Bike Sharing Demand
Forecast use of a city bikeshare system

Dong Zhu

Deakin University

(None)

Overview

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
Conclusion

Problem Definition
- Bike Sharing Demand Prediction

Data exploration
- Check for missing vaules
- Check for outliers

Data visualization
- Time characteristic analysis
- Weather characteristics analysis

Feature selection
- Correlation analysis
- Weather characteristics analysis
- Step Two - Outlying Degree Scoring
- Step Three - Outlying Aspects Identification

Evaluation Results
- Synthetic Dataset
- NBA Dataset

Conclusion

Problem Definition
Bike Sharing Demand Prediction
Data exploration
Data visualization
Feature selection
Evaluation Results
Conclusion

Problem Definition

Problem Definition
Bike Sharing Demand Prediction
Data exploration
Data visualization
Feature selection
Evaluation Results
Conclusion

Definition

Bike-sharing systems are a means of renting bikes, through which people can rent a bike from any place and return it when they arrive at their destination. The bike-sharing system clearly records the time of travel, the place of departure, the place of arrival and the time. Therefore, it can be used to study mobility in cities. In this project, historical usage patterns were combined with weather data to predict bike rental demand in Washington, D.C.

- Researchers can use bike sharing systems as a sensor network, which can be used for studying mobility in a city.
- This is a Supervised regression machine learning task.
- The training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month.

Problem Definition
Data exploration
Check for missing vaules
Check for outliers
Data visualization
Feature selection
Evaluation Results
Conclusion

Data exploration

Check for missing vaules

Problem Definition
Data exploration
Check for missing vaules
Check for outliers
Data visualization
Feature selection
Evaluation Results
Conclusion

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10886 entries, 0 to 10885
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    10886 non-null  object
 1   season      10886 non-null  int64
 2   holiday     10886 non-null  int64
 3   workingday  10886 non-null  int64
 4   weather     10886 non-null  int64
 5   temp        10886 non-null  float64
 6   atemp       10886 non-null  float64
 7   humidity    10886 non-null  int64
 8   windspeed   10886 non-null  float64
 9   casual      10886 non-null  int64
 10  registered  10886 non-null  int64
 11  count       10886 non-null  int64
dtypes: float64(3), int64(8), object(1)
memory usage: 1020.7+ KB
None
```

Figure 1: Training data information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6493 entries, 0 to 6492
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    6493 non-null   object
 1   season      6493 non-null   int64
 2   holiday     6493 non-null   int64
 3   workingday  6493 non-null   int64
 4   weather     6493 non-null   int64
 5   temp        6493 non-null   float64
 6   atemp       6493 non-null   float64
 7   humidity    6493 non-null   int64
 8   windspeed   6493 non-null   float64
dtypes: float64(3), int64(5), object(1)
memory usage: 456.7+ KB
None
```

Figure 2: Test data information

Check for outliers

Problem Definition

Data exploration

Check for missing vaules

Check for outliers

Data visualization

Feature selection

Evaluation Results

Conclusion

■ Statistical description

	count	mean	std	min	25%	50%	75%	max
season	10886.0	2.506614	1.116174	1.00	2.0000	3.000	4.0000	4.0000
holiday	10886.0	0.028569	0.166599	0.00	0.0000	0.000	0.0000	1.0000
workingday	10886.0	0.680875	0.466159	0.00	0.0000	1.000	1.0000	1.0000
weather	10886.0	1.418427	0.633839	1.00	1.0000	1.000	2.0000	4.0000
temp	10886.0	20.230860	7.791590	0.82	13.9400	20.500	26.2400	41.0000
atemp	10886.0	23.655084	8.474601	0.76	16.6650	24.240	31.0600	45.4550
humidity	10886.0	61.886460	19.245033	0.00	47.0000	62.000	77.0000	100.0000
windspeed	10886.0	12.799395	8.164537	0.00	7.0015	12.998	16.9979	56.9969
casual	10886.0	36.021955	49.960477	0.00	4.0000	17.000	49.0000	367.0000
registered	10886.0	155.552177	151.039033	0.00	36.0000	118.000	222.0000	886.0000
count	10886.0	191.574132	181.144454	1.00	42.0000	145.000	284.0000	977.0000

Figure 3: Data description

Check for outliers

Problem Definition
Data exploration
Check for missing vaules
Check for outliers
Data visualization
Feature selection
Evaluation Results
Conclusion

