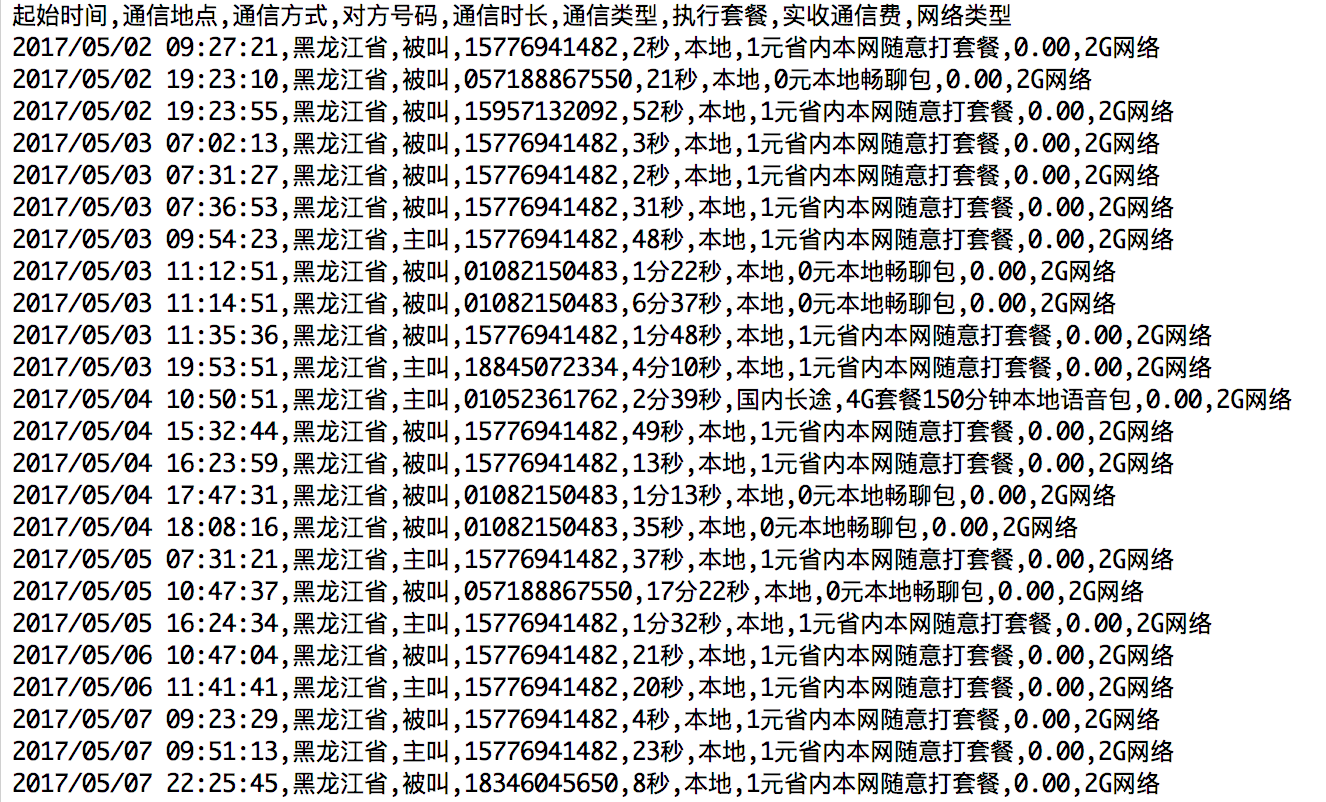
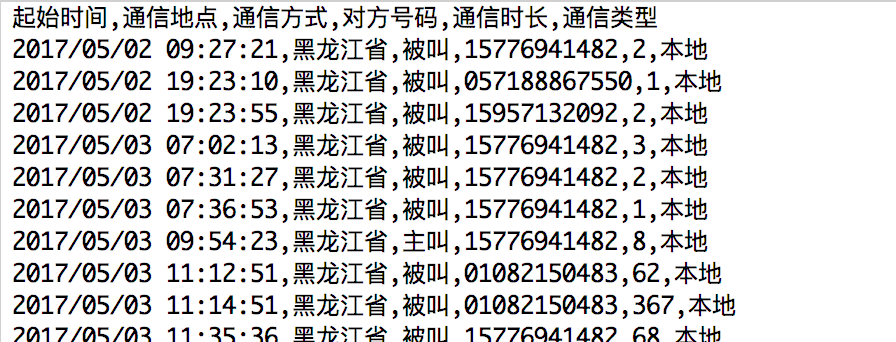
# 数据预处理

