

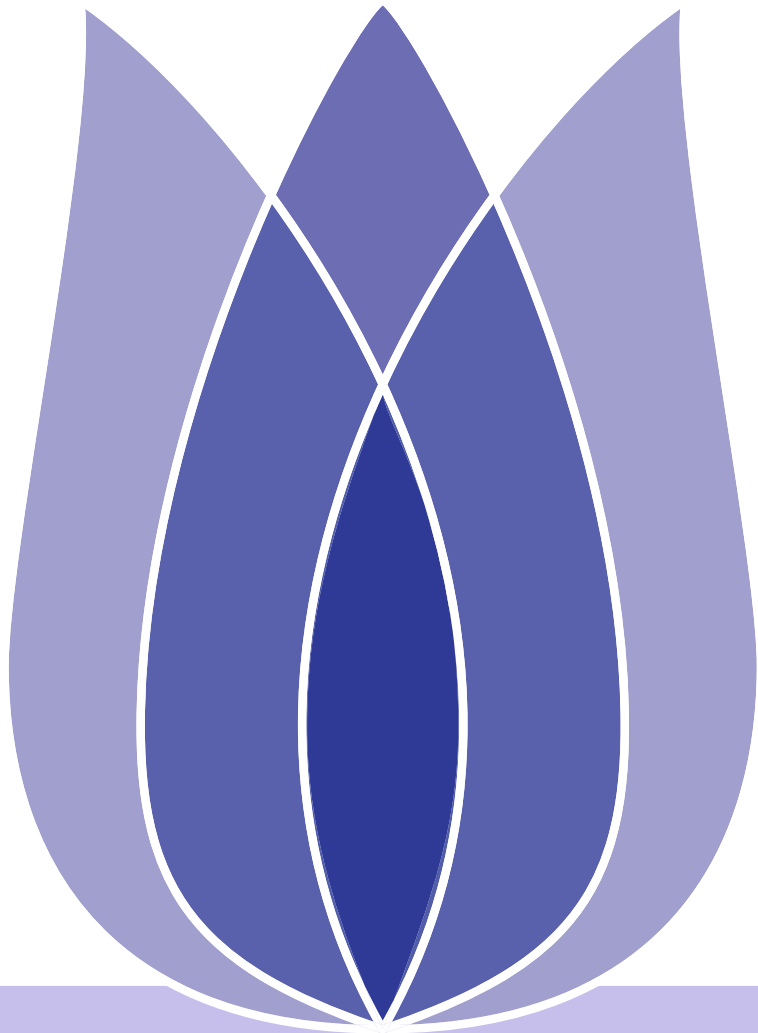


# Sales of Books Forecast

Lin Jiahong

Nanjing University of Science and Technology

(None)





# Overview

- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Conclusion](#)

## Problem Definition

Sales of Books Forecast

## Data Analysis

## Feature Extraction

Step One - Group Feature Extraction

Step Two - Outlying Degree Scoring

Step Three - Outlying Aspects Identification

## Model Train

Synthetic Dataset

NBA Dataset

## Conclusion



Problem Definition

Sales of Books Forecast

Data Analysis

Feature Extraction

Model Train

Conclusion

# Problem Definition



# Sales of Books Forecast

Problem Definition
Sales of Books Forecast
Data Analysis
Feature Extraction
Model Train
Conclusion

Defn

- Sales of Books Forecast aims to predict the sales of books in 2021 through the book sales data from 2017 to 2020.
- Data covers different countries and different stores.
  - There are cyclical and seasonal changes in book sales.

Data	row_num	date	country	store	product
train	70128	1461	6	2	4
test	17520	365	6	2	4



[Problem Definition](#)

[Data Analysis](#)

[Feature Extraction](#)

[Model Train](#)

[Conclusion](#)

# Data Analysis



# Overall data

- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Conclusion](#)

- Country - Belgium,France,Germany,Italy,Poland,Spain
- Product - [Kaggle Advanced Techniques],[Kaggle Getting Started],[Kaggle Recipe Book],[Kaggle for Kids: One Smart Goose]
- Stores - KaggleMart,KaggleRama
- Time line

Data	Earliest date	Latest date
<i>train</i>	2017 – 01 – 01	2020 – 12 – 31
<i>test</i>	2021 – 01 – 01	2021 – 12 – 31





# Monthly sales statistics

- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Conclusion

- the patterns in sales of all countries and stores are identical.the magnitudes of sales are different