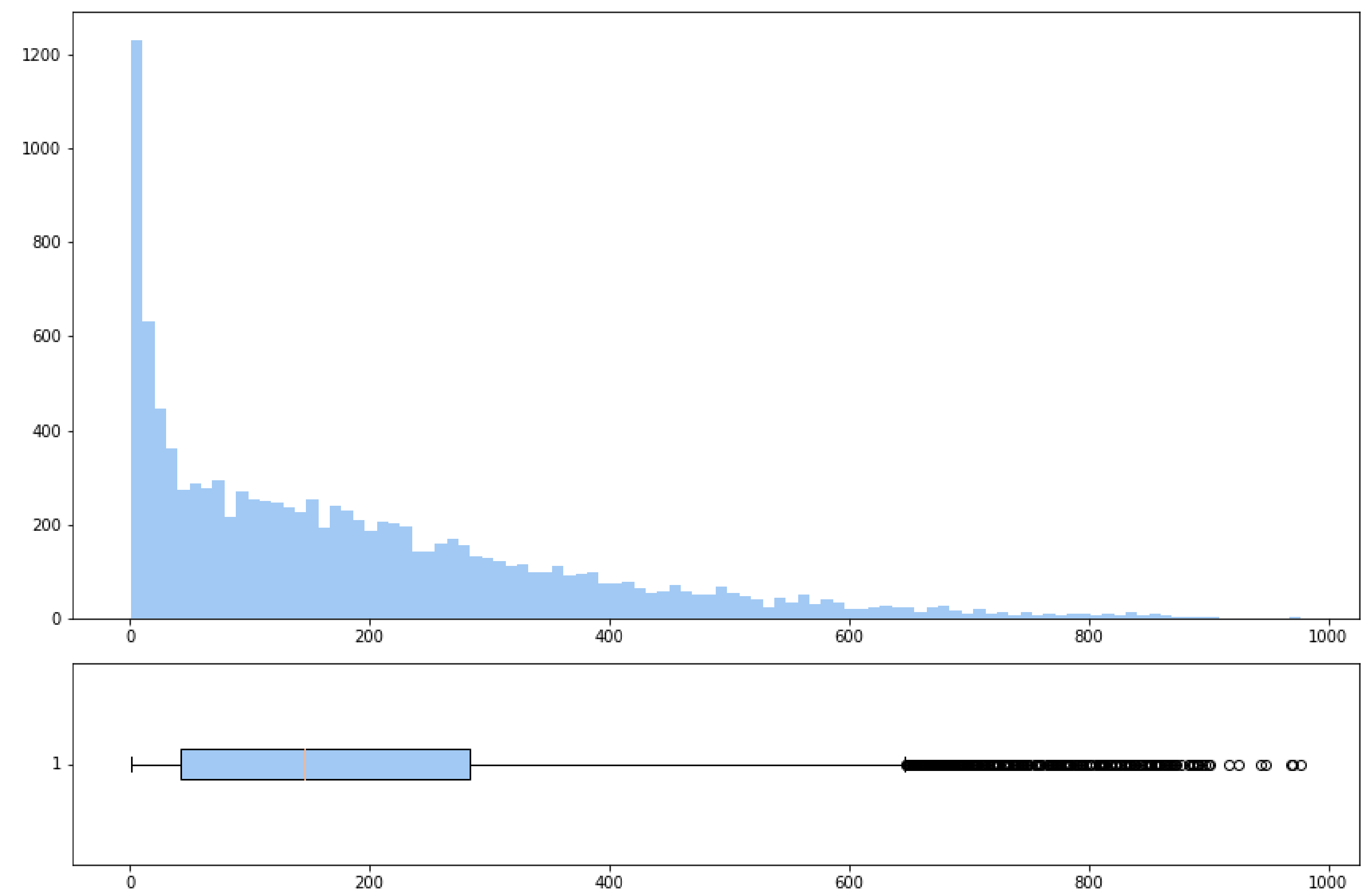


Figure 4: The distribution of the label "count"

Check for outliers

Problem Definition
Data exploration
Check for missing vaules
Check for outliers
Data visualization
Feature selection
Evaluation Results
Conclusion

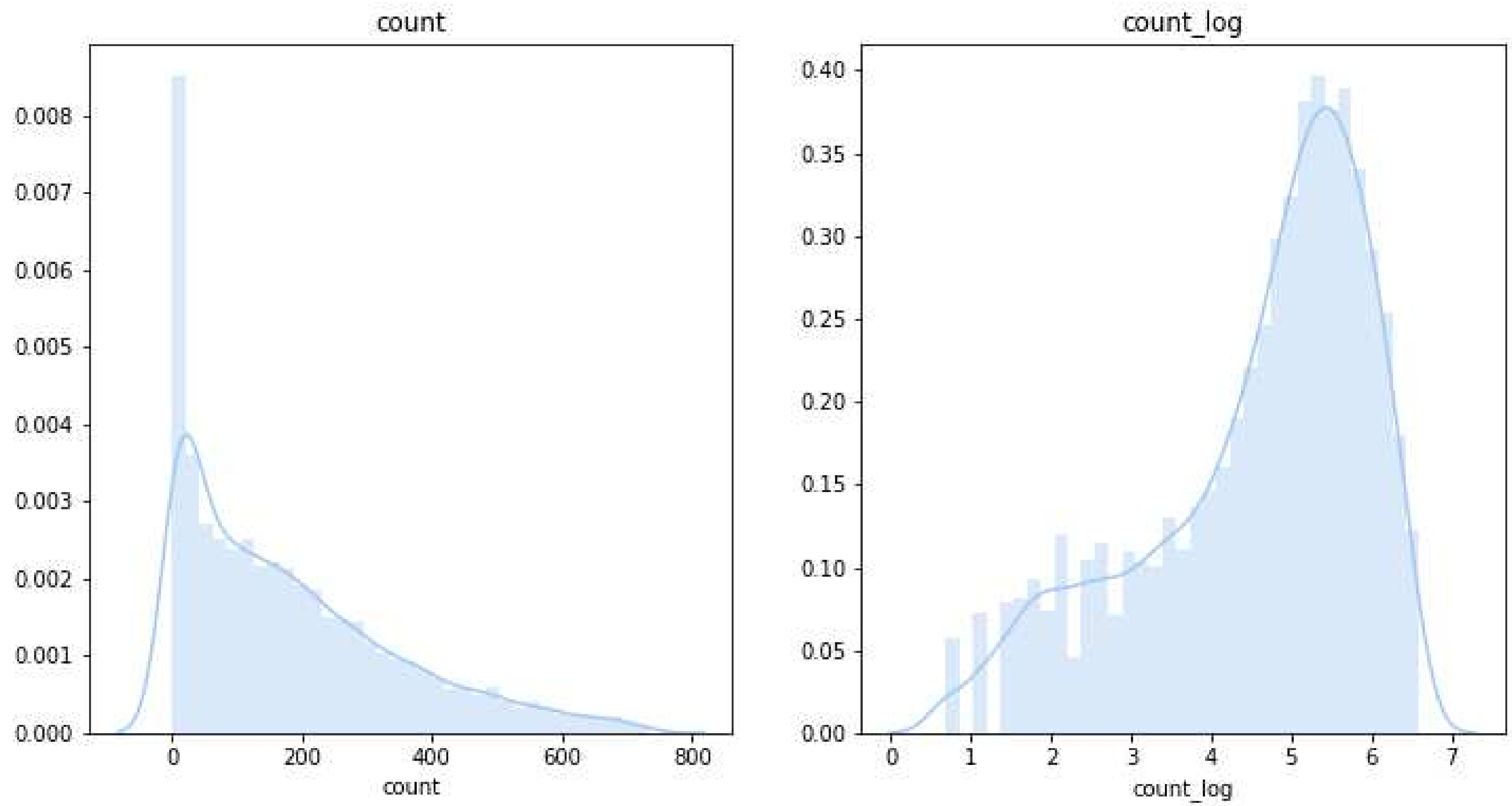


Figure 5: Count distribution compare

Check for outliers

Problem Definition
Data exploration
Check for missing vaules
Check for outliers
Data visualization
Feature selection
Evaluation Results
Conclusion

