Report: act_report - WeRateDogs – Insights into the @dog_rates Twitter page

Introduction

Real-world data rarely comes clean. The dataset wrangled for this project 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 – Insights into the @dog_rates Twitter page

This project works through the data wrangling process, focusing on the gathering, assessing, and cleaning of data. This report focuses on the visualizations and observations from tmy analysis provided as well.
The details of the wrangling process can be studied in Jupyter Notebook titled "wrangle_act.ipynb" on my Github account: https://github.com/darraghmerrick

/Twitter_API_Data_Analysis/blob/main/wrangle_act.ipynb.

The wrangling report can be read also: https://github.com/darraghmerrick /Twitter_API_Data_Analysis/blob/main/wrangle_report.ipynb

This project gathered data from the following sources:

1. The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students.
   WeRateDogs downloaded their Twitter archive and sent it to Udacity via email for NanoDegree students to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

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 was provided to Udacity students via a URL, which could be downloaded directly into the Jupyter ntebook using python commands.

3. Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite count, which was missing from the archive we were given.

In all 3 data sets, the 'tweet_id' field is the primary key that ties all the tables together. After the cleaning and tidying, the data sets can be merged together into one datasset.

Assess

Assessing data requires data analysts to evaluate a data set on quality and tidiness issues. The four (4) main data quality dimensions are: • Completeness: missing data? • Validity: does the data make sense? • Accuracy: inaccurate data? (wrong data can still show up as valid) • Consistency: standardization?

And there are three (3) requirements for tidiness: • Each variable forms a column • Each observation forms a row • Each type of observational unit forms a table

Clean The cleaning process involves three steps:

1. Define: determine exactly what needs to be cleaned, and how
2. Code: programmatically clean the code
3. Test: evaluate the code to ensure the data set was cleaned properly

Insights from Data after Wrangling

1. Looking at the data we can visualise which dog names are most popular.

The top 5 dog names are: a. Lucy b. Charlie c. Tucker d. Penny e. Oliver

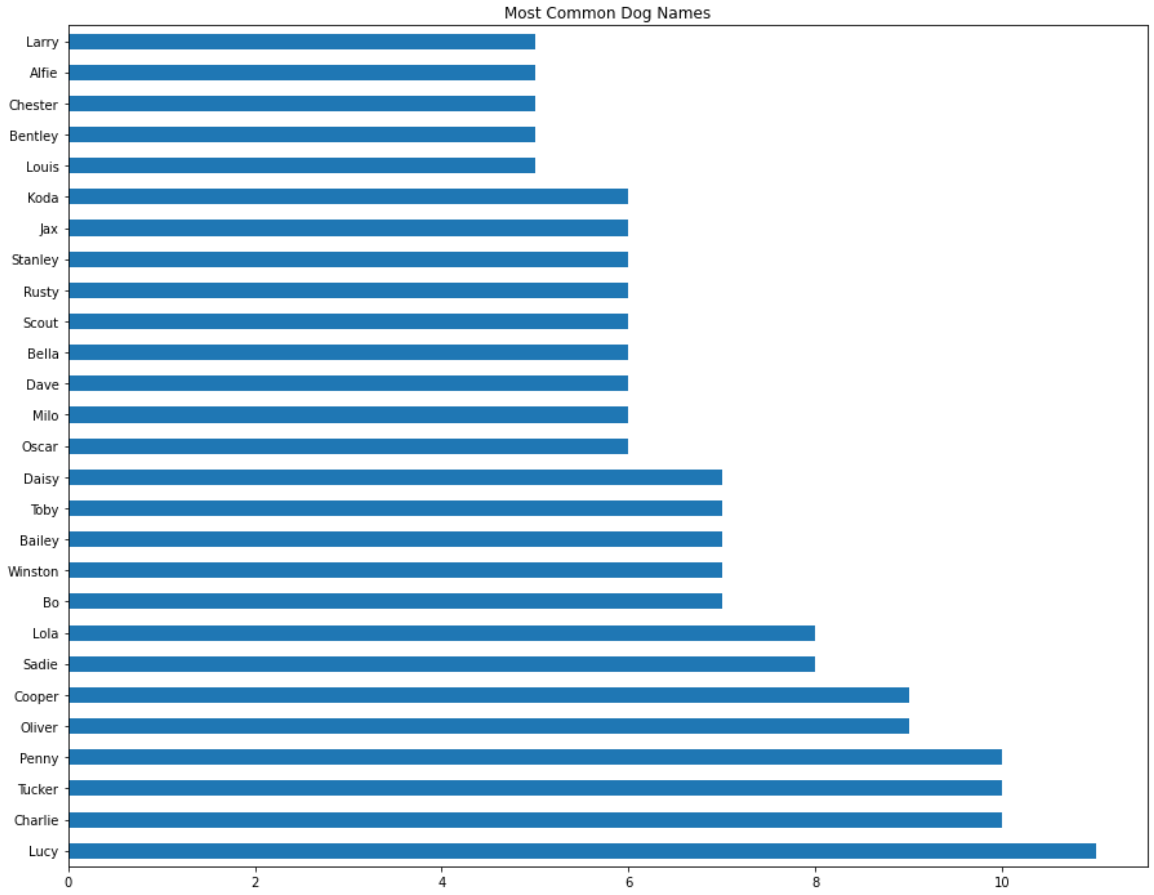
1. Next I looked at the DoggoLingo dog type and see which was favoutited the most. The most popular tags were; a. doggo, puppo b. puppo c. doggo, floofer

