**[WeRateDogs](https://en.wikipedia.org/wiki/WeRateDogs" \t "_blank)** 推特数据清洗项目报告

**项目简介**

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，如12/10、13/10。我的目标是清洗WeRateDogs推特数据，并创建有趣且可靠的分析和可视化。我的整个工作可以分为四步：数据搜集、数据评估、数据清理以及分析和可视化，下面将分别进行介绍。

**I、数据搜集**

项目中包含了三个数据集，包含了不同的内容，要通过不同方式进行搜集：

* WeRateDogs的推特档案：twitter\_archive\_enhanced.csv，由优达学城直接提供，点击链接就可以直接手动下载。
* 推特图像的预测数据：image\_predictions.tsv，是对出现在每个推特中狗的品种进行预测的结果，我使用了 [Requests](http://docs.python-requests.org/en/master/) 库和提供的URL编程下载得到。
* 每条推特的额外附加数据：tweet\_json.txt，使用Twitter API下载得到。由于网络原因，我直接使用了优达给的文件，并将JSON数据逐行读如pandas中，经整理的数据至少包含tweet\_id, retweet\_count,favorite\_count三列信息。

Ⅱ、**数据评估**

数据评估分为手动评估和编程评估两个阶段，分别使用的是直观观察数据和编程观察数据的方法，每种方法都有各自优缺点，相互结合使用是最好的。

* 手动评估主要是将数据集展开，使用经验进行观察的方法。
* 编程评估则是使用一些基本的python指令来展示数据的详细信息，并进行判别，常用方法如：.head( ), .sample( ), .info( ), .describe( ), .value\_counts( )等。

对数据集进行评估后，找出的主要问题包括：

Twitter\_df表格中的质量问题：

* 表格中包含retweets的冗余信息
* 'tweet\_id'数据类型显示为整型
* 'timestamp' 时间格式应该是datetime
* 'name'中包含很多None，而不是Nan
* 'name'中包含a，an等明显不是名字的信息
* reply相关的列只包含很少的信息，可以删除
* 'rating\_denominator'的分值不统一，除了10分之外，还有别的高分值，如120
* ‘rating\_numerator’ 给出的分数中，有些数值很低，有些太高，如420，165等，因此不具备统一性，可以创建一个比值来直观给出分数
* ‘doggo’ ，‘floofer’，‘pupper’，‘puppo’等列中的空值用了None，而不是Nan，并且三列都属于‘stages
* ‘source’列中的内容包含有很多冗余信息

整洁度问题：

* twitter\_df, images\_df 和tweet\_df 三个数据集描述的是同一个问题，可以合并
* 一个变量(dog stage) 用4列来描述 (doggo, floofer, pupper, and puppo)
* 'p1','p2','p3'等的描述方法不整洁，对于一幅图片，有一个预测为真即可

Ⅲ、**数据清理**

在数据清理过程中，主要包括三个步骤：定义，编码清理，测试是否清理成功。在进行清理之前，一个很重要的步骤是把所有的数据先进行备份，这样做是非常必要的，如果在清理过程中有损坏数据的步骤，就可以重新拿原始数据进行复原。

在具体清理过程，主要用到了以下方法：

* .merge( )
* .loc( )
* .extract( )
* .drop( )
* .isnan( )
* .astype( )
* .to\_datetime( )
* .islower( )
* .replace( )
* .rename( )
* .set\_option( )

清理的过程主要是针对在评估过程中发现的问题逐个解决。首先是针对整洁性的问题，将三个数据集合并成一个数据集，要注意保留有效的信息列，去除冗余的、无效的信息列。然后针对具体的问题，如数据类型的错误，大小写的错误，内容识别的错误等等一个个修改，并测试验证是否清理成功。最后还要根据需要对清理后的数据进行进一步分析，并用图表的方式显示出来。

**Ⅳ、结论**

此项目借助WeRateDogs的推特数据集，围绕如何得到高质量的数据，将数据清理的流程进行了内容翔实的演练。我从中受益颇多，在以下几个方面能力得到了增强：

* 数据获取能力：WeRateDogs的推特数据集由三部分构成，每一部分数据集获取的方式都不同，基本涵盖了获取数据的基本方式。
* 分析和处理大数据的能力：此次数据集有几千条数据，而且每条数据有几十列信息将其进行描述，数据也比较杂乱，在清理的过程中极大增强了数据处理能力。
* 增强了耐心和自信心：刚拿到数据的时候，其实是非常抵触的，觉得无从下手，但静下心来一步一步解决每一个问题，每一个小问题的解决是让我继续下去的动力，项目完成后极大增强了我的自信心和耐心。