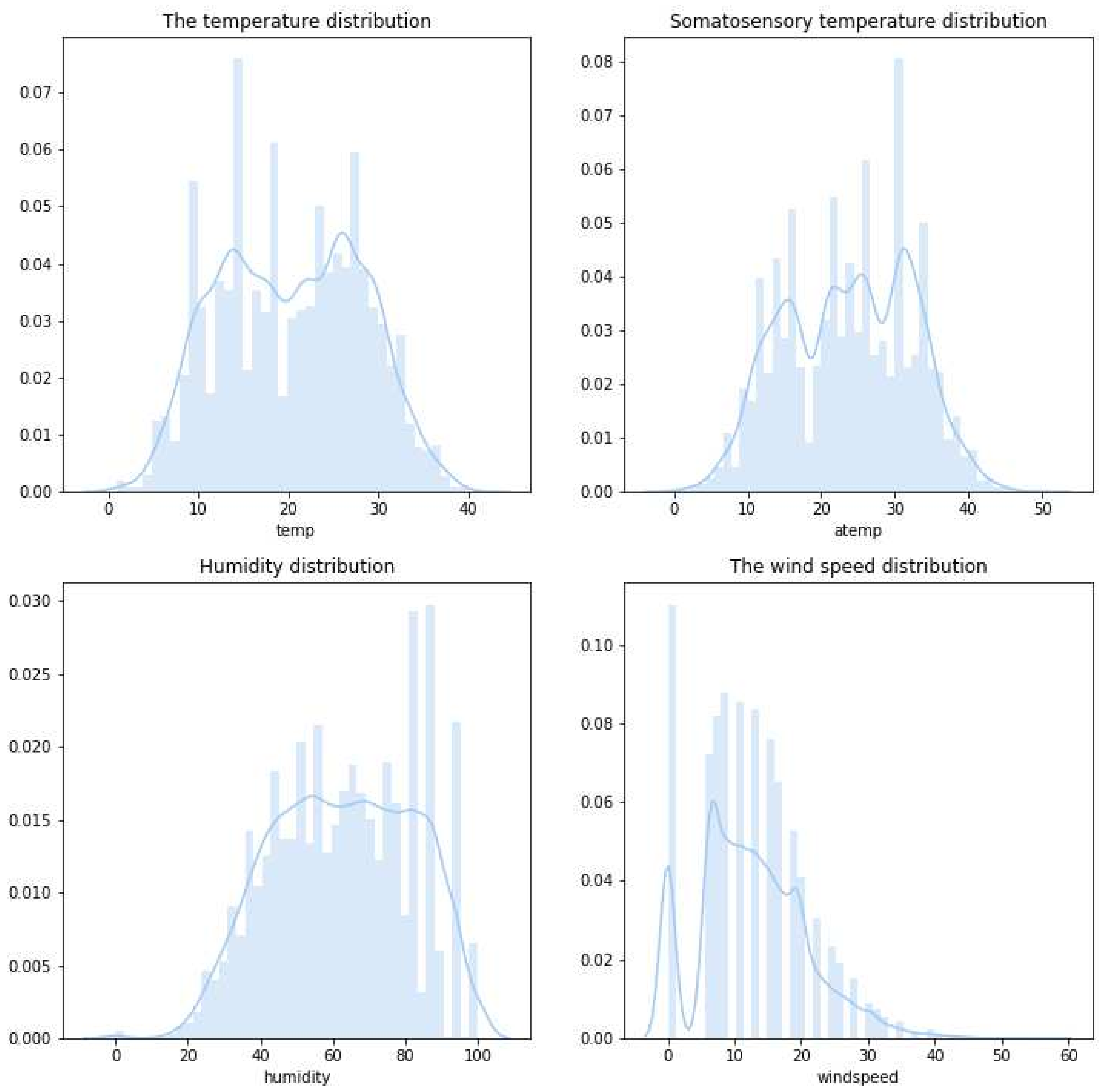


Figure 6: Main features distribution

Check for outliers

Problem Definition
Data exploration
Check for missing vaules
Check for outliers
Data visualization
Feature selection
Evaluation Results
Conclusion

