**石家庄铁道大学信息科学与技术学院**

**实习报告**

**2019年----2020年 第一学期**

专 业： 软件工程

课程名称： 软件工程综合实践

班 级：

姓 名： 学号：

指导教师：

**跨境电商红酒定价决策**

目录

[实验 3-1 数据预处理 2](#_Toc19226438)

[1. 处理数据格式 2](#_Toc19226439)

[2. 去重（去除重复数据） 3](#_Toc19226440)

[3. 过滤非红酒类商品 4](#_Toc19226441)

[4. 过滤非 750ml 红酒数据并处理价格 4](#_Toc19226442)

[5. 函数考虑 float 类型是为了函数复用，因为之后的处理过程仍需要该函数。 4](#_Toc19226443)

[6. 过滤无用属性 4](#_Toc19226444)

[实验 3-2 处理空值的属性数据 5](#_Toc19226445)

[实验 3-3 利用 pandaBI 查看数据各个维度的分布 6](#_Toc19226446)

[1. 可视化分析 6](#_Toc19226447)

[2. 处理特征 7](#_Toc19226448)

[实验 3-4 定义红酒的价格区间 8](#_Toc19226449)

[实验 4-1 利用决策树算法构建价格预测模型 9](#_Toc19226450)

[1. 读取数据 9](#_Toc19226451)

[2. 平衡数据 9](#_Toc19226452)

[3. 处理离散型特征 11](#_Toc19226453)

[4. 划分测试集训练集 11](#_Toc19226454)

[实验 4-2 基于模型辅助海关决策生成监测报表 13](#_Toc19226455)

[1. 基于模型预测 13](#_Toc19226456)

[实验 5-1 清洗微博数据 14](#_Toc19226457)

[1. 读取数据 14](#_Toc19226458)

[2. 处理性别 15](#_Toc19226459)

[3. 处理年龄 15](#_Toc19226460)

[4. 处理地区 16](#_Toc19226461)

[实验 5-2 利用聚类发现同级红酒 17](#_Toc19226462)

[1. 读取之前为微博数据分析准备的数据 17](#_Toc19226463)

[2. one-hot 编码 17](#_Toc19226464)

[3. 聚类 17](#_Toc19226465)

[4. 同级红酒发现 17](#_Toc19226466)

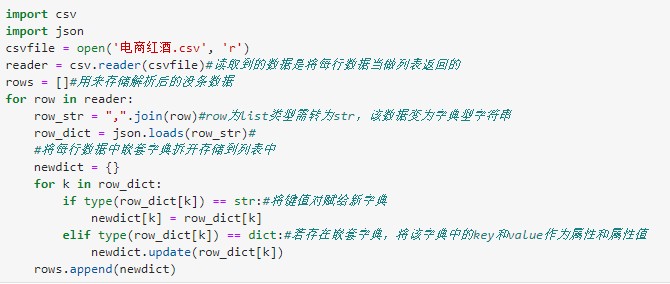
[5. 给微博数据加标签 20](#_Toc19226467)

[实验 5-3 利用 pandaBI 构建客户群体各维度分布图 22](#_Toc19226468)

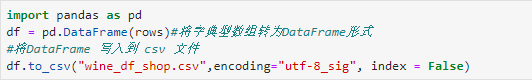
# 实验 3-1 数据预处理

1、处理数据格式

首先利用 json.loads 解决爬取数据中的中文编码问题：



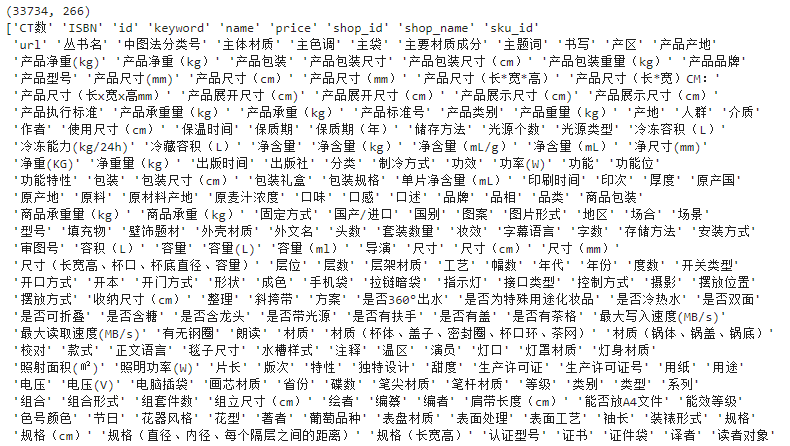
然后利用 pandas 将数据转化为 DataFrame 形式，并保存为csv文件。



