航空公司客户价值分析

## 挖掘背景：

略

## 挖掘目标：

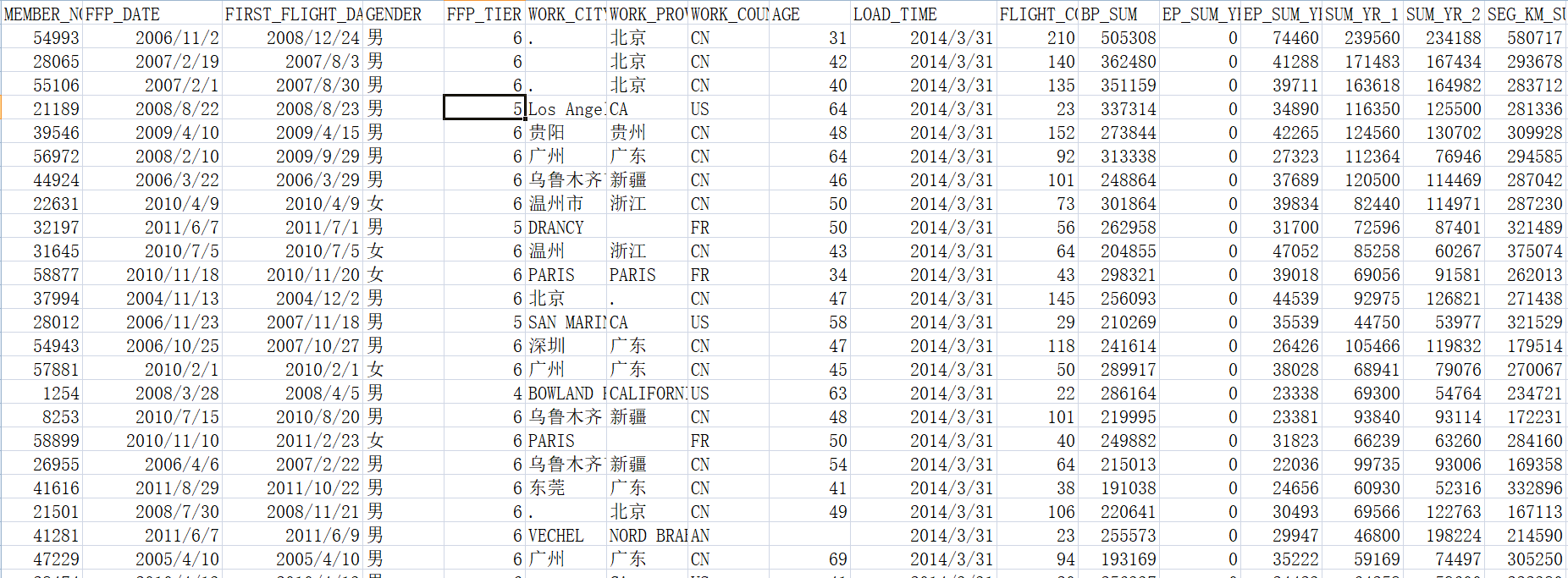
1. 借助航空公司客户数据，对客户进行分类；
2. 对不同的客户类别进行特征分析，比较不同类客户的客户价值；
3. 对不同价值的客户类别提供个性化服务，制定相应的营销策略；

## 分析方法与过程

识别客户价值应用最广泛的模型是通过三个指标（最近消费时间间隔（Recency）、消费频率（Frequency）和消费金额（Monetary））来进行客户细分，识别出高价值的客户，简称RFM模型。本案例中，增加两个指标L（会员入会时间距观测窗口结束的月数）和C（客户在观测窗口内乘坐舱位所对应的折扣系数的平均值），记为LRFMC模型。

