**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。

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

数据中包含10886行\*12列出行记录数据。数据包括时间、季节、节假日、工作日、天气、实际温度、感受温度、湿度、风速、未注册用户租借数量、注册用户租借数量、总租借数量 。由训练集和测试集组成，涵括了2011-2012年每小时的骑行数据。训练集由每个月的前19天组成，测试集是第20天到月末。计划预测测试集中每个小时内租用自行车的总数。

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时间 - 年月日小时

season季节 - 1 = spring春天, 2 = summer夏天, 3 = fall秋天, 4 = winter冬天

holiday节假日 - 0：否，1：是

workingday工作日 - 该天既不是周末也不是假日（0：否，1：是）

weather天气 - 1:晴天，2:阴天 ，3:小雨或小雪 ，4:恶劣天气（大雨、冰雹、暴风雨或者大雪）

temp实际温度

atemp体感温度

humidity湿度

windspeed风速

casual未注册用户租借数量

registered注册用户租借数量

count总租借数量  
**2 数据预处理**  
 2.1 检查缺失值

#观察训练集数据和各个属性的分布情况

**表1** 训练集数据描述统计结果

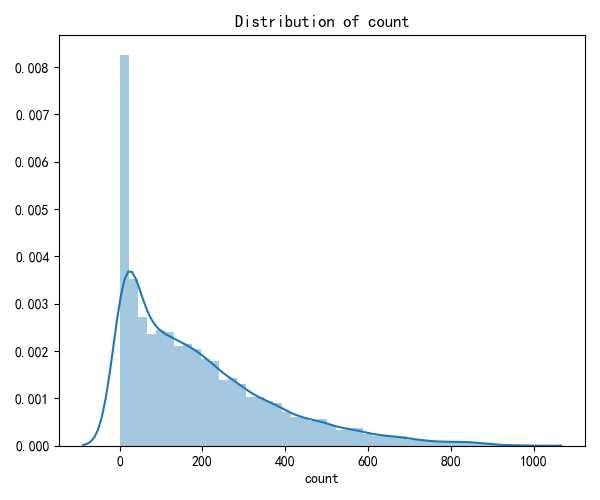
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| season | 10886 | 183 | 166.68 | 1 | 41 | 141 | 276 | 734 |
| holiday | 10886 | 0.03 | 0.17 | 0 | 0 | 0 | 0 | 1 |
| workingday | 10886 | 0.68 | 0.47 | 0 | 0 | 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，同时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**

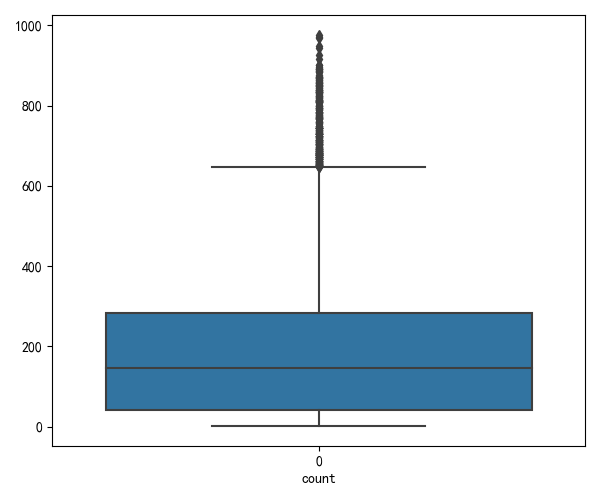


图2 count属性箱型图异常值检测结果图

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****”

