3.数据处理

3.1收集和整理

题目C提供的数据包括了2977所候选学校。而记分牌数据有7804条学校数据。因此我们将候选名单中的学校数据挑选出来。我们通过唯一的机构ID来对数据进行筛选。我们发现有41所学校是没有数据的。为了填补这些缺失数据,我们在IPEDS收集这些学校的数据。

IPEDS提供了详尽的各类报告和统计数据并且区分了年份,统计变量等。由于数据的类别难以区分和整理，我们下载了IPEDS的所有数据。然后我们再进行整理和提取。

将IPEDS所有的数据下载并解压后,共有913个数据记录文件,我们通过机构ID来检验数据的完整性,当数据不包含所有候选学校时认为数据是缺失的。我们编写了一个脚本完成这项工作。结果如下表格所示：

表格 IPEDS数据完整性报告

缺失 错误 完整 合计

903 7 3 913

我们分析了数据完整的3个数据文件。没有发现有效衡量学校状态的数据,最终我们决定将缺失数据的学校排除.

3.2基于主成分分析法的特征提取

我们提取了候选名单中拥有数据的学校的记分牌数据,共28936条。根据记分牌数据变量'CURROPER'，我们将不再运转的学校进行排除。最终实际参与到排名的学校有2936所。记分牌包括一部分描述了学校的地址,网站,ID,位置,机构类型等不可计算的数据。我们仅将能够参与计算数据分离出来,包括了描述了学校招生状况,学生毕业状况以及学生支出,就业后工作收入等98个变量。

我们利用主成分分析，整合相关性大的量。通过主成分分析降低变量维度，简化问题的分析。主成分分析步骤如下：

由主成分分析结果，我们选取14个主成分，贡献率达到99%。