# 作业2.3 编程实现数据的预处理

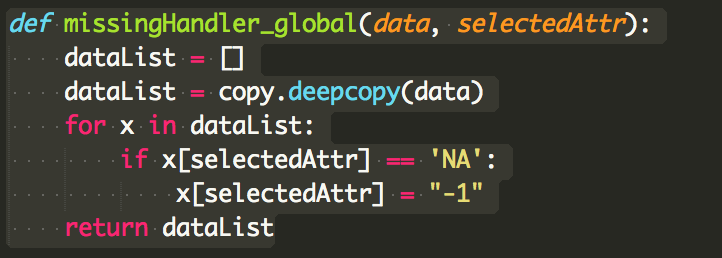
## 数据集1 ——空气质量

### 缺失处理

对缺失的数据采用两种方法：

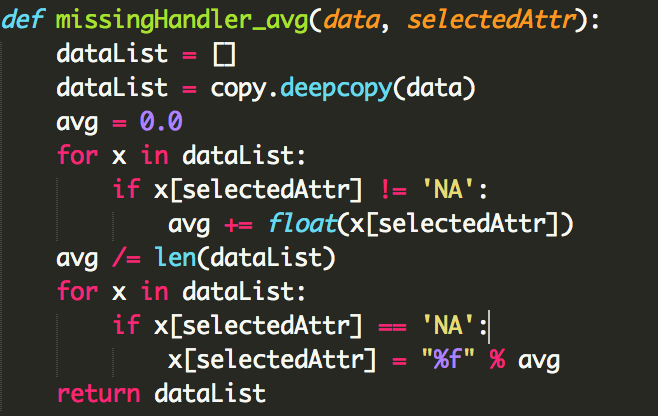
1. 对不需要处理的数据组用特殊值填充

对于特殊值NA，采用数据类型相同的-1来填充，使得在解析数据时不用对同属性的不同数据类型区别对待。代码如下：



1. 对需要处理的数据组用平均值填充