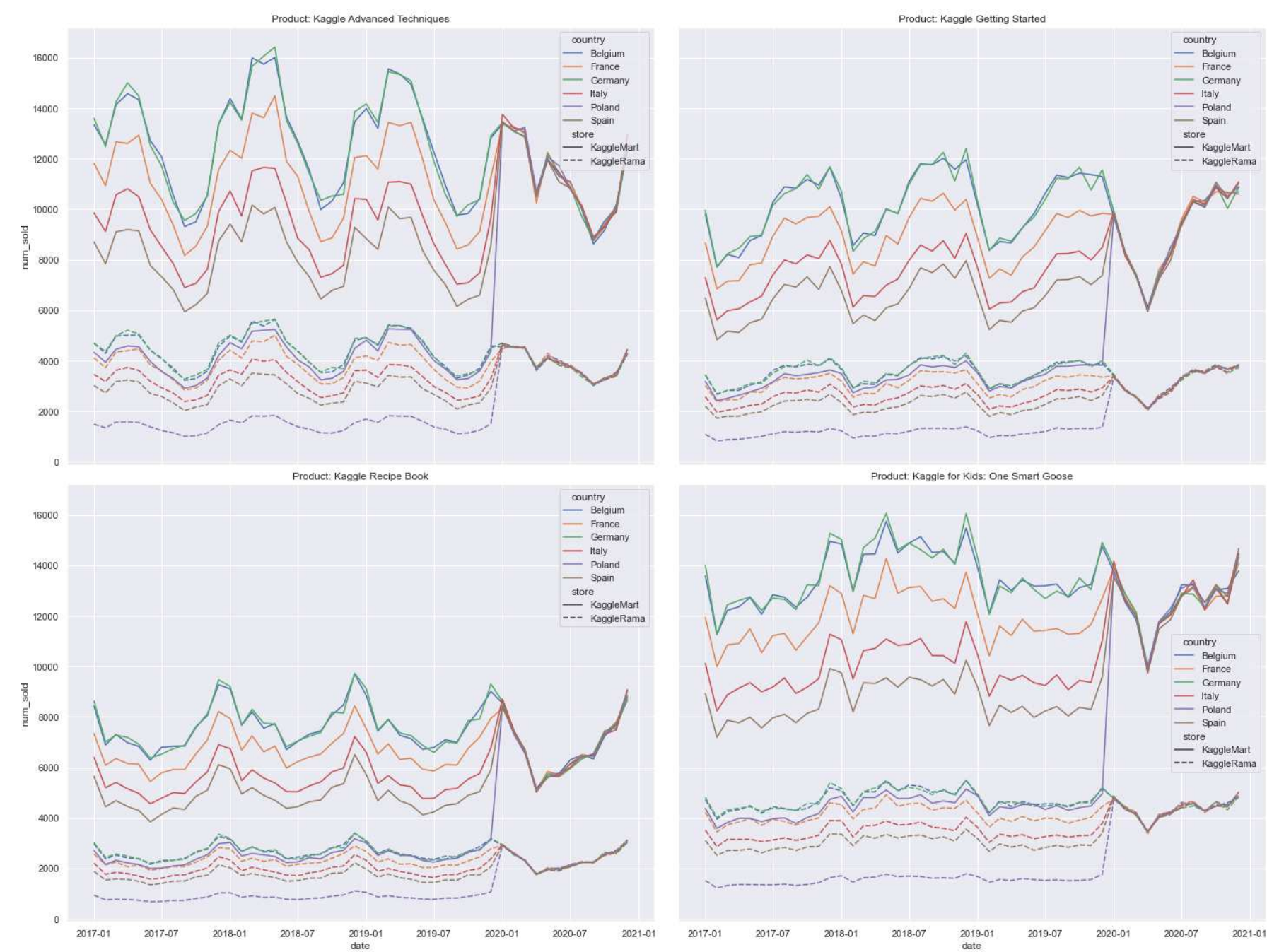


Figure 1: Monthly sales





# Aggregating Time Series(Store)

- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Conclusion

- Store-KaggleMart appears to consistantly have 74.25% of the total number of sales

Store	ratio
<i>KaggleMart</i>	0.742515
<i>KaggleRama</i>	0.257485

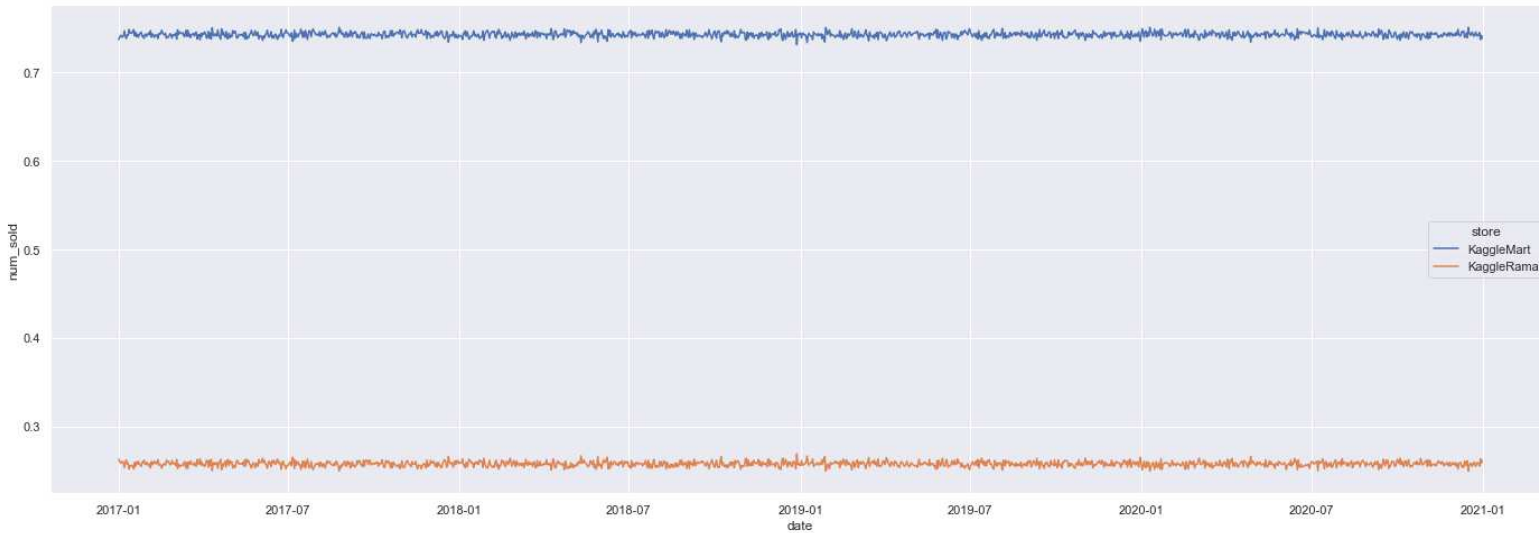


Figure 2: Stores ratio



# Aggregating Time Series(Store)

- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Conclusion

- To compare the trend of the two stores, multiply the sales data of the two stores by a constant.