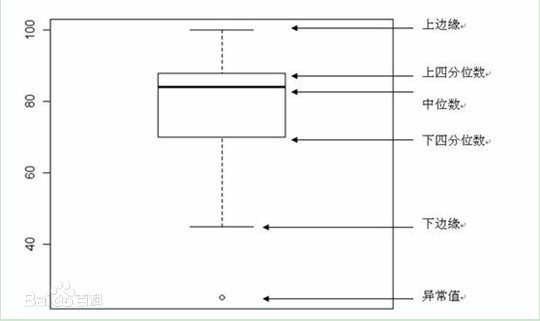


图4箱型图

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 |

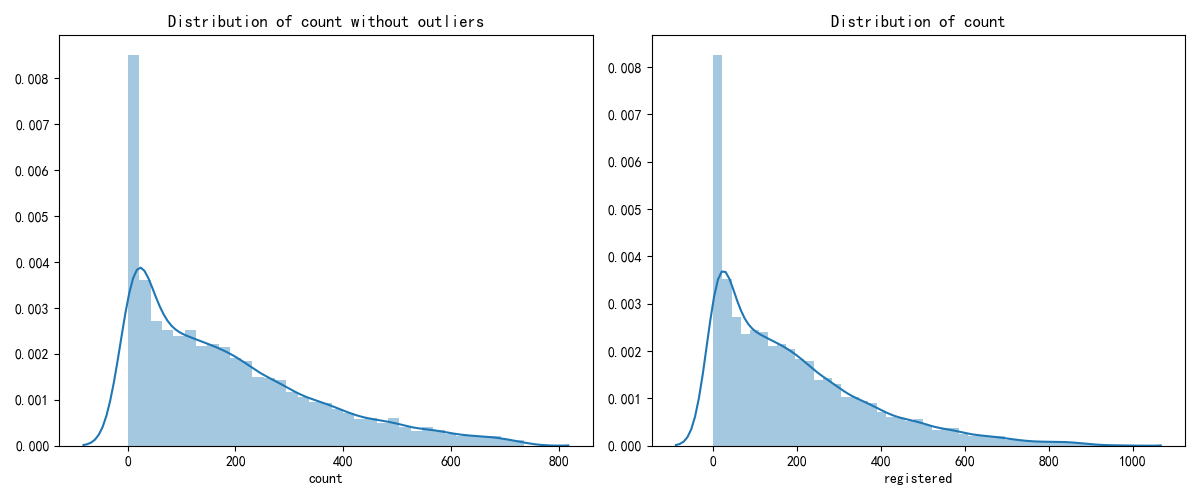


图2 **count属性去除3个标准差之后结果图**

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

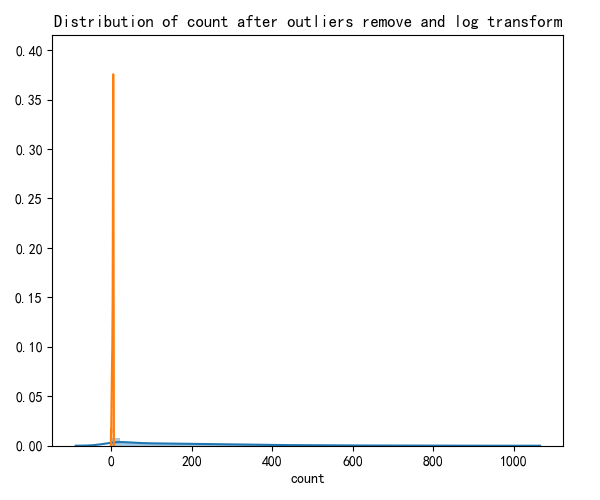


图3 log变换结果

对于casual、registered属性进行同理操作。

合并数据集。  
**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。

数值型数据humidity，windspeed，casual，registered的分布如下图所示。

