## 1 研究背景

## 2 工作准备

## 3 数据处理

### 3.1 数据来源

### 3.2 预测方法

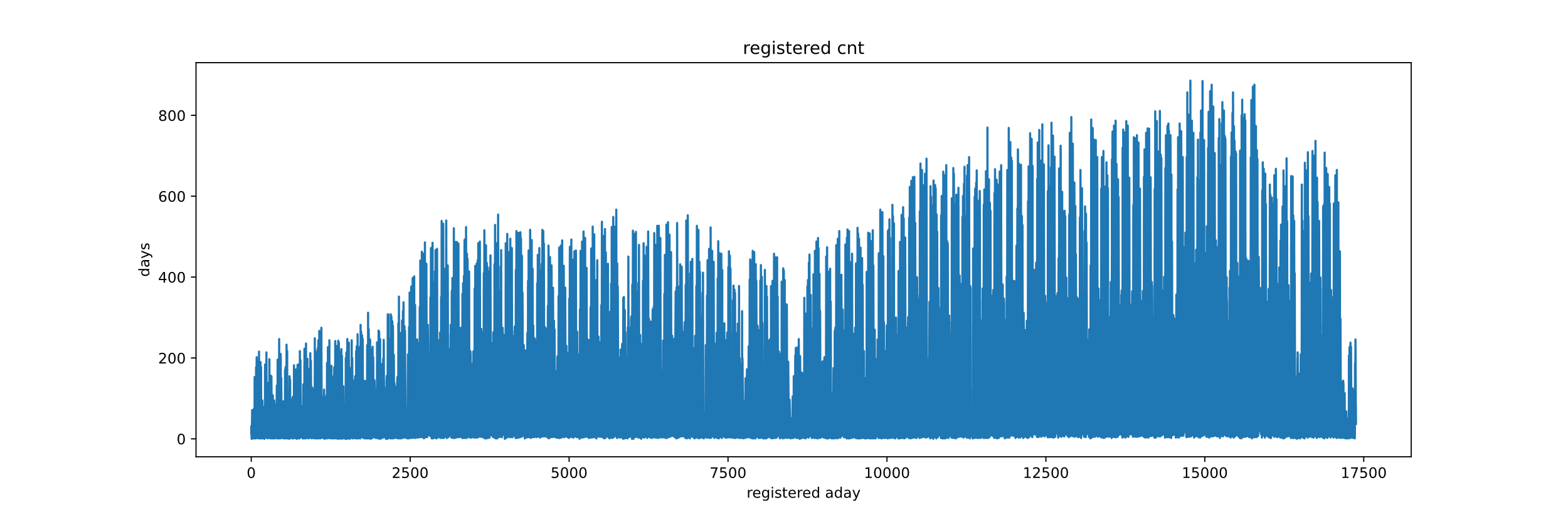
### 3.3 特征分析

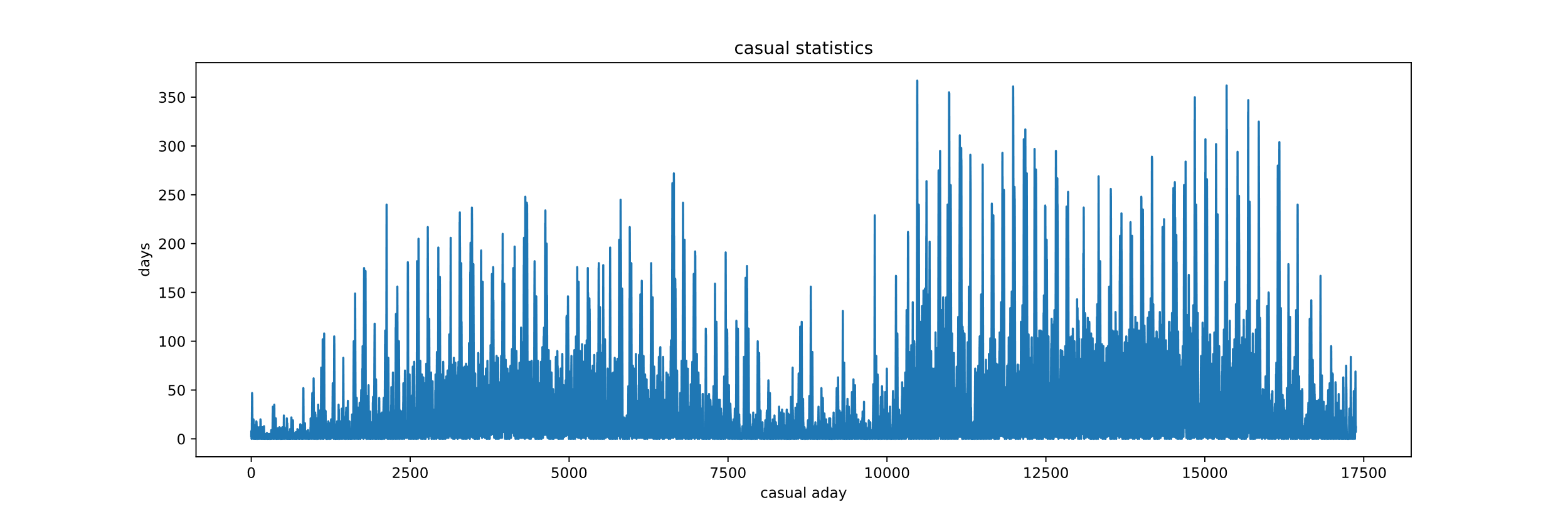
该数据集共有16个标签，但是不是所有的标签都适用于当做特征值。特征筛选的方法主要包括：Filter（过滤法）、Wrapper（封装法）、Embedded（嵌入法）。

|  |  |  |
| --- | --- | --- |
| 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。