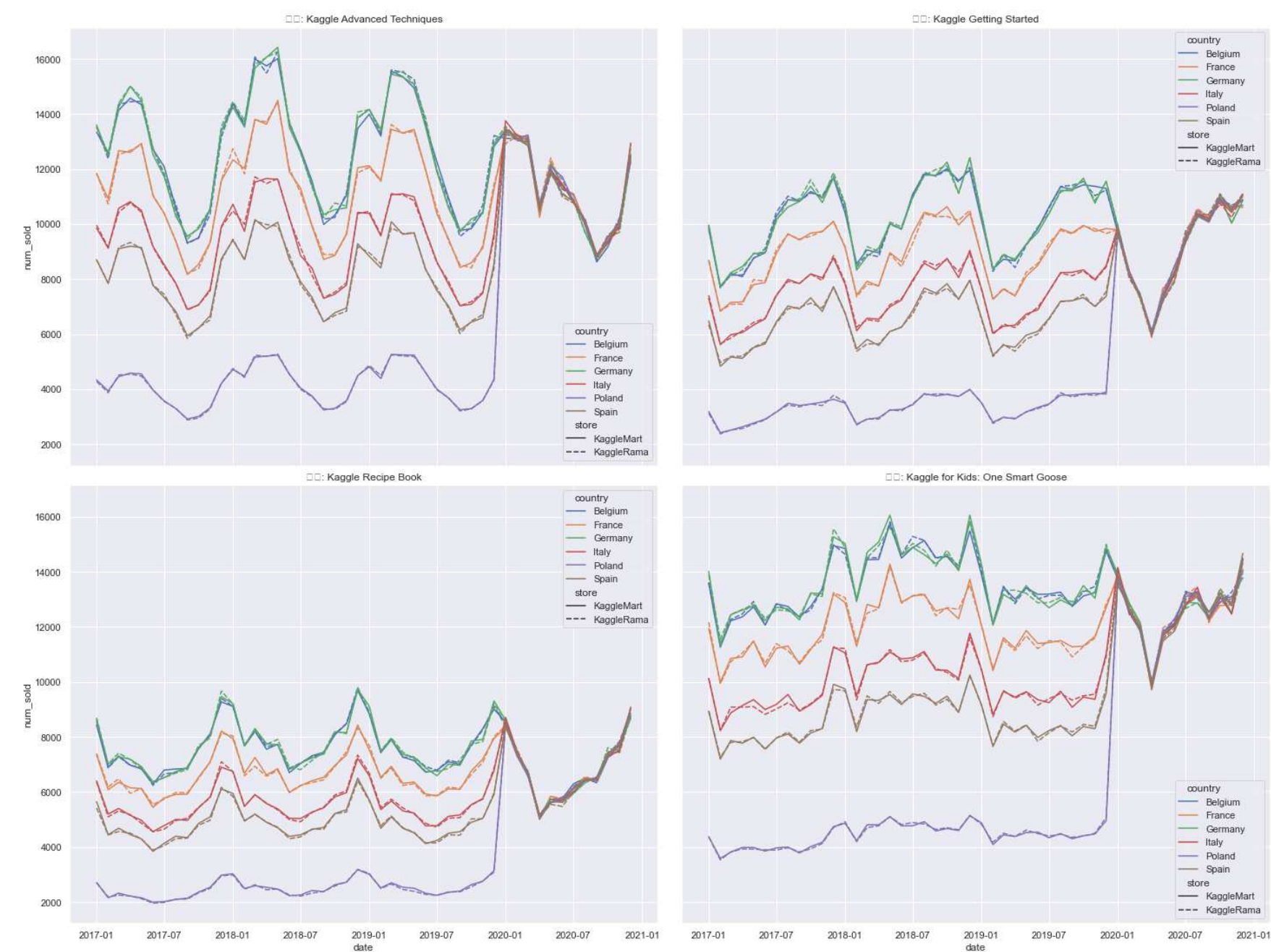


Figure 3: Stores ratio trend



# Aggregating Time Series(Country)

- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Conclusion

- Country-The ratio of total sales in different countries also fluctuates little.

Country	ratio
Belgium	0.218930
France	0.191360
Germany	0.219586
Italy	0.159383
Poland	0.071348
Spain	0.139393

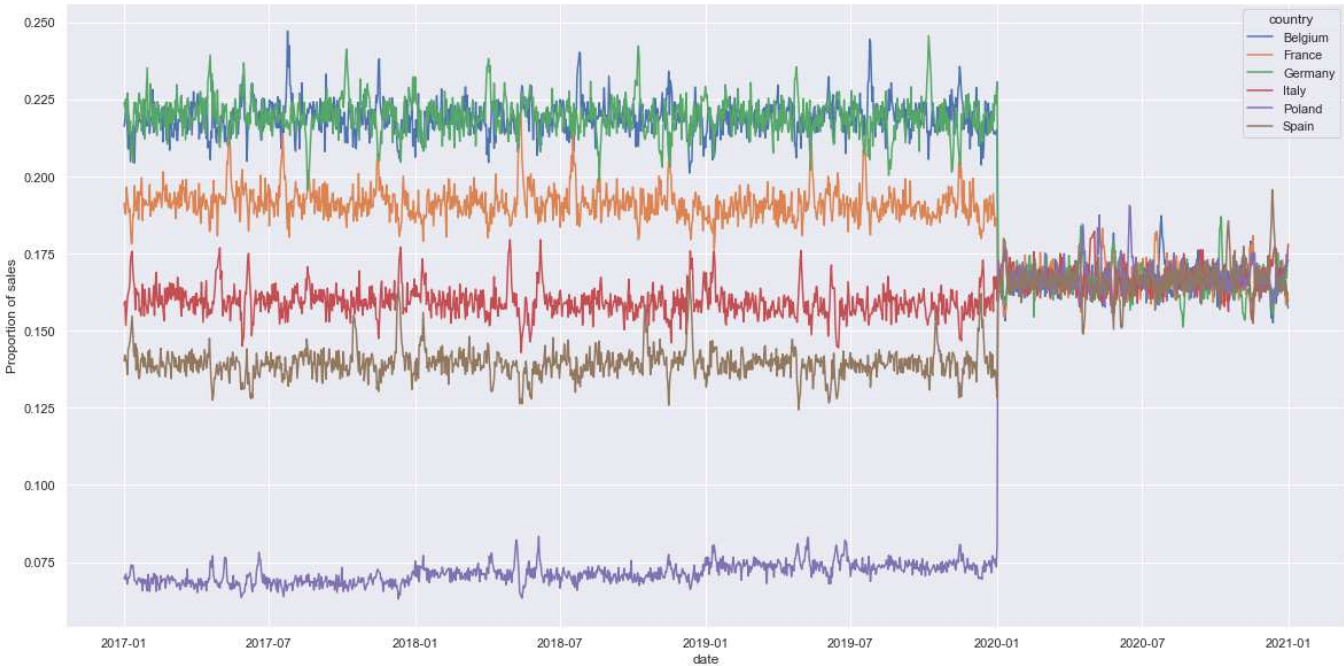


Figure 4: Countries ratio





# Aggregating Time Series(Country)

- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Conclusion

- Multiply all countries by a constant so they are comparable with Belgium.

