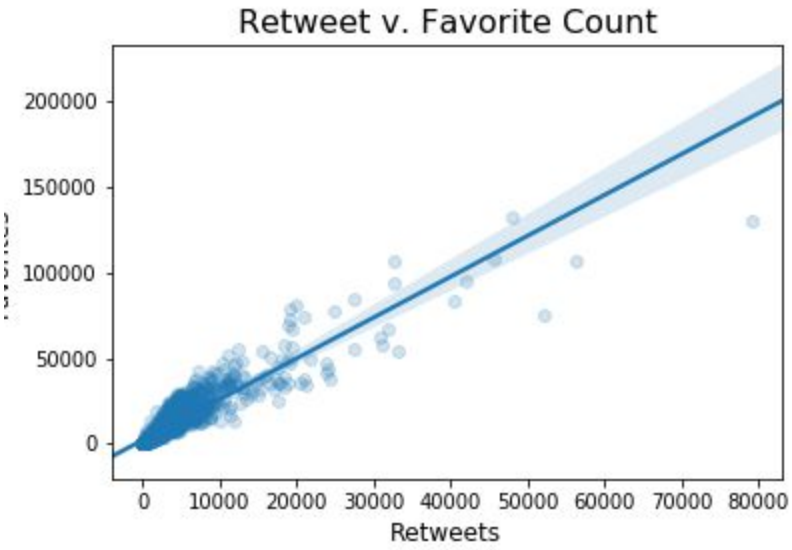
根据此次数据分析总结，主要分为三大步骤：gather，assess，clean。

真实世界的数据总是一种杂乱无章的形式存在，如果需要挖掘其中的价值，并需要对数据进行细致的梳理。这其中对数据的评估以及清洗是难点所在，本次数据清洗任务通过对we rate dogs 的推文进行分析，将推文整理成更加整洁，高质量的数据集。

此次收集数据，我并未使用Twitter API，而是直接使用优达学城所提供的csv文件，此外我还利用了request库从网上爬下了 image prediction。

我根据数据的质量以及整洁度评估了数据。
低质量的数据意味着内容有以下类似的问题：丢失数据，不准确的数据，不一直的数据，以及没有意义的数据。我通过删除不必要的列，转换数据类型，将狗狗名字变得更加一支，删除无用的行（retweet，不是狗狗的记录）来达到提升数据质量的目的。
低整洁度的数据意味这数据有结构上的问题。我将狗狗地位合为一栏，将三个dataframe根据TwitterID合并为一个。

三个结论，以及一个可视化图标
1.通过视图画，我了解到被点赞越多的推文，转发也就越多:



2. 幼年狗狗的图片有更大几率出现在we rate dogs的推文里。
3. 有一点奇怪的是，其中有相当大一部分的推文不是狗狗，而有可能是其他物种。

反思：
此次数据清理耗时很长，经历了反复修改，原因可以总结为以下

1. 对python语言特性不熟悉，导致写代码卡壳，需要搜索大量资料来学习
2. 对panda library不了解,不知道合适的function来使用。
3. 学习课程的时候，有许多细节没有注意，下次应该做更细致的笔记
4. 学习的时候，需要更多时间来学习数据分析过程的思路
5. 可以从kaggle寻找更多的练习，提高数据分析过程以及工具的熟悉度。