## 原始数据



该文件为 /代码/data/phone.csv。

## 清洗后数据

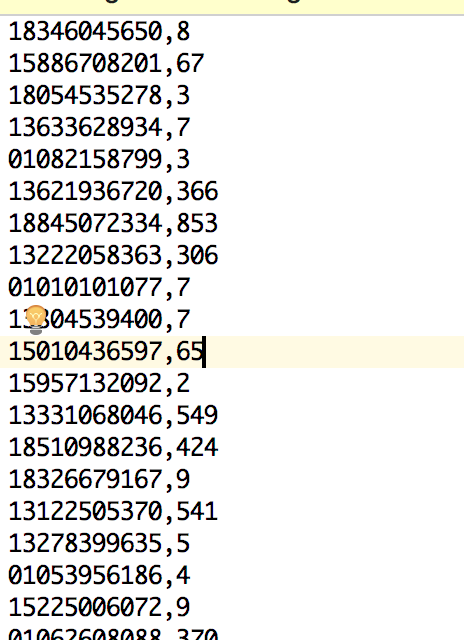


该文件为 /代码/data/phone\_clean.csv。

清洗的工具包为 /代码/utils/csv.py，该工具包中提供了众多接口用以解析、预处理、读取csv文件的数据。下图为选取的两个接口：



