### 3.1 数据来源

### 3.2 预测方法

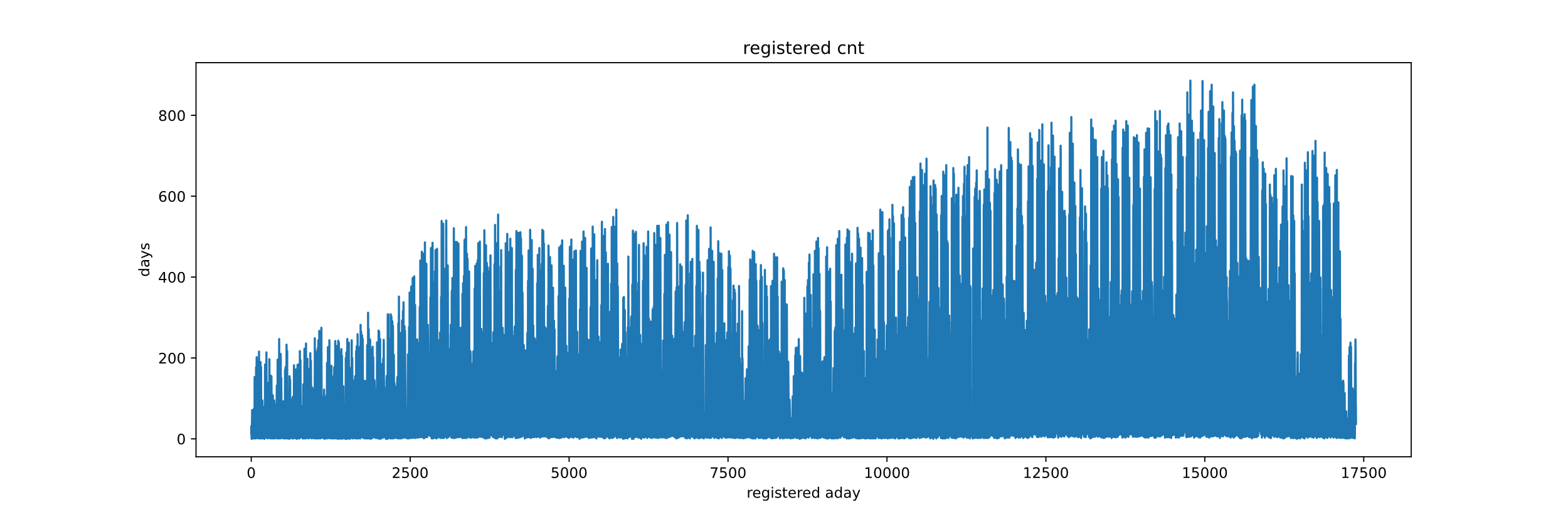
### 3.3 特征分析

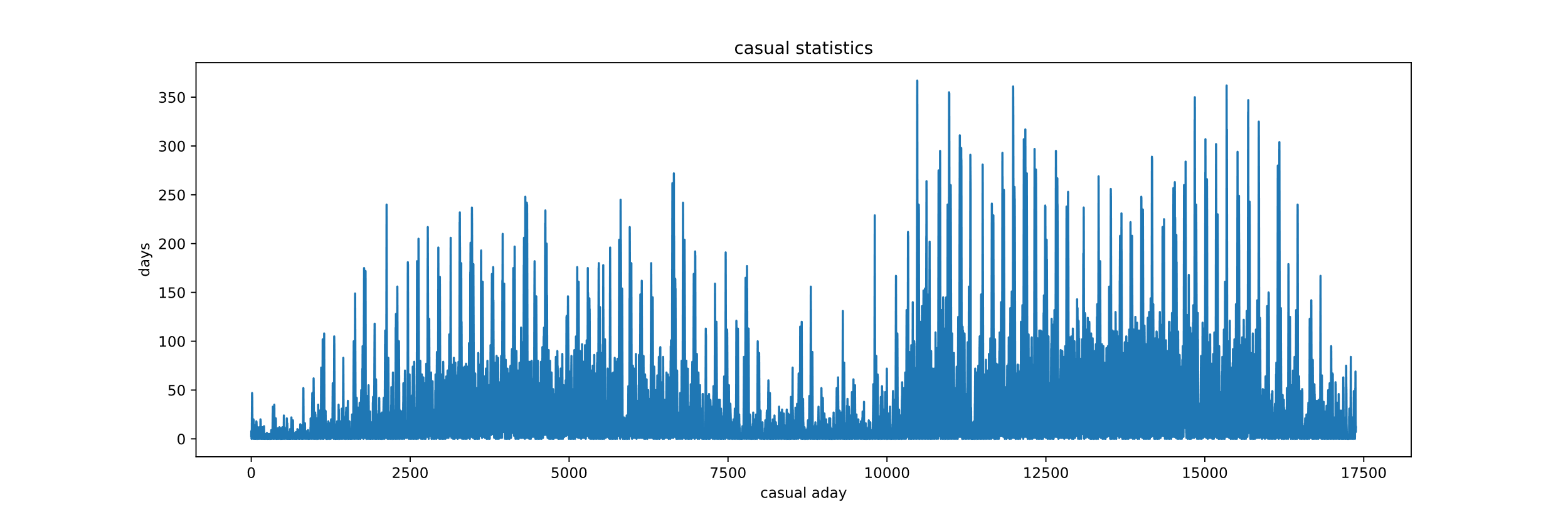
该数据集共有16个标签，但是不是所有的标签都适用于当做特征值。特征筛选的方法主要包括：Filter（过滤法）、Wrapper（封装法）、Embedded（嵌入法）。

