2023 美赛（春季赛）Y 题参考思路

Y 题是一道价格评估类题目，其实也可以看作是数据拟合题或者二手商品定价问题，这几个类型题目的方法都可能被用上。题目要求很清晰，就是要从给定的数据集以及自己收集的相关数据中尝试去拟合二手帆船价格，并对影响价格的一些因素进行分析。从需要使用的方法来看，该题存在的主要难点有两方面：第一：数据，虽然题目自带了一个价格相关的价格数据附件，但是其数据量是不足的，只是提供了一些要求我们必须考虑的因素，可以作为基础模型考虑的属性，题目中多次提到可以使用自己额外收集的数据，因此要做的更好就需要我们自己去额外收集相关的数据集，主要考虑两个大方面的影响要素数据，一个是宏观经济环境可能对定价有影响，比如地区 GDP，旅游行业的收入额等，人们的出游意愿也可能影响定价，另一个是帆船本身的数据，如不同型号，品牌，大小等因素，注意：*题目要求“识别并描述所使用的所有数据来源”*，因此大家也要注意数据的合理性和可信度；第二，需要参赛者有一些数据处理清洗的能力，同时对定价模型或预测模型有一些基础，能够利用一些统计方法分析属性变量对结果的影响程度。

针对问题一，要求我们开发一个数学模型，解释所提供的电子表格中每艘帆船的标价，并对每种帆船型号的价格估算精度进行讨论。对于这一问题，首先要对数据集进行预处理，附件数据还是比较完整的，但自己收集到的数据可能参差不齐，可能考虑的缺失值处理常用方法包括：最近邻算法、平均值填充等；拟合方法上这里提供两种思路：第一、最直接的方法是做回归分析，根据题目提出的各种属性和自己收集到的各种指标，建立多元回归模型，直接对价格进行拟合预测；第二、使用神经网络模型进行价格拟合，把问题当成一个预测类题目来做，输入是多个属性值，输出为价格，相对来说，这个方法可能会拟合地更精确。在得到价格拟合模型之后，再把我们估算的价格和实际价格做比较（要对每种帆船型号分类讨论），统计计算我们的估算精度，这里可以用绝对误差和均方误差等指标来衡量。

针对问题二，根据我们建立的模型来解释地区对市场价格的影响，并讨论是否有任何区域效应在所有帆船型号中都是一致的。问题二的题设要求我们在问题一的指标选取时，还需要尽可能地考虑地区数据，因此，就就需要我们通读题目再去进行问题一指标的选取工作。这里当我们问题一地回归预测模型的话，问题二就可以看作是对回归模型的分析，对单个某一自变量与因变量地关系可以进行讨论。流程为：首先根据不同帆船型号将数据集进行划分，再在不同的子数据集里进行分析，之后在比较各个子数据集的影响是否一致。方法可以采用显著性分析或者方差分析，对不同的帆船型号，通过分析地区相关属性对价格的影响，检验这些影响是否都显著，计算方差大小，即可看出其是否有区域效应以及区域效应是否一致。

针对问题三和问题四，相当于将我们的模型，代入香港这一实际例子进行分析。这里首先需要我们去收集一些香港相关的数据，包括 GDP，人均 GDP，旅游业占比等数据（见参考链接），然后代入问题一的模型中估计价格，再使用问题二的分析流程进行分析并从中总结出有特点的结论即可。

问题四，为香港（SAR）帆船经纪人准备一份一到两页的报告。包括一些精心挑选的图形，以帮助经纪人理解你的结论。根据前面分析的结论，言之有理即可，建议多插图增加美观性。

关键词：二手帆船，定价模型，多元回归分析，显著性检验，预测模型

由于二手帆船领域的专业性，相关数据信息的获取比较难，在这里推荐大家可以从一下方式尝试获取数据集：

1. 搜集更多二手帆船领域的网站，使用八爪鱼等软件或者自己编码爬虫进行数据爬取
2. 可以在咸鱼或淘宝上搜索相关的出售数据集
3. 香港相关：

   资料一线通：https://data.gov.hk/sc/

   统计数字 - 按主题：https://www.censtatd.gov.hk/sc/page_8000.html