**1 数据收集**  
1.1 项目说明

自行车共享系统是一种租赁自行车的方法，注册会员、租车、还车都将通过城市中的站点网络自动完成，通过这个系统人们可以根据需要从一个地方租赁一辆自行车然后骑到自己的目的地归还。

在这次比赛中，参与者需要结合历史天气数据下的使用模式，来预测D.C.华盛顿首都自行车共享项目的自行车租赁需求  
1.2 数据内容说明

比赛提供了跨越两年的每小时租赁数据，包含天气信息和日期信息，

训练集由每月前19天的数据组成，测试集是每月第二十天到当月底的数据。  
1.3 提出问题

通过测试集中的天气等特征值预测会员租赁数量，临时租赁数量和总租赁数量。  
1.4 变量说明  
 datetime（日期） - hourly date + timestamp

season（季节） - 1 = spring, 2 = summer, 3 = fall, 4 = winter

holiday（是否假日） - whether the day is considered a holiday

workingday（是否工作日） - whether the day is neither a weekend nor holiday

weather（天气等级） -

1. 清澈，少云，多云。

2. 雾+阴天，雾+碎云、雾+少云、雾

3. 小雪、小雨+雷暴+散云，小雨+云

4. 暴雨+冰雹+雷暴+雾，雪+雾

temp（温度） - temperature in Celsius

atemp（体感温度） - "feels like" temperature in Celsius

humidity（相对湿度） - relative humidity

windspeed（风速） - wind speed

casual（临时租赁数量） - number of non-registered user rentals initiated

registered（会员租赁数量） - number of registered user rentals initiated

count（总租赁数量） - number of total rentals  
**2 数据清洗**  
2.1 检查缺失值

#观察训练集数据描述统计

**表1** 训练集数据描述统计结果

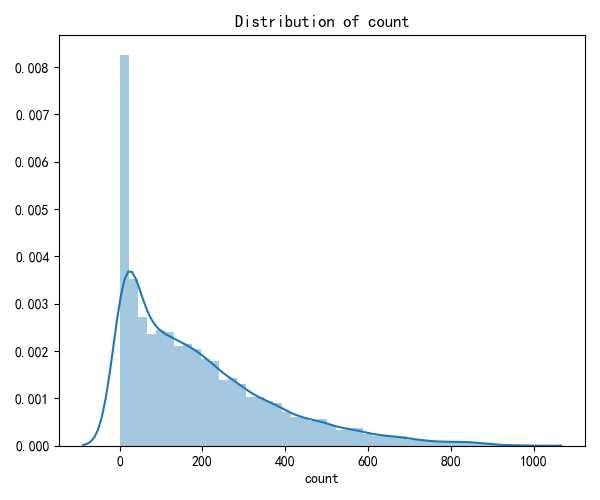
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| season | 10886 | 183 | 166.68 | 1 | 41 | 141 | 276 | 734 |
| holiday | 10886 | 0.03 | 0.17 | 0 | 0 | 0 | 0 | 1 |
| workingday | 10886 | 0.68 | 0.47 | 0 | 0 | 1 | 1 | 1 |
| weather | 10886 | 1.42 | 0.63 | 1 | 1 | 1 | 2 | 4 |
| temp | 10886 | 20.23 | 7.79 | 0.82 | 13.94 | 20.5 | 26.24 | 41 |
| atemp | 10886 | 23.66 | 8.47 | 0.76 | 16.67 | 24.24 | 31.06 | 45.45 |
| humidity | 10886 | 61.89 | 19.25 | 0 | 47 | 62 | 77 | 100 |
| windspeed | 10886 | 13 | 8.16 | 0 | 7 | 13 | 17 | 59.6 |
| casual | 10886 | 36.03 | 49.96 | 0 | 4 | 17 | 49 | 367 |
| registered | 10886 | 155.55 | 151.04 | 0 | 36 | 118 | 222 | 886 |
| count | 10886 | 191.57 | 181.14 | 1 | 42 | 145 | 284 | 997 |

从表可以观察得出，数值型数据不存在缺失值，但count、casual、registered属性差别较大，存在异常值。

2.2 检查并处理异常值

2.1.1 对于count属性进一步异常值分析和处理

1）密度分布情况



**图1 count属性密度分布图**

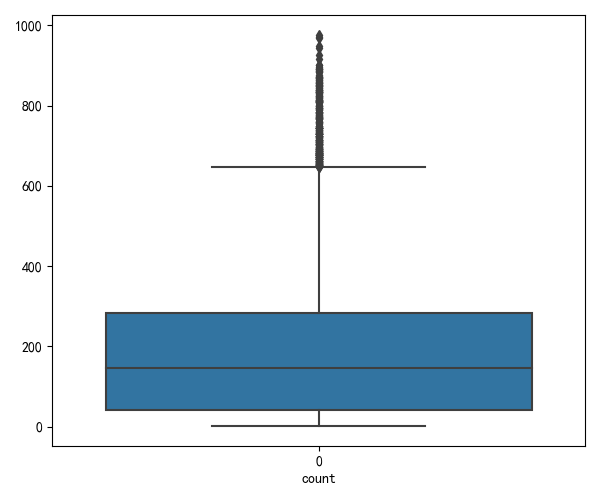


图2 count属性箱型图异常值检测结果图

注：在箱线图中，箱子的中间有一条线，代表了数据的中位数。箱子的上下底，分别是数据的上四分位数（Q3）和下四分位数（Q1），这意味着箱体包含了50%的数据。因此，****箱子的高度在一定程度上反映了数据的波动程度****。上下边缘则代表了该组数据的最大值和最小值。有时候箱子外部会有一些点，可以理解为数据中的“****异常值****”

