Store Sales Prediction

Xuechen Zhang

Nanjing University of Science and Technology
Deakin University
Chinese Academy of Sciences

2023-01-30

Overview

Problem Description

- Store Sales Prediction

Data Clean

- Data Describe
- Daily sales of 2015-2018
- Daily sales of 2015-2018
- Monthly sales of 2015-2018
- Monthly sales of 2015-2018
- Monthly sales of 2015-2018
- Sales per day of the week

Method analysis

Data processing

Model Training

Related Work and Challenges

- Related Work - Outlying Aspects Mining
- Challenges (1)

GOAM Algorithm

Problem Description

Store Sales Prediction

Describe

In this challenge, the title gives the daily sales volume of three products of two stores in three different countries from 2015 to 2018, and asks us to predict their sales volume in the next year.

Seasons and weekends will affect sales.

The country’s GDP will also have an impact on sales.

Data Clean

Data Describe

Problem Description

Data Clean

Data Describe

Daily sales of 2015-2018

Daily sales of 2015-2018

Monthly sales of 2015-2018

Monthly sales of 2015-2018

Monthly sales of 2015-2018

Sales per day of the week

Method analysis

Data processing

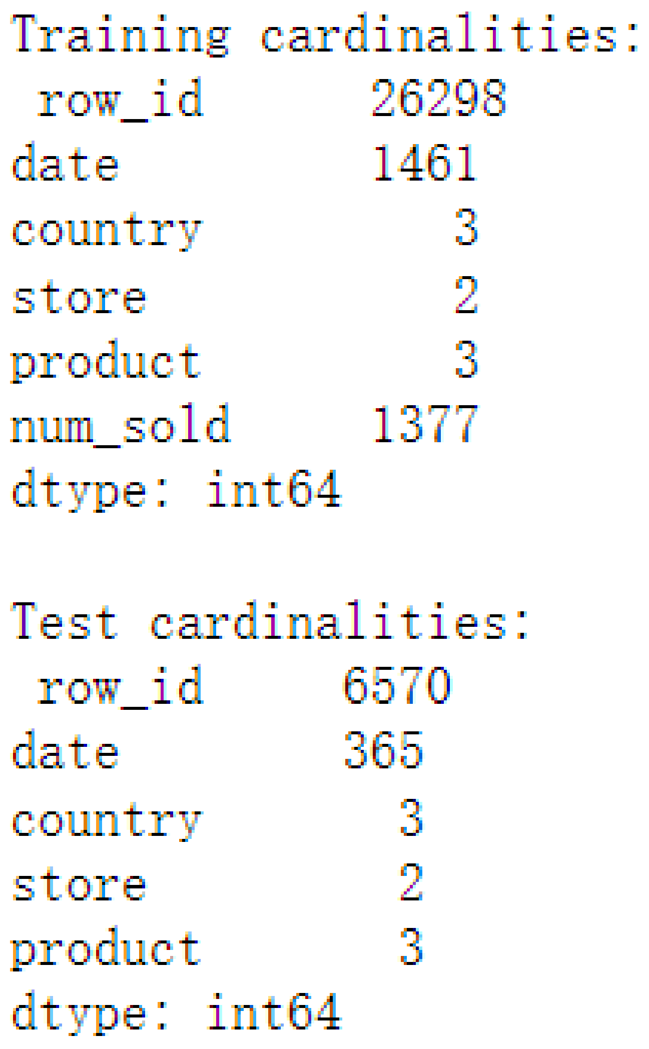
Model Training

Related Work and Challenges

GOAM Algorithm

Evaluation Results

Conclusion

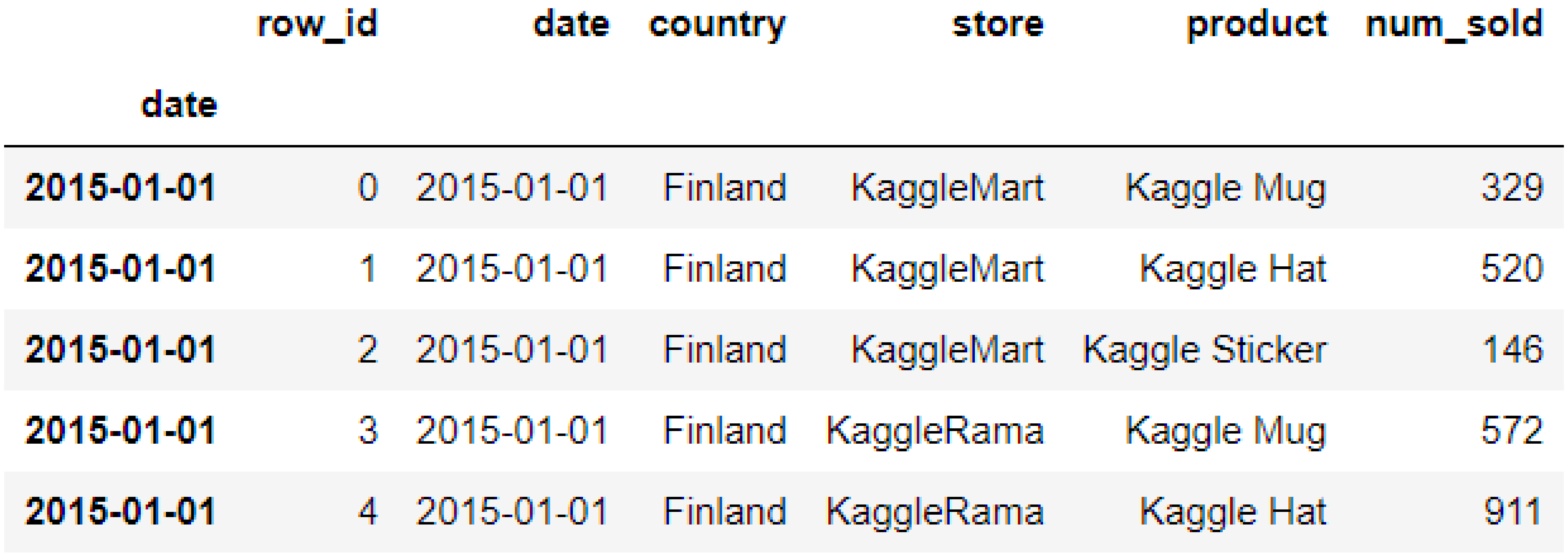


```
Training cardinalities:
 row_id      26298
date         1461
country         3
store           2
product         3
num_sold     1377
dtype: int64

Test cardinalities:
 row_id     6570
date        365
country       3
store         2
product       3
dtype: int64
```

Figure 1: Describe

Data Describe



	row_id	date	country	store	product	num_sold
date						
2015-01-01	0	2015-01-01	Finland	KaggleMart	Kaggle Mug	329
2015-01-01	1	2015-01-01	Finland	KaggleMart	Kaggle Hat	520
2015-01-01	2	2015-01-01	Finland	KaggleMart	Kaggle Sticker	146
2015-01-01	3	2015-01-01	Finland	KaggleRama	Kaggle Mug	572
2015-01-01	4	2015-01-01	Finland	KaggleRama	Kaggle Hat	911

Figure 2: Example

Data Describe

Problem Description
Data Clean
Data Describe
Daily sales of 2015-2018
Daily sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Sales per day of the week
Method analysis
Data processing
Model Training
Related Work and Challenges
GOAM Algorithm
Evaluation Results
Conclusion

Defn

To sum up, we can find that there are three countries, two stores and three products, so there will be 18 combinations. The training data covers 2015-2018, and the test data requires us to predict 2019. There is no missing value in training data and test data. Next, we will analyze the data by viewing the chart.