最后查看数据量和数据维度。



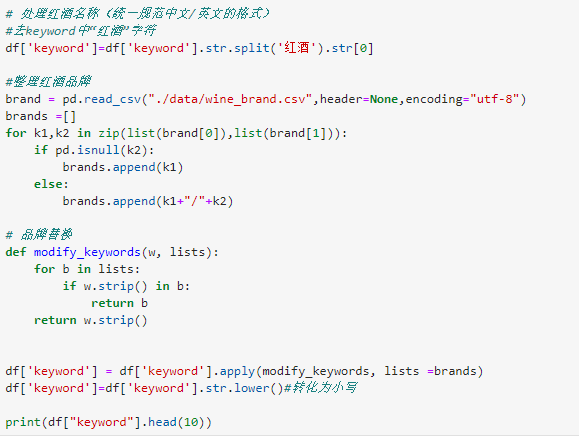
运行结果如下：



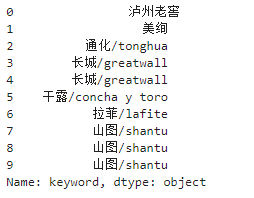
目前关键词字段存储的是"品牌 红酒"（如“拉菲 红酒”，“LAFITE 红酒”），处理流程如下：

首先去掉关键词中的“红酒”字符,只保留红酒品牌。

其次对于红酒品牌存在中英文两种表述的，统一归为形如“拉菲/LAFITE”形式。

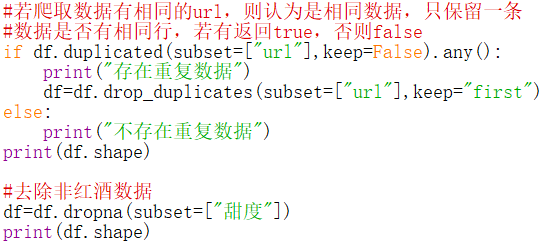


运行结果如下：



2、去重（去除重复数据）

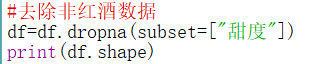
由于红酒品牌搜索时有中文和英文，可能会存在相同数据，根据 url 属性来判断两条数据是否重复。





可见，当前数据中无重复数据，接下来处理删除非红酒类商品。

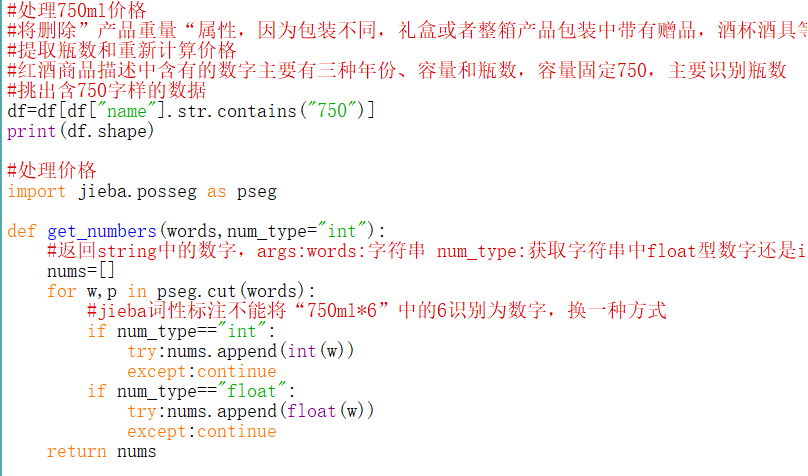
过滤非红酒类商品：发现数据中还存在红酒杯、醒酒器黄酒等噪声数据，而“甜度”是红酒的一般特性，若无该属性则很大程度上是非红酒数据，可能酒具以及黄酒白酒等数据。