1. 数据抽取

部分原始数据如下所示：

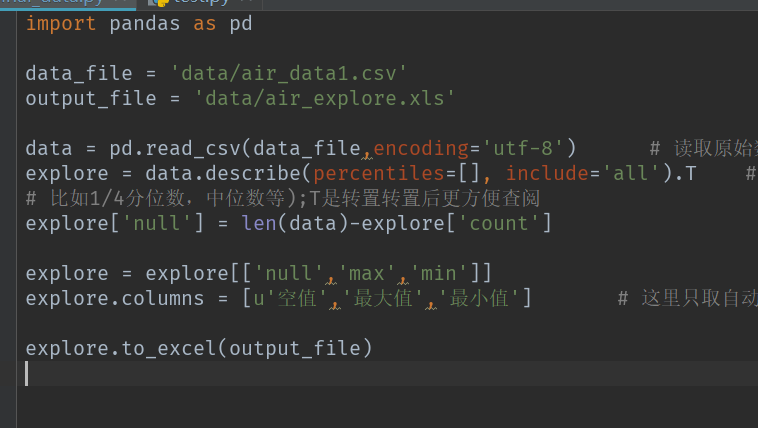


数据文件为data/air\_data.csv

1. 数据探索分析

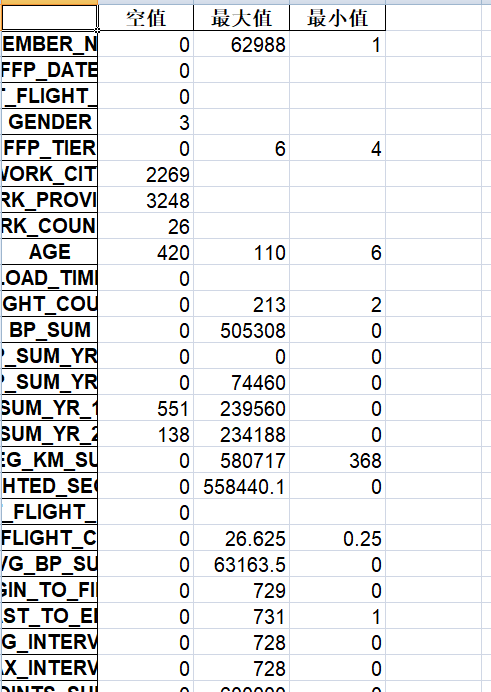
对数据进行缺失值分析与异常值分析，对每一列的数据查找最大值、最小值和空值的查找

代码如下所示：



代码详细见original\_data.py

输出的部分数据为：



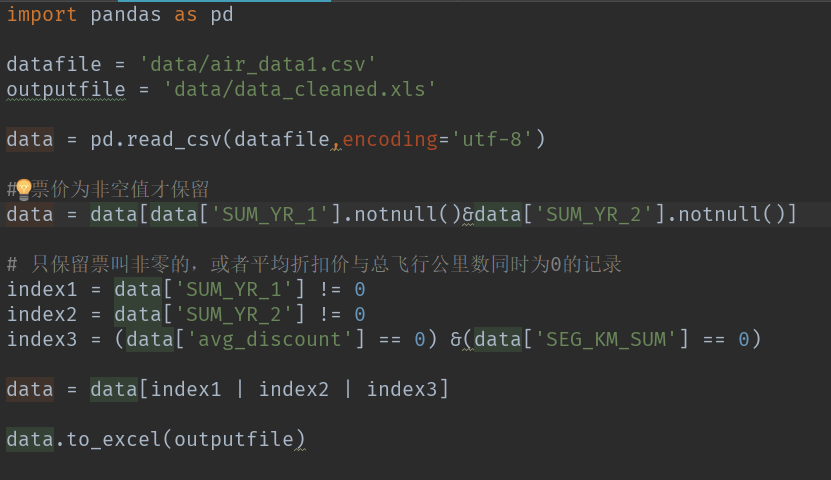
详细数据见data/ari\_explore.xls

1. 数据预处理

本例主要采用数据清洗、属性规约和数据变换的预处理方法。

* 1. 数据清洗
     1. 丢弃票价为空的记录。
     2. 丢弃票价为0、平均折扣率不为0、总飞行公里数大于0的记录（即只保留票叫非零的，或者平均折扣价与总飞行公里数同时为0的记录）。