Daily sales of 2015-2018

Problem Description
Data Clean
Data Describe
Daily sales of 2015-2018
Daily sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Sales per day of the week
Method analysis
Data processing
Model Training
Related Work and Challenges
GOAM Algorithm
Evaluation Results
Conclusion

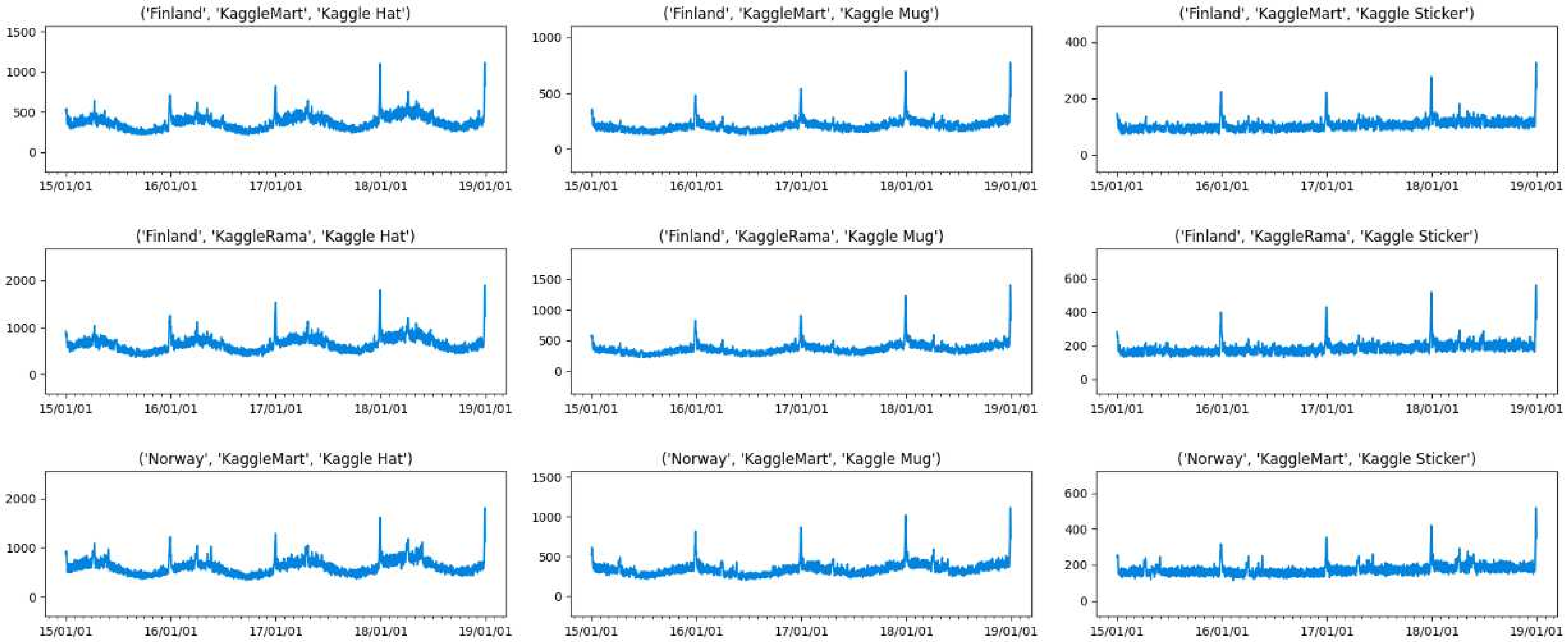


Figure 3: Daily sales of 2015-2018(1)

Daily sales of 2015-2018

Problem Description

Data Clean

Data Describe

Daily sales of 2015-2018

Daily sales of 2015-2018

Monthly sales of 2015-2018

Monthly sales of 2015-2018

Monthly sales of 2015-2018

Sales per day of the week

Method analysis

Data processing

Model Training

Related Work and Challenges

GOAM Algorithm

Evaluation Results

Conclusion

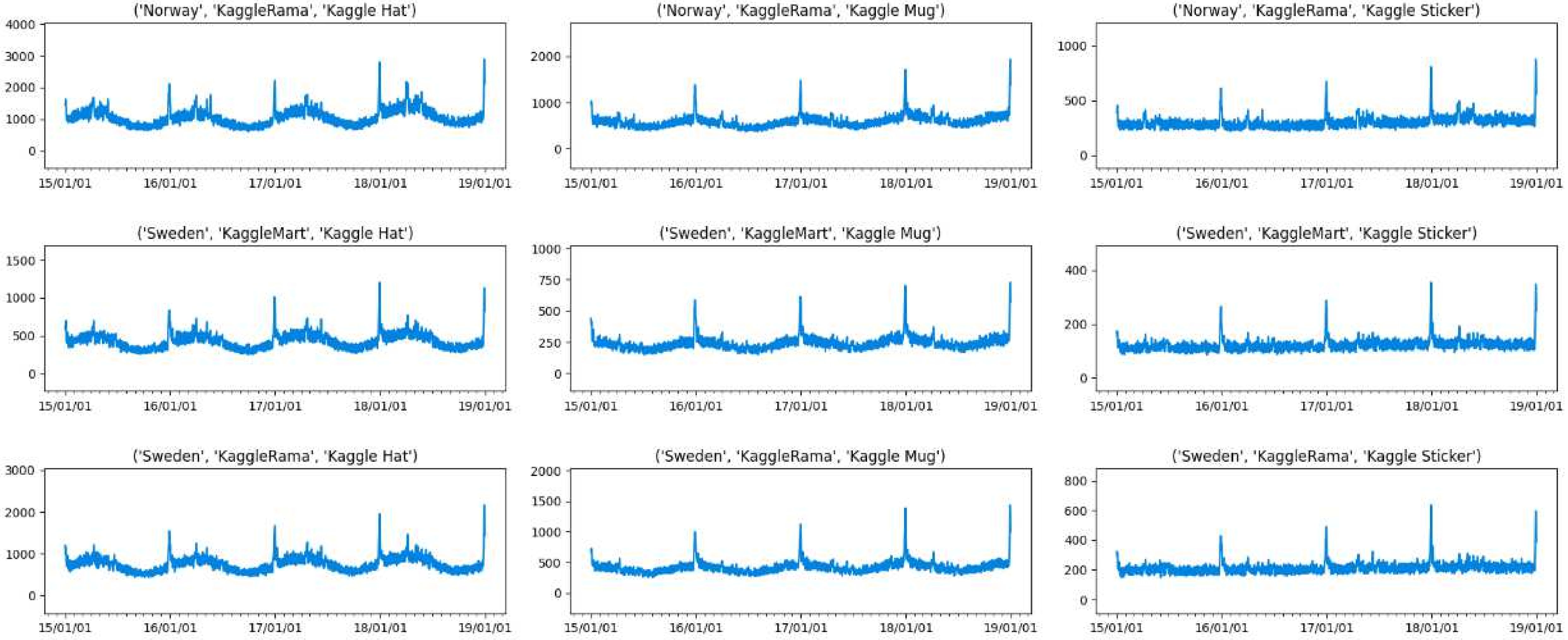


Figure 4: Daily sales of 2015-2018(2)

Daily sales of 2015-2018

Defn

From the above chart, we can see that the sales volume of each product at the end of each year is much higher than the average, and the sales volume of Kaggle Hat and Kaggle Mug seems to have seasonal characteristics, while the sales volume of Kaggle Sticker does not see obvious seasonal changes, so we should consider adding Fourier characteristics for Kaggle Hat and Kaggle Mug.