对于缺省的数据，我们采用该属性的平均值去填充，代码如下：

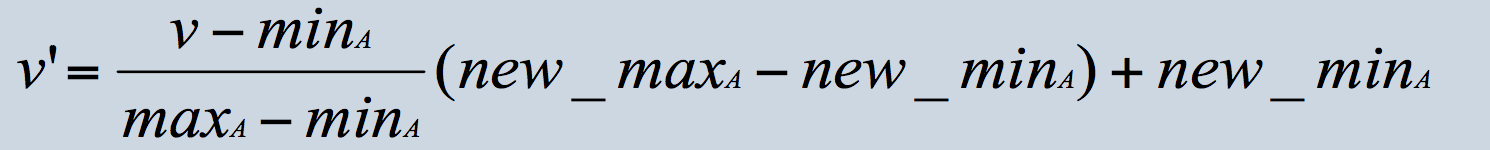


## 数据规范

对数据规范也使用两种手段：最小-最大规范化和小数定标规范化：

1. 最小-最大规范化：

使用如下公式进行数据规范化：

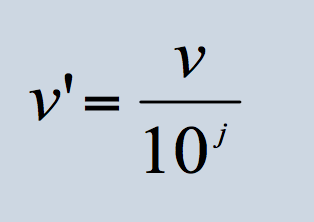


具体代码如下：



1. 小数定标规范化：

使用如下公式进行数据规范化：

