数据整理及可视化过程文档

1. 简介：

1.1 数据集介绍：

通过不同的方式收集推特用户 @dog_rates 的档案，也叫做 WeRateDogs。推特用户 WeRateDogs 以诙谐幽默的方式对人们的宠物狗评级。这些评级通常以 10 作为分母。但是分子呢？分子一般大于 10。 11/10、12/10、13/10 等，为什么呢？因为 “Brent 它们是好狗。”

1.2 任务：

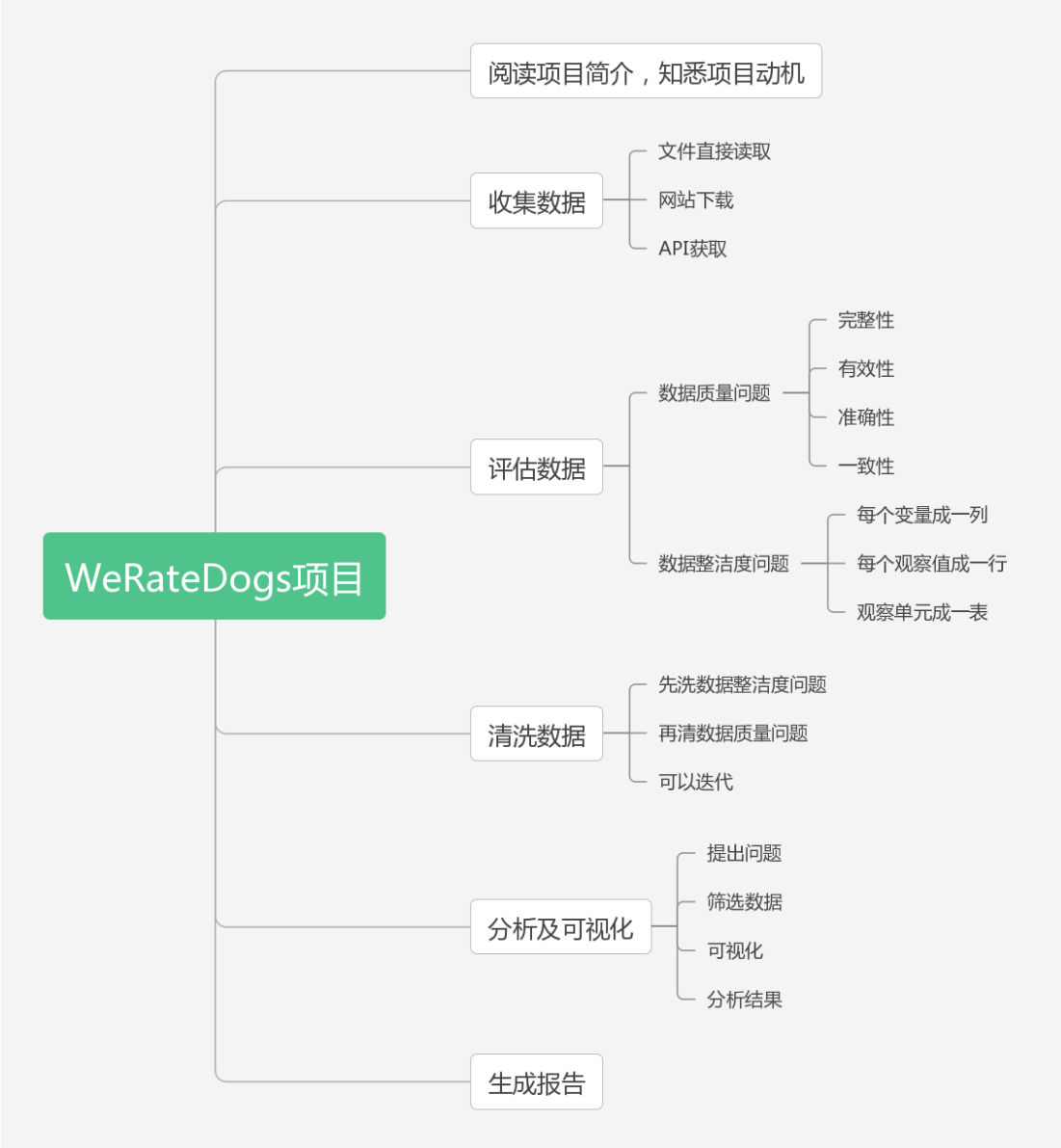
完成数据整理并进行分析及可视化！

1.3 要点：

- 我们只需要含有图片的原始评级（不包括转发）。
- 充分评估和清洗整个数据集需要巨大努力，所以只有一些问题（至少 8 个质量问题和 2 个清洁度问题）的子集需要进行评估和清洗。
- 根据 清洗数据 的规则，清洗包括合并数据的独立内容。
- 如果分子评级超过分母评级，不需要进行清洗。这个 特殊评级系统 是 WeRateDogs 人气度较高的主要原因。

2. 思路：

2.1 思维导图：



2.2 重点步骤解读：

2.2.1 评估数据

- ✧ 数据完整性：利用 pandas.DataFrame.head()查看数据构成；利用 pandas.DataFrame.info()或 pandas.Series.isnull().sum()查看是否有数据缺失;
- ✧ 数据有效性：利用 pandas.Series.value_counts()及 pandas.Series.unique()查看某列数据中，是否有不符合定义模式的数

据，例如 name 不可能为 a 或者是 am 等等；利用 type()查看某列的数据类型是否符合定义模式，例如时间的格式应该为 datetime 而非 object 等；利用 sum(pandas.Series.duplicated())查看是否有数据重复；

- ✧ 数据准确性：利用正则表达式，查看筛选出的狗狗评分、评级、名字是否准确；
- ✧ 数据一致性：在上述过程中，留心是否存在对同一事件的多种表述；
- ✧ 数据整洁度：在上述过程中，留心是否存在变量数与列数不对应，观察值数与行数不对应及观察单元与表格数不对应的问题。
- ✧ **注意 ：此外还应结合此项目的清洗要点进行评估。例如分子评分低于分母评分的数据需要清除等等。**

2.2.2 清洗数据

- ✧ 利用 pandas.DataFrame.loc[index,column]对数据进行修正；
- ✧ 利用 pandas.DataFrame.merge()将多表格整合为一个表格；
- ✧ 利用 pandas.DataFrame.drop()去除某列或某行；
- ✧ 利用 apply 和 split 筛选并提取某列中的数据；