数据清理过程阐述

总述：整个数据清晰和分析数据项目主要流程是：收集数据，评估数据，清洗数据，数据分析和数据可视化，形成分析报告。整个项目流程可以分为5步，而整个清洗过程中收集数据，评估数据，清洗数据是数据清理的主要过程。

以上所有步骤均围绕自己的研究目的进行，当然一些探索性的数据分析过程除外

第一步：收集数据

数据分析开始的第一步往往是数据收集，而收集过程更相当于数据分析的门槛，目前除了大部分公司有自身本身的数据库之外，可以通过SQL直接调取自己需要的数据，其实在更多的过程中个，我们需要从网络中爬取自己需要的数据，而这些数据往往会存在于各种形式中，因此掌握各种收集数据的方式很重要。比如从URL，API，以及JSON中读取数据。而在这个过程中，个人认为从文本中获取自己想要的信息，才是最难的。

第二步：评估数据

数据评估主要分为目测评估和编程评估两种评估方式，而评估内容主要是评估数据的质量和整洁度。而质量问题的评估我们可以按照以下思路来理顺自己的评估过程：完整性，有效性，准确性，一致性，避免在评估过程中遗漏一些问题。而整洁度的标准是：每个变量一列，每条记录一行，每种不同的观测一个表格。最常用的编程评估方式是：.info();.decribe();.value\_counts();sort\_values();head().tail()等。

第三步：清洗数据

数据清洗过程是对评估过程中，数据存在的问题进行清洗，最终得到质量好，整洁度高，可用的数据。而在数据清洗过程中，我们会使用到很多的处理函数问题，以及一些从文本中提取信息的问题。而在清洗结束之后，我们还需要对数据进行再次评估，确定我们得到的是干净整洁的数据。

收集-评估-清洗是一个不断迭代的过程，在整个数据整理的过程中，无论你出于那一阶段，如果出现问题都可以从新来时进行数据整理。