数据总概：

1. 训练集数据共478138行，总购买率0.01886694。
2. 测试集一共18371个样本，其中缺失值最多的性别共缺失469个数据，其他的缺失值个数一般为18.36或1，影响不大。

广告商品信息：

Item\_id：广告商品编号 有10075种

item\_category\_list：广告商品的类目信息，从大到小（例如：手机，全面屏手机）

item\_property\_list：广告商品的属性列表，没有顺序

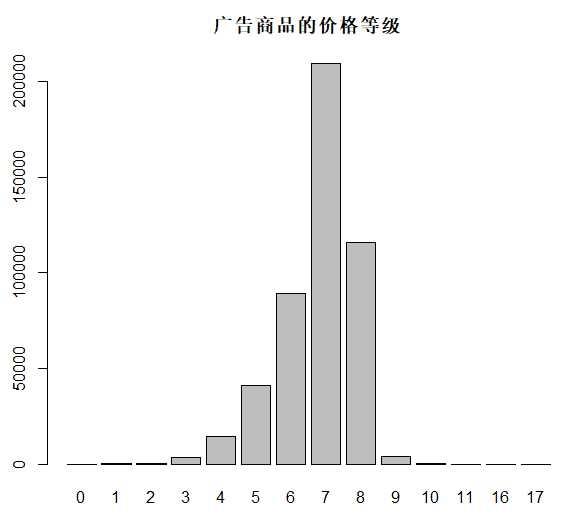
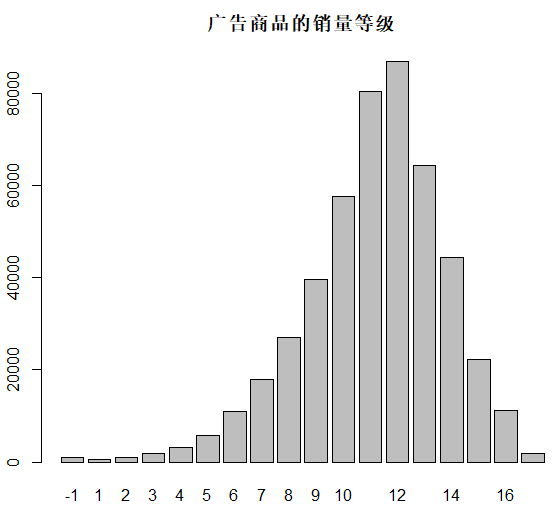
（例如：红色，4.7寸，公司：-1则表示没有公司信息）

item\_brand\_id：广告商品的品牌编号，共2055种（不同广告面向同一型号）

item\_city\_id：广告商品城市编号，共128个城市

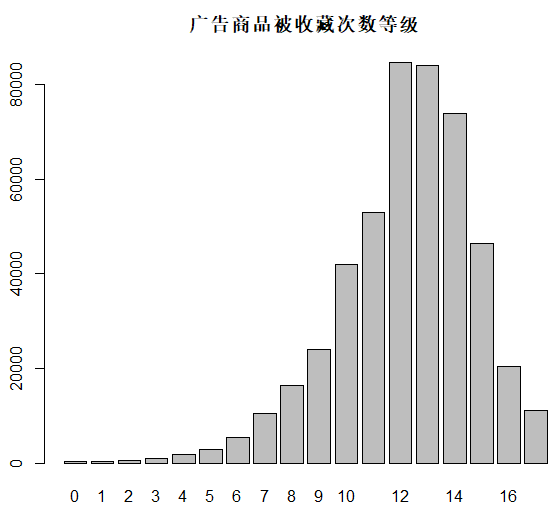
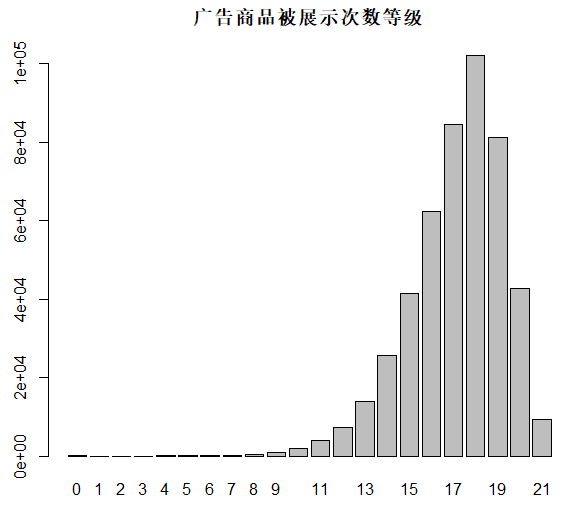
item\_price\_level：广告商品的价格等级