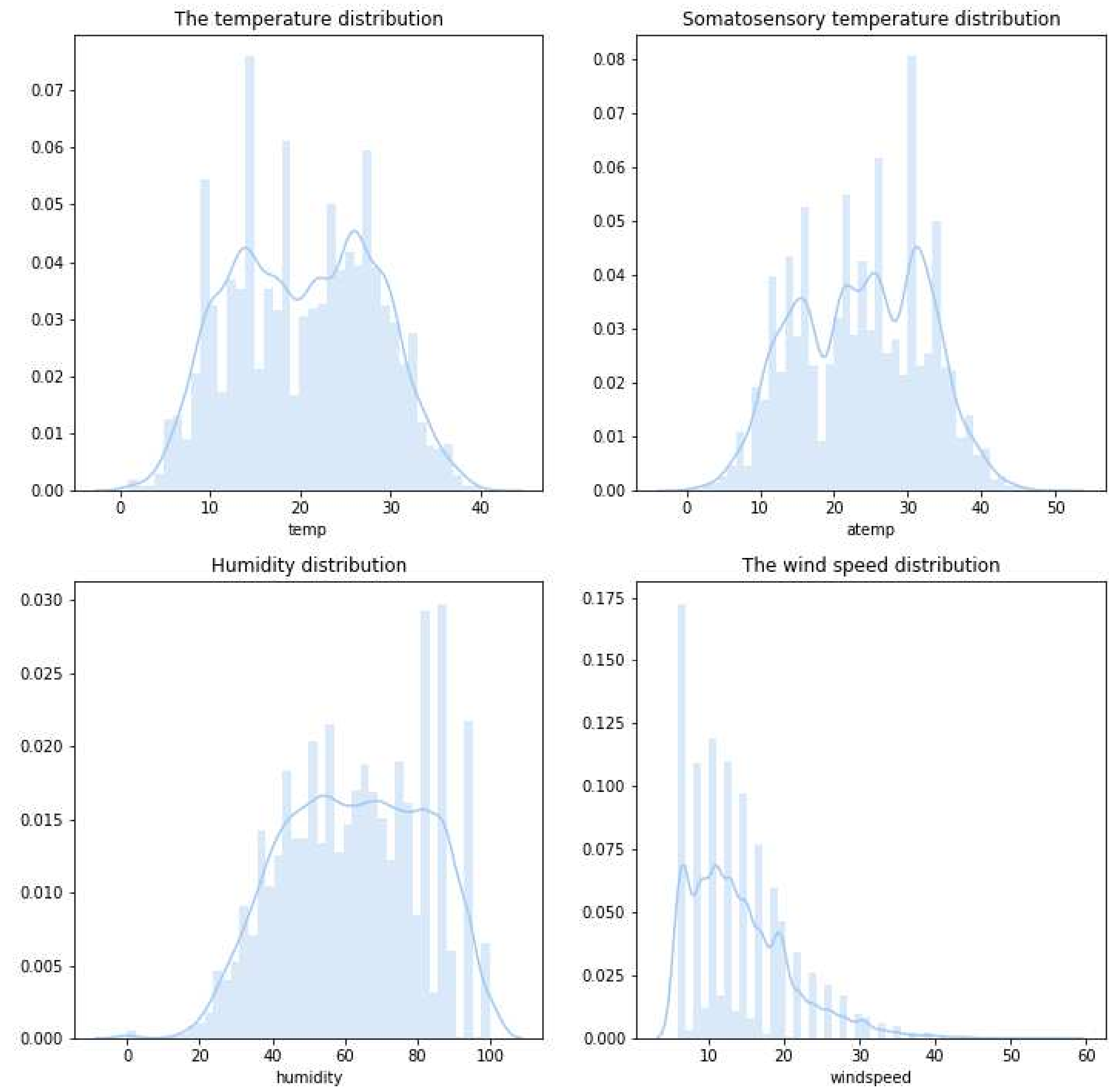


Figure 7: Main features distribution

Problem Definition
Data exploration
Data visualization
Time characteristic analysis
Weather characteristics analysis
Feature selection
Evaluation Results
Conclusion

Data visualization

Time characteristic analysis

Problem Definition
Data exploration
Data visualization
Time characteristic analysis
Weather characteristics analysis
Feature selection
Evaluation Results
Conclusion

- there are two peaks in the graph, one is from 7-8 in the morning, the other is from 5-6 in the afternoon, which is the morning peak and the evening peak respectively, which is in line with the actual situation.

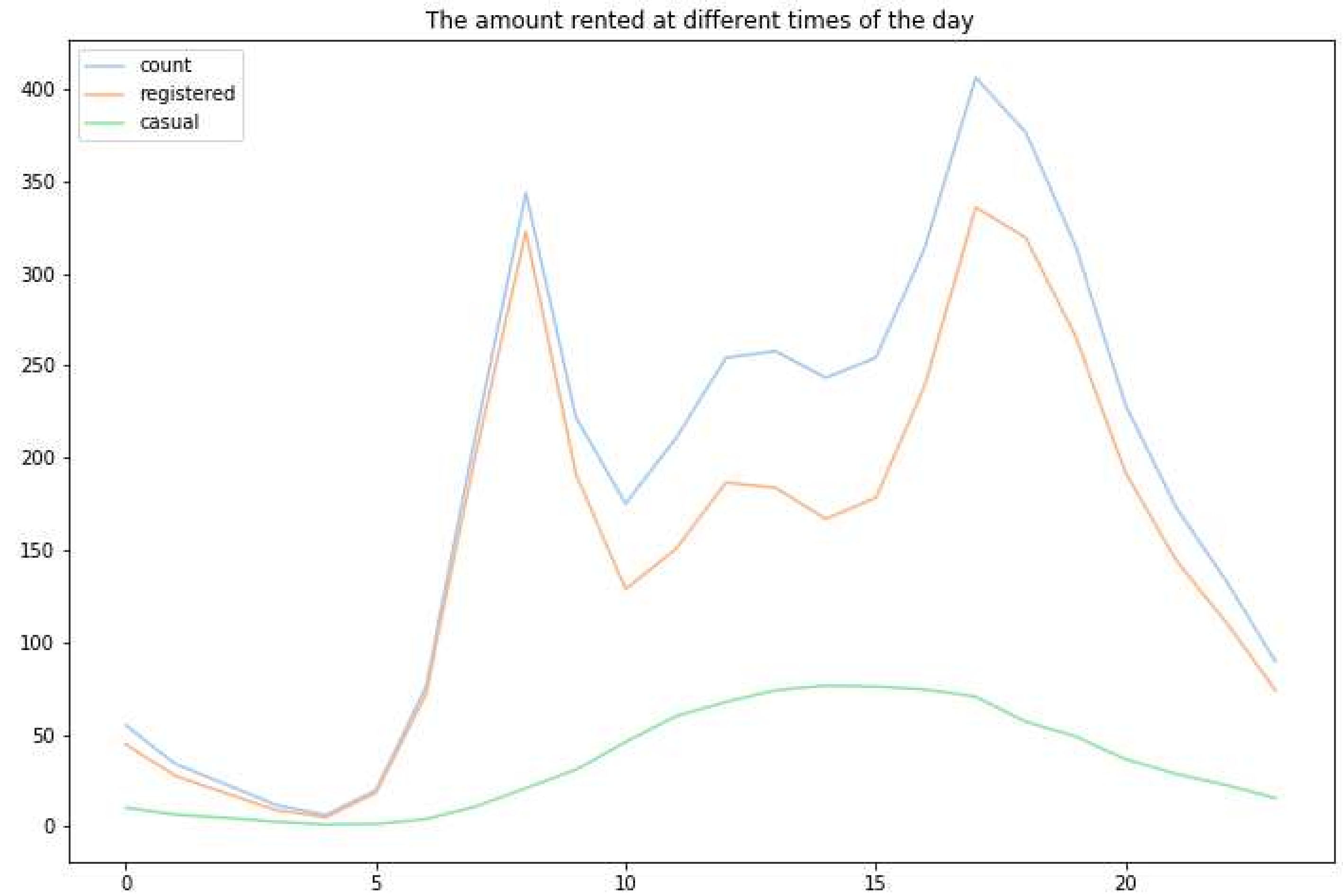


Figure 8: The amount rented at different times of the day

Time characteristic analysis

Problem Definition
Data exploration
Data visualization
Time characteristic analysis
Weather characteristics analysis
Feature selection
Evaluation Results
Conclusion

- from Monday to Friday, 8 in the morning of the day - 9 am and 5 to 7 PM, usage is more, may be caused by time going to work in the morning and evening after work time, include the reason of eating out at the same time, for the weekend, time is more focused, basic usage around 11 PM to 5 PM, This time is supposed to be everyone's leisure time.