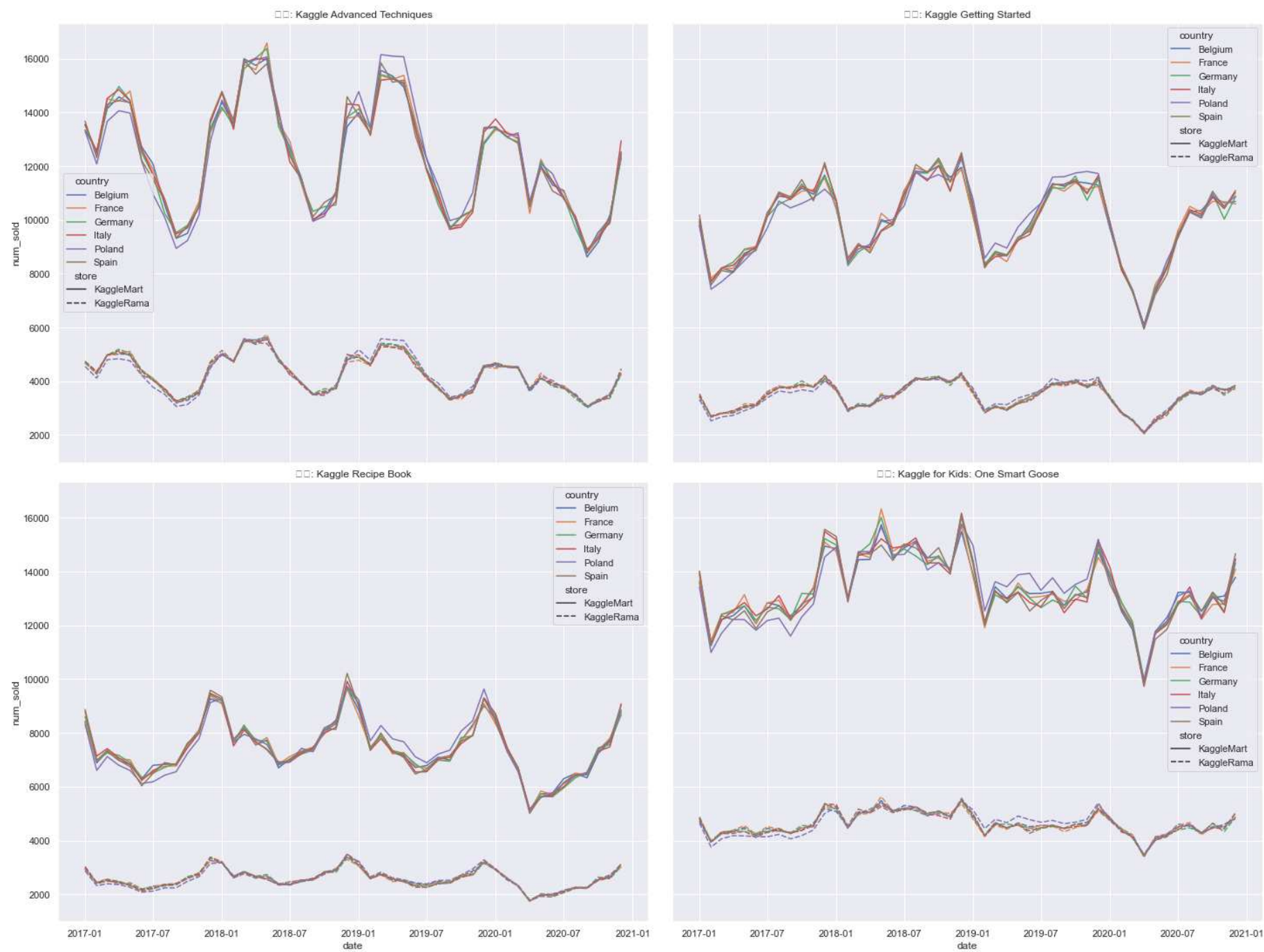


Figure 5: Countries ratio trend



# Aggregating Time Series(Country and Store)

- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Conclusion

- In the plots make all time series inline with the Belgium KaggleMart store by multiplying by a constant.

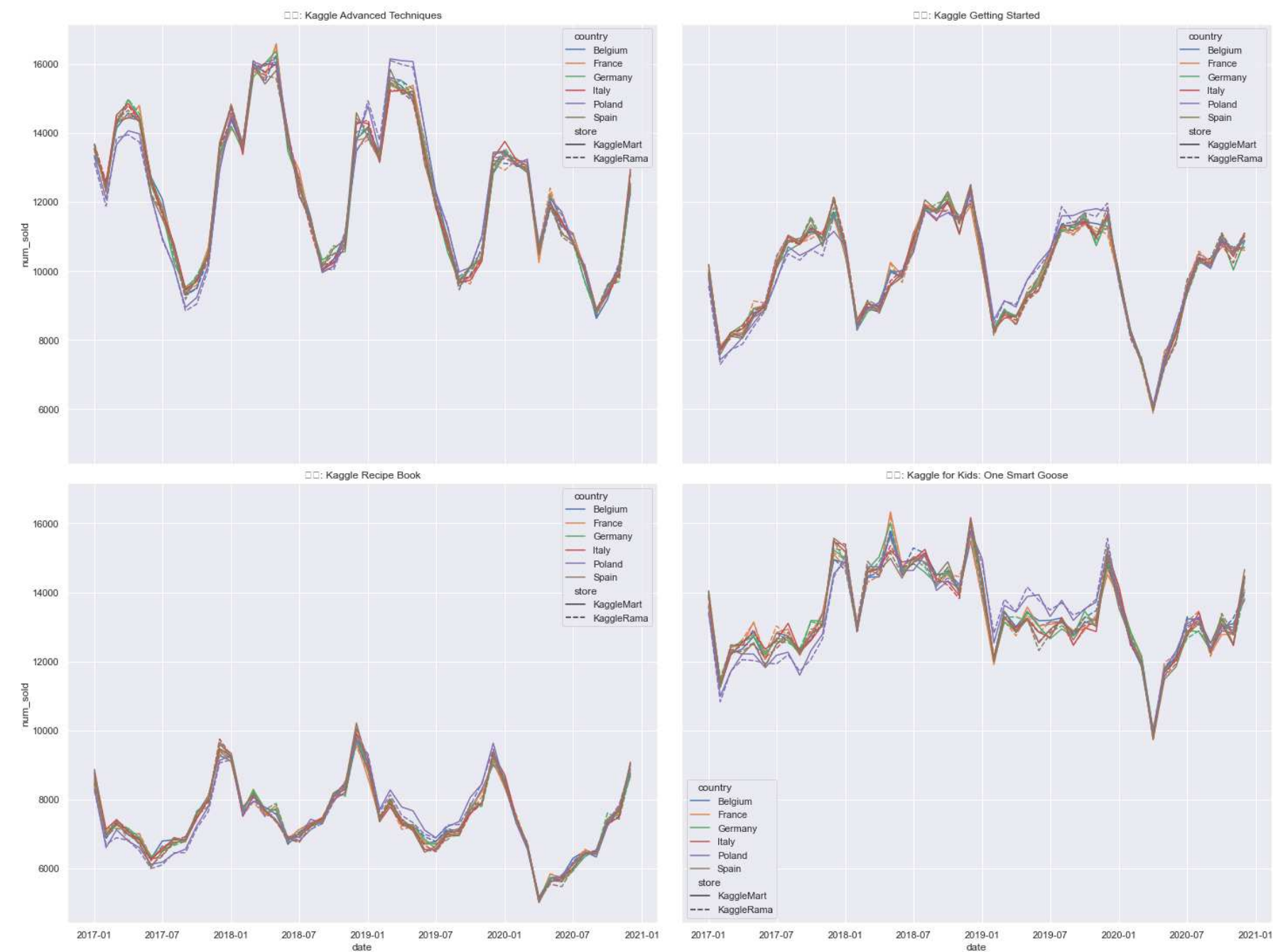


Figure 6: Countries and Store trend





# Aggregating Time Series(Product)

- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Conclusion

- The change trend of the sales volume of the four books is cyclical.

