Homework 2.1 Report

预处理数据

预处理数据经过了如下步骤的预处理

1. 删除如下列

- 执行套餐
- 实收通信费
- 网络类型

2. 将通信时长统一为秒为单位

3. 增加"亲密性", "性别"属性，并为每行手动填充数据

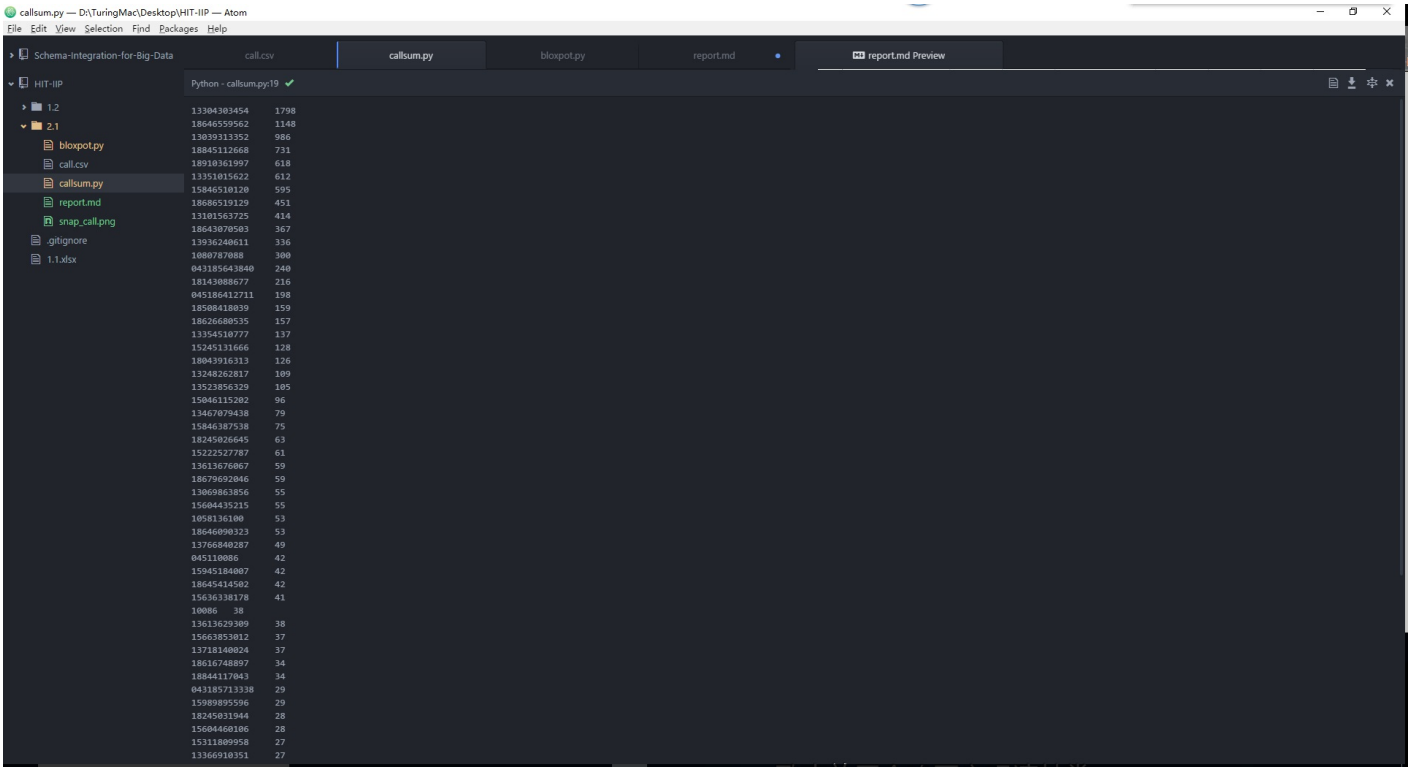
- 亲密性： 1：亲密 -1：不亲密
- 性别： 1：男 0：女

预处理后的部分数据如下

	A	B	C	D	E	F	G	H
1	起始时间	通信地点	通信方式	对方号码	通信时长	通信类型	亲密性	性别
2	2015/12/1 12:45	黑龙江省	被叫	15245131666	76	本地	1	1
3	2015/12/2 15:46	黑龙江省	被叫	15245131666	12	本地	1	1
4	2015/12/3 8:30	黑龙江省	被叫	45186412711	51	本地	-1	1
5	2015/12/3 10:50	黑龙江省	被叫	18761879740	24	本地	-1	1
6	2015/12/4 16:04	黑龙江省	被叫	13351015622	31	本地	1	1
7	2015/12/5 8:47	黑龙江省	被叫	18845112668	4	本地	1	0
8	2015/12/5 9:09	黑龙江省	被叫	18845112668	33	本地	1	0
9	2015/12/5 11:32	黑龙江省	被叫	15846387538	17	本地	-1	1
10	2015/12/5 11:34	黑龙江省	被叫	15846387538	58	本地	-1	1
11	2015/12/5 12:15	黑龙江省	被叫	13059020460	9	本地	-1	1
12	2015/12/5 12:26	黑龙江省	主叫	13101563725	33	本地	1	1
13	2015/12/5 12:45	黑龙江省	被叫	13101563725	19	本地	1	1
14	2015/12/5 15:36	黑龙江省	被叫	13936240611	64	本地	1	1
15	2015/12/6 22:07	黑龙江省	主叫	18845112668	440	本地	1	0
16	2015/12/6 22:15	黑龙江省	主叫	13039313352	791	国内长途	1	0
17	2015/12/9 11:44	黑龙江省	主叫	18679692046	45	国内长途	1	1
18	2015/12/9 12:23	黑龙江省	被叫	15846610908	11	本地	-1	1
19	2015/12/12 14:41	黑龙江省	被叫	13936240611	23	本地	1	1
20	2015/12/12 19:07	黑龙江省	被叫	17745620433	10	本地	-1	0
21	2015/12/13 20:10	黑龙江省	主叫	13351015622	15	本地	1	1
22	2015/12/13 21:02	黑龙江省	主叫	18245031944	28	本地	1	1
23	2015/12/15 9:54	黑龙江省	被叫	18646090323	12	本地	1	0
24	2015/12/15 12:19	黑龙江省	被叫	15344500278	20	本地	-1	1
25	2015/12/15 18:48	黑龙江省	被叫	13101563725	47	本地	1	1
26	2015/12/17 10:27	黑龙江省	被叫	13101563725	58	本地	1	1
27	2015/12/17 10:42	黑龙江省	主叫	13523856329	52	本地	1	1
28	2015/12/17 10:51	黑龙江省	被叫	13523856329	53	本地	1	1
29	2015/12/17 12:16	黑龙江省	被叫	13101563725	181	本地	1	1