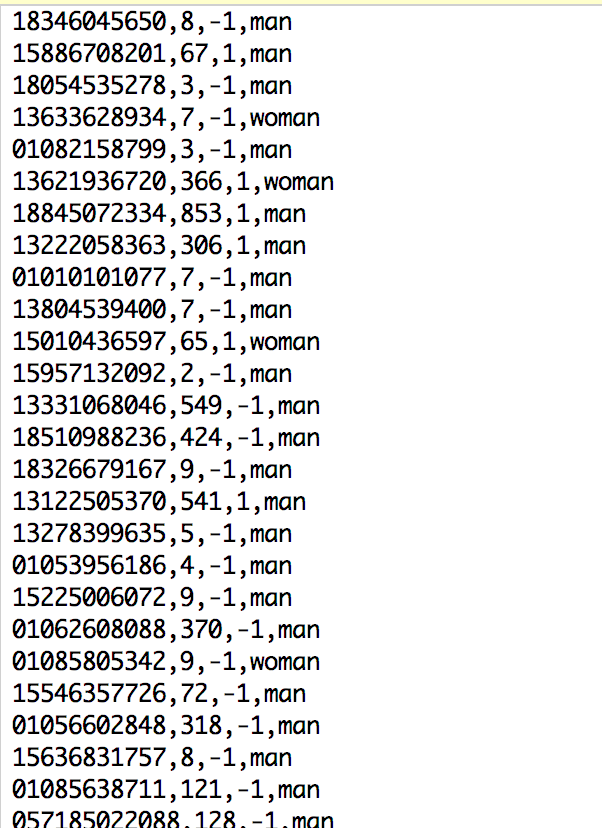
同时对号码的总通话时长进行统计，得出如下文件：



该文件为 /代码/data/phone\_and\_time.csv。

## 标注数据

我们需要对数据的亲密度和性别进行手动标注，标注后的结果为：



本文件在根目录下。

# 编程与计算

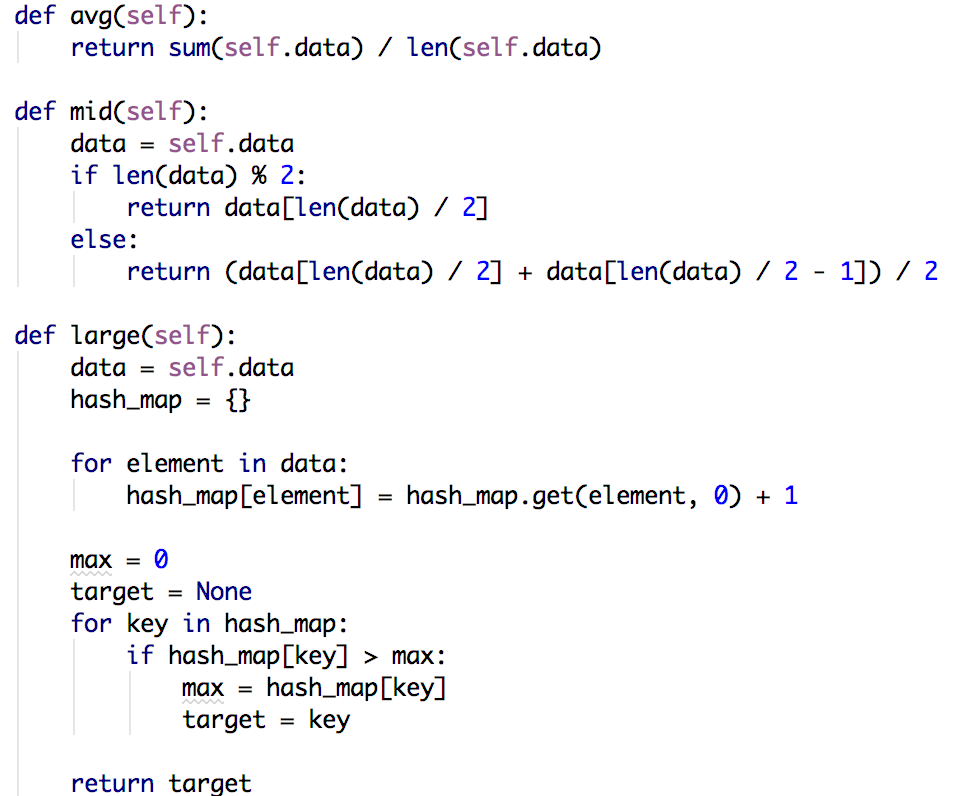
