# 寻找潜在的蜂蜜市场

10225501464 施长林

DaSE导论结课项目

1. 项目介绍

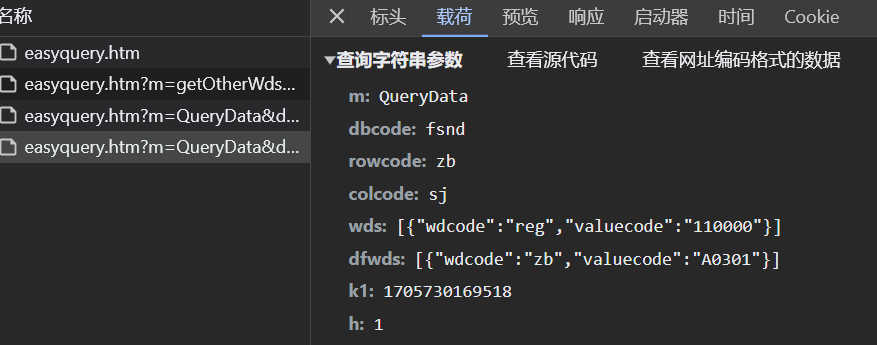
安徽岳西盛产高品质的天然蜂蜜，然而当地农民确苦于没有销售渠道，为了帮助农民销售，在调研了中国蜂蜜市场情况后，欲使用数据科学的方法，探究哪个省份可能存在潜在的未发掘的蜂蜜市场。

1. 数据获取

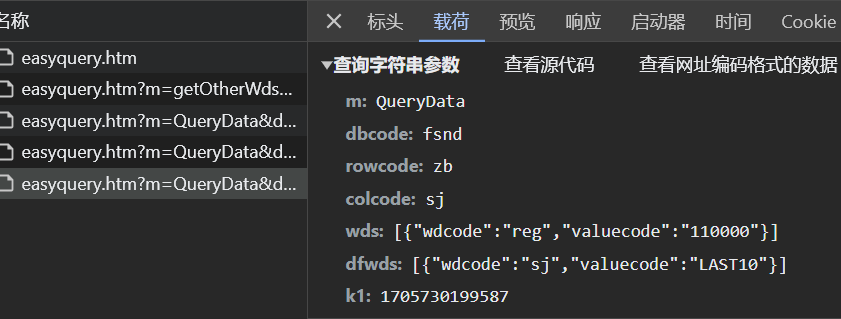
本项目中，主要采用了爬虫的方法获取数据，数据主要来自于国家数据网站的分省数据部分，对其中一些可能相关的指标的各省份近十年数据进行了爬取。

* 1. 请求网页

国家数据网站基于，使用开发者模式研究，例如点击指标-人口-总人口,有如下信息：



再点击最近十年，有如下变化：



即有两组，分别是指标代码和时间范围，而省份的选择则会影响中的，是时间戳。通过库发送这些信息进行爬虫。

* 1. 内容解析

解析上一步中保存至本地的网页文件，



可以看到实际的数据位于中。此处注意到，网站目前更新了的条目，但没有数据，所以爬取出来的都有一个空的项目，为了获得十年中的数据，我们需要请求，然后去除掉的条目。根据此的格式，可以用来提取，是节点的序号，类似于二维数组在内存中的排布，由列数（年份）和行数（指标）确定一个数据所在节点的序号。