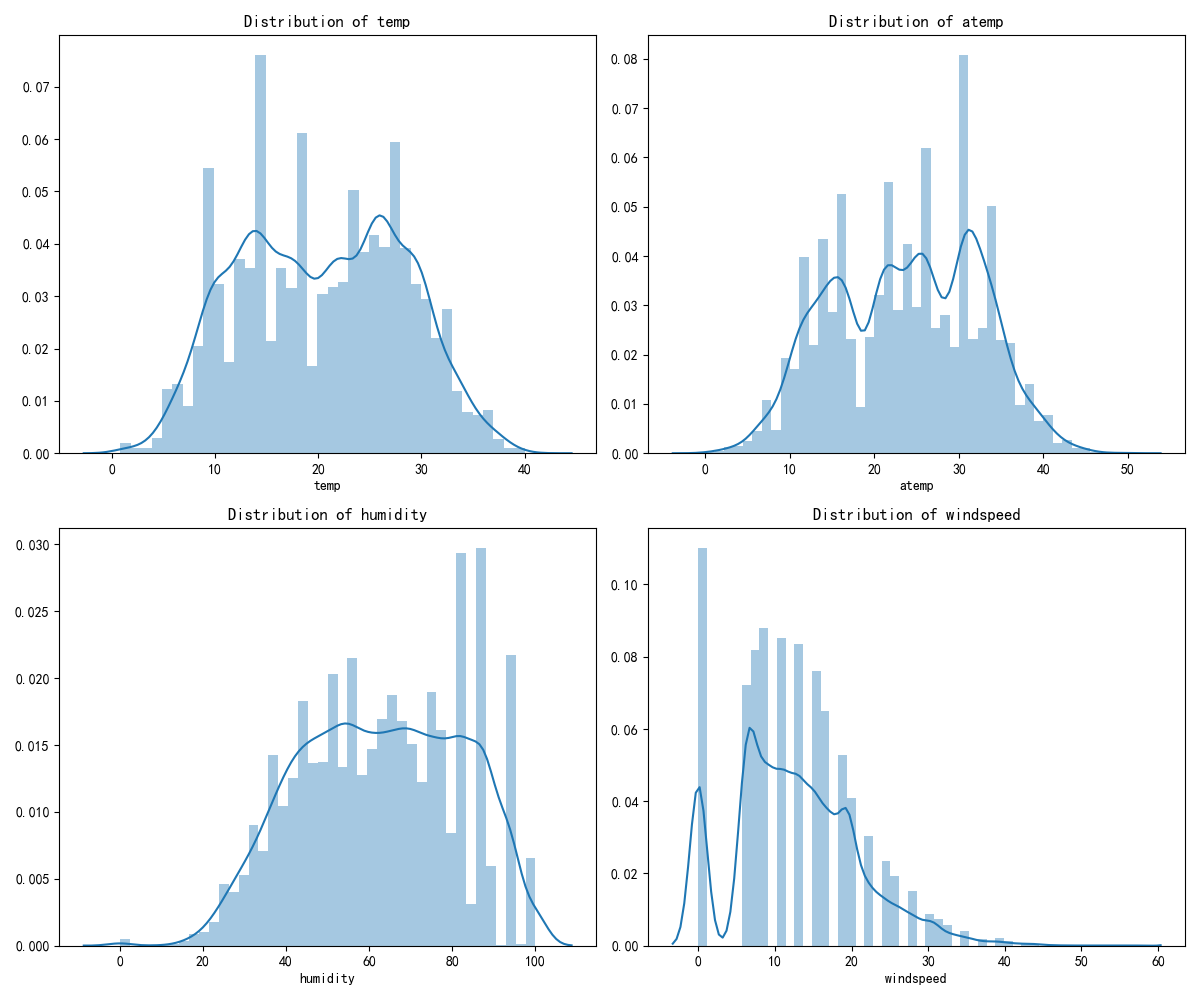


图4四个属性的分布图

通过这个分布可以发现一些问题，比如风速为什么0的数据很多，而观察统计描述发现空缺值在1--6之间，似乎可以推测，数据本身或许是有缺失值的，但是用0来填充了，但这些风速为0的数据会对预测产生干扰，使用随机森林根据相同的年份，月份，季节，温度，湿度等几个特征来填充一下风速的缺失值。

非零值的描述

表 2 风速的描述结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| windspeed | 15068 | 14.55 | 7.09 | 6 | 9 | 13 | 19 | 56 |

