二手帆船价格高昂，其挂牌价通常在十万美元数量级，个别甚至能达到百万级别，影响其定价的因素多种多样。为了更好地了解帆船市场，我们建立了一个**RadomForest-MultipleLinearRegression （RF-MLR）复合模型**来解释二手帆船挂牌价与衡量指标之间的关系、找出可能存在的区域效应、预测香港市场。

First of all，进行了**数据预处理**来降低建模过程的难度以及提高模型的准确度。从船体自身特质以及区域经济特点两方面分析确定帆船生产年份、地区人均消费支出等15个影响因素，通过**Python爬虫技术**扩充数据集。分析影响因素与挂牌价之间的相关性，从而去除相关性不大的净空高度等4个因素，对剩余的11个因素进行异常值分析和缺失值处理。

Secondly考虑到既有分类变量又有数值变量，且特征之间存在较强的相关性，我们选用**随机森林算法**分别对单体船和双体船进行模型训练，之后依据随机交换值后MSE最小的原则将各变量的重要性排序，选出前1/3个变量用于模型二的拟合。对于单体船和双体船选出的影响程度最大的特征，用**Multiple Linear Regression Model** 进行拟合，在此过程中将分类变量制造商品牌转化为**哑变量**参与回归，最终得出帆船挂牌价的数学表达式。并对每个变体进行预测准确度的评估，对于单体船、双体船而言，误差值在10%以内的变体分别占**96%**与**94%**， 模型预测效果较好。

Next，在分析帆船挂牌价是否存在区域效应时，经过统计可知帆船制造年份的数据在区域分布得较为平均，故而可以不考虑影响程度最大的特征生产年份，以所有二手单体帆船挂牌价的均值为标准，比较Caribbean、Europe、USA三大地理区域的帆船挂牌价平均值。得出的结论是**地区帆船平均挂牌价格USA >Total≈Europe > Caribbean。**这与RF-MLR模型中，单体船挂牌价与地区2020年人均消费支出成正相关、双体船挂牌价与地区近五年平均旅游收入成正相关相吻合。对于每个变体进行同样的比较，我们发现不一定与区域效应保持一致。最后分析了区域效应的统计意义与实际意义。

接下来，我们在**Multiple Linear Regression Model中添加实际挂牌价格**进行修改后，建立一个新的**区域效应评估函数RE。**将找到的香港地区二手帆船数据与原数据集进行比较，筛选出重合的部分，再应用区域效应评估函数，发现香港地区的挂牌价整体比另外三个地理区域更高，单体船和双体船区域效应一致。

It is worth mentioning我们还发现了两个有趣的结论。第一，在品牌上，德国、法国具有**本土品牌倾向性**，表格中本国品牌在交易市场其占比率为83.2%，75%；在船体选择上，**加勒比海地区倾向于选择双体船**。第二，**船体排水量与地区具有中等强度相关性**。

我们还对**Multiple Linear Regression Model**进行了灵敏度分析。我们去掉其中一个自变量时，R²将下降5%以上，且下降程度与其重要性程度正相关。我们将模型得到的系数分别变化1%，3%，5%时，随机抽取100个数据求RSS，将其归一化今后做散点图可以看出，系数变化量越大，散点图离散、偏离程度越大。

最后，我们向香港的帆船经纪人提供了一份报告，简单介绍了RF-MLR模型的原理和效果，以及我们据此得出的一些结论。

Keywords: Random Forest, Multiple Linear Regression, pricing model, sailboat