数据清洗的代码如下：



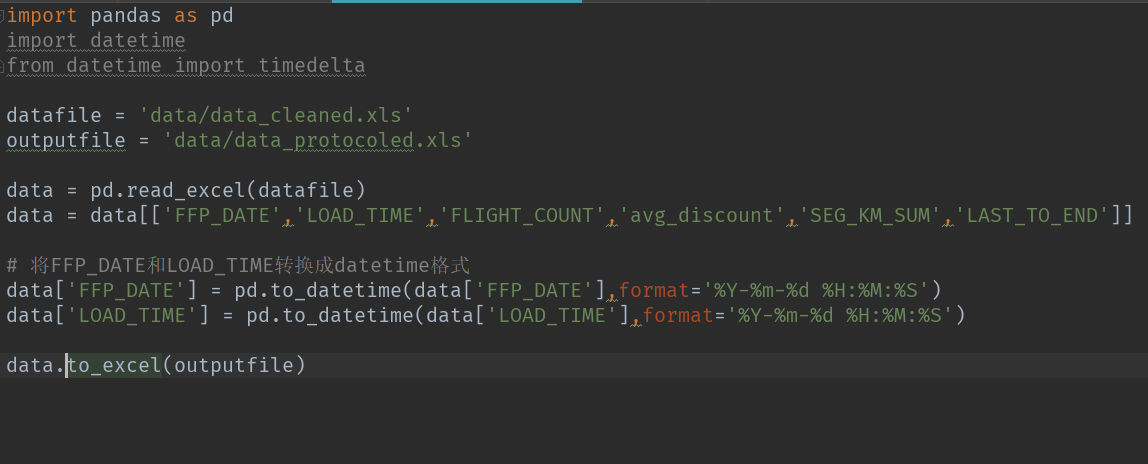
代码详细见data\_cleaning.py

输出的文件为data/data\_cleaning.xls

* 1. 属性规约

原始数据太多，则选择与LRFMC模型中的LRFMC指标相关的六个属性：FFP\_DATA（入会时间）、LOAD\_TIME（观测窗口的结束时间）、FLIGHT\_COUNT（观测窗口内的飞行次数）、AVG\_DISCOUNT（平均折扣率）、SEG\_KM\_SUM（观测窗口的总飞行公里数）、LAST\_TO\_END（最后一次乘机时间至观测窗口结束时长）。