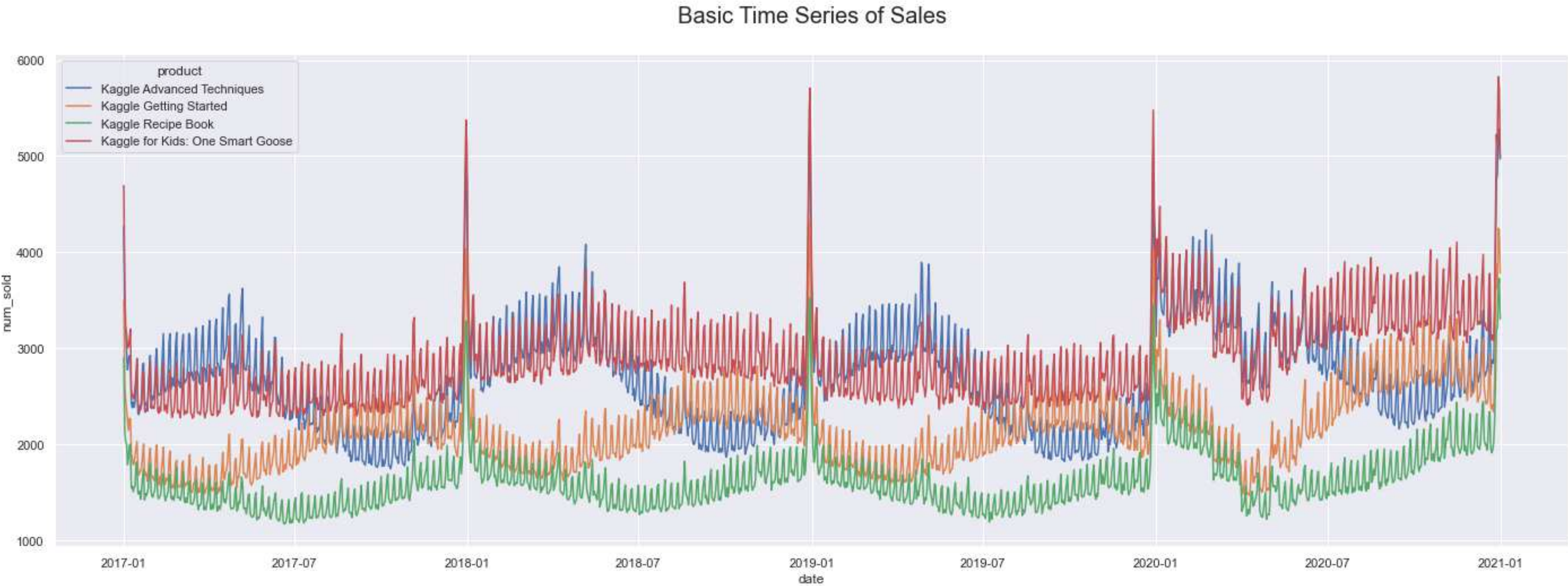


Figure 7: Sales of Product



# Aggregating Time Series(Product)

- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Conclusion

- The change trend of the sales proportion of the four books has rules.

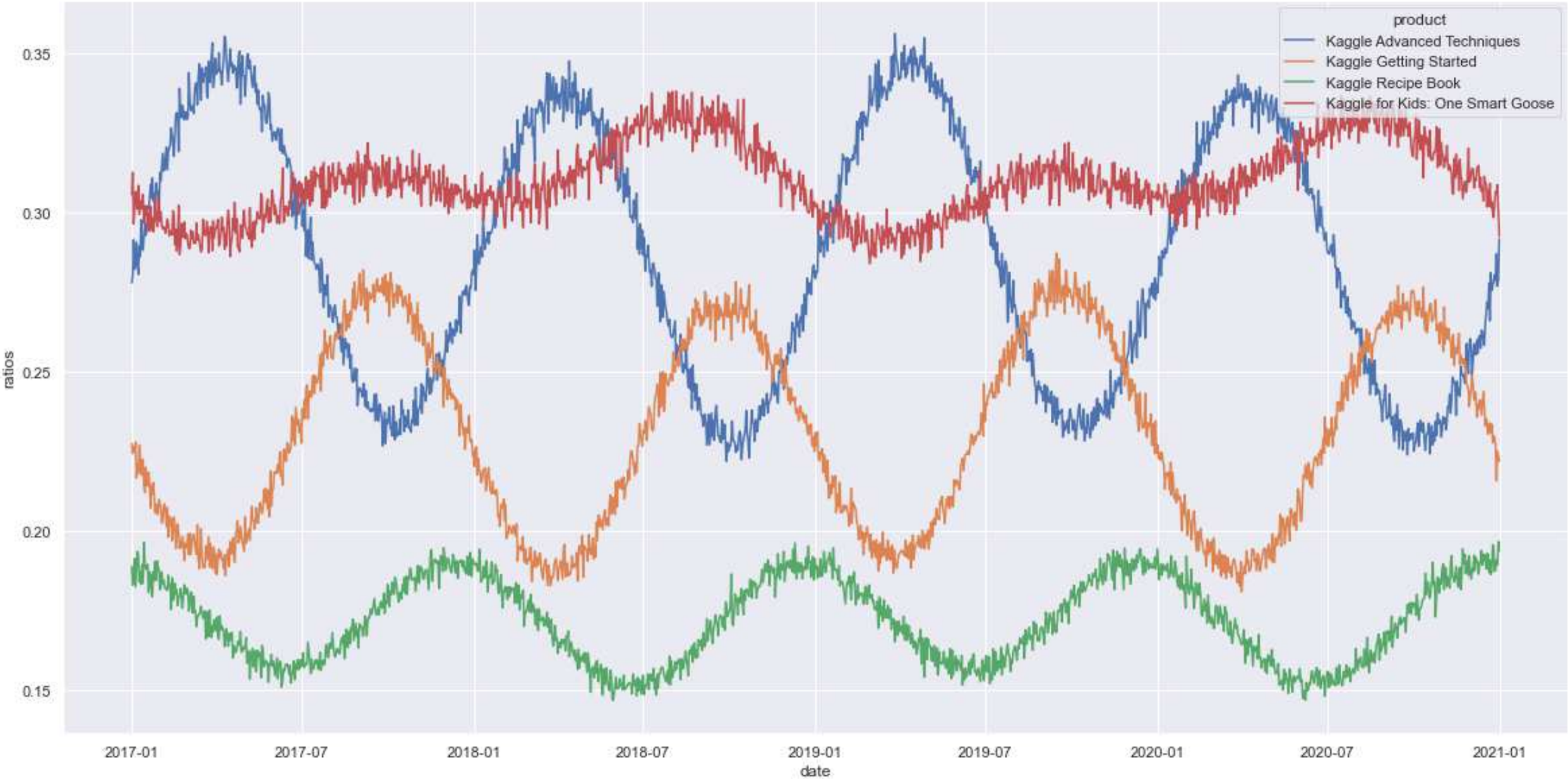


Figure 8: Product ratio trend





# Aggregated Time Series

- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Conclusion](#)

- aggregate the sales timeline to consider how to forecast the overall sales volume.

