数据收集

1. 从udacity项目细节的链接下载twitter_archive_enhanced.csv。
2. 根据提供的url，利用python下载image-predictions.tsv。
3. 按照要求配置API，并用python抓去Twitter数据，用json解析数据并保存其中有用的数据。

数据评估和清洗

Quality

1. twitter_archive表格中in_reply_to_status_id和in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id和retweeted_status_timestamp有缺失值，予以删去。
2. 把twitter_archive中的数据与image_predic的数据对比， 只保留含有图片的tweet。
3. 删除Twitter_archive中tweet_id为空的数据。
4. twitter_archive表格中tweet_id是整数格式，应该改成字符串格式；timestamp是字符串格式,应该改为时间格式。
5. 对于twitter_archive表格中rating_denominator，狗狗评分的分母都应该是10，对于分母不是10的数据认为是不符合规定, 通过逐条分解text得到正确的评分.一些tweet中含有对多个狗狗的评分， 通过正则查看所有text中含有多组评分的数据, 根据含义更正他们的值.
6. twitter_archive表格中对狗狗打分的分子的记录是有错误的，其中9.75被记录为了75，11.26被记录为了26，11.27被记录为了27，需要进行改正. 7.twitter_archive表格中expanded_urls表示了该条Twitter的原始网页，网页上的信息已经被包含在第三个表格tweet_data当中，所以可以删去这一条信息。
7. 在twitter_archive表格中，rating_numerator是浮点型，所以为了保持一致性，应该把rating_denominator也改成浮点型。
8. 在twitter_archive表中狗狗名字name信息中有一些是小写字母开头，这些是错误信息。对照text利用正则表达式找出来正确的名字，对于无法找出名字的替换为'None'。最后将所有的小狗名字为None的替换为np.Nan。
9. 在第二个表格image_predic中，tweet_id是整数格式，应该改为字符串格式。
10. 第二个表格image_predic中的img_num仅体现了该用户发的照片数量，可以删去这个信息。
11. 在第三个表格tweet_data中，点赞数favorite_count和转发数retweet_count是字符串格式，应该改为整数型。
12. 在第三个表格tweet_data中应该将id改为tweet_id，与其他表格保持一致。

Tideness

1. 在第一个表格twitter_archive中dog stage分成四列给出，可以合并为一列，对于多个种类的返回“multiple"。
2. 三个表格都是以tweet_id作为第一列，所以可以将tweet_id作为主键，将三个表格合并为一个表格。