Bike Sharing Demand

YAO YANG

Chongqing University of Posts and Telecommunications test

July 22, 2023

Overview

Problem Definition

- Bike Sharing Demand

Data Clean

- Data Describe
- Data Visualization Plot

Knowledge Discovery

- Variable Relationship Discovery
- Target Variable Analysis
- Fill In Zero Values

Model Solution

- Model Building
- Model Fusion Stacking
- Final prediction result

Problem Definition

Bike Sharing Demand

Defn

The goal of this project is to forecast bike rental demand given the input feature like the duration of travel, departure location, arrival location, and time elapsed.

Evaluation metrics: RMSLE (Root Mean Squard Logarithmic Error) is required to evaluate the model.

$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[log(p_i+1) - log(a_i+1)]^2}$

n is the number of test set samples, pi is the test value, and ai is the actual value. When the root mean square error is smaller, it means that the fitting effect of the data is better and the test value is closer to the actual value.

Data Clean

Data Describe

Defn

You are provided hourly rental data spanning two years. For this competition, 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. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.

- train.csv It contains a training set of target variables.
- test.csv It does not contain a training set of target variables.
- sampleSubmission.csv It is a properly formatted sample submission file.

Data Describe

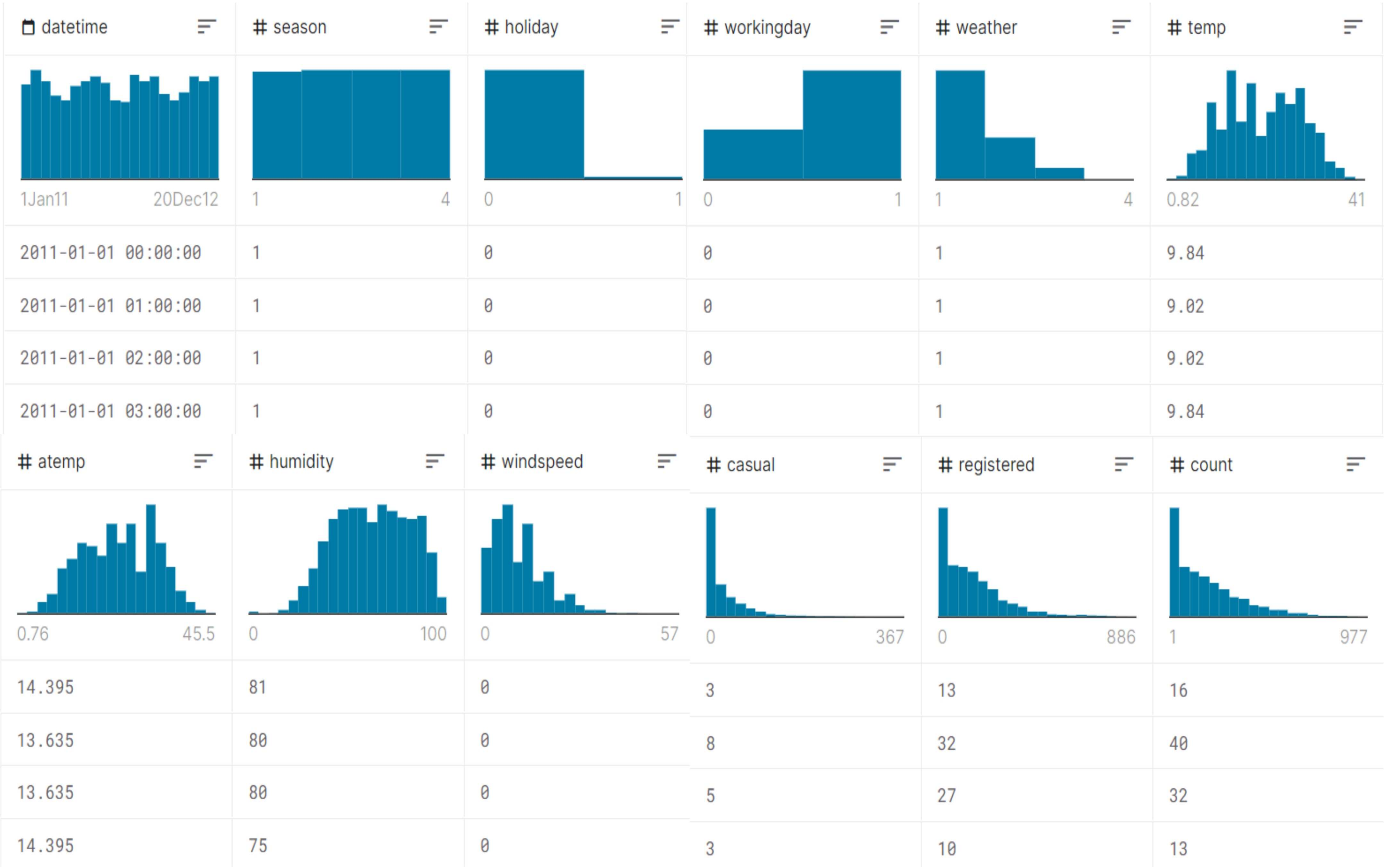
- datetime - hourly date + timestamp
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday
- workingday - whether the day is neither a weekend nor holiday
- weather - 1: Clear, Few clouds, Partly cloudy, Partly cloudy

