**Kaggle （Bike Sharing Demand）**

**题目：**[**https://www.kaggle.com/c/bike-sharing-demand**](https://www.kaggle.com/c/bike-sharing-demand)

**Github地址：**[**https://github.com/cqychen/mykaggle/tree/master/Bike%20Sharing%20Demand**](https://github.com/cqychen/mykaggle/tree/master/Bike%20Sharing%20Demand)

# 数据探探

这是一个关于自行车租赁预测的题目，相当于国内的ofo，摩拜单车啦。

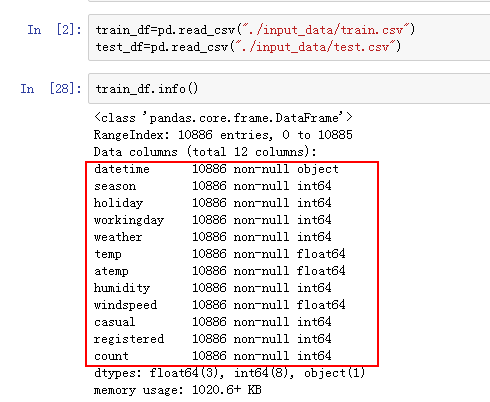
You are provided hourly rental data spanning two years. For this competition, the training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.

训练集提供了一个月的前19天的数据和使用情况，测试集提供后面20号以后的数据，我们主要的任务就是预测20号以后的使用量。

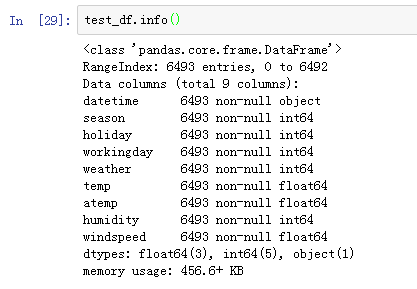
|  |  |  |
| --- | --- | --- |
| 列名 | desc | 中文描述 |
| datetime | hourly date + timestamp | 小时日期 和时间戳 |
| season | 1 = spring, 2 = summer, 3 = fall, 4 = winter | 1：春天 2：夏天 3：秋天 4：冬天 |
| holiday | whether the day is considered a holiday | 当天是否是节假日 |
| workingday | whether the day is neither a weekend nor holiday | 当天是否是工作日 |
| weather | 1: Clear, Few clouds, Partly cloudy, Partly cloudy  2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist  3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds  4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog | 1：晴，少云，部分多云，部分多云。 2：薄雾+多云，薄雾+破碎的云，薄雾+少量的云，雾 3：小雪，小雨+雷雨+散云，小雨+散云 4：大雨+冰盘+雷雨+雾，雪+雾 |
| temp | temperature in Celsius | 温度 |
| atemp | "feels like" temperature in Celsius | 感受 |
| humidity | relative humidity | 湿度 |
| windspeed | wind speed | 风速 |
| casual | number of non-registered user rentals initiated | 未注册用户的租赁数量 |
| registered | number of registered user rentals initiated | 注册用户的租赁数量 |
| count | number of total rentals | 总的租赁数量 |

