本项目数据来源于三个数据片段，分别是基于 WeRateDogs 的推特档案（twitter-archive-enhanced.csv）、推特图像预测数据（image_predictions.tsv）和有关狗狗推特转发和点赞的数据（tweet_json.txt），均以 dataframe 格式（twitter_archive、image_predictions 和 df_tweet_json）导入 JupyterNotebook 中，针对以下质量问题和整洁度问题进行整理。

整洁度问题做了如下整理：

1. twitter_archive 数据框中四个狗狗地位（doggo/floofer/pupper/puppo）以列名称出现，但它们是数值不是变量名。本项目采用两种处理方法：

   -采用融合函数 melt 将其转变为变量 stage 的数值。将涉及两种狗狗地位的 stage 列数值用’multiple’表示；（注：ipynb 文件中已用三引号标红）

   -重新提取 text 文本中狗狗 stage 信息，明确两种狗狗 stage，格式为中间以逗号隔开的字符串；

2. tweet_archive、image_predictions、df_tweet_json 三个数据框描述的是同一个观察单位，所以采用 merge 函数将三个数据框合并。（合并形式为’inner’，因为我们最终要得到非转发且有图像预测的数据集）

质量问题做了如下整理：

1. 数据并不都是狗狗的原始评级，也包括转发——提取非转发内容的行，删除与转发有关的所有列；
2. 数据并不都是狗狗评级，也包括别的物种——删除图像预测数据中三次预测都不是狗狗的行；
3. Timestamp 和 in_reply_to_status_id、in_reply_to_user_id 数据类型不正确——分别采用 to_datetime 函数和 astype 函数做类型转换；
4. 狗狗评分错误或者无效——利用 replace 或 loc 函数纠正错误的狗狗评分，利用 drop 函数删除无效的狗狗评分；
5. 狗狗名字错误或缺失（数据集中很多狗狗名字无效（例如'a','an'等），源于不全面的提取规则）——利用 df.index.str.extract()和正则表达式，将文本中'named'或'name is'等后面的的狗狗名字提取出来，再通过 replace 或 loc 函数替换为正确名字；
6. 狗狗地位分类有冲突——删除不正确的分类值，保留正确分类值；
7. 更新有关图像预测的列标题，使之描述更清楚一些——使用 rename 函数完成。