

数据整理报告

本次数据整理和分析项目是关于推特用户 [@dog_rates](#) 的档案数据，推特昵称为 [WeRateDogs](#)。WeRateDogs 是一个推特主，他以诙谐幽默的方式对人们的宠物狗评分。这些评分通常以 10 作为分母。但是分子则一般大于 10：11/10、12/10、13/10 等等。我们需要对采集到的推特数据进行整理，然后分析和可视化并得出一些结论。

1. 收集数据

主要涉及到三个数据集的收集，由于网络原因不便于直接对推特网站进行数据采集，因此三个数据集都是从文件读取数据到 `pandas.DataFrame` 中。

这三个数据集分别是：1. 主体数据集，是关于宠物狗的主要推特档案；2. 图片识别结果数据集，是推特中图片的识别结果；3. 额外数据集，是指从推特网络上采集到的关于 WeRateDogs 博主的额外的推特数据。

宠物狗的主要档案存放在一个 `csv` 文件中，关于推特的图片的识别结果存放在一个 `tsv` 文件中，这两个文件只需要通过 `pandas.read_csv` 就可以进行读取。对于额外数据集，需要通过 `json.loads()` 函数逐行读取数据。

2. 数据评估和整理

主要有观察法和编程法进行数据的评估，观察法通过将所有的数据打印出来，逐个观察发现有问题的数据，编程法通过 `DataFrame.describe()`, `DataFrame.info()`, `DataFrame.columns`, `Series.value_counts()`, `Series.sort_values()` 等函数发现问题。

而数据的问题分为质量问题和整洁度问题，质量问题一般是指数据类型错误、数据缺失、数据错误等问题，整洁度问题要求整个表格是一个对象的观察数据，每一行为一个数据，每一列为一个特征。

接下来对发现的问题进行整理，因三个表都是推特的观察数据，因此将三个表合并为一个表，删除不需要的列，为下一步数据分析做准备。

3. 数据分析

数据分析提出了以下几个方面的问题：

评分与dogtinary

- 是否某一类狗的评分更容易获得高分
- 是否某一类狗的评分更容易获得低分

评分与转发和点赞数量

- 评分高低与转发数量的关系
- 评分高等与点赞数量的关系

转发和点赞数量与dogtinary关系

- 是否某一dogtinary的狗转发数量更高
- 是否某一dogtinary的狗点赞数量更高

转发和点赞数量的关系

- 转发数和点赞数的关系

针对以上问题，从整理后的数据通过作散点图、作密度图、观察描述性统计数据得出了相关问题的结论。

4. 结论

1. 通过观察法和编程法，提出数据集中的质量和整洁度问题
2. 针对质量和整洁度问题，提出解决方案，并实施数据清理
3. 提出数据分析问题，利用清理后的数据完成数据分析问题的回答