1、 项目选题

本项目目标为研究口袋妖怪数据中隐含的联系，我打算使用样例数据提供的口袋妖怪各种数据，对表中连续的口袋妖怪进行数据分析，试图得出连续的口袋妖怪的进化关系，根据最后得出的进化关系，总结口袋妖怪中各属性的进化关系以及总体的进化方向和设计理念。

2、 数据预处理

数据包含了口袋妖怪的名称、攻防速度属性、出现代数、传奇、颜色、所属属性和种族，体态等共 23 个属性，预处理期望达到的目标是能正确分出进化种类，期初采用的方法是机器加人工的方式，根据对口袋妖怪了解的经验，主要根据相应 PM（口袋妖怪的英文缩写）的属性（TYPE 标签）、颜色和种族（EGG GROUP 标签）三个属性进行区分， 一般来说，同一进化链上的 PM 会有相同的种族属性和相对一致的颜色，在程序不确定时，让程序输出相应 PM 的名字，人工确定进化链。

在实际操作的过程中，发现数据收集的过程中存在同一进化链中的 PM 会分别在不同的时代出场，会导致同一进化链的不连续，故而需要进一步人工分类，工作量有些浩大，我找到了一个神奇宝贝 wiki 网站 wiki.52poke.com，通过爬取该网站的口袋妖怪进化数据来进行进化链的划分。最终，将在数据最后一列加上一个新的属性 prev，该属性表示进化到这个 PM 的前一阶段的编号。

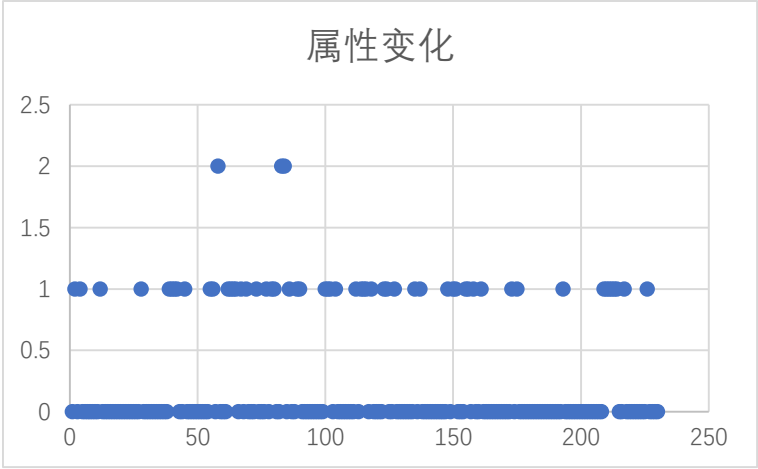
由于研究对象是 PM 的进化过程，故需要删去不会进化的 PM，然后将同一进化家族的 PM 放到一起进行研究，这一过程只需要根据 prev 字段进行两次扫描即可。

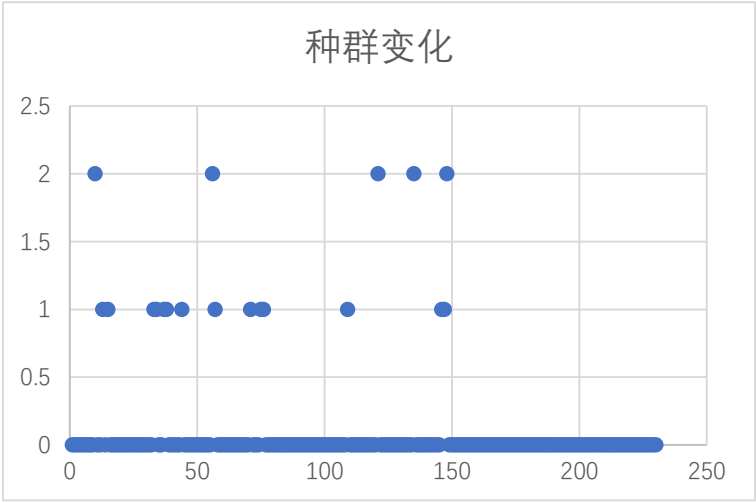
3、 数据分析

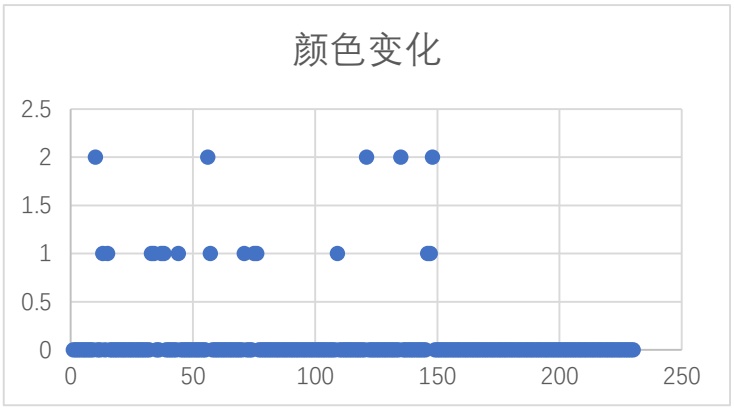
在预处理完毕之后，需要做的是提取同一进化家族的特征，然后对特征进行

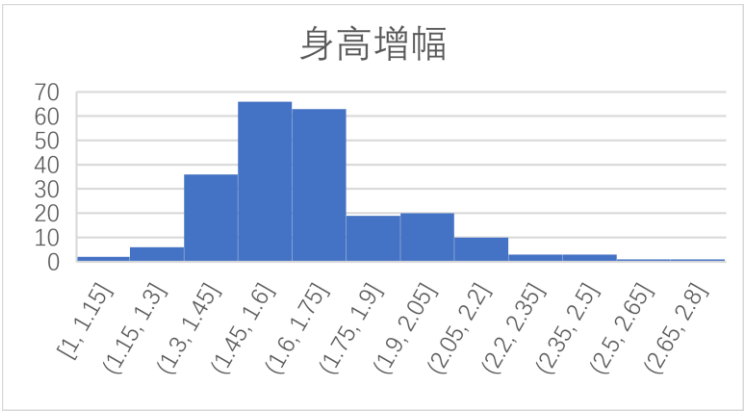
观察和相应的挖掘。这里我定义了几个特征，首先是 PM 的属性，分为增加、减少、发生变化三种情况，我定义不变为 0，变化一种为 1，两种为 2，增减也视为变化；然后是种群的改变，变化规则和属性一致；定义颜色的变化为一个特征；定义种族值为 total 字段即攻防血量特殊攻防 5 个字段的总和，以增加的幅度作为一个特征；分别定义身高体重的增长幅度作为一个特征；最后，将最为突出的一个属性的变化作为一个特征。如果存在进化次数不止一次的，定义特征值为两次特征的和，幅度特征定义为乘积。为方便后续算法，在计算出所有特征之后，以最终形态的 PM 编号作为一个进化家族的标签，并将特征全部归一化到 1 之后进行进一步的运算。

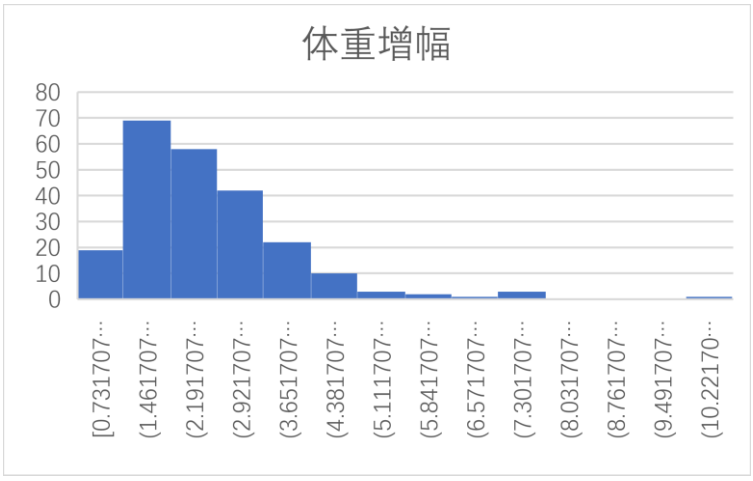
首先使用 excel 观察各个特征的分布情况，画出柱状图如下：

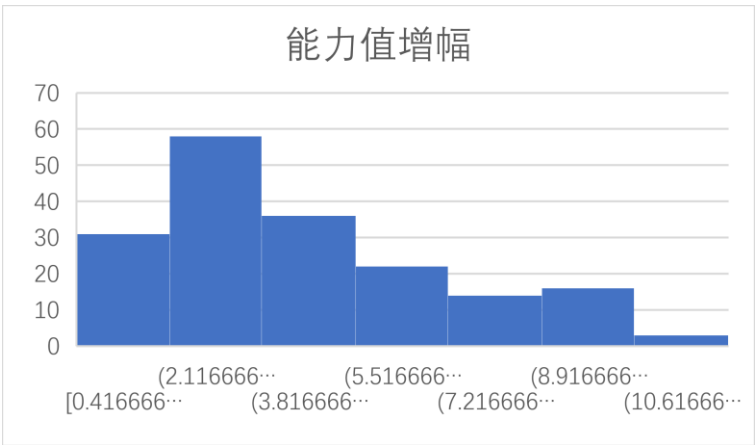





颜色变化
2.5
2
1.5
1
0.5
0
0
50
100
150
200
250

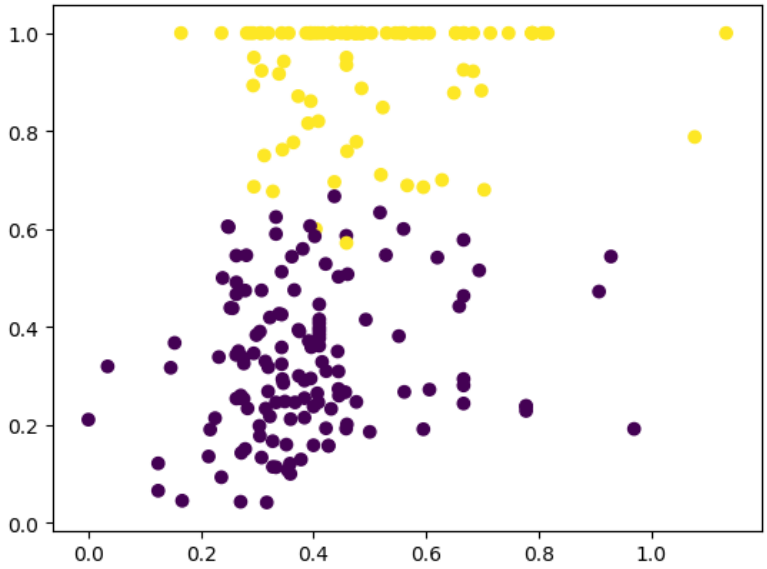

身高增幅
70
60
50
40
30
20
10
0
[1, 1.15]
(1.15, 1.3]
(1.3, 1.45]
(1.45, 1.6]
(1.6, 1.75]
(1.75, 1.9]
(1.9, 2.05]
(2.05, 2.2]
(2.2, 2.35]
(2.35, 2.5]
(2.5, 2.65]
(2.65, 2.8]


体重增幅
80
70
60
50
40
30
20
10
0
[0.731707…
(1.461707…
(2.191707…
(2.921707…
(3.651707…
(4.381707…
(5.111707…
(5.841707…
(6.571707…
(7.301707…
(8.031707…
(8.761707…
(9.491707…
(10.22170…


能力值增幅
70
60
50
40
30
20
10
0
[0.416666…
(2.116666…
(3.816666…
(5.516666…
(7.216666…
(8.916666…
(10.61666…

由于最强能力没有任何进化家族发生变化，故略去不看，而属性、种群和颜色的变化也属于变化较小的特征，下面对增幅特征进行归一化之后进行 K 聚类算法，经过若干次实验，用 Calinski-Harabasz Index 评估分数，发现分 2 类效果最好。此外，观察到增幅数据存在过于巨大的数据，我将大于一定值的直接限定到这个上限进行归一化，避免较小的数被直接忽略。

最终分类结果如图



其中 X 轴为身高，y 轴为能力值增幅，可见能力值增幅起了主要作用。

4、 总结

首先，可以看出，同一进化家族的 PM 在进化过程中，属性、种群、颜色的改变是较少的，总体来说同一家族的这些能力是一致的，而且同一家族的特色属性永远不会改变。这也符合自然界生物进化的理念。而身高体重的增幅大致符合正态分布，说明游戏作者在构建整个 PM 体型的体系的时候，是遵循了统计学规律的。而能力值增幅的分布则不然，在我去除了将近 1/4 的过大数据后，整体数据仍然向大的一边偏，这意味着作者对进化的定义是非常强的一种变强方式，比如其中最大的鬼斯通进化为耿鬼，能力值增幅超过 400 倍，这种设计方式可以让玩家对进化有更为强大的热情，而一部分进化不强的 PM 又可以满足喜欢收集和喜欢幼生期小精灵的玩家的需求。由于 PM 体型在游戏中得不到直观的感触，故其

科学的设定可以增加游戏的严谨性，而进化的设定可以刺激玩家追求进化和变强，同时契合 PM 这种源于自然又有超自然的生命的设定，我认为这是数据设计师的巧妙与强大之处，而通过聚类算法可以看出最终一个进化家族进化强度的评判，主要还是看其进化后能力值的增幅。

视频地址：链接: https://pan.baidu.com/s/1lQbfAXXcn4AKxjNn6Ks44Q 密码: z9iv（为了避免被度娘和谐，我将视频打包并使用加密链接，麻烦助教自己解压一下，谢谢!）