Udacity Data Analyst Nano Degree Project 04: WeRateDogs

Swadeep Kumar Singh

Introduction

Real-world data rarely comes clean. Using Python and its libraries, we have gathered 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. Wrangling efforts in a Jupyter Notebook, explained through analyses and visualizations using Python (and its libraries). The dataset for wrangling (and analyzing and visualizing) 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. WeRateDogs has over 4 million followers and has received international media coverage.

Our goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Method

The *Data wrangling Project* is organized into following parts:

1. Gathering data
2. Assessing data
3. Cleaning data
4. Storing, analyzing, and visualizing your wrangled data

I. Gathering Data

In this project we need to gather data from 3 resources.

1. The WeRateDogs Twitter archive. twitter_archive_enhanced.csv
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network.
   - This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
   - Alternatively the file can also be downloaded directly as described in tutorial.
3. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting.
   - Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.
   - We can also accessing Project Data without a Twitter Account even if we don't have twitter account as described in Twitter API tutorial.

II. Assessing Data

After gathering each of the above pieces of data, we assess them visually and programmatically for quality and tidiness issues.

Visual assessment for quality and tidiness issues

During the data gathering stage in order to get a feel for the data quality and tidiness, the csv files for each dataset were downloaded and opened in Excel. From excel, we can quickly identified:
8 quality issues can be identified based on completeness, validity, accuracy, & consistency.
2 tidiness issues which are with structure that prevent easy analysis.

Quality issues

1. df, df_image are different probably due to retweets and missing photos.

In "twitter-archive-enhanced.csv"

2. **name** of the dog is "None", or incorrect such as "a", "an", & "by"
3. Entries were missing or Null in **expanded_urls**
4. check if all four dog stages are None
5. 181 retweets and 78 replies which not needed
6. rating_denominator is not always 10 and rating_numerator is extremly high at some places
7. not all the data are in their most appropriate data type
8. consistent names and logical order to column data as final step in cleaning

Tidiness

1. Four dog stages columns in twitter-archived dataframe can be combined in one single column.
2. From three prediction columns in image-tweet dataframe, extract the best dog breed prediction including confidence value.
3. Select only columns important for the analysis from tweet_json dataframe.

Programmatically assessment for quality and tidiness issues

After visual assessment of gathered data we perform programmatically assessment using below commands and logic:

1. len(df)
2. df.value_counts()
3. df.isnull().value_counts()
4. df["doggo"].value_counts()
5. df.query(" rating_numerator > 13")["rating_numerator"]
6. df.dtypes
7. df.columns
8. df.sample(2)
9. df.info()
10. df.head()

III. Clean Data

In data cleaning phase we address the issues as described in *"accessing data"*. The results is of high quality and tidy master pandas DataFrame.

1. Combine dog stages in one single column,
2. Select only columns important for the analysis from dataframes,
    - Join data frames
    - Remove retweets and replies
    - Remove tweet that don't have image urls
    - Clean source column
    - Change timestamp datatype
3. Extract the best dog breed prediction & confidence value.
4. Clean remaining datatypes (tweet_id to string, rating_denominator & rating_numerator to float)

IV. Storing, Analyzing, and Visualizing Data

We store the final dataframe after cleaning in a CSV file as twitter_archive_master.csv. We also analysis and visualize the results using python libraries and Tableau. In the analysis we have tried to explore the top dog breeds, average ratings, tweet favorite count, and sources for the tweets (devices). On the other hand Tableau, which is data visualization tool for creating interactive plots and graphs is used to see the trend of dog names, and on which day of week favorite count & retweet count is more.

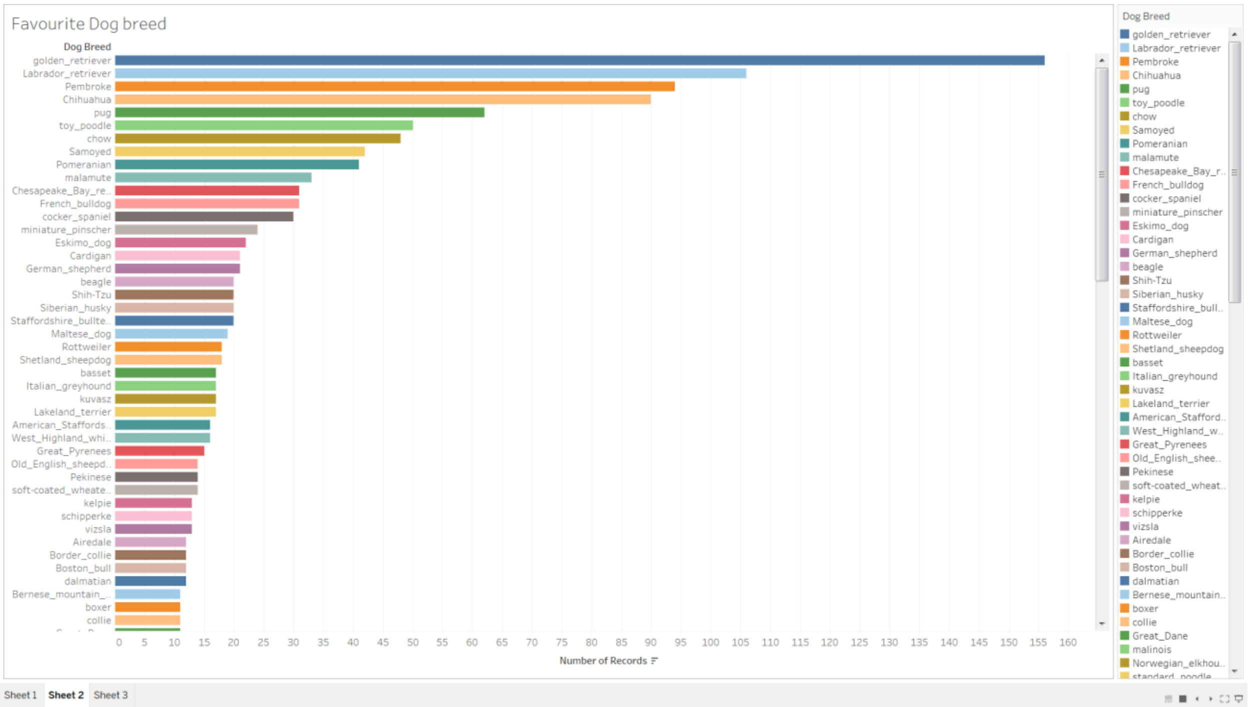
After performing the above analysis we can be able to provide top insights as follows:

Top insights after analysis

1. Most favorite dog breed which is tweeted on WeRateDogs are:
   - Golden retriever: 156
   - Labrador retriever: 106
   - Pembroke: 94
2. Most favorite count and retweet count comes on Monday.
3. Most favorite dog stage mentioned in tweet is pupper.
4. Famous dog names were Charlie, Cooper, Lucy, Oliver, and Tucker.
5. Most of the tweets came from iPhone.

Top visualization

Most favorite dog breed which is tweeted on WeRateDogs



Conclusion

The final master dataframe is store in CSV and shared as **twitter_archive_master.csv**. The wrangle_act.ipynb contains the code for data wrangling. In these report we have briefly described our wrangling efforts and also visualized the results for better communication.