填充好了以后再画图观察一下这四个特征值的密度分布。

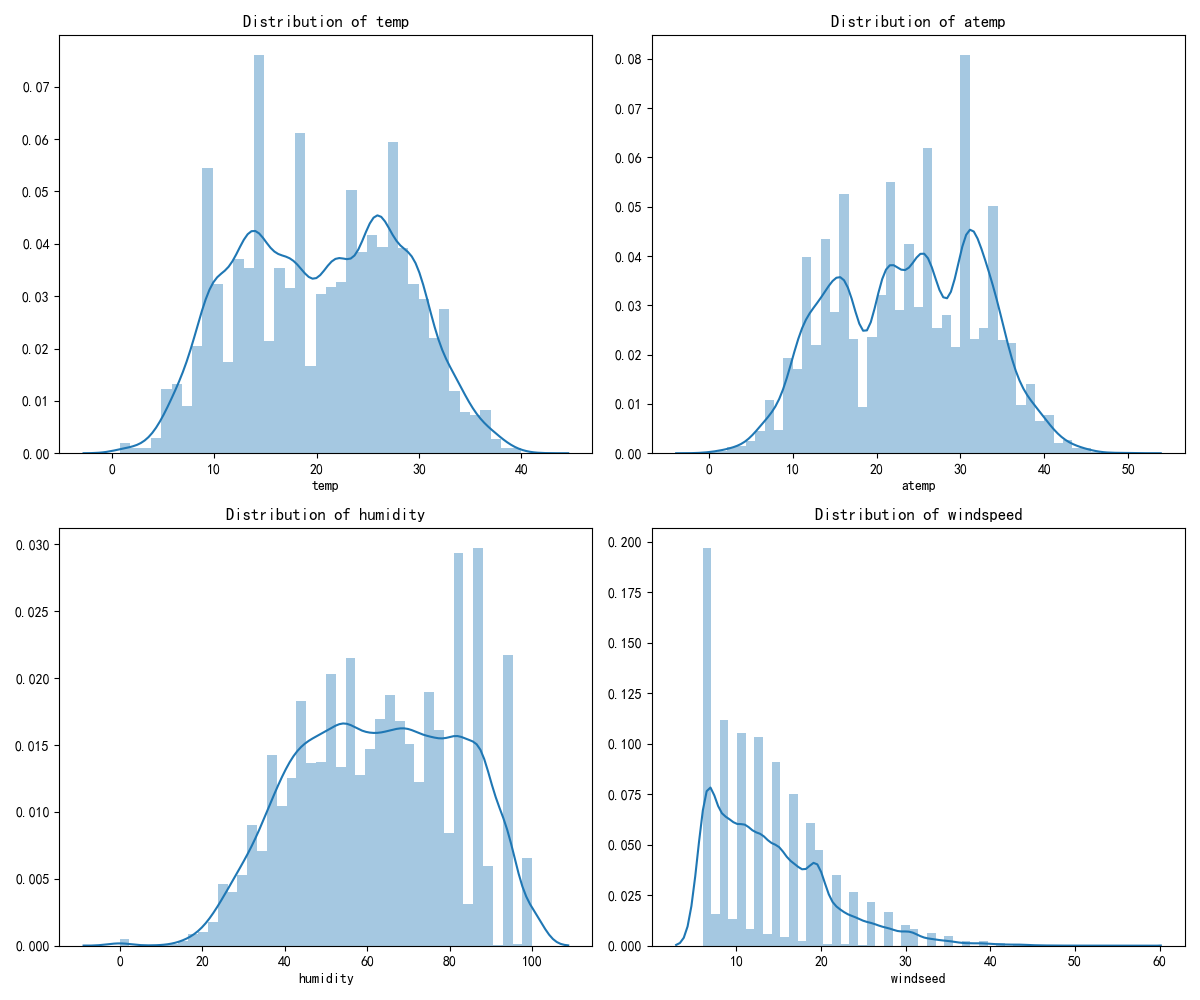
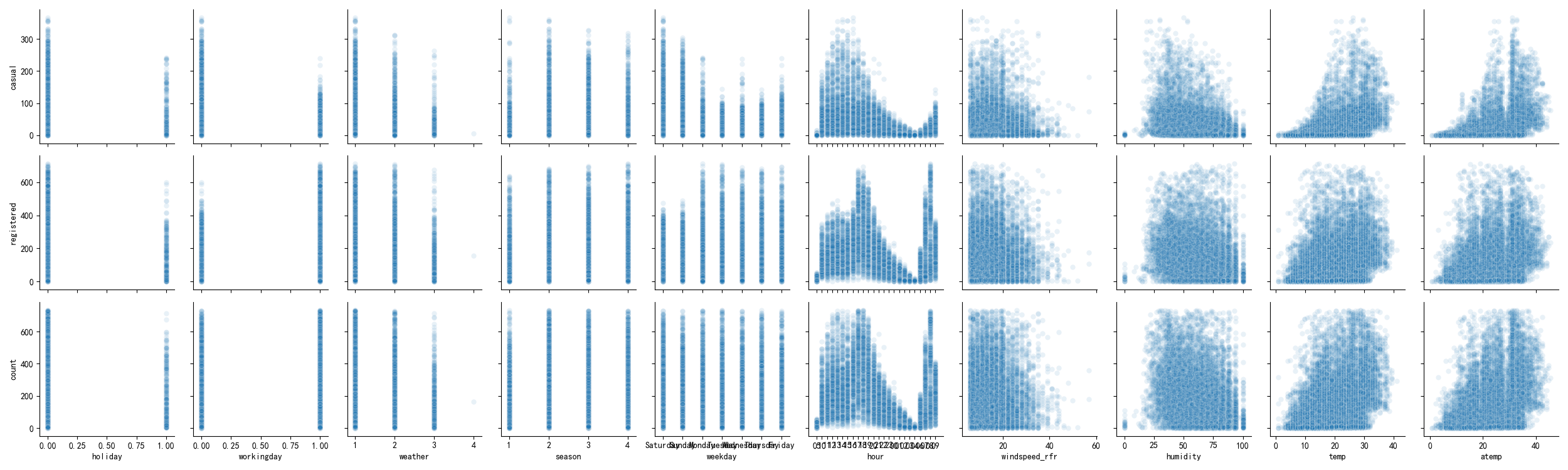


图5 填充后的分布

3.2 逐项展示

问题是希望预测每小时总租赁额，首先整体看一下租赁额相关的三个值和其他特征值的关系。



大致可以看出  
1.会员在工作日出行多，节假日出行少，临时用户则相反；  
2.一季度出行人数总体偏少；  
3.租赁数量随天气等级上升而减少；  
4.小时数对租赁情况影响明显，会员呈现两个高峰，非会员呈现一个正态分布；

5.租赁数量随风速增大而减少；  
6.温度、湿度对非会员影响比较大，对会员影响较小

查看各个特征与每小时租车总量（count）的相关性

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