|  |  |  |
| --- | --- | --- |
| 特征值 | 是否有空 | 类型 |
| season | 否 | int64 |
| mnth | 否 | int64 |
| hr | 否 | int64 |
| holiday | 否 | int64 |
| weekday | 否 | int64 |
| workingday | 否 | int64 |
| weathersit | 否 | int64 |
| temp | 否 | float64 |
| atemp | 否 | float64 |
| hum | 否 | float64 |
| windspeed | 否 | float64 |

其中，除了以上11列作为明显的特征，还有类似casual和registered这两个特征并未分析。

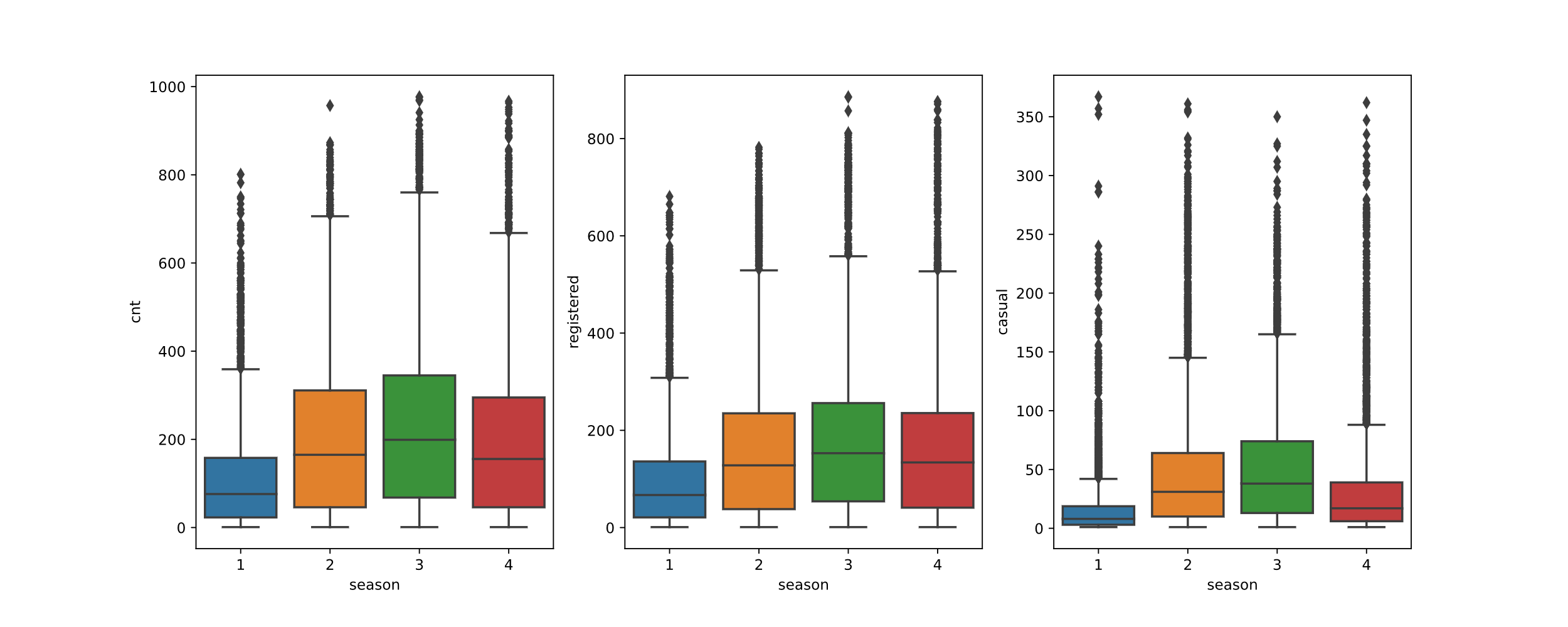
下方两图表示了注册会员与未注册会员使用共享自行车的情况。



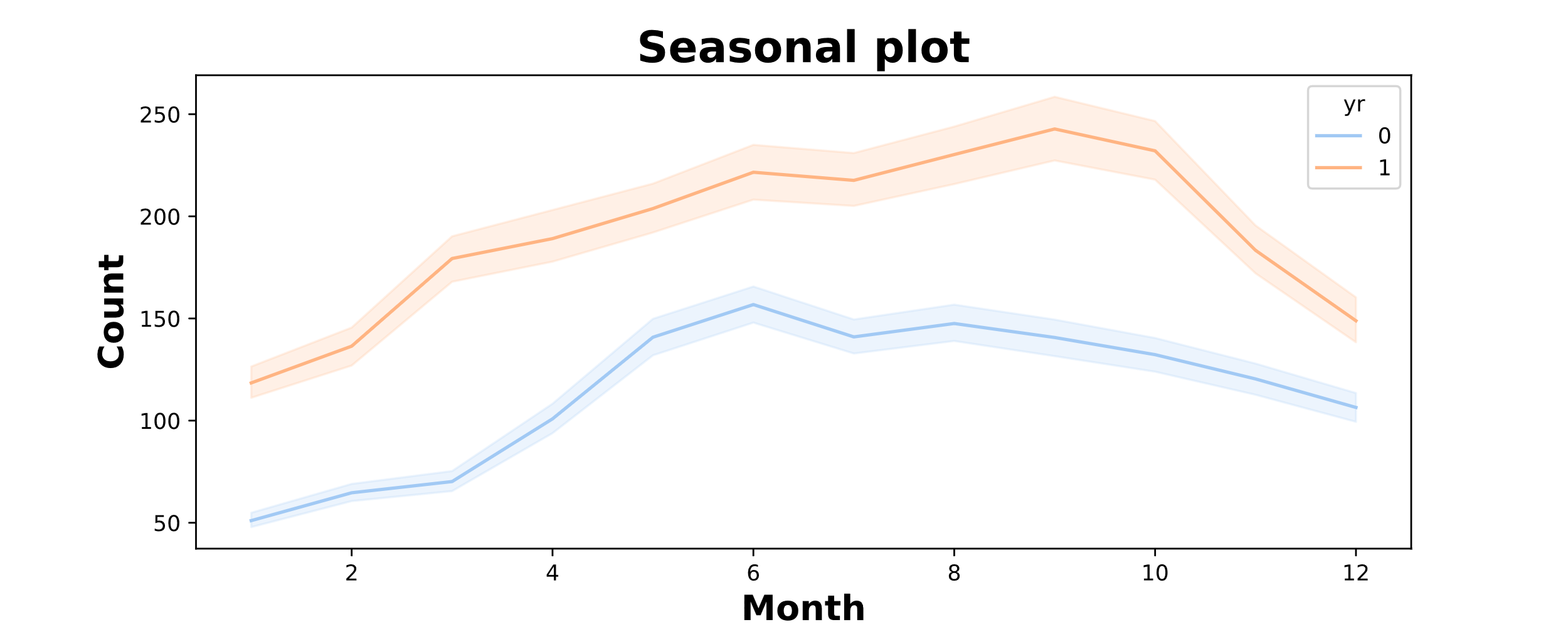


总的来看，注册会员使用人数随着时间增长呈现上升的趋势。

接下来看有关三个数量特征与季节之间的关系，不难看出，共享单车得到使用次数集中在夏季和秋季。但是在寒冷的冬季仍有不少的人们会选择它。



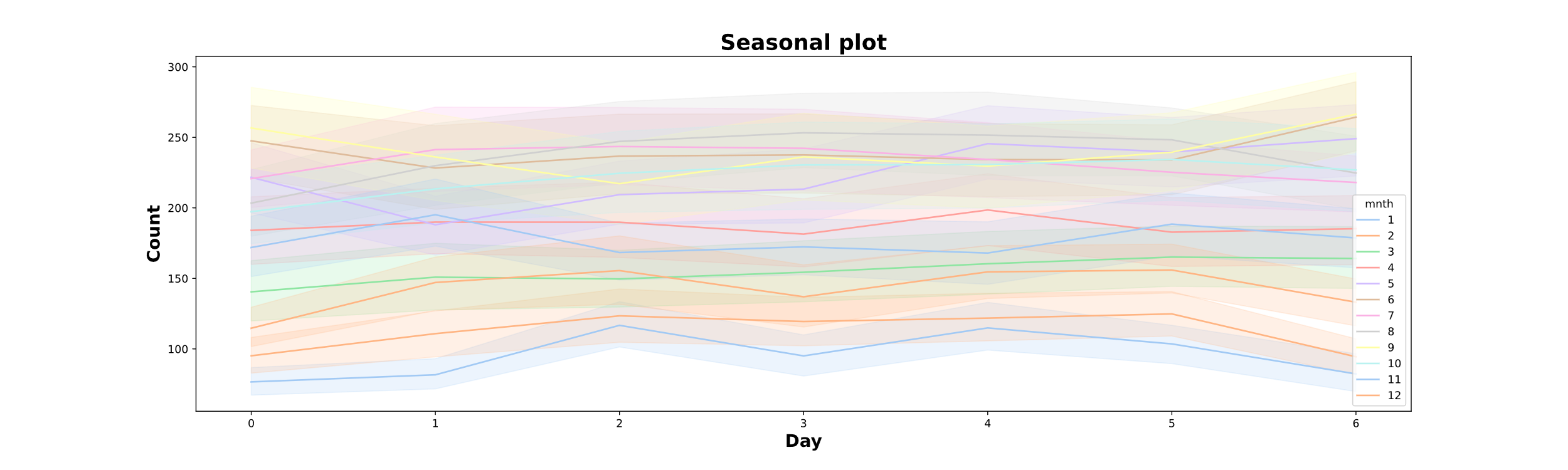
季节是影响共享单车使用情况很重要的一个因素，除此之外，共享单车业务的发展是十分迅速的，从下图中不难看出：