## 数据总览

读入数据，看看大致信息：



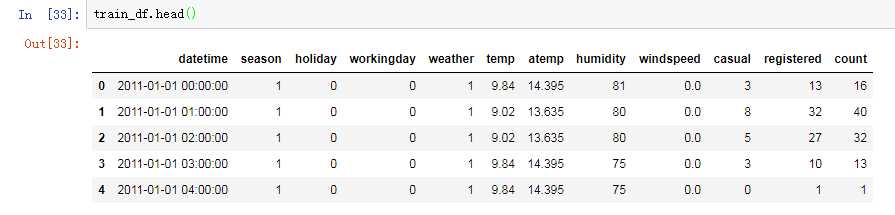
训练集数据共12列，没有数据缺失。哇咔咔



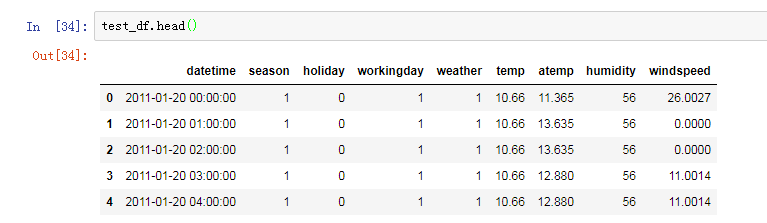
测试集数据共9列，没有数据缺失。

数据明细看看

训练集数据：

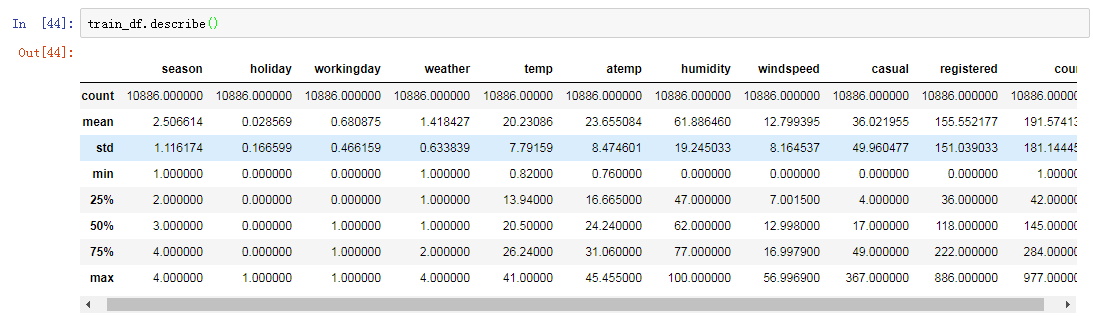


测试集数据：

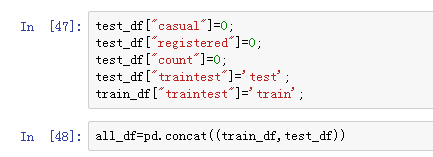


我们可以看到 在测试集中

Casual + register==count



将测试集和训练集合进行拼接，方便做特征工程：



### 日期

我们知道日期的格式是如下：yyyy-MM-dd hh:mm:ss

日期这个东东，基本是要构造出如下的特征：

年

周几

季度

月

小时

一年的第多少周

一年的第多少天

同时可以看到

代码如下：

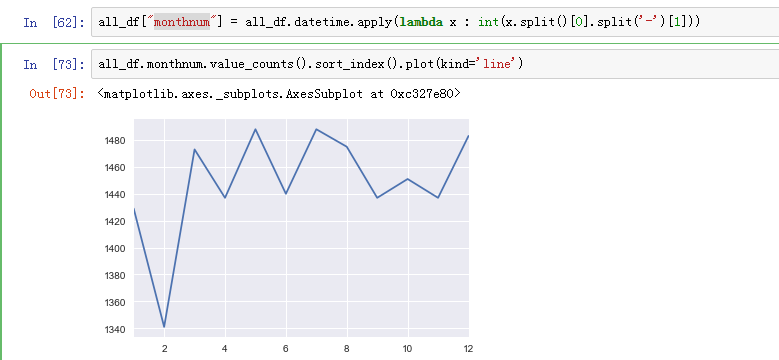
all\_df["date"] = all\_df.datetime.apply(lambda x : x.split()[0])

all\_df["monthnum"] = all\_df.datetime.apply(lambda x : int(x.split()[0].split('-')[1]))

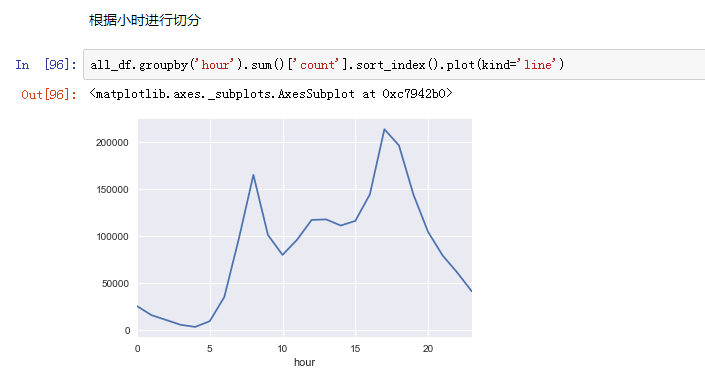
all\_df["daynum"]=all\_df.datetime.apply(lambda x : int(x.split()[0].split('-')[2]))

dailyData["hour"] = dailyData.datetime.apply(lambda x : int(x.split()[1].split(":")[0]))

dailyData["weekday"] = dailyData.date.apply(lambda dateString : calendar.day\_name[datetime.strptime(dateString,"%Y-%m-**%d**").weekday()])



根据小时用量排序查看：



可以看到根据小时用量大致可以分成5个时段：

0~6

6~10

10~15

15~20

20~24

代码处理下

### 季节

### 是否节假日

### 是否工作日

### 天气

### 温度

### 舒适度

### 湿度

### 风速

### 数据详情

# 特征工程

# 模型调优

# 模型融合

# 提交

# 待改进

# 参考连接

<https://www.kaggle.com/viveksrinivasan/eda-ensemble-model-top-10-percentile/notebook>