*注：*

*\**没有把取对数的那些变量列放在里面考虑，即只用了“原始数据”的数据列。原来想加date进去的，觉得也许可以搞天气，但是好像要换格式而且会议论文里没提，所以我想着后面再说吧。

\*因变量drivingscore



\*图都有清晰版本的，这个文档里可能看不清，就随便意思下

\*ipynb文件是python代码；然后STATA的log文件我就不发了，就几行

**\*因为是初步结果，随便看看那种，所以那个linearRegression我没有用python看系数和p-value啥的，而是直接丢进了STATA后面确定了一些东西之后可能还是要用同一个软件？（还是说就这样也行？）**

1. **数据处理**

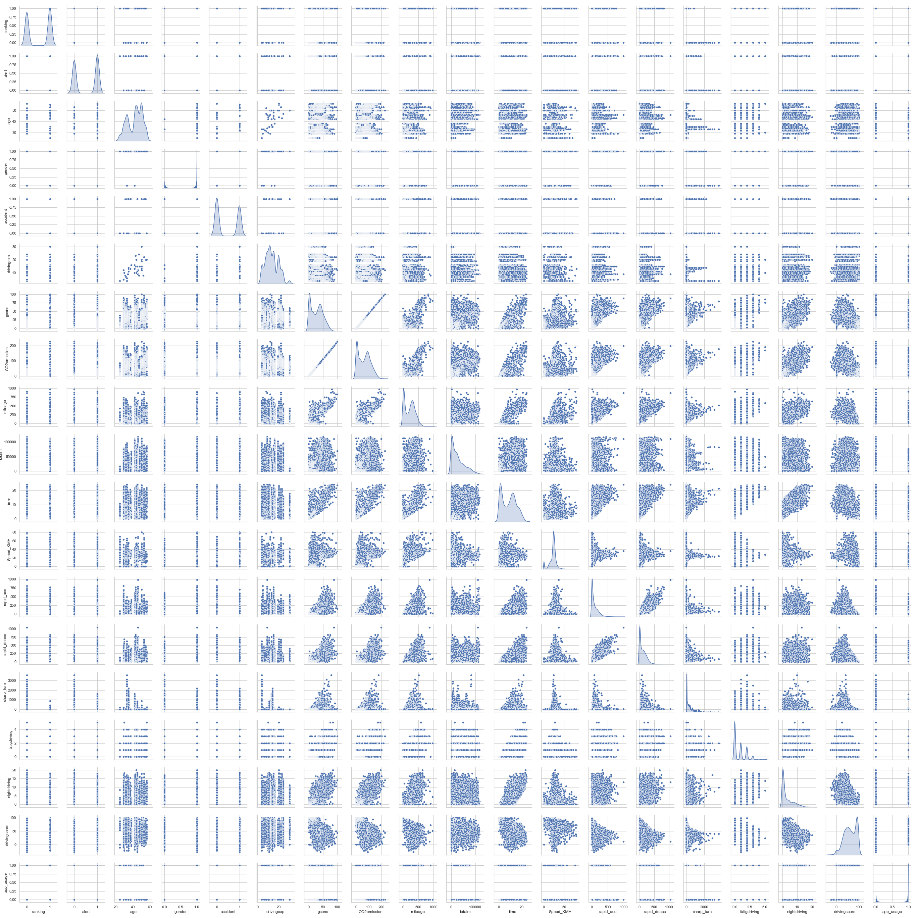
*eng371.xlsx*这个数据集是由最全的那个数据集经过如下处理：

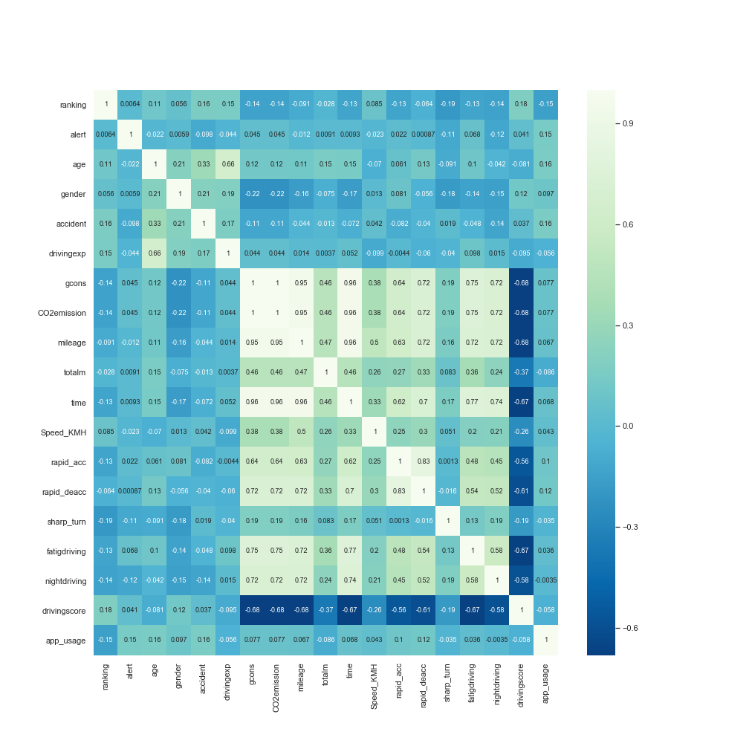
* 删除所有CO2emission为0的值——没出行
* 删除所有“闽xxxx”和“1.22E”乱码车牌——那些跟2019-2020那年的实验不是一批
* 所有变量名为英文

最终得到11187条数据

1. **特征选择（Python，最优特征子集选择）**
2. **观察df中所有变量相关性**

发现几个相关度太高的：CO2和gasconsumption、mileage、time强相关，这里仅选择gascons





于是只剩下15个自变量和一个因变量。

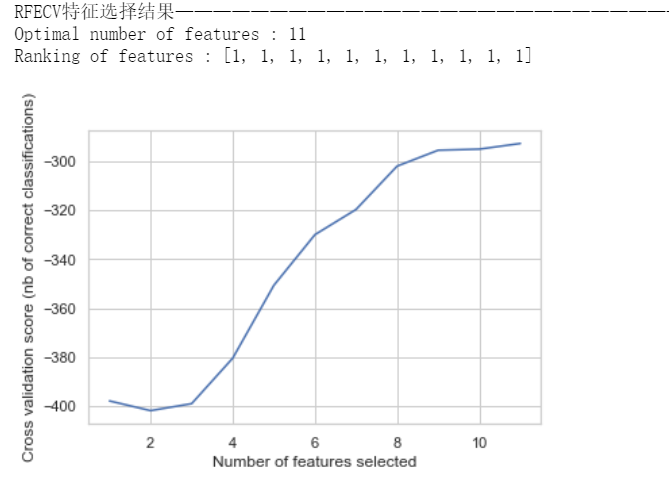
1. **利用RFECV方法进行特征选择（特征子集与交叉验证）**

这个方法不能容忍缺失值，所以这里我分了两种情况处理缺失值。

1. 直接删除有缺失值的变量

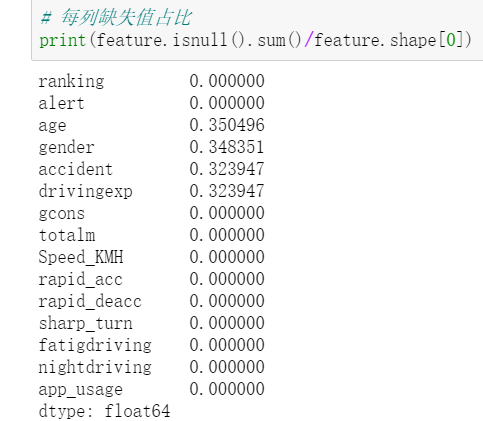
**能够直接删除的原因：**只有age, gender, accident, drivingexp存在缺失值（这玩意儿全都是那个问卷上录下来的）

**删除后特征选择结果：11个特征都保留**

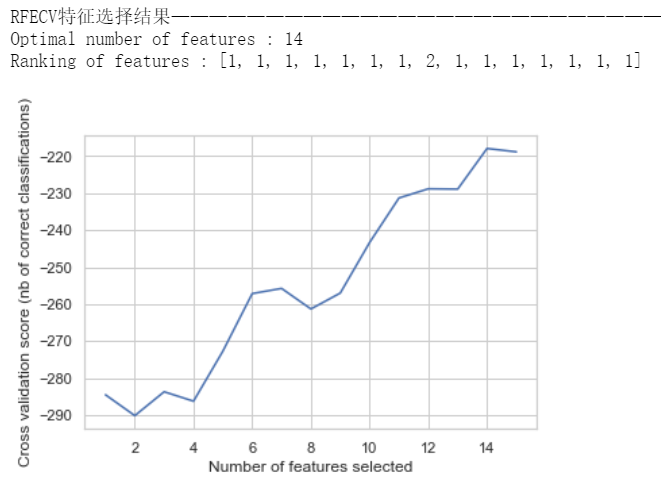


1. 删除所有含缺失值的行

**能够直接删除的原因：**只有age, gender, accident, drivingexp存在缺失值，且缺失值占比不大



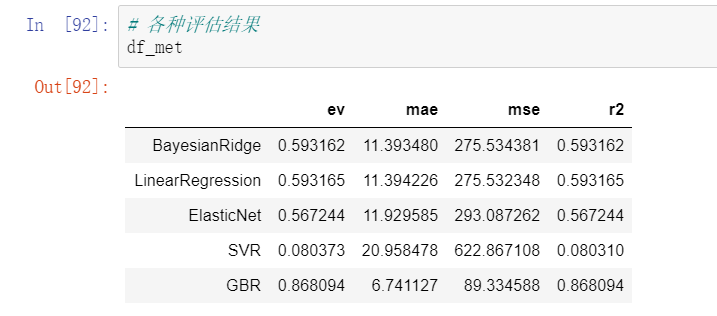
**删除后特征选择结果：14个特征保留，删除totalm**



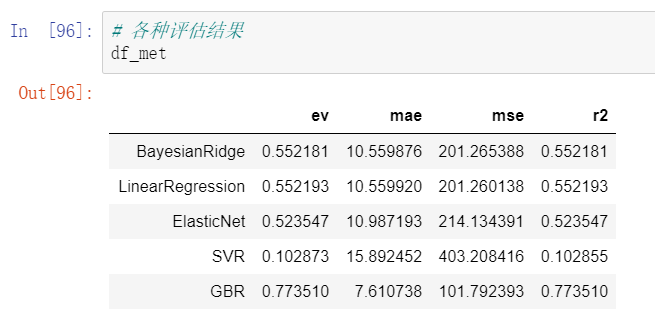
1. **跑模型与评估**
2. Python

**\*我看了五种回归模型，这个结果最好的是GBR-梯度提升线性回归，其次就是LinearR。但GBR这玩意儿系数、p-value啥的我还在寻找中，暂时不知道怎么显示**

删除有缺失值的列跑模型，即用11个features:



删除有缺失值的行后跑模型，即用14个features, 6993个obsevations:



1. STATA

**我好像不大会用STATA，老师你看下是不是有问题QAQ**