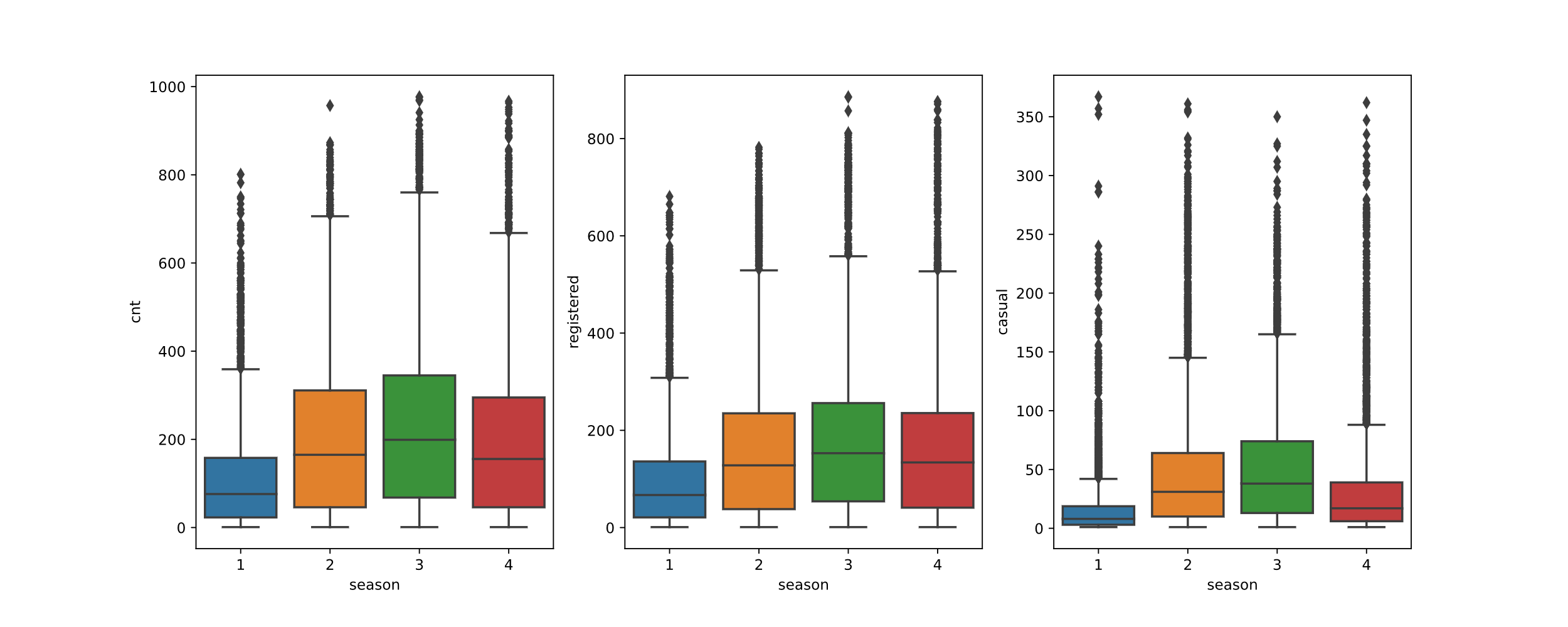
下方两图表示了注册会员与未注册会员使用共享自行车的情况。



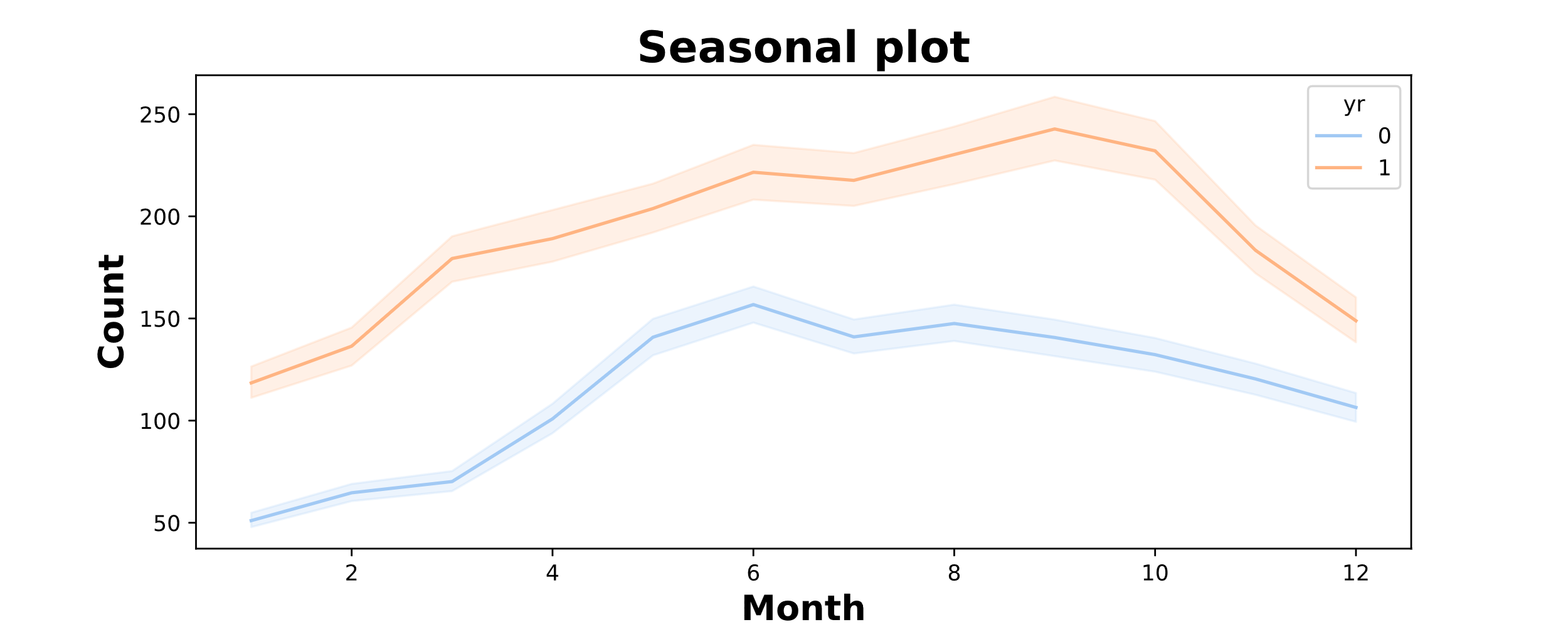


总的来看，注册会员使用人数随着时间增长呈现上升的趋势。

