数据预处理与统计分析 ¶

导入数据

In [27]:

```
import pandas as pd
f = open('附件1.csv')
data1 = pd.read_csv(f)
f.close()
print(data1.shape)
data1.head()
```

(70680, 9)

Out[27]:

	订单号	设备ID	应付金额	实际金额	商品	支付时间	地点	状态	提现
0	DD20170816749366361849990978 4	E43A6E078A07631	4.5	4.5	68g好丽友巧克力派2枚	2017/1/1 0:53	D	已出货未退款	已提现
1	DD20170816749366355581406116 4	E43A6E078A04172	3.0	3.0	40g双汇玉米热狗肠	2017/1/1 1:33	A	已出货未退款	已提现
2	DD20170816749357852689093988 6	E43A6E078A06874	5.5	5.5	430g泰奇八宝粥	2017/1/1 8:45	E	已出货未退款	已提现
3	DD20170816749368350718661583 7	E43A6E078A04228	5.0	5.0	48g好丽友薯愿香烤原味	2017/1/1 9:05	C	已出货未退款	已提现
4	DD20170816749375954861825200 6	E43A6E078A04134	3.0	3.0	600ml可口可乐	2017/1/1 9:41	B	已出货未退款	已提现

In [33]:

```
ff = open('附件2.csv')
data2 = pd.read_csv(ff)
ff.close()
print(data2.shape)
data2.head(10)
```

(315, 3)

Out[33]:

	商品	大类	二级类
0	100g*5瓶益力多	饮料	乳制品
1	100g越南LIPO奶味面包干	非饮料	饼干糕点
2	10g卫龙亲嘴烧香辣味	非饮料	肉干/豆制品/蛋
3	10g越南LIPO奶味面包干	非饮料	饼干糕点
4	110g顺宝九制话梅	非饮料	蜜饯/果干
5	120g达利园蔓越莓提子面包	非饮料	饼干糕点
6	12g劲仔小鱼（香辣味）	非饮料	肉干/豆制品/蛋
7	12g劲仔小鱼麻辣味	非饮料	肉干/豆制品/蛋
8	13g无穷烤鸡小腿蜂蜜味	非饮料	肉干/豆制品/蛋
9	145ml旺仔牛奶罐装	饮料	乳制品

查看应付金额和实付金额的汇总统计

In [2]:

```
data1.describe()
```

Out[2]:

	应付金额	实际金额
count	70680.000000	70680.000000
mean	4.060324	4.060324
std	3.357908	3.357908
min	0.000000	0.000000
25%	3.000000	3.000000
50%	3.500000	3.500000
75%	4.500000	4.500000
max	125.000000	125.000000

统计应付金额为0的订单数量，即免单数量

In [25]:

```
data1[data1.应付金额.values==0].count()
```

Out[25]:

```
订单号       202
设备ID      202
应付金额       202
实际金额       202
商品         202
支付时间       202
地点         202
状态         202
提现         202
dtype: int64
```

统计商品种类数量

In [26]:

```
data1.商品.nunique()
```

Out[26]:

303

统计各个地点的销售总额

In [32]:

```
data1[['实际金额','地点']].groupby('地点').sum()
```

Out[32]:

地点	实际金额
A	42542.6
B	53970.3
C	61572.1
D	33243.3
E	95655.4

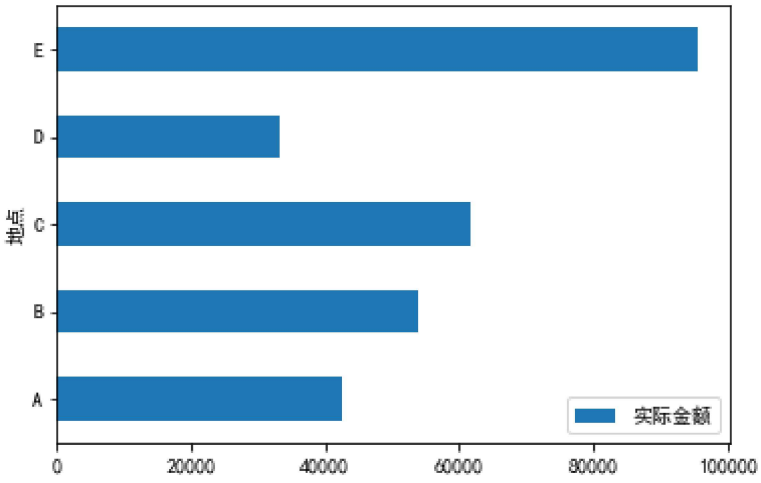
数据的可视化

In [36]:

```
import matplotlib.pyplot as plt
import matplotlib
matplotlib.rcParams['font.sans-serif']=['SimHei']   # 用黑体显示中文
matplotlib.rcParams['axes.unicode_minus']=False     # 正常显示负号
```

In [45]:

```
data = data1[['实际金额','地点']].groupby('地点').sum()
hist = data.plot(kind='barh')
```



地区画像

In [ ]:

销售额的预测

In [ ]:

撰写报告和提交文件

In [ ]: