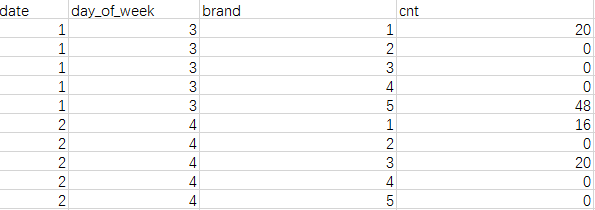
一 数据集说明

给你的数据集包括：train\_set\_revised，包括.txt和.csv两种格式; test\_A\_set\_revised,包括.txt和.csv两种格式;, test\_B\_set\_revised(.txt, .csv), 包括.txt和.csv两种格式.

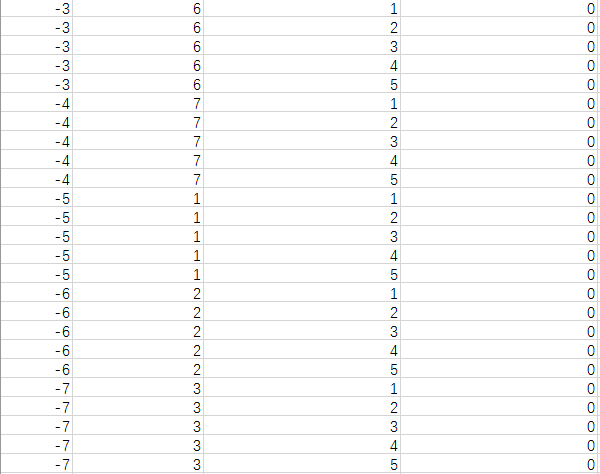
二 数据集补全

我将数据集中所有缺失的天数都补全了，而且是分品牌的。如下图所示：



如果这一天在给定的train\_20171215.txt中出现，则date就取train\_20171215.txt中给定的值，否则就将date设置为负数，分别取-1, -2,….如下图所示：





1. 通过date的正负，可以方便的区分自己补全的天数，和真实在train\_20171215.txt中出现的天数。
2. 这里我区分了不同的品牌，你可以自己聚合，先不考虑品牌。

三 字段说明

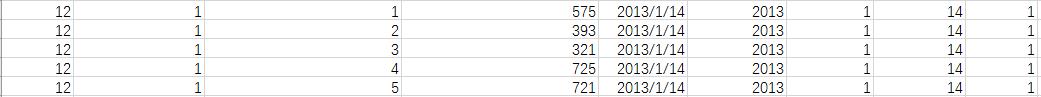
1 date, day\_of\_week, brand, cnt这4个字段的含义和赛题中的解释一样

2 其余字段说明如下

|  |  |
| --- | --- |
| Calendar | 这一天具体的日期 |
| Year | 这一天是哪一年 |
| Month | 这一天是哪一月 |
| Day\_of\_year | 这一天是这一年的第几天 |
| Gap | 这一天前面有几天是缺失的 |

比如：

1. train\_set\_revised



表明：date=12这一天是星期1，品牌1有575辆车上牌照（注意在test\_A\_revised或者test\_B\_revised中，cnt这一列总是为0，是需要我们预测的。）。它是2013-1-14号，年份为2013，月份为1，是2013年的第14天。Gap=1表明前面紧接着的有1天缺失。