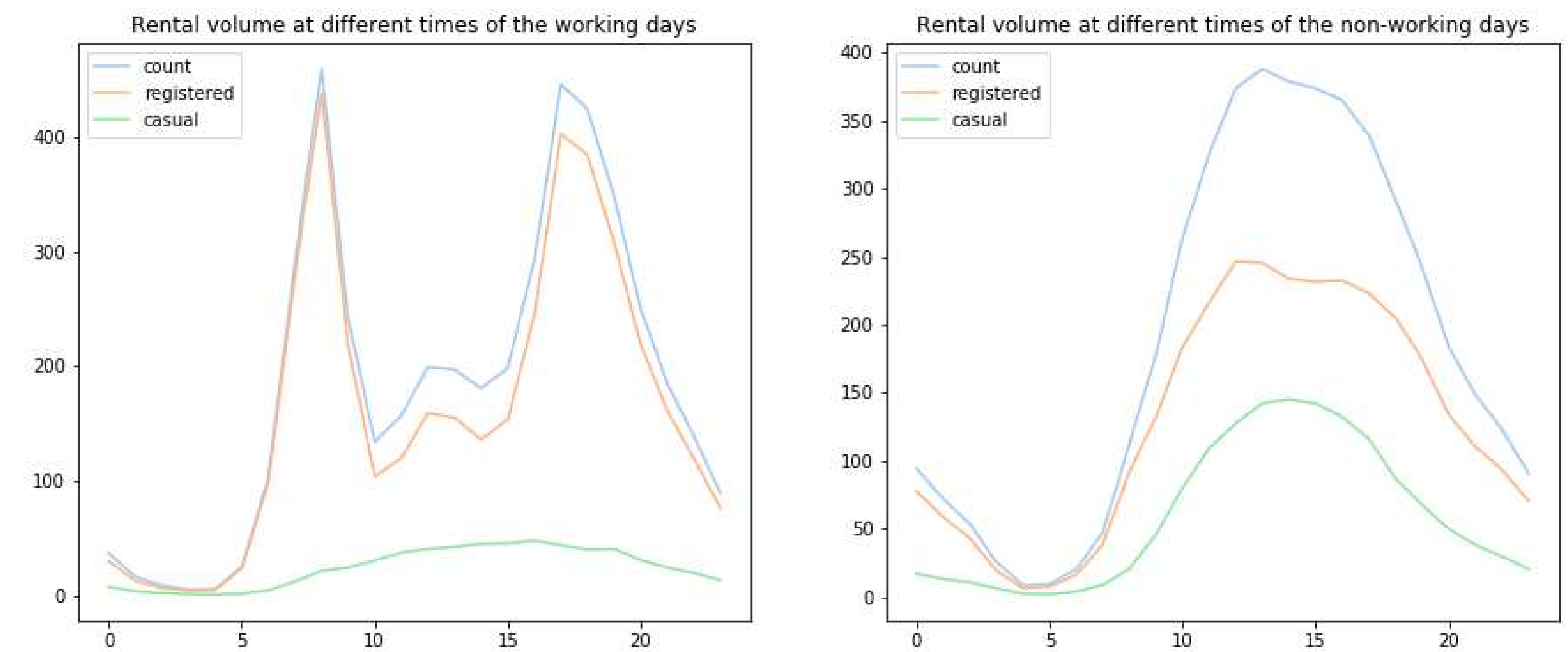


Figure 9: Rental amount at different times of the non-working days and the non-working days

Problem Definition
Data exploration
Data visualization
Time characteristic analysis
Weather characteristics analysis
Feature selection
Evaluation Results
Conclusion

- The usage is obviously lower in spring, probably due to the lower temperature.

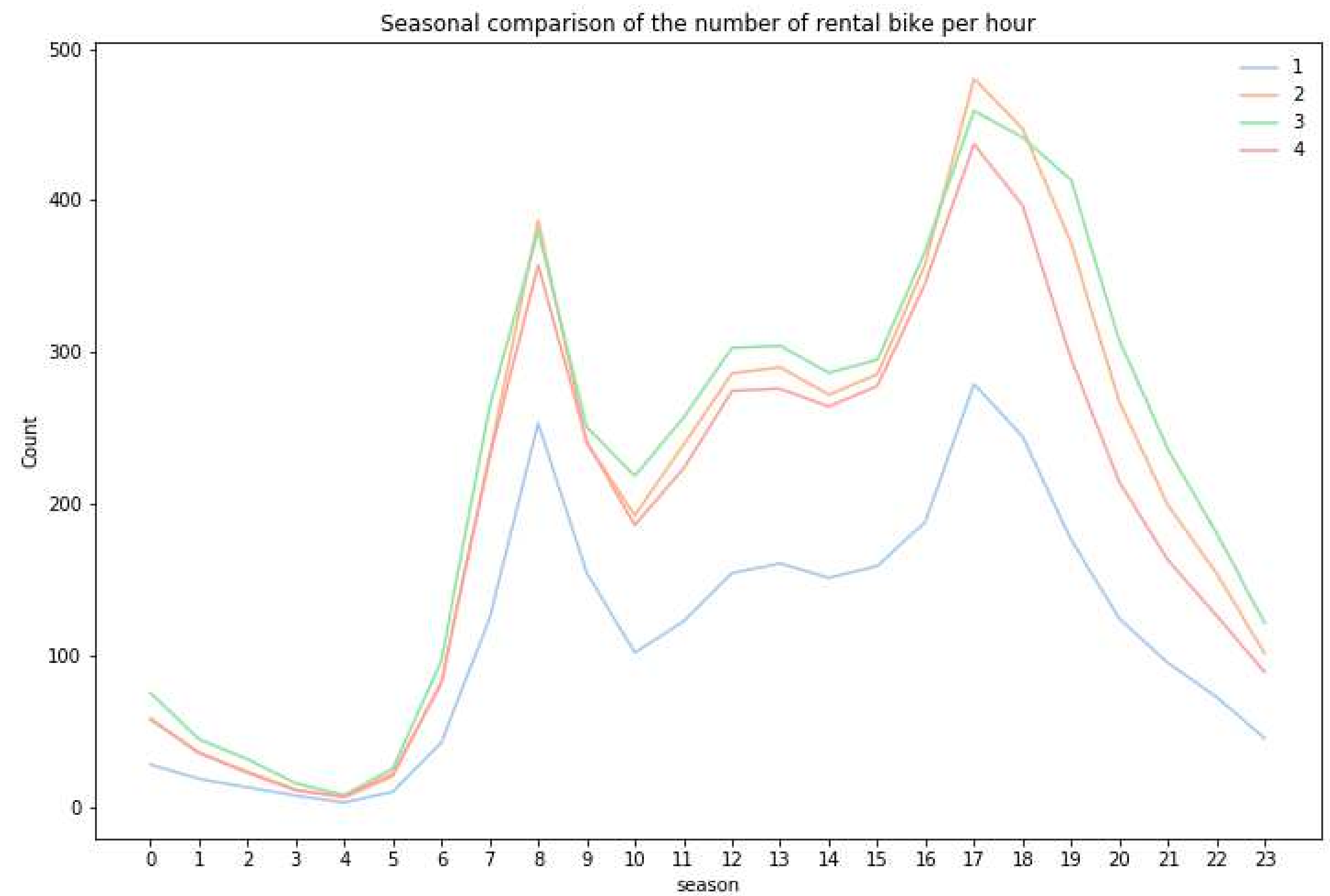


Figure 10: Seasonal comparison of the number of rental bike per hour

Weather characteristics analysis

Problem Definition
Data exploration
Data visualization
Time characteristic analysis
Weather characteristics analysis
Feature selection
Evaluation Results
Conclusion

