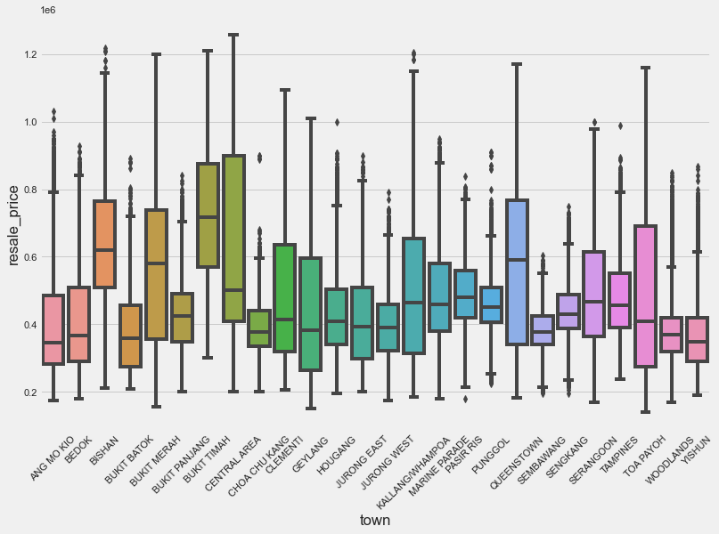
**1·数据预处理** Ordinary code编码

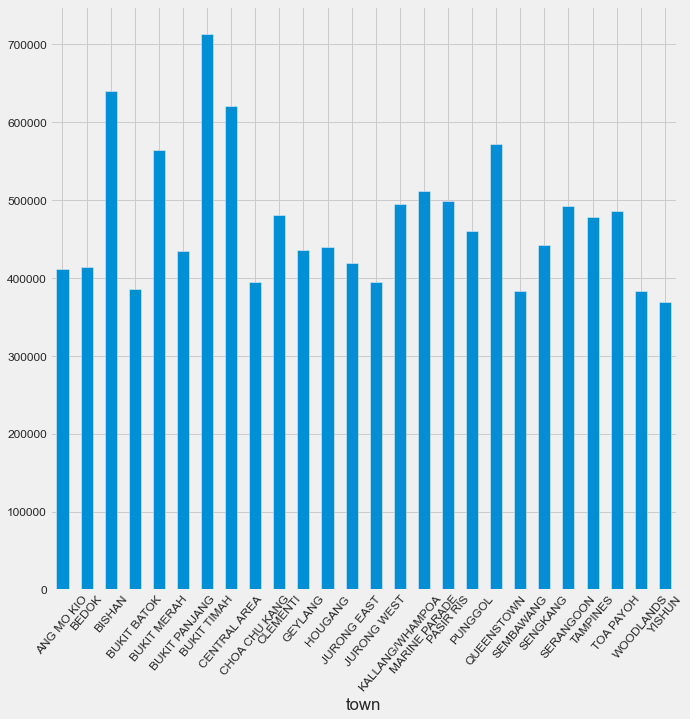
**1·1 Main Features**

**1·2 Town属性**

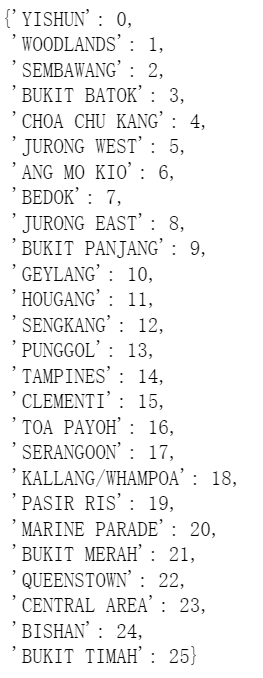
Town:按照每个Town平均价格的顺序进行编码。（横坐标为town,纵坐标为均价）。



