收集

通过 requests.get 函数以及 content 方法获取 WeRateDogs 推特档案和智能图片识别的信息的 bytes 对象，将其解码为 utf-8 格式，再存入 io.StringIO 中，最后通过 pd.read_csv 读取创建表格，分别命名为 df 和 img_predictions_df。

通过 Tweepy 库获取的每条 twitter 的信息，逐条读取 id，retweet_count 和 favorite_count 信息，并存入字典格式中，通过 DataFrame 函数创建表格，并命名为 appendix。

评估&清理

质量

通过目测评估可以发现:

在 archive 表格中，有三大类列缺失数据，包括宠物狗的 name 列，与“地位（stage）”变量相关的 doggo，floofer，pupper，puppo 列，以及与“回复”和“转发”相关的 in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp 列。 其中，每条 twitter 的 name 与 stage 变量中的信息可以通过 str.extract 函数以及相关的正则表达式从 text 提供的信息得到，但仍有不少缺失的数据，而 in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp 列的信息和这些缺失的数据一样，在现有条件下无法得到进一步的处理。 source 列中的信息是 HTML 标记语言格式，其中的链接到 twitter 官网的 iPhone 下载界面，而文本信息 包括“Twitter for iPhone”，除此之外没有过多信息，而且每条 twitter 的 source 信息都是一样的，所 以只需保留“Twitter for iPhone”，而不用将 source 信息拆分为两列甚至三列来展现 href 中的"http://twitter.com/download/iphone" 和 rel 中的"nofollow"。 在 appendix 表格中，id 的名称与 archive 和 predictions 表格中的 tweet_id 不同，需要统一。

通过编程评估可以发现: 在 archive 表格中， tweet_id，timestamp，in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retwe eted_status_user_id，retweeted_status_timestamp 列的数据类型错误，其中 timestamp 和 retweeted_status_timestamp 列的数据结构应该用 to_datetime 转化为 datetime 类型，而其它列的数 据结构应该用 str.astpye 转化为 str 格式。 在整洁度的操作中，doggo，floofer，pupper 和 puppo 合并为 stage 列，而 stage 列的数据结构为 str，应该用 str.astype 转化为 category 格式。 rating_numerator 和 rating_denominator 列的数值异常，按照 WeRateDogs 的标准，rating_denominator 应该恒为 10，而 rating_numerator 不会小于 10，但是不排除每条 twitte 都遵循 此要求，有时候会有给多条宠物狗或者非宠物狗评分等特殊状况出现。但为了方便数据处理，运用布尔 索引筛选出 rating_denominator 恒定为 10，rating_numerator 大于 10 且小于 20 的数据。

清洁度

通过目测评估可以发现： 在 archive 表格中，doggo，floofer，pupper 和 puppo 都是 stage 的分类变量，所以应该通过 str.extract 分离出每条 twitter 中的 text 的 stage 变量信息。

rating_numerator，rating_denominator，name 和 stage 列表示每条 twitter 中介绍的宠物狗的信息， 而其余列表示每条 twitter 的信息，包括文本内容，链接，图片，点赞与转发等信息，因此利用花式索引 将 archive 拆分为两个表格，一个包含宠物狗的信息，另一个包含每条 twitter 的信息，分别命名为 tweet_dog 和 tweet_source。 appendix 是 archive 表格附属信息，因为 archive 已经被分为两个新表格，所以应该通过 pd.merge 内 合并 appendix 和 tweet_source 两个表格。