DAND-project2-数据清洗

这个项目中首先我先将项目要求做了梳理，并在 wrangle_act文档的开头进行了后续工作计划的 action planning，这对我很有帮助，他们成为了我很好的指导后面数据清洗、数据评估、数据分析工作的纲领。

与之前所学DAND数据清洗课程的整体路线一致，我的项目也是从数据的下载获取开始的，我先通过编程的方式 download了git 上的相关项目文件，然后对于三个不同类型的文件分别进行了转换，使其成为可以操作的 DataFrame的形式，并对三个原始文件进行了复制，以便保持原始数据不受破坏。

然后分别对三个数据集的状态进行评估，这一部分我采用了编程评估和可视化评估两种方式，并且对于每一个数据集的相关问题都进行了记录和汇总。

之后针对上面部分提到的所有问题，基于我们的分析假设前提进行数据清洗。这里对于狗的地位的数据清洗部分耗去了较多时间，主要使用的方法也是for、if、else等控制流，这样使得我的代码运行速度不是那么快，由于时间有限，可能在后续时间充裕的情况下考虑使用矩阵或list的算法以节省时间，也希望reviewer能够提供更好的建议以提升我代码的运行速度。在数据清洗这部分我还利用正则表达式的方法对于狗的姓名进行了调整。观察发现很多狗名都隐藏在named之后，我通过程序进行了提取这部分名字，补充道数据集中，代替原来错误的代词a、an、the。

最后一部分我对清洗好的数据进行了可视化分析，得出了5大点12个小点的结论，可视化部分我采用seaborn的package来绘制，seaborn对于统计绘图有不错的可视化效果。

Update 2018_04_03

1st review存在的问题:
我们来观察下项目动机中的要点会发现，这里有2个问题是强制要求被标记且处理的： 1），只需要原始的数据 2），只需要包含图片的数据（新的提交中已经更改）

关于狗的地位的问题 多种地位的狗，其实最好是将他们保留下来 因为text中可能有些地位是以大写开头的，部分大写的地位没有提取（新的提交中已经更改）

狗名的提取还包括This is|Meet|name is|Say hello to|named 这几种不同类型的信号词（但我发现更正掉named的信号词的词条后，还有错误提取了冠词的现象，而这些现象往往都是以This is|Meet|等为提取信号词而错误提取的。 由于name本身不是我研究的主要对象，故还是保留了之提取named信号词的动作。其余的还是按照元数据提供的name信息进行分析。）