Tweet dog rates 数据的清理与分析

本项目的清理过程主要分为四个部分：数据收集、数据评估、数据清理、分析和可视化。

数据收集部分导入了 3 个文档, 因格式不同, 而分为 pd.read_csv()、pd.read_json()、以及应用 os 库对 URL 文件的数据下载和加载。

数据评估主要分为两大部分 :目测评估、编程评估。目测评估主要浏览数据概况，了解数据基本情况和标签含义，编程评估部分则对整个数据体系进行扫描，为了方便编程评估，在此部分之前，定义了 2 个函数，分别用来检测每个 column 的情况和数据中的 NaN 情况。在此部分的结尾，也对评估过程中发现的数据问题进行了总结，包括 8 个质量问题和 3 个整洁度问题。

数据清理部分也分为两大部分：清洁度和质量。在开始清理之前，对加载的 3 个表格做了备份处理。先进行了 3 个整洁度问题的处理，进行了数据集框架上的调整，在合并了 3 个数据集之后，再进行了 8 个质量清理，并最终将处理好的数据集存档到本地。

分析和可视化部分，先将清理过的数据集进行加载，并通过 scatter matrix 总体浏览数据之间的关系概况，之后根据分析情况，得出不同 column 数据之间的关系图，并尝试从中得出总结或结论。

20190830

崔传敏