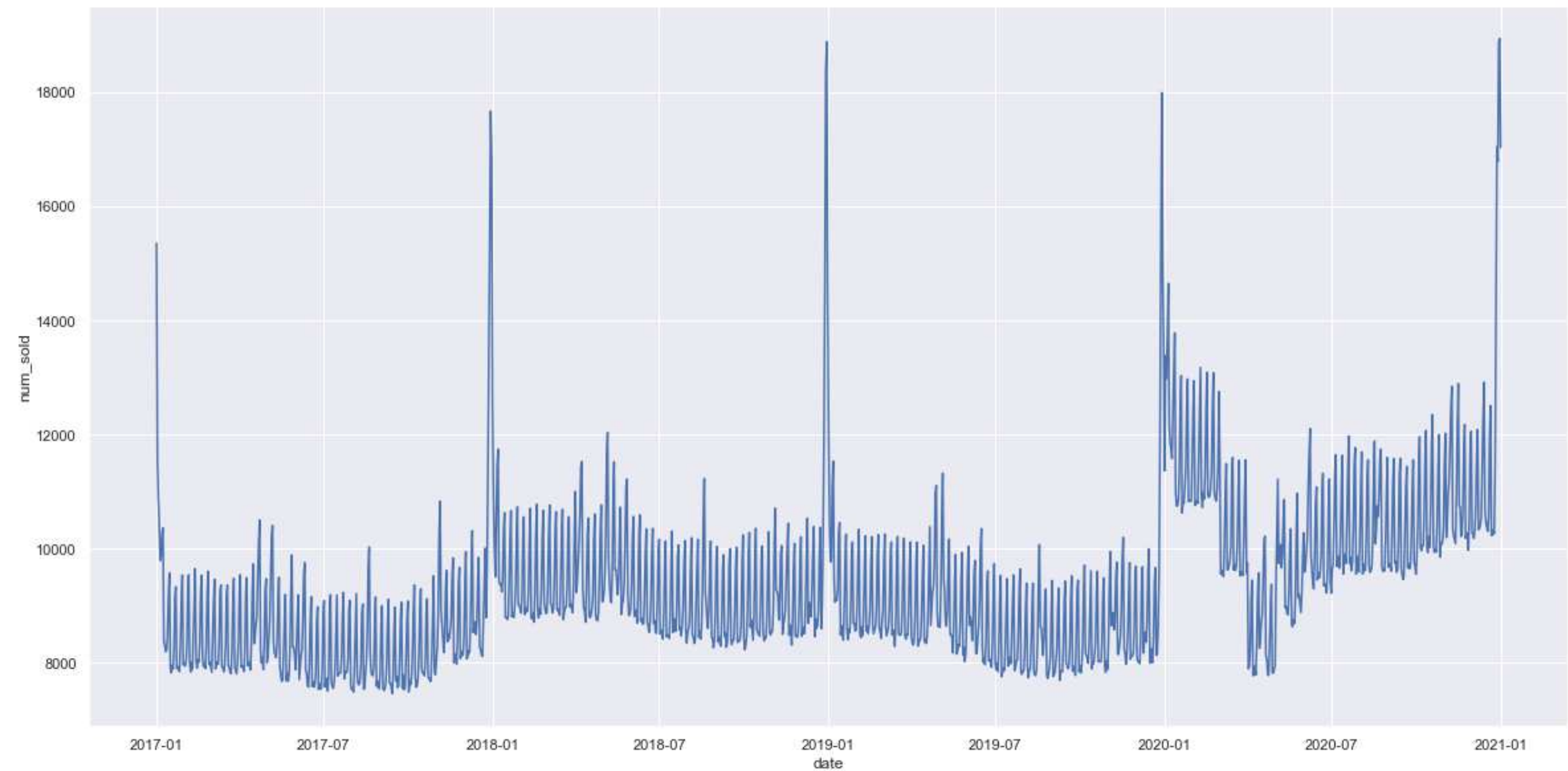


Figure 9: Aggregated time series



[Problem Definition](#)

[Data Analysis](#)

[Feature Extraction](#)

[Step One - Group Feature Extraction](#)

[Step Two - Outlying Degree Scoring](#)

[Step Three - Outlying Aspects](#)

[Identification](#)

[Model Train](#)

[Conclusion](#)

# Feature Extraction



- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
  - Step One - Group Feature Extraction
  - Step Two - Outlying Degree Scoring
  - Step Three - Outlying Aspects Identification
- [Model Train](#)
- [Conclusion](#)

# Framework of GOAM algorithm:

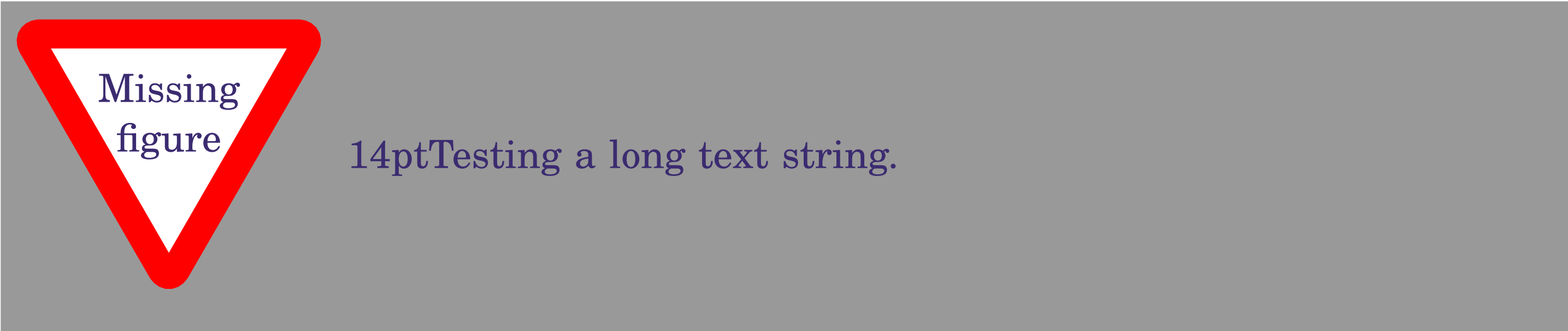


Figure 10: Framework of GOAM Algorithm



# Step One - Group Feature Extraction

- Problem Definition
- Data Analysis
- Feature Extraction
- Step One - Group Feature Extraction**
- Step Two - Outlying Degree Scoring
- Step Three - Outlying Aspects Identification
- Model Train
- Conclusion

■ Suppose  $f_1, f_2, f_3$  are three features of  $G_q$ .

$$f_1: \{x_1, x_2, x_3, x_4, x_5, x_2, x_3, x_4, x_1, x_2\}$$

$$f_2: \{y_2, y_2, y_1, y_2, y_3, y_3, y_5, y_4, y_4, y_2\}$$

$$f_3: \{z_1, z_4, z_2, z_4, z_5, z_3, z_1, z_2, z_4, z_2\}$$

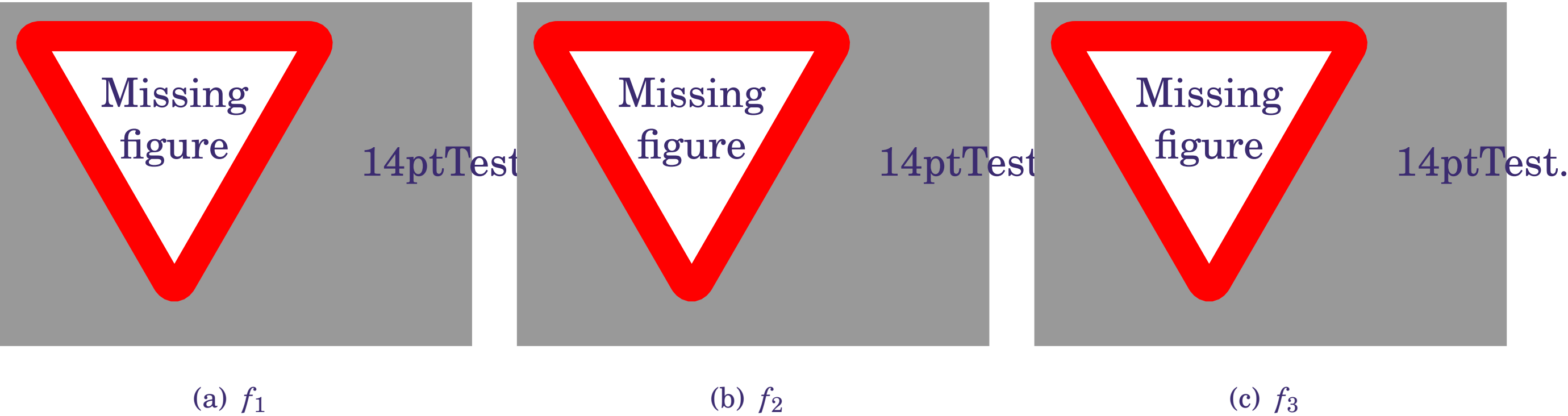


Figure 11: Histogram of  $G_q$  on three features



# Step Two - Outlying Degree Scoring

- Problem Definition
- Data Analysis
- Feature Extraction
  - Step One - Group Feature Extraction
  - Step Two - Outlying Degree Scoring
  - Step Three - Outlying Aspects Identification
- Model Train
- Conclusion