* 1. 数据存储

采取的策略是将每一个指标保存为一个文件，它的每一列代表一个年份，从年到年；而每一行则代表一个省份，按省份代码排序。

* 1. 数据处理

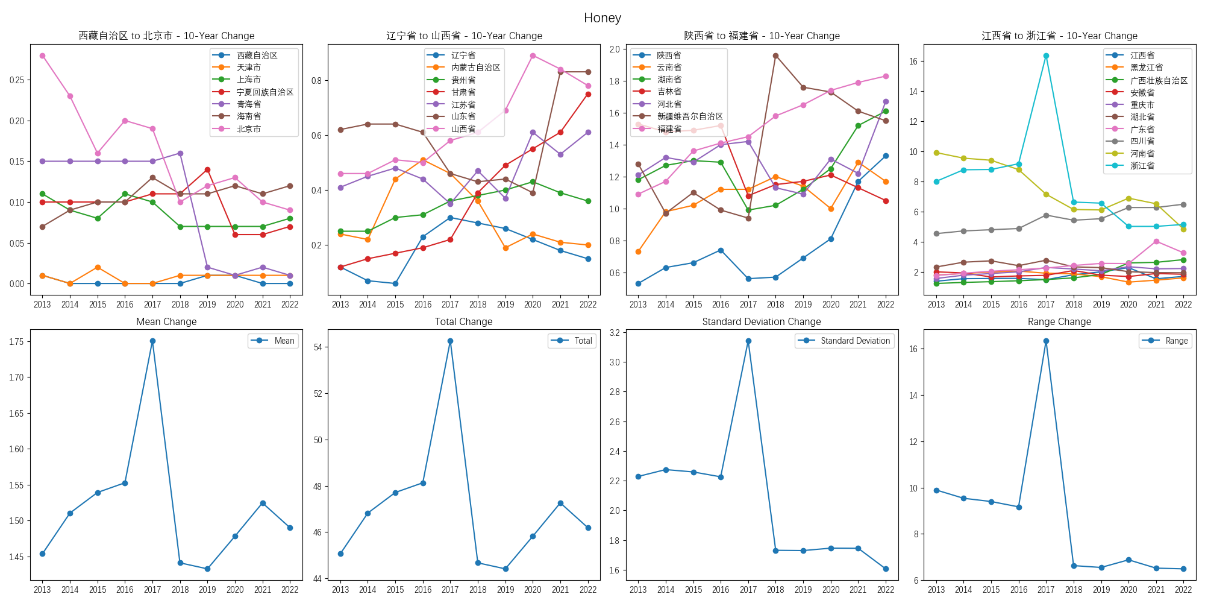
在进一步分析前需要处理爬取得到的数据中的空缺，此处采用线性插值的方法来填充缺失的数据；还需要对数据进行归一化，对某指标的某一年数据，有

* 1. 补充数据

利用直接爬取得到的数据，可以计算衍生指标。例如，由城市人口和总人口数，可以计算出城市人口率作为一项新的指标。除了从国家数据网站上爬取的这些指标的数据，另从行业报告中收集了个省份蜂蜜表观需求量的数据，以同样的格式用文件存储。

1. 基本统计信息与可视化

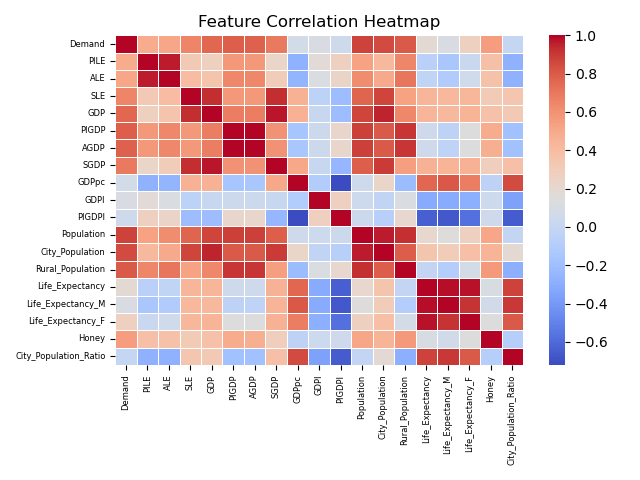
对于采集到的各项指标，各城市的十年变化，我们分析一些基本的统计指标，例如均值，总数，标准差和极差，进行可视化展示，例如蜂蜜产量：



