Fazer o data wrangling desse conjunto de dados do twitter ‘we_rate_dogs’ foi um grande desafio pois tivemos que trabalhar com muitos tipos de arquivos diferentes: arquivos em csv; tsv baixado diretamente da url; utilizar uma API do twitter para baixar os dados de retweets e favoritos de cada post; além de ter que trabalhar fazendo conversão de arquivos json para txt e txt para json.

Usar a API do twitter para baixar os dados faltantes foi a parte que mais demorou em tempo de execução pelas limitações de tempo de acesso e uso da própria API, além do tempo para criar e configurar a aplicação no site do twitter.

Na hora de escrever os json objects para arquivo txt foi a parte que mais tive dificuldade pela falta de conhecimento do retorno da função ‘get_status’. Em resumo, na parte de gathering a parte mais trabalhosa foi a de utilizar a API e gravar esses dados em um arquivo txt.

A fase de assessing foi bem tranquila e deu para achar rapidamente alguns erros de qualidade e de organização para serem anotados e resolvidos na parte de cleaning. A parte de cleaning foi a mais demorada e fez jus à teoria de que 80% do tempo da análise de dados é reservada para a parte de cleaning. Feita a cópia dos três principais arquivos a serem usados, eu dividi a fase de limpeza por dataframes, onde cada dataframe é um tópico e dentro dele estão os seus respectivos problemas. Cada problema está dividido em 3 subtópicos: define, code, e test. Isso facilita o entendimento e a organização. O Python facilita e muito essa etapa de cleaning pois já possui muitas funções prontas, e só precisei fazer uma função em toda etapa de cleaning, que foi a função para transformar as 4 colunas de tipos de cachorros em uma só. As funções melt, e get_dummies poderiam ser usadas mas preferi fazer uma função própria e usar o apply em cima do dataframe. A parte final do notebook contém um tópico chamado visualização onde eu faço o join(merge) de todos os 3 arquivos principais já limpos, o que resulta em um dataframe único para melhor agregar os dados. Depois faço alguns insights e faço o plot de alguns gráficos que serão usados para fazer o relatório ‘act_report.pdf’.