**箱型图**

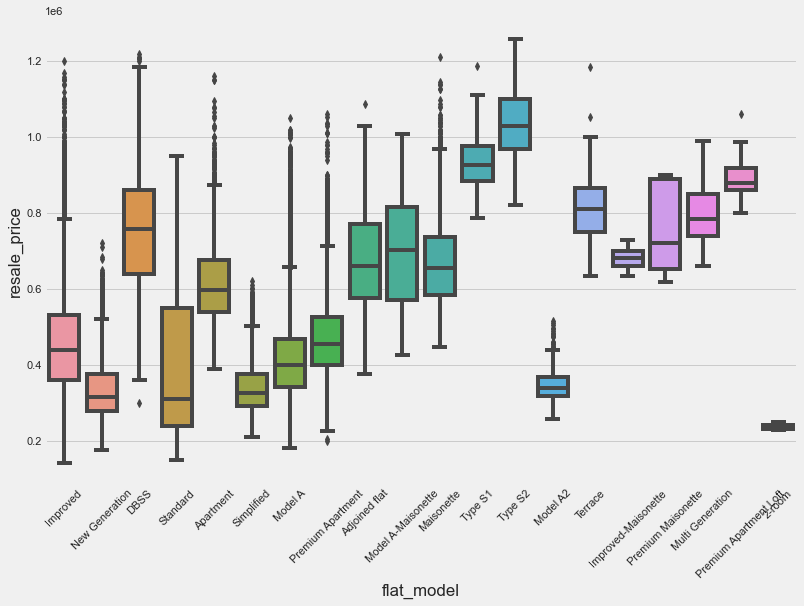


**柱状图**

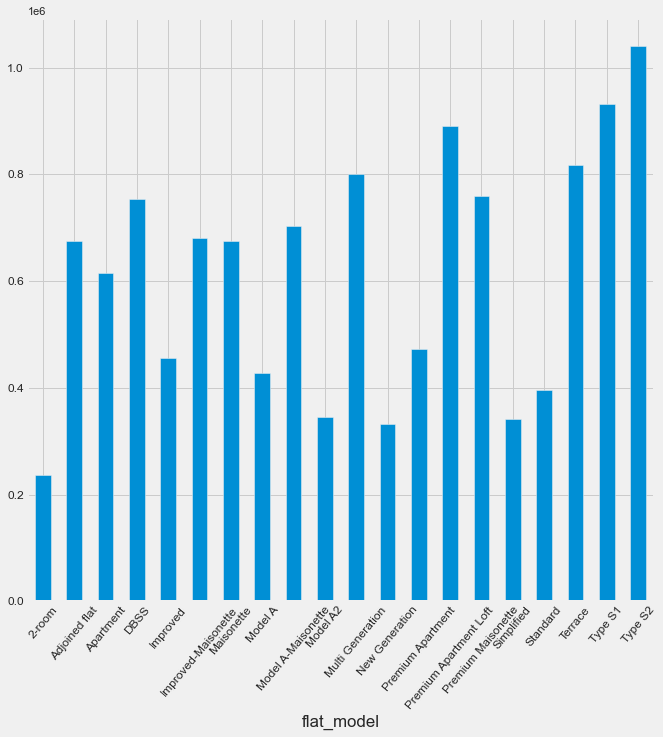
 

**各个Town均价及编码结果**

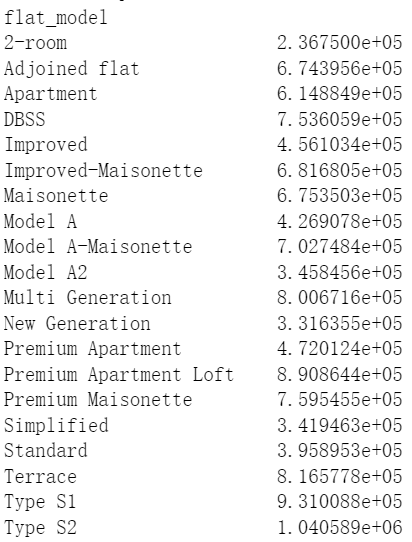
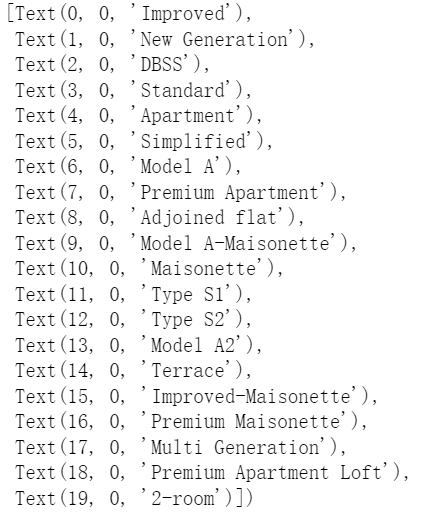
## 1·3 Flat\_Model属性（户型结构）