当前还剩余 18734 条数据。

1、过滤非 750ml 红酒数据并处理价格

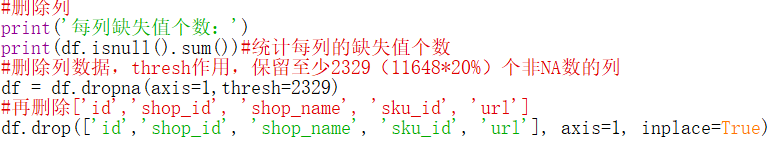


函数考虑 float 类型是为了函数复用，因为之后的处理过程仍需要该函数。

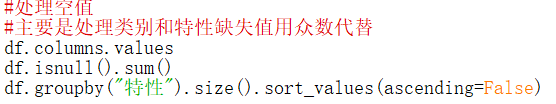


1、过滤无用属性

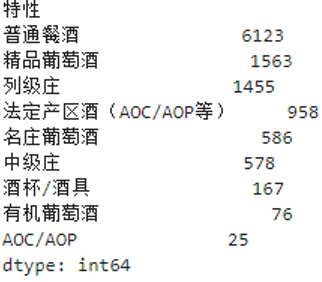
观察每列的缺失值数目，采用删除缺失值超过 80%的列的策略

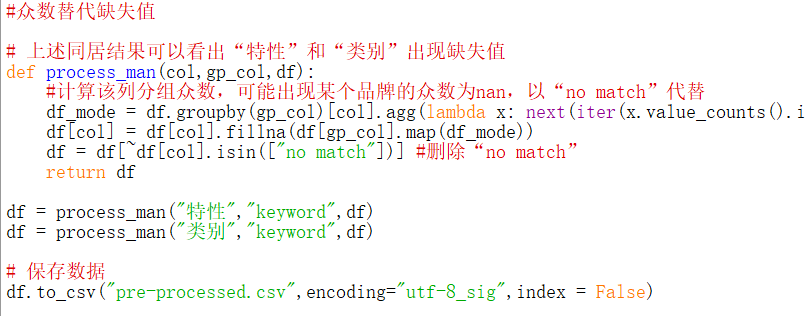


**实验 3-2 处理空值的属性数据**





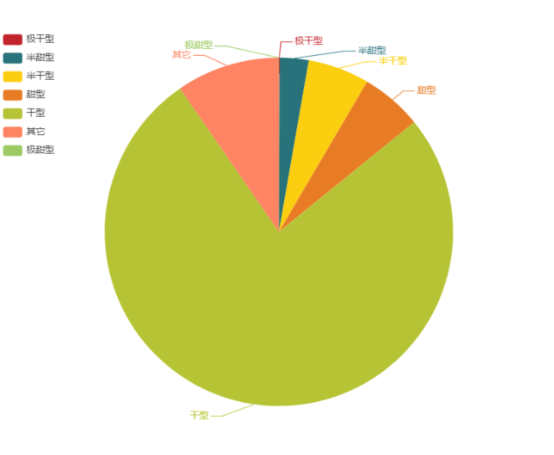




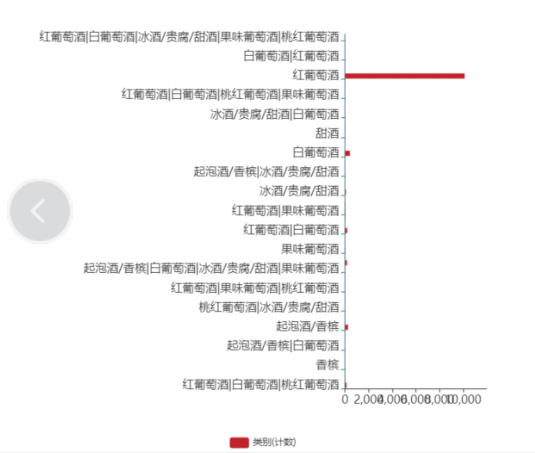
**实验 3-3 利用 pandaBI 查看数据各个维度的分布**

1、可视化分析

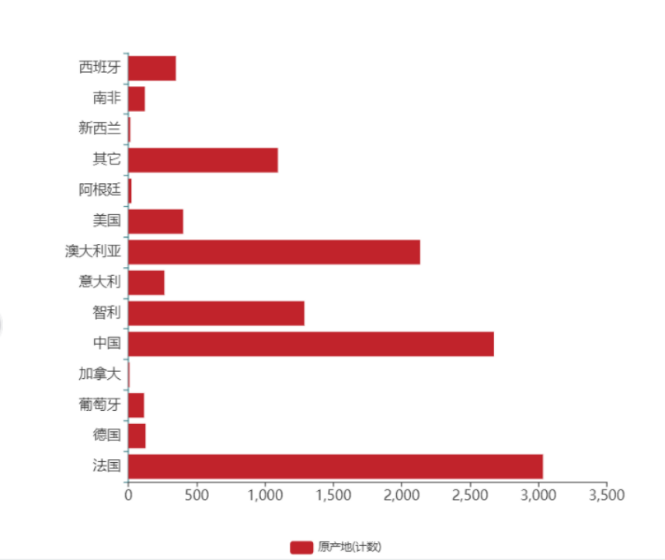
甜度：



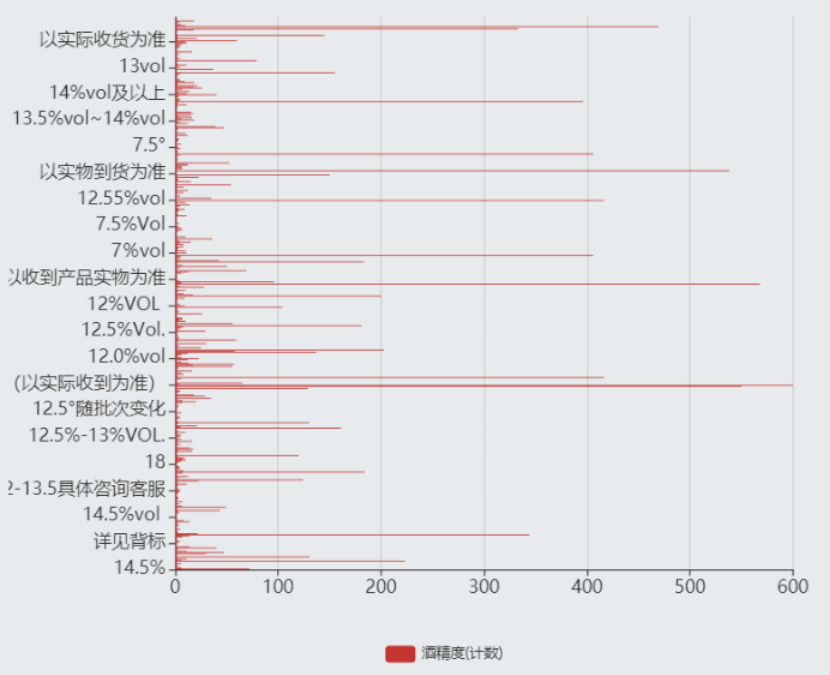
类别：



原产地：



酒精度：

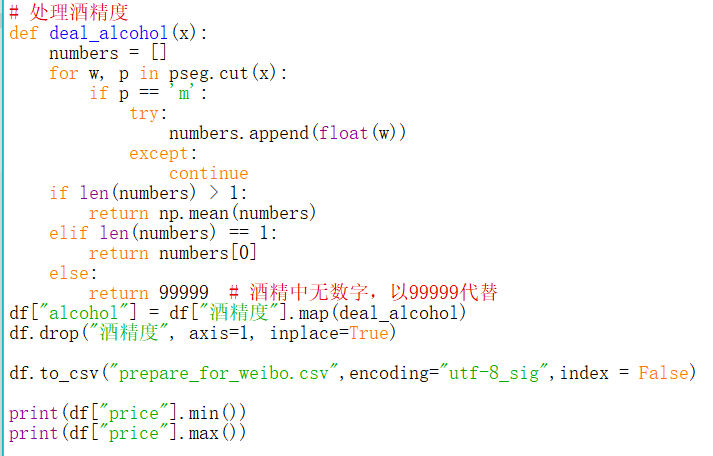


2. 处理特征

（1）处理葡萄品种的操作流程如下：



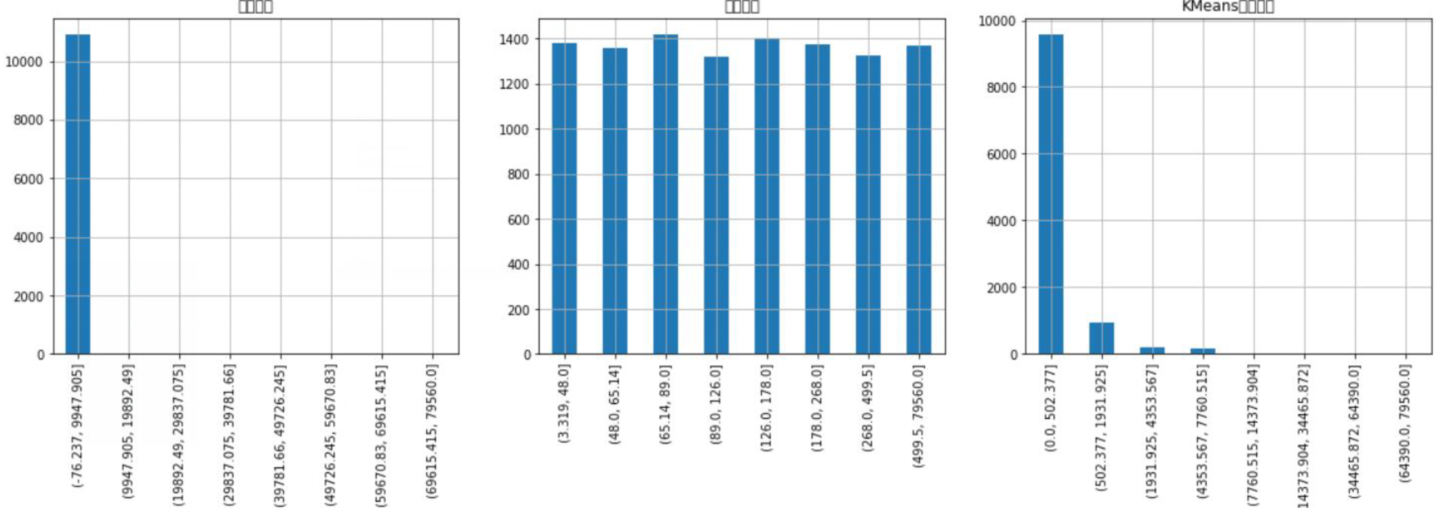
（2）处理酒精度的操作流程如下

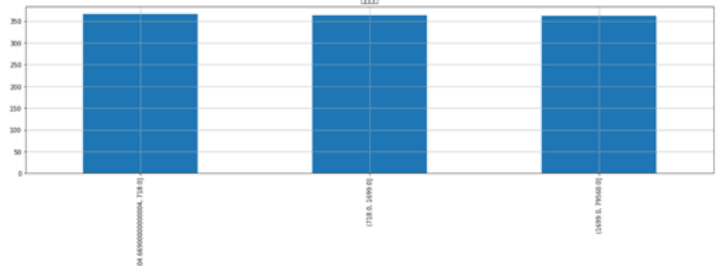
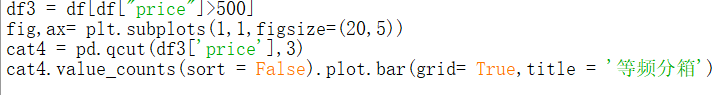


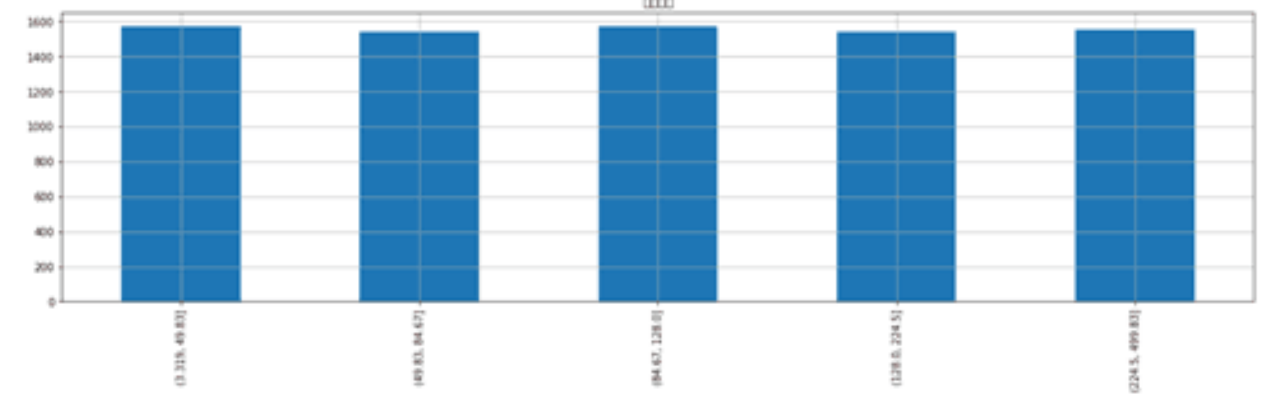
