# 实例：用聚类技术对顾客进行分群

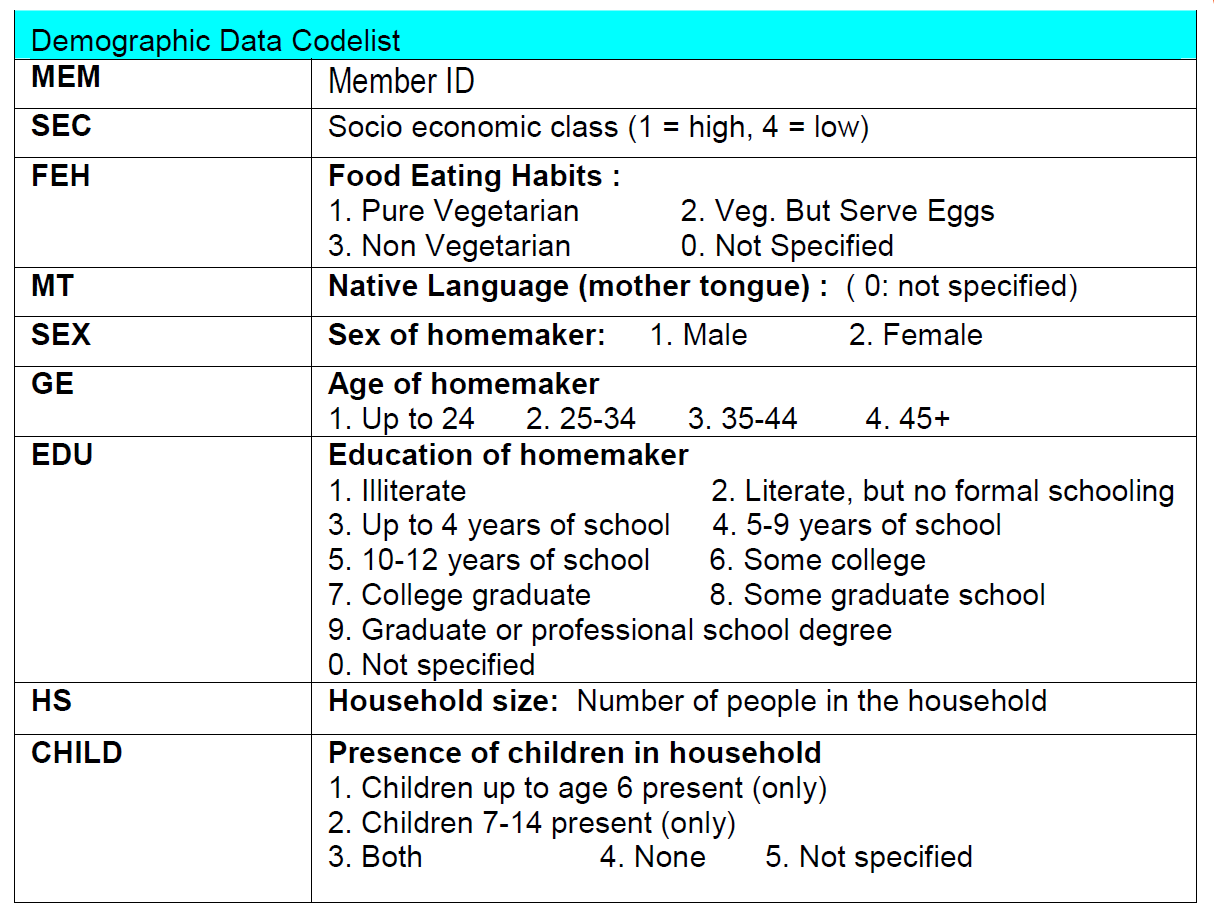
## 背景和概要说明

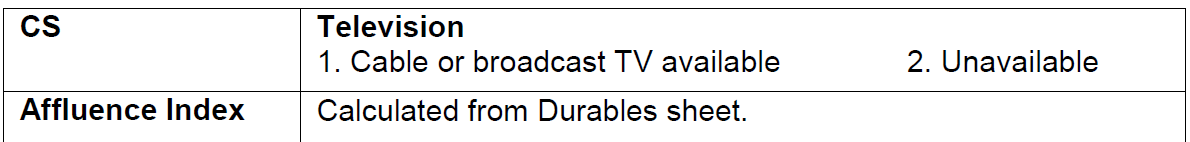
CRISA是一家亚洲市场研究机构，其专注于对顾客的消费行为（商品包括耐用品和非耐用品）进行跟踪分析。在一个研究项目中，商品方面CRISA跟踪30个产品大类（例如洗涤剂），每个大类有60至70种品牌，同时CRISA划定印度105个城镇中的5000个家庭作为待研究的客户群体，覆盖了印度80%的城市市场。其中研究选择的家庭是通过分层抽样（stratified sampling），层级是以社会经济基础和市场的条件为基础进行划分的，以保证抽取的样本能较好的反应总体市场的分布。**CRISA同时拥有交易数据（Transaction Data）和家庭数据（Household Data），其中交易数据每行是一条交易记录，家庭数据每行一个家庭为单位相关的数据**。我们这里主要关注家庭数据，其又可分为如下三大类别：