**箱型图**



**柱状图**

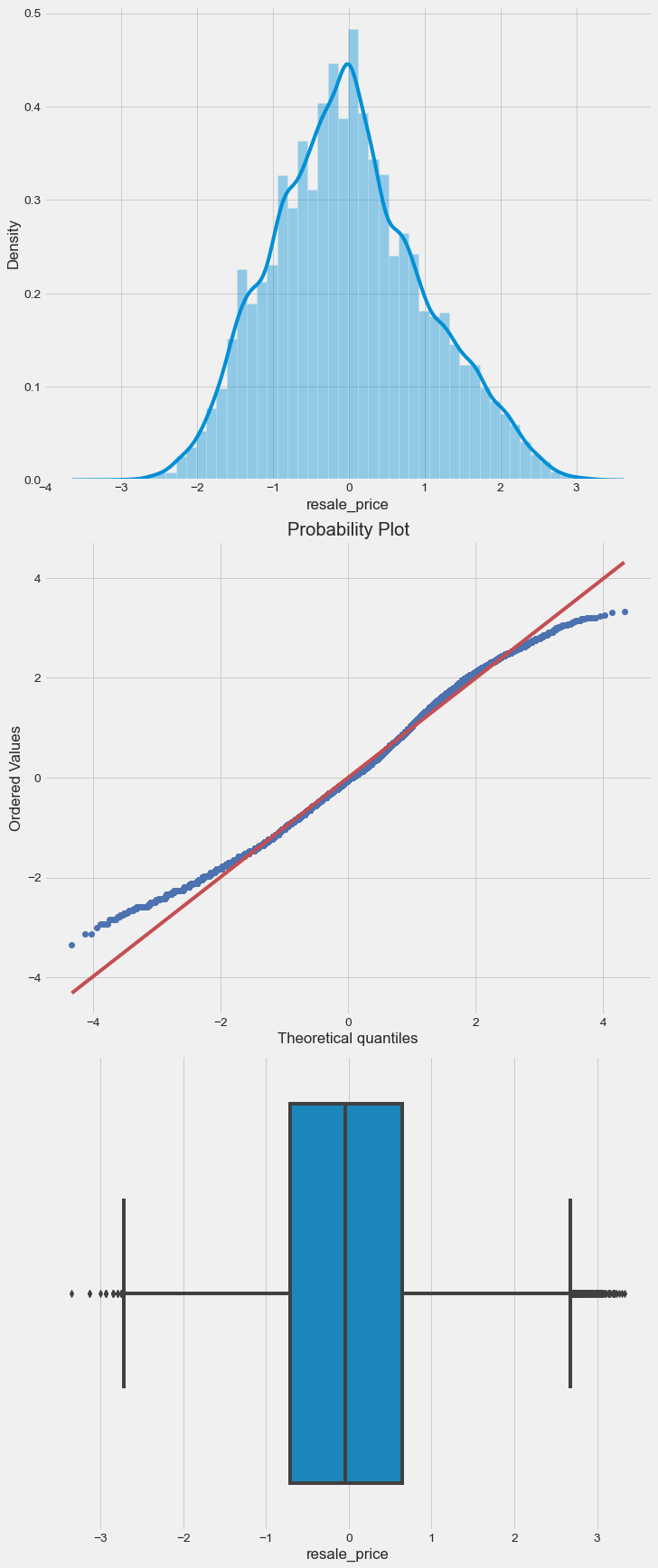
**各个Flat\_Type均价及编码结果**

**1·4 数据标准化（对数变换+标准化—减均值除以标准差）**

**df4['resale\_price'] = np.log1p(df4['resale\_price'])**

**df = (df-df.mean())/df4.std()**

**标准化之后查看概率分布函数图以及Q-Q分位数图，数据基本符合正态分布。**



**2·描述性统计（原始数据）**



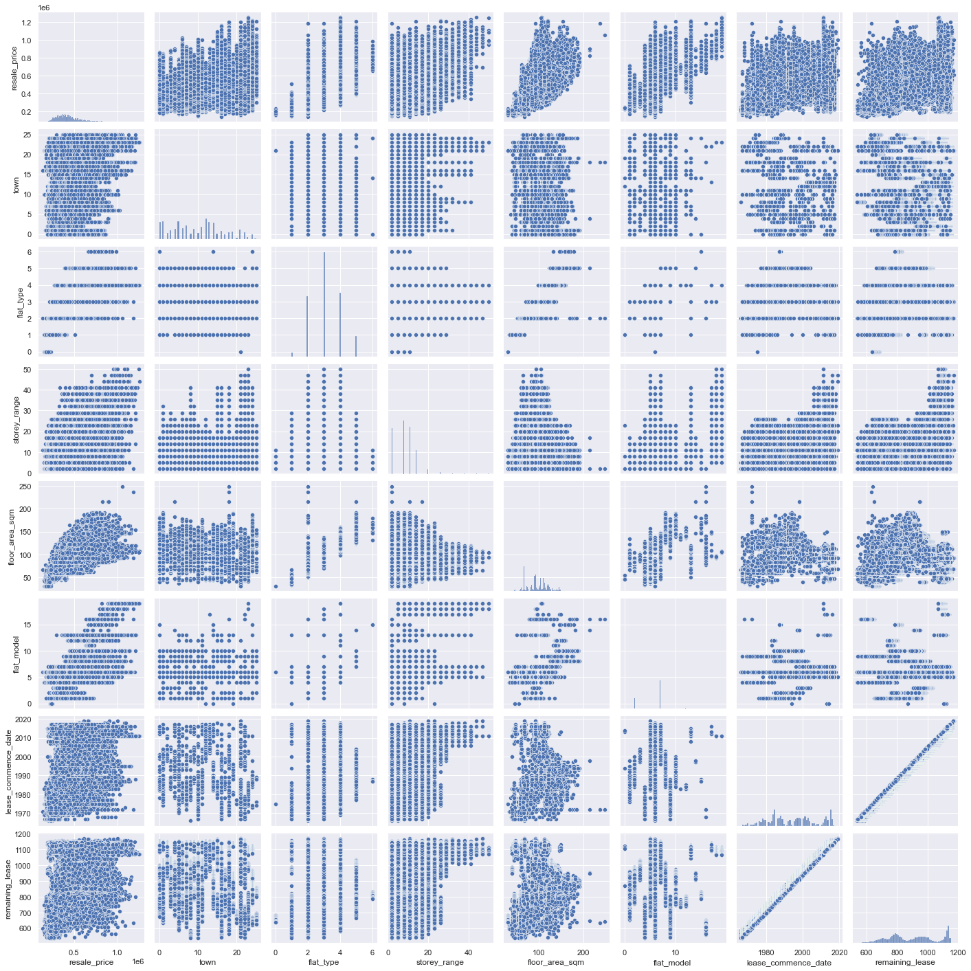
各个属性的描述性统计（计数、均值、标准差、分位数）

# 3·相关性

**3·1 计算相关系数**

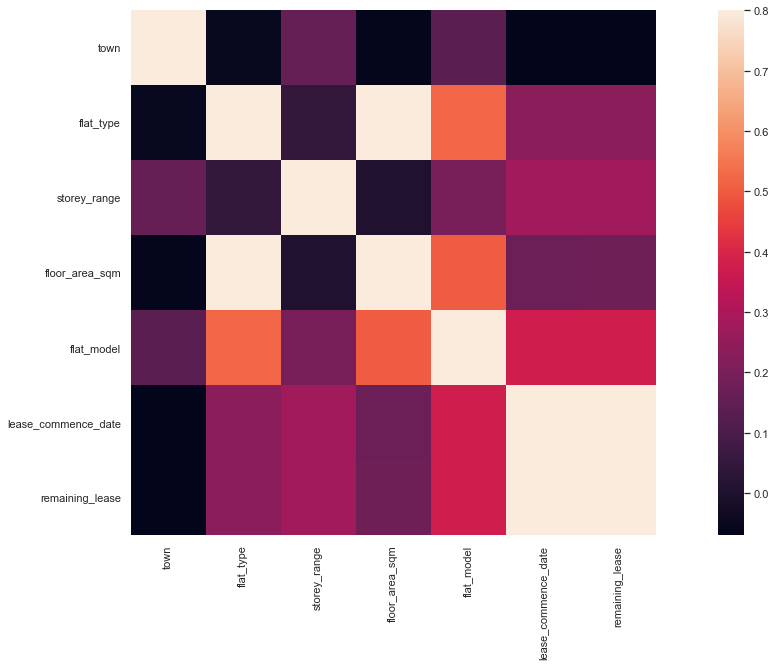
[售价，城镇，几室，楼层，面积，级别，出租日期，剩余租期]

['resale\_price','town','flat\_type','storey\_range','floor\_area\_sqm','flat\_model','lease\_commence\_date','remaining\_lease']

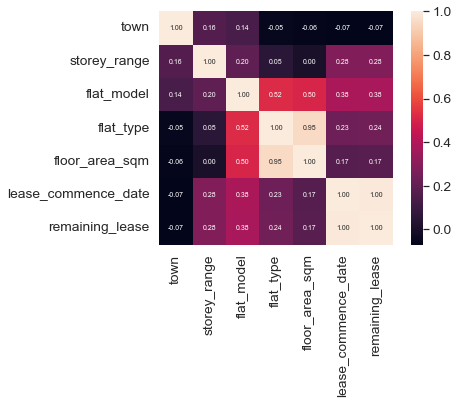


