学校代码： 10246

学　　号：19210980081

期 末 课 程 论 文

数据挖掘

美国房价分析与预测报告——基于 Zillow 经济数据

U.S. Housing Price Analysis Based on Zillow Economic Data

院　　系：　大数据学院

专　　业：　国际商务

姓　　名：　周嘉楠

指 导 教 师：　朱雪宁

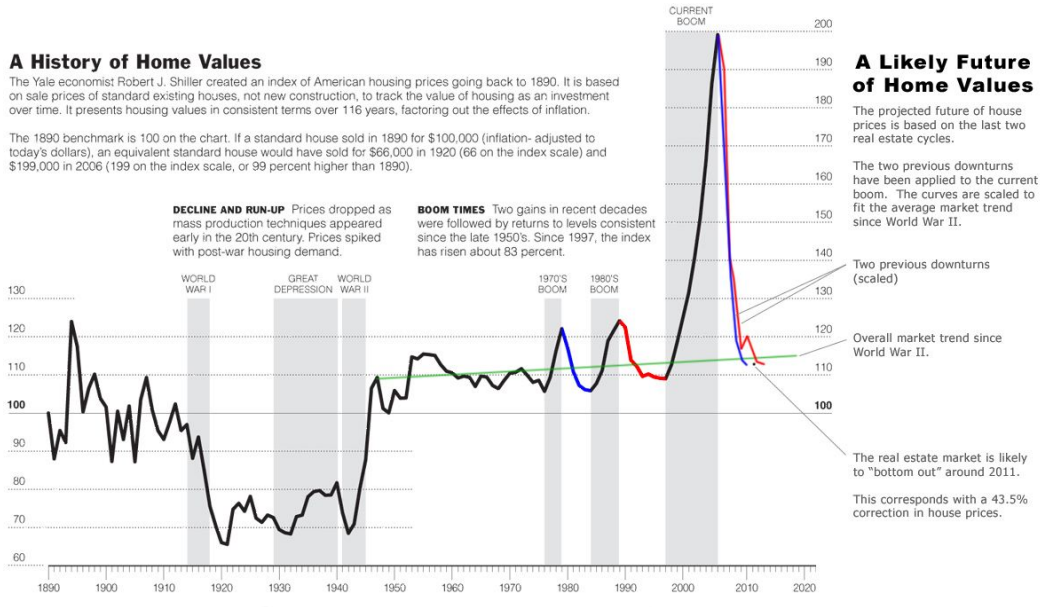
完 成 日 期：　2020 年 6 月 23 日

目 录

第一章 概 述

1.1 背景介绍

房地产业是美国国民经济的重要组成部分，是重要的基础产业。1992 年美国房地产业与整体国民经济的比例关系为：房地产业产值占所有产业产值之比为 0.95%,房地产业的工资总额与所有产业工资总额之和的比例为 1.2%，房地产业就业人口与所有产业总就业人口的比例为 1.3%。1997 年这些比例分别为房地产业产值占 0.94%, 工资额占 1.2%, 就业人口占 1.3%。虽然 1992 年与 1997 年的具体数字不同，但房地产业与整体国民经济的比例关系几乎完全一致，这不是偶然的，它说明了两个问题：一是房地产业发展与整体国民经济发展配合得非常协调，二是反映了房地产业与整体国民经济的合理比例关系，保持这种比例就足于满足生产和生活对房地产的需求。因为房地产业是作为生产和生活场所存在的，它的发展速度决定于经济增长和人们生活水平提高对其提出的需求。



1890-2020 年美国房价走势图

在过去的二十年里，美国房地产业伴随着社会的整体国民经济发展，基本保持平稳向上的发展趋势。美国房地产业也表现出周期性，一般是 18~20 年左右经历一个周期循环，但周期波动幅度较小，较少大起大落。美国主要房地产参考指数之一的“NCREIF 指数”从近 20 年来的表现情况来看，美国房地产投资回报呈现出平稳增长趋势，这也表明美国房地产市场是一个理性成熟的市场。从长线投资来看，在美国投资房地产具有风险低、收益稳定、回报率高等特点。

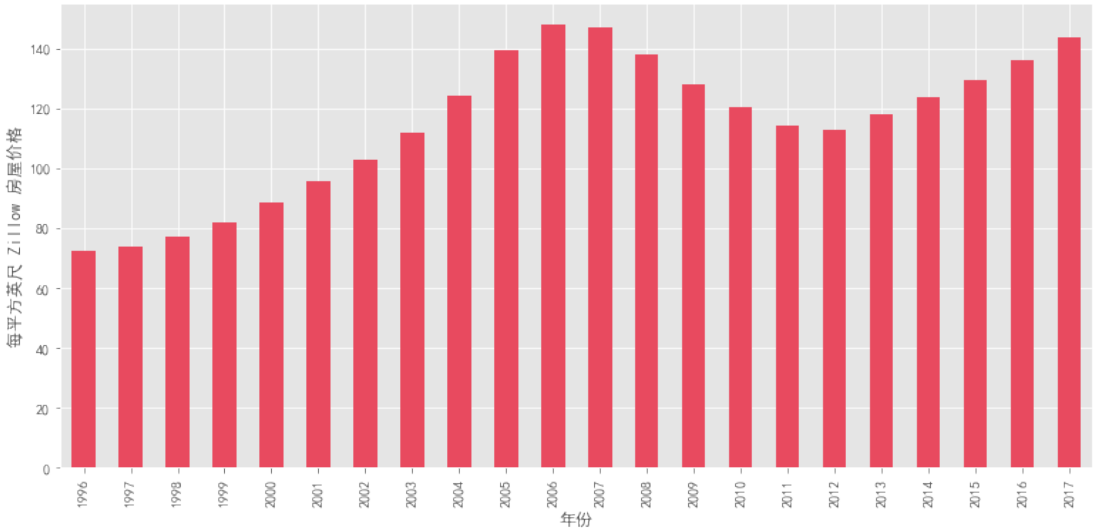
1.2 研究目标

本文将使用 Zillow 经济数据集分析美国房价具体情况，这是一份由 Zillow 的经济研究团队收集、整理和发布来自各种公共和专有资源的住房和经济数据，其中包括地方政府存档的不动产档案资料，包括契约合同、房屋登记信息和交易历史。文章将对房价数据集进行深入分析，探索美国房价的历史走势，并选取最具代表性的加利福尼亚州房价数据集，使用七种机器学习模型对房价进行预测。

第二章 数据描述

2.1 不同年份房屋每平方英尺价格变化

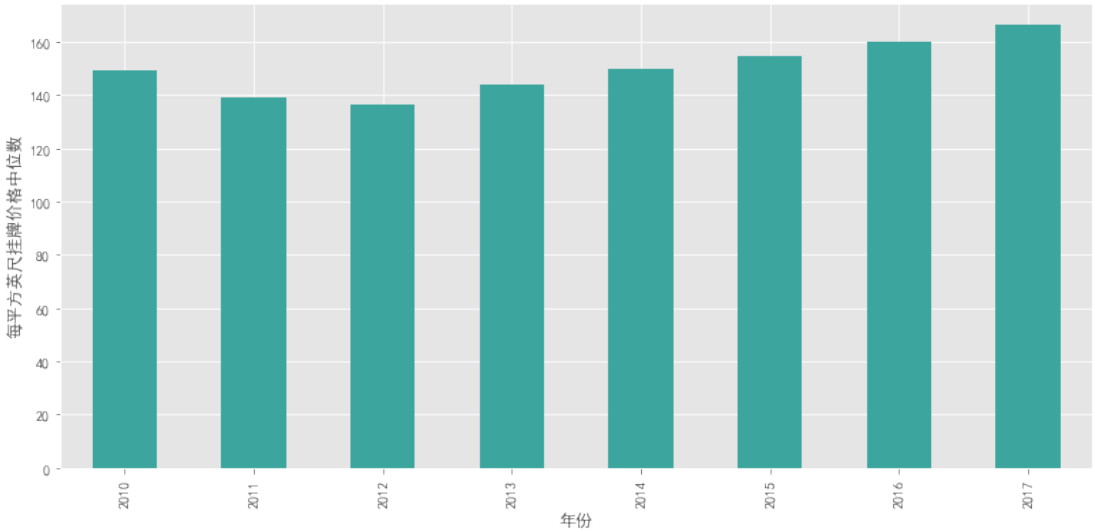
图1：不同年份房屋每平方英尺价格变化



从图 1 可以看到，在 2007 年之前，美国房屋平均价格一路飙升，Zillow 房价指数(ZHVI)从 1996 年的 78 上涨到 2006 年的 143，这 10 年间涨幅高达 83.3%，直到 2008 年因房价引起的全球金融危机的爆发，ZHVI 一直下跌到 2012 年才重新恢复上涨，而这 5 年间由 2007 年的 142 下跌至 2012 年的 112，跌幅为 21.13%，截止到 2017 年，ZHVI 指数达到了 142，已经恢复到金融危机前的最高点。

2.2 不同年份每平方英尺挂牌价格的中位数

图2：不同年份每平方英尺挂牌价格的中位数

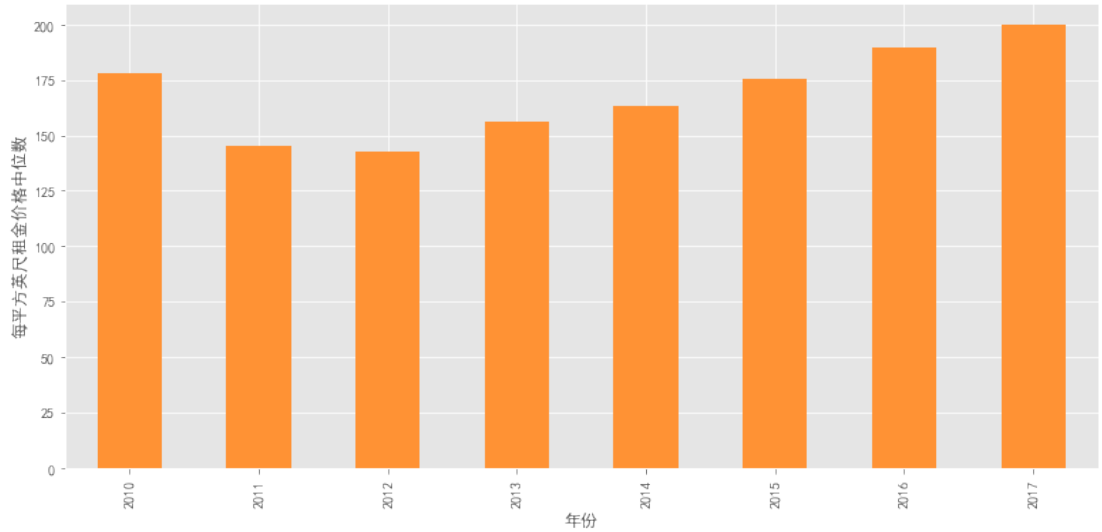


从图 2 中我们可以得知，不同年份每平方英尺挂牌价格的中位数在 2010 年至 2017 年之间变化不大，其变化趋势也是与美国整体的经济形势走势相同，值

得注意的是， 不同于全国均价，价格中位数已经超越了 2010 年的前期高点，创下了历史新高。

2.3 不同年份每平方英尺租金的中位数

图3：不同年份每平方英尺租金的中位数



从图 3 中可以看出，房屋的平均租金的 ZHVI 指数要高于平均售价的 ZHVI 指数，同时从 2010 年到 2012 年每年的变化幅度也大于同时期平均售价的变化幅度，其整体 走势与平均售价的走势是相同的。