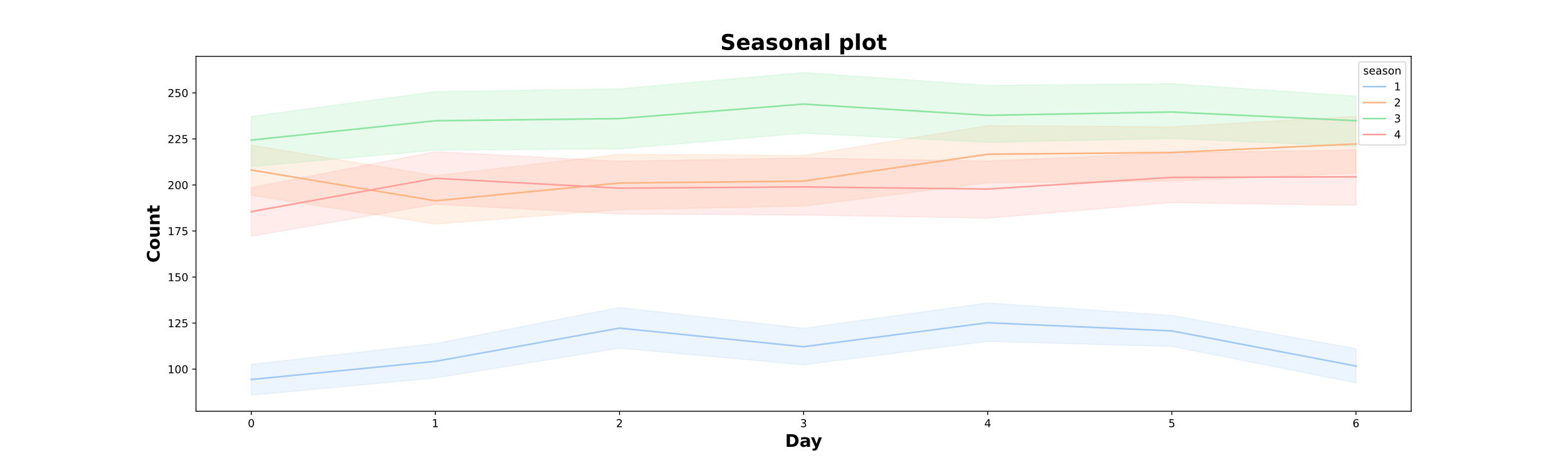


(yr= 0代表2011年，yr = 1代表2012年)

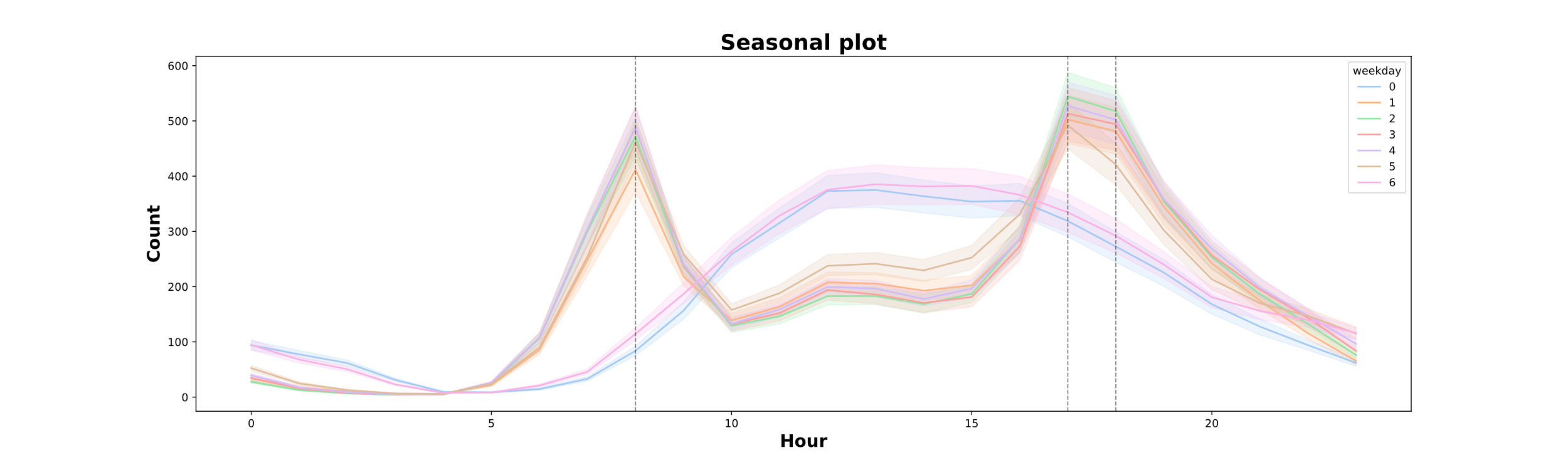
除了2011年末和2012年末的差距相对较小之外，其他时间2012年的共享单车使用数量有着明显的优势，在3月份，2012年共享单车使用量较2011年翻了一番还要多。