## 统计量计算

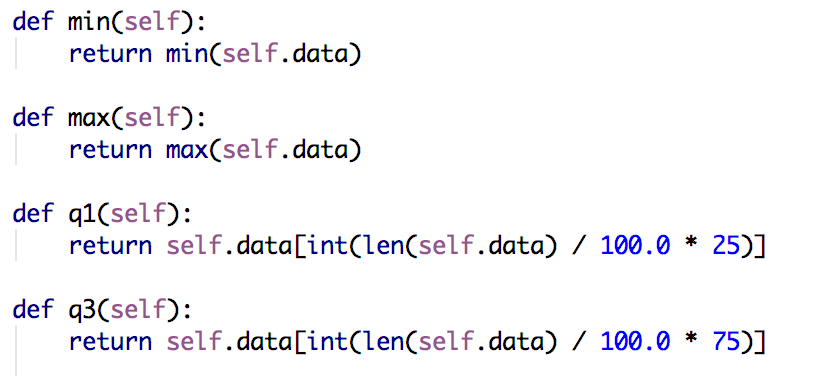
在运行main.py后，本程序会进行一系列文件处理与计算。

其中统计量结果在控制台显示，以下为我个人通话记录的统计量信息：

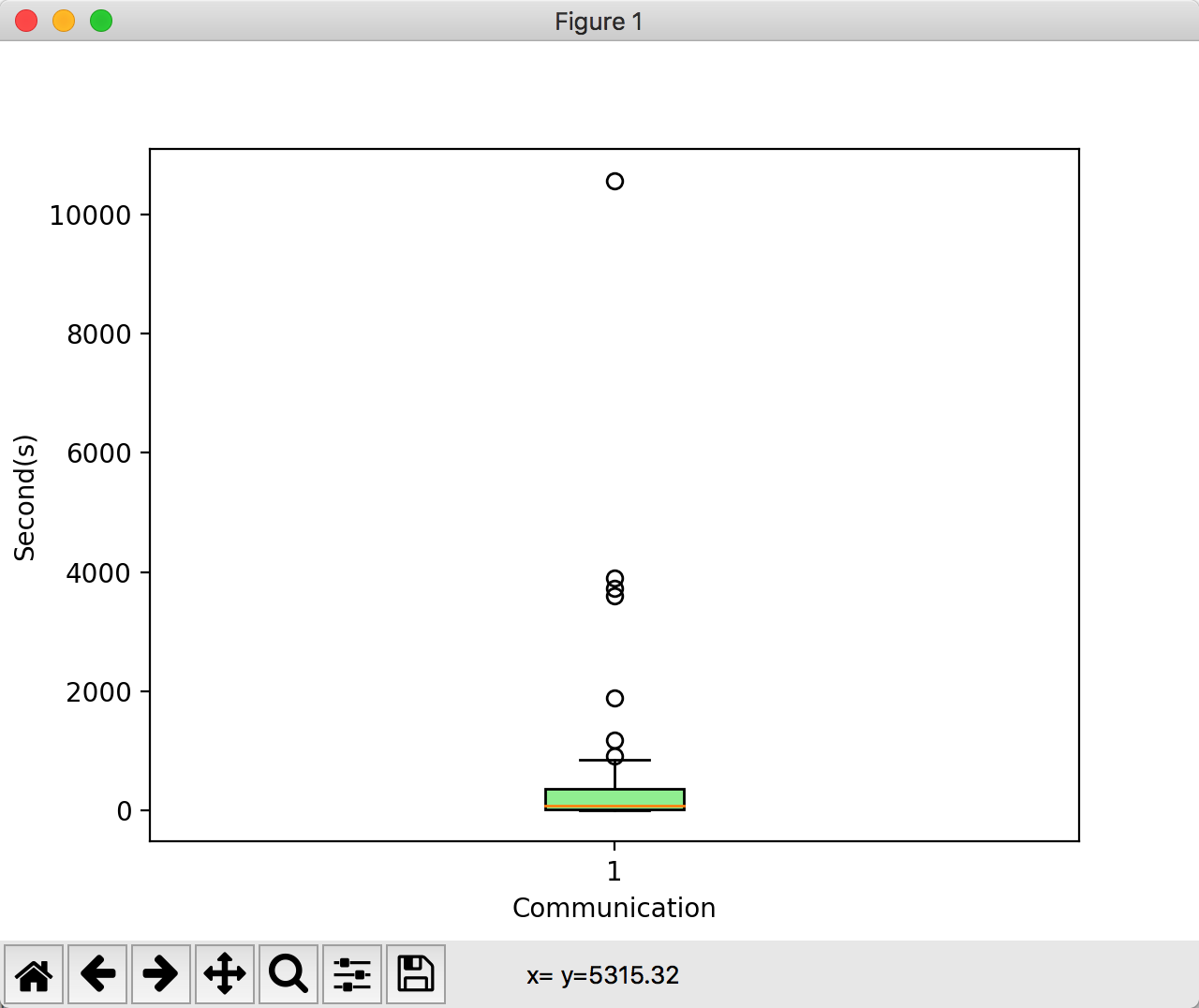


负责计算统计量结果的文件为 /代码/utils/cal.py。