Monthly sales of 2015-2018

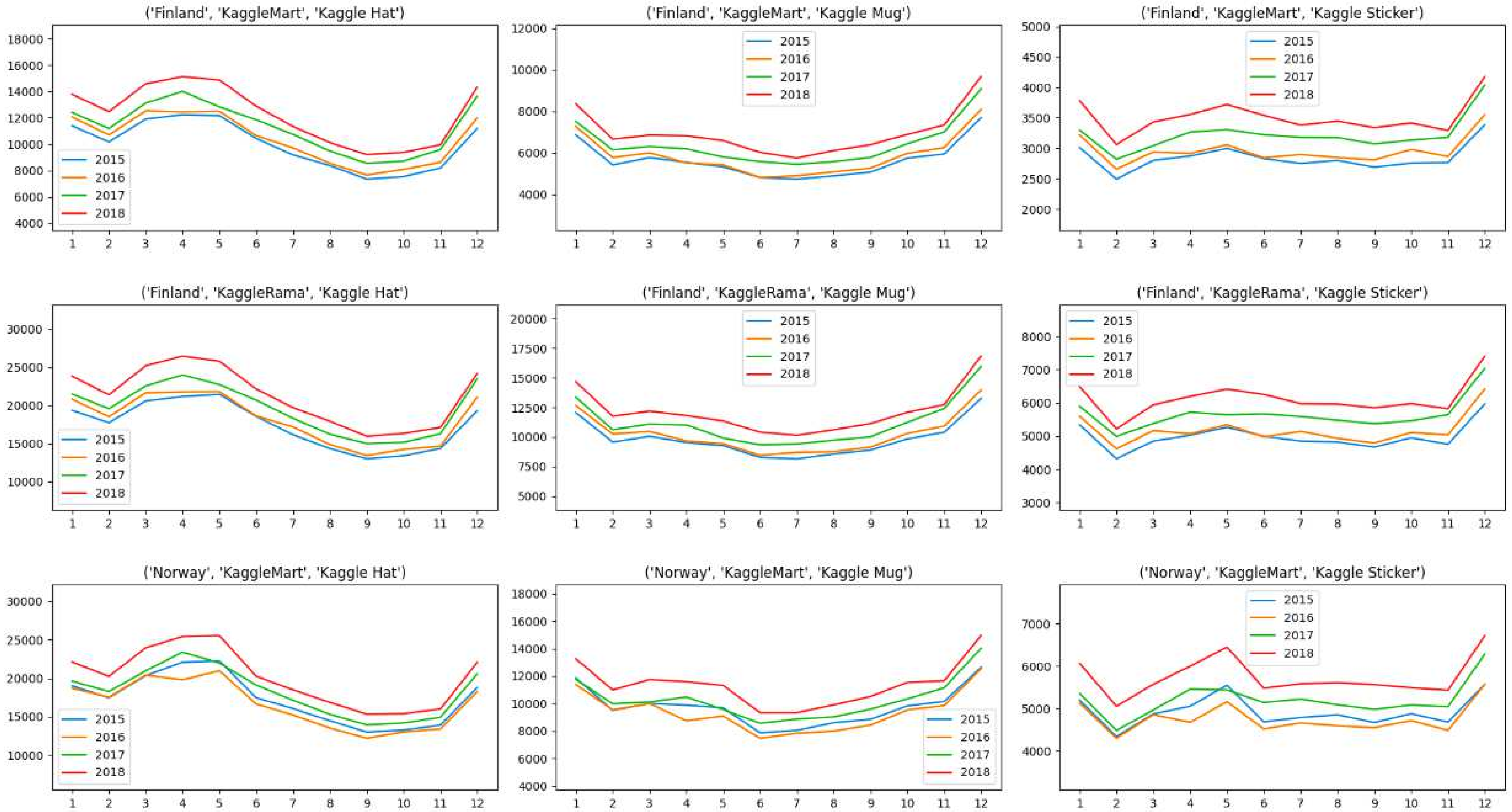


Figure 5: Monthly sales of 2015-2018(1)

Monthly sales of 2015-2018

Problem Description
Data Clean
Data Describe
Daily sales of 2015-2018
Daily sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Sales per day of the week
Method analysis
Data processing
Model Training
Related Work and Challenges
GOAM Algorithm
Evaluation Results
Conclusion

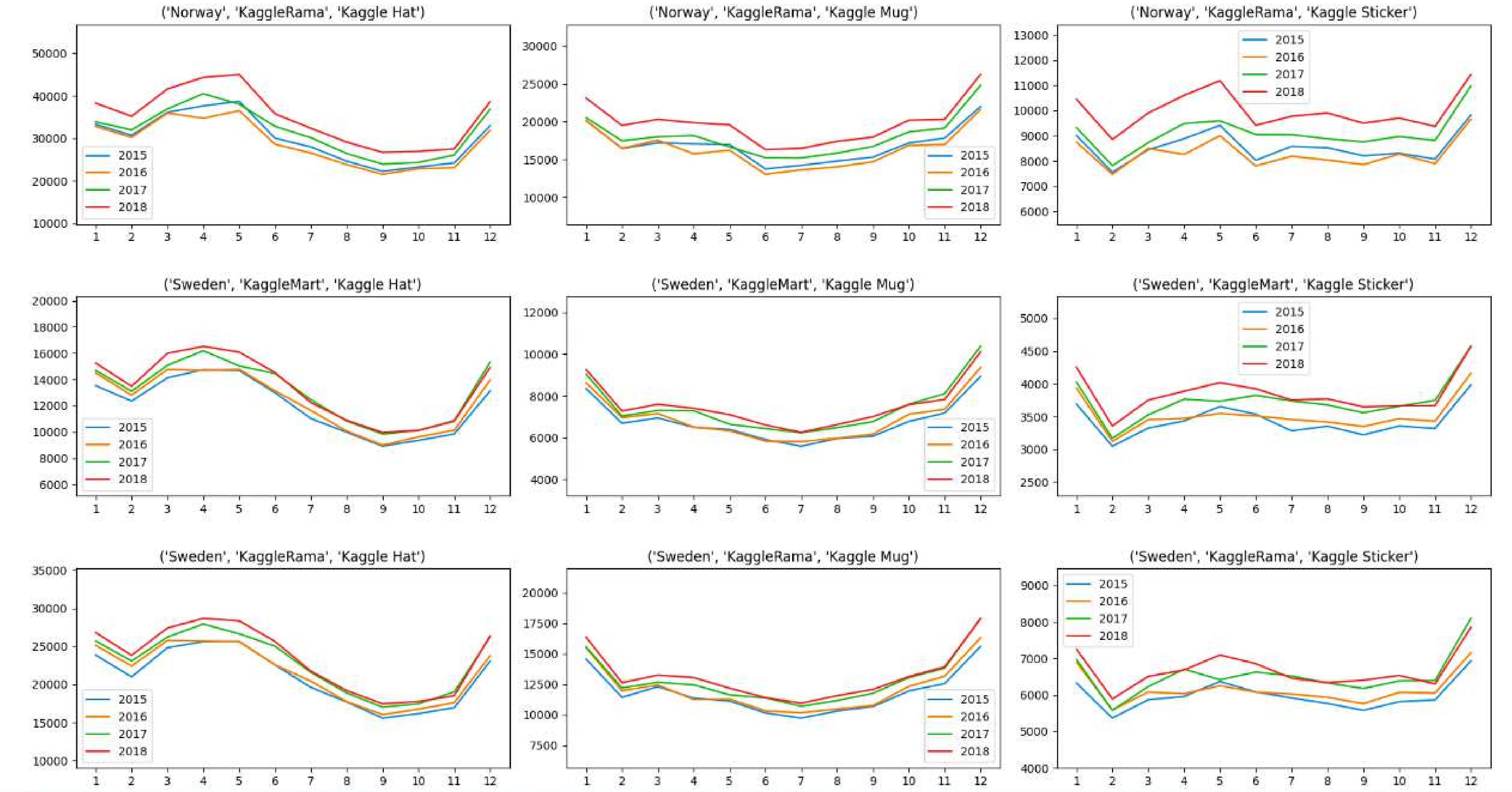


Figure 6: Monthly sales of 2015-2018(2)

Monthly sales of 2015-2018

Problem Description
Data Clean
Data Describe
Daily sales of 2015-2018
Daily sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Sales per day of the week
Method analysis
Data processing
Model Training
Related Work and Challenges
GOAM Algorithm
Evaluation Results
Conclusion

- Analysis Chart
  - Observe the phenomena in the chart and draw corresponding conclusions