下图展示了12个月和一个星期七天内数量的变化情况。不难从图中观测出来，在8月份的星期三，使用共享单车人数到达了顶峰。



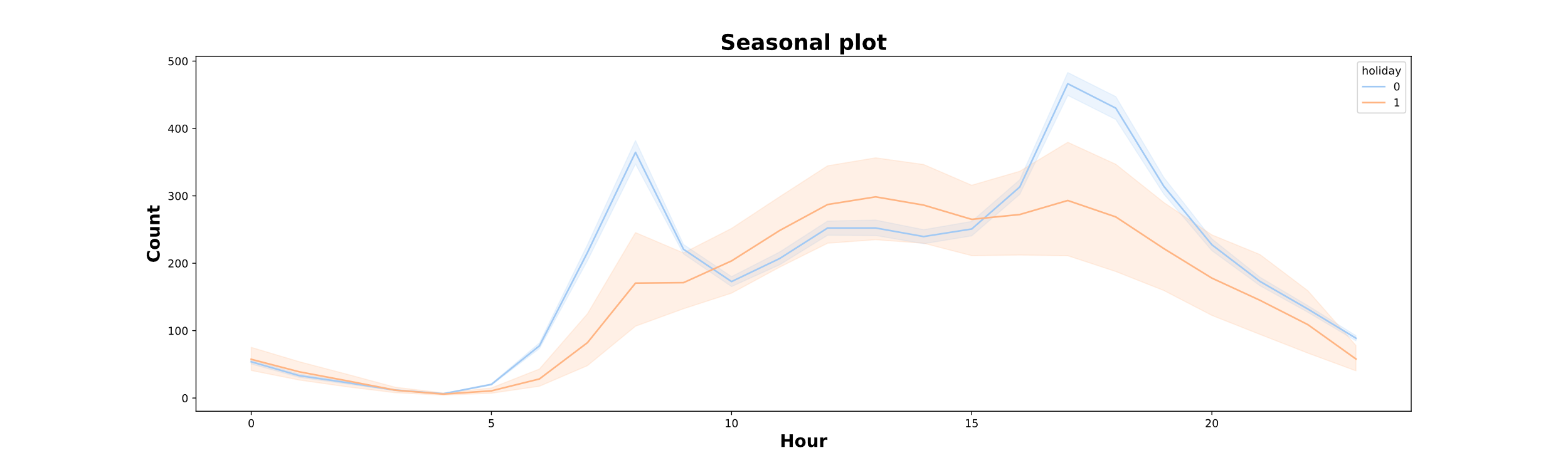


比较完了星期与季节月份之间的关系，一天之中使用共享自行车的时间也尤为关键。比如，早上8:00之前普遍人们的休息时间。使用量很可能与日间有较大的差异。



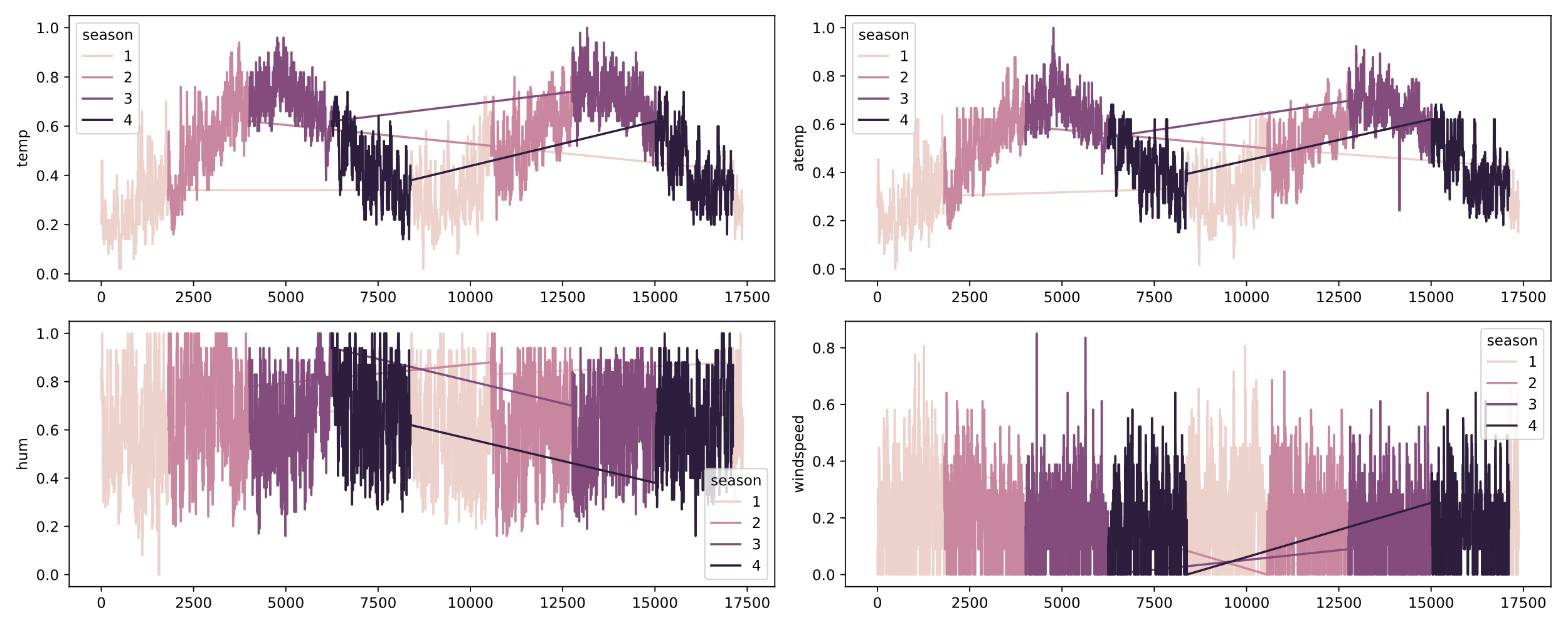
显然，数据的呈现符合我们的猜想，共享自行车的使用量与人们活跃时间有很大的关系。

虽然休息时间与工作时间会有很大的差异，但事实上有些日子可能会缩小这样的差距，例如假日(holiday)。在假日的自行车使用量中，没有普通工作日那么尖锐的波峰了。

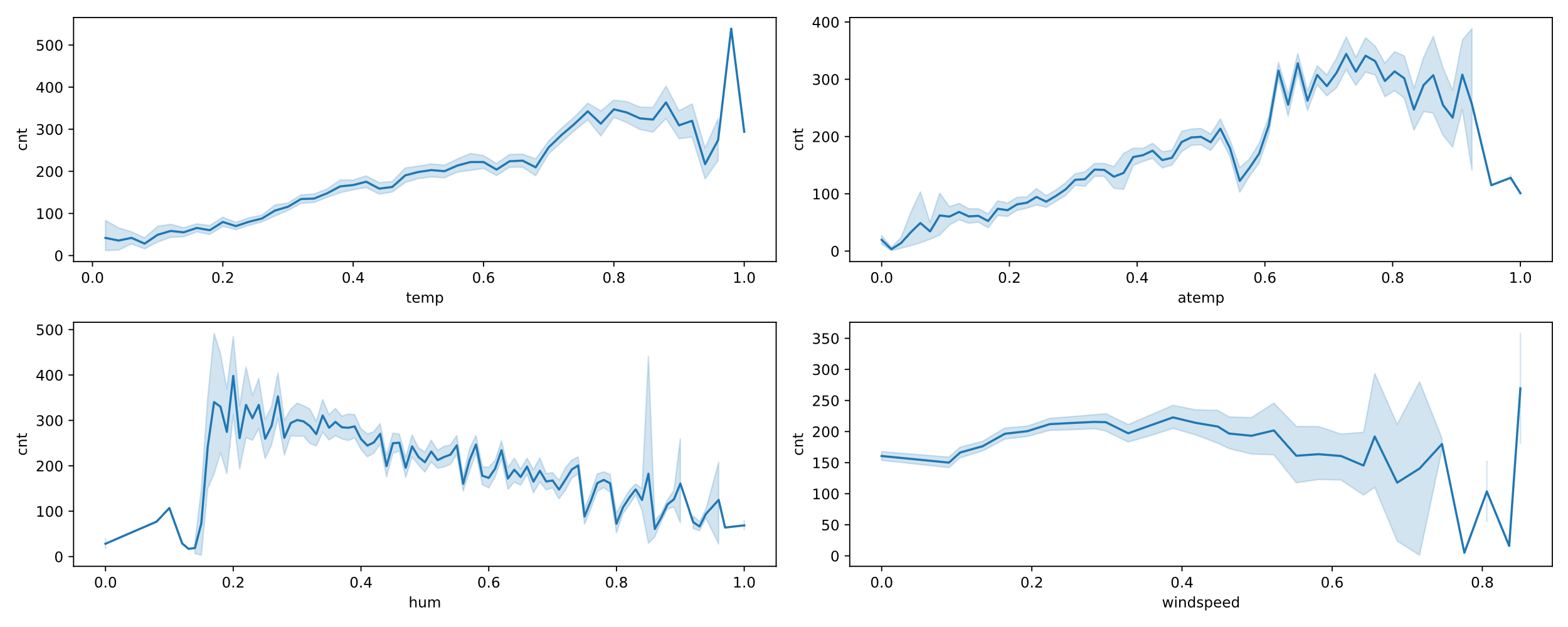


### 3.4 特征筛选

虽然上一个section比较了那么多个特征与共享自行车使用量之间的关系。但是不是所有特征都有决定性的特点。比如假日与工作日之间的关系。季节与星期之间的关系。共享单车业务发展迅速的影响。仔细管擦，有几个额外的特征游离于时间关系之外——空间也是很关键的一个要素。



下图展示了四个环境因素影响下，共享自行车使用量的变化。



由此可见，环境对共享自行车使用量的影响是十分大的。其曲线变化无法通过日常经验进行推测。

**特征选择方法：**

1. 过滤法：

去掉取值变化小的特征，此方法虽然简单但是其实在很多情况下都不太适用，如果机器的资源充足，并且希望尽量保留所有信息，可以把阈值设置比较高，或者过滤离散特征只有一个取值的特征。