2.4 不同户型房屋的每年价格变化情况

图4：不同户型房屋的每年价格变化情况

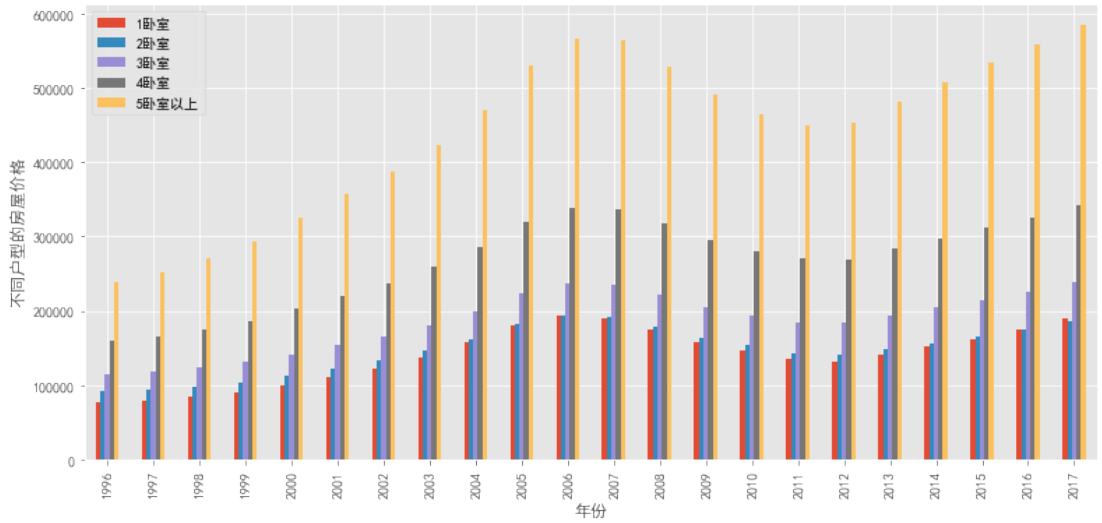


图 4 展示了不同户型房屋的每年价格变化情况，显而易见的是随着卧室数量的增多，房屋价格有非常明显的上升，其中从 4 个卧室到 5 个卧室之间价格相差幅度最大，但也可以看到当价格发生波动时，5 卧室的房屋价格也是振幅最大的，整体而言他们的变化方向和趋势是呈相同方向的。值得注意的是，1 卧室房屋和

2 卧室房屋价格 走势相当有趣，由于 1 卧室和 2 卧室是人们最常居住的户型，所以市场需求量较大，在 1996 年-2006 年之间，1 卧室房屋与 2 卧室房屋的价格相差越来越小，直到 2006 年两者的价格居然相等了，这也是证明房地产市场的白热化，导致人们对 1 卧室房屋投资和居住需求明显增大，但后来的金融危机使得 1 卧室房屋与 2 卧室房屋的价格又回归了正常差距，直到 2012 年房地产市场复苏后，1 卧室房屋价格又呈现了猛烈的上升态势，2016 年两者价格又一次持平，而 2017 年 1 卧室房屋价格甚至超越了 2 卧室房屋的价格，这是否又意味着房地产市场的过热呢？是一个值得深入思考的问题。

2.5 不同州的房屋每平方英尺价格中位数的分布情况

图 5.Median of the value of all homes per square foot in different states

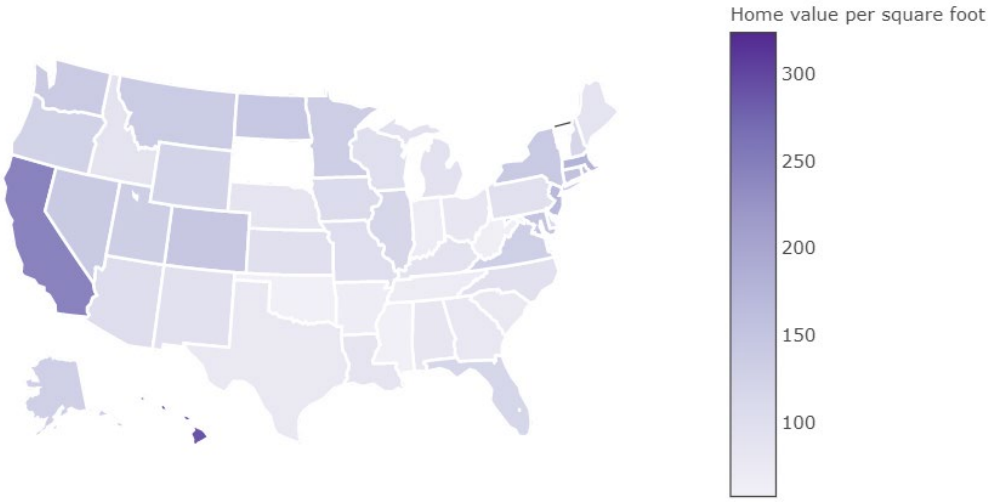
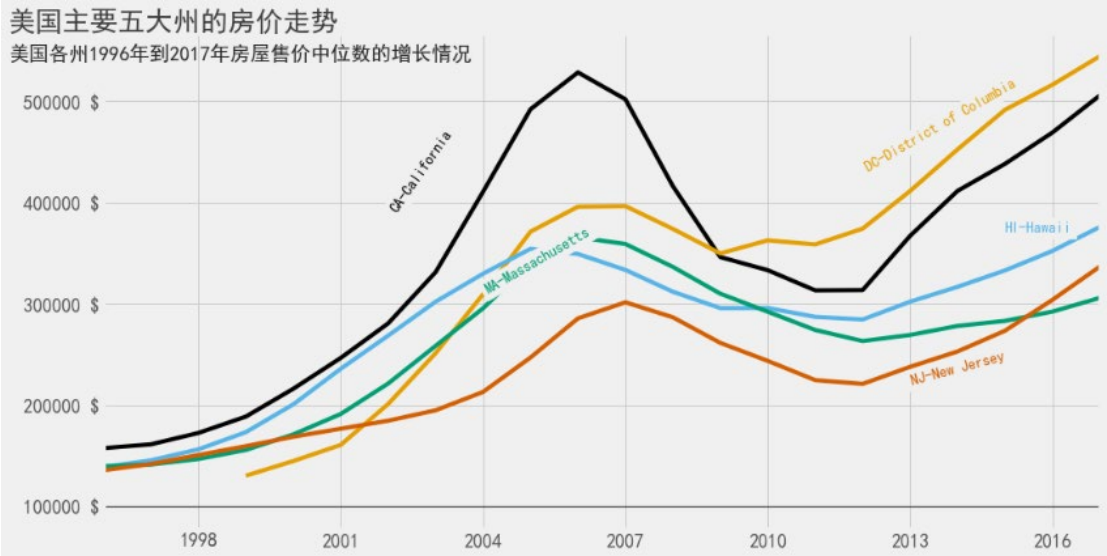


图 5 是一个动态交互的地图型可视化组件，由 Plotly 制作，可以看到美国所有州的房屋每平方英尺价格中位数的分布情况，从图中可以清晰地看到，CA（加利福尼亚州） 和 HI（夏威夷州）的房价中位数是最高的，其次是 NY、VA、MA、NJ 等美国东北部地区，这非常符合美国的经济发展分布情况，其中加利福尼亚州和夏威夷是经典的度假胜地，以高档别墅和度假酒店为主要房地产业态，是全球房价最高的地点之一；而纽约、曼哈顿这种地区是美国经济发展的核心区域，也是寸土寸金，房价当然是理所应当的高。

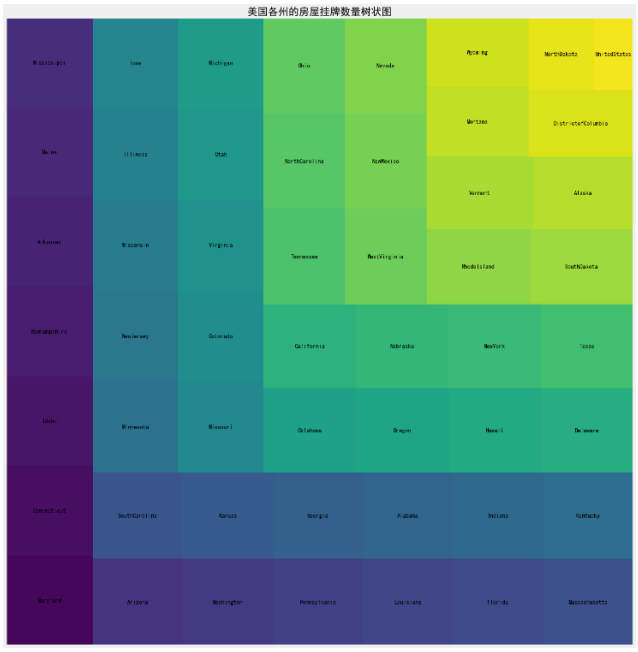
第三章 数据分析

3.1 美国主要州的房价走势



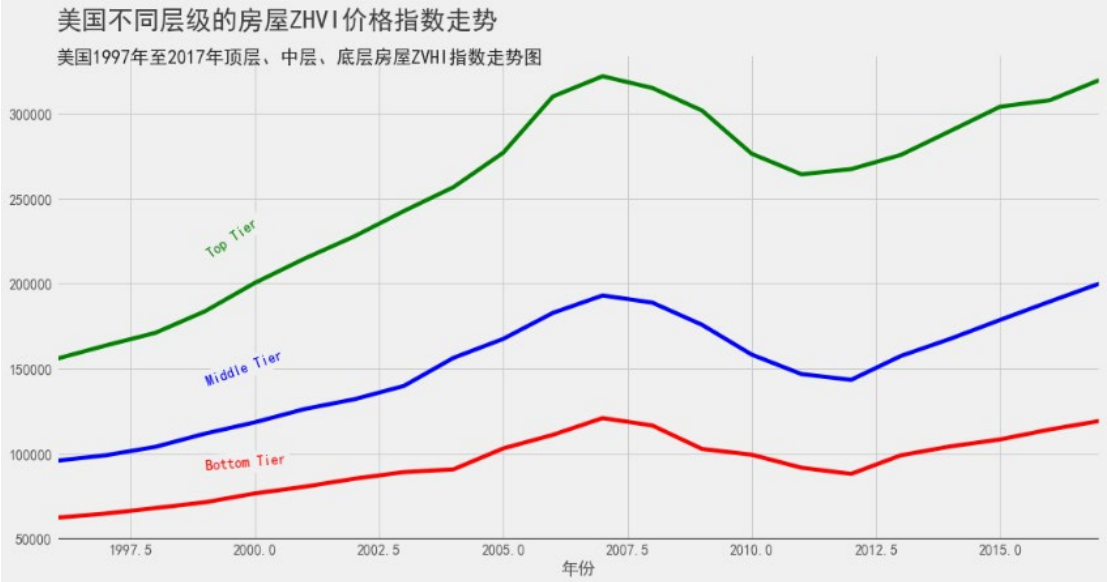
从美国主要五大州的房价走势可以看出，这与第一部分的最后一小节分析相一致，图中的五条折线代表了美国房价最高的五大洲，分别是 CA、DC、HI、NJ、MA，可以看到其中加利福尼亚州的房屋售价自 1998 年至 2007 年金融危机之前都是占据了第一名的位置，而在房产泡沫最严重的 2006 年，其涨幅也是全国第一，随后的危机给房价带来的重创是非常严重的，使得加利福尼亚州的房价下跌了 66.67%，跌幅同样是全国第一，而后随着房地产市场的复苏，加州再也没有回到昔日的辉煌状态，目前全美房价最高的是哥伦比亚州，加州名列第二。