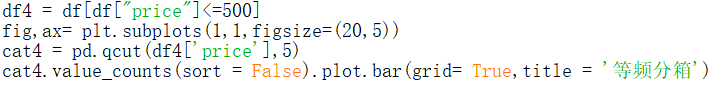
**实验 3-4 定义红酒的价格区间**

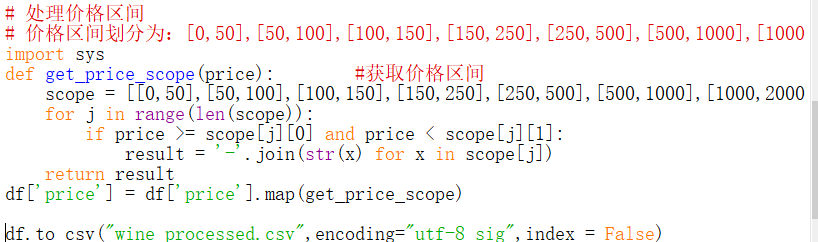
查看数值分布数量的常见方式有三种：等宽分箱、等频分箱和基于聚类的分箱。







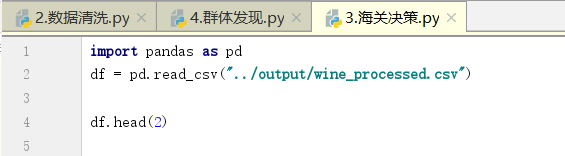






**实验 4-1 利用决策树算法构建价格预测模型**

1、读取数据

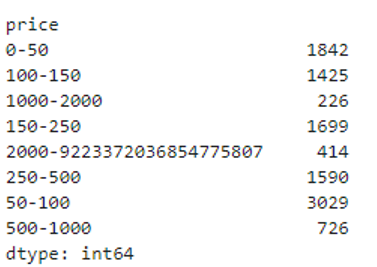




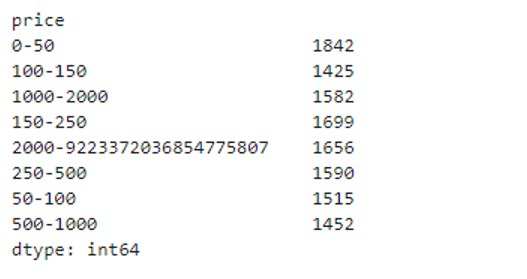
2、平衡数据

观察每个类别的数据，采用欠采样/过采样的方法对数据进行平衡处理。

df.groupby("price").size()