Phenomena

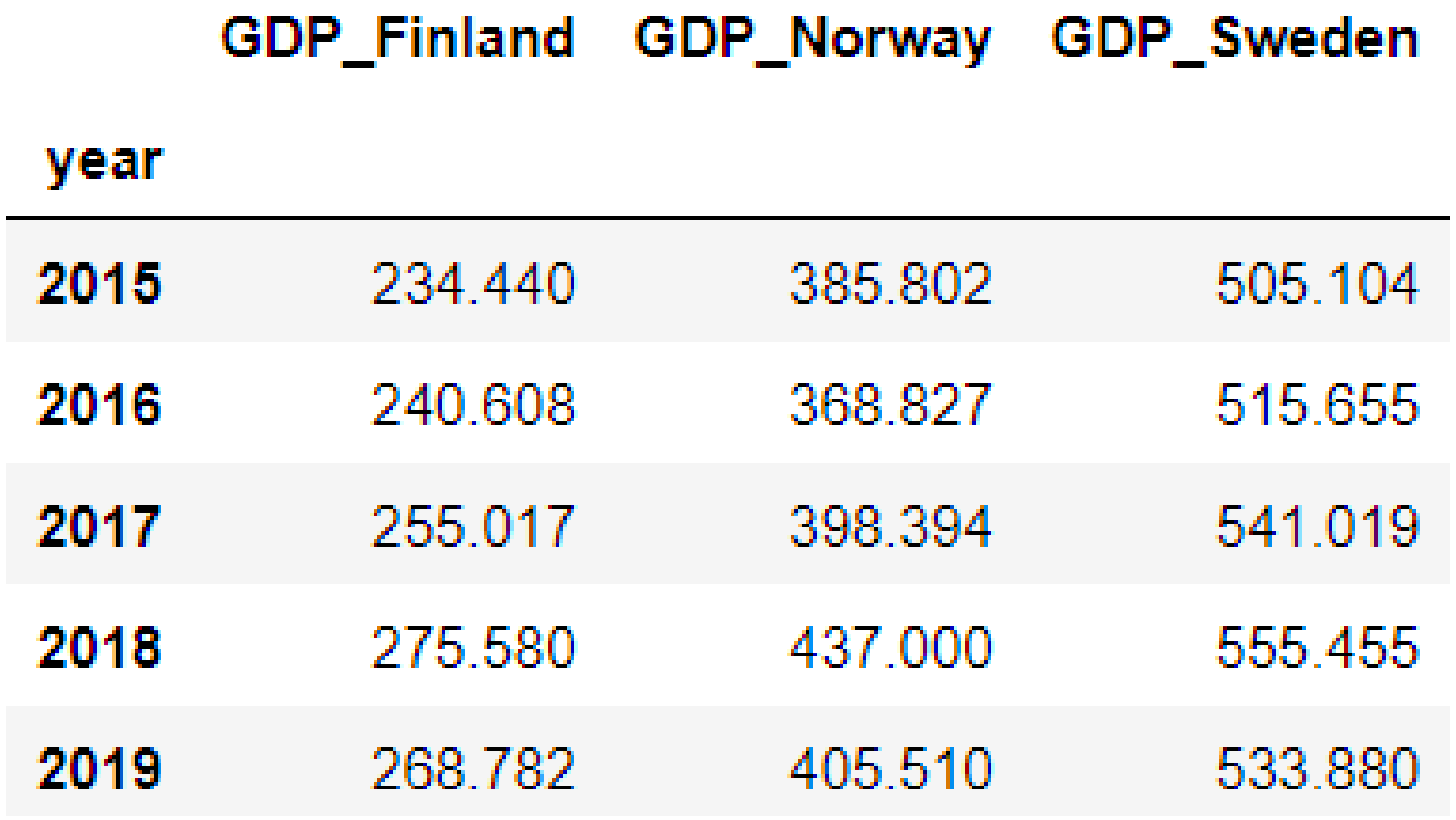
- The monthly fluctuations in different years of the same portfolio are similar.
- The sales volume in most portfolios is increasing year by year.
- Norway’s sales are not increasing year by year.

Conclusions

- The sales volume of each month is seasonal.
- The annual sales volume is related to other factors (the guess is GDP)

Monthly sales of 2015-2018

- GDP data
  - We have found the GDP of the three countries in 2015-2018.
  - Norway's GDP in 2015 is higher than that in 2016.



year	GDP_Finland	GDP_Norway	GDP_Sweden
2015	234.440	385.802	505.104
2016	240.608	368.827	515.655
2017	255.017	398.394	541.019
2018	275.580	437.000	555.455
2019	268.782	405.510	533.880

Figure 7: GDP of 2015-2018

Sales per day of the week

Problem Description
Data Clean
Data Describe
Daily sales of 2015-2018
Daily sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Monthly sales of 2015-2018
Sales per day of the week
Method analysis
Data processing
Model Training
Related Work and Challenges
GOAM Algorithm
Evaluation Results
Conclusion

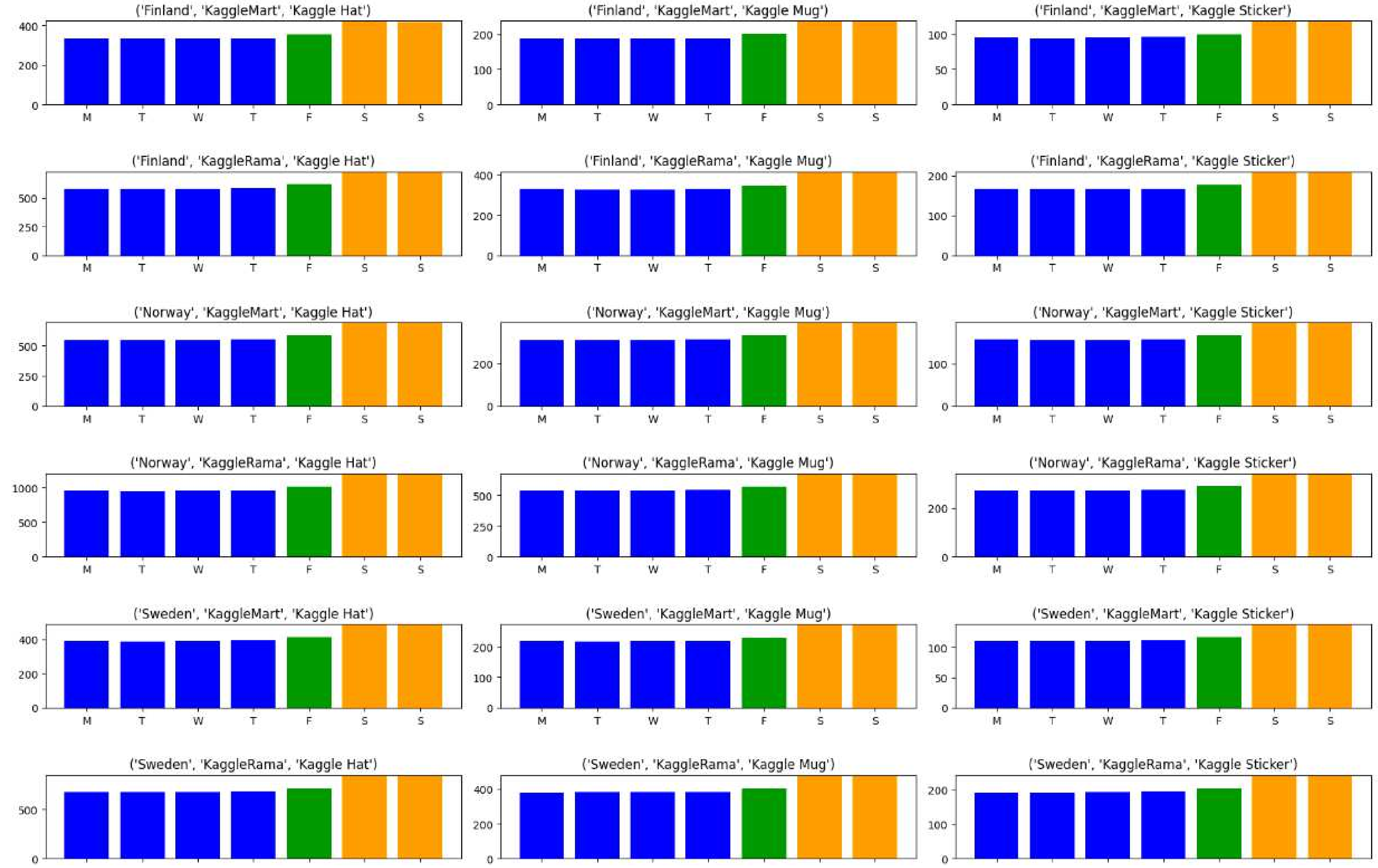


Figure 8: Sales per day of the week

Sales per day of the week

Defn

From the Sales per day of the week, we can see that the sales volume on the weekend is higher than that on the weekday, which means that the week also has seasonal characteristics. For such a short period of time, we should consider adding seasonal indicators.