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。



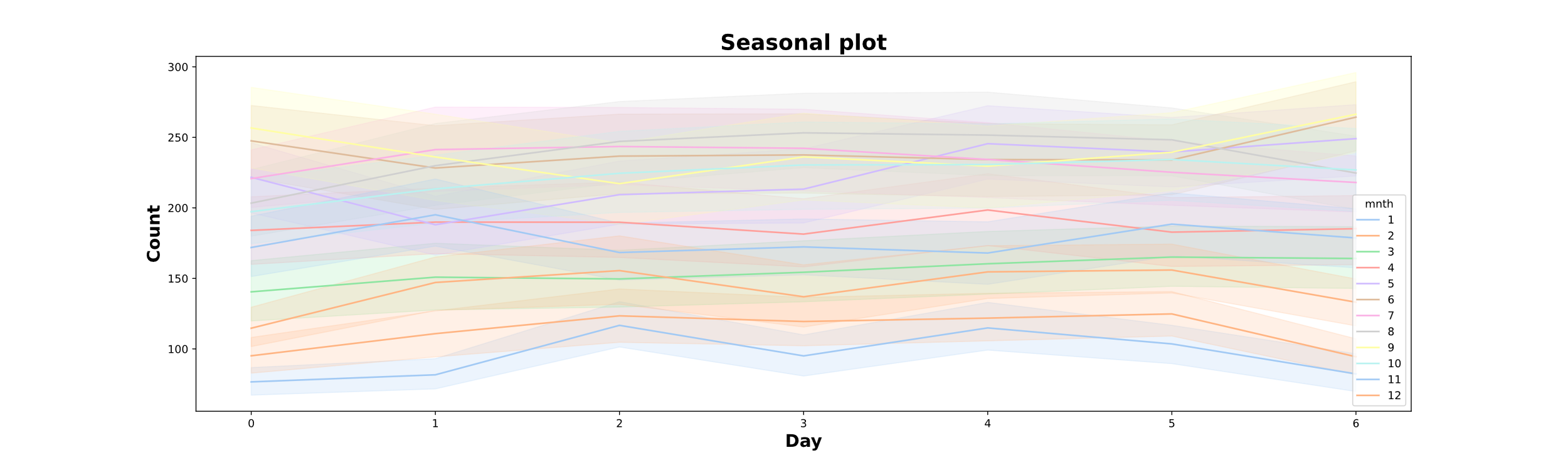
季节是影响共享单车使用情况很重要的一个因素，除此之外，共享单车业务的发展是十分迅速的，从下图中不难看出：

