Wrangle_Report

- 数据来源：
  1. 首先根据网页tsv使用request获取了image-predictions.tsv的预测数据；
  2. 下载twitter-archive-enhanced.csv的推特档案数据，读取json格式的；
  3. 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/点赞数数据；
- 在数据评估部分，分别使用目测评估和编程评估发现了**8+个质量问题和2个结构问题**，其中第6个质量问题包含大量数据准确性的问题；

  质量问题

  1. timestamp格式为str
  2. “地位”几列（doggo,floofer,pupper,puppo）空缺数据以None填充
  3. 存在181个转发的记录
  4. expanded_urls存在数据重复
  5. name列存在数据缺失，填写为None或a的记录
  6. 评分问题
  7. 部分图片第一个识别项目预测不是狗狗照片
  8. df_image共有2075条记录，df_archive和df均有2352记录。

  结构问题

  1. df_archive和包含评论/点赞数的df在两个表中
  2. “地位”（doggo,floofer,pupper,puppo）在不同列中

- 保存原始表备份，随后对10个问题进行了清洗；
- 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；
- 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。