PROJETO DATA WRANGLING: UMA BREVE ANÁLISE DOS DADOS DA CONTA DO TWITTER WE RATE DOGS

WeRateDogs é uma conta do *Twitter* que classifica os cachorros de seus usuários de uma forma muito bem humorada. Para este projeto foi cedido dados de aproximadamente 3000 *tweets* dos seus usuários.

Para a realização desta análise foi unificado em um só *DataFrame* o resultado da coleta, avaliação e limpeza de três outros *DataFrames*, que contiam informações como ids dos *tweets,* o texto, estágio do cachorro, possível raça, a probabilidade dessa possível raça, quantidade de *likes*, quantidade de *retweets*, entre outros. Para a execução da análise foi utilizado o *Notebook Jupyter*.

Através da análise dos dados, podemos observar com a ajuda da função *.describe()* que nas colunas:

- **confiabilidade_1** temos caso onde a confiabilidade máxima foi de 100% e a mínima de aproximadamente 4%;
- **confiabilidade_2** temos caso onde a confiabilidade máxima foi de aproximadamente 49% e a mínima de aproximadamente 1e-5%;
- **confiabilidade_3** temos caso onde a confiabilidade máxima foi de aproximadamente 27% e a mínima de aproximadamente 2e-8%.
- **qtd_likes** temos caso onde a quantidade máxima de *likes* foi de 164915.00000 e o mínimo 51.00000
- **qtd_retweets** temos caso onde a quantidade máxima de *retweets* para aquele *tweet* foi de 84101.000000, enquanto há caso em que não ocorreu nenhum *retweet.*

Com essa análise acima, podemos tirar algumas conclusões interessantes em relação ao que tais dados podem nós falar, como por exemplo, se uma empresa está procurando um cachorro para vender seu produto, o interessante seria contratar um cachorro que tenho bastantes likes e retweets, pois seus “fãs” poderia identificar aquele cachorro que eles gostam com o produto vendido pela a empresa e posteriormente poderia adquirir tal produto.

Em relação a que cachorro possui a maior quantidade de likes e qual possui a maior quantidade de retweets, foi notado que o mesmo cachorro possui as quantidades máximas de likes e retweets, o tweet_id 744234799360020481.

No caso da confiabilidade, foi encontrado um caso onde ela era de 100%, ao olhar o caso, observamos que se tratava de um caso onde o este verdade apontou ‘falso’ para esse caso, ou seja, a análise não acertou em 100% a raça do cachorro, ou até possivelmente o animal na foto nem seja um cachorro.

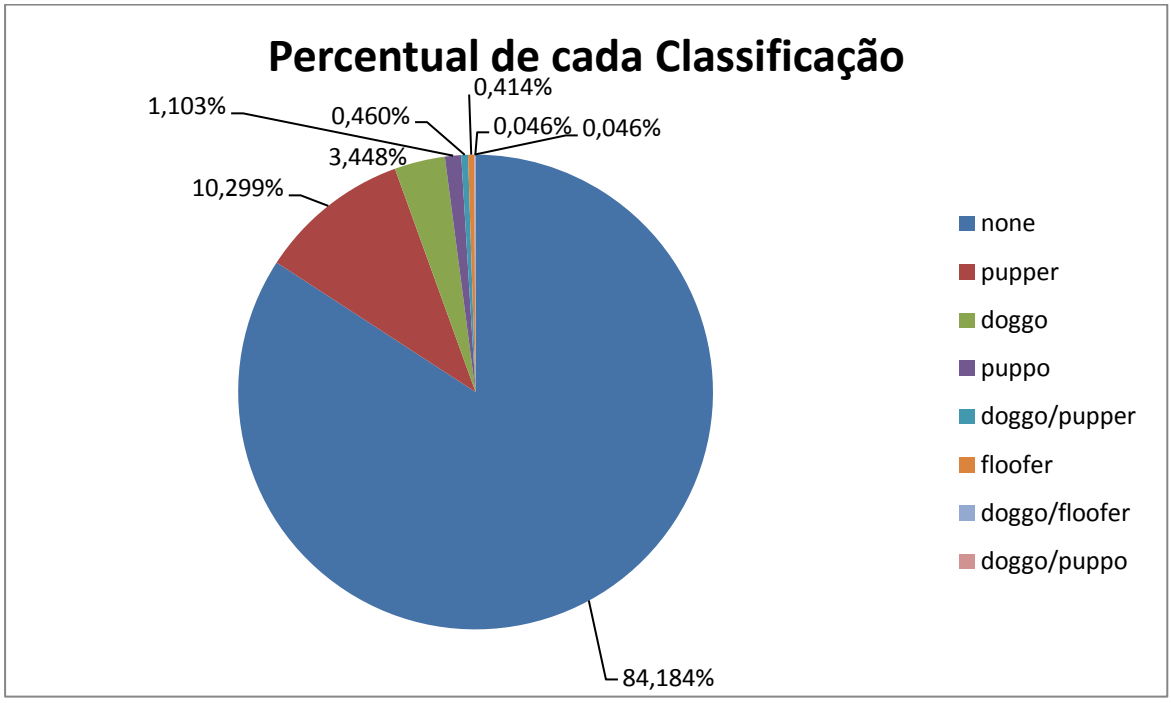
Através do agrupamento de dados por anos, conseguimos chegar em uma tabela que relaciona o ano do tweet, o qtd de tweets naquele ano e a média das principais medidas presentes no *DataFrame:*

TABELA MÉDIAS POR ANO						
data_tweet	qtd	confiabilidade_1	confiabilidade_2	confiabilidade_3	qtd_likes	qtd_retweets
		média	média	média	média	média
2015	580	0,570612	0,135882	0,632250	1642,166994	704,697446
2016	971	0,590240	0,134917	0,059729	6343,155474	2354,510580
2017	359	0,641745	0,134100	0,055755	21597,558897	5470,824561

Podemos notar que:

- A confiabilidade_1 é a que possui as maiores médias, porém como visto anteriormente, uma confiabilidade alta não significa que certeza da raça;
- No ano de 2017 foi obtida a maior média de *likes* e *retweets*, porém nesses dados foi considerado parte do ano de 2015 e 2017.

Em relação ao estágio de vida que o cachorro está, foi utilizada quatro classificações: *pupper, doggo, puppo, floofer*. Porém, nos dados há casos onde não houve classificação (, e outros casos onde houve mais de uma classificação (um possível motivo, é que a foto possuía mais de uma cachorro). Abaixo, temos um gráfico exemplificando a relação das quantidades de cada classificação.



Podemos notar que, para os dados que receberam classificação, os cachorros classificados como pupper estão presentes na maior quantidade de tweets, seguidos pelos doggo e puppo. Porém, os que não foram classificados, possuem aproximadamente 85% dos dados

Por conseguinte, está foi uma análise básica, sem a utilização “pesada” da estatística. Porém, podemos retirar algumas conclusões em relação ao algoritmo que classifica as raças e o estágio de vida. Tais algoritmos não foram tão precisos, possivelmente, por estarmos retirando tais informações de textos que muitas vezes não são padronizados. Uma possível melhoria seria o administrador da conta criar uma regra de como deve ser os comentários de apresentação do cachorro, criando um script a ser usado, contudo isso tiraria o humor espontâneo dos comentários.