清理与数据分析：推特用户 WeRateDogs 对宠物狗的评分

一、数据分析背景

整理推特用户 WeRateDogs 对宠物狗评分，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。为什么会有这样的评分？因为 "They're good dogs Brent." WeRateDogs 拥有四百多万关注者，曾受到国际媒体的报道。

二、数据分析步骤

a) 数据分析环境

- 在 PC 上使用 Jupyter Notebook 操作。
- 可能需要文本编辑器、文字处理软件、表格处理软件，如 word、excel。
- 需要下列 Python 库：

Pandas、numpy、requests、tweepy、json、os、glob、matplotlib.pyplot

b) 收集数据

i. 观察数据源格式

1. 主数据：twitter-archive-enhanced.csv
   来源：课程提供下载地址
2. 辅助数据：image-predictions.tsv
   来源：互联网下载，URL：
   https://raw.githubusercontent.com/udacity/new-dand-advanced-china/master/%E6%95%B0%E6%8D%AE%E6%B8%85%E6%B4%97/WeRateDogs%E9%A1%B9%E7%9B%AE/image-predictions.tsv
3. 辅助数据：tweet_json.txt
   来源：可从推特 API 获取或课程提供下载地址

c) 评估数据

i. 目测评估

ii. 编程评估

iii. 数据评估小结

1. 阐述对数据集的理解（列的含义及列内数据来源）
2. 提出准备分析的问题
3. 根据问题提出数据清理任务清单

d) 数据清理

i. 质量问题

1. Q1-清理评分数据
2. Q2-将 id 列转换字符串类型
3. Q3-处理 tweet_json.txt 数据集中`display_text_range`列
4. Q4-狗名字中的‘None’变量转变为空值
5. Q5-狗名字变量中大小写不统一且小写姓名不符合取名常识
6. Q6-狗种类名称大小写不一致

7. Q7-狗体型分类中的'None'处理为空值
8. Q8-清理主数据集中的推特转发数据
9. Q9-清理主数据集中不需要的数据
10. Q10-备份清理质量问题后的数据集

ii. 清洁度问题

1. Q1-加载数据
2. Q2-整合三个数据集
3. Q3-狗体型分类合并
4. Q4-更改数据类型
5. Q5-保存清洁度清理后的数据

e) 探索性数据分析

i. 加载数据

ii. Q1-哪种体型的宠物狗最受欢迎？

1. 直接通过清理后数据集得出结论。

iii. Q2-哪种品种的宠物狗获得较高评分？

1. 通过对整合后的宠物狗体型列进行分析，得出结论

iv. Q3-宠物狗推特转发量和喜爱量之间的关系

1. 使用时间戳索引数据集
2. 使用折线图、散点图、频谱图。

三、分析结论

- pupper 体型的宠物狗最受欢迎。
- 得分前三名的宠物狗分别为 pomeranian（博美犬）、clumber（克伦博猎犬）、kuvasz（匈牙利库维斯犬）。
- 通过时间戳折线图，2016 年 4 月以前喜爱量是转推量的两倍，之后喜爱量的增长速度明显高于转推量，推测导致这种现象的原因是喜爱比转推操作上更加方便。
- 通过散点图，转推数量在 0-10000 之间，转推数量与喜爱数量的相关性很强，转推数量超过 20000 后相关性越来越弱。