变量间相关关系图

很明显'lease\_commence\_date'与'remaining\_lease'之间有很强的相关关系，后面考虑将两个变量合并。



**相关性热力图**



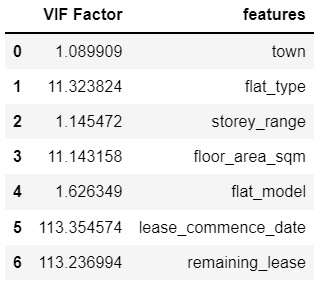
**皮尔森相关系数图**

颜色越浅相关系数越强，很明显'floor\_area\_sqm'与’flat\_type’也有很强的相关性，说明在后续多元线性回归中两变量有很强共线性，考虑合并。

**3·2 计算方差膨胀因子**

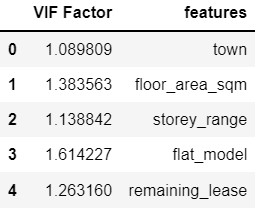
**变量合并前**

计算方差膨胀因子 ，我们可以通过计算方差膨胀系数VIF来检验回归模型是否存在严重的[多重共线性](https://baike.baidu.com/item/%E5%A4%9A%E9%87%8D%E5%85%B1%E7%BA%BF%E6%80%A7/10201978?fromModule=lemma_inlink)问题。VIF值越接近1，多重共线性越轻，反之越重。可以发现’flat\_type’与'floor\_area\_sqm','remaining\_lease'与’lease\_commence\_date’两对变量的VIF值很大，需要进一步处理。