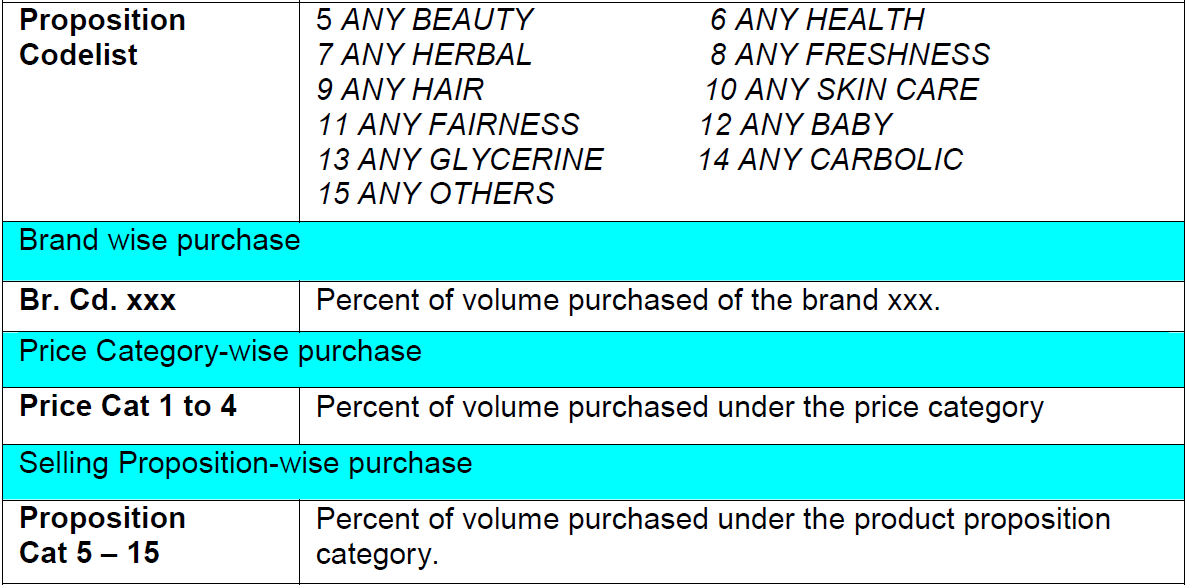
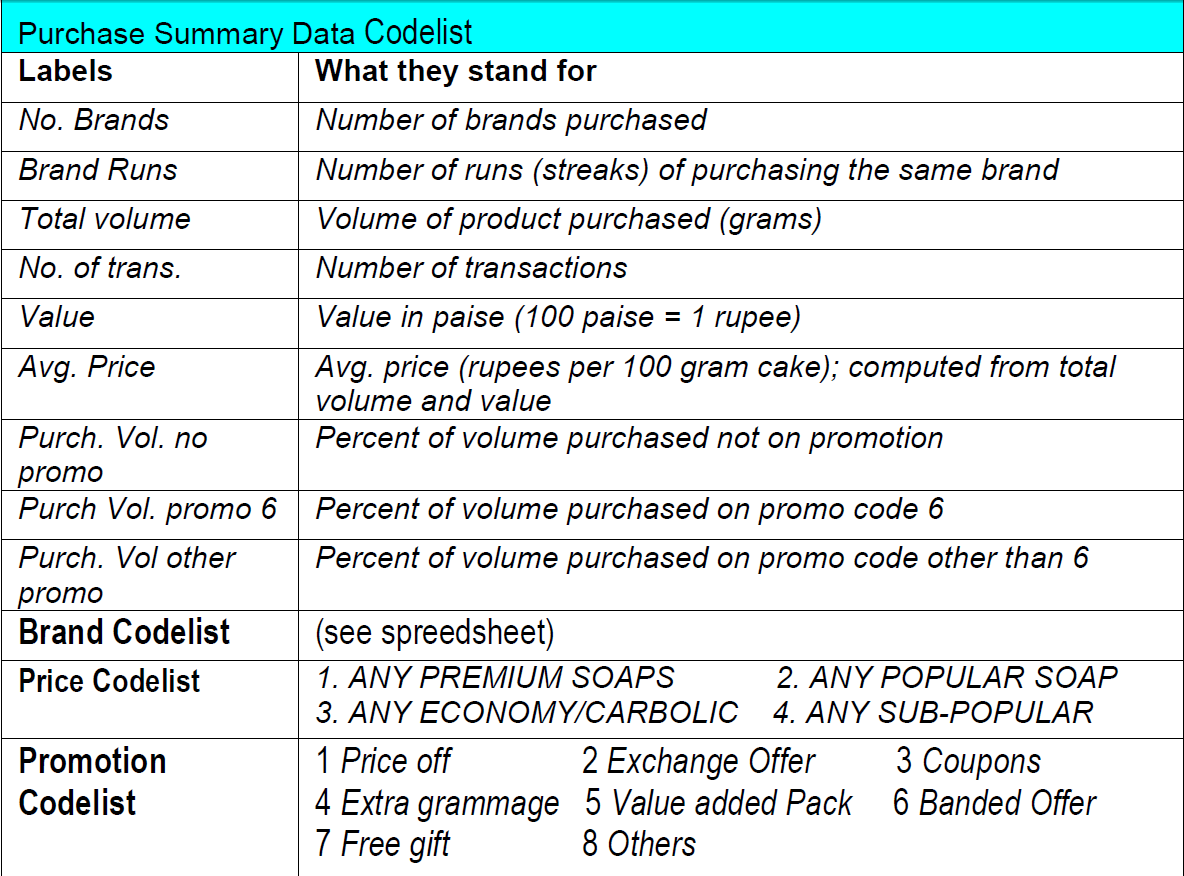
* **家庭的人口统计信息（以年度进行更新）；**
* **耐用品的拥有情况（汽车，洗衣机，等以年度进行更新；**
* **基于以上信息会计算一个富足程度（affluence index）；**
* **产品类别和品牌的购买数据（月度进行更新）。**