- Temperatures below 10 degrees, above 30 degrees, and fewer bike rentals – too cold or too hot will damper rental demand.
- The higher the wind, the fewer bike renters - high winds dampen rental demand.
- The higher the humidity in the air, the fewer people who hire bikes - it's more comfortable to ride on dry days.

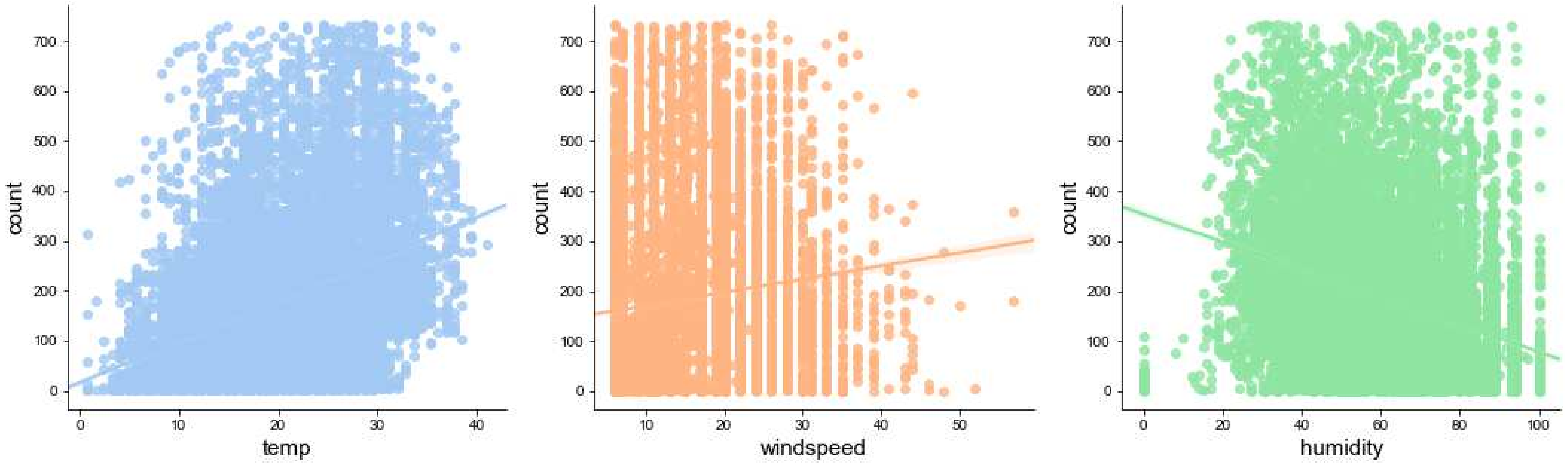


Figure 11: The effect of weather on rental amount

Problem Definition
Data exploration
Data visualization
Feature selection
Correlation analysis
Weather characteristics analysis
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
Conclusion