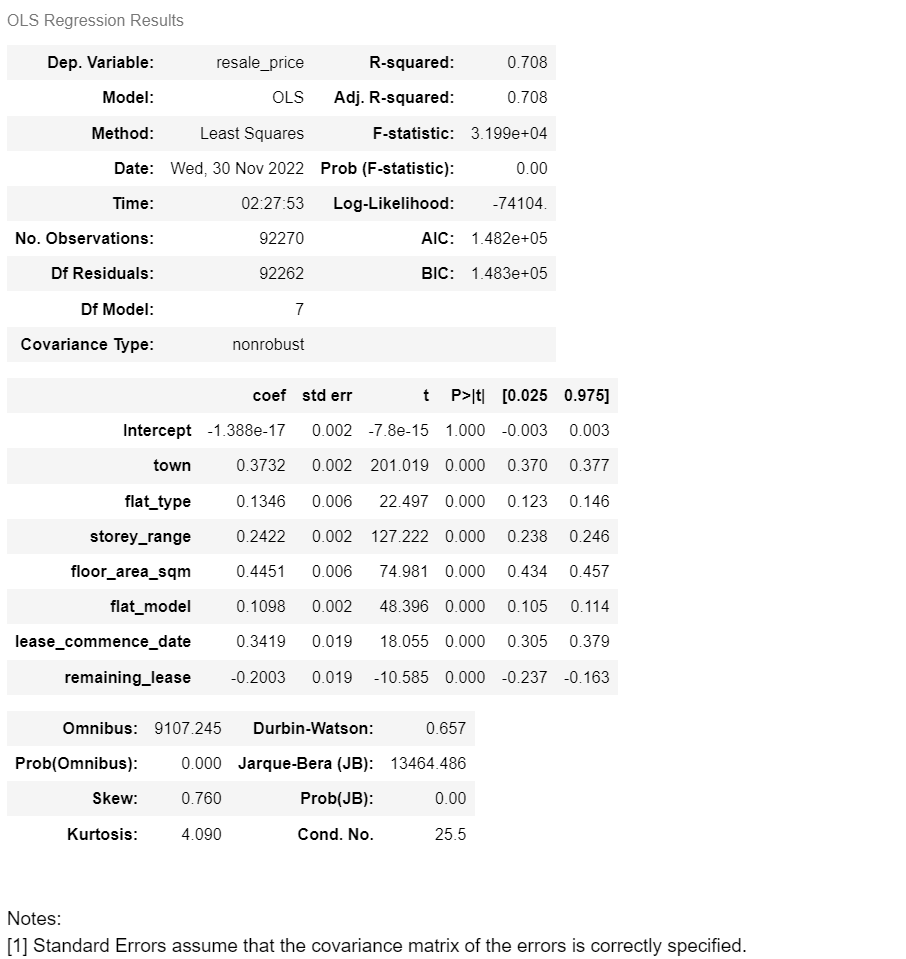
**变量合并后**

保留'town','floor\_area\_sqm','storey\_range','flat\_model','remaining\_lease'五个特征变量，将’flat\_type’与'floor\_area\_sqm'合并，'remaining\_lease'与’lease\_commence\_date’合并。再次计算VIF，发现VIF值减小，回到可以接受的范围。

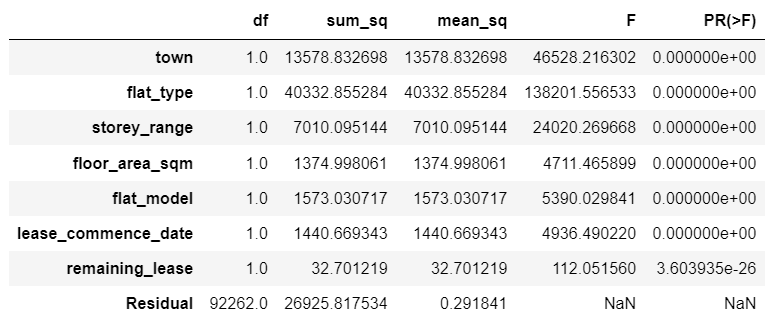


**4·多元线性回归**

**4·1·1 多元线性回归（变量合并前）**



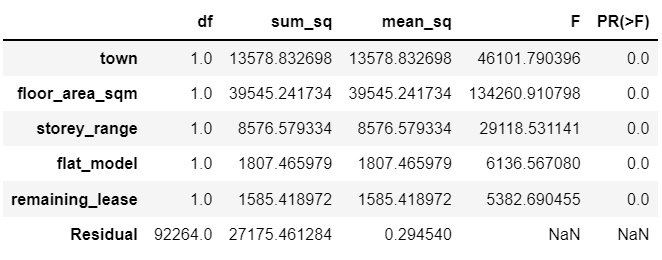
**4·1·2方差分析**



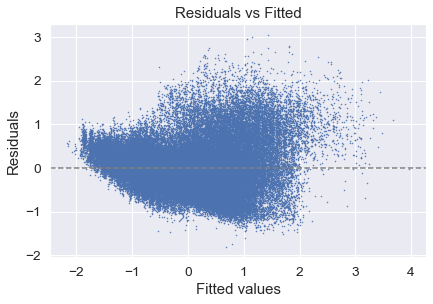
**4·2·1 多元线性回归（变量合并后）**



**4·2·2 方差分析**



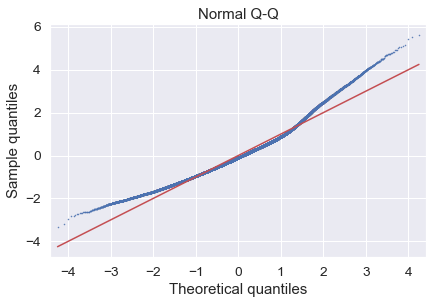
**4·2·3 残差分析**



残差拟合图

这张图中横轴是y值（Fitted value），纵轴是残差（Residuals）。在这幅图中，我们看到残差的分布是比较均匀的，这样就代表误差分布符合Guaasian-Markov Condition。