item\_sales\_level：广告商品的销量等级

item\_collected\_level：广告商品被收藏次数等级

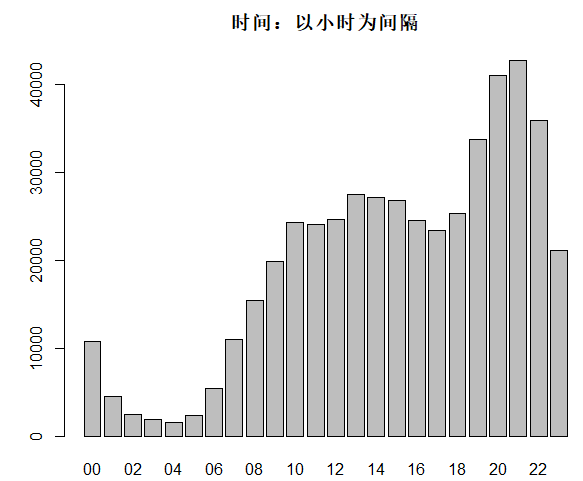
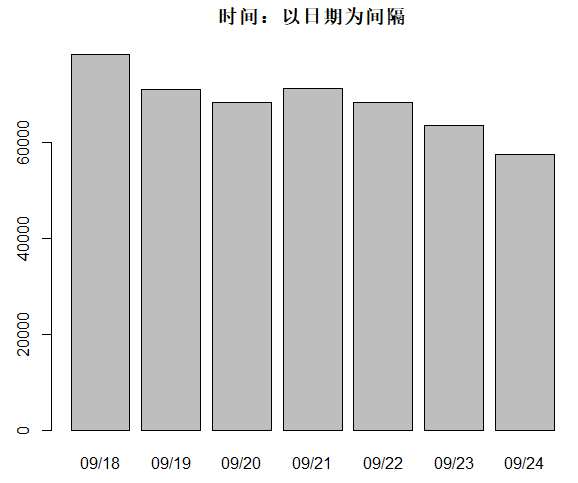
item\_pv\_level：广告商品被展示次数等级（怎么把它和销量价格收藏次数结合?）

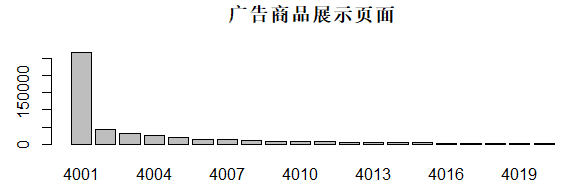
上下文信息：

context\_id： 上下文信息的编号，共478111种（比总样本量少27个）

context\_timestamp：广告商品的展示时间（只有七天的数据，日期有处理）



context\_page\_id：广告商品的展示页面数据（一般都会选择在第一页展示）



predict\_category\_property：根据查询词预测的类目属性列表

（“；”后为一个新的类目“，”后为一个类目对应的属性，属性之间没有顺序）

用户信息

user\_id:用户编号 有197694个用户

user\_gender\_id：用户预测性别，0为女性，1为男性，2为家庭用户

-1 0 1 2

12902 360817 94070 10349

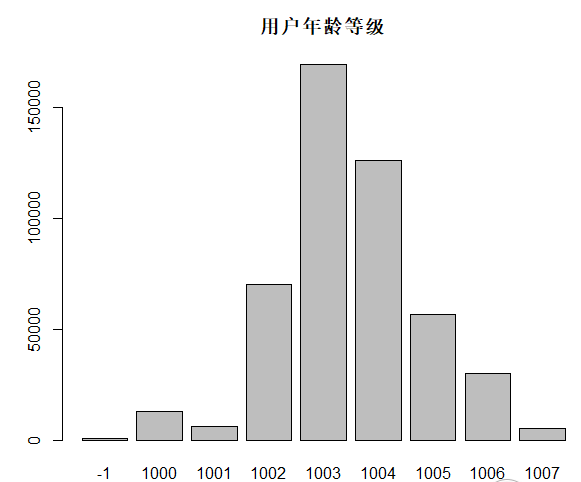
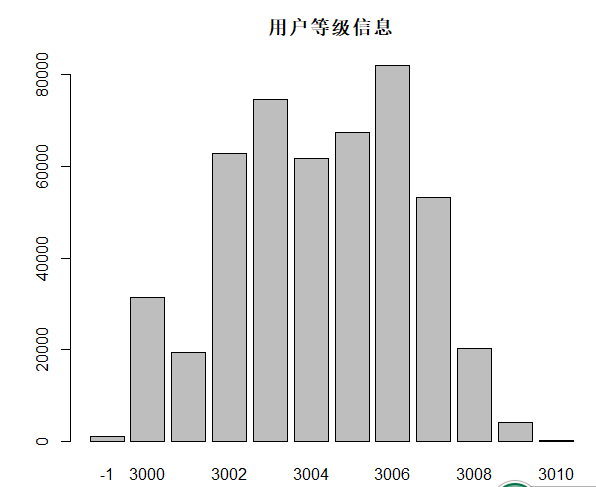
user\_occupation\_id：用户预测职业编号

-1 2002 2003 2004 2005

964 146917 6418 23803 300036

user\_age\_level：用户预测年龄等级

user\_star\_level：用户星级编号（除性别外用户信息缺失都为964，怀疑是新用户）



店铺信息：

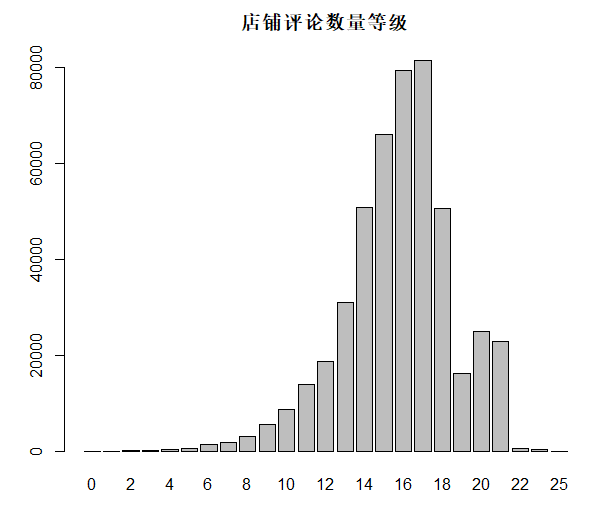
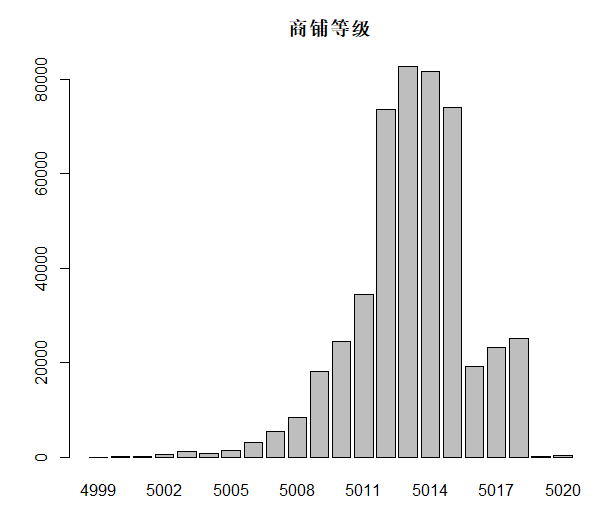
shop\_id：店铺编号 共有3959家店铺的信息

（最多的一家店铺有11278个样本信息，有457家店铺只有一条信息）

shop\_review\_positive\_rate：店铺好评率，绝大部分都在0.9+，缺失值7个

shop\_review\_num\_level：店铺的评价数量等级

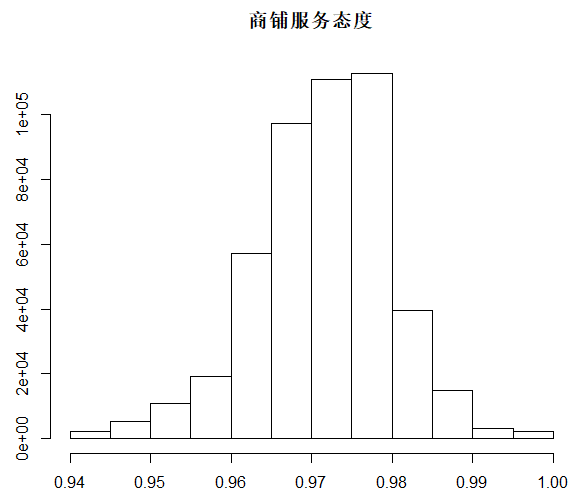
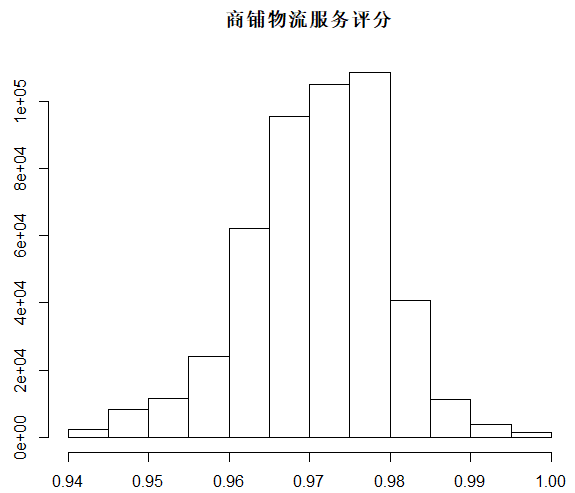
shop\_star\_level：商铺星级等级

shop\_score\_service：商铺服务态度评分（缺失值59个，小于0.94的有3017个）

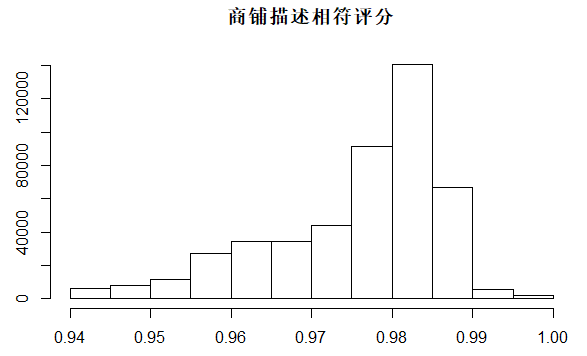
shop\_score\_delivery：商铺物流服务评分

（缺失值也是59个，小于0.94的有3055个）

shop\_score\_description：商铺的描述相符评分

（缺失值59个，小于0.94的6387个）

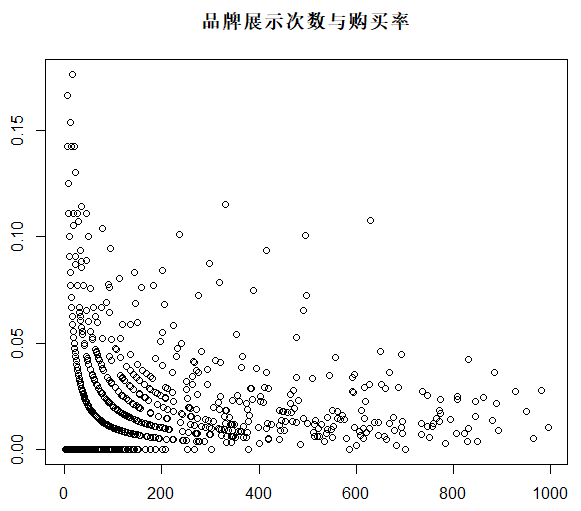


|  |  |  |  |
| --- | --- | --- | --- |
| 商铺评分之间的相关系数 |  | 物流服务 | 描述相符 |
| 服务态度 | 0.994 | 0.938 |
| 物流服务 |  | 0.931 |

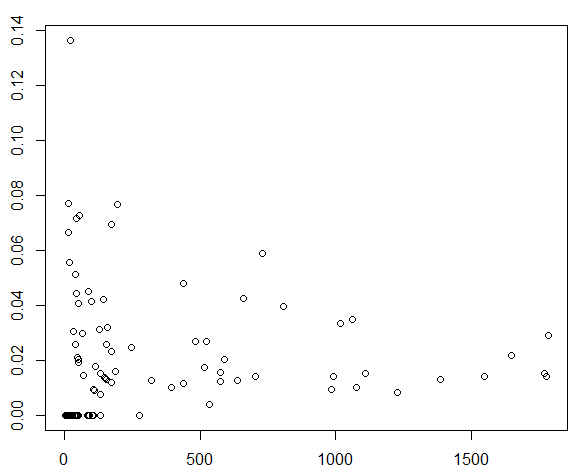
Part3: 与购买率的关系

3.1 广告商品信息与购买率的关系

item\_brand\_id:图横轴为品牌的样本数，纵轴为购买率，图显示的是样本数少于1000（最多的有接近7w次）并且购买率小于0.2（为1的需要额外处理），感觉购买率与品牌还是有一些关系的，在频数大于200后分布比较均匀。

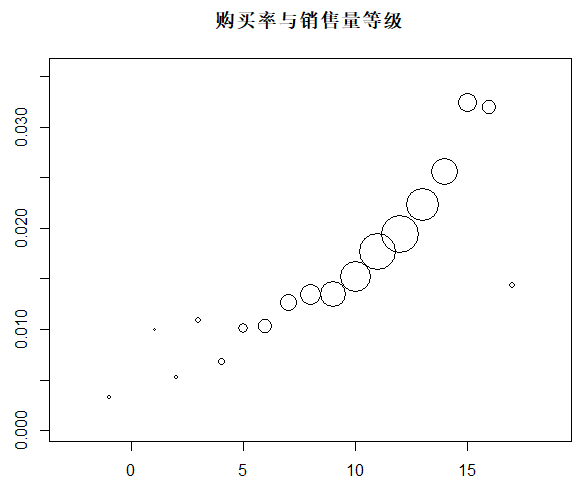
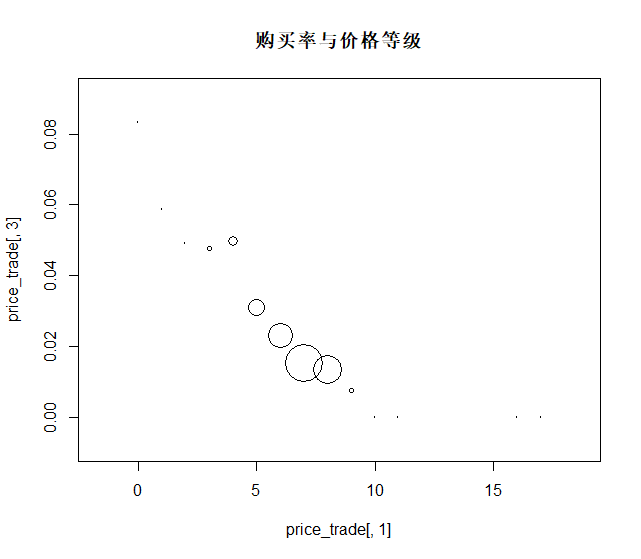


item\_city\_id：图显示的是样本数少于2000（最大为15w+），购买率小于0.15（最大为0.3+且展示次数较少），随着频数增大分布越来越稳定。



item\_price\_level：横轴为价格等级，纵轴为购买率，标签为样本量，可以认为价格等级和购买率的关系特别大！

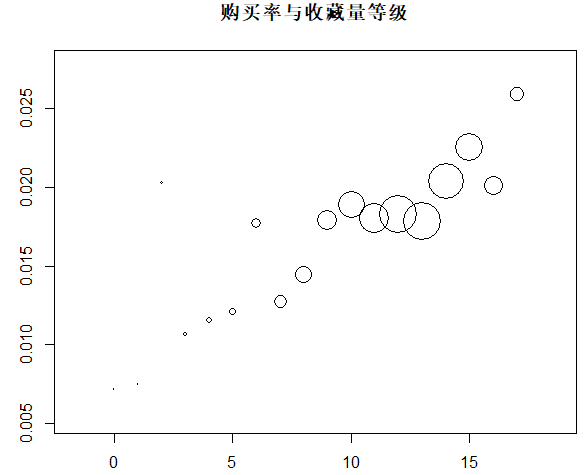
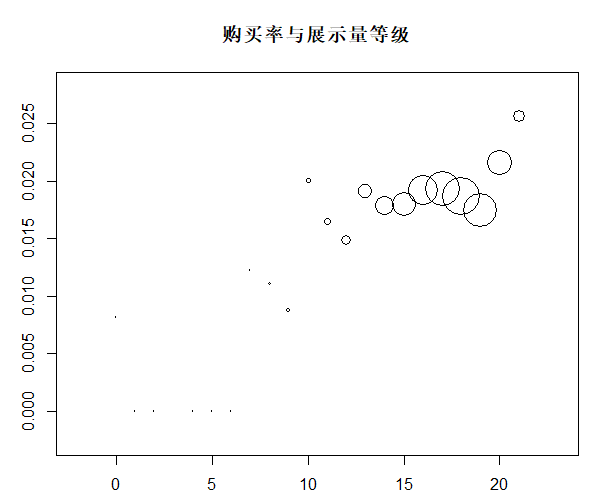
item\_sales\_level：销售量等级存在913个缺失值，气泡的大小代表样本量的多少，可以认为销量等级也和购买率关系很大



item\_collected\_level：基本上也是正向相关

item\_pv\_level：相比之下展示量等级对购买率的影响就没有那么大了，

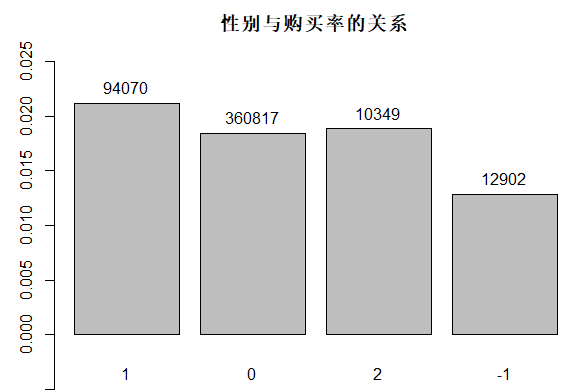
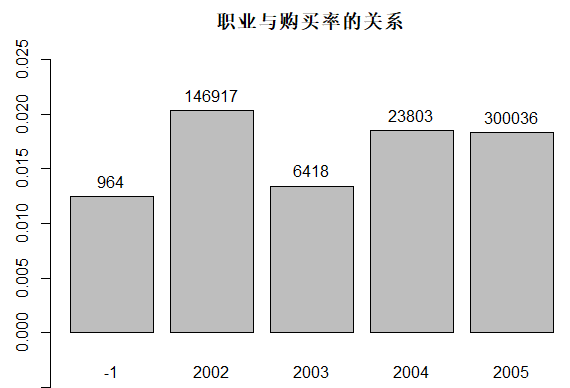
收藏等级为0，销量等级一定为缺失值

3．2 用户信息与购买率的关系

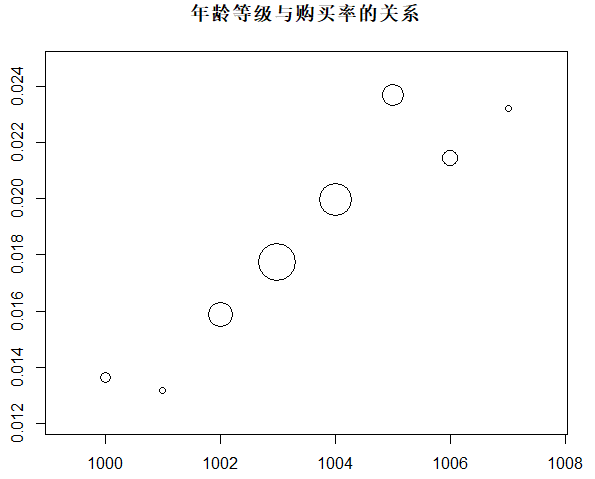
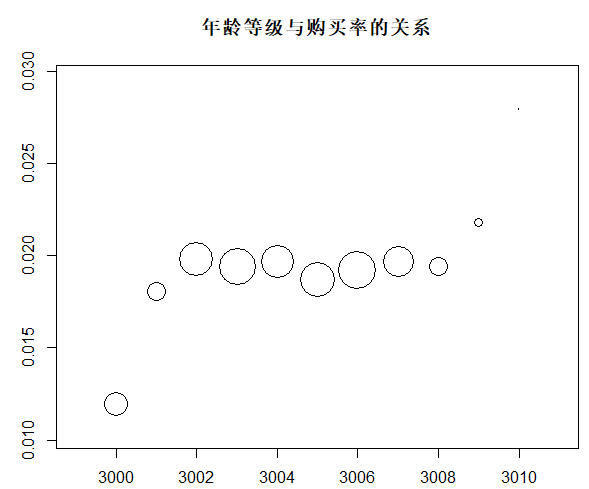
user\_gender\_id：发现不同性别的购买率差别不大，但是缺失值却相差很大，这是为什么呢？

user\_occupation\_id：缺失值也是964个，职业对于购买率的影响还不明显

user\_age\_level：缺失值964个相应购买率为0.0124，从整体看年龄等级与购买率的相关程度也很大，可能年龄越大越相信广告

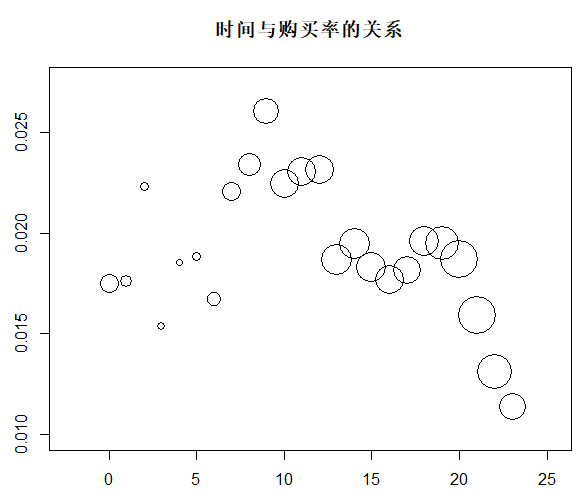
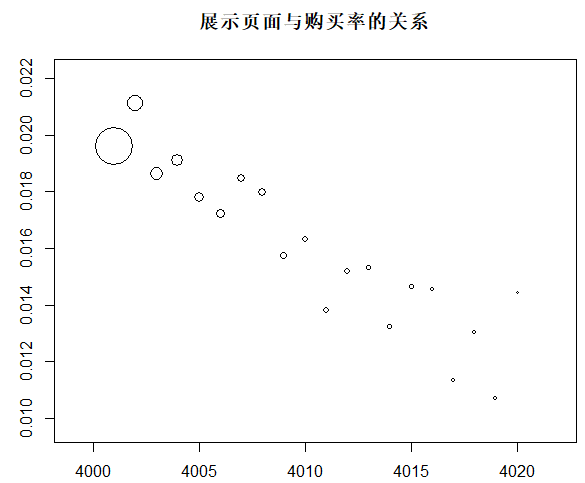
user\_star\_level：缺失值964个，可以认为存在一定关系，但星级最低的明显异常，可能需要额外提出（猜想有可能是注册新用户恶意搜索），星级最高的用户购买率最高可能是因为样本来太少导致的偏差，需要降权

3.3 上下文信息与购买率之间的关系

context\_timestamp：可以将1-6点的数据合并然后购买率基本上就呈现一个逐步下降的趋势了

context\_page\_id：页面与购买率负相关！！

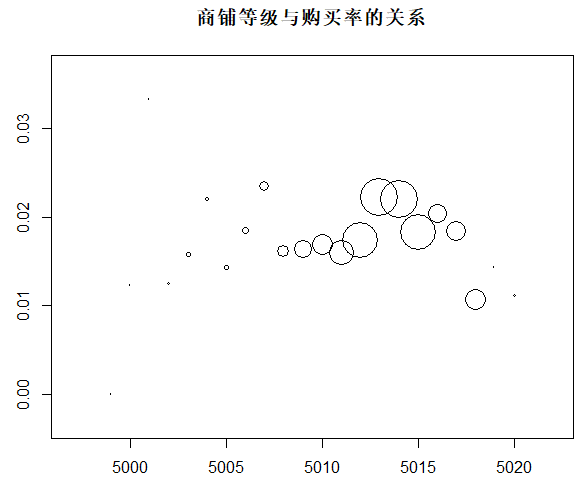
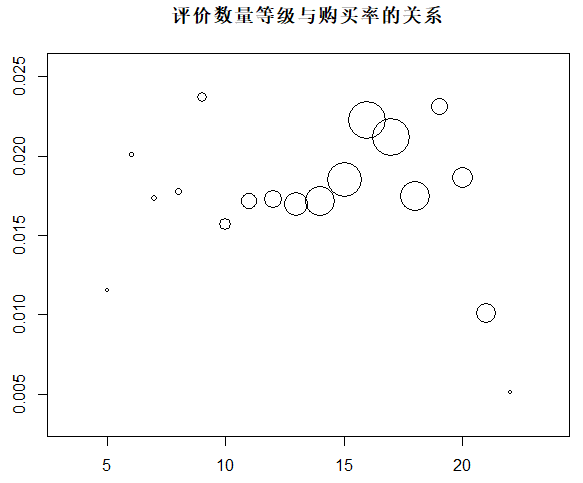
 

3.4 店铺信息与购买率的关系

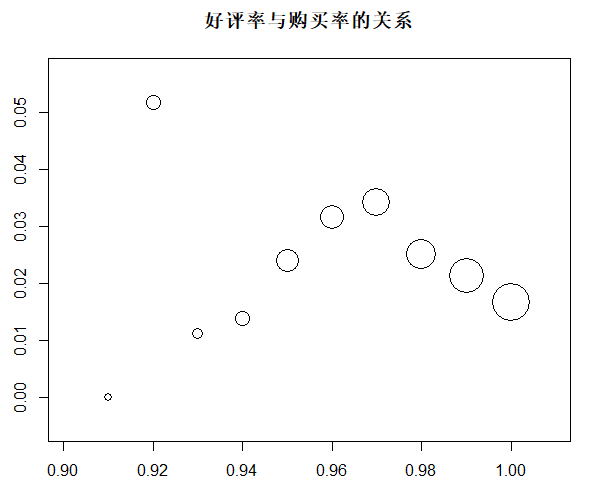
shop\_review\_num\_level：当等级为0,1时的样本（27个）广告都没有购买行为。

将0-5合并成5,22-25合并成22感觉关系不明显

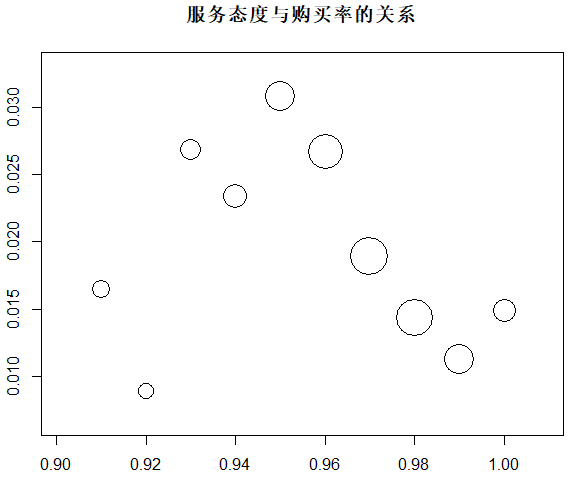
shop\_star\_level：商铺等级与购买率关系也不是很大，但是要注意当等级4999时应该是新注册的店铺，样本有七个，全为0



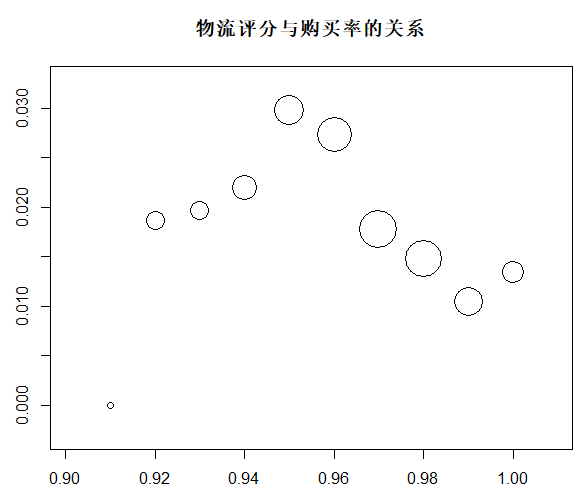
shop\_review\_positive\_rate：缺失值7个，气泡大小进行了对数化处理，好评率小于0.9的购买行为为0 ，而且好评率基本上都聚集在了1附近，有刷好评的嫌疑，而在0.96附近的则更为真实



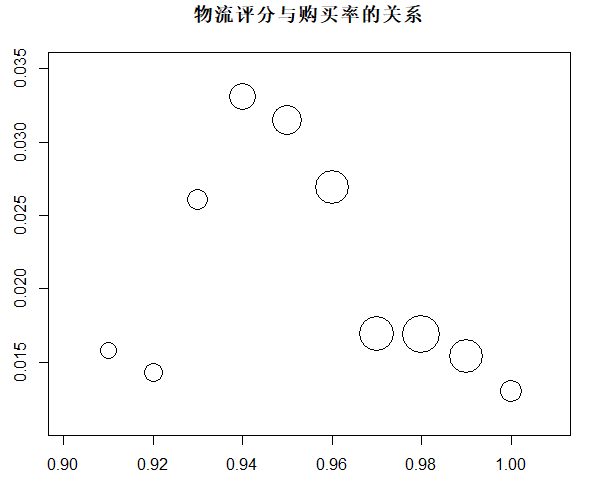
shop\_score\_service：缺失值59个，气泡大小对数化，形状与上图相似



shop\_score\_delivery：缺失值59个，气泡大小对数化，形状与上图相似



shop\_score\_description：缺失值59个，气泡大小对数化，形状与上图相似



深刻揭示了物极必反的道理，好评不要刷太高，但是也有可能是几个样本量特别多的店铺导致这样的情况