- Calculate Earth Mover Distance
  - ◆ Represent one feature among different groups
  - ◆ Purpose: calculate the minimum mean distance

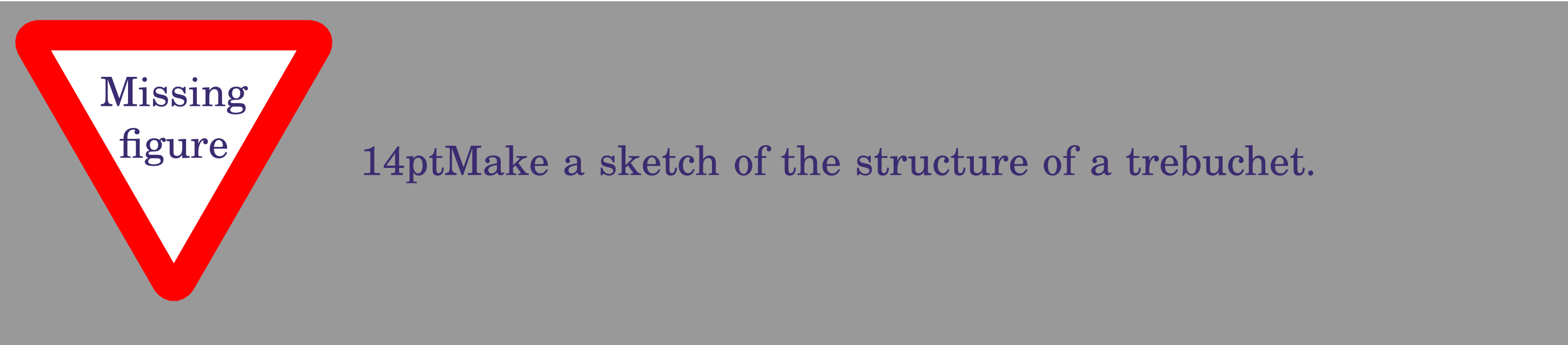


Figure 12: EMD of one feature



# Step Two - Outlying Degree Scoring

Problem Definition

Data Analysis

Feature Extraction

Step One - Group Feature Extraction

Step Two - Outlying Degree Scoring

Step Three - Outlying Aspects

Identification

Model Train

Conclusion

## ■ Calculate the outlying degree

$$OD(G_q) = \sum_1^n EDM(h_{q_s}, h_{k_s})$$

- ◆  $n \Leftrightarrow$  the number of contrast groups.
- ◆  $h_{k_s} \Leftrightarrow$  the histogram representation of  $G_k$  in the subspace s.





# Step Three - Outlying Aspects Identification

- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
  - [Step One - Group Feature Extraction](#)
  - [Step Two - Outlying Degree Scoring](#)
  - [Step Three - Outlying Aspects Identification](#)
- [Model Train](#)
- [Conclusion](#)

- Identify group outlying aspects mining based on the value of outlying degree.
- The greater the outlying degree is, the more likely it is group outlying aspect.





# Pseudo code

Problem Definition

Data Analysis

Feature Extraction

Step One - Group Feature Extraction

Step Two - Outlying Degree Scoring

Step Three - Outlying Aspects Identification

Model Train

Conclusion

## ■ Pseudo code of GOAM algorithm



14ptTesting a long text string



# Illustration

- Problem Definition
- Data Analysis
- Feature Extraction
  - Step One - Group Feature Extraction
  - Step Two - Outlying Degree Scoring
  - Step Three - Outlying Aspects Identification
- Model Train
- Conclusion

Table 1: Original Dataset

$G_1$	$F_1$	$F_2$	$F_3$	$F_4$	$G_2$	$F_1$	$F_2$	$F_3$	$F_4$
	10	8	9	8		7	7	6	6
	9	9	7	9		8	9	9	8
	8	10	8	8		6	7	8	9
	8	8	6	7		7	7	7	8
	9	9	9	8		8	6	6	7
$G_3$	$F_1$	$F_2$	$F_3$	$F_4$	$G_4$	$F_1$	$F_2$	$F_3$	$F_4$
	8	10	8	8		9	8	8	8
	9	9	7	9		7	7	7	9
	10	9	10	7		8	6	6	8
	9	10	8	6		9	8	8	7
	9	9	7	9		8	7	9	8



# Illustration

- Problem Definition
- Data Analysis
- Feature Extraction
  - Step One - Group Feature Extraction
  - Step Two - Outlying Degree Scoring
  - Step Three - Outlying Aspects Identification
- Model Train
- Conclusion

Table 2: outlying degree of each possible subspaces

Feature	Outlying Degree	Feature	Outlying Degree
$\{F_1\}$	4.351	$\{F_2, F_3\}$	4.023
$\{F_2\}$	2.012	$\{F_3, F_4\}$	4.324
$\{F_3\}$	1.392	$\{F_2, F_4\}$	2.018
$\{F_4\}$	2.207	$\{F_2, F_3, F_4\}$	2.012

■ Search process:

$OD(\{F_1\}) > \alpha$ , save to  $T_1$ .  
 $OD(\{F_2\}) < \alpha$ , save to  $C_1$ .  
 $OD(\{F_3\}) < \alpha$ , save to  $C_2$ .  
 $OD(\{F_4\}) < \alpha$ , save to  $C_3$ .

$OD(\{F_2, F_3\}) > \alpha$ , save to  $N_1$ .  
 $OD(\{F_3, F_4\}) > \alpha$ , save to  $N_2$ .  
 $OD(\{F_2, F_4\}) < \alpha$ , remove.  
 $OD(\{F_2, F_3, F_4\}) < \alpha$ , remove.