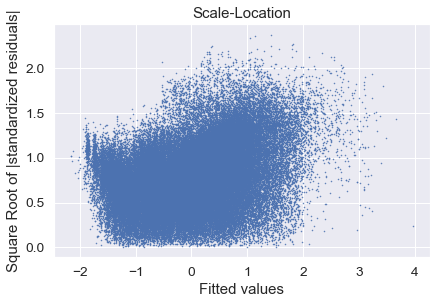
R2=0.762，大于0.5，统计模型拟合效果较好，残差相抵。



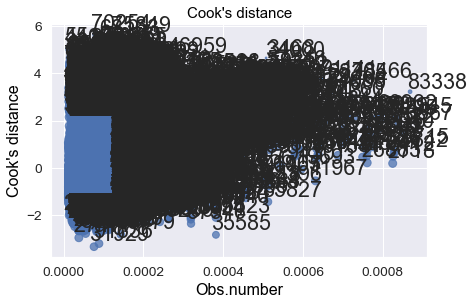
QQ图

残差QQ图，用来描述残差是否符合正态分布。通过QQ图清晰可见，残差大致呈正态分布。

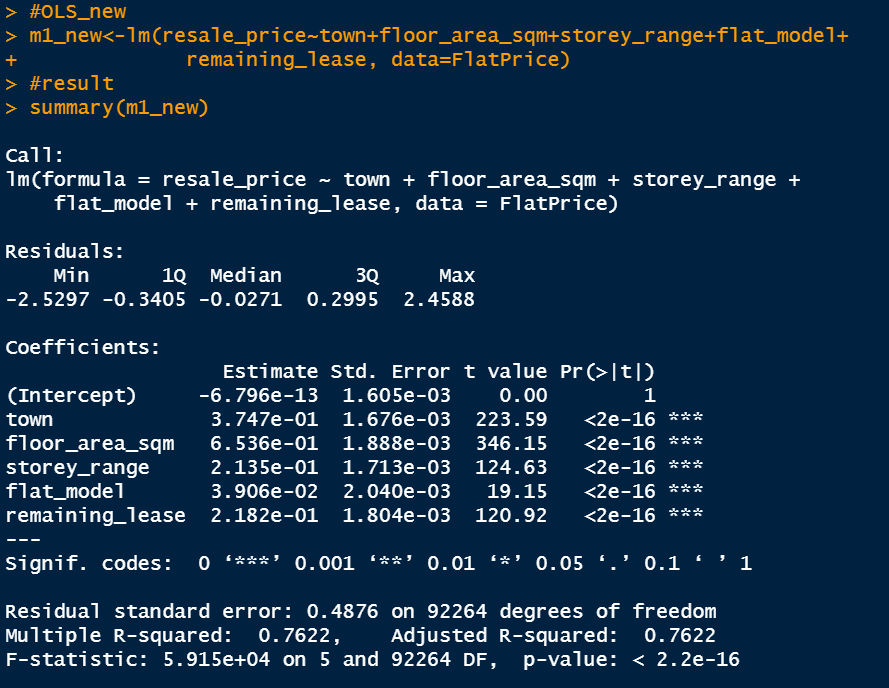
Q-Q图，全称Quantile-Quantile图，是把两个分布的quantile放在一起进行比较，来判断这两个分布是不是相似的。这幅图的作用就是检验误差是不是服从正态分布。如果是，这张图上的点将会贴近 y=x 这条直线。



