UDACITY

DATA ANALYTICS FUNDAMENTALS NANODEGREE

Project 04: Wrangle and Analyze Data
Data Wrangling Report

Hernán Adasme

1 Introduction

Real-world data rarely comes clean. This project wrangles, cleans, analyzes, and visualizes data from the tweet archive of Twitter user WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comments about them. By Using Python and its libraries like Numpy, Pandas and Matplotlyb, the data has been gathered, cleaned and analyzed. The following sections describe the "wrangling effort" performed on data.

2 Data Gathering

The project gathered data from three different sources:

• The WeRateDogs Twitter archive. The twitter archive enhanced.csv provided by Udacity for its Students, contains tweet data for all 5000+ of their tweets as they stood on August 1, 2017.

• The tweet image predictions (what breed of dog or other object, animal, etc.) that is found in each tweet according to a neural network. The file image-predictions.tsv is hosted on Udacity's server. It has been downloaded programmatically using the Requests library and the following URL:
https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad-image-predictions/image-predictions.tsv

• Each tweet's retweet count and favorite ("like") count at minimum. This data was gathered by using the tweet IDs in the WeRateDogs Twitter archive. The data was requested through a query from the Twitter API, for each tweet's JSON data using Python's Tweepy library. Each tweet's entire set of JSON data has been stored in a file called tweet-json.txt file. Each tweet's JSON data should be written to its own line. Finally this txt file was read line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. This stage was specially hard, mostly because of the unfamiliarity of the method.

3 Assessing

Once the data was successfully gathered, data has been assessed in order to find both quality and tidiness issues. Messy and untidy data won´t allow the researchers to perform valid analysis.

• **Quality.** The data assessing process helped to find out the following quality issues:

> Missing values on columns in-reply-to-status-id, in-reply-to-user-id, retweeted-status-id, retweeted-status-user-id, retweeted-status-timestamp and expanded-urls

> Predictions with not enough confidence percentage

> Erroneous datatyopes on several columns

> None values in place of missing value tag: NaN

> No values on columns: doggo, floofer, pupper, puppo

> Predicted breeds names with lower and upper case

> Three different predictions with different highest confidence percentage

> Tables could be just in one big DataFrame: merge df-tweets with archive, by 'tweet-id

> Tables could be just in one big DataFrame: merge df-tweets with archive, by tweet-id.

• **Tidiness.** In a tidy data frame, each variable forms a column, each observation forms a row, each type of observational unit forms a table. The data assessing discovered the following two tidiness issues:

> Doggo, floofer, pupper, puppo column exist in different columns: they should be combined into a single column as this is one variable that identify stage of dog

> Information about tweets is spread across three different files/dataframes. These three dataframes should be merged as they are the same observational unit

4 Data Cleaning

The data cleaning process has been performed into the wrangle-act.ipynb. Quality issues have been assessed based on the following structure: define, code and test. Tidiness issues were also corrected by following the same three-step-method. The final result of the cleaning process is a tidy DataFrame, ready to analyze, and realiable as a source of information. This process was arduous but rewarding. By cleaning data I incorporated important elelemnts from Numpy and Pandas libraries.