**业务需求**：CRISA面向如下两类客户提供数据分析服务，其一是广告公司订阅上述数据库服务，以向下游客户提供广告和营销策略建议；其二是产品生产商订阅其数据库以监测其产品的市场占有情况。

### 1.1.1 数据字段说明







### 1.1.2 核心问题

* + 数据集
    - BathSoap.xls数据集，包括CodeList、DM\_Sheet
    - DM\_Sheet中包含601个家庭的Demographic、Purchase Summary等信息，商品主要是印度香皂，涵盖639个品牌
  + 如何进行客户分群？
    - CRISA以往都是基于消费者的人口统计信息去做。
  + 如何进行更加精细的客户分群？
    - 是否可以基于购买过程中的信息进行客户分群，从而制定更加精细化的客户分群，针对更加细化的不同的客户群体采用采用不同的促销手段（如针对折扣敏感型和非敏感型用户使用不同的策略），更广义的即不同的销售主张（Selling Proposition）。
    - **Purchase Behavior**购买行为相关特征变量：购买数量、购买频率、折扣敏感性、品牌忠诚度等，其中品牌忠诚度的衡量可包括：1）购买的所有不同品牌的总数；2）在不同品牌之间切换的频率；3）在不同品牌上消费的占比。**具体涵盖：BathSoap.xls的DM\_Sheet中Purchase Summary和Brand wise Volume相关变量。**
    - **Purchase Basis**购买基础相关特征变量：价格（Price）、销售主张（Selling Proposition）等。**具体涵盖：BathSoap.xls的DM\_Sheet中Price Category-wise Volume和Selling Proposition-wise Volume的相关变量。**

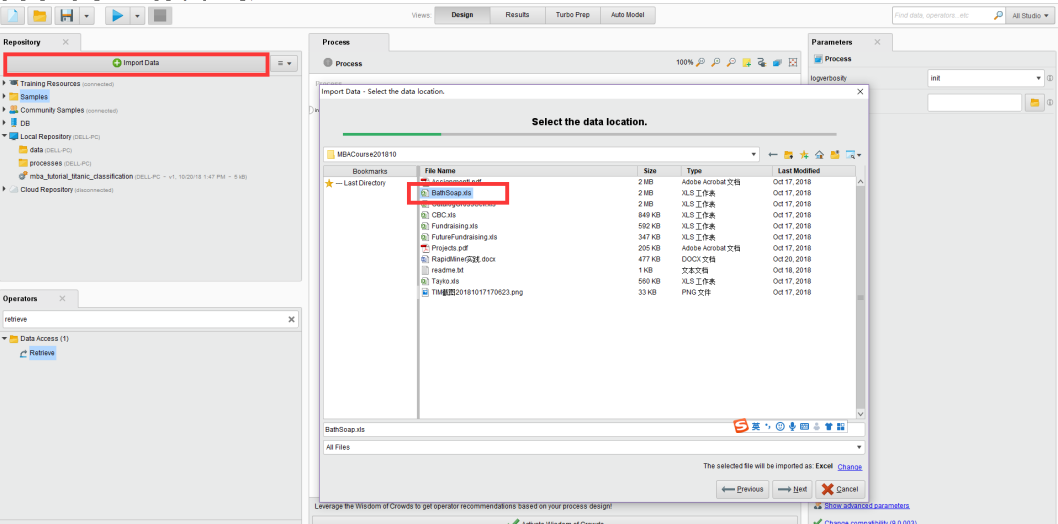
### 1.1.3 分析目标

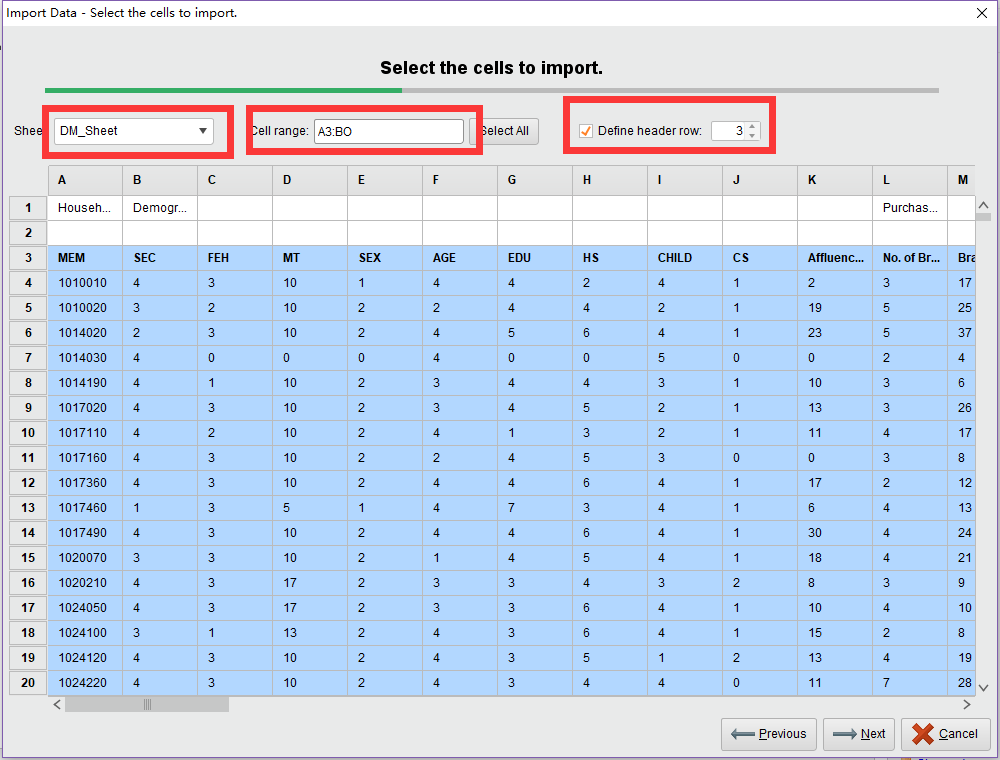
* 分别基于不同类型的变量，使用聚类技术识别细分的家庭聚簇，变量可以是：1）**Purchase Behavior**；2）**Purchase Basis**。
* 分析基于不同变量划分的聚簇的特点，以指导后续的广告和营销活动。