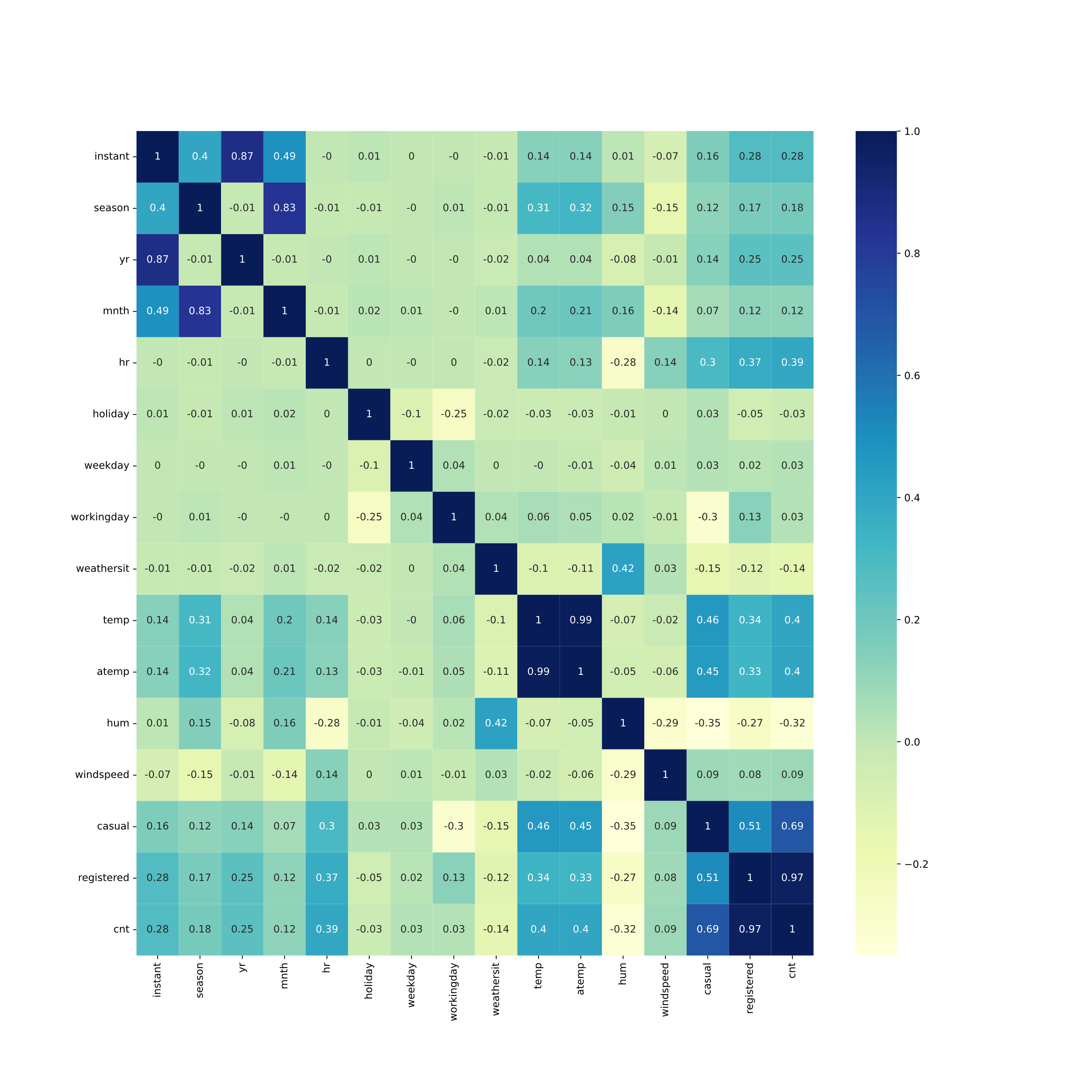
（2）单变量特征选择

单变量特征选择方法独立的衡量每个特征与响应变量之间的关系，单变量特征选择能够对每一个特征进行测试，衡量该特征和响应变量之间的关系，根据得分扔掉不好的特征。

该方法简单，易于运行，易于理解，通常对于理解数据有较好的效果（但对特征优化、提高泛化能力来说不一定有效）

单变量特征选择可以用于理解数据、数据的结构、特点，也可以用于排除不相关特征，但是它不能发现冗余特征。

先看一下特征之间的联系度。



这图能够很好地和上面系列图联系一起，而且大部分都证实了我们的猜想。环境天气的变化很容易就影响了共享自行车使用数量。但是值得注意的是，部分联系度超过了0.6甚至0.8以上的特征我们需要筛选一下。

(1)casual，registered和cnt，casaul + registered = cnt，所以普通用户和注册用户对本次共享单车使用量并没有什么很重要的影响，单纯是拆分出来的部分人数。

(2)atemp和temp，atemp代表的是体感温度，temp代表的是地面以上1m的温度，这两者的之间的关系也是如同是从同一个特征中分离出来，不太能够作为两个特征去使用，剔除。

