Group Outlying Aspects Mining

Jincai Ma

Xi'an Shiyou University, China

Introduction

Project Background Bike-sharing is not new to us. This report mainly analyzes the data of bike-sharing in Washington, US from 2011 to 2012.

The Data Source The data comes from Kaggle https://www.kaggle.com/c/bike-sharing-demand

Project Purpose This project is mainly about the prediction of relevant data, and the description and analysis of relevant factors are presented here.

Related Field Name Interpretation

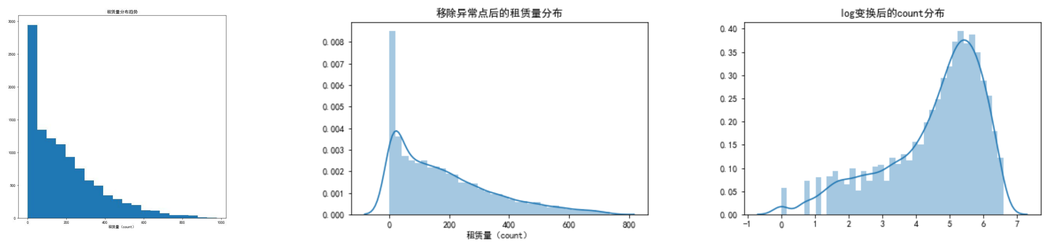
- datetime season holiday workingday weather temp atemp humidity windspeed casual registered count

Data Analysis

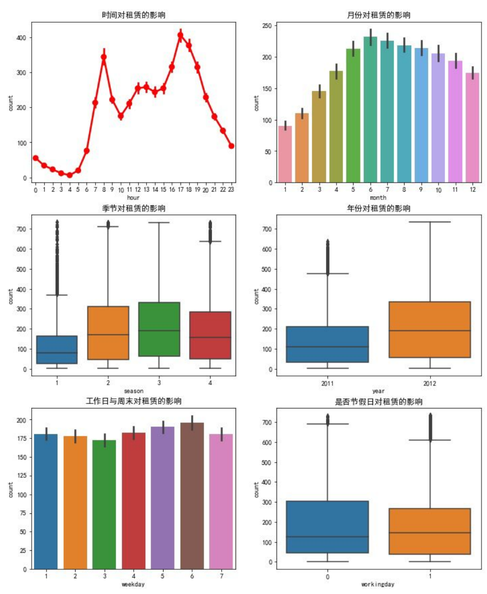
- Descriptive statistics of the data

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	coun
count	10886.000000	10886.000000	10886.000000	10886.000000	10886.00000	10886.000000	10886.000000	10886.000000	10886.000000	10886.000000	10886.00000
mean	2.506614	0.028569	0.680875	1.418427	20.23086	23.655084	61.886460	12.799395	36.021955	155.552177	191.57413
std	1.116174	0.166599	0.466159	0.633839	7.79159	8.474601	19.245033	8.164537	49.960477	151.039033	181.14445
min	1.000000	0.000000	0.000000	1.000000	0.82000	0.760000	0.000000	0.000000	0.000000	0.000000	1.00000
25%	2.000000	0.000000	0.000000	1.000000	13.94000	16.665000	47.000000	7.001500	4.000000	36.000000	42.00000
50%	3.000000	0.000000	1.000000	1.000000	20.50000	24.240000	62.000000	12.998000	17.000000	118.000000	145.00000
75%	4.000000	0.000000	1.000000	2.000000	26.24000	31.060000	77.000000	16.997900	49.000000	222.000000	284.00000
max	4.000000	1.000000	1.000000	4.000000	41.00000	45.455000	100.000000	56.996900	367.000000	886.000000	977.00000

- The standard deviation of the number of leases you have to predict at the end is very large.So let's look at the distribution by drawing it.
- Exclude data other than three standards,log of count



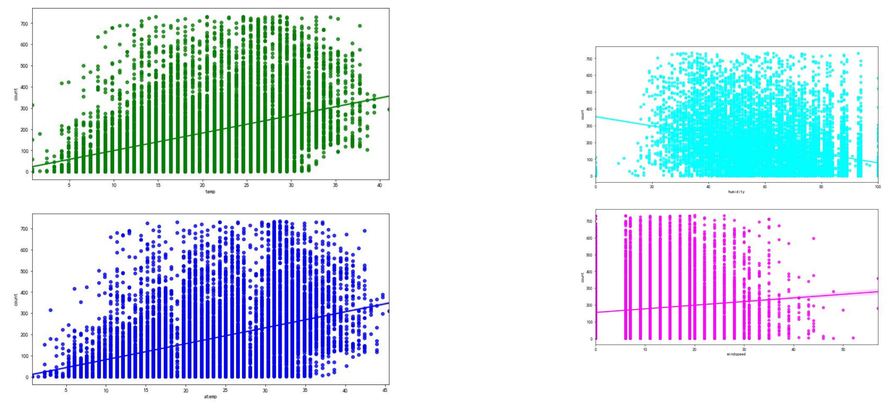
- The impact of hour,month,season,year,weekday,workingday



- The impact of weather

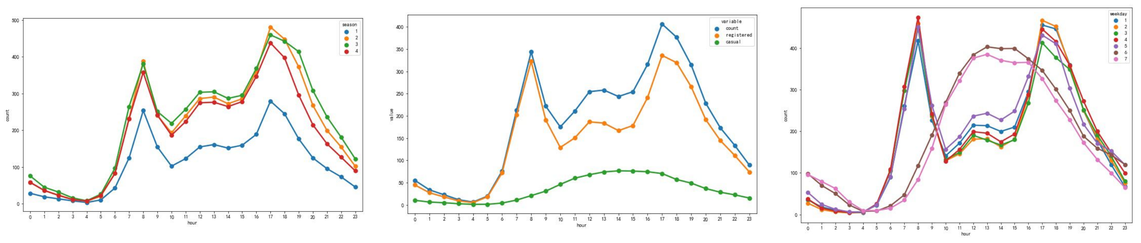
Data Analysis

- The impact of temp,atemp,humidity,windspeed



Data Analysis

- Impact of season,week,registered and non-registered users on cycling usage trends



- Draw the thermal diagram of the correlation coefficient

It can be seen that the correlation from large to small is:registered casual hour temp atemp year month season windspeed weekday holiday workingday weather humidity

Build Model

1.Separate the training set and test set.
2.Remove unwanted eigenvalues:'casual','count','datetime','registered','date','atemp','mo
3.Cross validation is used to determine the optimal parameters.
4.View the selected optimal parameters:max depth: 20, n estimators: 150
5.Apply the optimal parameters to the model, it can be obtained
Accuracy on test set : 0.6945996275605214

Conclusion

Through this Kaggle project, I practiced by myself to have a deeper underst
of data visualization and to explore the structure and rules of

logos/tulip-wordmark-eps-converted-to.pdf

The 11^{th} International Conference on Knowledge Science, Engineering and Management (KSEM 2018), 17-16/20/2020,Xi'an, China

oster