## 操作步骤

### 1.2.1 导入数据

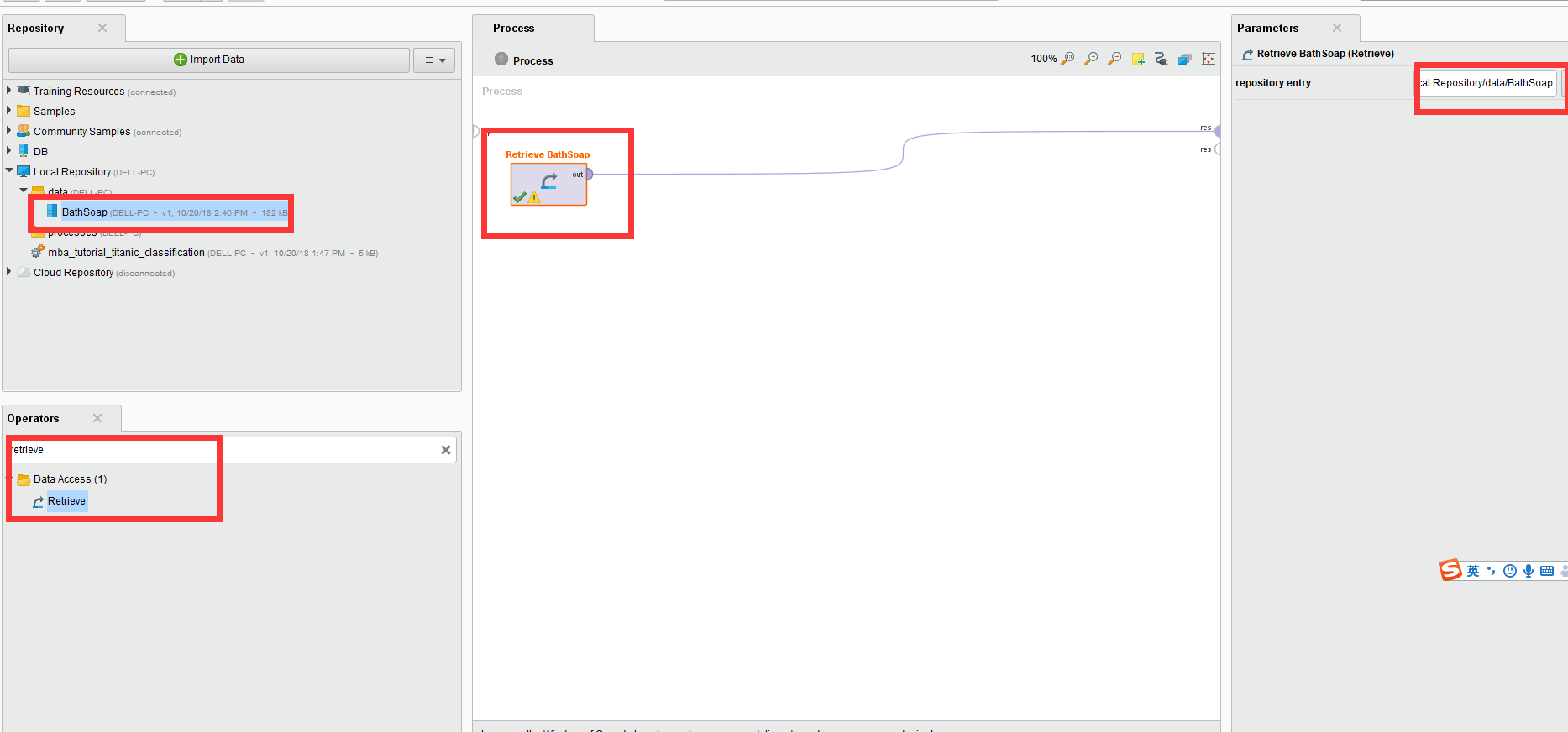
首先从本地电脑导入Excel数据，存到Local Repository中的Data目录下





