此次项目数据分析主要步骤如下：
1. 首先要搜集数据
搜集方式：爬虫、服务 API、问卷调查等。此次项目通过原始下载文件获取数据以及推特 API 获取基础的推特档案（因为国内网站访问推特 API 不方便，所以直接使用项目提供的文件）
2. 数据评估
首先加载搜集数据，理解每个字段所表达的含义。
数据评估主要从以下几个点考虑：
a. 数据完整性：主要评估数据缺失程度，分为数据信息缺失和数据字段缺失。
b. 数据准确性：主要是评估数据是否包含异常值，可能由于人员手工录入错误等，比如身份证号码是否满足校验规则等。
c. 数据有效性：主要是评估数据的取值的类型、值域、格式是否正确，比如数据是有关驾驶人员的信息，那么年龄就必须满足（小型汽车 18-70 岁）范围内。
d. 数据实效性：主要是指信息仅在一定时间范围内对决策有所作用，如果时间太久，这个信息就没有价值。
e. 数据一致性：主要是当多表关联时，主表与子表关联字段是否一致。

主要考察了数据的质量和清洁度
质量问题：

twitter_archive 表格

- rating_numerator 存在异常值，max=1776
- source 中还保留了 a 标签
- 'in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id', 'retweeted_status_user_id','retweeted_status_timestamp' 缺失值过多，且对此次分析无影响，可以删除相关列
- 一部分 name 中值为 *a an the*
- expanded_urls 存在缺失

image_predict 表格

- jpg_url 有 66 个重复值

tweet_data 表格

- tweet_id 应与前两个表的类型一致，为 int64

清洁度：

- *twitter_archive* 表中 *timestamp* 字段需要处理成常用日期格式 yyyy-MM-dd hh:mi:ss

- *twitter_archive* 表中 doggo,floofer,pupper,puppo 都属于'stage'，应该合并成一列

3. 数据分析及可视化

常用可视化图形：

1.直方图(Histogram) 主要展示数据的分布

2.柱状图(Bar) 主要比较数据的大小

3.饼图(Pie) 主要探索变量的分布占比

4.热图(Heatmap) 主要探索多个变量两两之间的相关性

5.散点图(Scatte) 主要用于探索两个变量间的相关性