# BOX OFFICE FORECAST

JINCAI MA

Abstract. Bike-sharing refers to the bicycle sharing services provided by enterprises in campuses, subway stations, bus stations, residential areas, business districts and public service areas. It is a time-sharing model and a new green and environment-friendly sharing economy. In essence, bike-sharing is a new type of transportation rental business – bicycle rental business, which mainly relies on the carrier of (bicycle) bicycles.Can make full use of the city due to rapid economic development caused by the sluggish bicycle travel;Maximize the public road pass rate. The purpose of this project is to predict the demand for bike rental in the D.c. D.C. bike-sharing program by combining historical weather data on bike-sharing usage patterns.

## Contents



*Date*: 2020-10-15.
1991 *Mathematics Subject Classification.* Forecast use of a city bikeshare system.

## 1. Introduction

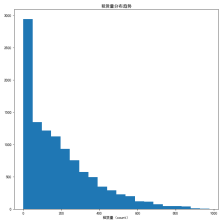
(1)Bike-sharing is not new to us. This report mainly analyzes the data of bike-sharing in Washington, US from 2011 to 2012.
(2) The data comes from Kaggle https://www.kaggle.com/c/bike-sharing-demand.
(3) This project is mainly about the prediction of relevant data, and the description and analysis of relevant factors are presented here.
(4) Related elements: datetime season holiday workingday weather temp atemp humidity windspeed casual registered count

## 2. Data Analysis

First of all, our work can be divided into the following steps:
(1) Descriptive statistics of the data

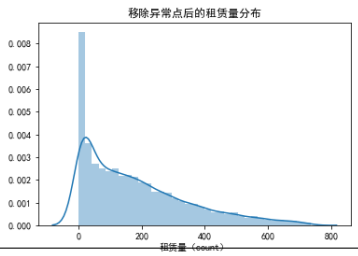
| | season | holiday | workingday | weather | temp | atemp | humidity | windspeed | casual | registered | coun |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 10886.000000 | 10886.000000 | 10886.000000 | 10886.000000 | 10886.00000 | 10886.000000 | 10886.000000 | 10886.000000 | 10886.000000 | 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 |

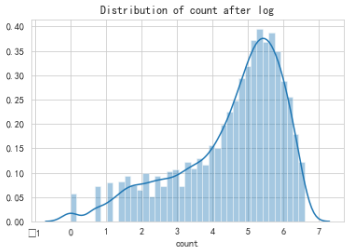
(2) The standard deviation of the number of leases you have to predict at the end is very large.So let's look at the distribution by drawing it.



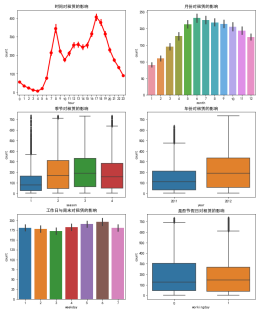
(3) Exclude data other than three standards,log of count

| | season | holiday | workingday | weather | temp | atemp | humidity | windspeed | casual | registered | cour |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ount | 10617.000000 | 10617.000000 | 10617.000000 | 10617.000000 | 10617.000000 | 10617.000000 | 10617.000000 | 10617.000000 | 10617.000000 | 10617.000000 | 10617.00000 |
| nean | 2.499294 | 0.029104 | 0.676180 | 1.421871 | 20.073588 | 23.490210 | 62.138363 | 12.779423 | 34.301309 | 142.816144 | 177.11745 |
| std | 1.121325 | 0.168107 | 0.467954 | 0.636097 | 7.779602 | 8.466483 | 19.238023 | 8.175715 | 47.716238 | 128.456579 | 158.26198 |
| min | 1.000000 | 0.000000 | 0.000000 | 1.000000 | 0.820000 | 0.760000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.00000 |
| 25% | 1.000000 | 0.000000 | 0.000000 | 1.000000 | 13.940000 | 16.665000 | 47.000000 | 7.001500 | 4.000000 | 34.000000 | 40.00000 |
| 50% | 2.000000 | 0.000000 | 1.000000 | 1.000000 | 20.500000 | 24.240000 | 62.000000 | 12.998000 | 16.000000 | 114.000000 | 139.00000 |
| 75% | 4.000000 | 0.000000 | 1.000000 | 2.000000 | 26.240000 | 31.060000 | 78.000000 | 16.997900 | 46.000000 | 212.000000 | 271.00000 |
| max | 4.000000 | 1.000000 | 1.000000 | 4.000000 | 41.000000 | 45.455000 | 100.000000 | 56.996900 | 355.000000 | 652.000000 | 663.00000 |