Problem Description

Data Clean

Method analysis

Data processing

Model Training

Related Work and Challenges

GOAM Algorithm

Evaluation Results

Conclusion

Method analysis

Method analysis

Defn

Because the structure of the data set is not very complex and there are few influencing factors, the linear regression model is selected for this model, and the method of combining time series with linear regression model is used. Add some elements of time series, such as Fourier characteristics, seasonal indicators, and real world GDP data.

Data processing

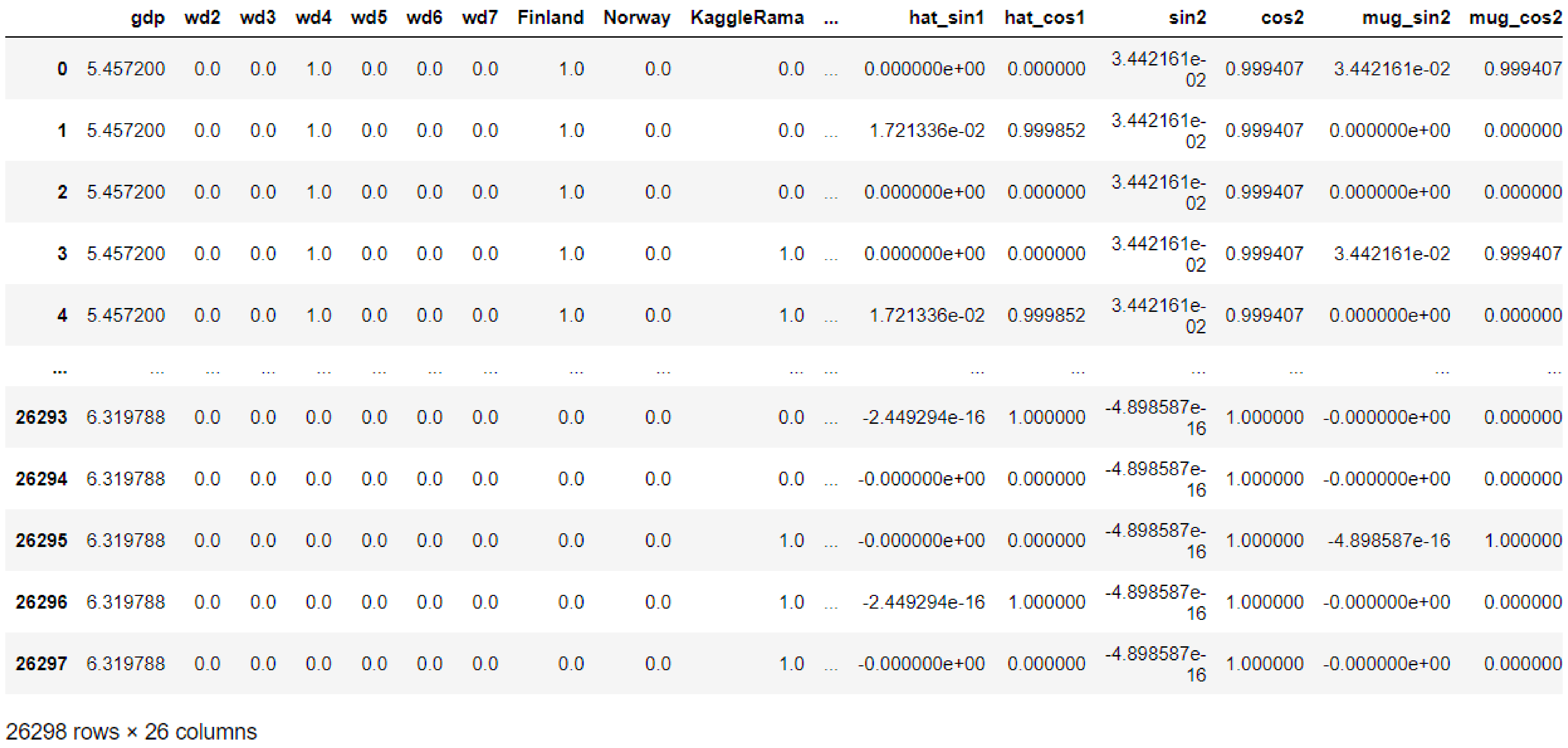
Data processing

Problem Description
Data Clean
Method analysis
Data processing
Model Training
Related Work and Challenges
GOAM Algorithm
Evaluation Results
Conclusion

Defn Use Pandas database to operate the data, add GDP information, and add seasonal indicators every week. Unique coding for commodities, countries and stores. At the same time, Fourier feature is added.

Data processing

- The processed data are as follows:



	gdp	wd2	wd3	wd4	wd5	wd6	wd7	Finland	Norway	KaggleRama	...	hat_sin1	hat_cos1	sin2	cos2	mug_sin2	mug_cos2
0	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.000000e+00	0.000000	3.442161e-02	0.999407	3.442161e-02	0.999407
1	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	1.721336e-02	0.999852	3.442161e-02	0.999407	0.000000e+00	0.000000
2	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.000000e+00	0.000000	3.442161e-02	0.999407	0.000000e+00	0.000000
3	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	...	0.000000e+00	0.000000	3.442161e-02	0.999407	3.442161e-02	0.999407
4	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	...	1.721336e-02	0.999852	3.442161e-02	0.999407	0.000000e+00	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26293	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	-2.449294e-16	1.000000	-4.898587e-16	1.000000	-0.000000e+00	0.000000
26294	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	-0.000000e+00	0.000000	-4.898587e-16	1.000000	-0.000000e+00	0.000000
26295	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	-0.000000e+00	0.000000	-4.898587e-16	1.000000	-4.898587e-16	1.000000
26296	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	-2.449294e-16	1.000000	-4.898587e-16	1.000000	-0.000000e+00	0.000000
26297	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	-0.000000e+00	0.000000	-4.898587e-16	1.000000	-0.000000e+00	0.000000