由此可见，[1000-2000],[2000-MAX],[500-1000]的价格区间需要增长一定的量， 通过复制的方式增加该价格区间的数据量；[50-100]的数据量较多，欠采样该价格区间的数据。处理代码如下：



从结果看，各个价格区间的数据量持平。

3、处理离散型特征

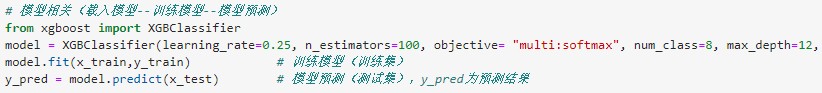
算法不允许“绿色”这种非数值型特征，需要做转换处理。



## 划分测试集训练集

## 

5、模型构建



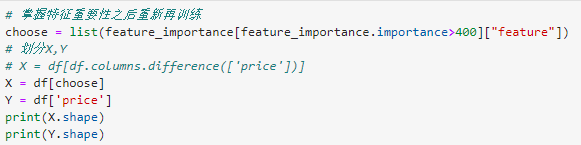
6、模型效果

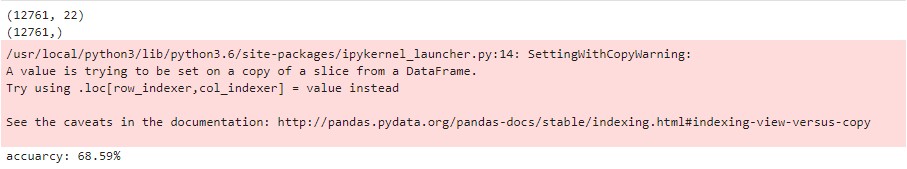


7、特征选择与再训练

首先查看模型各个维度的重要性：



掌握特征的重要性排列之后，我们可以选择部分特征再做训练，对比之前的效果，结果如下：



**实验 4-2 基于模型辅助海关决策生成监测报表**

1. 基于模型预测

本节讲述的是如何拿现有模型对单条数据进行预测，有此基础就可以对海关的

红酒数据进行价格区间预测，跟报关单上的价格做对比审核。



读取模型，并对数据进行预测

注意此处需要对数据进行转换，利用上一个小节保存下来的 encoders.dict 文件对数据做相同的转换。



结果如下：

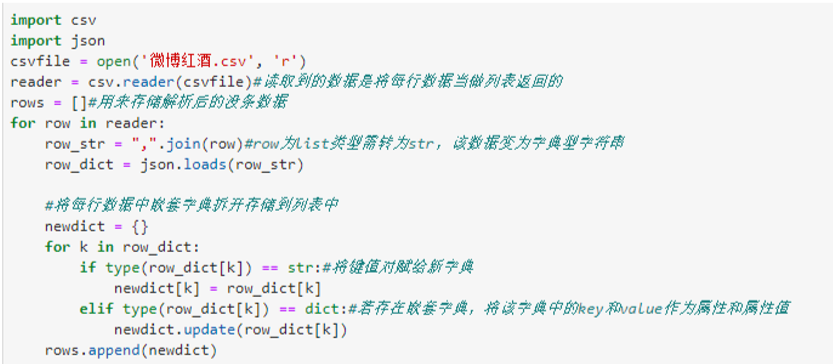


**实验 5-1 清洗微博数据**

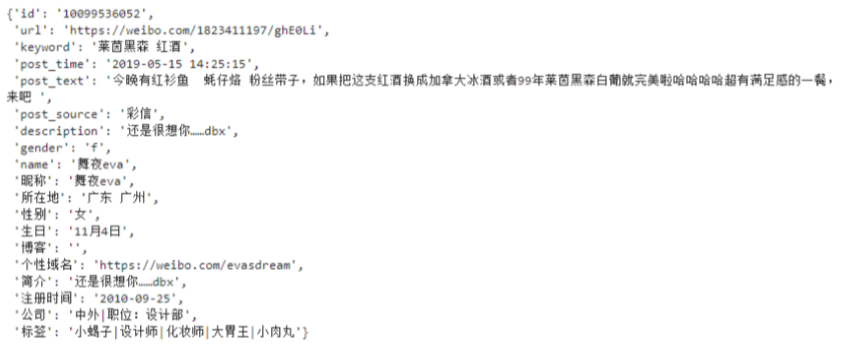
1. 读取数据

读取数据的方式和第二章的读取数据一致，都是采用 json.loads() 函数读成字典类型。

代码如下：



可以得到 24611 条数据，示例如下：



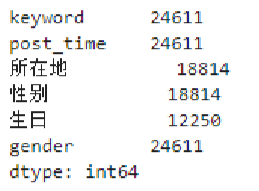
处理搜索关键词，即 keyword 字段，处理红酒品牌的中英文表示，代码同上。



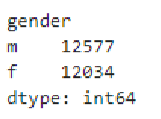
接下来我们整理想要的属性，包含"keyword","post\_time","所在地","性别","生日","gender"这六种属性，保存临时处理文件，用以观察。

2、处理性别

用 df.count 函数查看各维度的非缺失值，结果如下：



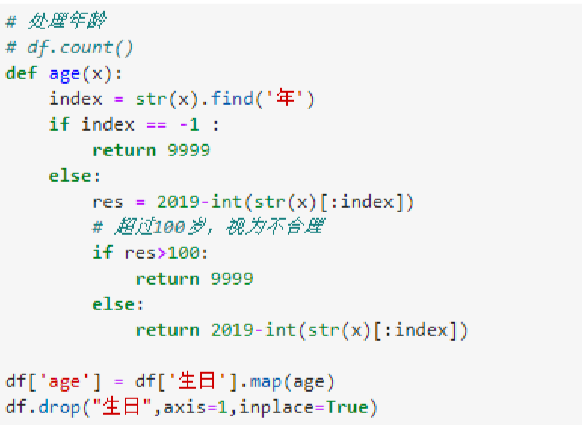
可见 gender 字段比性别可行度高，使用 df.groupby 验证 gender 中是否存在异常值。