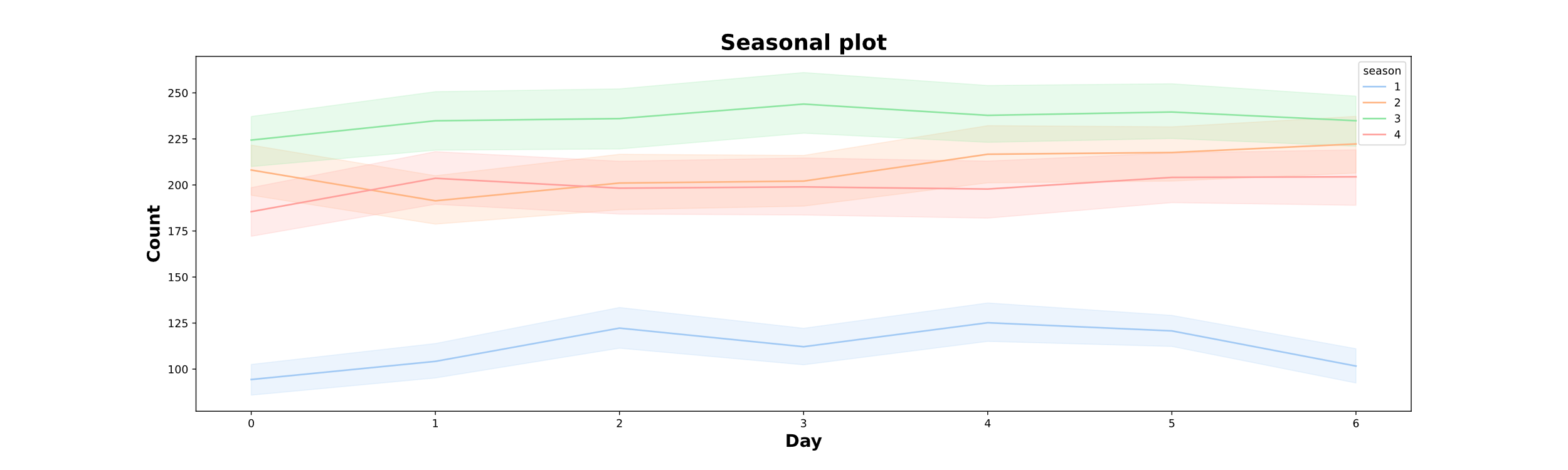


(yr= 0代表2011年，yr = 1代表2012年)

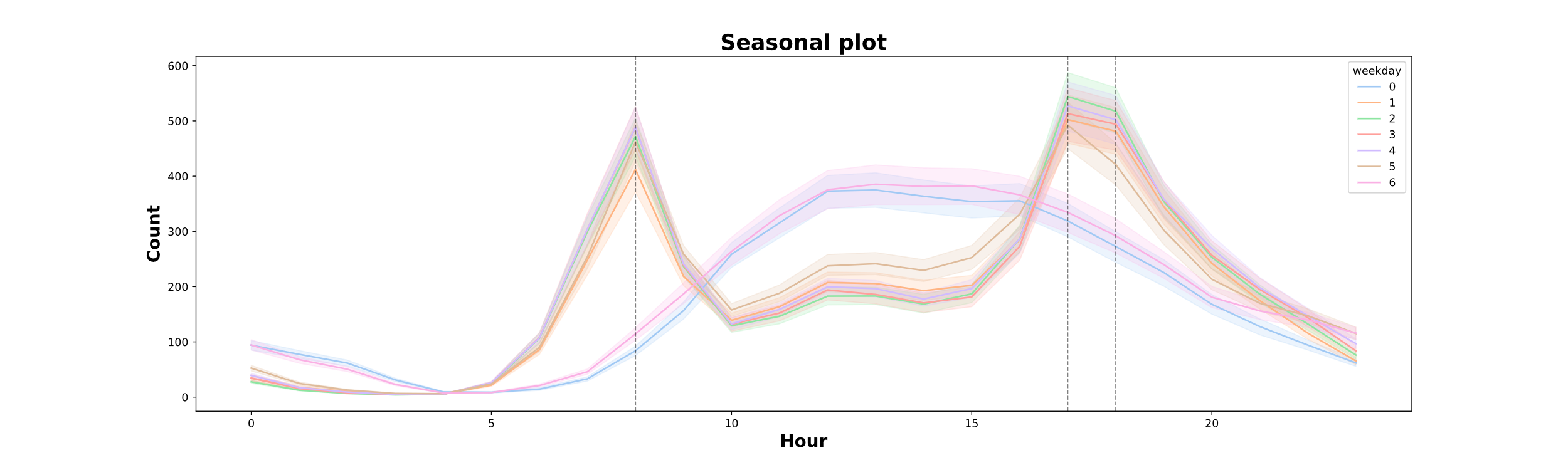
除了2011年末和2012年末的差距相对较小之外，其他时间2012年的共享单车使用数量有着明显的优势，在3月份，2012年共享单车使用量较2011年翻了一番还要多。

