Wrangle Report

September 8, 2018

1 Data Wrangling Report

In this project, we wrangled over 2000 dog tweets that were that came from multiple sources of data. We gather the data via twitter api, requests and had data handed to us. We assessed the data for several quality and tidiness issues. We cleaned the data based on the issues we discovered and we organized the data into a database.

1.0.1 Gathering the Data

During the gather phase we had to use data from 3 different sources.

1. **Data was provided to us in a .tsv file called image_predictions.**
2. **Data had to be pulled from an external URL using the requests package. This was called twitter_archive.csv**
3. **Data had to be extracted from the twitter api via tweepy package and stored in a text file called tweet_json.txt**

The first two steps were fairly straight forward (using pd.read_csv and the request package), however the third datasource was a little more complicated. In order to extract the data from the third dataset we had to first query the twitter api using the get_status method and then extract all the json into a file line by line using the ._json method. We also had to make sure we were catching some errors with try and except since not all the data has every field we want. Once we had the tweet_json.txt file we had to read it line by line into a list of dictionaries so we could organize it into a dataframe which we called tweet_info.

1.0.2 Assessing the Data

During our assessment phase we had to look at the 3 different dataframes we created.

1. **twitter_archive**
2. **tweet_info**
3. **image_predictions**

With the first two data frames we discovered several quality and tidiness issues. Tweet_info had the wrong datatypes and was also missing some values for the media_type and media_url fields. It also had an error column that had two observations in it and needed to be stored in a seperate tables. Twitter_archive had several datatype issues, it also had issues where we had to use regex functions in order to parse out fields and seperate them. Image_predictions had very few issues and in the end the only issue we addressed is the mix of capatalized and lower case letters in the p1, p2, p3 columns.

1.0.3 Cleaning the Data

We broke up our cleaning phase into four steps, one for each data set and one for combining and organizing the data once each were clean.

We had to go back to some gather steps and had to re-assess many times during this phase. We had to gather more data from tweet_json.txt after we discovered that the media_type and media_url fields could be stored in embedded dictionaries. We had to create a recursive function to gather that data and were able to populate 98 values that were previously null. When we changed datatypes we had to sometimes fill null values so we would not get errors, we also had to do this when we parsed through data using regex to seperate the fields. In our cleaning phase we also discovered several tidiness issues which caused us to break up the 3 original datasets we had into 5 tables in our dog_twitter.db.

1.0.4 Conclusions

It is a long and iterative process to gather, clean and assess data. There were many times in the cleaning phases where we had to go back and re-assess and re-gather data. Overall we were able to create a clean database ther we could use for our Analysis phase.