3.2 美国各州的房屋挂牌数量树状图



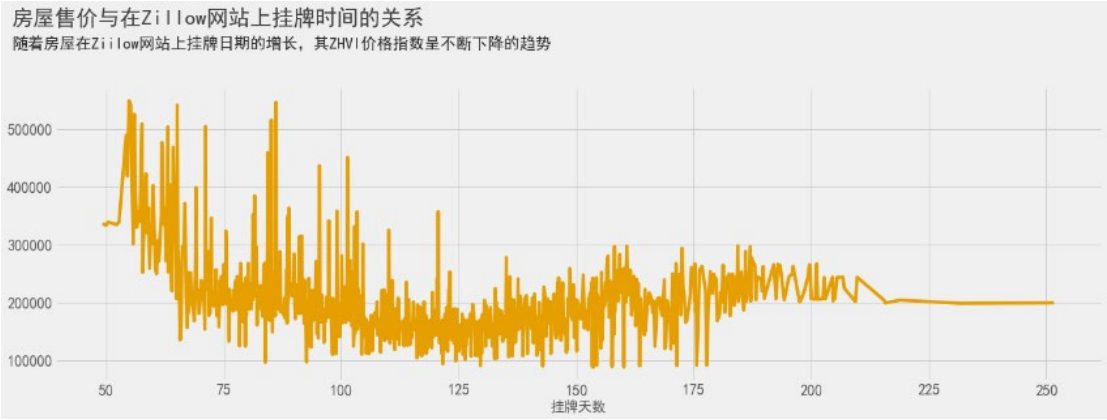
上图是美国各州的房屋挂牌数量树状图，以美观且直观的方式展现了全美五十个州房屋在 Zillow 网站上的挂牌数量，面积越大、颜色越深代表数量越多，而面积越小、颜色越浅代表数量越少，可以看到全美五十个州之间的差距并不算大，除了最右上角浅黄色的北达科他州数量相对较少，证明 Zillow 在选取数据时并不存在样本失衡的情况。

3.3 美国三种层级的房屋价格走势



从美国三种层级的房屋价格走势的折线图可以看出，三种层级的房屋整体走势的方向是相同的，都跟随美国整体经济的变化而波动，在 1997 年初，顶层房屋价格指数在 150000 美元附近，中层房屋价格指数在 100000 美元附近，而底层房屋价格指数在 60000 附近，在房地产泡沫最鼎盛的 2007 年 5 月份，三种层级房屋的价格指数差距达到了最大化，顶级房屋价格指数冲到了 330000 美元左右，而底层房屋价格指数仅在 130000 美元左右，两者差距从一开始的 90000 美元增加到了 200000 美元。

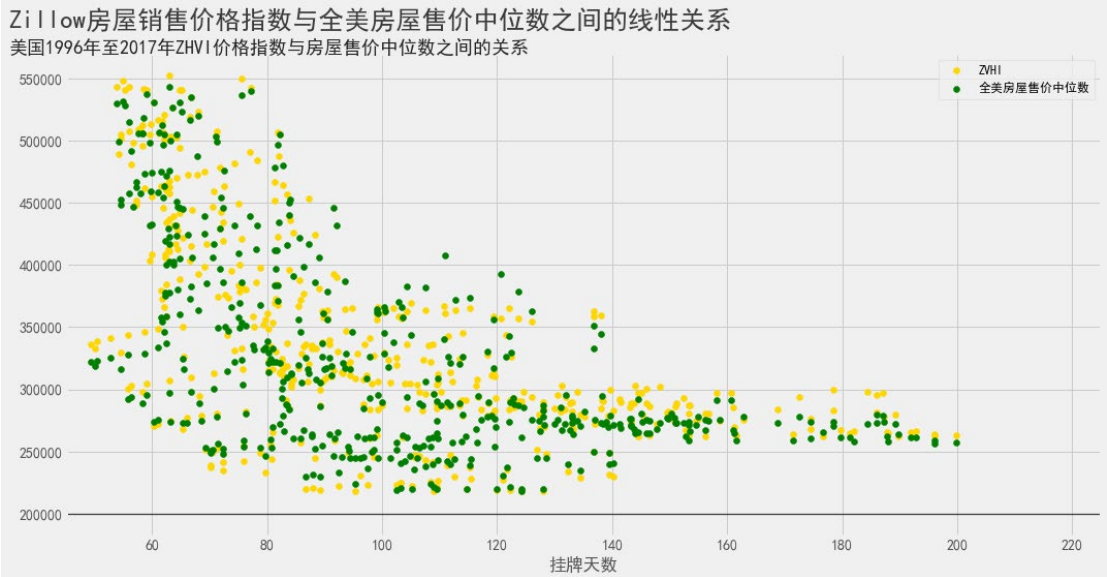
3.4 房屋售价与在 Zillow 网站上挂牌时间的关系



房屋售价与在 Zillow 网站上挂牌时间的折线关系图清晰地展示了随着房屋

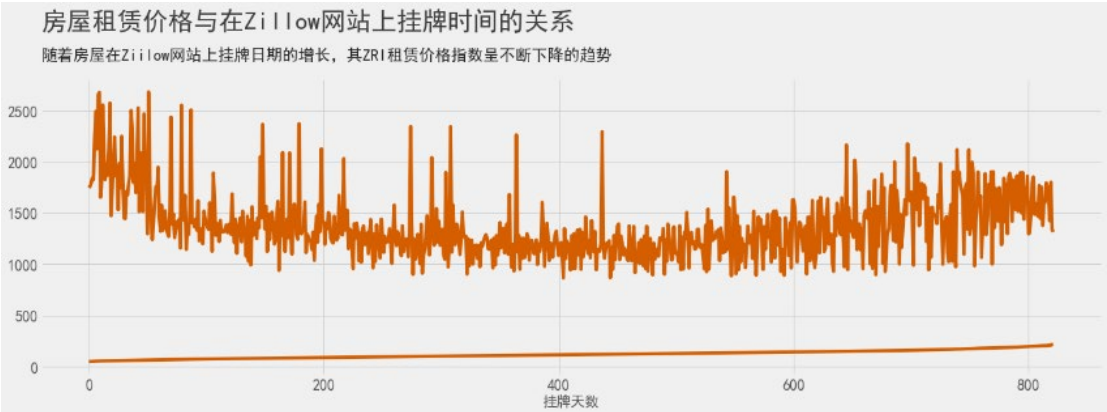
在 Ziilow 网站上挂牌日期的增长，其 ZHVI 价格指数呈不断下降的趋势。有趣的是，如果仔细观察，在最初的 55 天内，房屋价格是向上增长的，这表明了人们在出售房屋时最初总是乐观的，希望自己的房屋卖一个好的价钱，甚至会后悔自己最初挂牌价比预期的要少，但是随着时间的推移，久久不能出售的房屋在第 75 天时价格达到了低谷，其降价幅度参考最高点时达到了 65%-80%，而后会有一定的小幅反弹，如果房屋在 200 天时依然无法出售，则房屋主人将会逐渐对房屋失去信心，甚至不再维护 Zillow 上的价格，在图上的直观表现即为 210 天后价格变成了一条直线。

3.5 ZHVI 指数与全美房屋售价中位数之间的线性关系



本部分使用散点图探索了随着挂牌日期的增加，Zillow 房屋销售价格指数与全美房屋售价中位数之间的关系。

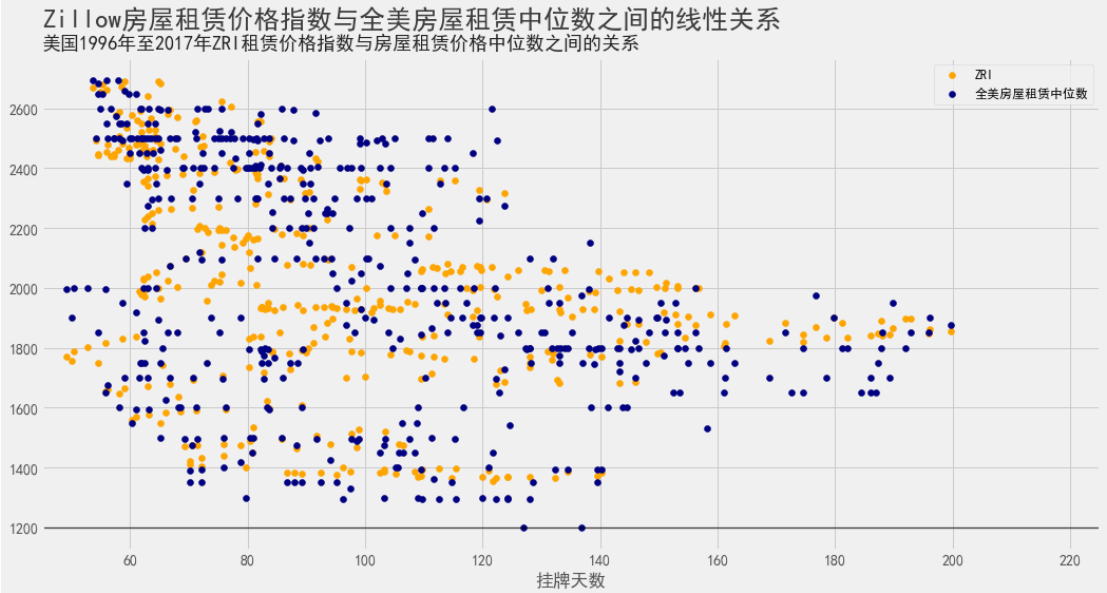
3.6 房屋租赁价格与在 Zillow 网站上挂牌时间的关系



房屋租赁价格与在 Zillow 网站上挂牌时间的关系图展示了美国房屋自挂在 Zillow 网站上随着天数的增加其 ZRI 租赁价格指数走势，与销售房屋不同，房屋租赁在美国是更普遍的现象，而挂牌日期最长也达到了 800 天，证明 Zillow 的房屋出租是一项比较长久的业务。同样地从最初的 50 天内，房屋价格会有剧烈的向上跳价现象，后期随着时间的推移逐渐下降，租赁价格在 250 天左右达到了

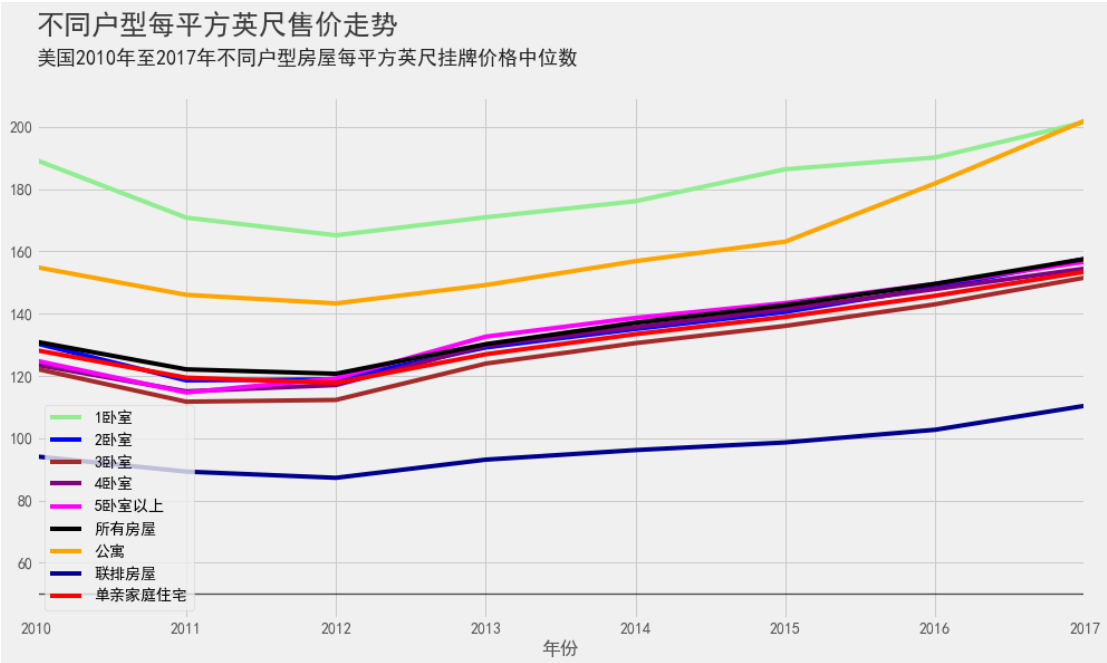
最低点，比最高点下降了 60%左右，随后在第 600 天时开始缓慢回升，在 780 天附近达到高点，价格比最低点回升 50%左右。

