基于ARMA模型对二手房销售预测

# 目标

销售数据是随着时间变化的序列，通过对未来的销售进行预测，方便对人员、物料等各种资源投入的把控，控制好库存，减少浪费，也可以制定未来的营运策略，提高管理效率。

由于房地产普遍具有较强的异质性、不易流动性、不可分割性等特征,致使对二手房资产的定价研究难度较大。时间序列分析是处理动态问题的一种重要的数学工具,而二手房的价格波动显然与时间相关，现搜集的北京地区的2020年1月-2020年12月的二手房房价数据,利用时间序列分析作为工具建立了二手房价格预测的随机性模型;并把预测的价格和实际价格作了对比分析，研究发现两者的误差较小，对二手房交易的买卖双方能够提供一些合理的参考价值。  这里使用ARMA算法，不仅与前P期的序列值有关，也与前q期的随机扰动有关

# 模型

ARMA模型(auto regressive moving average model)自回归滑动平均模型，模型参量法高分辨率谱分析方法之一。这种方法是研究平稳随机过程有理谱的典型方法，适用于很大一类实际问题。它比AR模型法与MA模型法有较精确的谱估计及较优良的谱分辨率性能，但其参数估算比较繁琐。

ARMA模型参数估计的方法很多：

如果模型的输入序列{u(n)}与输出序列{a(n)}均能被测量时，则可以用最小二乘法估计其模型参数，这种估计是线性估计，模型参数能以足够的精度估计出来；

许多谱估计中，仅能得到模型的输出序列{x(n)}，这时，参数估计是非线性的，难以求得ARMA模型参数的准确估值。从理论上推出了一些ARMA模型参数的最佳估计方法，但它们存在计算量大和不能保证收敛的缺点。因此工程上提出次最佳方法，即分别估计AR和MA参数，而不像最佳参数估计中那样同时估计AR和MA参数，从而使计算量大大减少。

( Percival, Donald B.; Walden, Andrew T. (1993). Spectral Analysis for Physical Applications. Cambridge University Press. ISBN 052135532X.)

# 采集数据

现在通过一定的途径，获得了2020年全年的二手房销售数据，记录了如下内容——

1. 合同编号
2. 成交公司
3. 成交时间
4. 所在区县
5. 街　　道
6. 街道(或小区) 院门牌号
7. 楼　　号
8. 建筑年代
9. 层　　数
10. 建筑面积
11. 户　　型
12. 装修标准
13. 用　　途
14. 成交单价
15. 房屋性质
16. 房屋类型

