Nanodregree Data Science II

Project 2: Wrangle and Analyse Data

Lucas Amorim Bonini
15/02/2019

Introduction

Real-world data rarely comes clean. Using Python and its libraries, gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The dataset analyzed is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Gather

The data are gathered from 3 different sources. a part comes from a file in the format csv, another comes from a download of a file in tsv format and lastly the API (Application Programming Interface) of the tweeter is used. All this using python libraries like pandas and tweepy.

Assess

The data is accessed using some built-in python functions that show the parcels of data collected, format and other variables where errors and probrems can be detected visually or programmatically.

Clean

A quick cleanup of the collected data is done only to demonstrate the "dirt" that comes with the data. Some of the detected issues related to quality and tidiness are cleaned and then the final file is stored in a csv format file.

Summary/Conclusion

At the end are made 3 very simple insights that can be made demonstrating some of the possibilities that can be made when analyzing tweets.

Insight 1 shows the kind of dog that has more retweets and favorites.
Insight 2 shows the correlation between retweets and favorites.
Insight 3 shows which type of dog is most prevalent.