I didn't expect for there to be 2 values for 1 dog, but there were multiple rows, with 2 values.

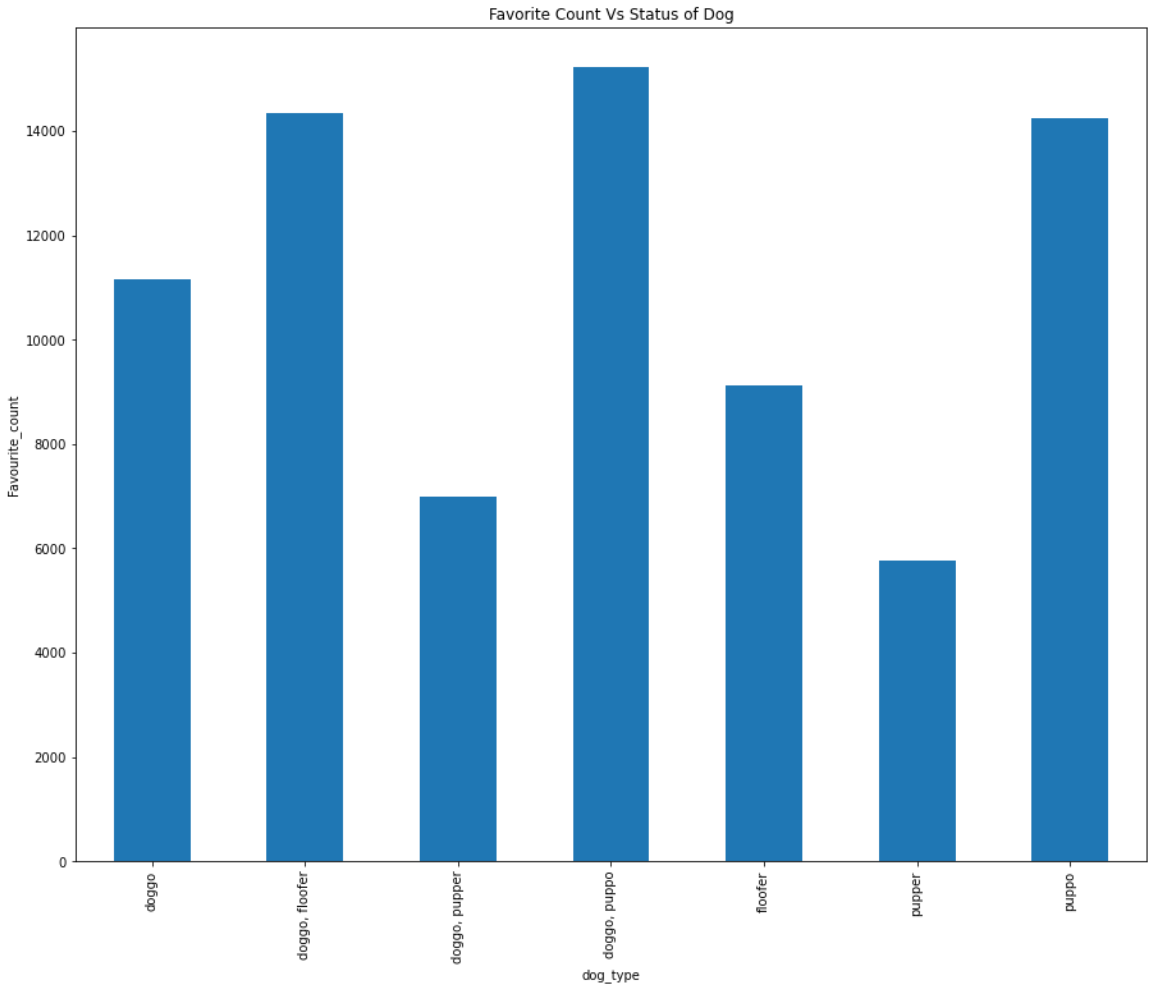
1. From the image predictions file, I counted the image file, which had the most occurance.
2. I found out which dogLingo type is most frequent and visualised it in a pie chart.
3. Now that we've fixed the datatype from string to datetime, we can use the datetime datatype to find the ealiest and latest dates of the timeframe the twiiter achive spans and plot it, monthly to see the WeRateDogs Tweets over time.

