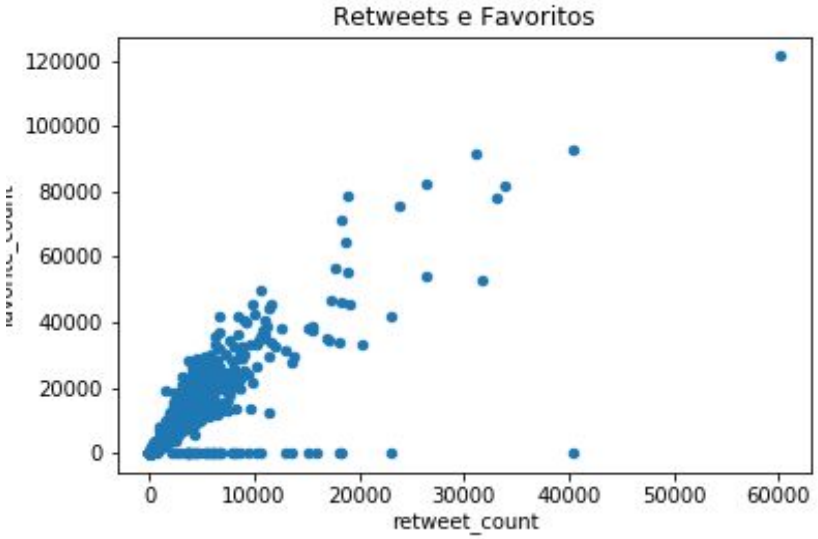
Projeto: Preparar e analisar dados
Data Wrangling - WeRateDogs

Este projeto foi um grande desafio para mim do começo ao fim. Foi muito bom e importante aprender sobre o processo de Data Wrangling e ter uma noção básica sobre como consultar dados da API do Twitter por meio de uma conta autenticada de desenvolvedor e da biblioteca Tweepy.

Depois de ter selecionado e corrigido os problemas de qualidade e arrumação dos três conjuntos de dados, foi necessário fazer as análises e visualizações com algumas intuições acerca dos arquivos. Comecei com uma análise sobre a correlação entre as contagens entre os tweets favoritos e a contagem de retweets. Utilizei a função scatter plot da biblioteca Pandas e obtive a seguinte visualização:



A análise estatística mostra uma grande distribuição positiva à direita em ambas as categorias indicadas pelo grande desvio padrão. Os resultados também indicam que as pessoas vão preferir um tweet com mais frequência, então elas retweetam o tweet original, como indicado pela maior contagem de favoritos.

A partir da visualização, podemos ver uma forte correlação entre os dados favoritos e retweets com um coeficiente de correlação de Pearson, r, igual a 0,92. A forte correlação faz sentido lógico porque quanto mais popular um tweet, maior a possibilidade da contagem de retweet e favoritos crescer.

Outras análises pertinentes a esse conjuntos de dados foram feitas, como as questões: Quais são os 10 nomes de cachorros mais comuns?, e Qual o cachorro mais ‘retweetado’? Usando a função Count nos nomes, foi possível descobrir os 10 nomes mais comuns. São eles: Oliver, Winston, Tucker, Bailey, Penny, Cooper, Lucy, Bella e Toby.

Para descobrir o cachorro campeão de retweets foi necessário ter a descrição estatística do conjunto de dados com a contagem dos favoritos, retweets e classificadores. Com a descrição, podemos ver que o cachorro mais “retweetado” foi o Stephen, da raça Chihuahua.