使用的工具包：

Numpy, Pandas, Json, Matplotlib

项目分析：

这个项目要使用到三个数据集：tweet 数据，image 数据，json 数据，json 数据已经由项目提供，所以此项目省去的 API 的使用。

评估过程：

1. 由于三个数据集各自包含不同的数据信息，所以如果要进行比较全面客观地狗狗分析，需要将三个数据集合并在一起，这样也有利于观察数据集。这一步解决了比较大的清洁度问题。
2. 另外，对狗狗评价描述的选项词汇各占一列，是数据集显得臃肿，可以单独建一列评价描述，然后将这些不同的评价描述抓取出来，对应的填充到相应的行。

清洗过程：

1. 对于格式不统一的，使用 capitalize 函数将字符串统一转换成首字母大写，其他字母小写的格式，方便后期调用分析。
2. 合并之后的数据集很多列都存在大量的缺失数据，而这里面的很多列队分析来说并无太大的用处，这一部分的数据可以选择删除，进一步精简数据集。
3. 很重要的一点，tweet 数据集中有一部分是转发的 Twitter，并非该博主原始发布的数据，需要丢弃，降低噪声，除此之外，根据项目的要求，需要将那些虽然是原始发布，但并不包含图片的数据一并删除。

分析及可视化：

我们能想到的最常见的问题：比如狗狗的名字，什么样的名字用的最多？人们一般喜欢养什么样的狗狗？通过对这些数据的可视化，可以客观地展示现实生活中人们养狗的一些偏好。在对狗狗 status 分析的时候，由于原始数据集中存在大量的缺失值，所以可视化呈现的结果有一定的局限性。