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。



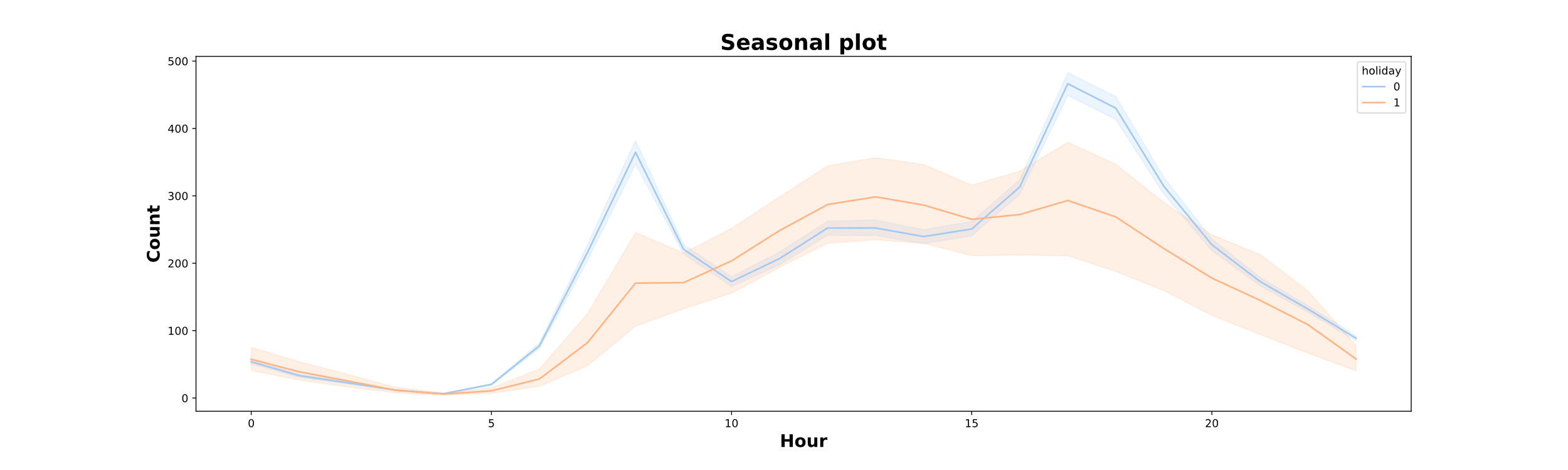


比较完了星期与季节月份之间的关系，一天之中使用共享自行车的时间也尤为关键。比如，早上8:00之前普遍人们的休息时间。使用量很可能与日间有较大的差异。



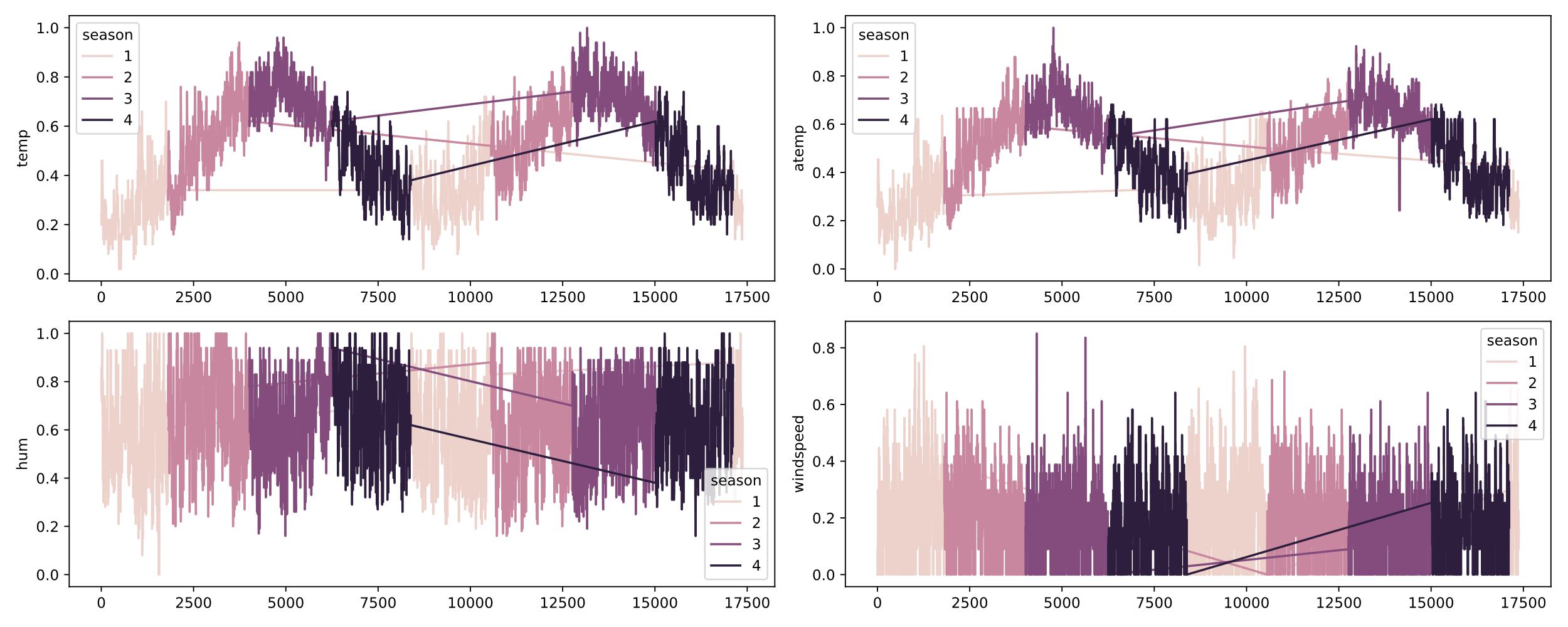
显然，数据的呈现符合我们的猜想，共享自行车的使用量与人们活跃时间有很大的关系。

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。

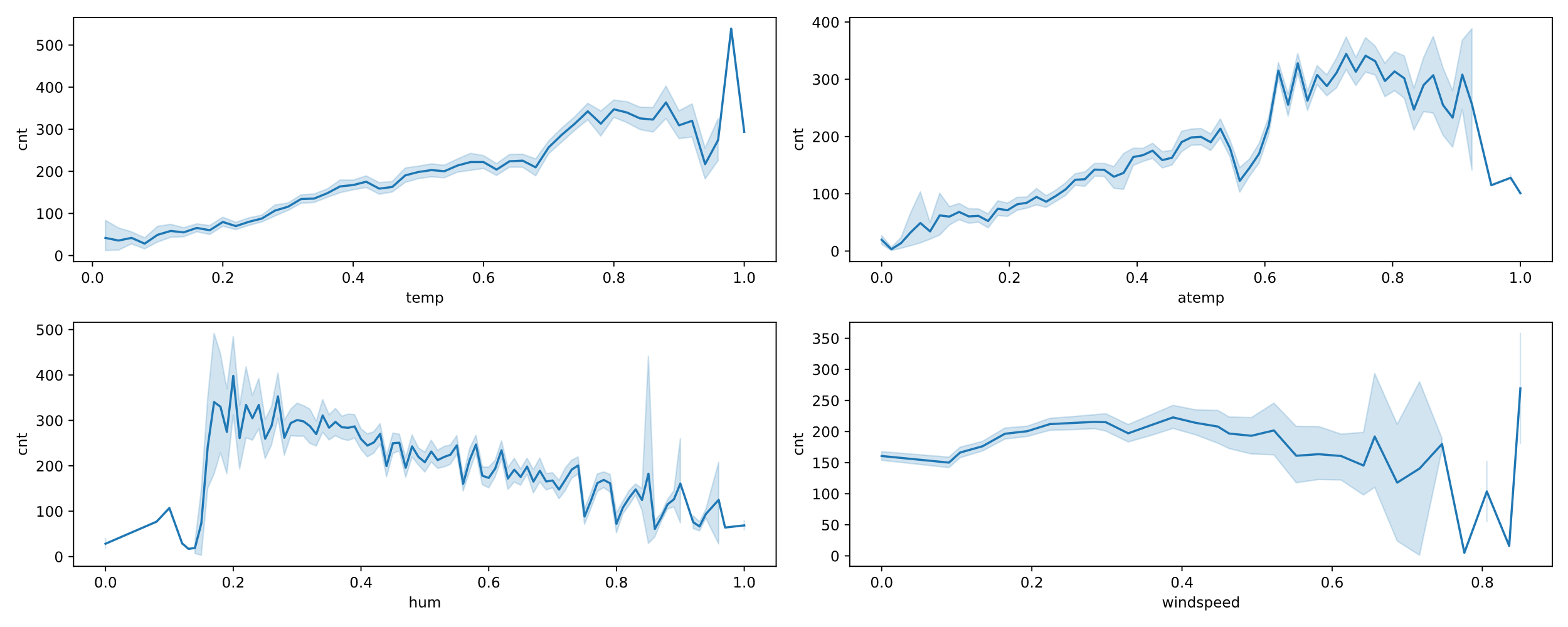


### 3.4 特征筛选

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。

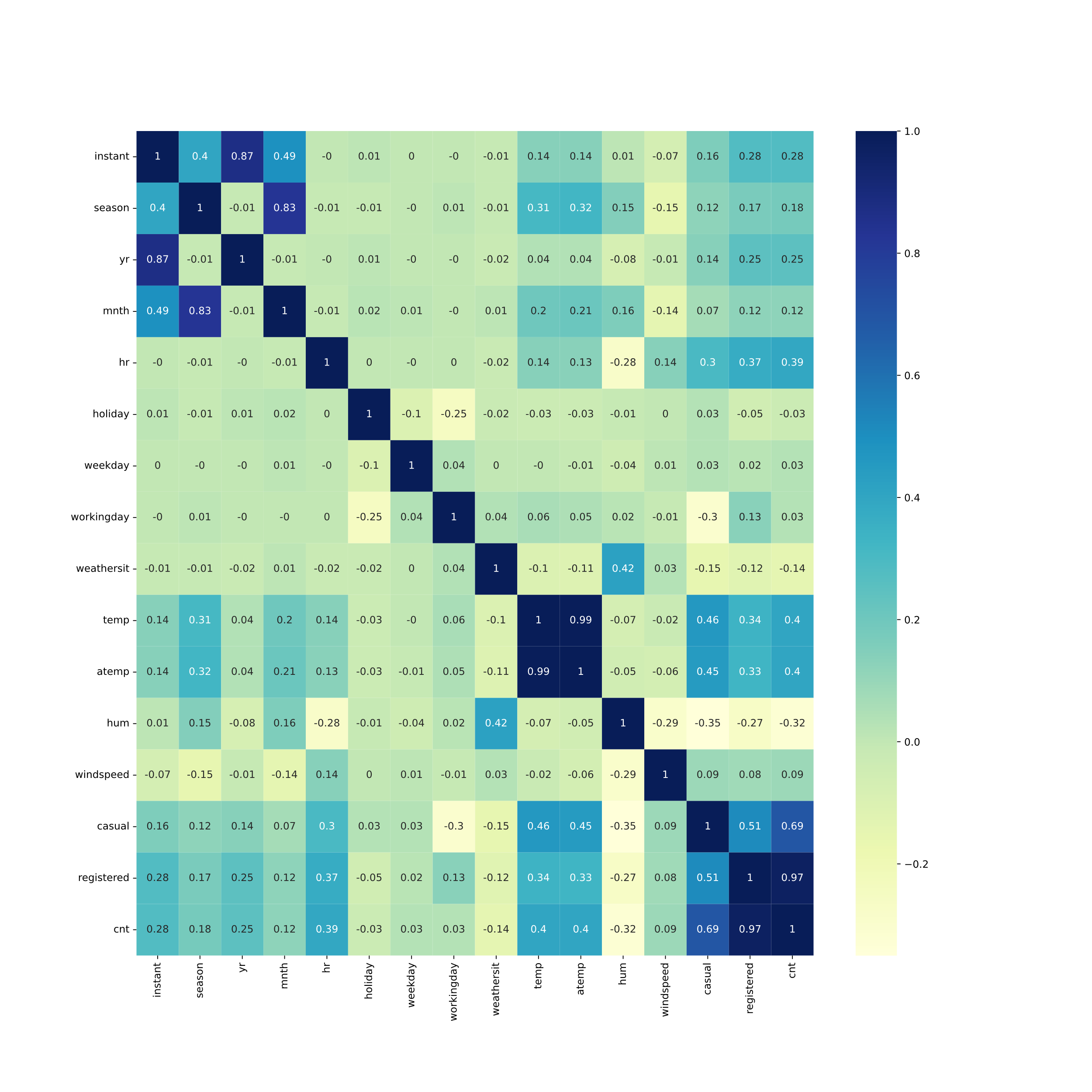
（2）单变量特征选择

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。

本次模型选择将会根据不同模型的表现选择出性能较为出色的模型。

###### 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

**基础环境**

|  |  |
| --- | --- |
| 名称 | 版本 |
| 操作系统 | Windows 11 家庭中文版 |