3.7 ZRI 价格指数与全美房屋租赁中位数之间的线性关系



本部分使用散点图探索了随着挂牌日期的增加，Zillow 房屋租赁价格指数与全美房屋租赁价格中位数之间的关系。

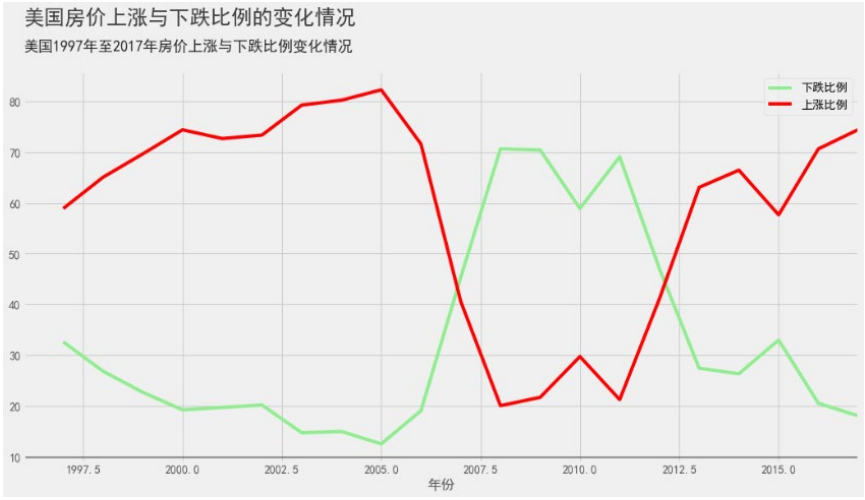
3.8 不同户型每平方英尺售价走势



以上折线图展示了美国 2010 年至 2017 年不同户型房屋每平方英尺挂牌价格中位数的变化趋势，可以看到其中 1 卧室房屋、公寓是最热门的两种户型，其价格走势一直领先于其他户型的房屋，而在近期公寓的价格走势已经超越了 1 居室房屋，成为全美国最热门且昂贵的户型，而所有户型中价格最低的则是联排房

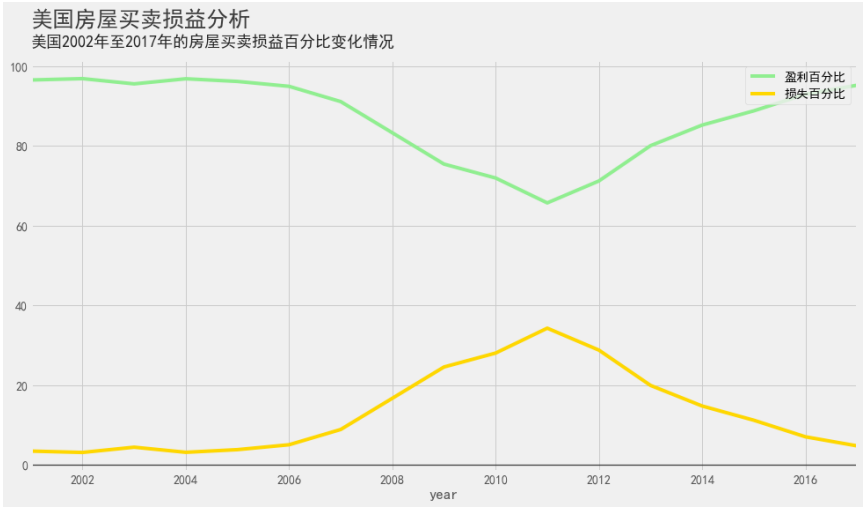
屋，这可能是由于美国的住房文化决定的，联排房屋属于别墅的一种，而大部分别墅是建立在郊区，郊区的房价普遍比市区要低很多，其中联排别墅又是别墅中比较廉价的一种，故其价格是所有户型中最低的。根据 2017 年最新数据，最昂贵的公寓每平方英尺单价要比最便宜的联排房屋高 82%。

3.9 美国房价上涨与下跌比例的变化情况



从美国房价上涨与下跌比例的变化情况图可以看出自 1997 年-2015 年在美国投资房产的大环境变化趋势，在金融危机之前，美国每年有 60%以上的房屋价格是呈上涨趋势的，而每年仅有不到 35%的房屋价格是下跌的，这一数字在 2005 年达到了最低点，2005 年仅有不足 15%的房屋价格处于下跌状态，而当年涨价的房屋高达 80%，房地差泡沫破裂后，房价下跌的比例在 2007 年中旬达到了峰值，为 70%，而当年上涨房屋的比例仅为 20%，随后即引起了全球经济的大崩盘。有趣的是房地产市场也是先于经济复苏的，自 2011 年中旬即恢复了上涨趋势，比经济全面复苏早了 6 个月左右。

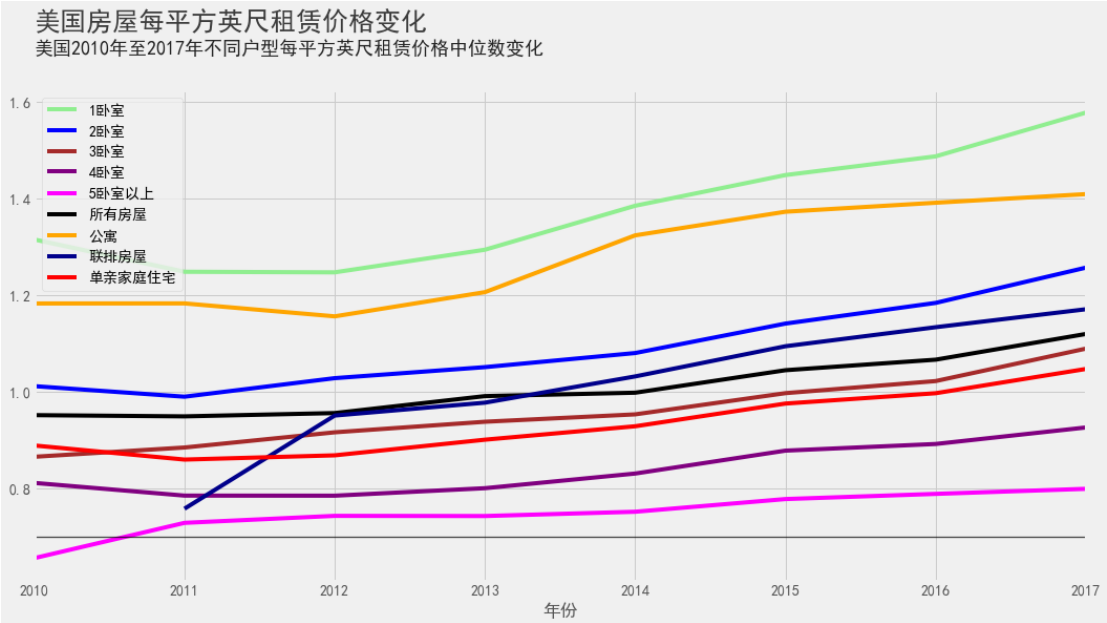
3.10 美国 2002 年至 2017 年的房屋买卖损益分析



上面的美国房屋买卖损益分析图展示了 2002 年至 2017 年的房屋买卖损益百分比变化情况，也是比较直观地显示了投资美国房地产近 15 年的整体走势情

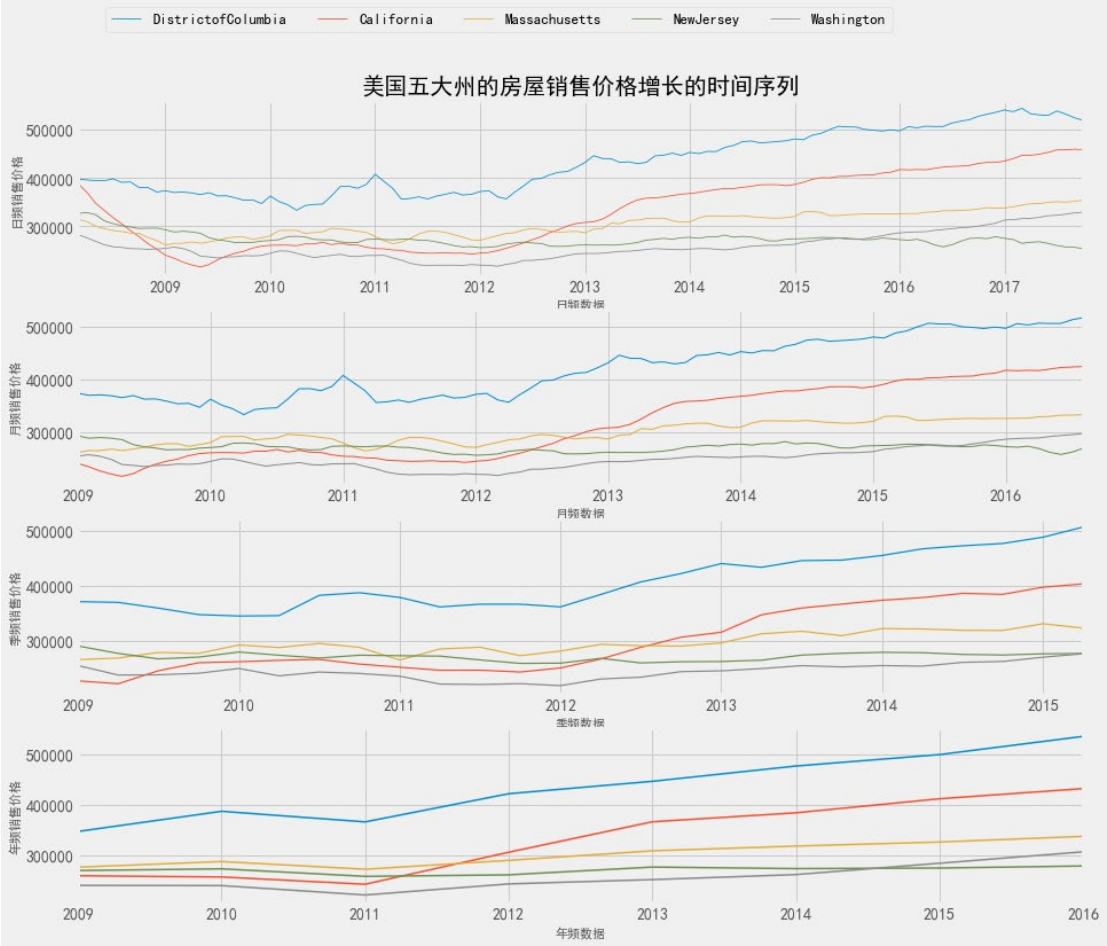
况，与美国房价上涨与下跌比例的变化情况相一致，其中 2008 年金融危机开始后一直到 2011 年复苏之前，投资房地产的最大亏损达到了 36%左右，后面用了 6 年左右的时间才恢复到前期历史水平。

3.11 美国各州不同户型房屋租赁价格中位数



从上面美国房屋每平方英尺租赁价格变化图中可以清晰地看到 2010 年至 2017 年不同户型每平方英尺租赁价格中位数变化趋势，租赁价格与出售价格既有相似之处也有不同之处，相似之处是 1 卧室和公寓这两种户型依然是所有户型的房屋里最昂贵的两种，而 1 居室房屋一直没有被公寓的租金所超越，而租金单价最便宜的居然是 5 居室，这似乎与违反了常理，按人们正常的理解，卧室数量越多的房屋无论是售价还是租金都应该是最昂贵的，但是我们需要注意这里考虑的是每平方英尺的价格，因为 5 居室的房屋面积非常大，即使整套的租赁价格很昂贵，平均到每平方英尺的价格就便宜了很多，如果 5 居室的租赁单价与 1 居室相同的话，那么整体租赁价格将会过于昂贵，超出了人们所能接受的范围。还有一点值得注意的是，联排房屋的租金在 2011 至 2012 年间有一个较大的提升，而且其综合排名一直在 4 名左右，这也是符合美国房地产租赁市场的现状，即许多生活在郊区的人们并不会选择自己购置房产，而是租赁联排房屋，所以其租赁价格并非如售价一样排名最后。