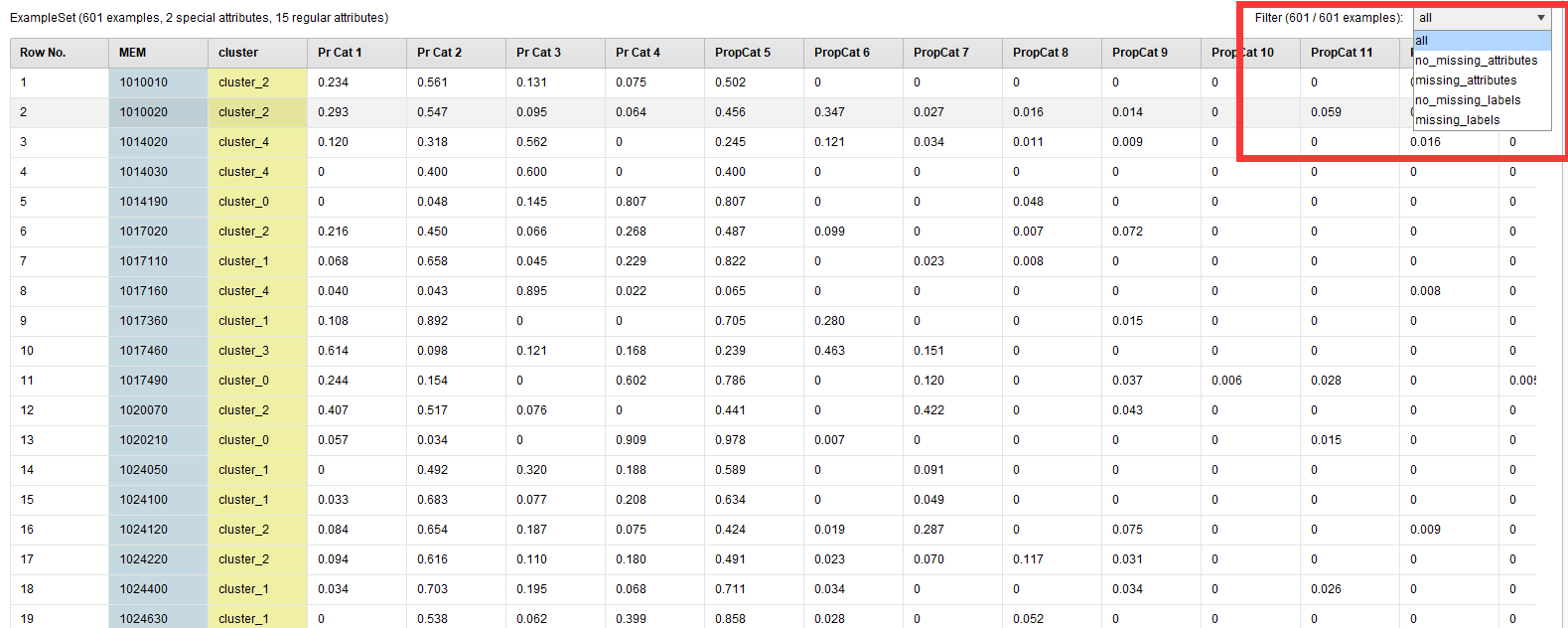
如上图需要注意字段名的选择以及Cell range的选择。

### 1.2.2 加载数据

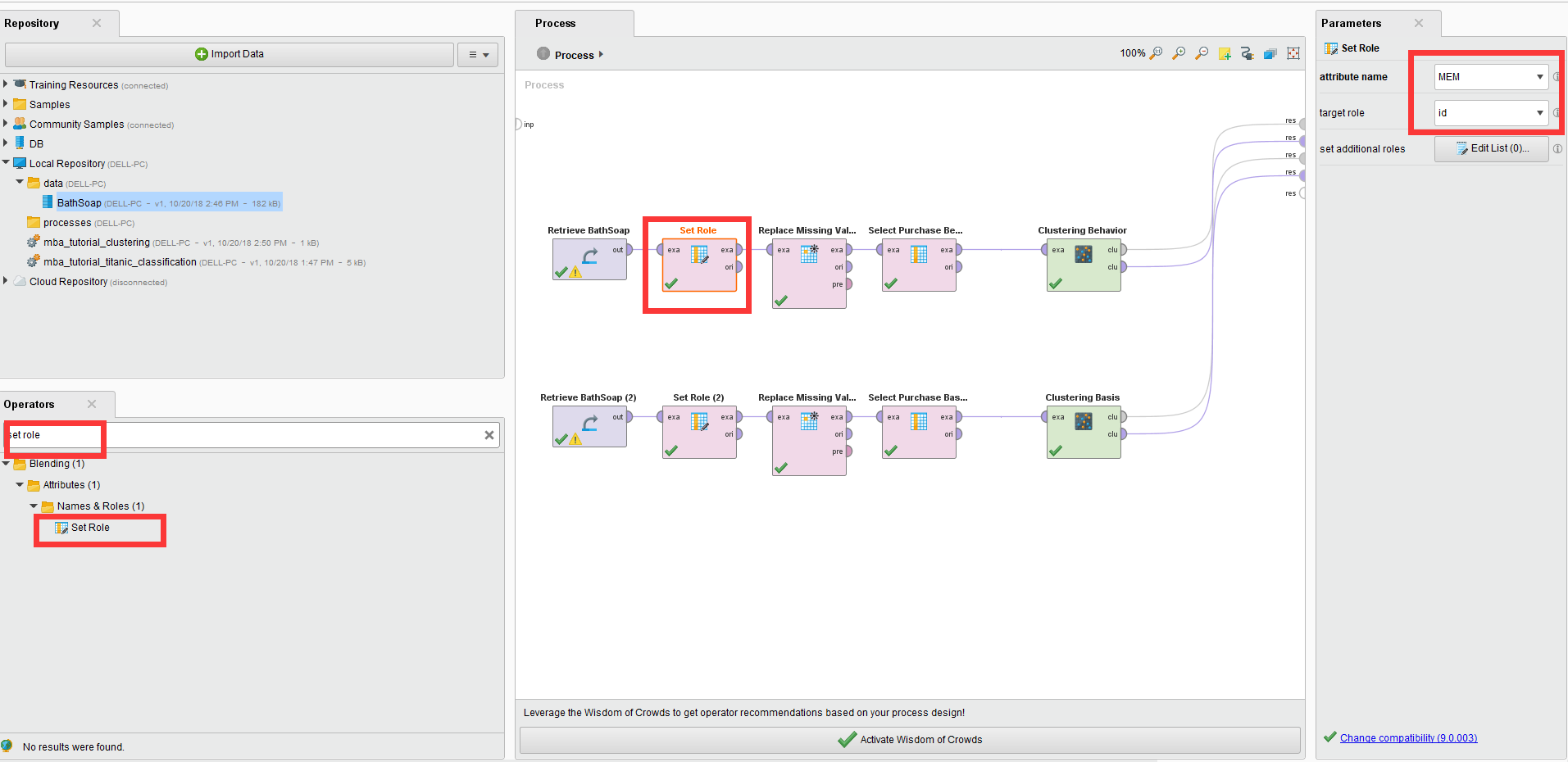


### 1.2.3 视觉评估与数据预处理

加载数据后，对数据进行简单的视觉评估：缺失、失真数据、异常值等。

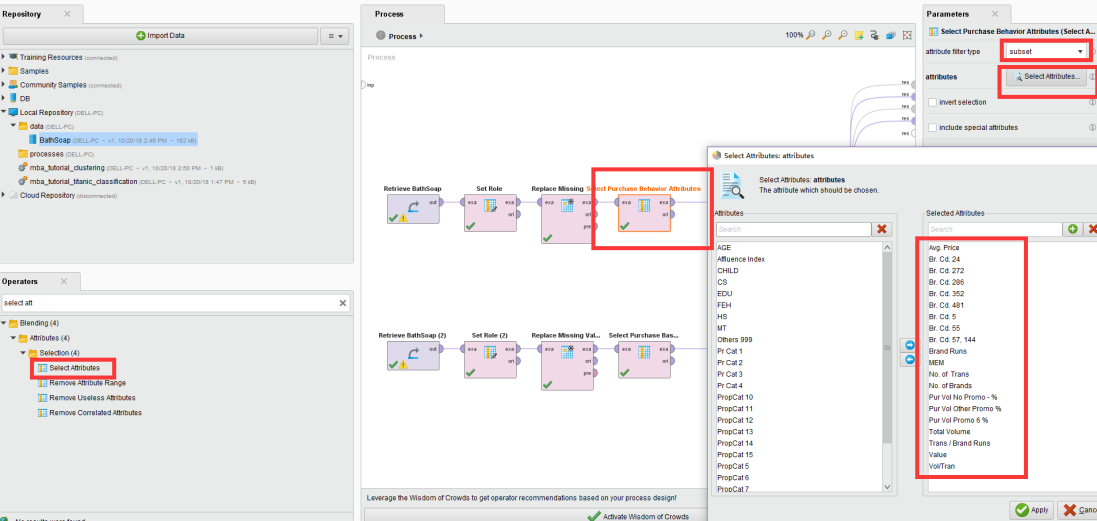


MEM是用来唯一标示家庭ID的，对于模型预测分析没有关系，因此它不应该作为变量包含在模型里面，我们把MEM的角色设置为id。调用“Set Role”操作符，在参数设置界面里，属性名称选为“MEM”，“target role”设置为id角色。

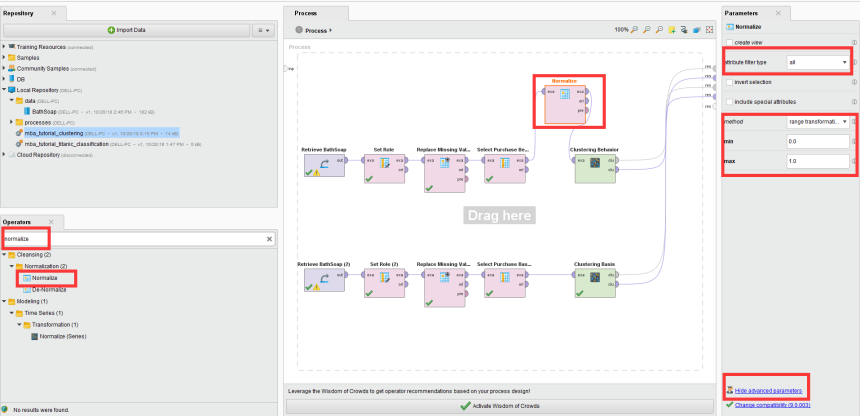