即对性别的处理，保留 gender 字段即可。

3、处理年龄

据观察，出生字段有人写的是日期，有人写的是星座，有人写的是年月日，因而采用提取年份前面的数字再与当前年份相减得到年龄值。



结果如下所示：



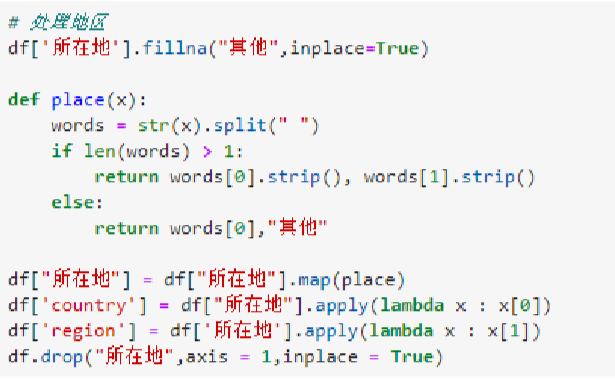


4. 处理地区

据观察，所在地的取值有三种，分别是其他，大地区，大地区加小地区

处理方法： 替代缺失值为其他，延伸出两维数据，分别表示为大地区和小地区，没有小地区则表示为其他。

代码如下：

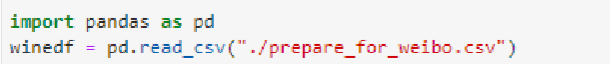


结果如下所示：



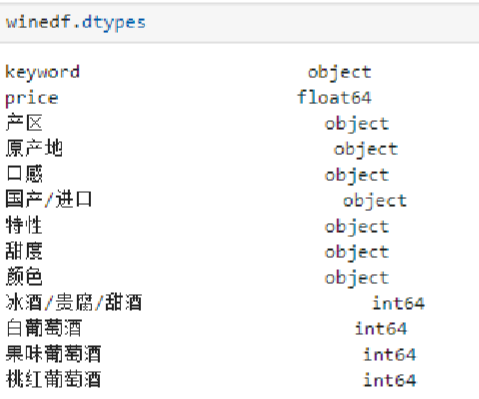
**实验 5-2 利用聚类发现同级红酒**

1、读取之前为微博数据分析准备的数据



2、one-hot 编码

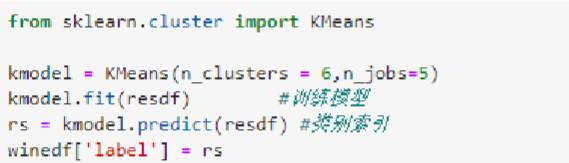
查看 winedf 的 dtypes：



由于 kmeans 接受数值型数据，所以需要采用 one-hot 编码对 object 类别的数据进行转换。



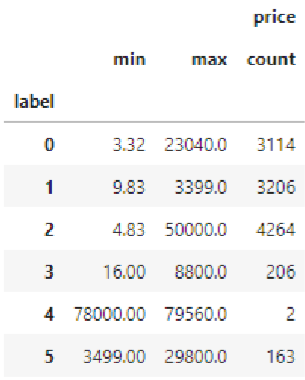
3、聚类



4、同级红酒发现