3.12 金融危机后美国五大州的房屋销售价格增长的时间序列



上图是 2008 年金融危机之后美国五大州的房屋销售价格增长的时间序列，其中我们分别按照日频、月频、季频、年频的数据频率进行了统计，可以看到随着时间跨度的增长，价格曲线越来越平滑。这里我们选取了房价最高的五个州，分别是哥伦比亚特区、加利福尼亚州、马萨诸塞州、新泽西州、华盛顿，与文章中本部分最开始统计的五大洲的房屋销售价格增长趋势相对应，其中哥伦比亚特区的房价一直领先，而金融危机前一直领先的加利福尼亚州先是遭到了重创，知道 2011 年之后才开始重新恢复上涨态势，并在 2013 年重新恢复到全国排名第二的位置。从本数据集中也可以侧面看出，哥伦比亚特区的房子一直没有遭到过度投机性的炒作，所以即使是金融危机也没有使其房价有较大的跌幅，从整体来看，哥伦比亚特区的房价自 2009 年至 2016 年的复合涨幅达到了 4.27%，与美国极低的银行利率相比，是非常稳健且健康的投资标的。

第四章 建立模型

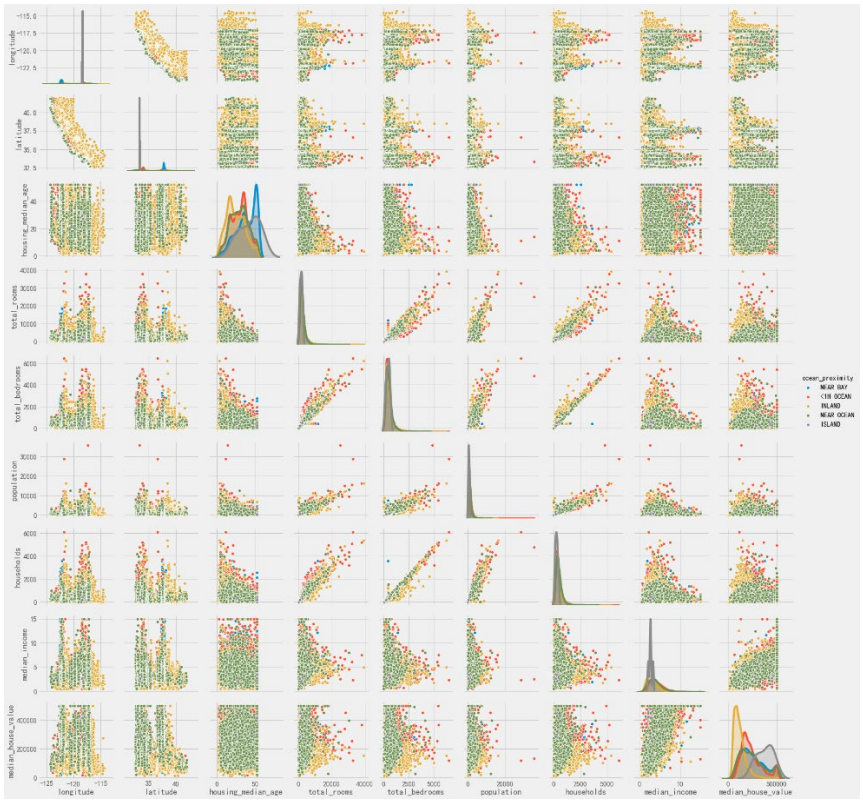
由于在上述分析中大多为时间序列数据，为了建模便利，我们在本部分选取 Zillow 经济数据库中最热门的加利福尼亚州的房价数据集，其中包含了 10 个数据维度，共计 20640 条房价数据。

4.1 数据纵览

4.1.1 变量描述

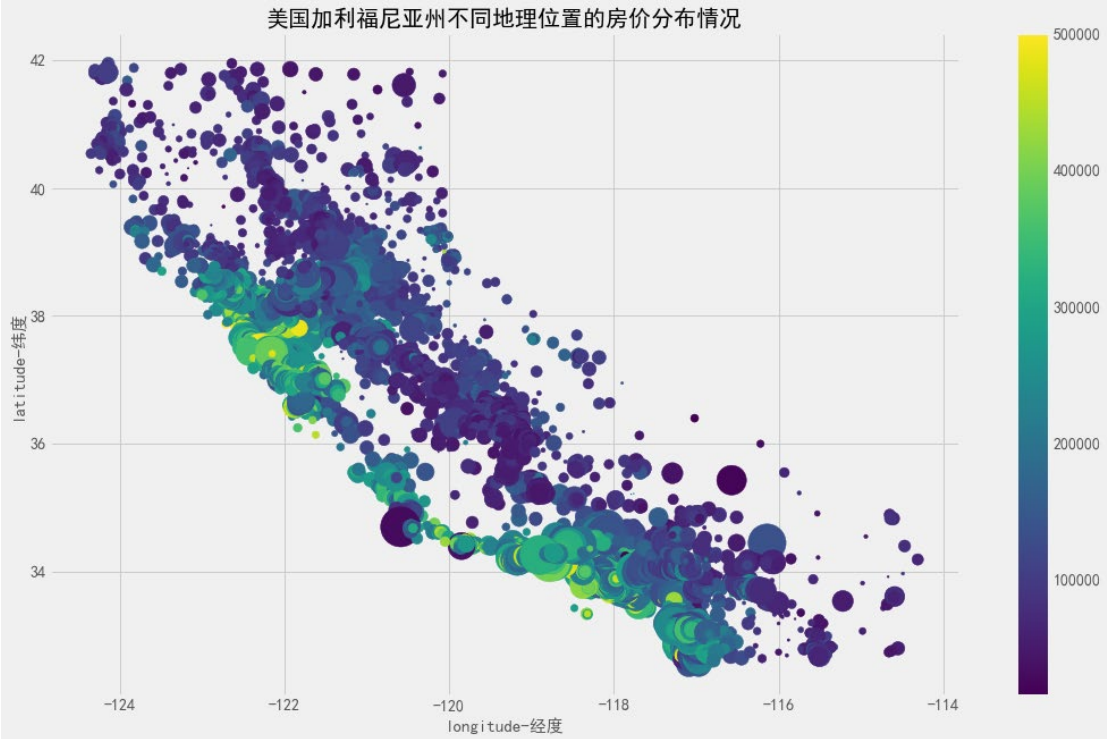
变量类型	变量名	详细说明	取值范围
因变量	median_house_value	房价中位数	14999~500001
自变量	longitude	经度	-124.35~-114.31
	latitude	纬度	32.54~41.95
	housing_median_age	房屋年龄中位数	1~52
	total_rooms	房间总数	2~39320
	total_bedrooms	卧室总数	1~6445
	population	人口数量	3~35682
	households	家庭数量	1~6082
	median_income	收入中位数	0.4999~15.001
	ocean_proximity	与大海的距离	Near Bay, <1H Ocean Inland, Near Ocean,Island

4.1.2 特征分布直方图



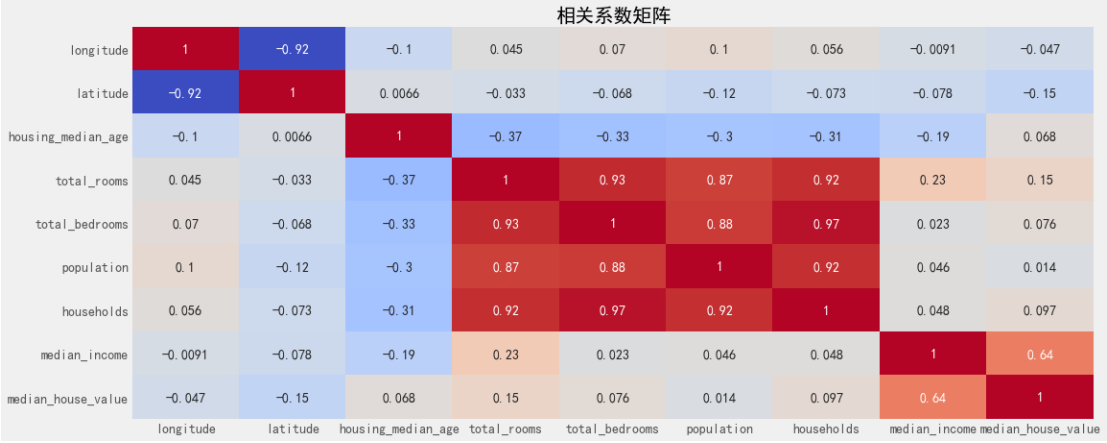
由于我们建模的因变量是 median_house_value，所以需要保证这部分数据不会有很大的“偏移性”，通过上述的直方图中可以看出，在岛屿（Island）上的房屋价值偏高，而且大部分集中在 50 万美金的区域，所以我们需要将大于 50 万美金部分的数据移除。

4.1.3 房价按地理位置分布情况



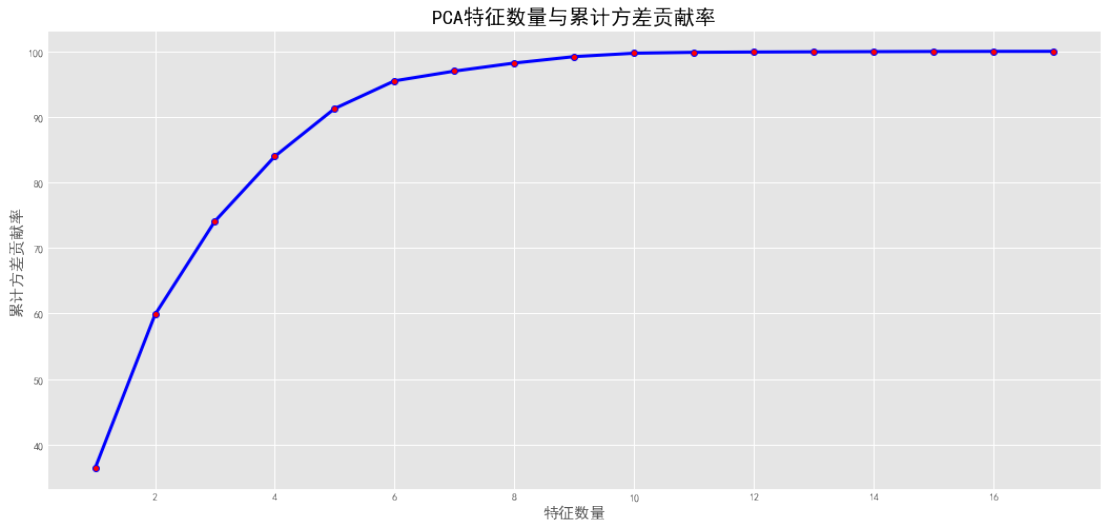
从上图中可以清晰地看出美国加州大部分高端奢侈住宅都分布在经度-124 至-121、纬度 36 至 38 以及经度-120 至-117、纬度 33 至 35 这两大地区，而这两大地区分为别有“人间天堂、度假胜地”之称的旧金山和洛杉矶。

4.1.4 相关系数矩阵



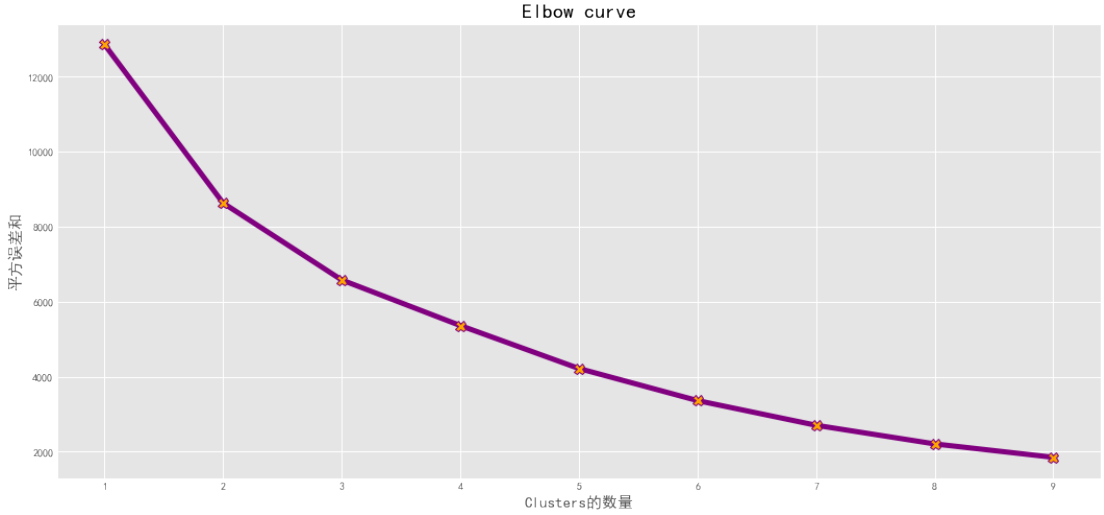
4.2 数据预处理

4.2.1 PCA 降维

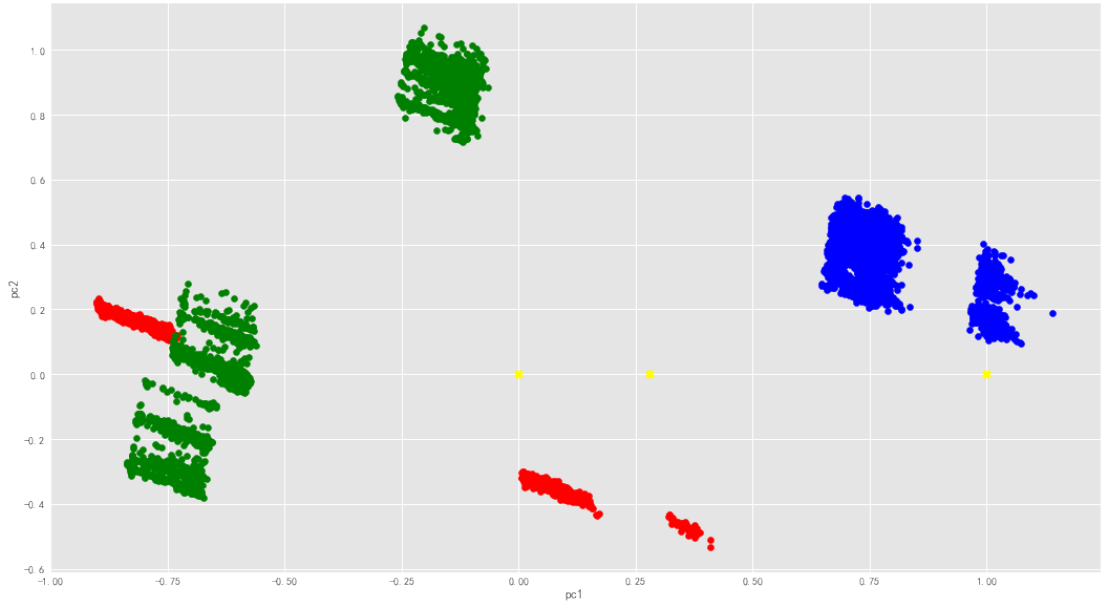


从上图中可以看到前两个主成分已经提供了 60%的方差贡献率。

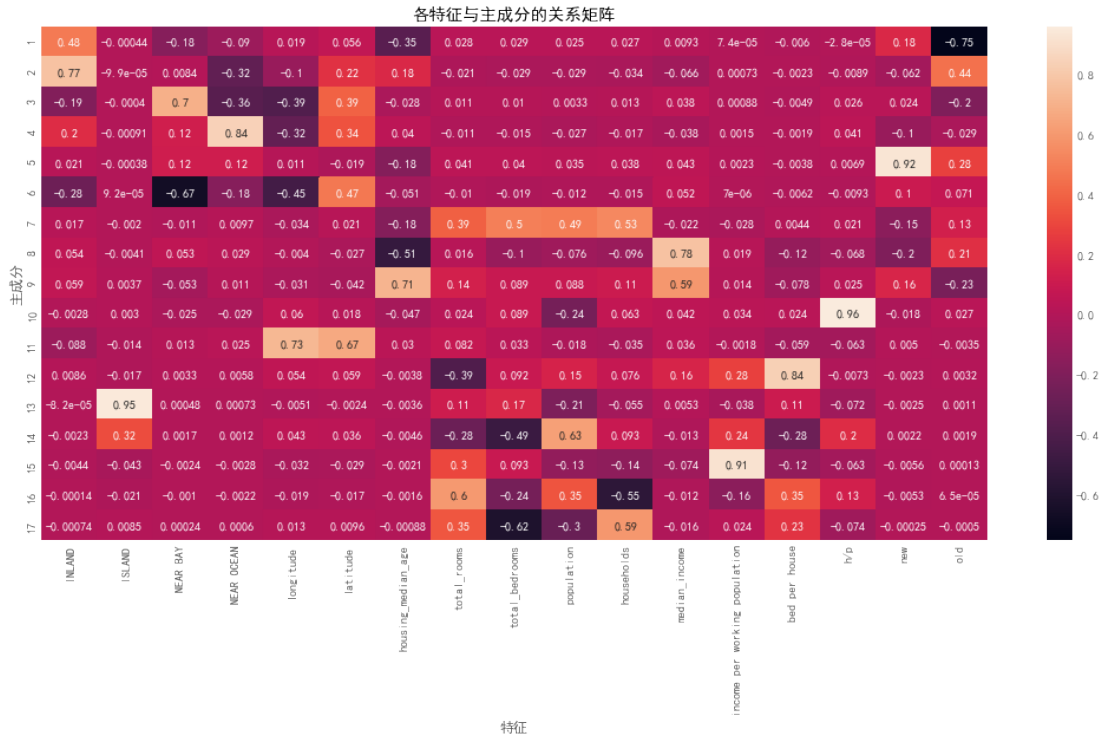
4.2.2 K-means 聚类



可以看到聚类的最佳数量值是 3。

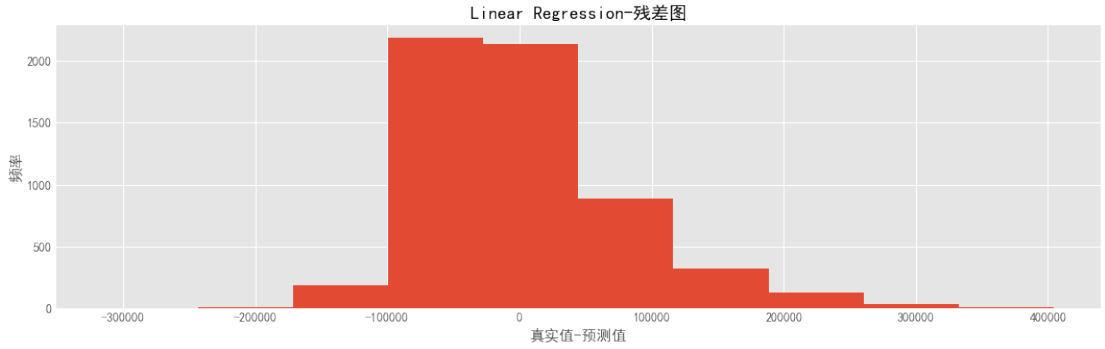


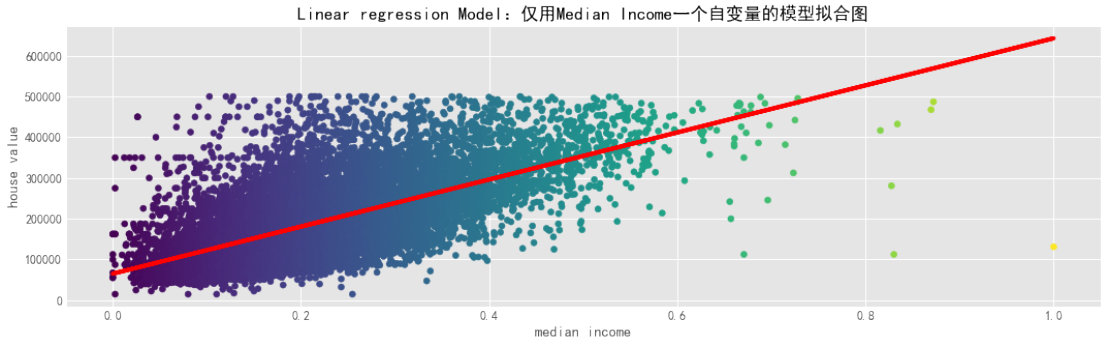
4.2.3 特征与主成分关系图



4.3 建立模型

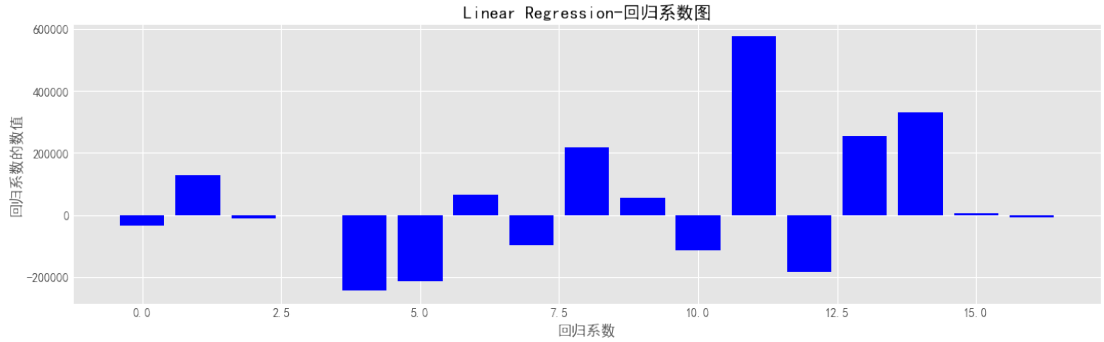
4.3.1 线性回归模型：Linear Regression Model（仅用 Median Income 自变量）

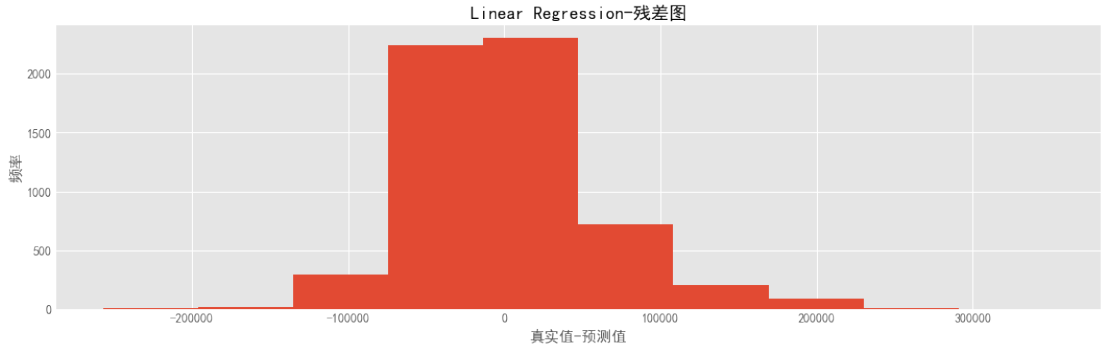




交叉验证准确率：41.3506511798424
交叉验证的方差：1.9235500099556033
Linear Regression-RMSE:76318.44298565741