Feature selection

Correlation analysis

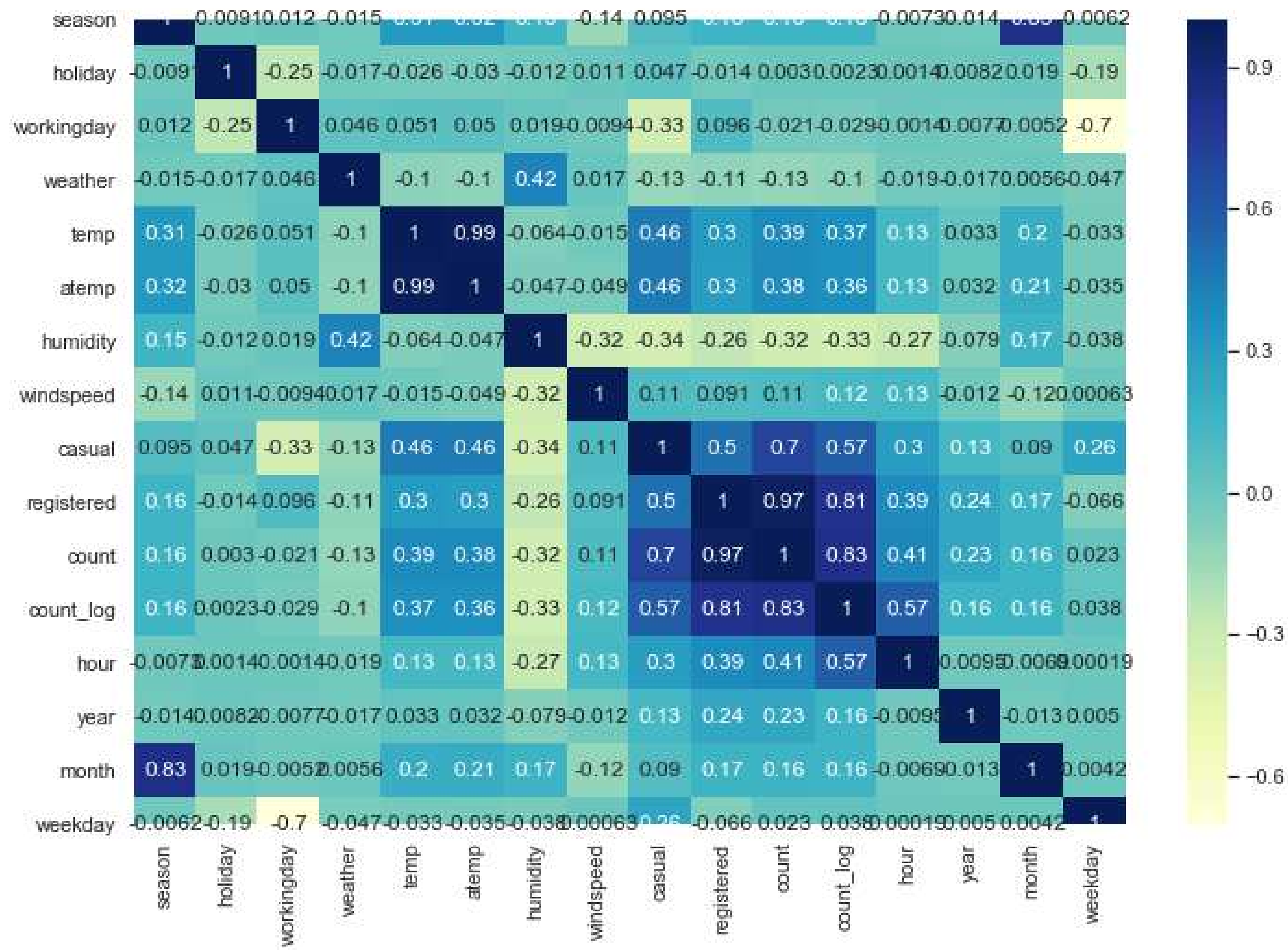


Figure 12: Correlation analysis

Correlation analysis

Problem Definition
Data exploration
Data visualization
Feature selection
Correlation analysis
Weather characteristics analysis
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
Conclusion

- The influence of characteristics on count is as follows: hour>temp>atemp>humidity>month>season>year>weather>windspeed>workingday>wee

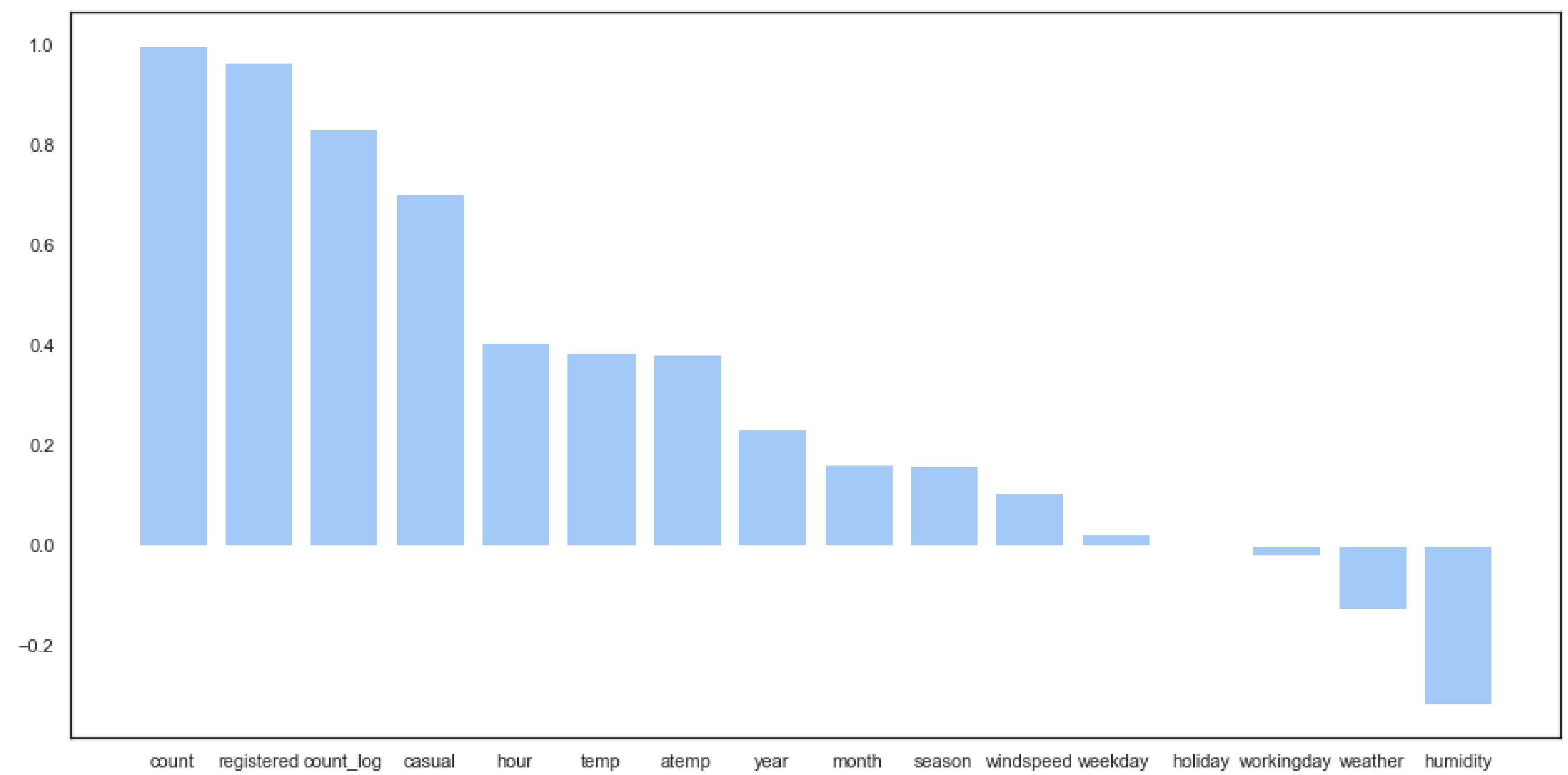


Figure 13: Correlation rank

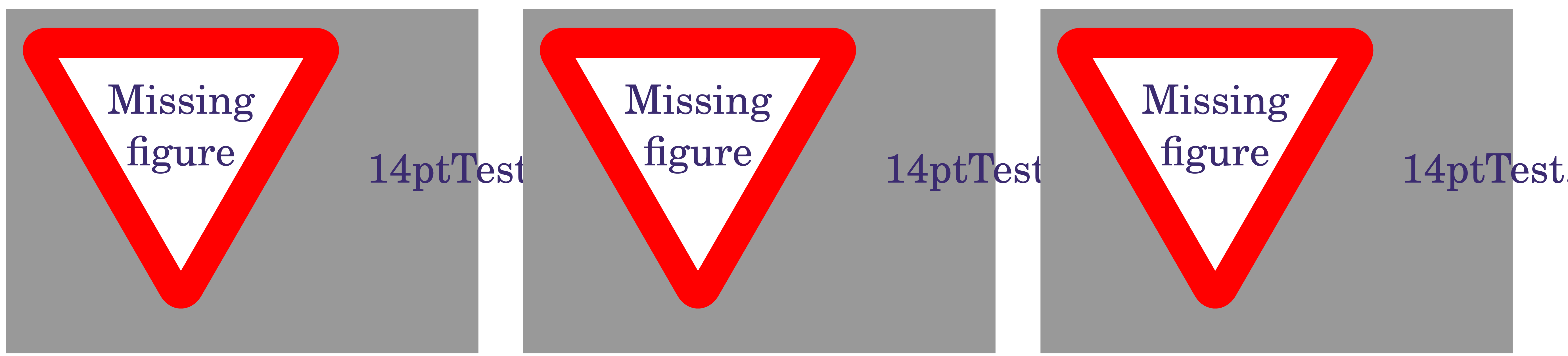
Weather characteristics analysis

- 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 14: 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