# 数据处理

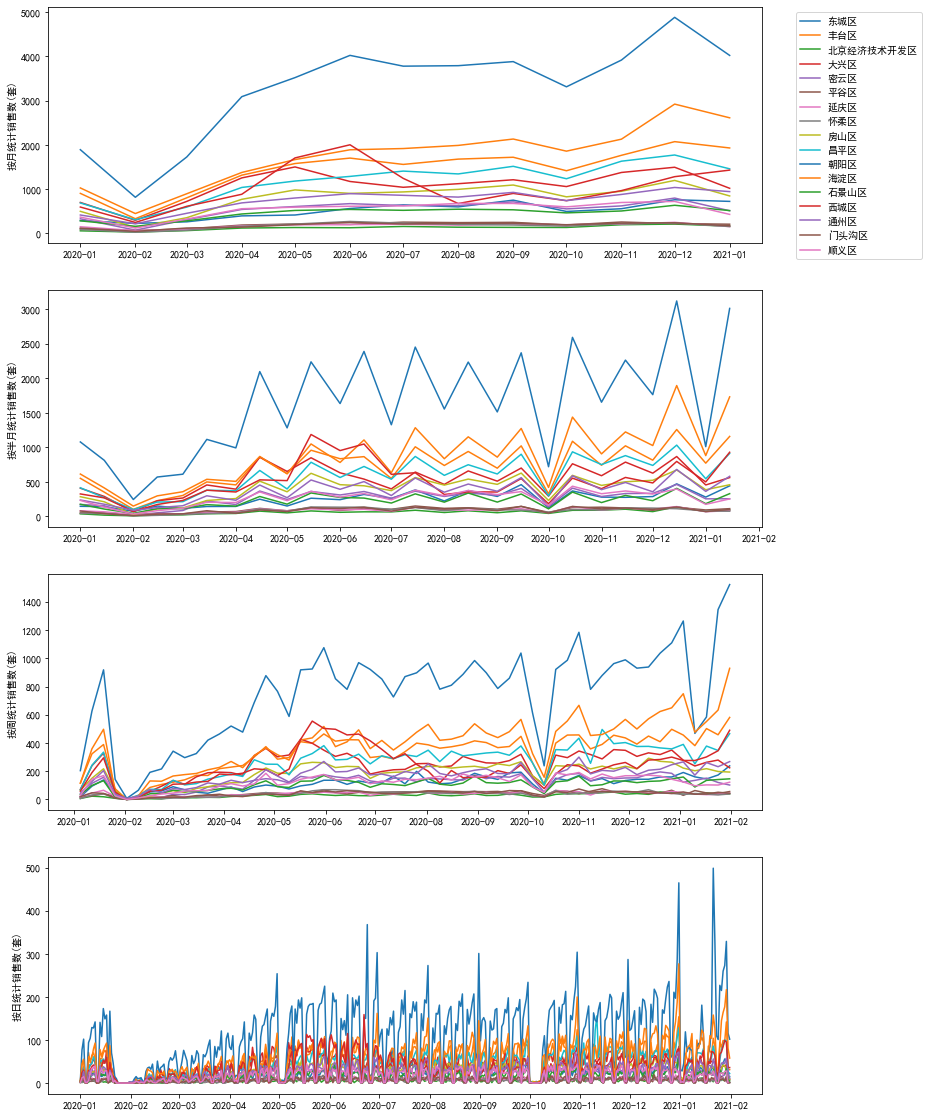
1. 把收集的数据合并
2. “用途”只记录“普通住宅”
3. 按照“月”、“半月”、“周”、“日”的方式统计数据



从数据上看，2月过年期间和10月国庆期间成交量很低

过年后遭遇怡情，成交量受限，随着金三银四慢慢提升

因为不动产非快销品，成交量小，周期长，所以不宜按日统计。只有一年的数据，按月统计量不足，不方便做时间序列分析，最后我们锁定按周统计。



交易量看，东城的交易量最大

# 平稳性检测

平稳ARMA(p, q)的均值和方差都是常数，用什么模型主要看自相关系数和偏自相关系数。如果自相关拖尾，偏自相关截尾，即p≠0 且q=0 则用AR，反之MA，如果p阶拖尾，q阶也拖尾，则ARMA。p代表周期、q代表噪音。

## 这里采用两种方法来检验：

### 1、ACF(自相关函数)

y(t, s)=E(Xt - µt)(Xs - µs)

