Wrangle Report
Trevor Stuart
5.19.2018

I. Introduction

The following report will briefly go over the steps taken to wrangle the data for this project. Three steps were taken before analysis and visualizations could be made, those steps were: gathering data, assessing the data, and then cleaning the dataset.

II. Gathering Data

There were three sets of data that went into this project.

The first set was a CSV file that was given to us to simply download and then read into our jupyter notebook workspace using the pandas command – *read.csv.* This was the simplest part of the gathering process.

The second set was an image prediction file that had to be programmatically downloaded from a given URL.

The final data source was taken from Twitter's API using Python's Tweepy module. We then read that data into a pandas dataframe using the id column.

III. Assessing Data

While assessing the data, I used the panda functions *.head()* and *.tail()* to find most of the quality and tidiness issues. I also created some quick visualizations in some cases to find outliers. Looking at *.info()* and *.describe()* information also showed a lot of issues with the three datasets.

IV. Cleaning Data

Once the data was assessed, it was time to clean the dataframes. There were a number of issues that had to be programmatically fixed.

1) Quality Issues
   a) Removed '&amp' and replaced with '&' in the text column of the wrd dataframe.
   b) All unnecessary retweet data columns were removed.
   c) Changed 'id' column to 'tweet_id' in wrd_tweetinfo dataframe to match the other dataframes.
   d) Removed underscores and replaced with spaces for all of the dog breed names in the image prediction dataframe.
   e) Also, capitalized all of the dog breed names using the title function.
   f) Removed all non-name words in the name column of the wrd_dataframe.
   g) Removed duplicate tweets in the wrd_tweet.
   h) Converted all time and date info from string type data to datetime objects.

2) Tidiness Issues
   a) Combined the ‘type’ columns (doggo, floofer, pupper, and puppo) into a single column.
   b) Got rid of all unnecessary columns in the image prediction dataframe.