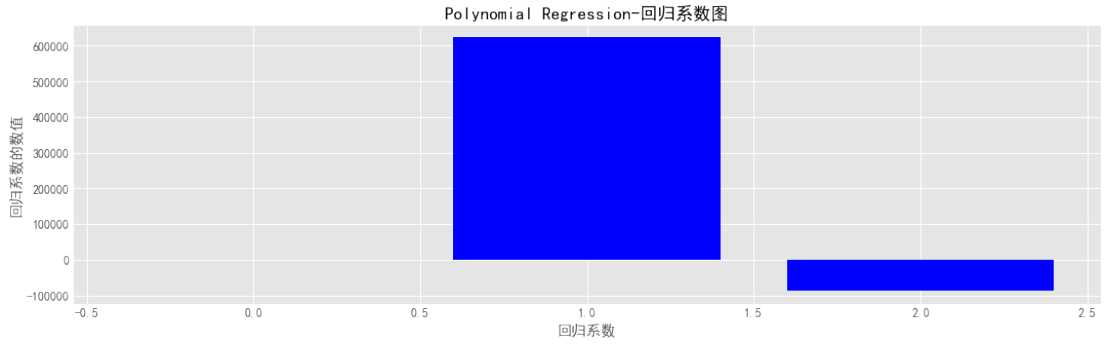
4.3.2 线性回归模型：Linear Regression Model（使用全部变量）

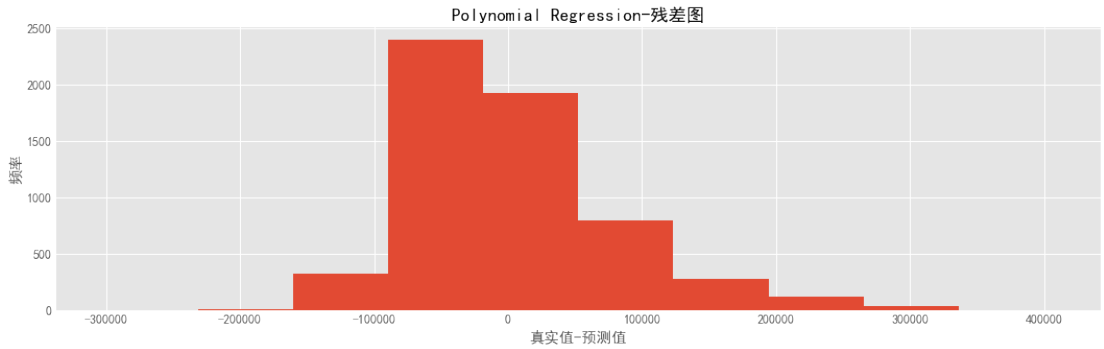




交叉验证准确率：63.275904340783654
交叉验证的方差：1.8152421769594154
Linear Regression-RMSE:58680.28133084016

4.3.3 多项式回归模型：Polynomial regression Model（仅用 Median Income 自变量）



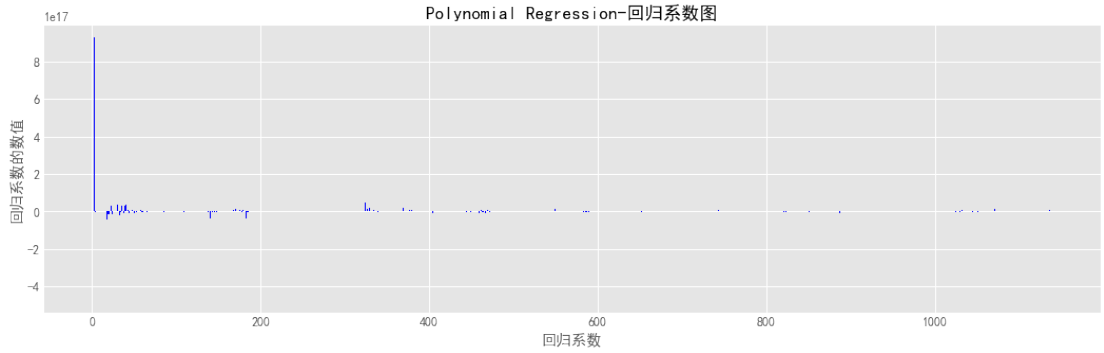


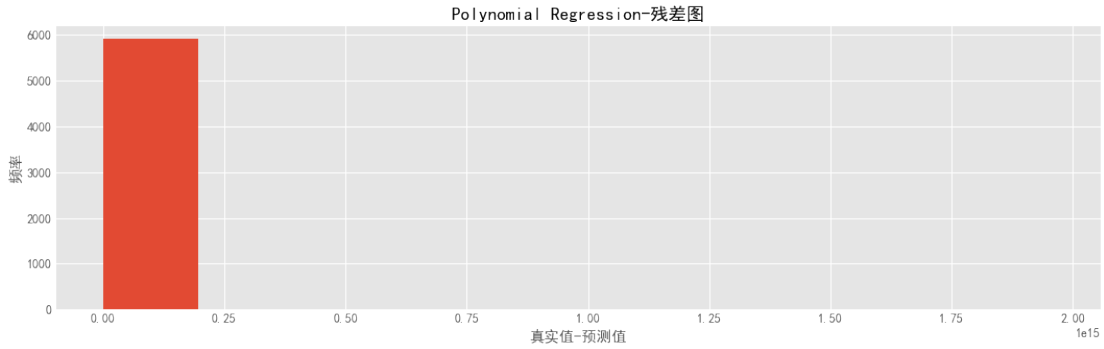
交叉验证准确率：41.35096135301548

交叉验证的方差：1.8965677094607423

Polynomial Regression-RMSE:76253.76750278792 vs 使用全部变量

4.3.4 多项式回归模型：Polynomial regression Model（使用全部变量）



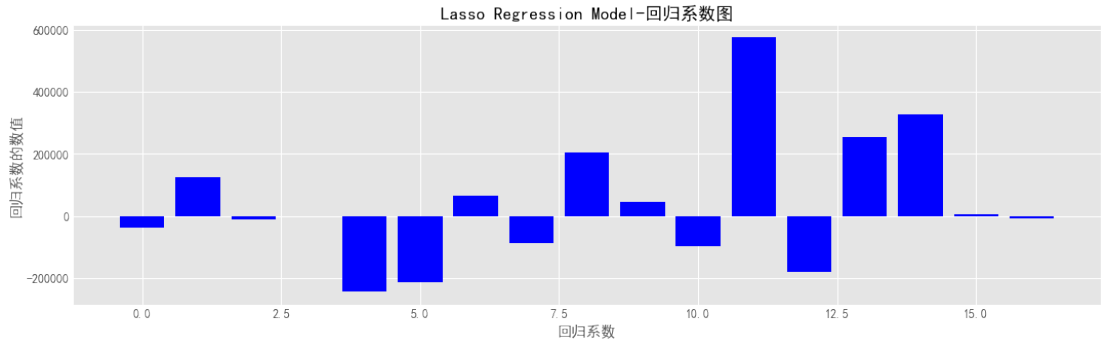


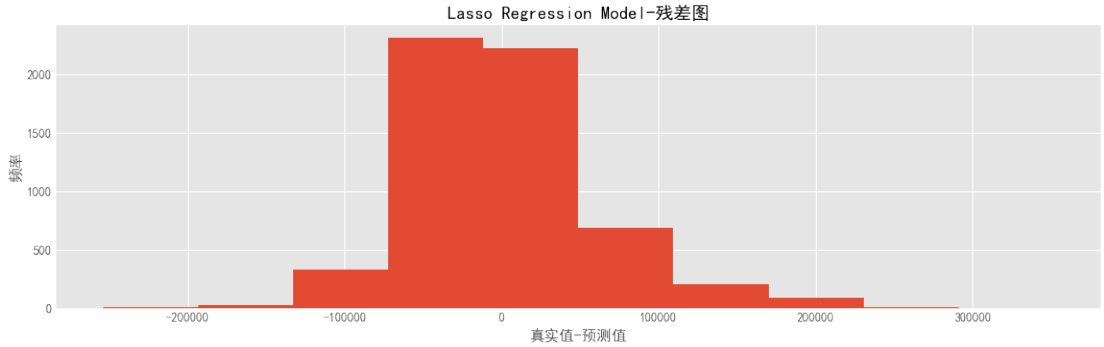
交叉验证准确率：-1.1871174120461946e+21

交叉验证的方差：3.561282144480787e+21

Polynomial Regression-RMSE:25516240115247.438

4.3.5 Lasso 回归模型：Lasso Regression Model

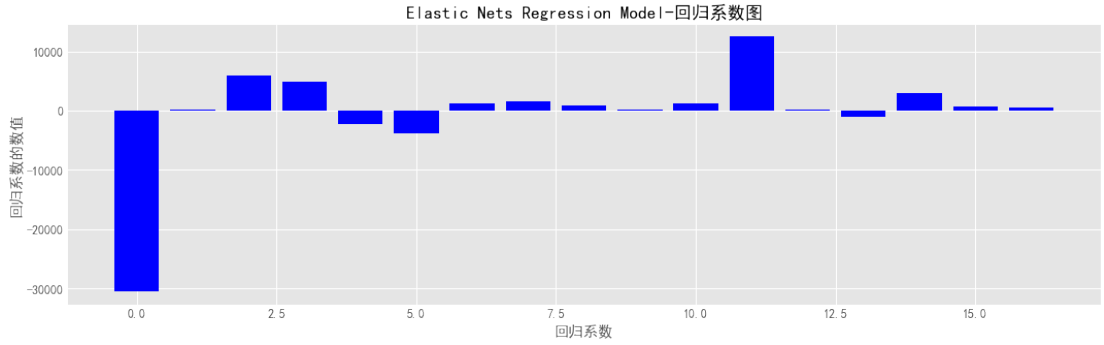


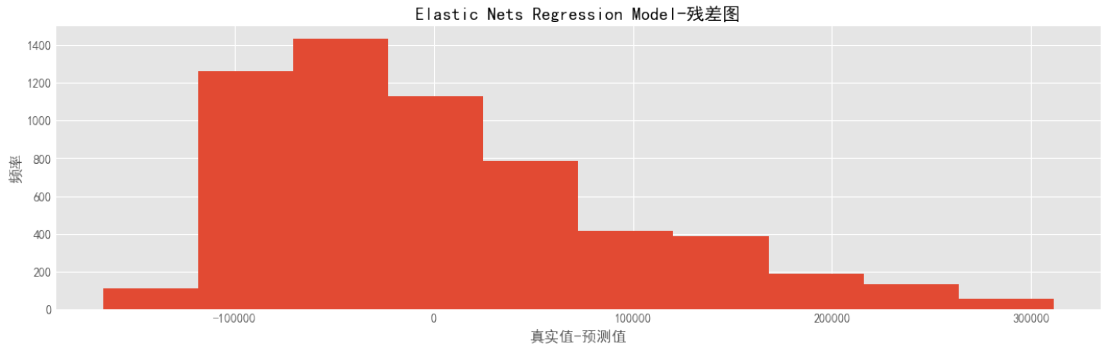


交叉验证准确率：63.28180454731174

交叉验证的方差：1.8048910525748456
Lasso Regression Model-RMSE:58685.20604947836

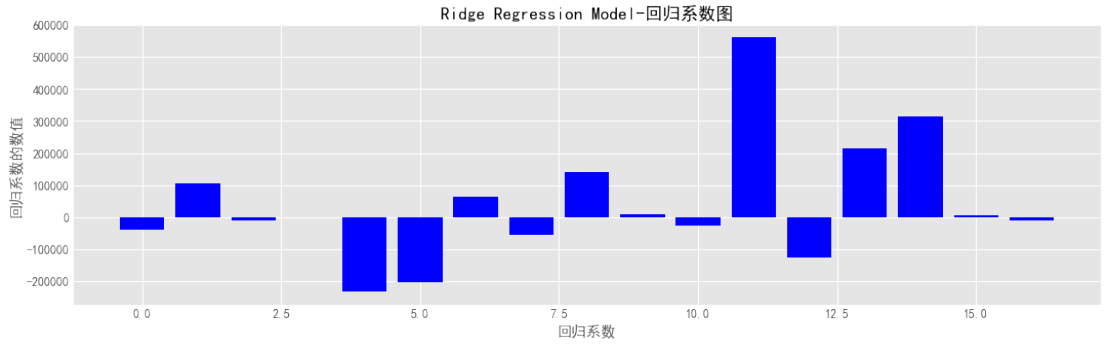
4.3.6 弹性网回归模型：Elastic Nets Regression Model

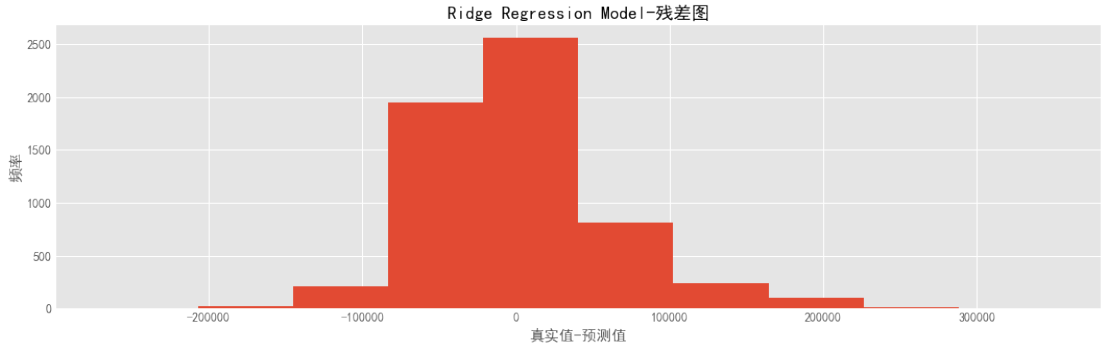




交叉验证准确率：15.148033620292273
交叉验证的方差：0.5460544261385081
Elastic Nets Regression Model-RMSE:91407.11085990786

4.3.7 岭回归模型：Ridge Regression Model

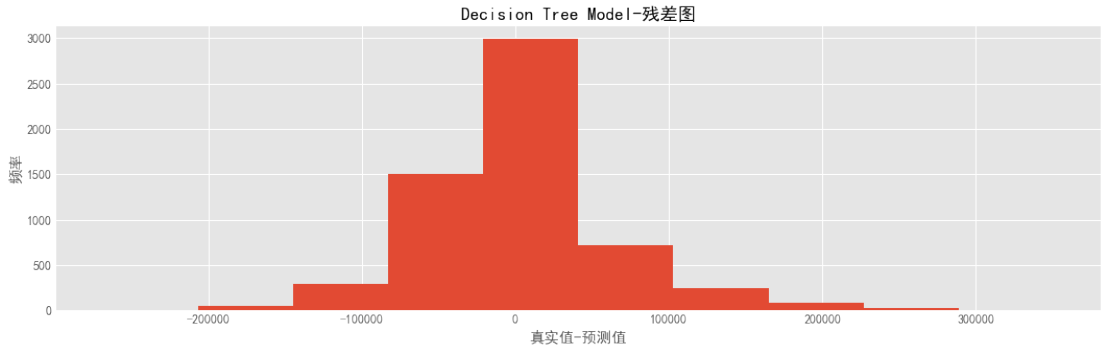


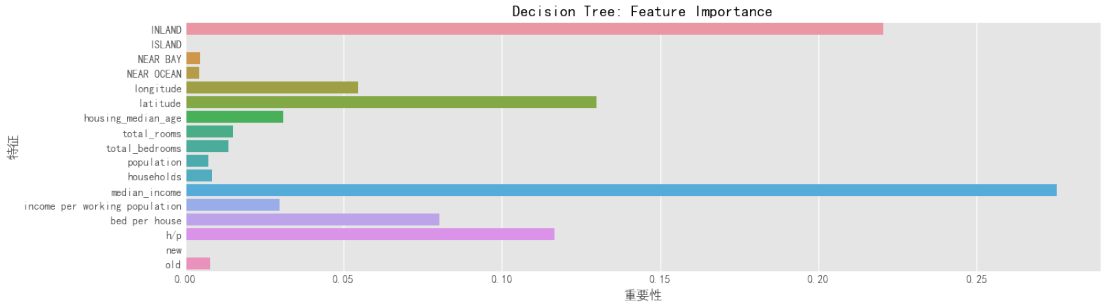


交叉验证准确率：63.36267143046281

交叉验证的方差：1.5945979590003365
Ridge Regression Model-RMSE:58797.98570139773

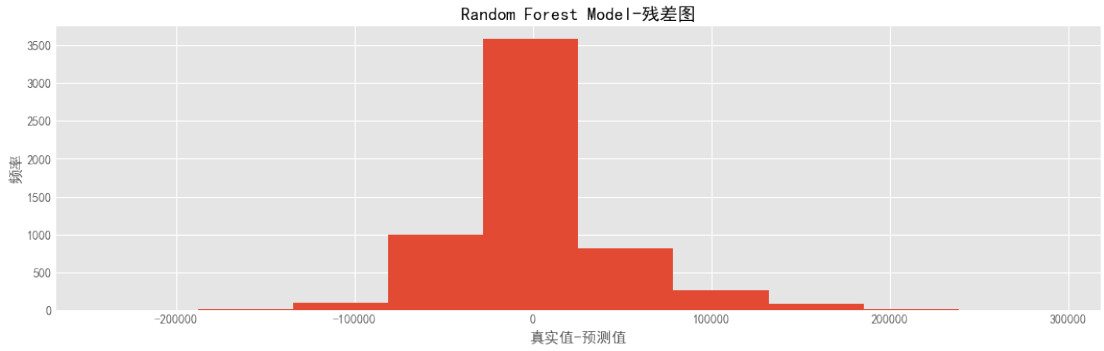
4.3.8 决策树模型：Decision Tree Model

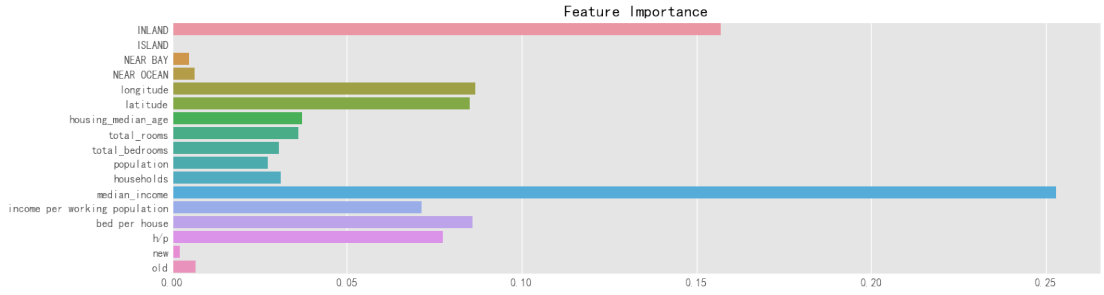




Decision Tree Model-最优 R-Squared：0.653694134651589
Decision Tree Model-RMSE:59631.25161453401

4.3.9 随机森林模型：Random Forest Model





Random Forest Model-最优 R-Squared：0.7850066322924291
Random Forest Model-RMSE:46031.20036790459

4.3.10 模型综合对比

编号	模型	自变量	R 方	RMSE	排名
1	Linear regression Model	使用所有特征	0.6328	58680.28	第 4 名
2	Linear regression Model	仅用 Median Income 自变量	0.4135	76318.44	第 6 名
3	Polynomial regression Model	仅用 Median Income 自变量	0.4135	76253.77	第 7 名
4	Polynomial regression Model	使用所有特征	-0.0120	25516240115247.40	第 9 名
5	Lasso Regression Model	使用所有特征	0.6328	58685.21	第 5 名
6	Elastic Nets Regression Model	使用所有特征	0.1515	91407.11	第 8 名
7	Ridge Regression Model	使用所有特征	0.6336	58797.99	第 3 名
8	Decision Tree Model	使用所有特征	0.6537	59631.25	第 2 名
9	Random Forest Model	使用所有特征	0.7850	46031.20	第 1 名

可以看到，R-Squared 最大的模型是 Random Forest 随机森林模型，达到了 0.7850，说明对该模型的解释程度达到了 78.5%以上，其 RMSE 也是所有模型中最小的一个，这确实证明了集成模型强大的学习能力，紧随其后的是 Decision Tree 决策树模型，R-Squared 为 0.6537，即模型解释度为 65.37%，但其 RMSE 比随机森林高了许多，甚至比线性模型族也稍高一些。在广义线性回归模型族中，Ridge Regression 岭回归模型的表现最好，其 R-Squared 为 0.6336，模型解释度达到了 63.36%，仅比决策树模型低了 2.01%，而其 RMSE 低于决策树模型。令人惊讶的是最简单的线性模型表现与 Lasso 回归模型一样好，其 R-Squared 均为 0.6328，而表现最差的是使用了所有特征作为自变量的 Polynomial regression 模型。综上所述，使用强大的集成模型——随机森林模型对数据进行拟合，可以很好地达到预测效果。

第五章 结论与建议

在本文中，我们首先对 Zillow 经济数据库中的城市房价时间序列数据进行了综合分析，了解了美国 1996 年至 2017 年的整体房价走势。其中我们对销售价格和租赁价格两大类别分别绘制折线图，可以看到 2008 年美国金融危机对房价产生了比较大的影响，直到 2012 年左右才重新恢复上涨趋势，最严重的时候房价跌幅达到了 40%以上，而这其中房价最高的加利福尼亚州受到的影响最大，最深跌幅达到了 67%，导致 2012 年房地产市场恢复元气之后其房价依然没有再次回到全国第一的位置，而哥伦比亚特区却坐上了全美房价的头把交椅。

除了房屋销售价格外，我们还对房屋租赁价格进行了分析，与房产销售不同，房屋租赁市场受到经济周期的影响较小，这是由于美国人民的生活习惯决定的，而且不同户型的租赁价格也完全不同，1 居室房屋以及公寓这两大户型无论在售价还是租金上都是全国排名前二，而后面户型的排名却有所不同，联排房屋的销售价格是倒数第一，而 5 居室以上房屋的租赁价格排名倒数第一，这说明在购房上联排房屋是最经济实惠的选择，而在租房上 5 居室由于面积分摊的原因导致其租金单价最低。

最后我们还对 Zillow 网站上房屋挂牌时间与其售价和租金的关系进行了研究，发现人们最初都会后悔自己的挂牌价格过低，但是随着时间的推移价格会有较大的降幅，在卖房市场上如果一个房屋超过 225 天依然没有被出售，则房东会失去信心而不再维护网站上的价格，而租房市场则不同，由于美国人民对租房的需求量较大，Zillow 的出租挂牌房屋一直处于比较活跃的状态，而且租赁其价格没有出售价格的波动那样剧烈。

在建模部分，由于时间序列数据建模较为复杂，所以我们选用了 Zillow 经济数据库中的加州房价数据进行训练和预测，加州房价数据集包含了多个特征，比较适合建立模型进行拟合。在模型选取部分，我们使用了 7 种机器学习模型，在线性函数族里面选择了比较具有代表性的线性回归模型、多项式回归模型、Lasso 回归模型、弹性网回归模型以及岭回归模型，还选取了决策树模型和集成学习模型中的随机森林模型。经过检验，集成学习模型的代表——随机森林模型表现最优，模型解释度达到了 75.6%，而其他模型解释度均低于 66%，所以如果需要对房价进行预测，则优先选取随机森林模型。

作为宏观经济的重要组成部分，房地产市场总体走势与经济周期运行密不可分。通过分析美国房价历史走势并建立模型预测房价，是十分具有现实意义的。同时我们也需要注意美国因房地产市场的过度炒作而给经济带来的严重后果，这对中国房地产市场的健康良好发展是非常宝贵的经验。