Analysis of 'WeRateDogs' Twitter Datasets:

A project of the Data Analyst Nanodegree

by Gerrit Tombrink

The analysis of the 'WeRateDogs' Twitter datasets is a part of the Data Analyst Nanodegree, created by Udacity. The goal of this project is to wrangle, analyse, and visualize 'WeRateDogs' Twitter data in combination with additional prediction and Twitter datasets, downloaded via a URL and a Twitter API. The datasets consist of three parts:
The first dataset contains tweet data for all 5000+ of the 'WeRateDogs' Twitter archive, which was importedas a .csv file.
The second dataset was programmatically downloaded via a URL and imported as a .tsv file. It contains a table full of image predictions about breeds of dogs, which is based on the images of the 'WeRateDogs' Twitter archive.
The third dataset contains additional information on the tweets, which were programmatically downloaded via the Twitter API, and saved and imported as a .txt file.
After the gathering, assessing, and cleaning processes, the datasets were analysed. This data analysis aimed to answer the three following questions:

1. What is the average value (mean and median) for the predictions p1_conf, p2_conf and p3_conf?
2. How are the data records of p1_conf, p2_conf and p3_conf distributed?
3. What are the names of the dogs with the highest true prediction rates of p1_conf, p2_conf, and p3_conf?

Based on the column p1_conf it could be clarified that the mean of p1-confident is 0.593672800402211 and the median is 0.5874. The distribution of this data (p1_conf) is left-skewed (negatively-skewed) and the dog names with the highest true prediction rates are Buddy, Lenox and Shaggy. Additionally, some "None" values were found.
Based on the column p2_conf, it could be illustrated that the mean of p2-confident is 0.13436420311714425 and the median is 0.1175. The distribution of this data (p2_conf) is right-skewed (positively-skewed) and the dog name with the highest true prediction rate is Jiminus.
The additional column p3_conf showed that the mean of the p3-confident is 0.06021704374057313 and the median is 0.0495. The distribution of this data (p3_conf) is right-skewed (positively-skewed) and dog name with the highest true prediction rate is Bluebert.

In summary, we clarified that the higher probabilities of true predictions were found in the dataset of p1_conf. Furthermore, the histograms showed that the data of p1_conf were very well distributed in comparison to p2_conf and p3_conf (see Figure 1). The maximum value of p1_conf was found at the value of 1.0. There exist several dog names (Buddy, Lenox and Shaggy) with the highest true prediction rate of p1_conf.

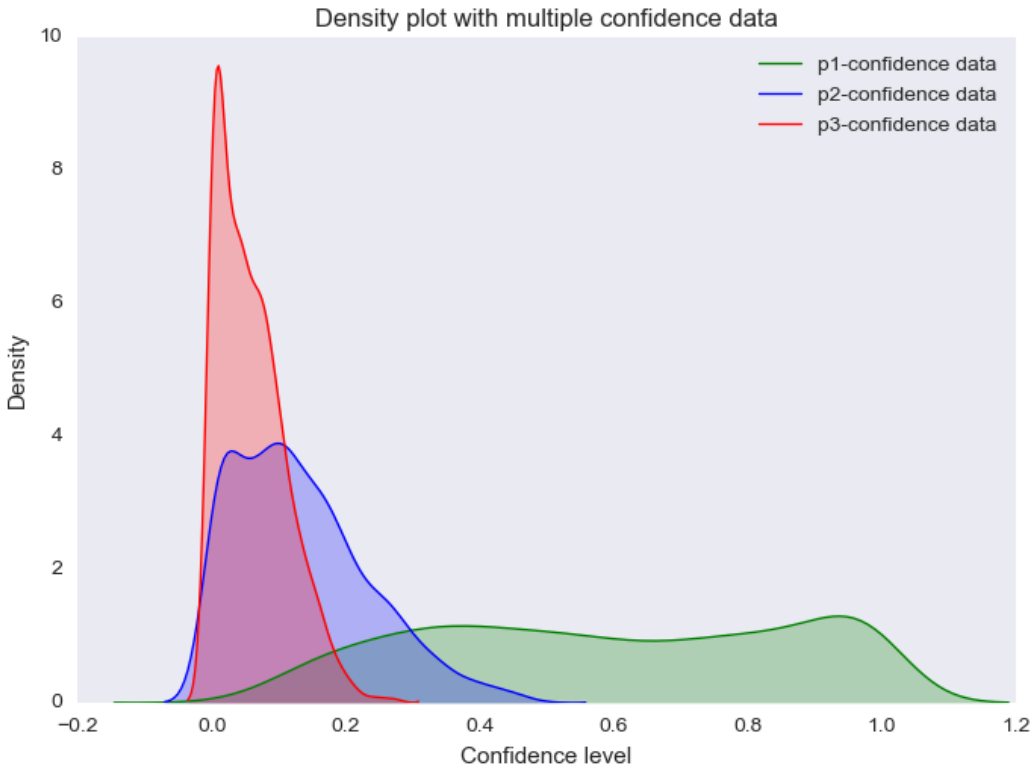

Figure 1: Histograms of the columns p1_conf, p2_conf and p3_conf.