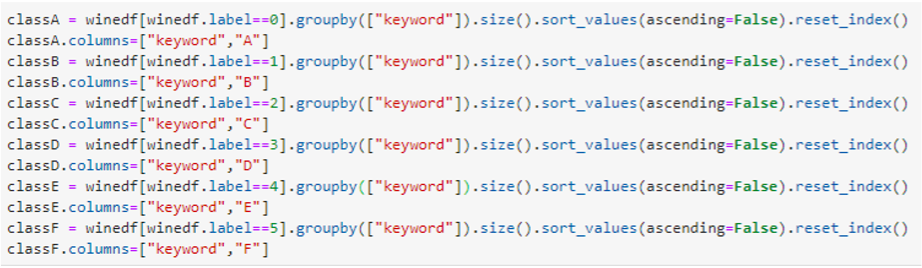
看一下每个类别中数目较多的红酒类别：





每个类中的价格最小值和最大值，可见分类界限没那么明确，造成这样的原因。主要是两部分：一是聚类数目没有选择好，二是数据的准确度，之前的红酒价格是根据数据规范算出来的，但有可能提取的瓶数是错的，导致数据是不正确的，所以聚类上对品牌会造成影响。

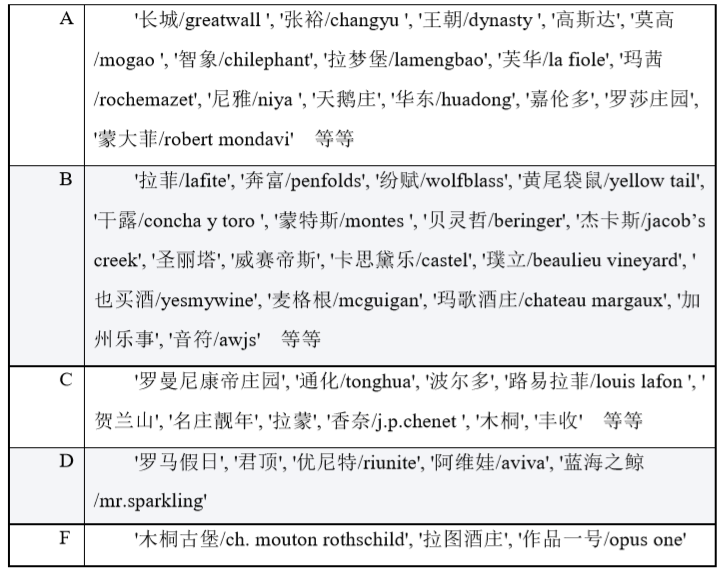
但我们可以查看一下每个类中数目占比就多的红酒品牌：



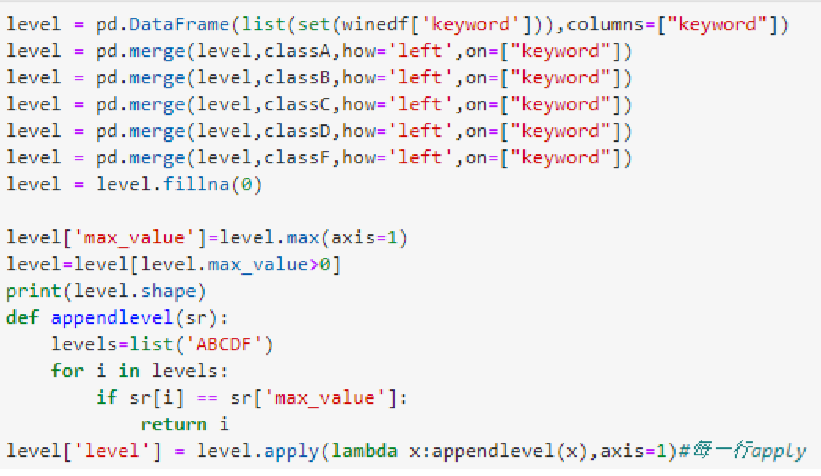


从结果上看，同样的红酒品牌会在多个类中出现，可以由数量上判断该品牌属

于哪个类别，由此整理得到：



代码如下：







5. 给微博数据加标签

把清洗微博数据的结果和同级红酒的处理结果做合并



结果如下：



# 删除level为空的数据result.dropna(axis='index',inplace=True)  
print(result.shape)

#保存为 level.csv 文件

result.to\_csv("../output/level.csv",encoding="utf-8\_sig",index = False)



# 实验 5-3 利用 pandaBI 构建客户群体各维度分布图