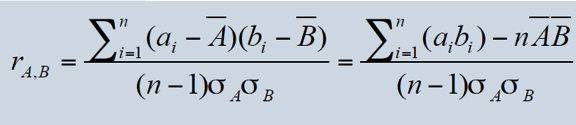


具体代码如下：

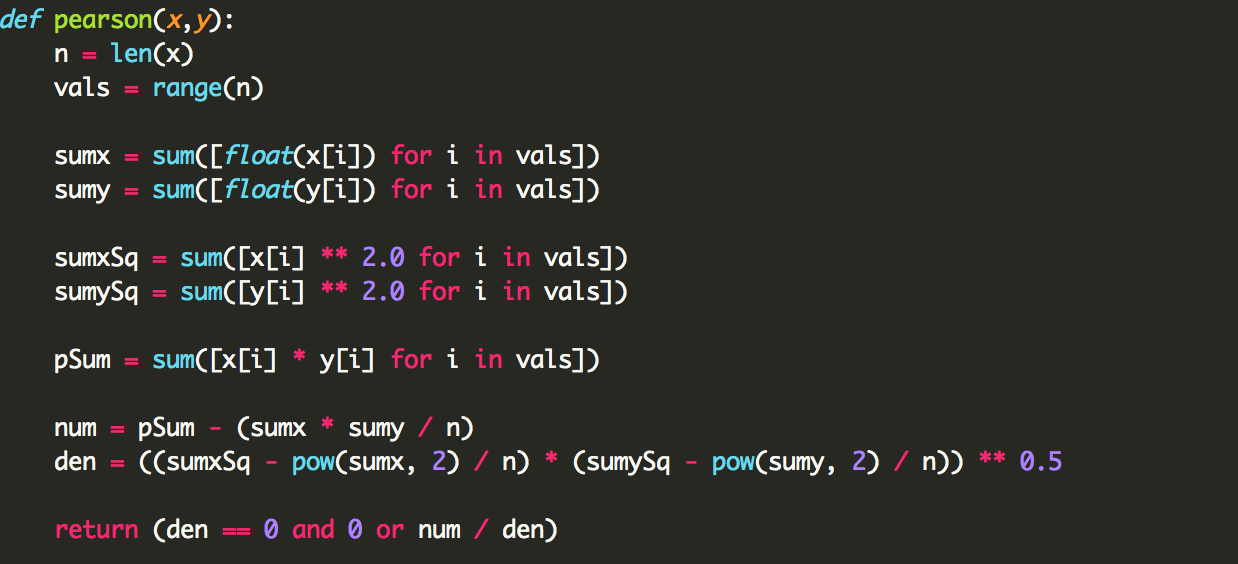


## 相关系数分析

使用相关系数的相关公式如下：

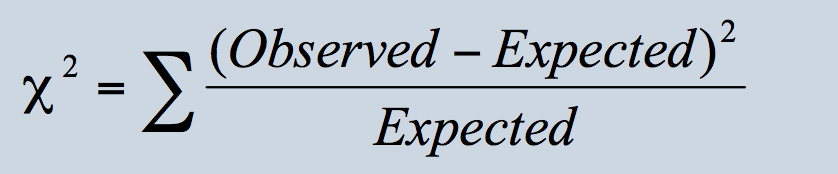


相关代码如下：



## 卡方分析

使用的公式为：



相关代码如下：

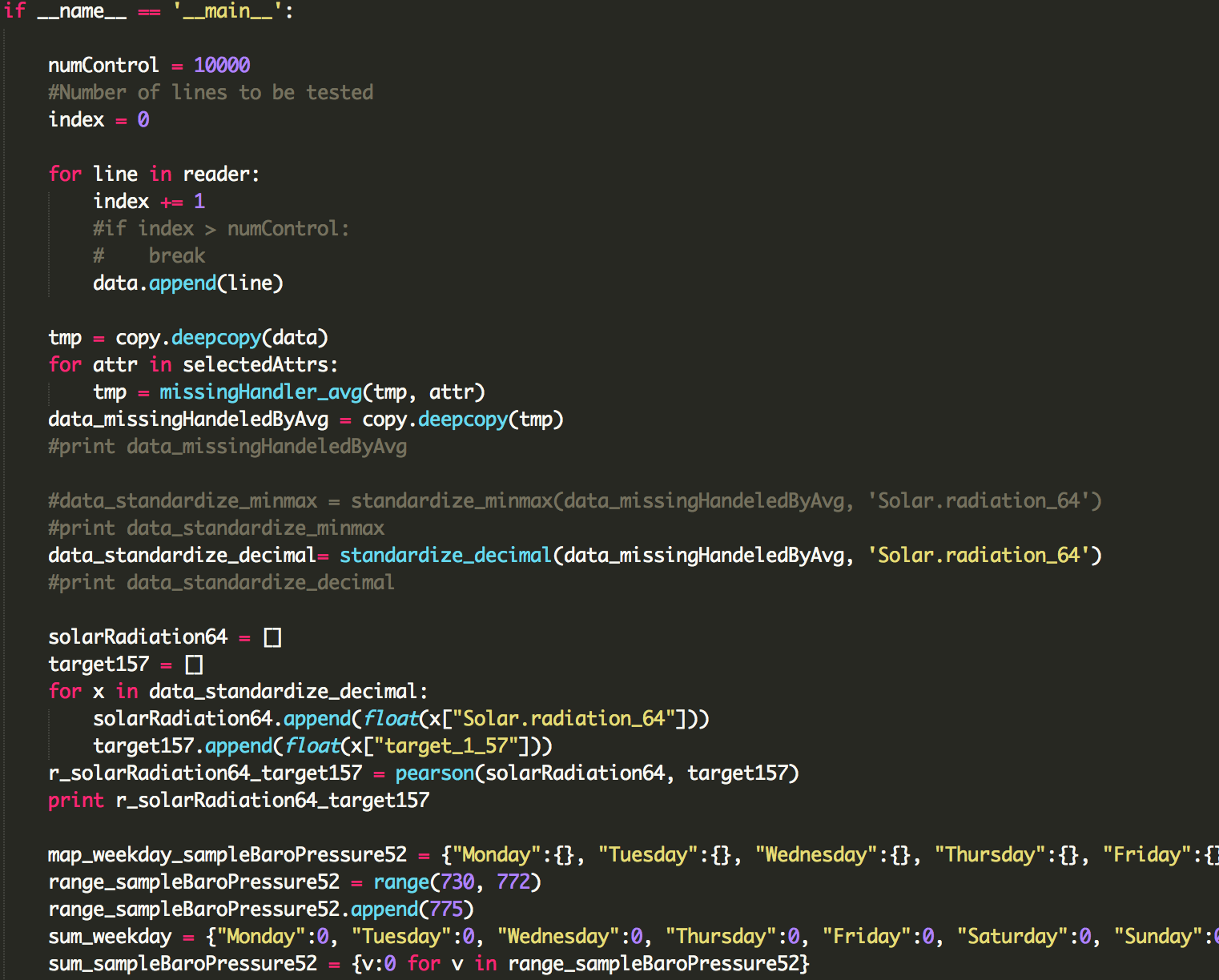