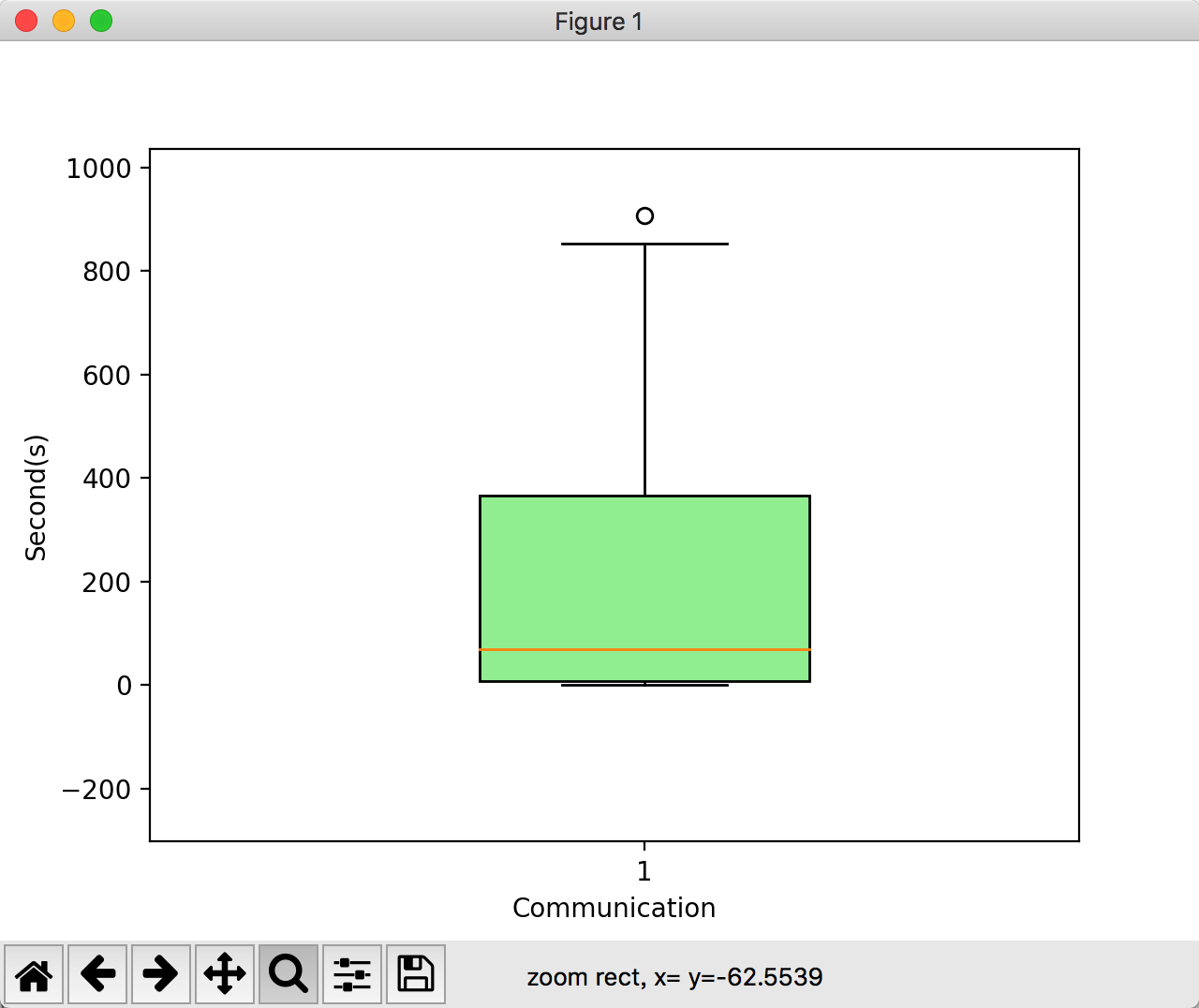




计算统计量后，我们对通话时间属性进行盒图可视化：



如图所示，我个人的通话记录有几个通话时间明显偏多的离散点，在图中均得到了显示。我们也可以放大局部，观察盒图盒子附近的结果。



可以看出盒图是根据五数概括画出，同时也根据中间四分位数极差得出了偏差过大的离散点。在我个人的信息中，离散点均是过大的数据，并没有过小的数据。