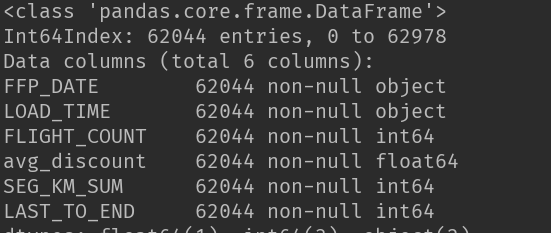
数据规约的代码如下：



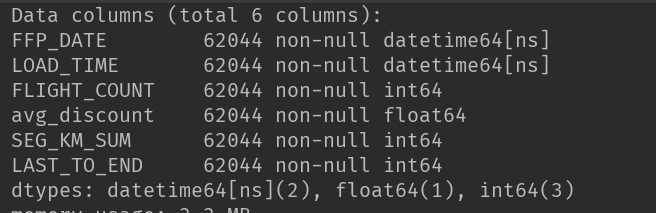
代码详细见attribute\_protocol.py

需要注意的是要把日期格式转换成datatime格式，才能在数据变换是进行相减操作

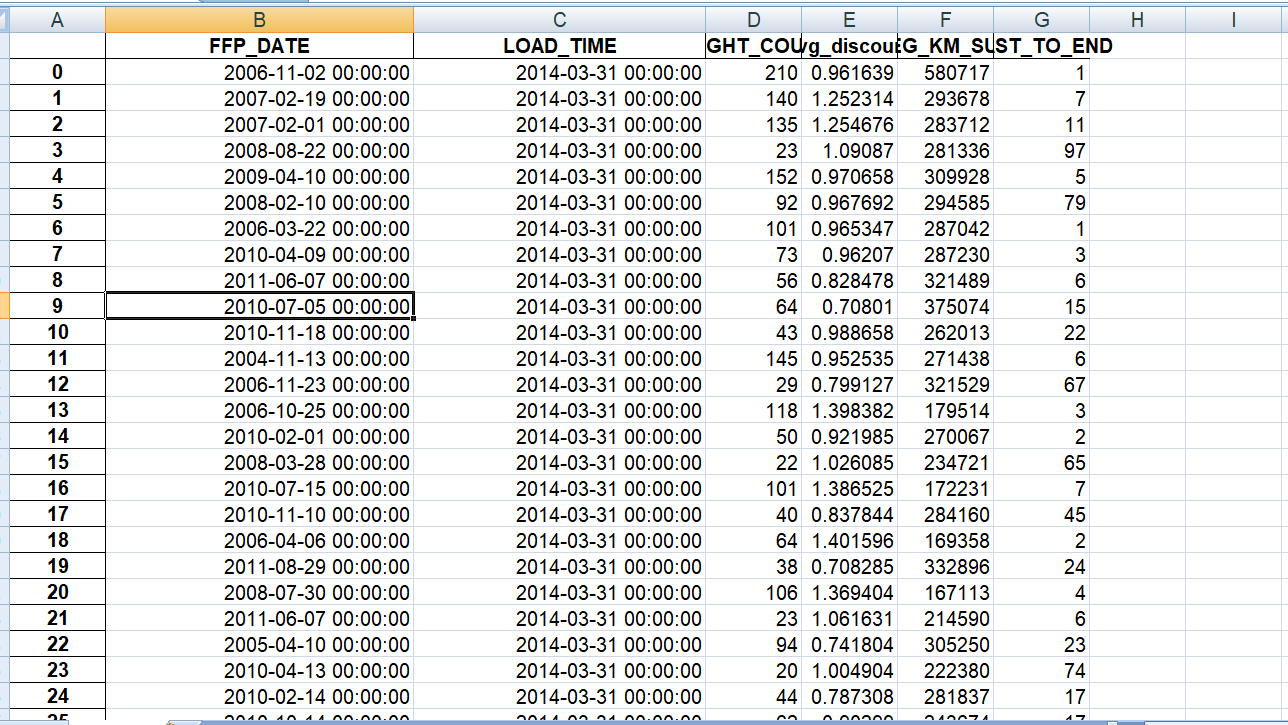
可以用data.info()函数来查看各属性的类型，类型转变前输出的类型如下图所示：