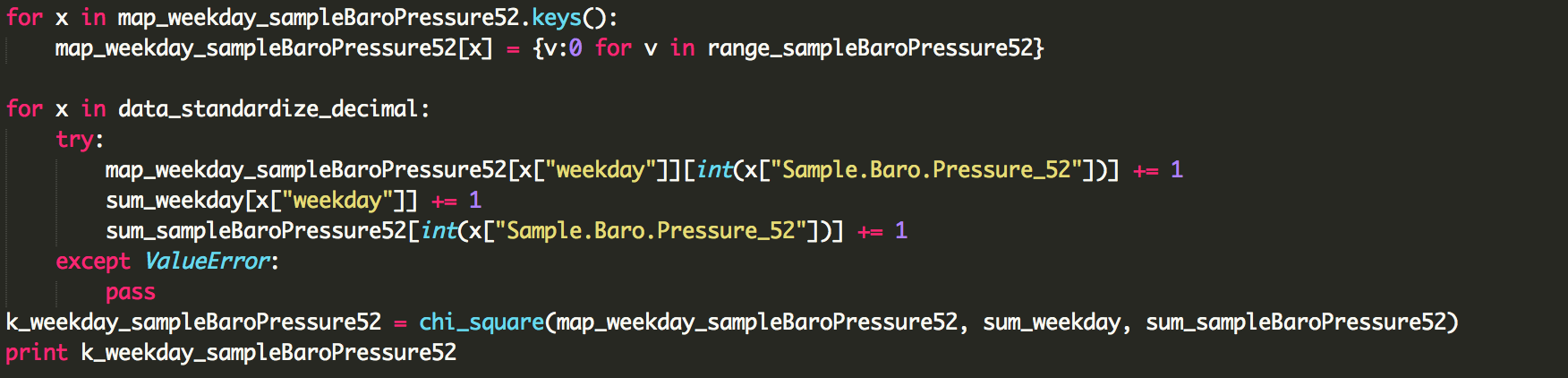
## 最终结果

最终结果如下：



调用部分的代码如下：



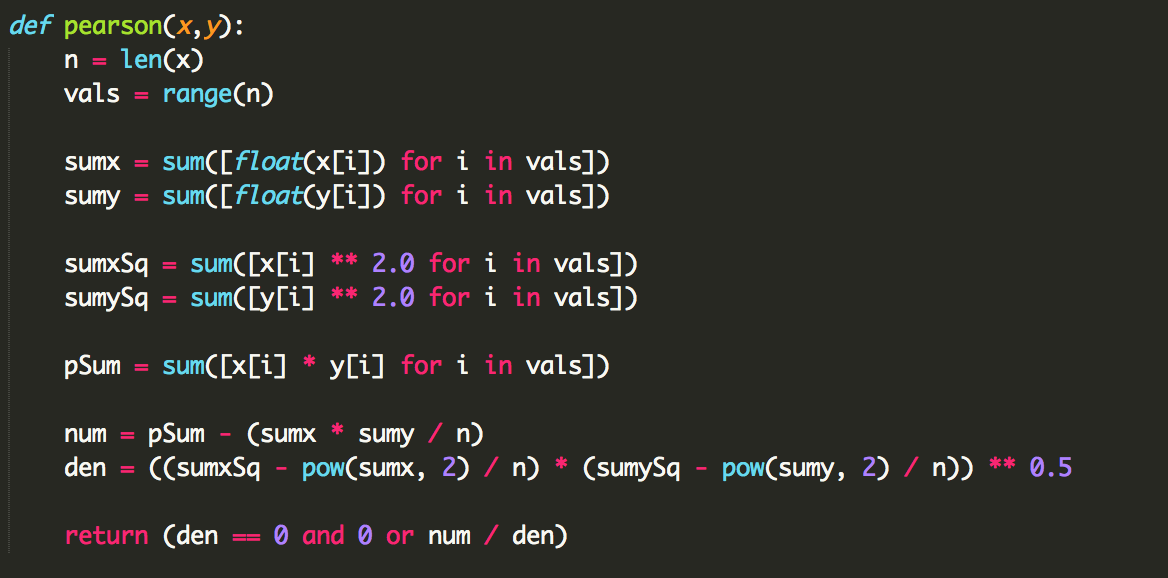


结果分析：根据相关系数和卡方检测的结果，我们可以"Solar.radiation\_64"和"target\_1\_57"属性很弱（几乎无）的负相关关系，"weekday"和"Sample.Baro.Pressure\_52"属性有正相关关系。