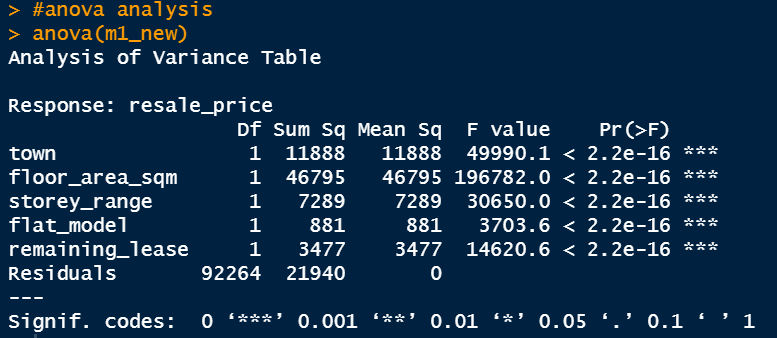
标准化残差对杠杆值:通常用Cook距离度量的回归影响点。



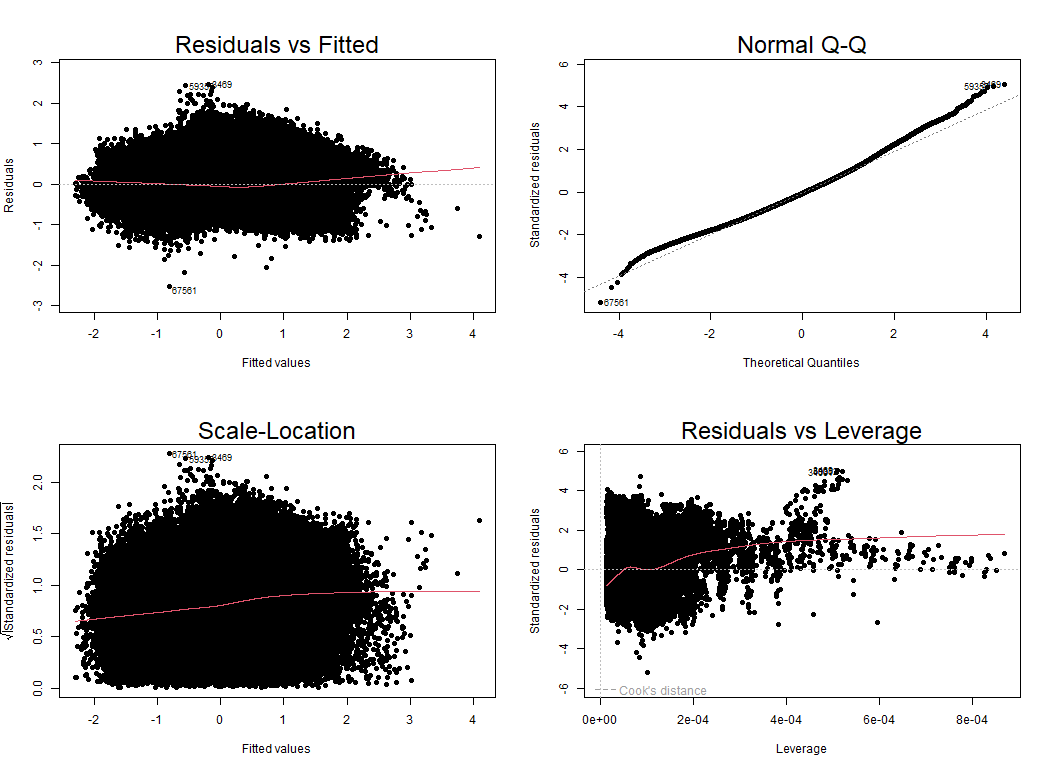
**R语言多元线性回归：**



**R语言方差分析：**



**R语言残差分析：**



左上图是残差对拟合值作图，整体呈现出一种先下降后下升的模式，显示残差中可能还存在未提炼出来的影响因素。

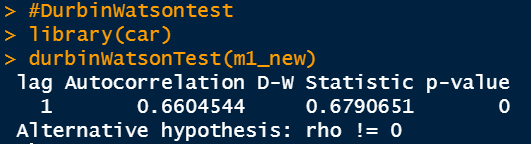
右上图残差QQ图，用以观察残差是否符合正态分布。5935与3469两点偏离。

左下图是标准化残差对拟合值，用于判断模型残差是否等方差。5935、3469、87561为离群点。

右下图是标准化残差对杠杆值，虚线表示的cooks距离等高线。我们发现5935与3469两点偏离，对样本有较大的影响

the residual plots and the leverage graph

DurbinWatson 检验：（检验残差之间的相关性）



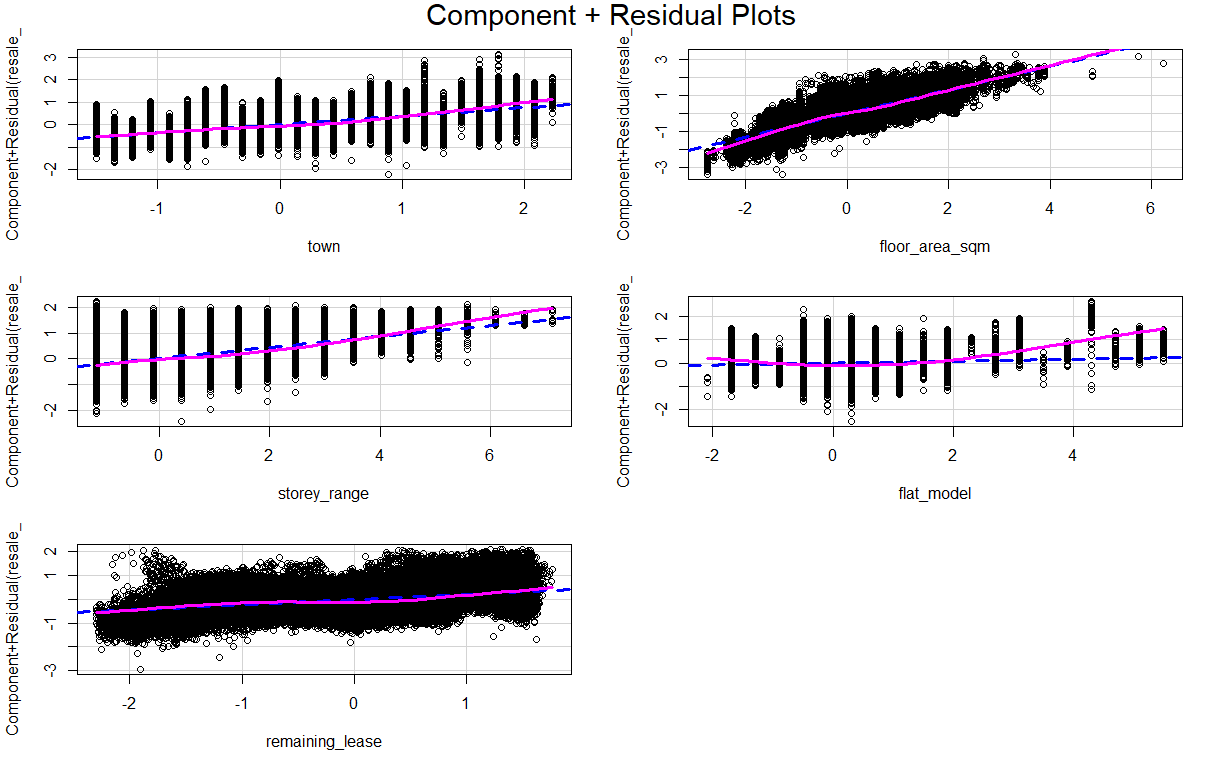
H0:There is no correlation among the residuals.

Ha:The residuals are auto correlated.

From the output we can see that the test statistic is 0.6790651 and the corresponding p-value is 0. Since this p-value is smaller than 0.05, we can reject the null hypothesis and it is possible that there is a correlation between the residuals.