  2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

  3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

  4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp - temperature in Celsius
- atemp - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated
- registered - number of registered user rentals initiated
- count - number of total rentals

Data Describe



datetime	season	holiday	workingday	weather	temp
1Jan11 20Dec12	1 4	0 1	0 1	1 4	0.82 41
2011-01-01 00:00:00	1	0	0	1	9.84
2011-01-01 01:00:00	1	0	0	1	9.02
2011-01-01 02:00:00	1	0	0	1	9.02
2011-01-01 03:00:00	1	0	0	1	9.84

atemp	humidity	windspeed	casual	registered	count
0.76 45.5	0 100	0 57	0 367	0 886	1 977
14.395	81	0	3	13	16
13.635	80	0	8	32	40
13.635	80	0	5	27	32
14.395	75	0	3	10	13

Figure 1: Describe

Data Visualization Plot

Problem Definition

Data Clean

Data Describe

Data Visualization Plot

Knowledge Discovery

Model Solution

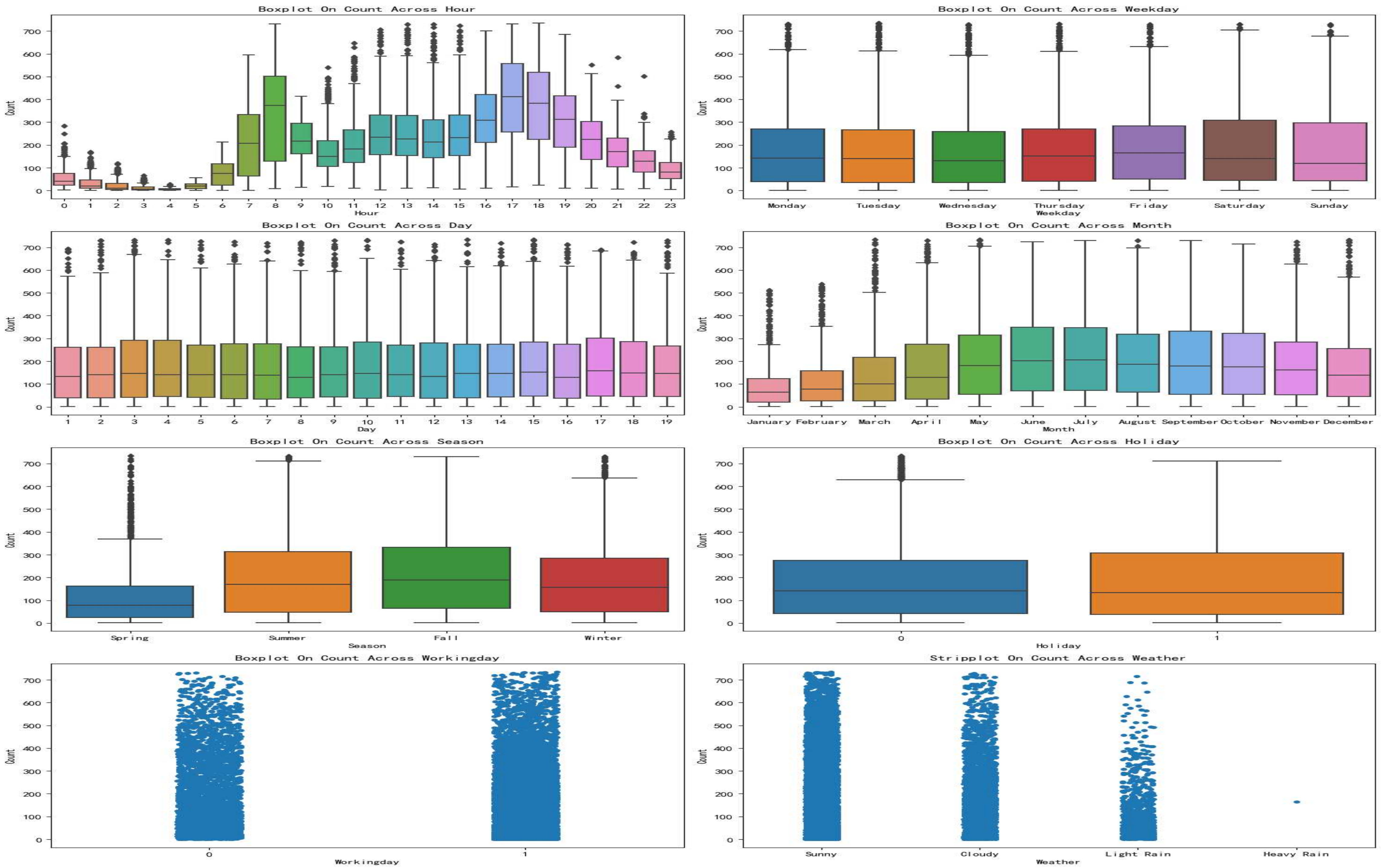


Figure 2: Box Plot and Scatter Plot

Data Visualization Plot

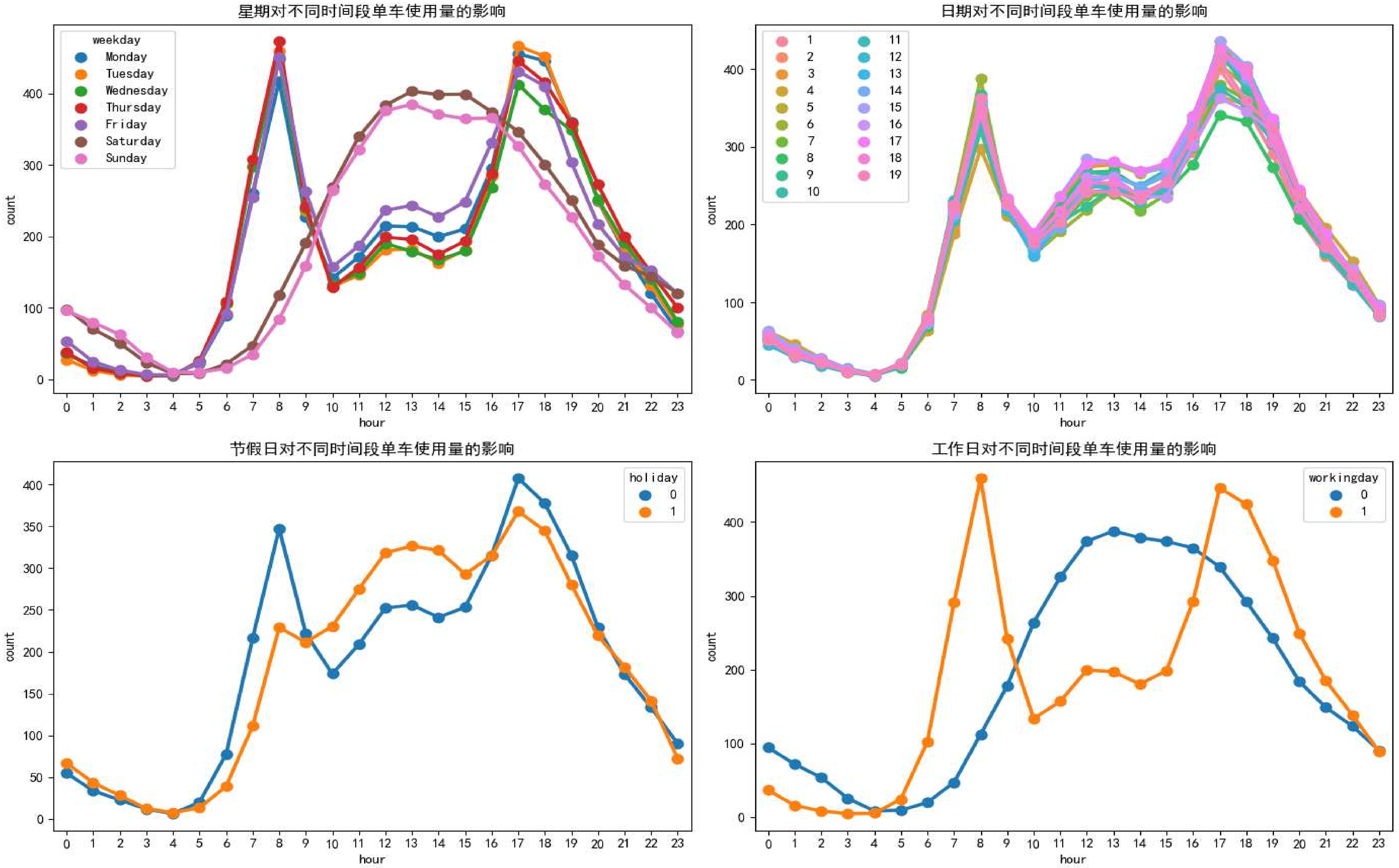


Figure 3: Line Chart

Data Visualization Plot

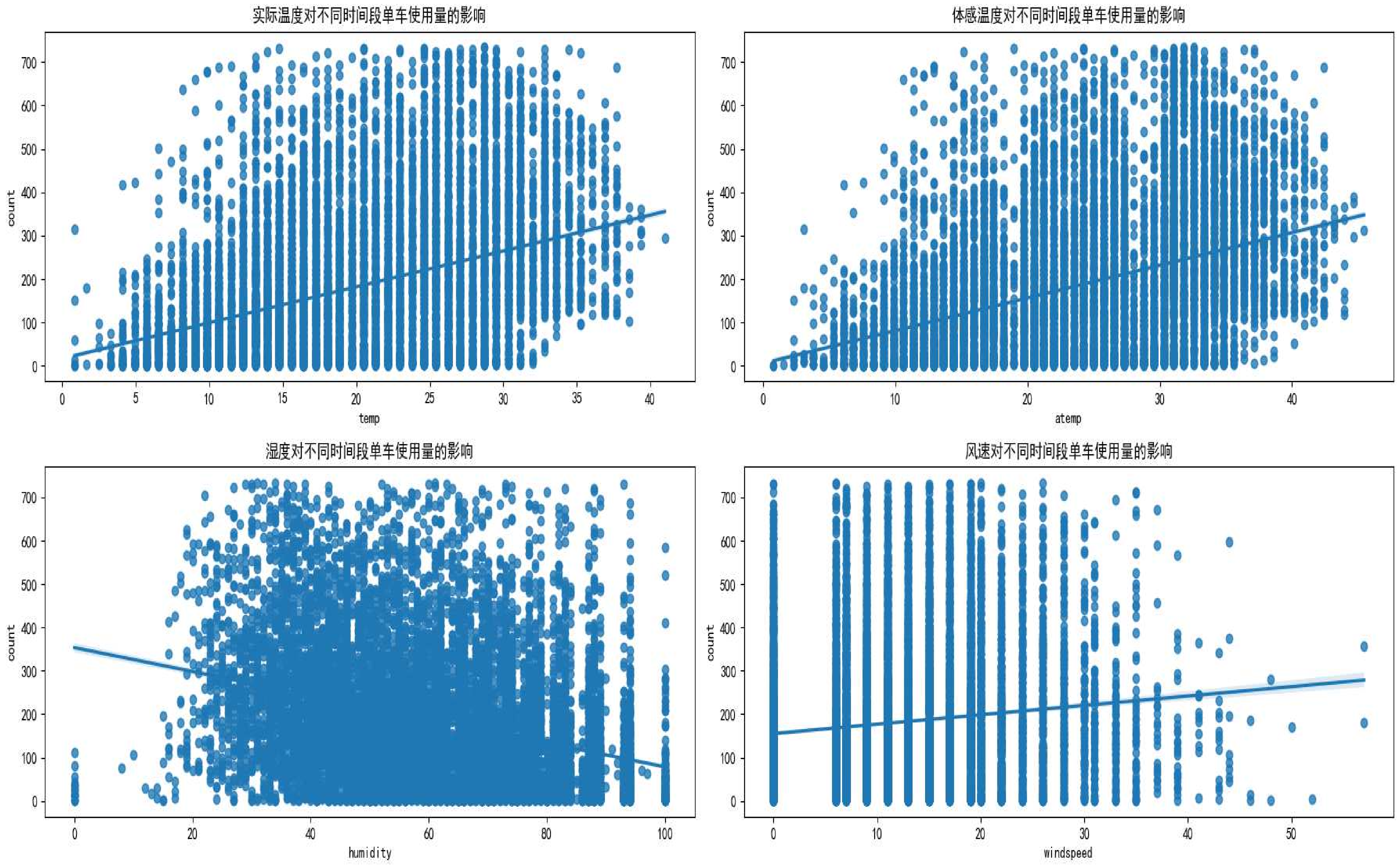


Figure 4: Scatter Plot

Data Visualization Plot

Problem Definition
Data Clean
Data Describe
Data Visualization Plot
Knowledge Discovery
Model Solution