| CPU | 12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz |
| 内存 | 16G |

**开发环境**

|  |  |  |
| --- | --- | --- |
| 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

###### 4.3.1随机森林模型

在第三节我们用了线性回归的模型来大概预测共享单车的使用情况，可以看到并不理想，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，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，可能随机森林的方法能够提高预测的准确率。为此，本文提出了一种基于随机森林的自行车租赁需求预测模型。

(1) 模型结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 训练集数据量 | 测试集数量 | MSE | MAE | RMSE | 解释方差分 | 袋外分数 | r2\_score | 平均运行时间 |
| 16659 | 720 | 0.2 | 0.311 | 0.447 | 0.775 | 0.917 | 0.756 | 10.9s |

(2) 预测结果概览图

### 

### 4.3.2 xgboost模型

**Xgboost回归模型调参过程**：

(1) 选择较高的学习速率（learning rate）。一般情况下，学习速率的值为0.1。但是对于不同的问题，理想的学习速率有时候会在0.05到0.3之间波动。选择对应于此学习速率的理想决策树数量。 Xgboost有一个很有用的函数“cv”，这个函数可以在每一次迭代中使用交叉验证，并返回理想的决策树数量。

(2) 对于给定的学习速率和决策树数量，进行决策树特定参数调优（max\_depth，min\_child\_weight，gamma，subsample，colsample\_bytree）。在确定一棵树的过程中，我们可以选择不同的参数。

(3) Xgboost的正则化参数的调优。（lambda，alpha）。这些参数可以降低模型的复杂度，从而提高模型的表现。

(4) 降低学习速率，确定理想参数。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 参数 | 起始值 | 终止值 | 最佳分数值 | 最佳分数 |
| n\_estimators | 20 | 300 | 140 | 0.769 |
| max\_depth | 3 | 10 | 9 | 0.81 |
| min\_child\_weight | 1 | 6 | 3 | 0.81 |
| gamma | 0 | 10 | 0 | 0.81 |
| colsample\_bytree | 0 | 10 | 0.7 | 0.825 |
| subsample | 0 | 10 | 0.8 | 0.825 |
| reg\_alpha | 3 | 5 | 3 | 0.832 |
| reg\_lambda | 3 | 5 | 5 | 0.832 |
| learning\_rate | 0.05 | 0.5 | 0.1 | 0.832 |

