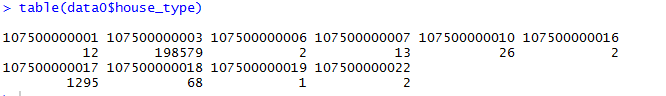
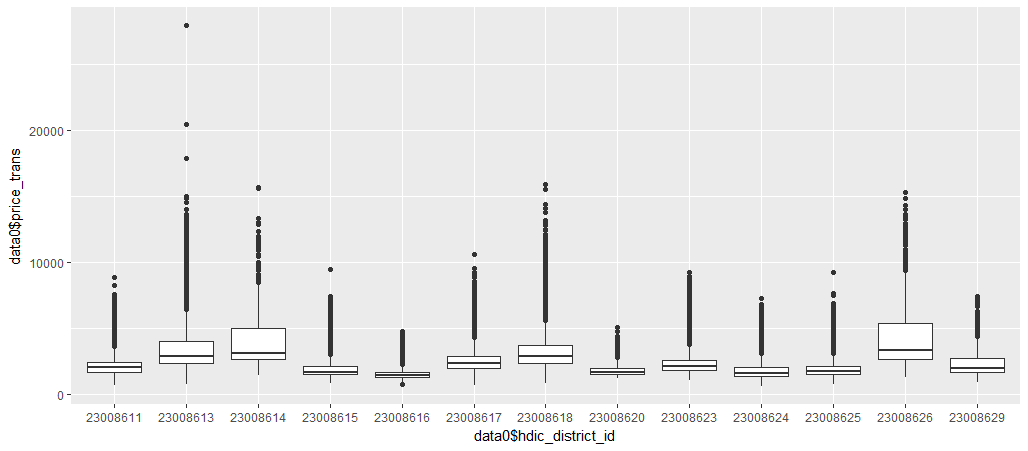
house\_type 这个字段的统计信息如下：





所以我只考虑：规划类型为住宅的房源。



考虑在决策树模型中放入的特征：

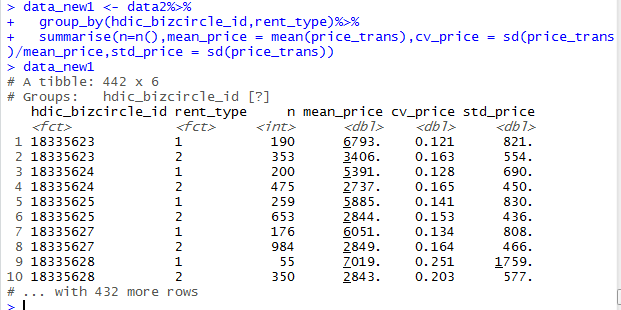
商圈编号 hdic\_bizcircle\_id 共有223个商圈

租赁类型 rent\_type 1 整租 2 合租

房源面积 house\_area

出租面积 rent\_area

## 对商圈和租赁方式求租赁价格均值（标准差、变异系数）

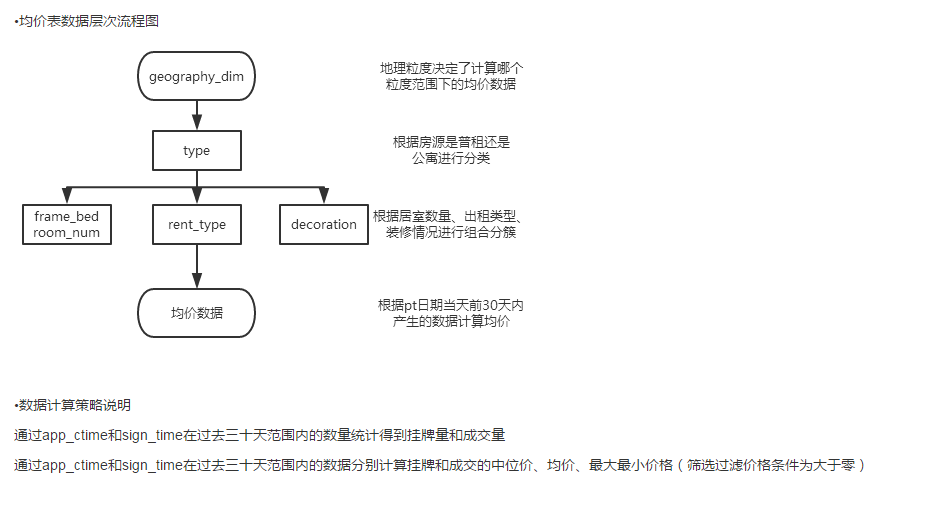


能看出什么来吗？

之前的均价模型的策略;

3.计算成交或者挂牌数据时，通过对地理粒度、房源来源、居室数量、出租类型、装修情况进行group by,取价格大于零和时间符合范围的数据进行统计.

这个已经在大数据平台上面显示为：rentplat\_dw\_avg\_price\_di 这个表。



App\_crme : Pt-30

仔细体会 rentplat\_dw\_avg\_price\_di 这个表：

可以继续探索的地方有：1. 由均价得到一个均价区间（从点估计到区间估计该怎么做?）

2.如何评估得到的这些均价的效果？一种想法是：在均价加减5%的区间，看有多少样本点能落入？另一种想法是：求组内方差（实际价格-均值）之后？

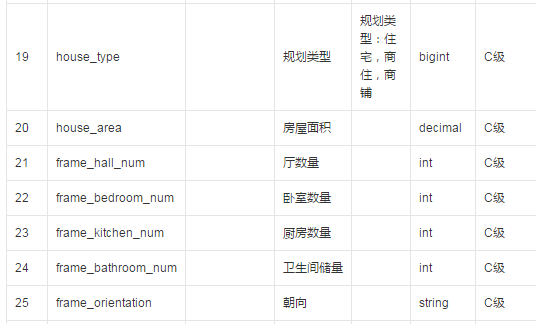
3.除了求均价用到的分类字段之外，还有哪些字段可以进一步对同一组内的房源价格做进一步的划分呢？

我自己觉得可能会有内部区分力的字段有;

楼栋完成时间；房屋面积；出租区域面积。









马上要实践的事情;

1. 对 rentplat\_dw\_avg\_price\_di 这个表select \* ,以此来看一下这个表是如何提炼均价结果的。
2. 参考之前wiki上小区均价策略留下的反思问题：

•待改进的点

1.某些小区在30天内可能没有成交数据或挂牌数据，此时考虑用该小区上一个pt的数据来填补null值

2.目前均价采用的是房屋总价，日后考虑使用房屋总价/出租面积，这样更为精确.

进展：对于合租而言，房屋单价没有意义。因为出租面积和成交价格不相关（pearson=0.1）

3.目前房价未对离群价格进行过滤筛选，日后考虑可以结合第二点进行房屋单位面积价格进行范围限定筛选

4.采用join导致逻辑复杂计算一天pt需要半小时左右，考虑将计算中途数据存为一个临时表以供调用来减少重复计算，来代替join逻辑

## 合租房源的出租面积不是影响其成交价格的显著变量，二者之间没有相关性。

## 将每个商圈的整租交易数据视为一个总体，则每个总体的出租面积与成交价格呈高度正相关关系。