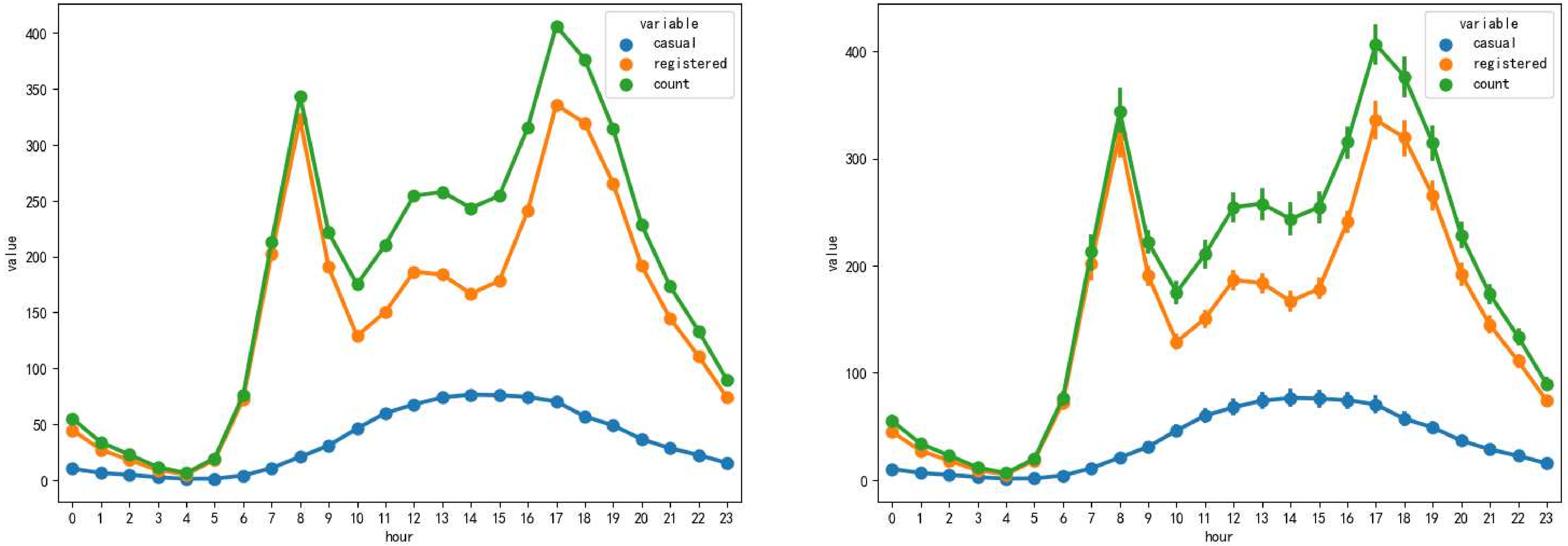


Figure 5: Line Chart

Knowledge Discovery

Variable Relationship Discovery

- Problem Definition
- Data Clean
- Knowledge Discovery
- **Variable Relationship Discovery**
- Target Variable Analysis
- Fill In Zero Values
- Model Solution