26298 rows × 26 columns

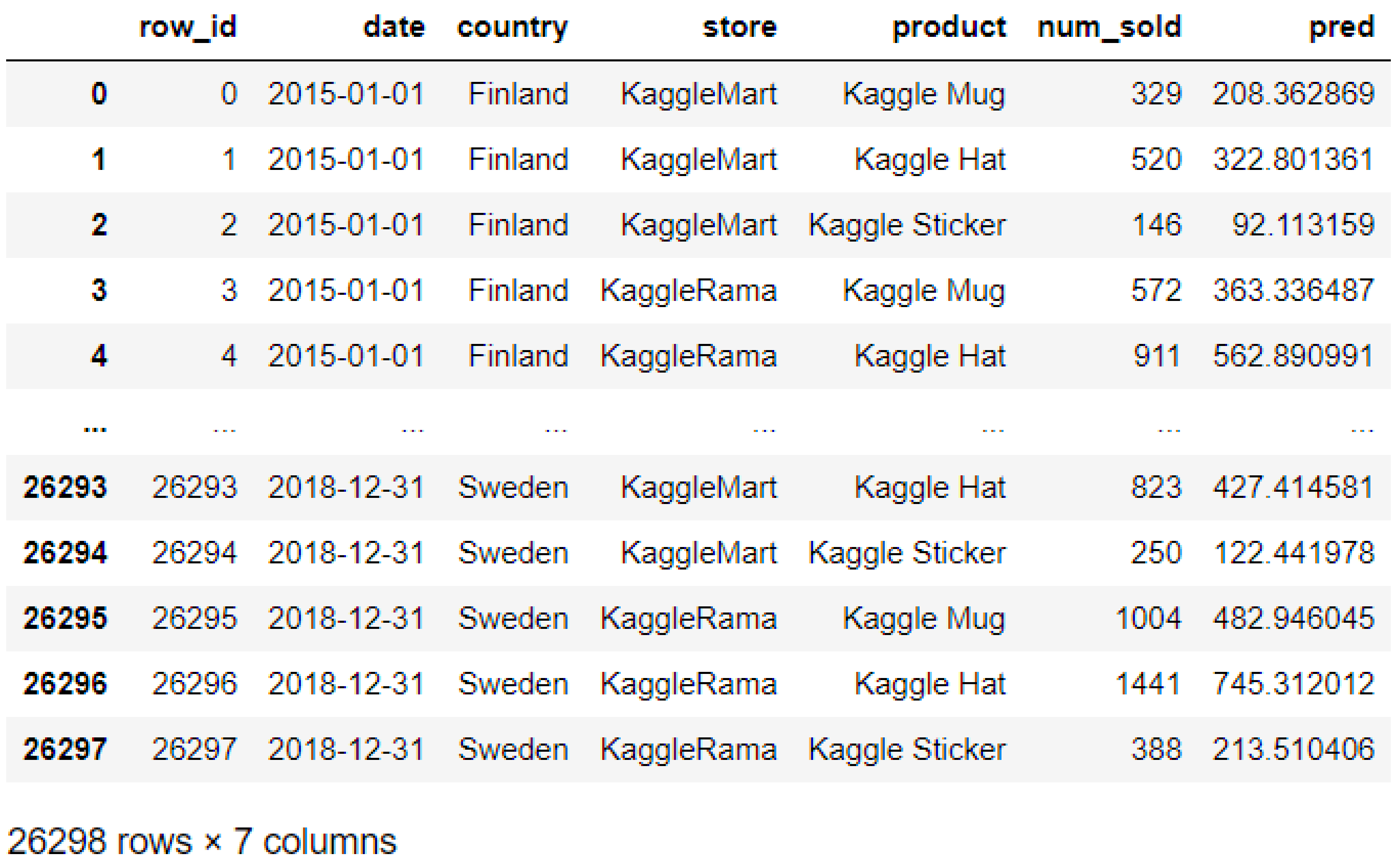
Figure 9: GDP of 2015-2018

Model Training

Model Training

Defn

The model uses the linear regression model in sk-learn and uses SMAPE as the loss function.The training results of the model are as follows.:



	row_id	date	country	store	product	num_sold	pred
0	0	2015-01-01	Finland	KaggleMart	Kaggle Mug	329	208.362869
1	1	2015-01-01	Finland	KaggleMart	Kaggle Hat	520	322.801361
2	2	2015-01-01	Finland	KaggleMart	Kaggle Sticker	146	92.113159
3	3	2015-01-01	Finland	KaggleRama	Kaggle Mug	572	363.336487
4	4	2015-01-01	Finland	KaggleRama	Kaggle Hat	911	562.890991
...	...	...	...	...	...	...	...
26293	26293	2018-12-31	Sweden	KaggleMart	Kaggle Hat	823	427.414581
26294	26294	2018-12-31	Sweden	KaggleMart	Kaggle Sticker	250	122.441978
26295	26295	2018-12-31	Sweden	KaggleRama	Kaggle Mug	1004	482.946045
26296	26296	2018-12-31	Sweden	KaggleRama	Kaggle Hat	1441	745.312012
26297	26297	2018-12-31	Sweden	KaggleRama	Kaggle Sticker	388	213.510406

26298 rows × 7 columns

Figure 10: Model Training

Model Training

Defn

The predicted loss value chart is as follows:

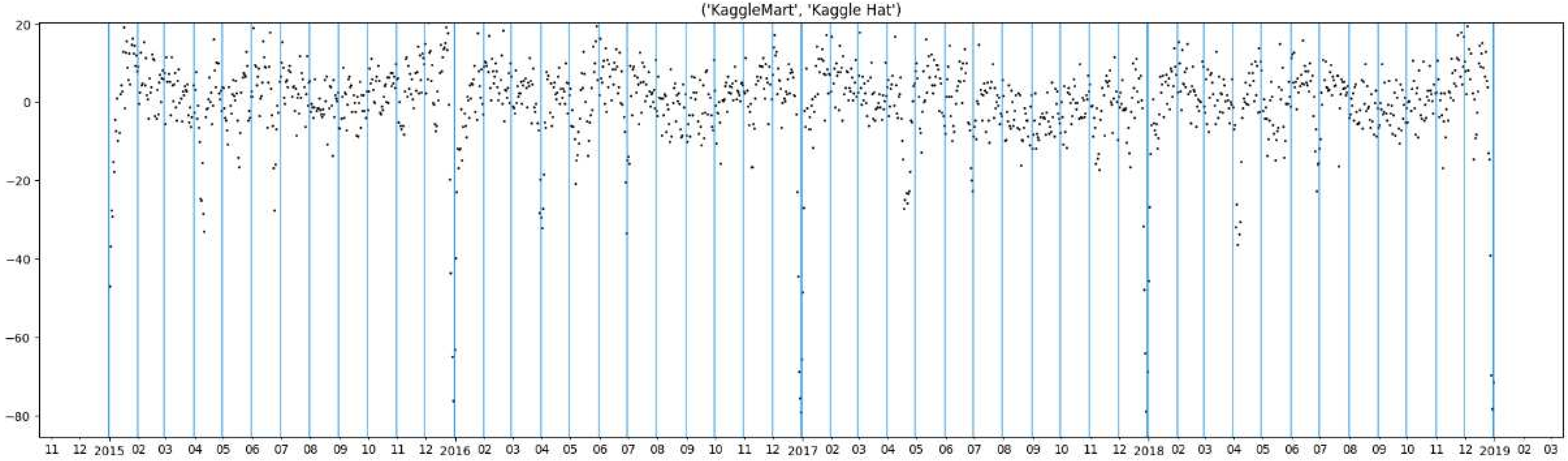


Figure 11: Model Training

Term Definition

- Non-Trivial Outlying Subspaces
  - Multi-dimension subspaces.
  - G_q's outlying degree $\rho(\cdot) > \alpha$.

Problem Description

Data Clean

Method analysis

Data processing

Model Training

Related Work and Challenges

Related Work - Outlying Aspects Mining

Challenges (1)

GOAM Algorithm

Evaluation Results

Conclusion

Related Work and Challenges

Related Work - Outlying Aspects Mining

- Existing Methods - Feature selection
  - To distinguish two classes: the query point (positive) & rest of data (negative)

Disadvantages

Advantages

Related Work - Outlying Aspects Mining

- Existing Methods - Score-and-search
  - Define an outlying score function.
  - Search subspaces.

Disadvantages

- Dimensionality bias.
- Search efficiency is Not high (dataset is large).
- Not identify group outlying aspects.

Advantages

- Quantify the outlying degree correctly.
- High Comprehensibility.

Group Outlying Aspects Mining

- Focus on differences between groups.
- Multiple points.