随后选择missing\_attributes，发现有一条数据缺失了很多属性，需要做一些处理。对缺失值可以进行取均值的处理，通过左侧的‘Cleansing’数据清洗操作符：“Peplace Missing Value”操作符进行处理，默认回填均值。如果没有缺失值，可以不处理，不添加此操作符。

紧接着，我们选择用于聚类的属性，通过左侧的“Select Attributes”操作符选择Purchase Behavior相关的变量：

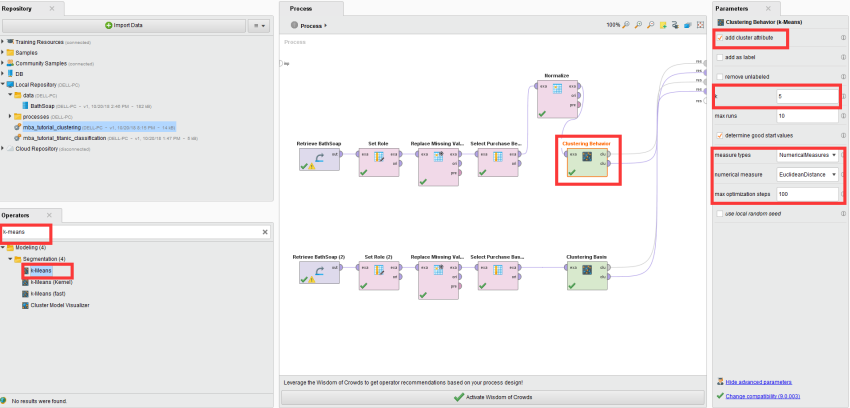


由于不同属性的Scale不一样，为避免较大值属性对模型产生未知的影响，需要对属性值做normalization，我们选用range-transformation将所有属性值normalize到0-1之间。

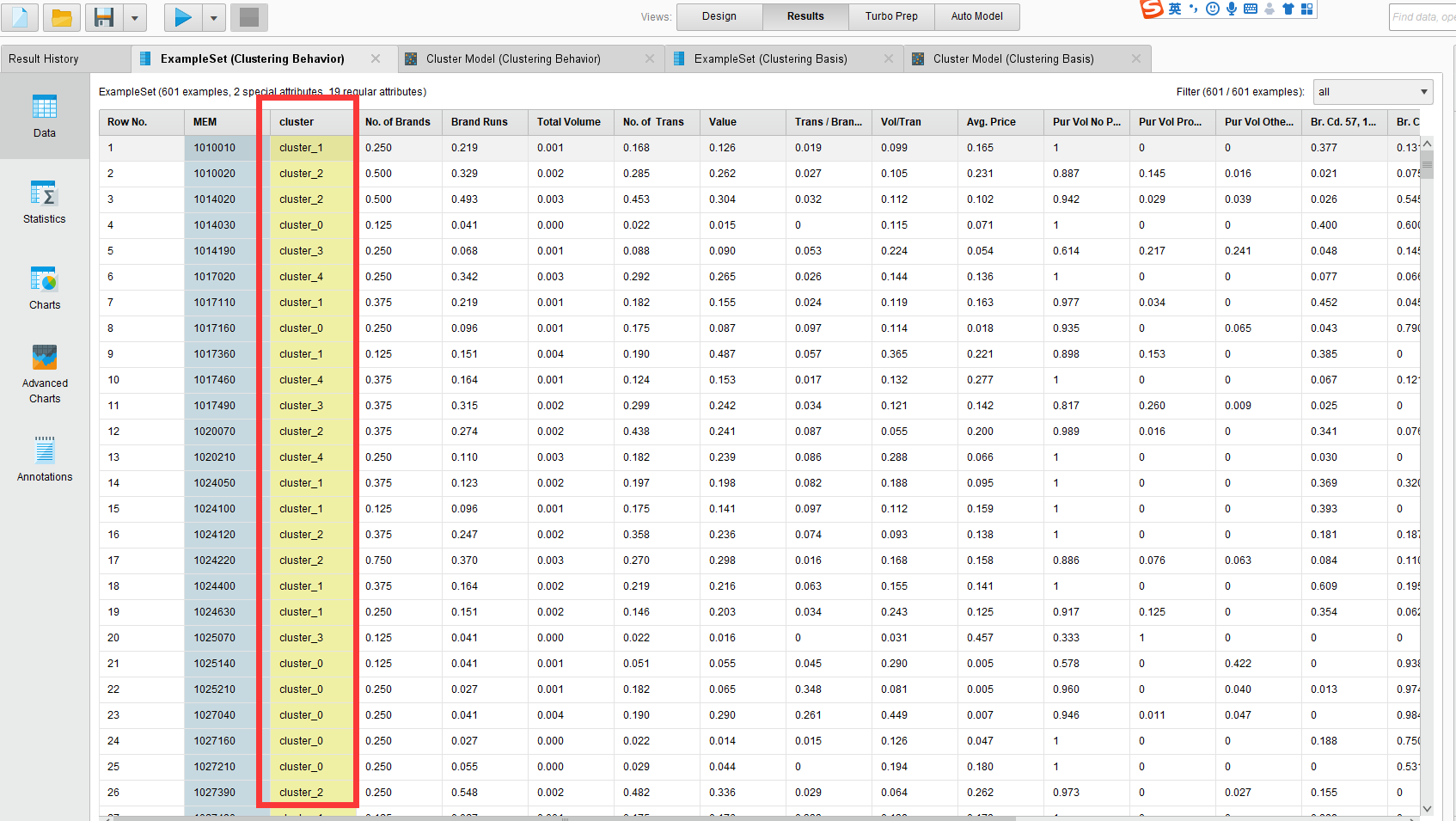


### 1.2.4 基于Purchase Behavior的聚类分析

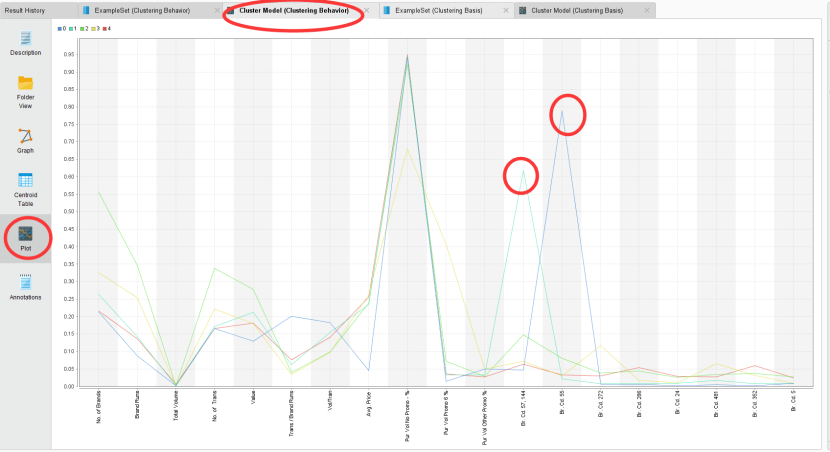
添加k-means clustering操作符，设置聚类数量为5



首先可以看到每个家庭都会被标注出类簇编号：

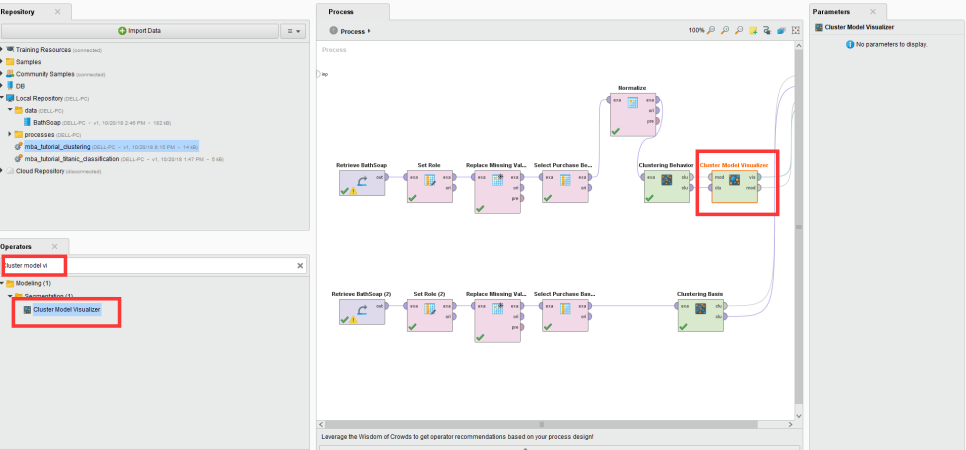


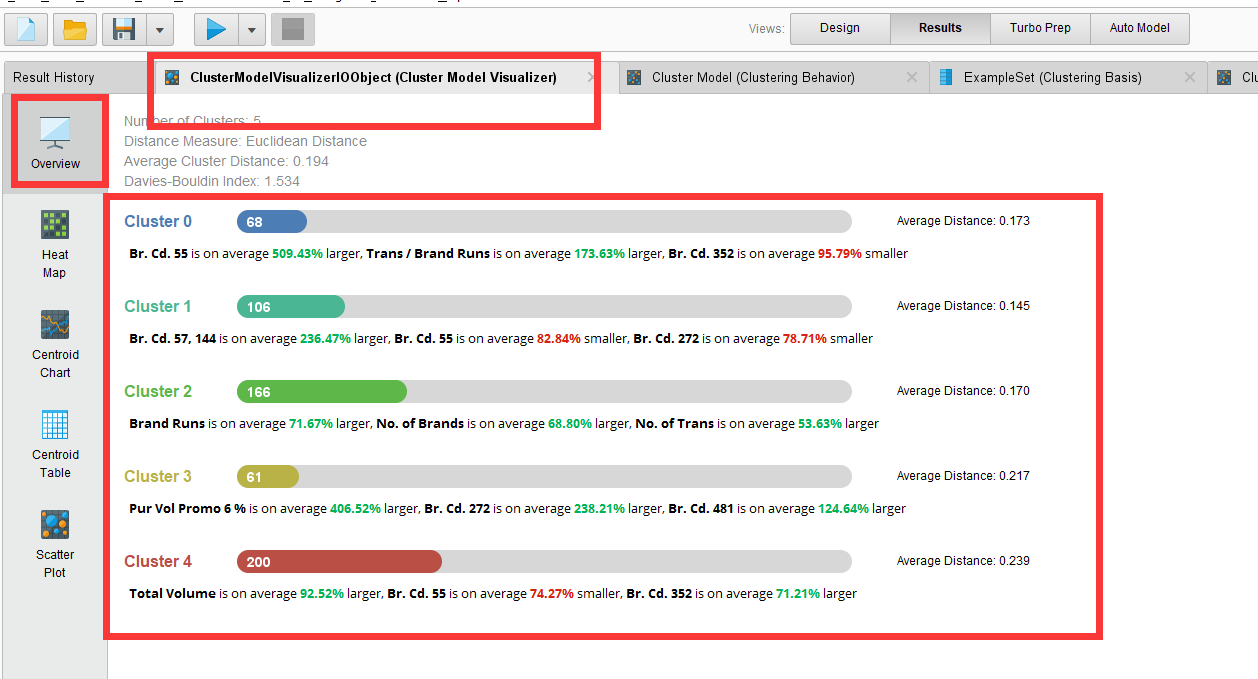
其次查看各个类的类中心，并分析其特点：



可以从上图发现Cluster 0：偏好于Brand 55；Cluster 2：偏好于Brand 57,144

此外，可添加“Cluster Model Visualizer”操作符对聚类结果进行可视化：

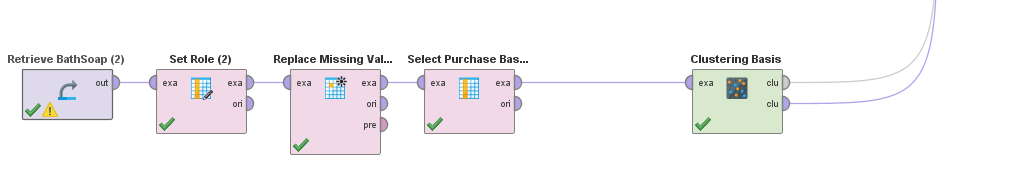






### 1.2.5 基于Purchase Basis的聚类分析

可使用复制Purchase Behavior的操作符的方式进行快速构建（在operator的Select Attribute时候稍加修改即可），最终结构如下图所示：



## 个人作业

每个人完成课堂演示内容，并在基础上完成客户分群案例中的后续分析，提交包含原始数据说明、数据初探、操作流程、数据分析结果等在内的课程报告。

