing component corrects for trends and seasonality, and the autoregressive component models the link between the current and previous observations.

The model can be trained using previous sales data and used to predict sales in the future.



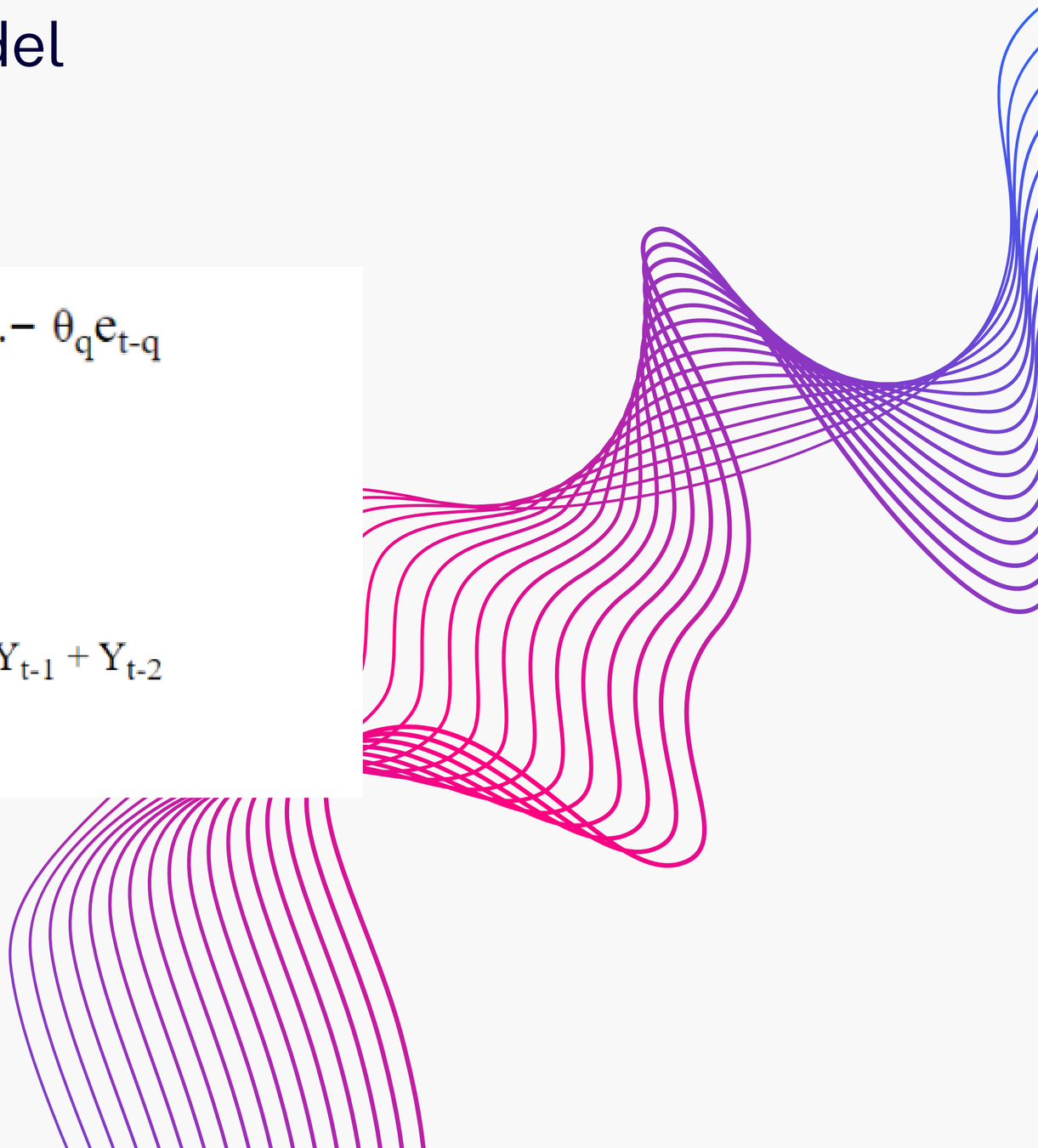
Time Series Forcasting - ARIMA Model

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

If d=0: $y_t = Y_t$

If d=1: $y_t = Y_t - Y_{t-1}$

If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$



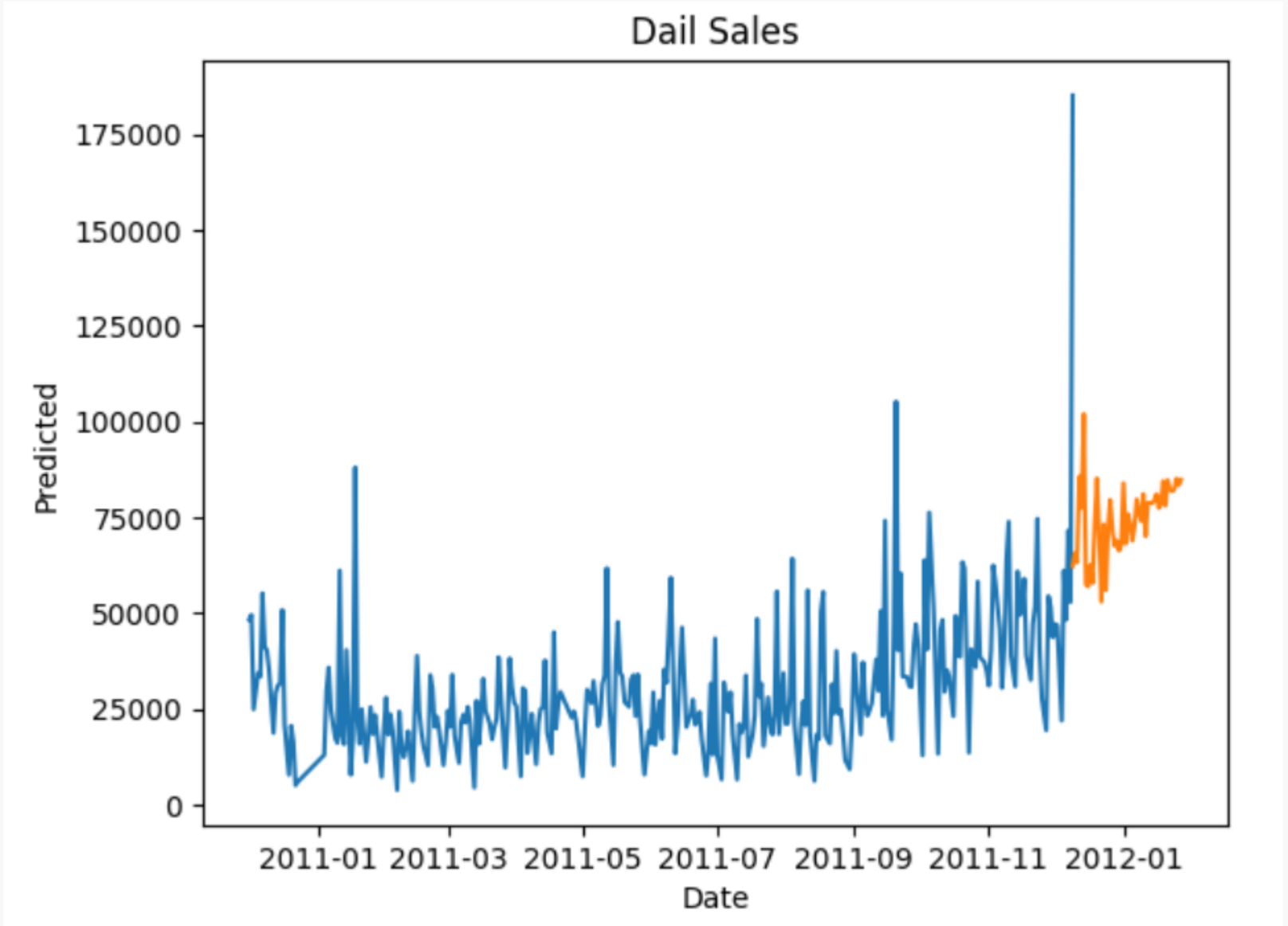
Performance Measurement Index

To choose the best from the results of the clustering algorithms, we pass the clusters formed by each one of them to the Apriori algorithm. Now, we check the lift values for common item sets. Higher lift values validate that the clusters were created appropriately. Finally, clusters formed from **K means Clustering** are choosen.

To choose the best from the results of the association algorithms, we find the zhang metric for all common association rules formed by Apriori and fp growth algorithm. Higher the zhang value, means items are more associated. Finally, association rules given by **apriori algorithm** is chosen.

For evaluating the time series forecasting model, rmse (root mean square error) is used.

Test RMSE: 6456.943


Dail Sales
175000
150000
125000
100000
75000
50000
25000
0
Predicted
2011-01
2011-03
2011-05
2011-07
2011-09
2011-11
2012-01
Date

References:-

Dataset

Apriori Algorithm-AN IMPROVED APRIORI ALGORITHM FOR ASSOCIATION RULES Mohammed Al-Maolegi1 , Bassam Arkok2

Time series forecasting-Time Series Clustering: A Superior Alternative for Market Basket Analysis Swee Chuan Tan & Jess Pei San Lau

Clustering