Visualisations

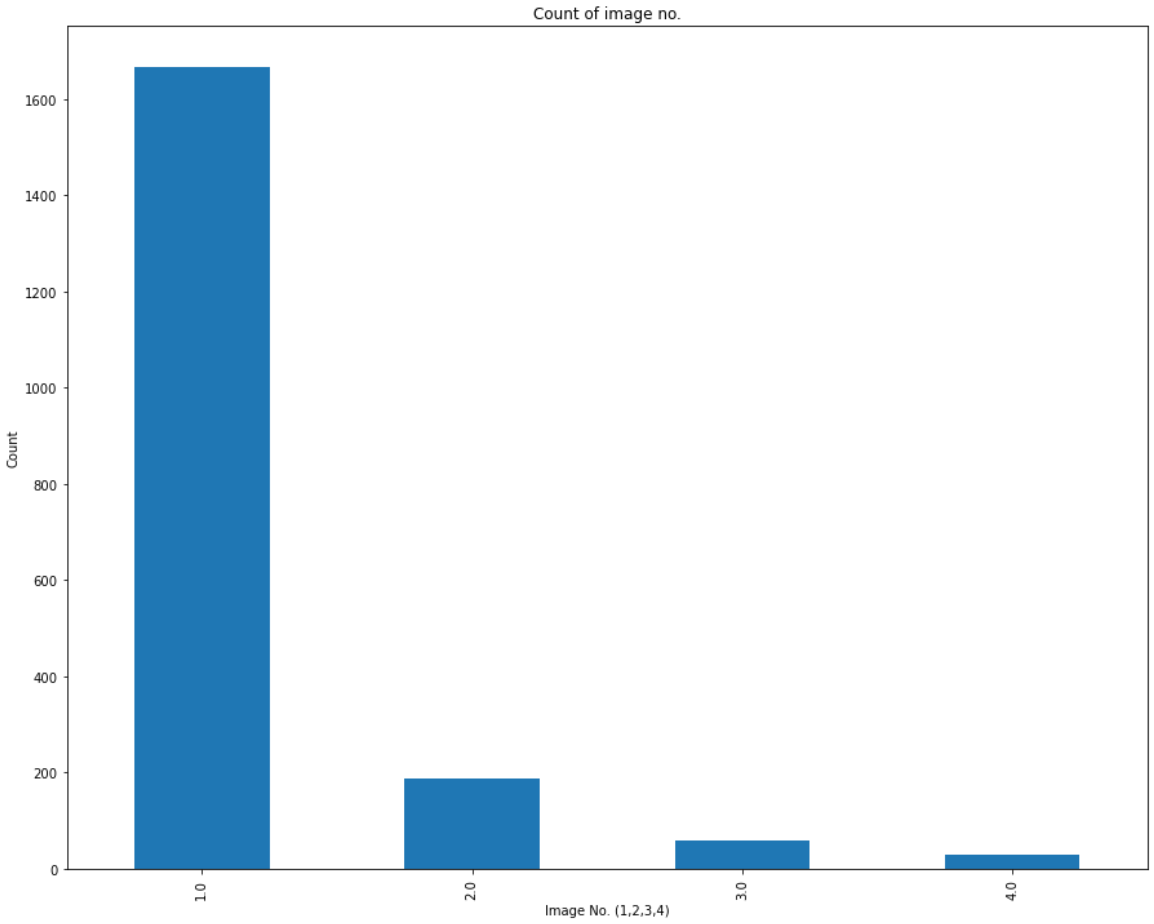
Visualisation 1 - Most Common Dog Names



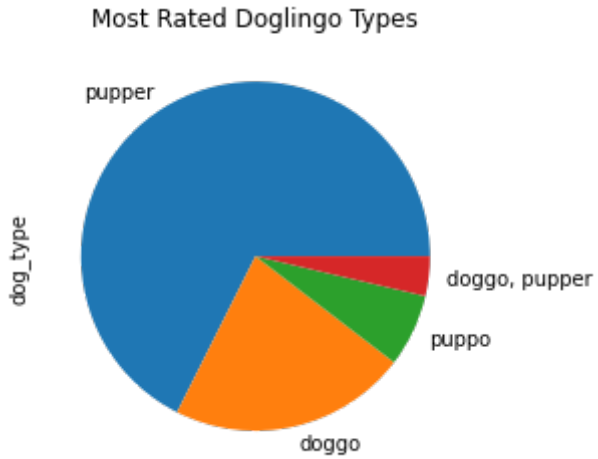
Visualisation 2 - Most Favourited DogLingo type



