Wrangle Report

Iniciei o projeto juntando os dados dos arquivos ‘twitter_archive_enhanced.csv’ e ‘image_predictions.tsv’. Esta parte do trabalho foi bastante rápida. Em seguida foi necessário acessar a API do twitter, uma etapa bastante demorada, em que várias coletas foram feitas até que eu entendesse o funcionamento da API e gravasse os arquivos corretamente em um json.Na etapa de avaliação programática pude observar vários erros de qualidade que não havia visto com apenas a avaliação visual, este tipo de avaliação foi bastante útil.Encontrei nos dataframes criados erros como: dados que deveria ser string e estavam como int ou float, problemas de arrumação nas colunas que definiam o estágio do cachorro, tweets que não tinham foto, retweets(que não deveriam ser considerados), notas de avaliação incorretas, nomes de cachorros incorretos, entre outros erros de qualidade ou arrumação. Comecei a limpeza pelos erros de arrumação, pois é mais fácil limpar dados arrumados. Juntar as informações de estágios de cachorros de quatro colunas para uma foi uma das partes mais trabalhosas, onde verifiquei também que alguns cachorros possuíam mais de uma definição de estágio.Na parte de limpeza, arrumar os nomes de cachorros também não foi nada fácil, precisei utilizar regex.Verifiquei que manter apenas os nomes que começavam com letra maiúscula era uma estratégia efetiva. Nas etapas seguintes o trabalho foi um pouco mais tranquilo, arrumar as notas de avaliação não apresentou muitas dificuldades depois que analisei os denominadores superiores a dez com o método ‘value_counts’.O conjunto ‘twitter_archive_enhanced.csv’ foi o que apresentou a maior parte dos problemas, por ser também o maior conjunto de dados. No conjunto criado a partir da API muitas colunas acabaram sendo descartadas, pois continham dados repeticos com relação ao conjunto no formato ‘csv’.No conjunto ‘image_predictions.tsv’ o número de erros foi menor, a escolha de qual algoritmo era o mais confiável ficou bastante evidente com o método ‘describe’. Optei por utilizar apenas as classificações de raça do primeiro algoritmo, pois os outros tinham um nível de confiança muito

baixo. Desta forma obtive um número menor de dados, mas com uma qualidade maior do que teria considerando previsões dos outros dois algoritmos.