修改过后的个属性的类型为：



输出的部分数据如下：



详细数据文件为data/data\_protocoled.xls

* 1. 数据变换和标准化

通过规约后的属性求得LRFMC模型的5个指标

* + 1. L = LOAD\_TIME – FFP\_DATE

会员入会时间距观测窗口结束的天数 = 观测窗口结束时间-入会时间【单位：天】

* + 1. R = LAST\_TO\_END

客户最近一次乘坐公司飞机据观测窗口结束的月数 = 最后一次乘机时间至观察窗口末端时长【单位：月】

* + 1. F = FLIGHT\_COUNT

客户在观测窗口内乘坐公司飞机的次数=观测窗口的飞行次数【单位：次】

* + 1. M = SEG\_KM\_SUM

客户在观测时间内在公司累计的飞行里程 = 观测窗口的总飞行公里数【单位：公里】

* + 1. AVG\_DISCOUNT

客户在观测时间内乘坐舱位所对应的折扣系数的平均值 = 平均折扣率【单位：无】

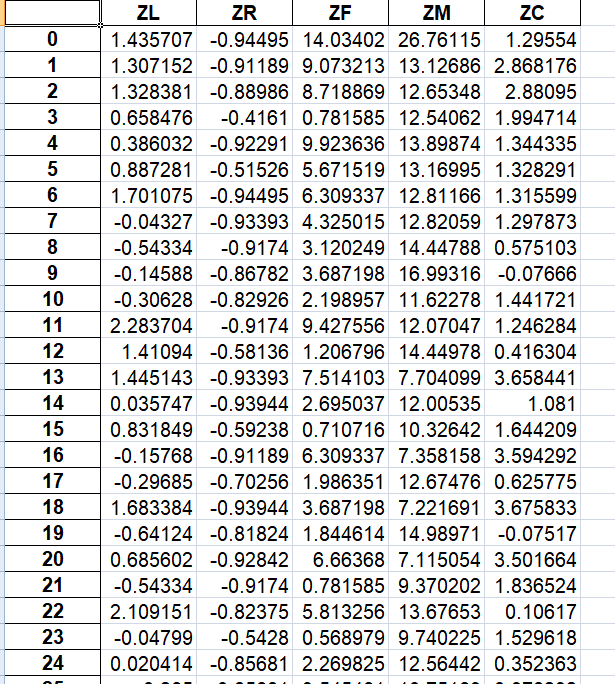
然后对变换后的数据进行标准化处理，列名都加了一个Z

数据变换和标准化的代码如下图所示：



详细代码见data\_changing.py

部分输出文件如下图所示：

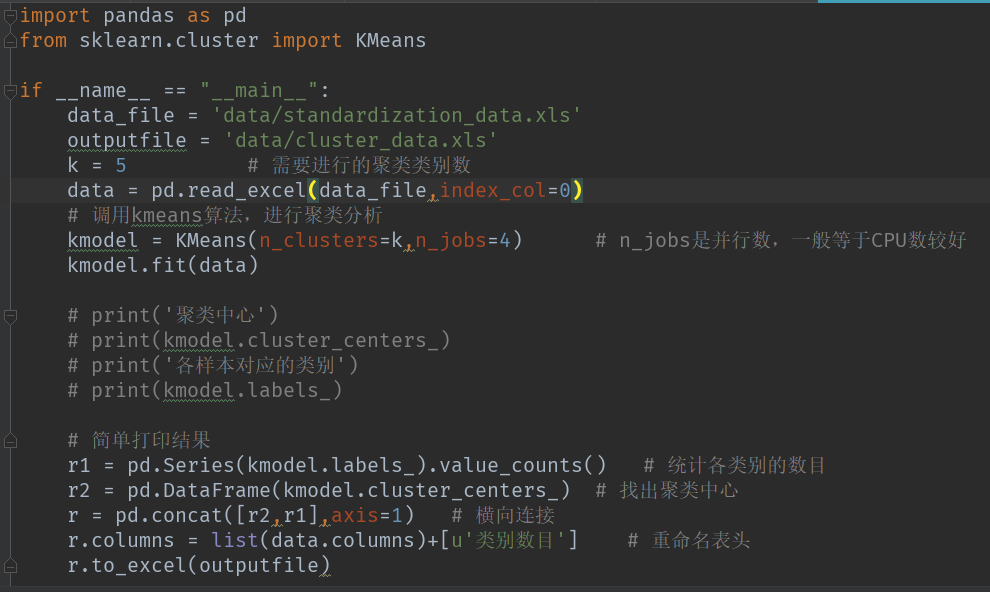


数据文件为data/standardization.xls

1. 模型建立

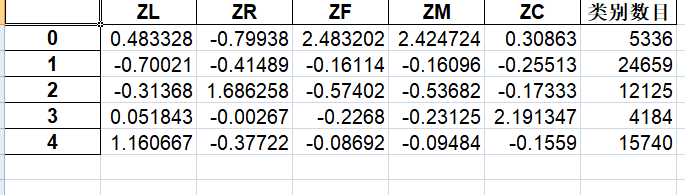
客户聚类，用K-Means聚类算法对客户数据进行客户分群，聚成5类（需要结合业务的理解与分析来确定客户的类别数量）

代码如下所示：



详细代码见K-Means\_cluster.py

聚类后的数据如下所示



数据详见data/cluster\_data.xls