30	2015/12/18 12:11	黑龙江省	主叫	18508418039	129	国内长途	1	1
31	2015/12/18 13:40	黑龙江省	主叫	18679692046	14	国内长途	1	1
32	2015/12/18 17:20	黑龙江省	被叫	18645414502	28	本地	1	1
33	2015/12/18 19:15	黑龙江省	被叫	18508418039	30	本地	1	1
34	2015/12/18 19:28	黑龙江省	被叫	15636338178	41	本地	1	1
35	2015/12/21 18:10	黑龙江省	被叫	17745620433	12	本地	-1	0
36	2015/12/22 21:41	黑龙江省	被叫	18686519129	451	本地	1	1
37	2015/12/24 9:53	黑龙江省	被叫	13936240611	76	本地	1	1
38	2015/12/24 10:46	黑龙江省	被叫	18910361997	207	本地	1	1
39	2015/12/25 14:30	黑龙江省	被叫	1058136108	12	本地	-1	0
40	2015/12/25 18:14	黑龙江省	被叫	18910361997	58	本地	1	1
41	2015/12/26 11:38	黑龙江省	被叫	13304303454	158	本地	1	1
42	2015/12/26 12:08	黑龙江省	被叫	13304303454	51	本地	1	1

统计每个号码通话时间

将csv数据使用python读入，并以电话号码作为key存入字典中，将多次通话时长累加并排序，得到通话时间的统计。



绘制盒图

使用python在matplotlib库的基础上对通话时间这一数据进行盒图绘制。

求统计量

首先将通话时间读入并存入数组、排序，并计算如下统计量：

- avg 均值

```
time_len = len(time)
time_avg = 0.0
for t in time:
  time_avg += t * 1.0 / time_len
```

- median 中位数

```
if length % 2 != 0:
  return l[length/2-1]
else:
  return float(l[length/2-1] + l[length/2]) / 2
```

- mode 众数

```
time_count = {}
for t in time:
  time_count[t] = time_count.get(t,0) + 1
sorted_count=sorted(time_count.items(), key = itemgetter(1), reverse = True)
time_mode = sorted_count[0][0]
```

- min 最小值

```
time.sort()
time_min = time[0]
```

- max 最大值

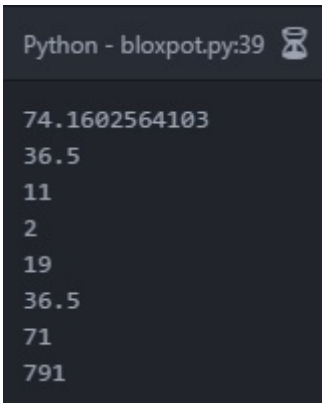
```
time.sort()
time_q1 = time[int(time_len * 0.25)]
```

- q1 四分位数：第25个百分位数

```
time_q1 = time[int(time_len * 0.25)]
```

- q3 四分位数：第75个百分位数

```
time_q1 = time[int(time_len * 0.75)]
```



绘制盒图

使用matplotlib库绘制坐标轴，axvline和axhline方法绘制盒图的直线，plot方法绘制离群点，并对图像进行额外的修饰。

盒图如下