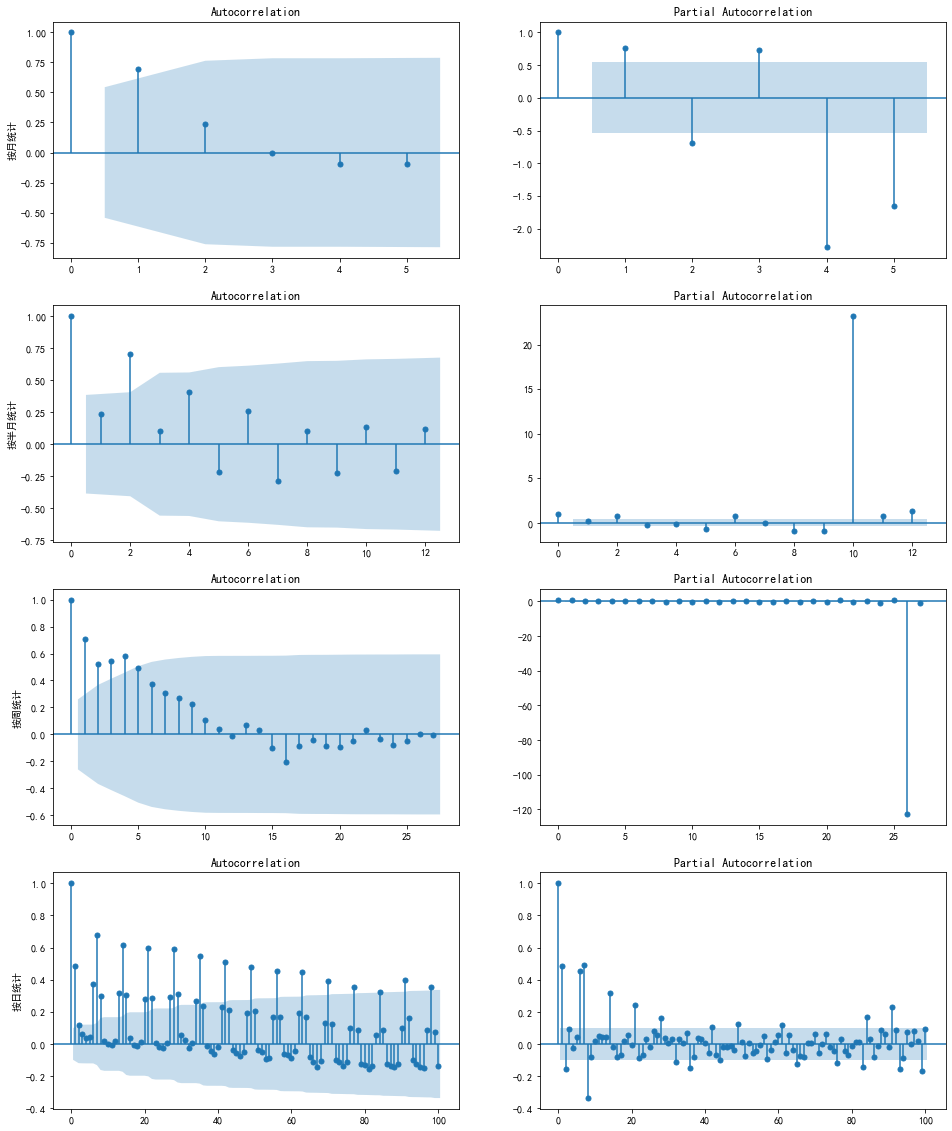
定义ρ(t,s)为时间序列的自相关系数，为ACF

ρ(t, s)=y(t, s) / sqrt(DXt \* DXs)

E为期望，D为方差

### 2、PACF(偏自相关函数）

自相关系数ρ(t,s)并不是只有两个点t和s的数据决定的。而是还包含了t-1 ~ s+1时间段值的影响。而PACF是严格这两个变量之间的相关性。



x 轴表示滞后值，y 轴上 -1 和 1 之间则表现了这些滞后值的正负相关性。蓝色区域为置信区间，默认情况下，置信区间这被设置为95％，这表明，这段代码之外的相关值很可能是相关的，而不是统计上的意外。滞后值为 0 相关性为 1 的点表示观察值与其本身 100% 正相关。自相关系数会很快衰减向0，所以可认为是平稳序列

我们发现10月份的噪声很大

# 建模

## 预测（forecast）

对尚未发生或目前还不明确的事物进行预先的估计和推测，是在现时对事物将要发生的结果进行探讨和研究，简单的说就是指从已知事件预测未知事件。

## 时间序列预测（Time Series Forecasting）

通过分析时间序列，根据时间序列所反映出来的发展过程、方向和趋势，进行类推或延申，借以预测下一段时间或以后若干时间段内可能达到的水平。

## 平稳型时间序列（Stationary Time Series）

平稳时间序列，其统计特征不随时间变化二变换，一般采用均值、方差或者协方差来作为统计特征。

## 差分（Integrated）

差分是一种让时间序列数据平稳的常用手段，异界差分的公式为：。n阶差分，是在n-1阶差分的基础上，按照一阶差分的公式进行计算。

## 常用的时间序列预测模型：

AR(p)模型（Autoregressive Model）：自回归模型描述的是当前值与历史值之间的关系

MA(q)模型（Moving Average Model）：移动平均模型描述的是自回归部分的误差累计

ARMA模型：所谓ARMA模型，是指将非平稳时间序列转化为平稳时间序列，然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型。

# 预测

## 单位根检验（Dickey-Fuller test）

|  | **value** |
| --- | --- |
| **Test Statistic Value** | -1.323177 |
| **p-value** | 0.618503 |
| **Lags Used** | 3 |
| **Number of Observations Used** | 53 |
| **Critical Value(1%)** | -3.560242 |
| **Critical Value(5%)** | -2.91785 |
| **Critical Value(10%)** | -2.596796 |

p-value的值过大，序列不平稳

所以我们做一阶差分

|  | **value** |
| --- | --- |
| **Test Statistic Value** | -7.229506 |
| **p-value** | 0.0 |
| **Lags Used** | 2 |
| **Number of Observations Used** | 53 |
| **Critical Value(1%)** | -3.560242 |
| **Critical Value(5%)** | -2.91785 |
| **Critical Value(10%)** | -2.596796 |

现在p-value小于0.005，序列平稳

获取一下最佳的p和q

{'aic': 0 1 2

0 NaN 920.490848 914.225882

1 922.937106 917.508865 916.173512

2 915.979224 913.228796 914.307500

3 911.144204 913.142954 915.136162

4 913.143041 915.127565 917.111637,

'bic': 0 1 2

0 NaN 924.541552 920.301937

1 926.987810 923.584920 924.274919

2 922.055279 921.330203 924.434258

3 919.245611 923.269713 927.288272

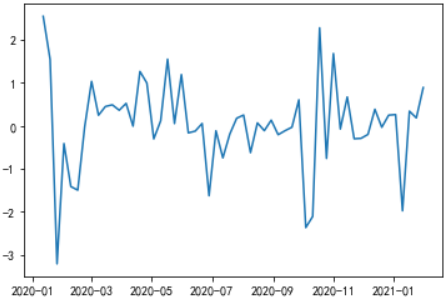
4 923.269800 927.279675 931.289099,

'aic\_min\_order': (3, 0),

'bic\_min\_order': (3, 0)}

构建模型

1. 根据刚才得到的order值设置(3, 0)
2. ARMA模型建模和训练
3. 得到模型评分
4. 绘图得到拟合曲线



观察图形，模型曲线与一阶差分曲线拟合性较好

我们总共是一年的数据，预测3个月的数据应该在可控范围内

p = model.predict(start='2021-02-01', end='2021-5-31')

2021-02-07 -748.399412

2021-02-14 -448.188481

2021-02-21 180.621879

2021-02-28 582.109958

2021-03-07 65.559721

2021-03-14 -248.217686

2021-03-21 -97.990004

2021-03-28 222.680142

2021-04-04 202.251191

2021-04-11 -2.250430

2021-04-18 -74.038174

2021-04-25 46.721330

2021-05-02 133.900483

2021-05-09 86.041075

2021-05-16 7.996484

2021-05-23 13.648458

2021-05-30 67.951916

2021-06-06 83.949756

但是预测是根据我们的一阶差分处理之后的，所以我们需要还原

最后还原的数据为——

| **还原数据** |
| --- |
| **2021-02-07** | 4756.600588 |
| **2021-02-14** | 4308.412107 |
| **2021-02-21** | 4489.033987 |
| **2021-02-28** | 5071.143945 |
| **2021-03-07** | 5136.703666 |
| **2021-03-14** | 4888.485980 |
| **2021-03-21** | 4790.495977 |
| **2021-03-28** | 5013.176119 |
| **2021-04-04** | 5215.427310 |
| **2021-04-11** | 5213.176880 |
| **2021-04-18** | 5139.138705 |
| **2021-04-25** | 5185.860035 |
| **2021-05-02** | 5319.760518 |
| **2021-05-09** | 5405.801593 |
| **2021-05-16** | 5413.798077 |
| **2021-05-23** | 5427.446535 |
| **2021-05-30** | 5495.398451 |
| **2021-06-06** | 5579.348207 |

我们再输出图像看看结果