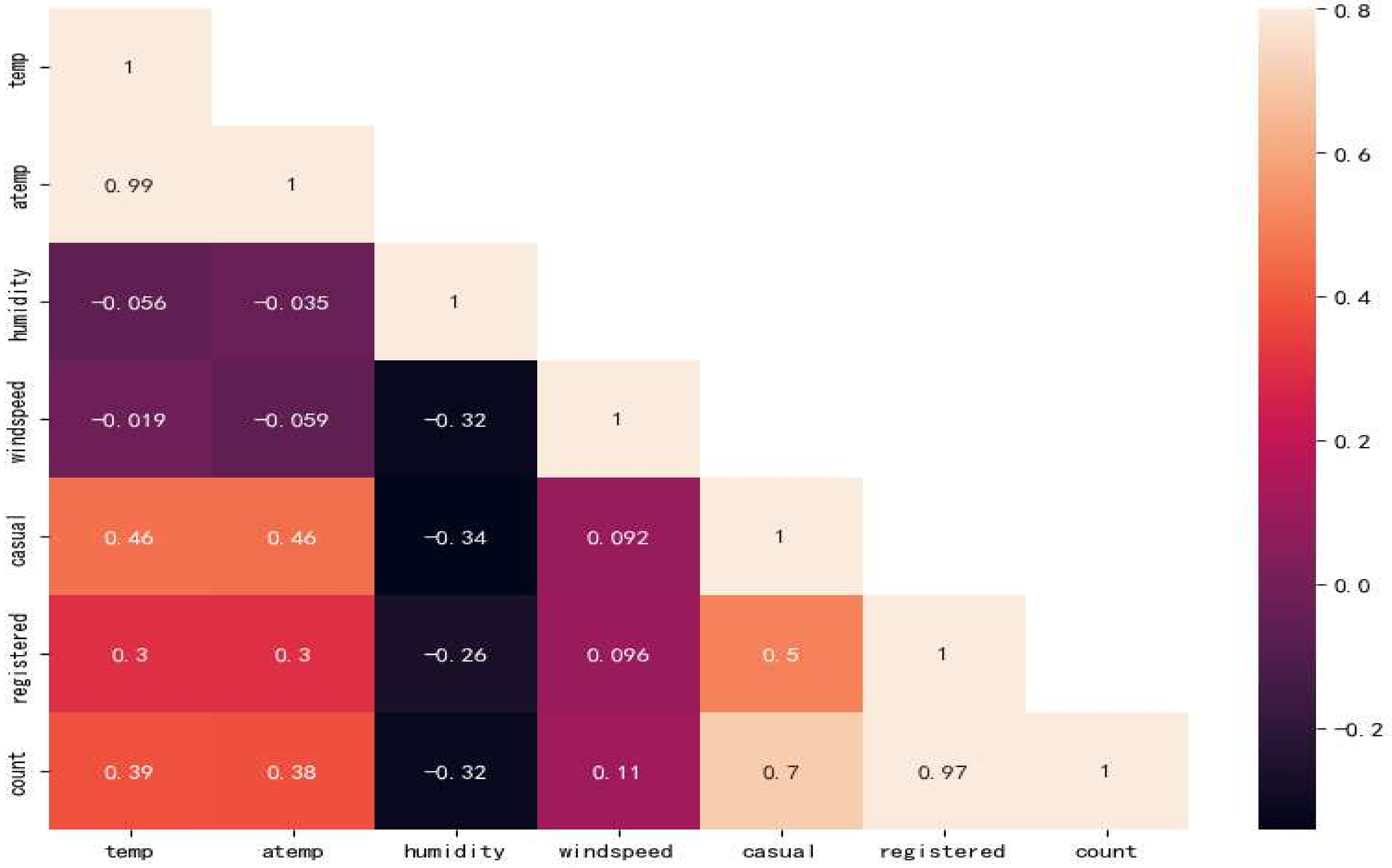


Figure 6: Hot Map

Target Variable Analysis

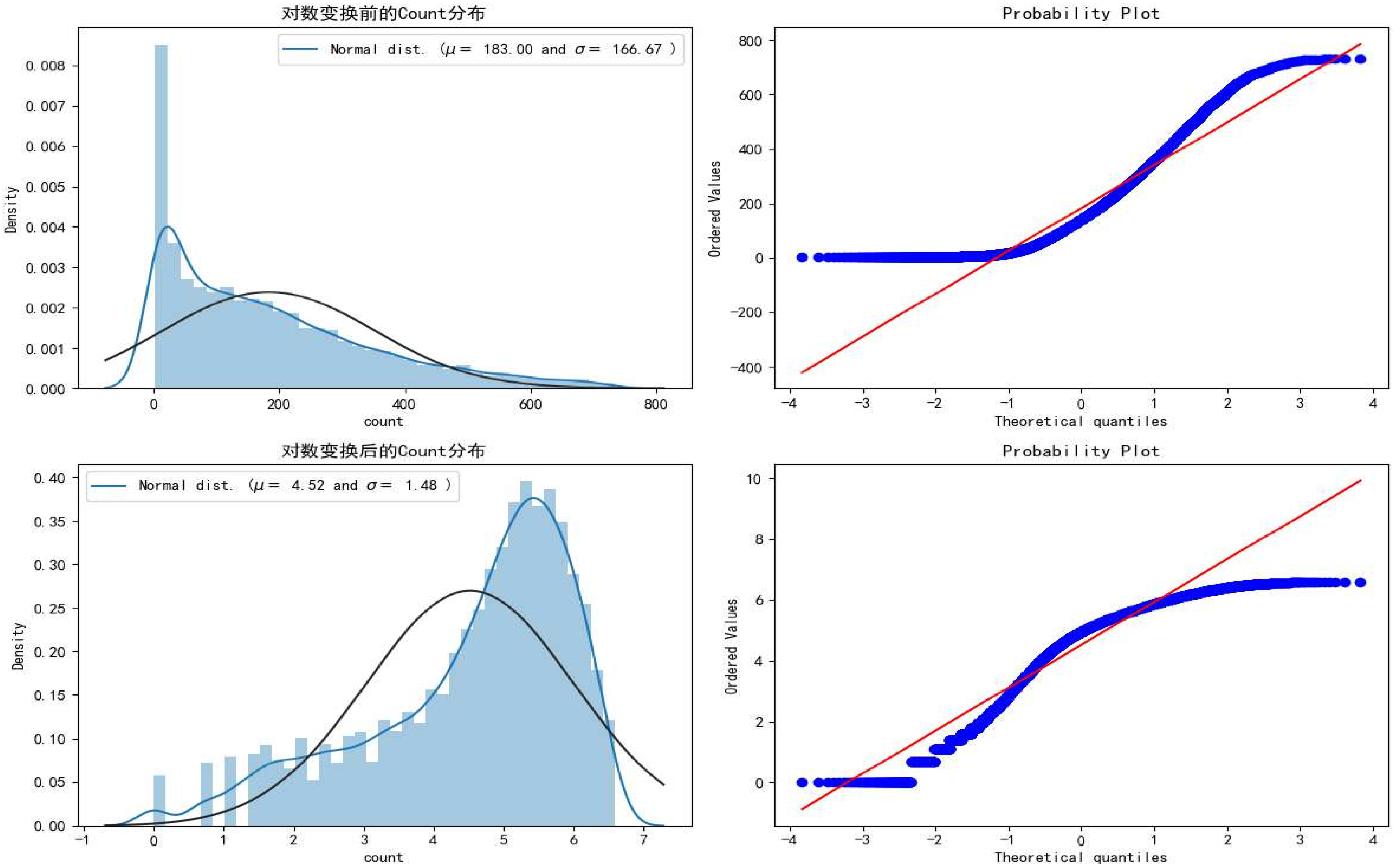


Figure 7: Variable Conversions

Fill In Zero Values

The random forest model will be used to fill the zero values in the windspeed feature.

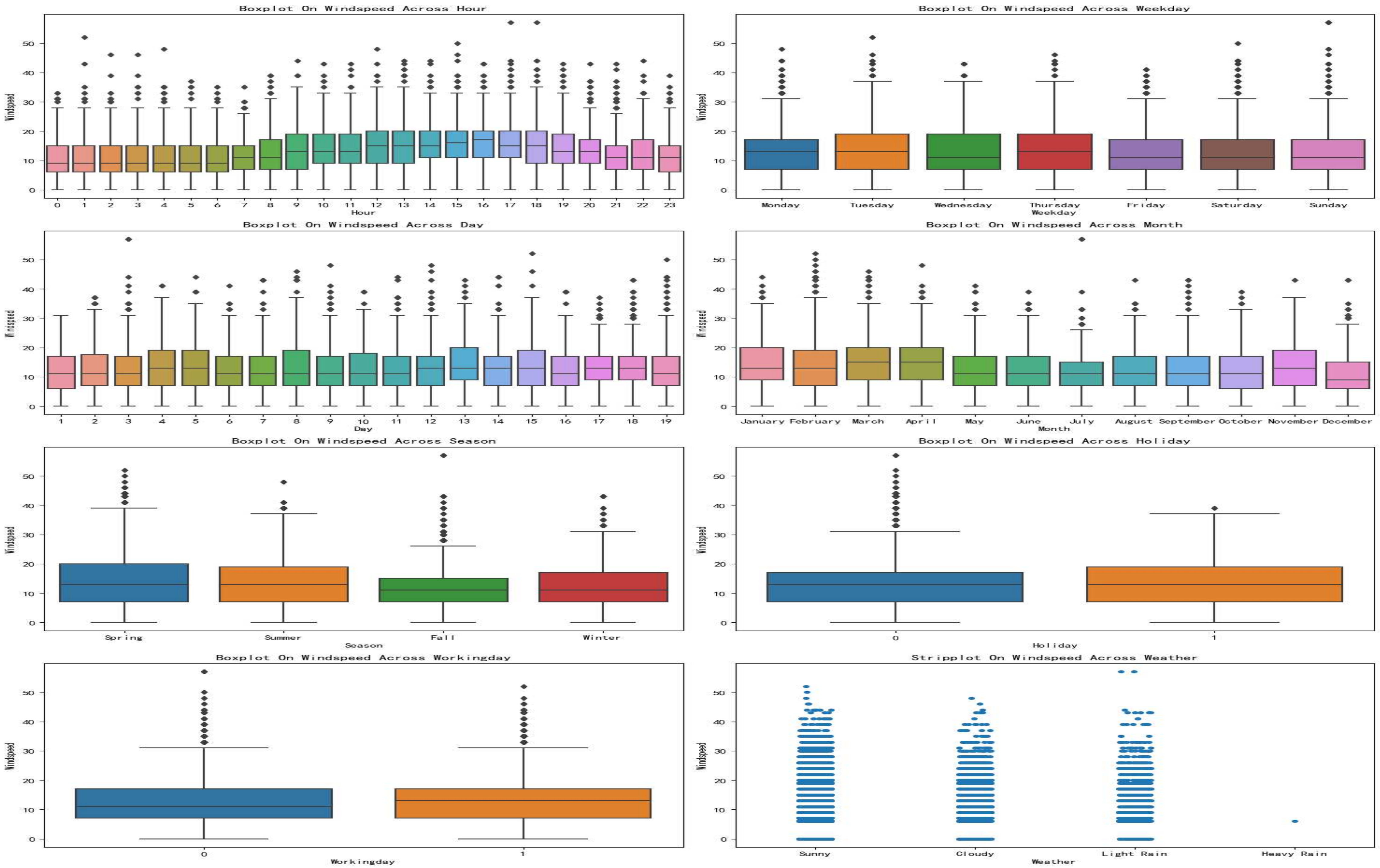
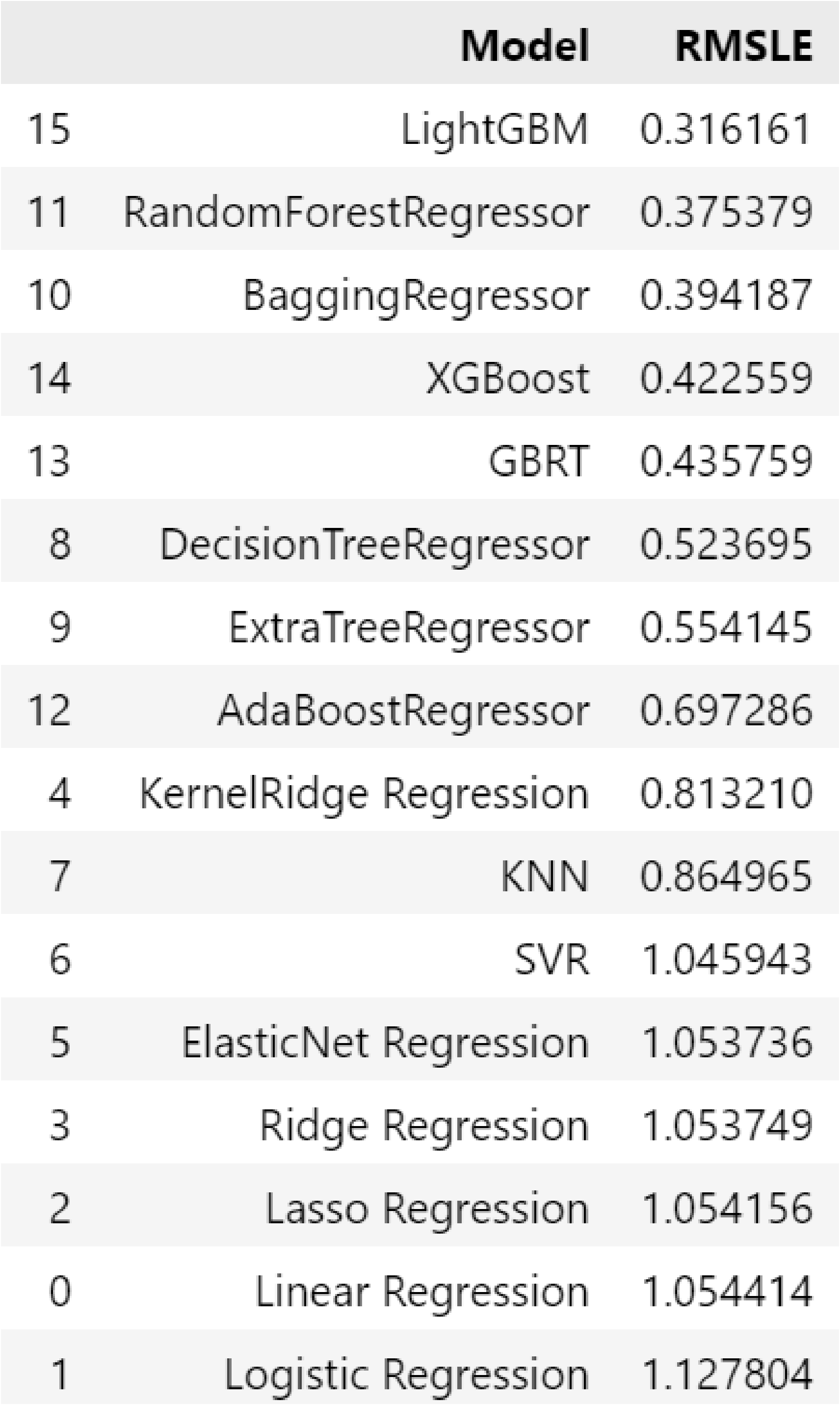


Figure 8: Relationship Between Features and Windspeed

Model Solution

Model Building

Summary of RMSLE scores for the 16 models:



	Model	RMSLE
15	LightGBM	0.316161
11	RandomForestRegressor	0.375379
10	BaggingRegressor	0.394187
14	XGBoost	0.422559
13	GBRT	0.435759
8	DecisionTreeRegressor	0.523695
9	ExtraTreeRegressor	0.554145
12	AdaBoostRegressor	0.697286
4	KernelRidge Regression	0.813210
7	KNN	0.864965
6	SVR	1.045943
5	ElasticNet Regression	1.053736
3	Ridge Regression	1.053749
2	Lasso Regression	1.054156
0	Linear Regression	1.054414
1	Logistic Regression	1.127804

Figure 9: RMSLE Scores

Model Fusion Stacking

RMSLE For Stacking: 0.3144

Final prediction result

Result

The best two models Stacking and LightGBM are weighted and the final prediction is saved.

ensemble = stacking_pred * 0.60 + lgb_pred * 0.40