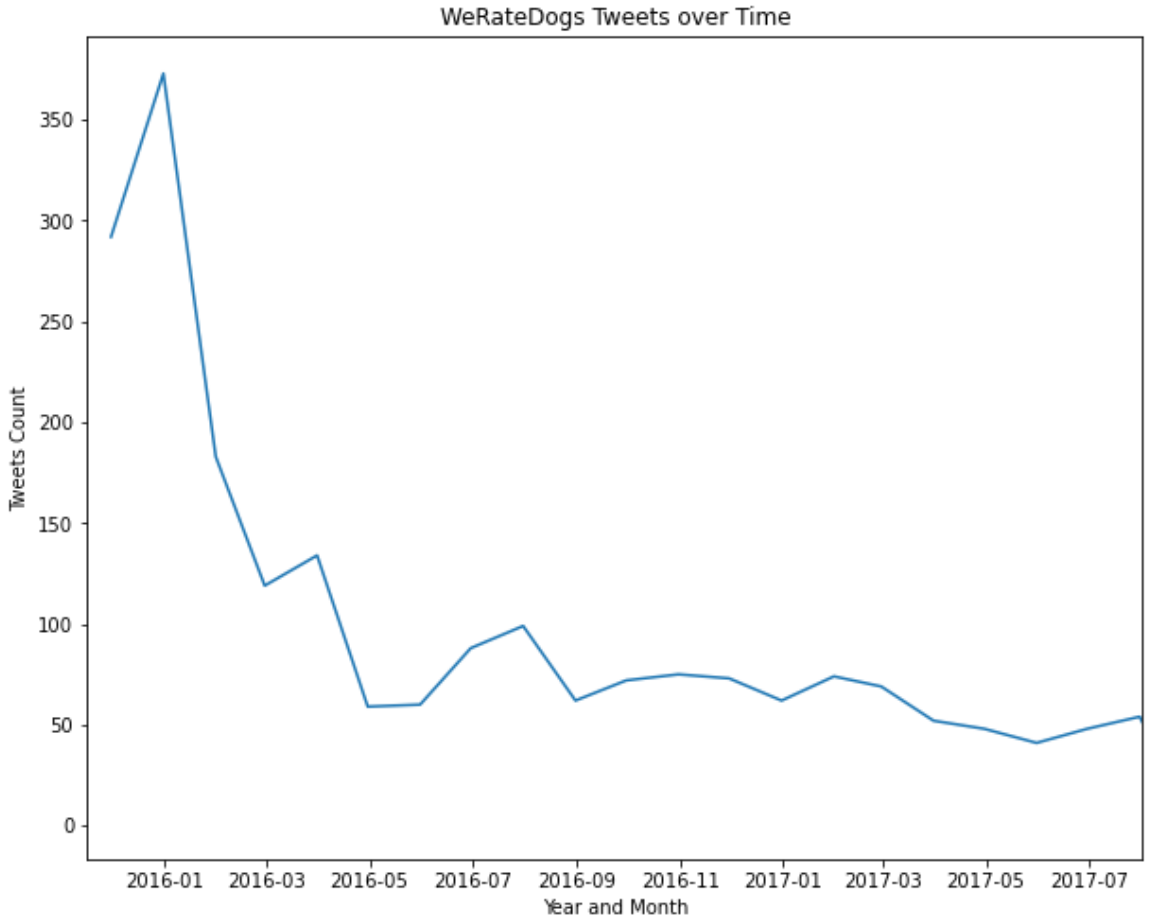
Visualisation 3 - Most Frequent Image number from the prediction neural network data set.



Visualisation 4 - Most rated DogLingo types



Visualisation 5 - WeRateDogs Tweets over time



Project Summary

This was a difficult project. It took a significant amount to get set up as a twitter developer to access the Twitter api, to analyse the 3 data sets, find the quality and tidiness issues, deliver insight from the data and visualise them. Python is essential to iterate the data and correct it. Visual assessment is prone to human error and does not scale to handle large data sets. Not having the correct data types makes the data useless. For example the time column was a string, and until we changed it to date time, we could not have viewed the WeRateDogs

Tweets over time.

This project was an eye opener, that when compiling data together from multiple sources, it will not be clean and tidy and will require wrangling to get it into the required state. For work and reporting purposes we need to turn these long complex data sets into simple visualisations to tell a story or deliver a message about trends figures to our readers. This is why the data analyst is so important into today's data driven society.

In [ ]: