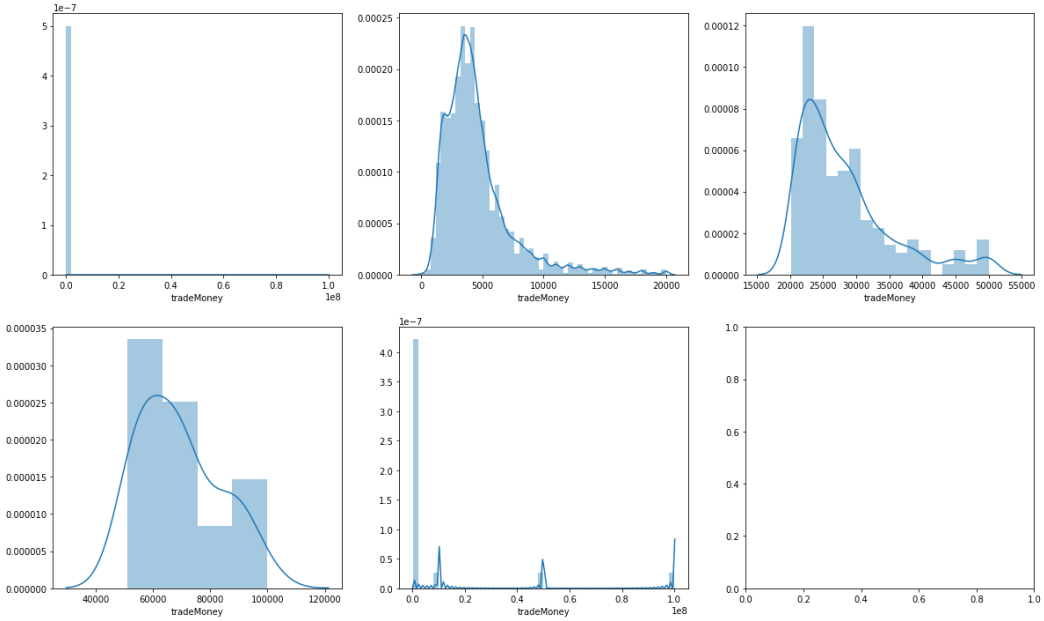
总结

任务一:赛题分析,这个任务主要是对题目和数据分布的理解

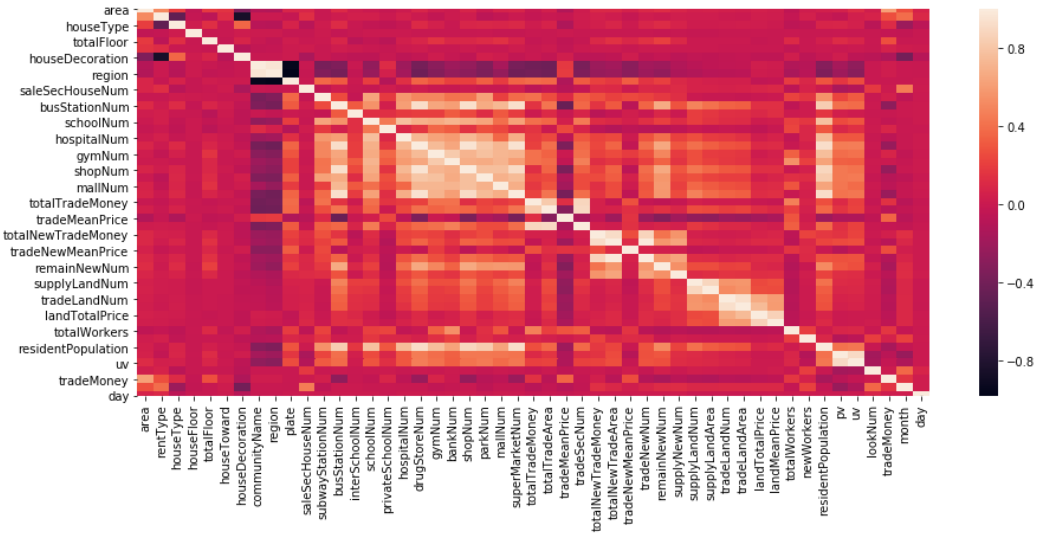
- 对各个属性,数值型和非数值型进行区分
- 对数值型数据进行describe,查看均值等属性
- 对于地区导致的学校医院等公共措施的分布,发现在45地区最发达
- 对缺失值的查看:发现缺失值集中在pv和uv的那18行,是人为删除
- 查看非数值型特征的分布
- 对目标target的分布进行查看



- 尝试对数据进行处理

任务2:数据清洗

- 对rentype进行清洗,使用众数方式填充
- 对于buildyear使用均值进行填充
- 对于pv和uv这18行的数据使用均值进行填充
- 对交易时间进行分割
- 删除city,tradeTime,ID这几列
- 对训练集合的target使用孤立森林删除离群点
- 对训练集合中的area同样进行数据清洗
- 对训练集合的area和target进行处理
- 对每一个region进行数据查看,通过箱型图进行深度数据清洗



H1 任务3: 特征工程

- 对房间描述几室几厅几卫进行分割,添加房-卫比
- 对租房方式中的未知方式进行查看
- 分割交易时间
- 计算统计特征
- 分组特征
- 聚类特征
- 上面这几种方法的意思就是不断生成新的特征
- 平滑,为了更好的下降
- 线性,相关系数,递归特征消除,基于惩罚项的特征选择,基于随机森林的处理

H1 任务4:模型选择

- lightgbm方法的处理
- 参数的选择,K折交叉验证

H1 任务5:模型融合

- 将特征放进模型中预测，并将预测结果作为新的特征加入原有特征中再经过模型预测结果
- 以鸢尾花数据进行处理,对多种算法进行实际处理