## 数据集2 —— 通话数据

分析通话数据的相关系数和卡方检验与前面提到的类似。

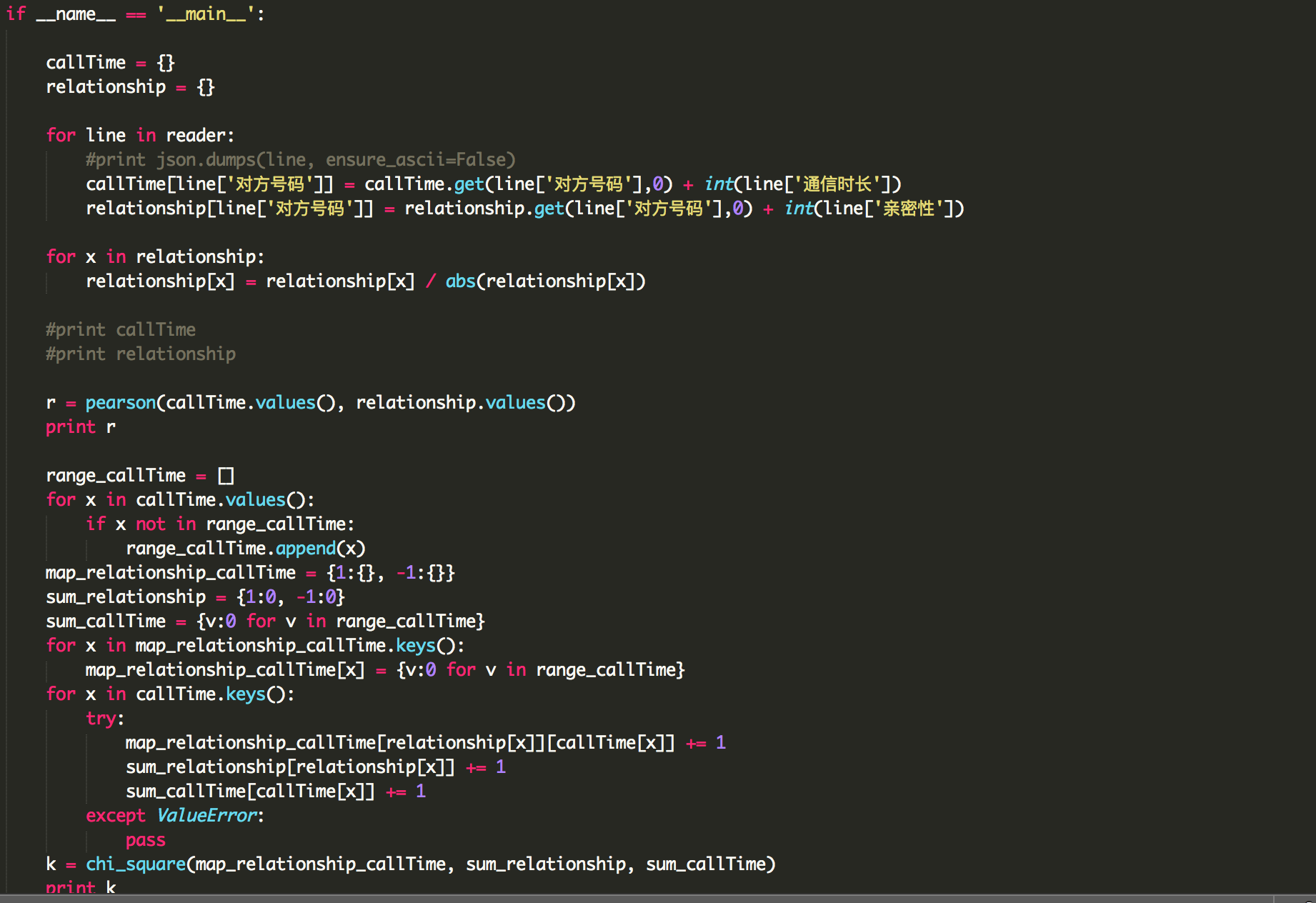
相关系数方法：



卡方检测方法：



调用过程：



运行结果：



结果分析：

根据相关系数和卡方检测的结果，我们可以认为总通信时长和亲密度有一定的正相关关系，但是这个关系并不足够强。