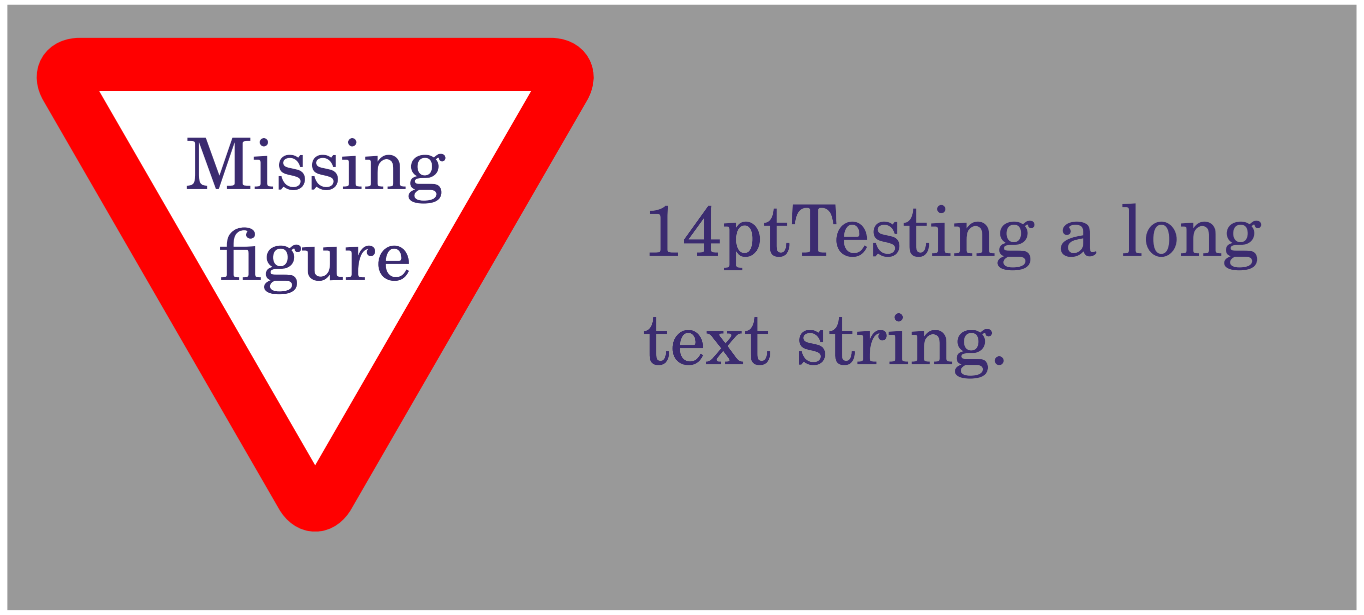


Figure 12: Group Outlying Aspects Target

Outlying Aspects Mining

- Concentrates on differences between objects.
- One point.

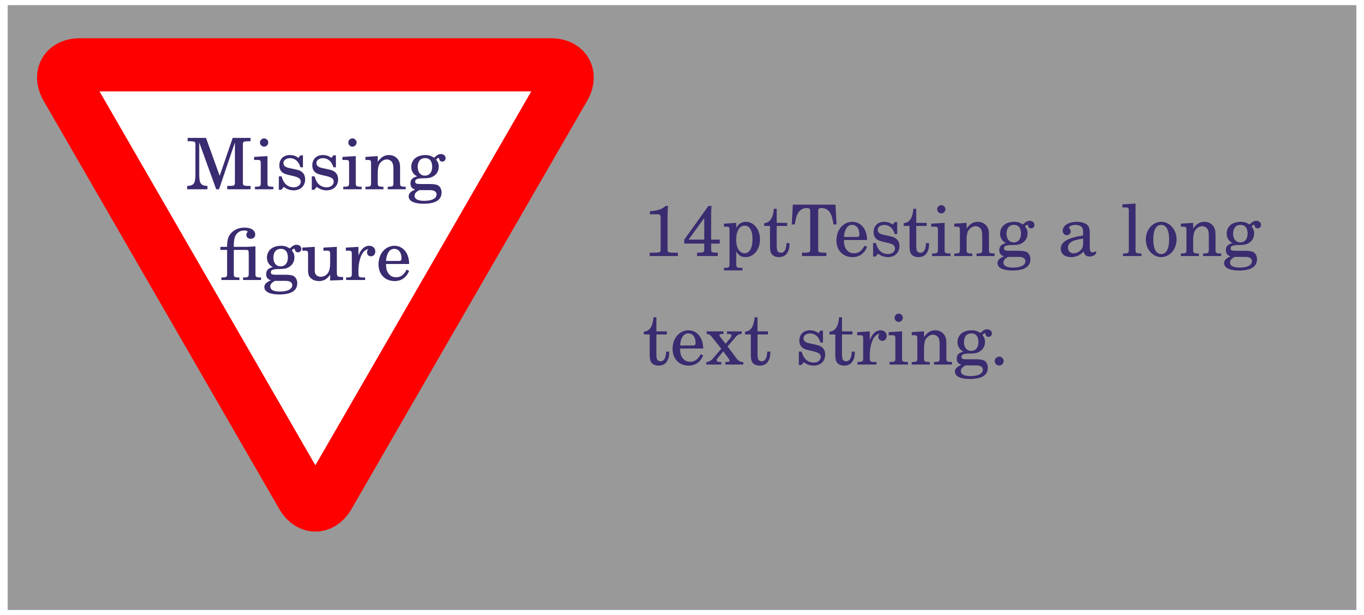


Figure 13: Outlying Aspects Target

Challenges (1)

- How to represent the group features.
    - Can be affected by outlier values.
    - Can Not reflect the overall distribution of group features.

Challenges (2)

Problem Description
Data Clean
Method analysis
Data processing
Model Training
Related Work and Challenges
Related Work - Outlying Aspects Mining
Challenges (1)
GOAM Algorithm
Evaluation Results
Conclusion

- How to evaluate the outlying degree in different aspects.
  - Need design a scoring function when necessary.
  - Adopting an appropriate scoring function (without dimension bias) remains a problem.

Challenges (3)

Problem Description
Data Clean
Method analysis
Data processing
Model Training
Related Work and Challenges
Related Work - Outlying Aspects Mining
Challenges (1)
GOAM Algorithm
Evaluation Results
Conclusion

- How to improve the efficiency.
  - When the dimension of the data is high, the candidate subspace grows exponentially.
  - It will easily go beyond the limits of the computation resources.

Problem Description

Data Clean

Method analysis

Data processing

Model Training

Related Work and Challenges

GOAM Algorithm

Step One - Group Feature Extraction

Step Two - Outlying Degree Scoring

Step Three - Outlying Aspects Identification

Evaluation Results

Conclusion

GOAM Algorithm

Framework of GOAM algorithm:

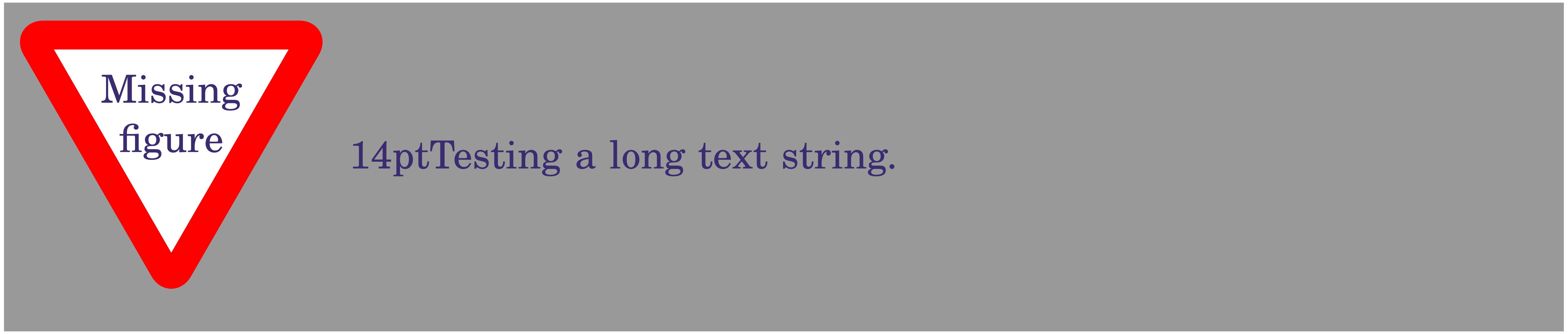


Figure 14: Framework of GOAM Algorithm

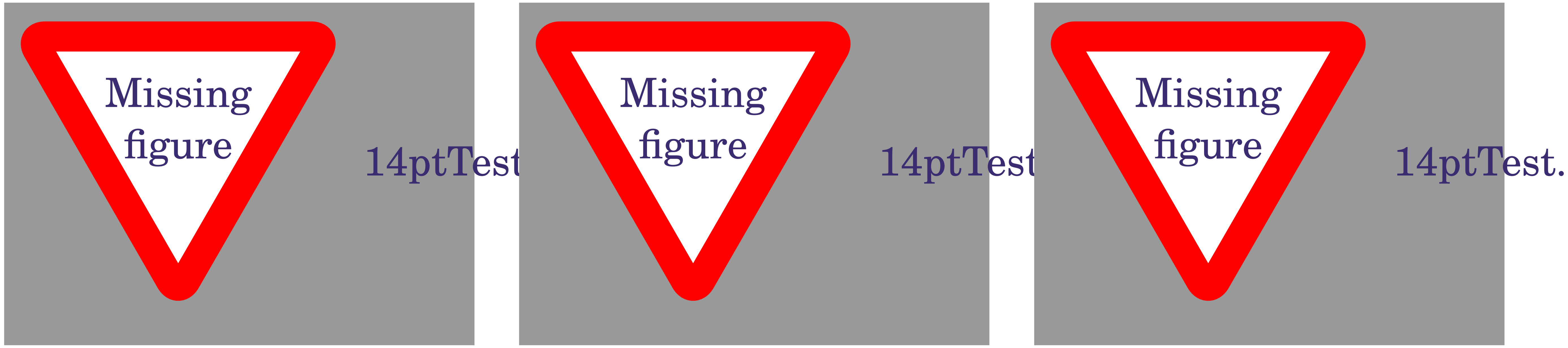
Step One - Group Feature Extraction

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



(a) f_1 (b) f_2 (c) f_3

Figure 15: Histogram of G_q on three features

Step Two - Outlying Degree Scoring

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

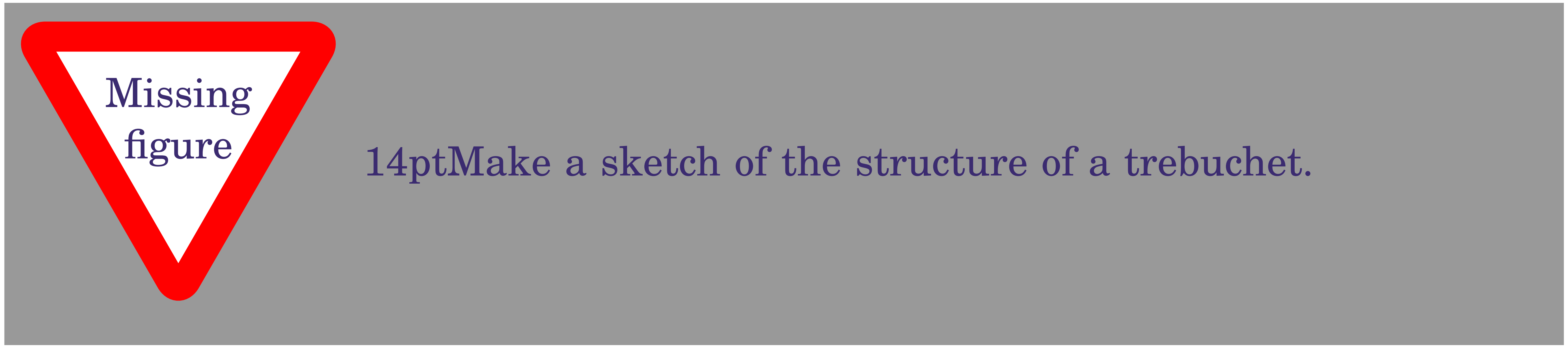


Figure 16: EMD of one feature

Step Two - Outlying Degree Scoring

- Calculate the outlying degree

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

- n ⇔ the number of contrast groups.
- h_{k_s} ⇔ the histogram representation of G_k in the subspace s.

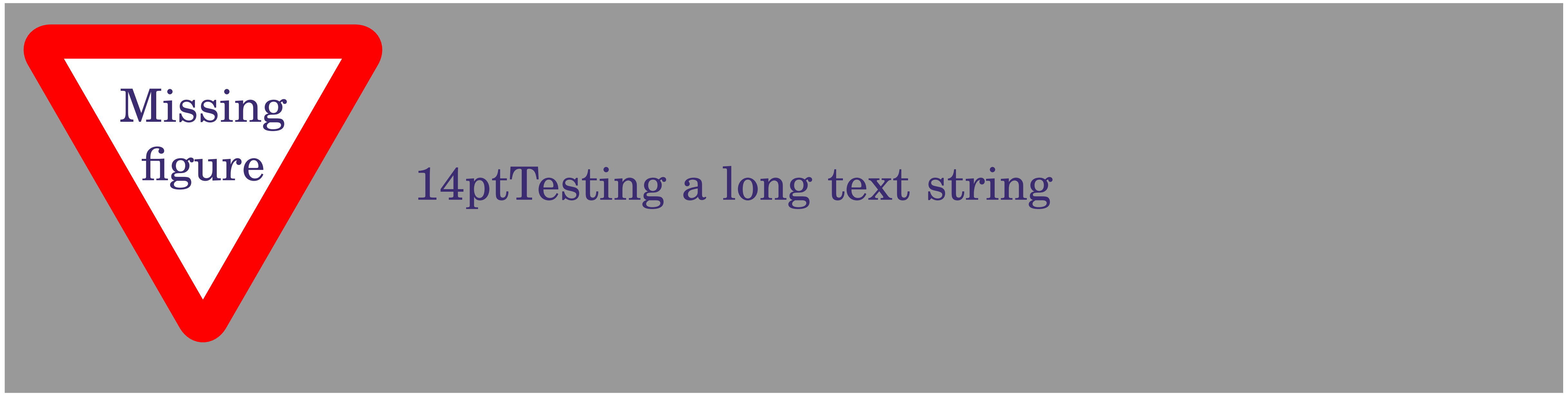
Step Three - Outlying Aspects Identification

Problem Description
Data Clean
Method analysis
Data processing
Model Training
Related Work and Challenges
GOAM Algorithm
Step One - Group Feature Extraction
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
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

- Pseudo code of GOAM algorithm



Illustration

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

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

Reduction of Complexity

- Bottom-up search strategy.
- Reduce the size of candidate subspaces.

Efficiency

- Before: $O(2^d)$
  Now: $O(d * n^2)$

Problem Description

Data Clean

Method analysis

Data processing

Model Training

Related Work and Challenges

GOAM Algorithm

Evaluation Results

Synthetic Dataset

NBA Dataset

Conclusion

Evaluation Results

Evaluation

- $Accuracy = \frac{P}{T}$

  P: Identified outlying aspects

  T: Real outlying aspects

Synthetic Dataset

- Synthetic Dataset and Ground Truth

Table 3: Synthetic Dataset and Ground Truth

Query group	$\mathbf{F_1}$	$\mathbf{F_2}$	F_3	$\mathbf{F_4}$	F_5	F_6	F_7	F_8
i_1	**10**	**8**	9	**7**	7	6	6	8
i_2	**9**	**9**	7	**8**	9	9	8	9
i_3	**8**	**10**	8	**9**	6	8	7	8
i_4	**8**	**8**	6	**7**	8	8	6	7
i_5	**9**	**9**	9	**7**	7	7	8	8
i_6	**8**	**10**	8	**8**	6	6	8	7
i_7	**9**	**9**	7	**9**	8	8	8	7
i_8	**10**	**9**	10	**7**	7	7	7	7
i_9	**9**	**10**	8	**8**	7	6	7	7
i_{10}	**9**	**9**	7	**7**	7	8	8	8

Synthetic Dataset Results

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%

NBA Dataset

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*).

NBA Dataset

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

NBA Dataset

- 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,+∞]

NBA Dataset Results

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}

Conclusion

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.

Questions?

Contact Information

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

✉ GANGLI@TULIP.ORG.AU

🏠 TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING