# Relatório

## 1 - Coleta:

- Tempo coleta dos tweets através da Tweepy foi demorado. Tive que dividir a coleta em várias etapas.
- O modo solicitado para salvar os arquivos dos tweets, em um txt, não era prático. Optei por manter o formato json ao salvar, pois o Pandas já tem um método pronto para ler esse tipo de arquivo.

## 2 - Avaliação:

- Utilizei os métodos de avaliação clássicos - info, head, value_counts - além de checar se os valores da coluna estavem dentro do domínio correto - caso, por exemplo, da coluna *retweed_status_timestamp que possuia letras nas datas.*
- *Dados faltantes: Há dados faltantes nas colunas retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp - eu optei por retirar as linhas com dados faltantes nessas colunas, já se tratavam de tweets não originais - e nas colunas dos tipos de cachorro - nesse caso, há a string "None" em vez NaN.*
- *Não entendi as 4 ultimas colunas, as que possui os tipo de cachorro como nome. A maioiria das linhas possui dados faltantes nas 4 colunas. Fiquei em dúvida se precisava fazer alguma classificação própria.*

## *3 - Limpeza:*

- *Foi complicado manter o esquema do notebook apresentado nas aulas, com as marcações em markdown corretas.*
- *Tentei checar se o nome da linha batia com que estava no texto do tweet utilizando regex, porém sempre retornava erro. Utilizei algumas ferramentas online para testar o meu regex. Nessa feramentas o regex funcionava, ao contrário de quando eu utilizava ele no notebook, retornava sempre a string inteira.*
- *A tabela csv obtida através de download não possui documentação. Isso dificultou entender a tabela.*
- *Optei por juntar todas as tabelas em uma só, já que todas elas pertencem a mesma unidade observacional.*
- *Não utilizei o método melt pra tirar as colunas com variáveis como nome pois dificultaria na hora de retirar os dados faltantes - linhas com "None" nas 4 colunas, que são a maioria, iriam desaparecer.*