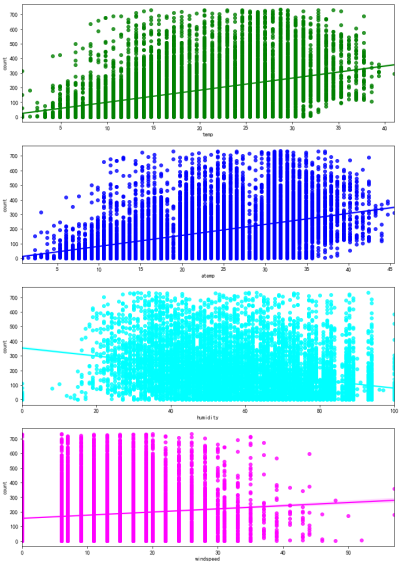




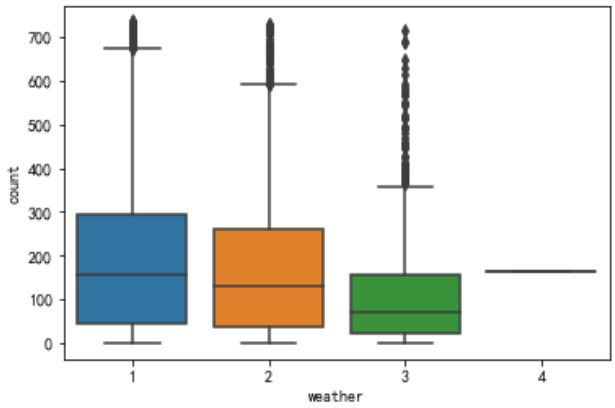

(4) The impact of hour,month,season,year,weekday,workingday



(5) The impact of temp,atemp,humidity,windspeed

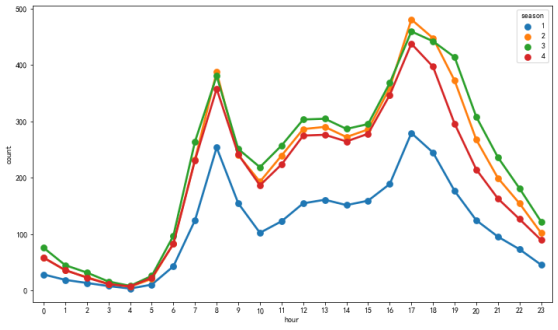


(6) The impact of weather

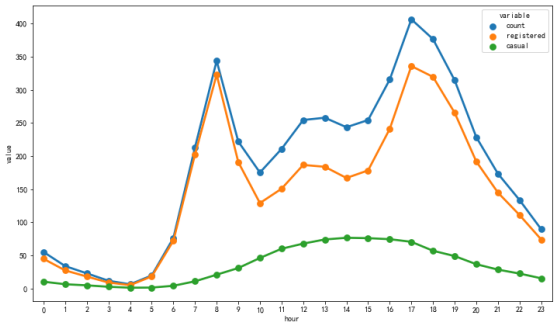



(7) Impact of season,week,registered and non-registered users on cycling usage trends

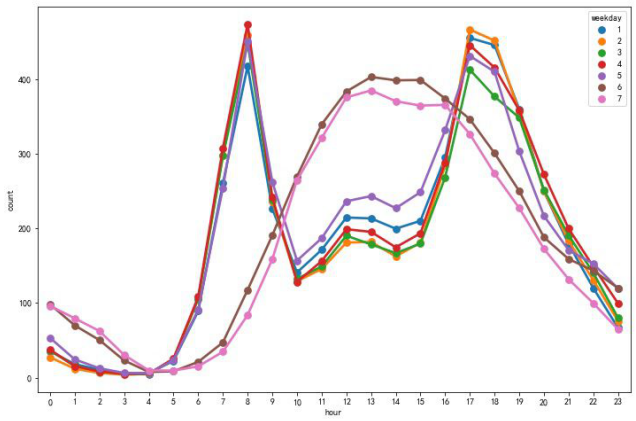
a.For different times of the day, there is a clear trend in the use of Shared bikes, with two distinct peaks, in line with people's understanding of morning peak and evening peak.The trends were the same for all four seasons, except that usage in spring was slightly lower than in the other three.



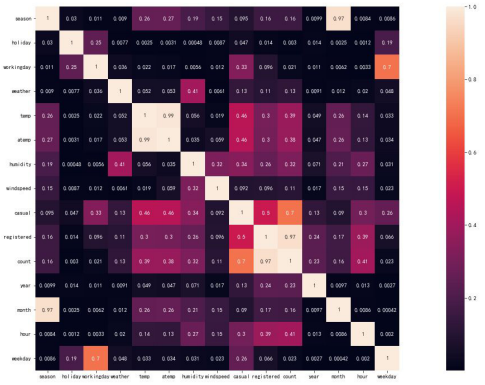

b.The usage of registered users accounts for the majority of the total usage, and the trend is consistent with the total usage trend, rather than that of registered users. The usage at different times of the day does not change much, and the trend is similar to the usage trend at weekends.




c.From Monday to Friday, there are two peak usage periods, while on weekends, the usage trend is completely different from that on weekdays. The usage trend changes from bimodal to flat unimodal, and the peak usage period is concentrated at 11-17 o 'clock.




(8) Draw the thermal diagram of the correlation coefficient



cost

## 3. Build Model

1.Separate the training set and test set.
2Remove unwanted eigenvalues:'casual','count','datetime','registered','date','atemp','month','year','season','weat
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
Recall rate on test set: 0.7379725915789399

## 4. Conclusions

Through this Kaggle project, I practiced by myself to have a deeper understanding of data visualization and to explore the structure and rules of data by means of drawing and tabulating.

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
*Email address*, A. 1: xxx@tulip.academy