Udacity data wrangling

Foi realizado um projeto de data wrangle para finalizar a fase de data wrangle do curso da Udacity fundamentos em Data Science II.

O inicio do projeto se deu com a coleta de dados. Manualmente foi baixado o arquivo *twitter-archive-enhaced.csv*. Outro arquivo, chamado de *imagem-predictions.tsv* foi baixado de forma programático. Ambos os arquivos foram carregados em um *dataframe* específico para cada um.

De forma programática, utilizando a API *Tweepy*, foi acessado cada *tweet* e armazenado as informações que vinham em formato *json* em um arquivo texto chamado *tweet_json.txt*. Alguns dados tiveram retorno de erro, informando da inexistência de alguns *tweets*, dessa forma, não foram possíveis de armazenar seus dados. Após coletar todos os dados dos *tweets*, foram carregados em um *dataframe*.

Com os dados carregados, inicia-se a avaliação de forma a encontrar problemas de qualidade ou de estrutura. Foram encontrados onze problemas de qualidade e quatro de estrutura.

Antes de iniciar a limpeza dos dados é feito uma cópia dos dataframes, de forma preservar os dados originais e motivos de comparação.

Dos problemas encontrados, três deles não foi possível ter ação para solucioná-los, referente a ausência de dados que tem acesso. Outros problemas de qualidade estavam relacionados a utilização de tipos de variável inadequados para representar os dados. Para solucioná-los, apenas foi modificado os tipos de variável.

Teve problema com variáveis com valores que não representam efetivamente o conteúdo que deveriam conter, ação tomada para esse problema foi substituir esses valores pelo valor nulo. Outro caso em quem avaliação dada ao cão na postagem tinha denominador igual a zero, ocorrendo erro ao dividir valores com zero. A correção para esse problema foi excluir da tabela *twitter_archive* a linha com o valor zero.

Os demais erros de qualidade foram corrigidos, agora é foco nos problemas de estrutura.

O primeiro problema a ser percebido foi que existia duas colunas, uma de numerador e outra de denominador na tabela *twitter_archive*. O correto seria ter apenas uma coluna com o resultado da divisão. A correção desse problema de estrutura foi criado uma nova coluna chamada de *rating* que tem como valor o resultado da dita divisão, e deletado as duas colunas originais.

Outro problema de estrutura encontrado foi que os dados que coletamos da API *Tweepy* não deveriam formar uma tabela a parte, mas sim deveriam fazer parte da tabela *twitter_archive*. Tivemos então que mesclar essas duas tabelas, de modo que toda a informação fique reunida.

Ainda quanto à estrutura, tínhamos que solucionar o problema de que a ordem das predições do algoritmo (1ª, 2ª e 3ª predição) não estava representada em uma coluna própria, estava espalhada por diferentes colunas. Foi preciso criar uma nova coluna só para a variável, além de outros procedimentos adicionais.

Resolvidos os problemas de qualidade e estrutura, tem-se um conjunto de dados pronto para ser explorado e analisado. Salvo o resultado do data wranling em dois arquivos CSV *twitter_archive_clean* e *image_predictions_clean*. Com isso, encerra-se a fase de wrangling.