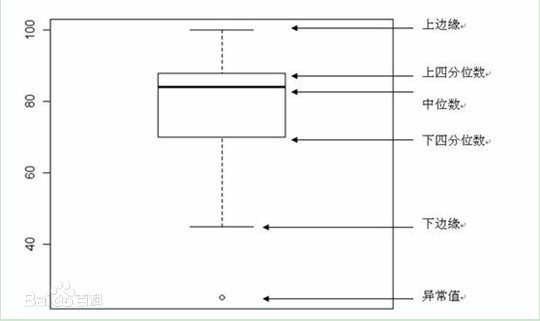


图4箱型图

从图和表可知count:均值191，标准差181，50%分位数是145，75%分位数是284，最大值977，说明右侧存在长尾。去除掉异常值，并取log处理，观察结果。

1. 去除异常值：将大于μ＋3σ的数据值作为异常值（拉依达准则）

去除3个标准差之后的count属性统计描述

**表2 count属性去除3个标准差之后的统计描述结果**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| count | 10739 | 183 | 166.68 | 1 | 41 | 141 | 276 | 734 |

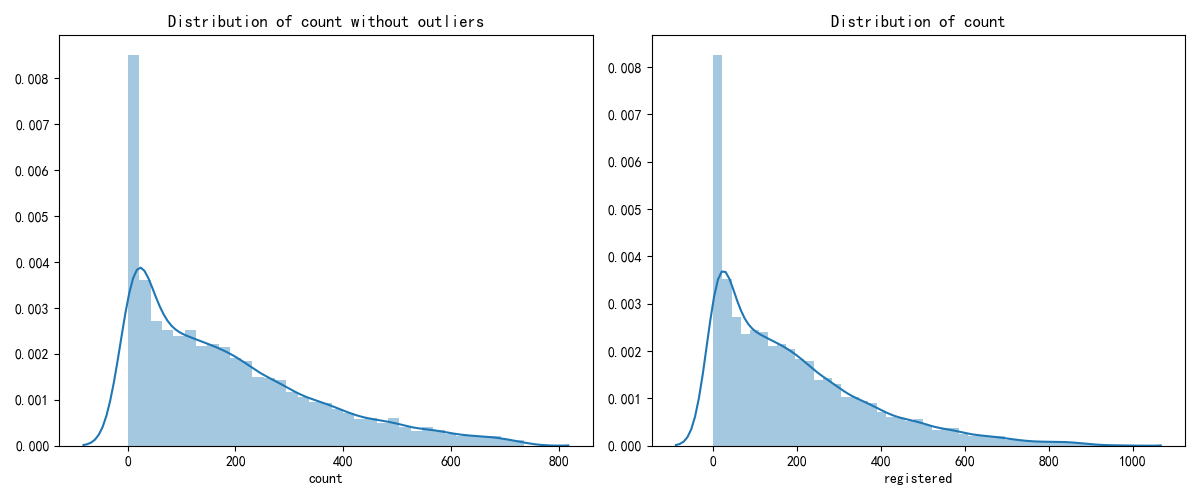


图2 **count属性去除3个标准差之后结果图**

从图中可以看到数据波动依然很大，而我们希望波动相对稳定，否则容易产生过拟合，所以希望对数据进行log变换，使得数据相对稳定。结果如图3所示。

1. 去异常值后的log变换

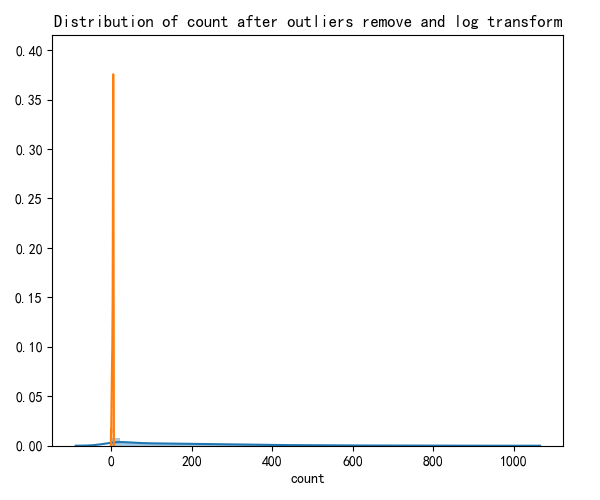


图3 log变换结果

对于casual、registered属性进行同理操作。  
**3 分析数据**  
3.1 整体观察

接下来对其余的数值型数据进行处理，由于其他数据同时包含在两个数据集中，为方便数据处理先将两个数据集合并。

对于时间数据进行详细提取

3.2 逐项展示  
  
**4. 选择特征值**  
  
**5. 选择模型、训练模型**  
  
**6. 预测测试集数据**