基于RapidMiner后续客户分群分析的方向可以包含但不限于：

* K-means聚类最重要的参数是聚类数量K，而课堂演示是直接指定K=5，因此可考虑研究不同的聚类个数（如K=2-10）对聚类效果的影响，建议的方案是能否将K值选择融入RapidMiner的processes中，使用聚类内部评价指标实现自动的聚类个数选择；如若不能，可以手动调整K值，对不同K下聚类效果进行视觉评估与对比，以选择较优参数。
* 课堂演示案例中使用了如下三个指标对家庭的品牌忠诚度进行衡量：No. Brands, Brand Runs, Brand wise Purchase，其中Brand wise Purchase是不同品牌的消费占比（7个品牌的比例），这里的问题是能否用一个单独的数值表征Brand wise Purchase，例如由多个品牌占比计算出来的熵。可结合新计算的变量重新进行聚类分析。
* 基于不同类的变量组进行聚类分析，课堂演示展示的是基于1）Purchase Behavior和2）Purchase Basis的聚类，课下可以对Purchase Basis聚类结果进行分析，同时进一步使用3）家庭人口统计特征进行聚类分析，以及三类变量的组合进行聚类，例如1）和2）、2）和3）、1）和3）等，可以观察从不同视角进行聚类的聚簇的不同特点，分析对制定差异化广告或营销策略有用的聚簇模式。