从输出中我们可以看出检验统计量为0.6790651，对应的p值为0。由于这个p值小于0.05，我们可以拒绝零假设，残差之间可能存在相关性。





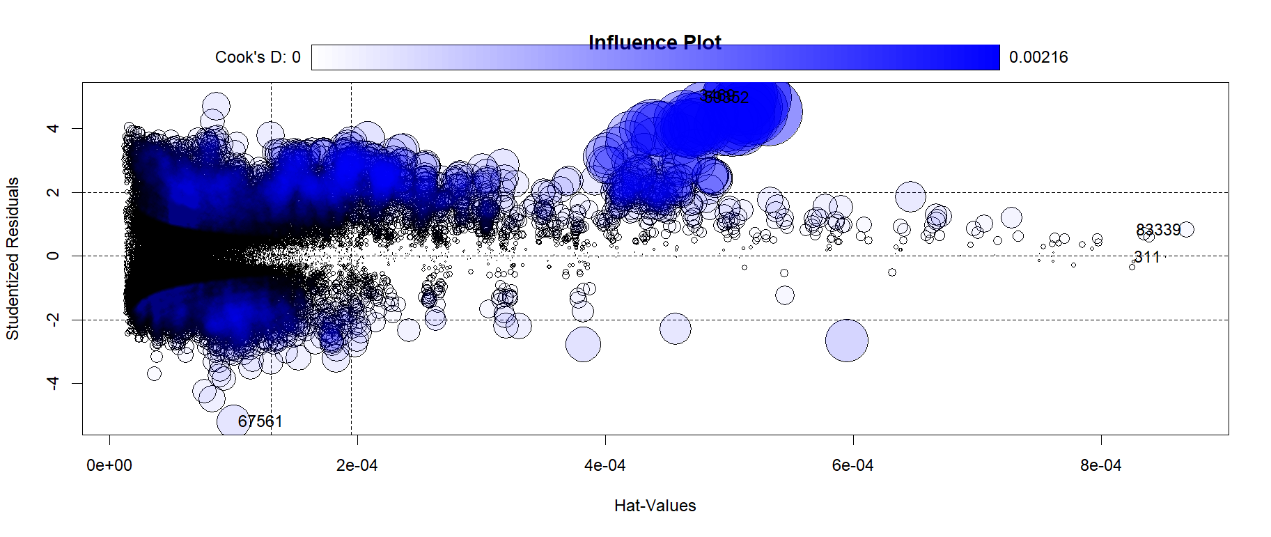
H0:the model is linear.

Ha:The model is non-linear.

From the residuals vs fitted diagram, it can be found that the model is roughly linear, showing a uniform distribution above and below the 0 line. The same conclusion can be drawn by observing the principal component residual plot through the crPlots(m1) function. The model is roughly linear, if it is nonlinear you may need to add polynomials, or transform variables (such as log transformation), or use other regression variants.

从residuals vs fitted diagram可以发现模型是大致线性的，表现为均匀分布在0线上下。 通过crPlots(m1)函数观察主成分残差图也可以得出同样的结论。 模型是大致线性的，如果是非线性可能需要添加多项式，或变换变量（如对数变换），或使用其他回归变体。

Influence Plot：



IRLS模型：