至此，我们的特征选择完毕了，剔除了几乎同样的特征，剔除了结果集分离出来标签影响。

### 3.5 数据预处理

### 3.6 模型选择

在刚刚的数据预处理中，我们选择了sklearn的线性回归模型进行对比，可以看到RMSE(均方根误差)在数据完全未处理的情况下，是20000，而在数据预处理后下降到了12000，这仅仅是在开方的情况下进行的结果分析，也由此可以知道，线性回归模型并不是一个适合本次数据集的模型。同时更不是一个易于调整和分析的模型。因为数据集特征集和结果集之间的关系很明显不是单纯线性关系。

###### 3.6.1随机森林模型

由传统的多元线性回归方法的自行车租赁需求预测模型,我们发现,传统的方法是不适合自行车租赁需求预测〇再次查看数据,我们发现的因素包含在季节,天气是这样一个哑变量因素,如季节,是1、2、3、4这样的表达式，使得线性回归分析不准确，根据这样数据的特点，让我想想随机森林的方法。森林中的每棵树都依赖于一个随机向量，森林中的向量都是独立同分布的。最终的决策树是基于随机向量势树上的“投票”生成的，即随机森林的分类选择获得的票数最多。

###### 3.6.2 Xgboost

Xgboost作为一种新型的集成学习方法，优点颇多，Xgboost支持并行处理，众所周知，决策树的学习最耗时的一个步骤是对特征的值进行排序，Xgboost在训练之前预先对数据进行了排序。然后保存结构，后续计算中重复这个结构，大大减少了计算量，给使用者很大的调参空间。

###### 3.6.3 LSTM

长短期记忆模型（long-short term memory）是一种特殊的[RNN](https://so.csdn.net/so/search?q=RNN&spm=1001.2101.3001.7020)模型，是为了解决RNN模型梯度弥散的问题而提出的；在传统的RNN中，当时间比较长时，需要回传的残差会指数下降，导致网络权重更新缓慢，无法体现出RNN的长期记忆的效果，因此需要一个存储单元来存储记忆，因此LSTM模型被提出；LSTM模型是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。

###### 3.6.4 人工神经网络

人工神经网络（简称神经网络）是一种受人脑的生物神经网络启发而设计的计算模型。人工神经网络非常擅长从输入的数据和标签中学习到映射关系，从而完成预测或者解决分类问题。人工神经网络也被称为通用拟合器，这是因为它可以拟合任意的函数或映射。

前馈神经网络是我们最常用的一种网络，它一般包括3层人工神经单元，即输入层、隐含层和输出层，其中，隐含层可以包含多层，这就构成了所谓的深度神经网络。

## 4 结果分析

### 4.1 实验环境

|  |  |  |
| --- | --- | --- |
| 工具 | 版本 | 用途 |
| Python | 3.7.7 | 主要实现的编程语言 |
| jupyter-core | 4.11.1 | 网页文档式编程 |
| keras | 2.10.0 | 开源人工神经网络库 |
| Markdown | 3.4.1 | 文档编辑器 |
| matplotlib | 3.5.3 | Python 的 2D绘图库 |
| numpy | 1.21.6 | 开源的数值计算扩展 |
| seaborn | 0.11.2 | 对matplotlib进行二次封装的可视化工具 |
| scikit-learn | 1.0.2 | 机器学习库 |
| tensorflow | 2.10.0 | 深度学习框架 |
| torch | 1.12.1 | 深度学习框架 |

### 4.2评价指标

预测值：

真实值：

###### 4.2.1 MSE均方误差（Mean Square Error）

MSE是真实值与预测值的差值的平方然后求和平均。通过平方的形式便于[求导](https://so.csdn.net/so/search?q=%E6%B1%82%E5%AF%BC&spm=1001.2101.3001.7020)，所以常被用作线性回归的**损失函数**。范围[0,+∞)，当预测值与真实值完全吻合时等于0，即完美模型；误差越大，该值越大。

###### 4.2.2 RMSE 均方根误差

均方根误差（Root Mean Square Error），其实就是MSE加了个根号，这样数量级上比较直观，比如RMSE=10，可以认为回归效果相比真实值平均相差10。衡量观测值与真实值之间的偏差。常用来作为[机器学习](https://so.csdn.net/so/search?q=%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0&spm=1001.2101.3001.7020)模型预测结果衡量的标准。

###### 4.2.3 ****MAE（Mean Absolute Error）****

**平均绝对误差**是绝对误差的平均值。可以更好地反映预测值误差的实际情况。范围[0,+∞)，当预测值与真实值完全吻合时等于0，即完美模型；误差越大，该值越大。

###### 4.2.4 r2\_score

搞清楚R2\_score计算之前，我们还需要了解几个统计学概念。

**回归平方和SSR**

即估计值与平均值的误差，反映自变量与因变量之间的相关程度的偏差平方和。

**残差平方和：SSE**

**总离差平方和：SST**

**SST = SSR + SSE**

即估计值与真实值的误差，反映模型拟合程度

**R2\_score计算公式**

R^2 score，即决定系数，反映因变量的全部变异能通过回归关系被自变量解释的比例。计算公式：

进一步化简

（1） 一般用在线性模型中（非线性模型也可以用）

（2）不能完全反映模型预测能力的高低,某个实际观测的自变量取值范围很窄，但此时所建模型的 很大，但这并不代表模型在外推应用时的效果肯定会很好。

### 4.3 实验结果分析与总结