(5) 模型结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 训练集数据量 | 测试集数量 | MSE | MAE | RMSE | 解释方差分 | r2\_score | 调参时间 | 训练时间 |
| 16659 | 720 | 1472.4 | 16.9 | 38 | 0.79 | 0.771 | 13min | 2.3s |

(6) 模型分析

Xgboost可以利用调参来进行梯度提升，但是在多次调参的结果后，模型都无法达到预期，但是在了解每一个参数背后的意义，能够有助于训练更好的模型。

###### 4.3.3 人工神经网络

在前面的讨论中，我们看到，只要能够调节神经网络中各个参数的组合，就能得到任意想要的曲线。可问题是，我们应该如何选取这些参数呢？答案就在于训练。要想完成神经网络的训练，首先要给这个神经网络定义一个损失函数，用来衡量网络在现有的参数组合下输出表现的好坏。

1. 搭建神经网络

单个神经元可以形成一条有规律的曲线，在多个神经元的加持下，可以很好的解决机器学习的问题，一般地，只要变换参数组合，我们就可以用两个隐含层神经元拟合出任意具有单峰的曲线。因此我们在使用线性回归模型时，不用再担心模型曲线与预期不符合的问题了。相比决策树，拟合能力更强，深度学习可以自动提取原始数据的特征

由前文数据预处理可知，独热编码下，数据被划分为56个特征，所以神经网络的输入层是56个神经元，隐含层是人为预估，本次调整为10个神经元。

1. 数据的分批处理

在进行训练循环的时候，我们还会遇到一个问题。在前面的例子中，在每一个训练周期，我们都将所有的数据一股脑地儿输入神经网络。这在数据量不大的情况下没有任何问题。但是，现在的数据量是16 875条，在这么大数据量的情况下，如果在每个训练周期都处理所有数据，则会出现运算速度过慢、迭代可能不收敛等问题。

解决方法通常是采取批处理（batch processing）的模式，也就是将所有的数据记录划分成一个批次大小（batch size）的小数据集，然后在每个训练周期给神经网络输入一批数据，如图3.22所示。批量的大小依问题的复杂度和数据量的大小而定，在本例中，我们设定batch\_size=128。

下图为训练时平均误差随训练周期快速下降。

形状

中度可信度描述已自动生成

1. 模型结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 训练集数据量 | 测试集数量 | MSE | MAE | RMSE | 解释方差分 | r2\_score | 训练时间 |
| 16659 | 720 | **0.2** | 0.28 | 0.42 | 0.84 | 0.793 | 74s |

1. 预测结果概览图

图表, 条形图

描述已自动生成

三个模型的对比

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型名称 | | 训练集数据量 | 测试集数量 | MSE | MAE | RMSE | 解释方差分 | r2\_score | 平均运行时间 |
| 随机森林 | 16659 | | 720 | 0.2 | 0.311 | 0.447 | 0.775 | 0.756 | 10.9s |
| Xgboost | 16659 | | 720 | 1472.4 | 16.9 | 38 | 0.79 | 0.771 | 13min |
| 人工神经网络 | 16659 | | 720 | 0.2 | 0.28 | 0.42 | 0.84 | 0.793 | 74s |

###### 4.3.4模型结果分析

本次预测使用的模型有五个，线性回归模型，随机森林决策树，LSTM长短时记忆模型，xgboost模型，人工神经网络模型，五个中最后呈现效果较好的只有三个，线性回归模型由于一开始并不符合本次数据集的预测关系，效果较差。

时间序列预测问题中数据形式的特殊性，导致了搭建模型过程中会遇到各种各样的坑。如果是单点预测，LSTM就能取得非常好的效果。如果是长周期预测，那就必须加一些注意提示才能更好的进行对齐，这个时候可以考虑LSTM外加注意提示。另一个考虑因素是数据噪声大不大，往往噪声越大的数据集，用越简单的模型往往能取得更好的效果，而用了复杂模型反而会因为过拟合导致效果较差。

本次预测的时间段是在2012年最后一个月，可以看到预测的准确率都没有达到到0.8，原因是12月份最后十天属于美国的圣诞节，而在数据集中没有集中体现节日的重要性，同时美国圣诞节最为隆重，具有额外的节日特殊性，在最后一部分时间预测线状图无法拟合。

## 5 总结

本次研究主要通过对现实生活较为普遍的一种出行方式进行了一定程度的分析，在共享出行理念的影响下，社会对“两轮出行” 方式的接受度逐渐提高。预测就得从搭建模型入手，从最初的机器学习模型到神经网络搭建，不但可以了解到模型搭建过程函数拟合对于预测结果的重要性，更多的是在预测方法层面提高了对数据集的认知。

本次模型主要通过对比来筛选出各自的优势，取长补短，可以观察到xgboost是一个较为理想的预测模型，它在预测的过程中有较高的预测分数，相对较短的训练时间。在共享单车预测中有比较良好的性能。