# Strengths of GOAM Algorithm

- Problem Definition
- Data Analysis
- Feature Extraction
  - Step One - Group Feature Extraction
  - Step Two - Outlying Degree Scoring
  - Step Three - Outlying Aspects Identification
- Model Train
- Conclusion

- Reduction of Complexity
  - ◆ Bottom-up search strategy.
  - ◆ Reduce the size of candidate subspaces.
- Efficiency
  - ◆ Before:  $O(2^d)$   
Now:  $O(d * n^2)$



[Problem Definition](#)

[Data Analysis](#)

[Feature Extraction](#)

**[Model Train](#)**

[Synthetic Dataset](#)

[NBA Dataset](#)

[Conclusion](#)

# Model Train



# Evaluation

- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Synthetic Dataset](#)
- [NBA Dataset](#)
- [Conclusion](#)

- $Accuracy = \frac{P}{T}$   
P: Identified outlying aspects  
T: Real outlying aspects



- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Synthetic Dataset
- NBA Dataset
- Conclusion

## ■ Synthetic Dataset and Ground Truth

Table 3: Synthetic Dataset and Ground Truth

Query group	<b>F<sub>1</sub></b>	<b>F<sub>2</sub></b>	<i>F<sub>3</sub></i>	<b>F<sub>4</sub></b>	<i>F<sub>5</sub></i>	<i>F<sub>6</sub></i>	<i>F<sub>7</sub></i>	<i>F<sub>8</sub></i>
<i>i<sub>1</sub></i>	<b>10</b>	<b>8</b>	9	<b>7</b>	7	6	6	8
<i>i<sub>2</sub></i>	<b>9</b>	<b>9</b>	7	<b>8</b>	9	9	8	9
<i>i<sub>3</sub></i>	<b>8</b>	<b>10</b>	8	<b>9</b>	6	8	7	8
<i>i<sub>4</sub></i>	<b>8</b>	<b>8</b>	6	<b>7</b>	8	8	6	7
<i>i<sub>5</sub></i>	<b>9</b>	<b>9</b>	9	<b>7</b>	7	7	8	8
<i>i<sub>6</sub></i>	<b>8</b>	<b>10</b>	8	<b>8</b>	6	6	8	7
<i>i<sub>7</sub></i>	<b>9</b>	<b>9</b>	7	<b>9</b>	8	8	8	7
<i>i<sub>8</sub></i>	<b>10</b>	<b>9</b>	10	<b>7</b>	7	7	7	7
<i>i<sub>9</sub></i>	<b>9</b>	<b>10</b>	8	<b>8</b>	7	6	7	7
<i>i<sub>10</sub></i>	<b>9</b>	<b>9</b>	7	<b>7</b>	7	8	8	8





# Synthetic Dataset Results

- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Synthetic Dataset](#)
- [NBA Dataset](#)
- [Conclusion](#)

Table 4: The experiment result on synthetic dataset

Method	Truth Outlying Aspects	Identified Aspects	Accuracy
GOAM	$\{F_1\}, \{F_2F_4\}$	$\{F_1\}, \{F_2F_4\}$	100%
Arithmetic Mean based OAM	$\{F_1\}, \{F_2F_4\}$	$\{F_4\}, \{F_2\}$	0%
Median based OAM	$\{F_1\}, \{F_2F_4\}$	$\{F_2\}, \{F_4\}$	0%



- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Synthetic Dataset](#)
- [NBA Dataset](#)
- [Conclusion](#)

## Data Collection

### Source

*Yahoo Sports* website (<http://sports.yahoo.com.cn/nba>)

### Data

- Extract NBA teams’ data until March 30, 2018;
- 6 divisions;
- 12 features (eg: *Point Scored*).



- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Synthetic Dataset
- NBA Dataset
- Conclusion

The detail features are as follows:

Table 5: Collected data of Brooklyn Nets Team

Pts	FGA	FG%	3FA	3PT%	FTA	FT%	Reb	Ass	To	Stl	Blk
18	12	42	2.00	50	7.00	100	0	4	3	0	0
15.7	14.07	41	5.45	32	3.05	75	3.98	5.1	2.98	0.69	0.36
14.5	11.1	47	0.82	26	4.87	78	6.82	2.4	1.74	0.92	0.66
13.5	10.8	42	5.37	37	3.38	77	6.66	2	1.38	0.83	0.42
12.7	10.59	39	5.36	33	3.37	82	3.24	6.6	1.56	0.89	0.31
12.6	10.93	40	6.94	37	1.70	84	4.27	1.5	1.06	0.61	0.44
12.2	10.39	44	3.42	35	2.70	72	3.79	4.1	2.15	1.12	0.32
10.6	7.85	49	4.51	41	1.35	83	3.34	1.6	1.15	0.45	0.24



- Problem Definition
- Data Analysis
- Feature Extraction
- Model Train
- Synthetic Dataset
- NBA Dataset**
- Conclusion

## ■ Data Preprocess

Table 6: The bins that used to discrete data of each feature

Labels	Pts	FGA	FG%	3FA	3PT%	FTA
low	[0,5]	[0,4]	[0,0.35]	[0,1.0]	[0,0.2]	[0,1.0]
medium	(5,10]	(4,7]	(0.35,0.45]	(1.0,2.5]	(0.2,0.3]	(1.0,1.5]
high	(10,15]	(7,10]	(0.45,0.5]	(2.5,3.5]	(0.3,0.35]	(1.5,2.5]
very high	(15,+∞]	(10,+∞]	(0.5,1]	(3.5,+∞]	(0.35,1]	(2.5,+∞]
Labels	FT%	Reb	Ass	To	Stl	Blk
low	[0,0.6]	[0,2.0]	[0,1.0]	[0,0.6]	[0,0.2]	[0,0.25]
medium	(0.6,0.65]	(2,5]	(1,2]	(0.6,0.9]	(0.2,0.5]	(0.25,0.5]
high	(0.65,0.75]	(5,6]	(2,4]	(0.9,1.7]	(0.6,0.75]	(0.5,0.7]
very high	(0.75,1]	(6,+∞]	(4,+∞]	(1.7,+∞]	(0.75,+∞]	(0.7,+∞]



- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Synthetic Dataset](#)
- [NBA Dataset](#)
- [Conclusion](#)

Table 7: The identified outlying aspects of groups

Teams	Trivial Outlying Aspects	NonTrivial Outlying Aspects
Cleveland Cavaliers	{3FA}	{FGA, FT%}, {FGA, FG%}
Orlando Magic	{Stl}	None
Milwaukee Bucks	{To}, {FTA}	{FGA, FTA}, {3FA, FTA}
Golden State Warriors	{FG%}	{FT%, Blk}, {FGA, 3PT%, FTA}
Utah Jazz	{Blk}	{3FA, 3PT%}
New Orleans Pelicans	{FT%}, {FTA}	{FTA, Stl}, {FTA, To}



- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Conclusion](#)**

# Conclusion





# Conclusion

Problem Definition

Data Analysis

Feature Extraction

Model Train

Conclusion

- Formalize the problem of *Group Outlying Aspects Mining* by extending outlying aspects mining;
- Propose a novel method **GOAM algorithm** to solve the *Group Outlying Aspects Mining* problem;
- Utilize the pruning strategies to reduce time complexity.



**TULIP**

*Team for Universal Learning and Intelligent Processing*





# Questions?

- [Problem Definition](#)
- [Data Analysis](#)
- [Feature Extraction](#)
- [Model Train](#)
- [Conclusion](#)



# Contact Information

Associate Professor Gang Li  
School of Information Technology  
Deakin University, Australia



GANGLI@TULIP.ORG.AU



TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING