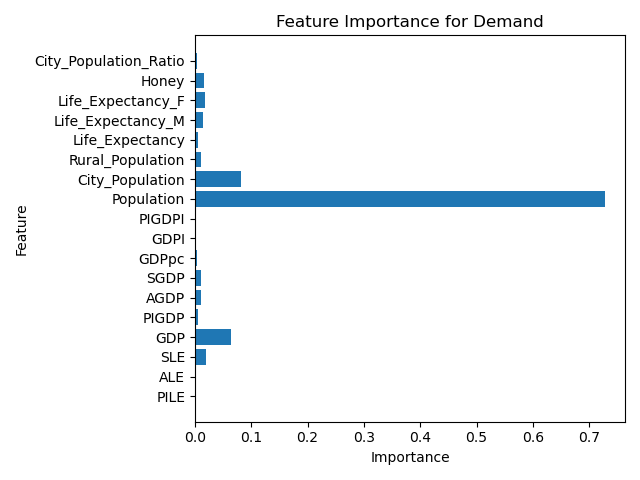
各种指标的图标位于

1. 特征工程
   1. 相关性分析

各指标的相关系数如下：



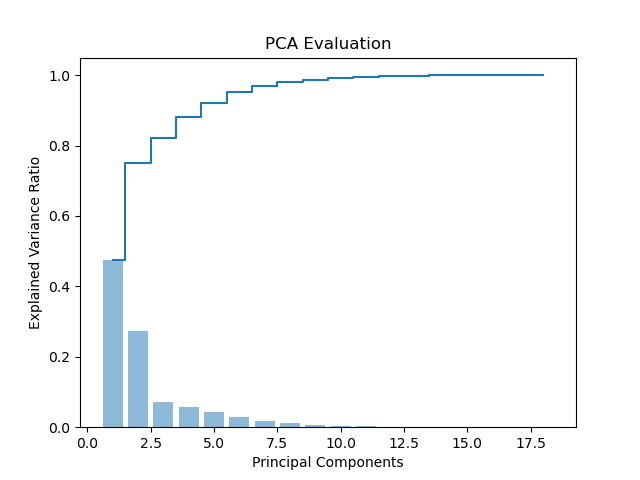
随机森林模型得出的特征重用性如下：



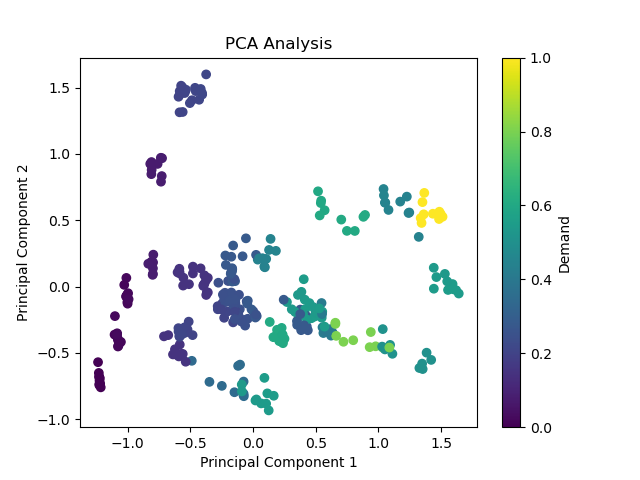
综合两者考虑，可以选择下列特征用于建模

* 1. 降维

除了直接选取可能相关的特征，还可以用对数据进行降维，如下图所示：

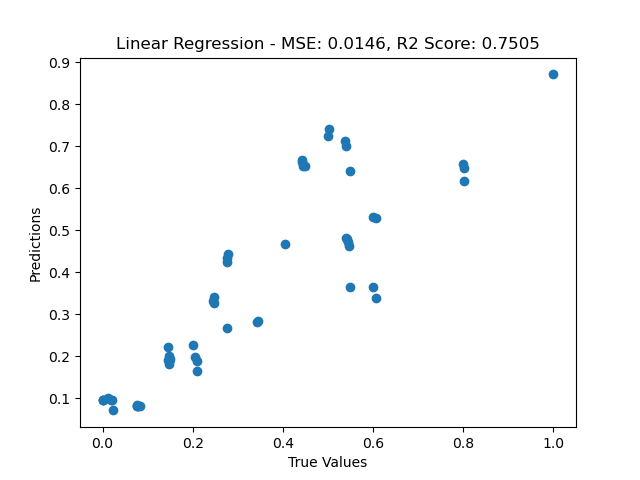


尝试将数据由降至两维：

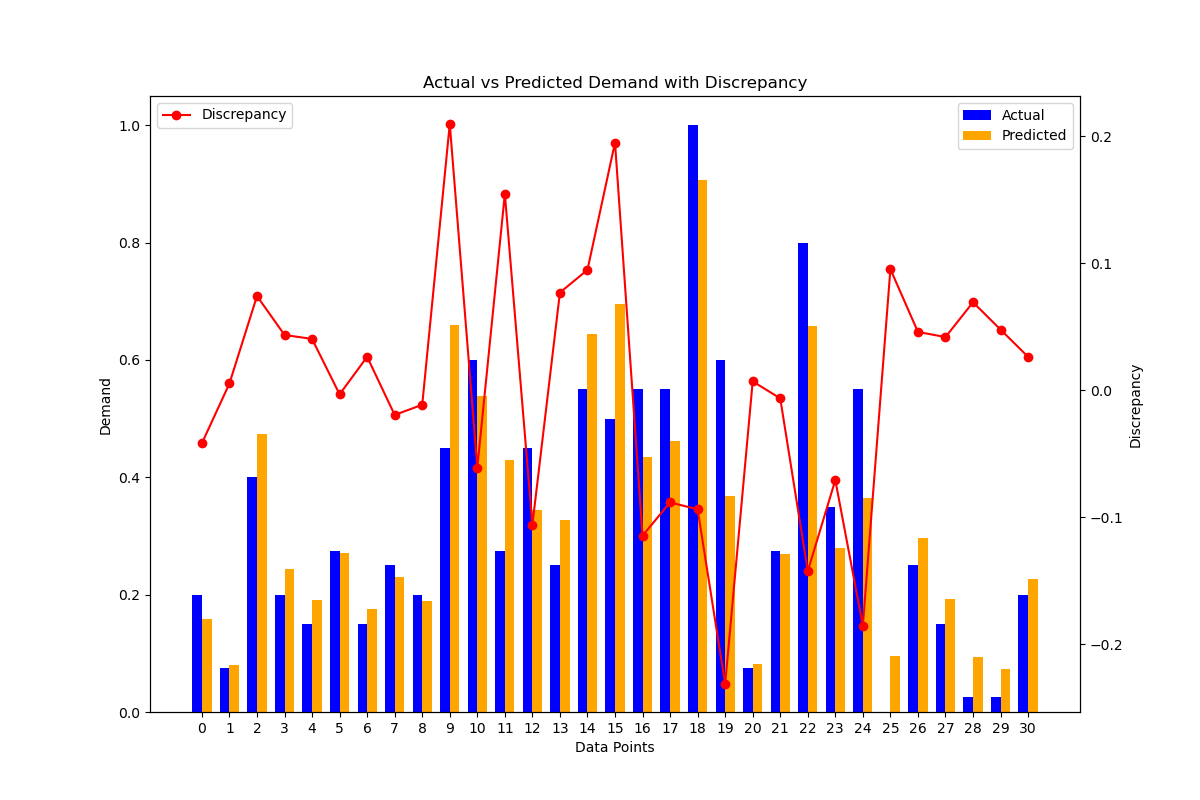


1. 数据建模

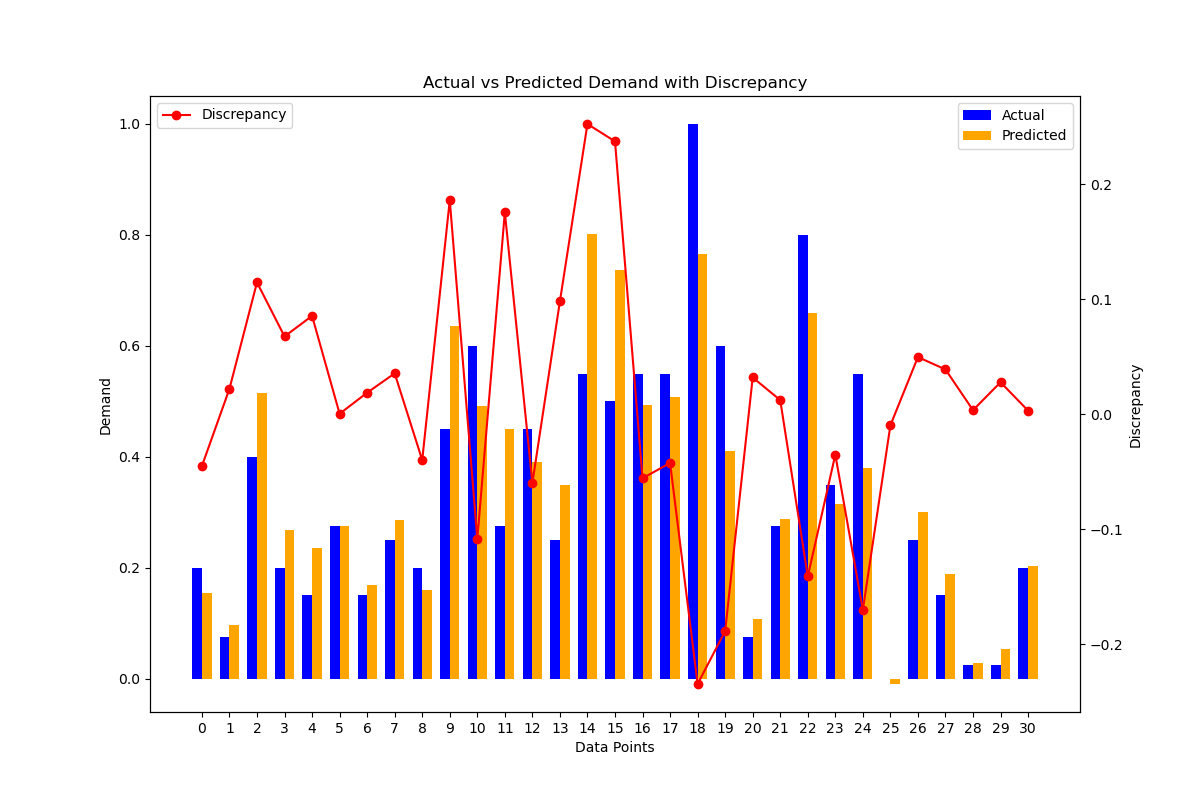
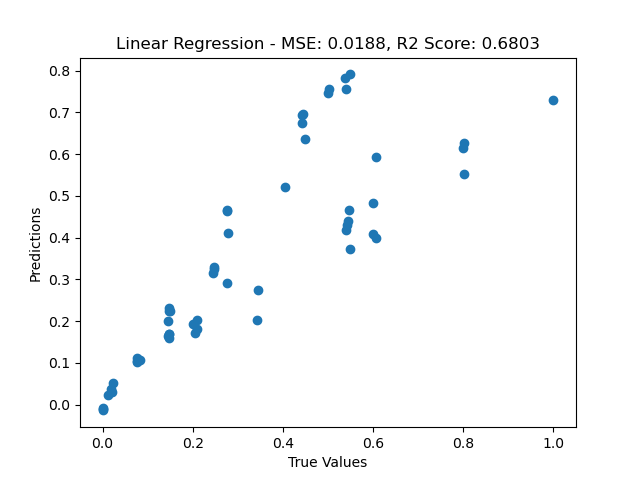
尝试使用线性回归模型进行建模，用的数据训练模型，评估模型效果。例如，用中分析出的个可能比较重要的特征建模：



通过和可知，模型效果尚可，用模型对年的数据进行预测，再与年的实际值比较，以预测值减去实际值为差异，如果较大，说明该省份可能有潜在的蜂蜜市场，如下图所示：



还可以用中获得的降维后的数据进行回归分析，类似可得：



1. 总结
   1. 初步结论

根据以上两种方法处理后的数据的回归结果可知，有几个省份有比较明显的实际值小于预测值，包括江苏省，安徽省，山东省，河南省。这些省份作为经济，人口大省，目前可能仍有一定的蜂蜜市场余量没有被发掘，可以加入蜂蜜销售地点的备选范围中。

* 1. 反思与展望

本项目中还有诸多不足之处，主要包括采集的指标项目不够多，以及建模比较简单，效果可能不够理想等。未来将在这两个方面继续深化，用更多的数据，更精密的模型进行更精准的预测。