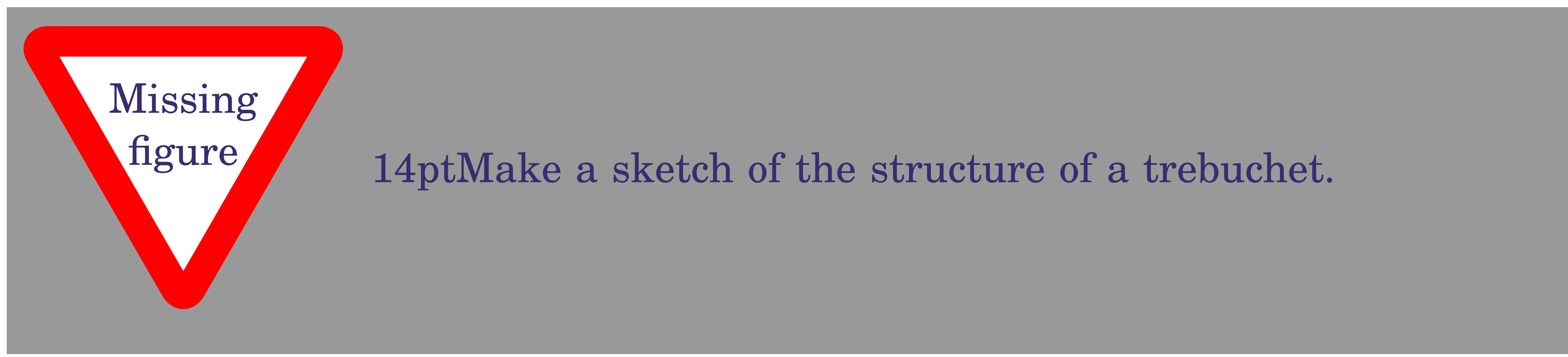


Figure 15: EMD of one feature

Step Two - Outlying Degree Scoring

Problem Definition
Data exploration
Data visualization
Feature selection
Correlation analysis
Weather characteristics analysis
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
Conclusion

- 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

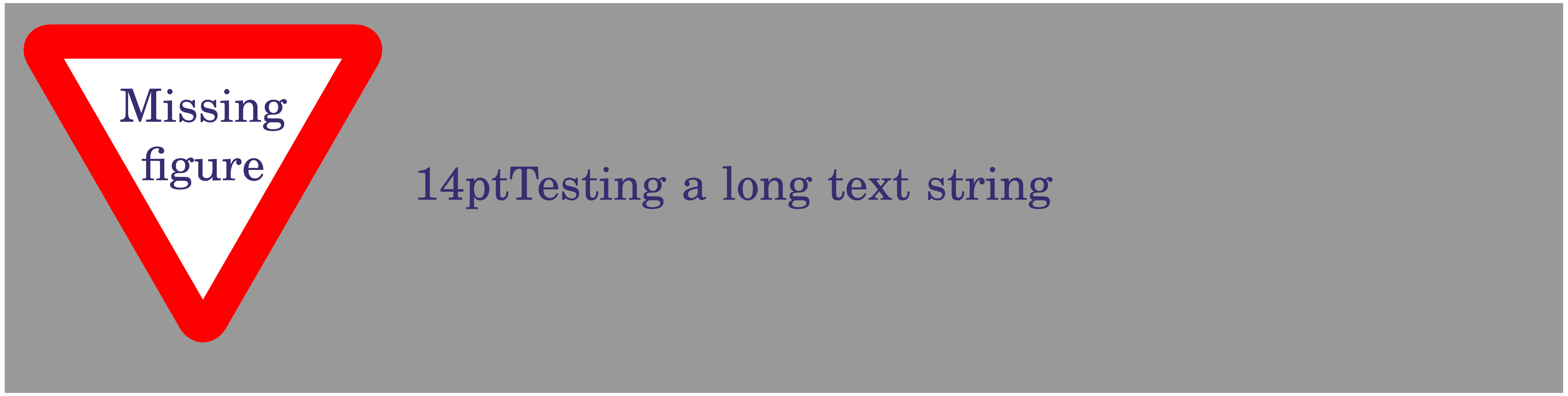
Problem Definition
Data exploration
Data visualization
Feature selection
Correlation analysis
Weather characteristics analysis
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
Conclusion

- 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

Problem Definition
Data exploration
Data visualization
Feature selection
Correlation analysis
Weather characteristics analysis
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
Conclusion

- Pseudo code of GOAM algorithm



Illustration

Problem Definition

Data exploration

Data visualization

Feature selection

Correlation analysis

Weather characteristics analysis

Step Two - Outlying Degree Scoring

Step Three - Outlying Aspects Identification

Evaluation Results

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 exploration
Data visualization
Feature selection
Correlation analysis
Weather characteristics analysis
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
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 exploration
Data visualization
Feature selection
Correlation analysis
Weather characteristics analysis
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
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 exploration
Data visualization
Feature selection
Evaluation Results
Synthetic Dataset
NBA Dataset
Conclusion

Evaluation Results

Evaluation

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
Synthetic Dataset
NBA Dataset
Conclusion

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

  P: Identified outlying aspects

  T: Real outlying aspects

Synthetic Dataset

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
Synthetic Dataset
NBA Dataset
Conclusion

- 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

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
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%

NBA Dataset

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
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*).

NBA Dataset

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
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 exploration
Data visualization
Feature selection
Evaluation Results
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,+∞]

NBA Dataset Results

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
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 exploration
Data visualization
Feature selection
Evaluation Results
Conclusion

Conclusion

Conclusion

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
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?

Problem Definition
Data exploration
Data visualization
Feature selection
Evaluation Results
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