Wrangle_Report

- 首先下载并打开了image-predictions.tsv的预测数据和twitter-archive-enhanced.csv的推特档案数据；
- 没有使用tweet的API接口爬取数据，而是直接使用了网站提供的tweet_json.txt的转发数，喜欢数数据；
- 在数据评估部分，分别使用目测评估和编程评估发现了**10个质量问题和3个结构问题**，其中质量问题还有更多（如相当多的错误的评分分子和分母），由于时间和精力关系，没有逐个查看；
- 随后对13个问题进行了清洗，主要清洗内容：

    1. timestamp格式为时timestamp
    2. “地位”几列（doggo,floofer,pupper,puppo）标记为None的内容以空值填充
    3. 删除181个转发的记录
    4. 删除expanded_urls重复项中的时间较早的记录
    5. name列标记为None或a的内容以空值填充
    6. index=313，rating_numerator、rating_denominator分别为13，10
    7. index=1068，rating_numerator、rating_denominator分别修改为14,10
    8. index=1662，rating_numerator、rating_denominator分别修改为10，10
    9. 删除第一个识别项目预测结果为非狗狗的记录
    10. 只保留df_archive和df中和df_image匹配的记录。
    11. df_archive.source列拆分为href，rel和'内容'三部分
    12. 合并df_archive,df_image和df表
    13. “地位”（doggo,floofer,pupper,puppo）内容合并到一列中
- 清洗后最终得到一个表并储存为twitter_archive_master.csv；
- 最后对数据进行了简单的分析和可视化，source列虽然进行了结构上的拆分为三列，但实际用处不大，因此没有进一步分析；
- 可以进行的进一步探索：

    1. 对text列的文本分析；
    2. 对转发和原发的数据的对比分析；
    3. 对等级的进一步分析；
    4. 可以基于时间和评分的关系，探索两者是否有一定的相关性。