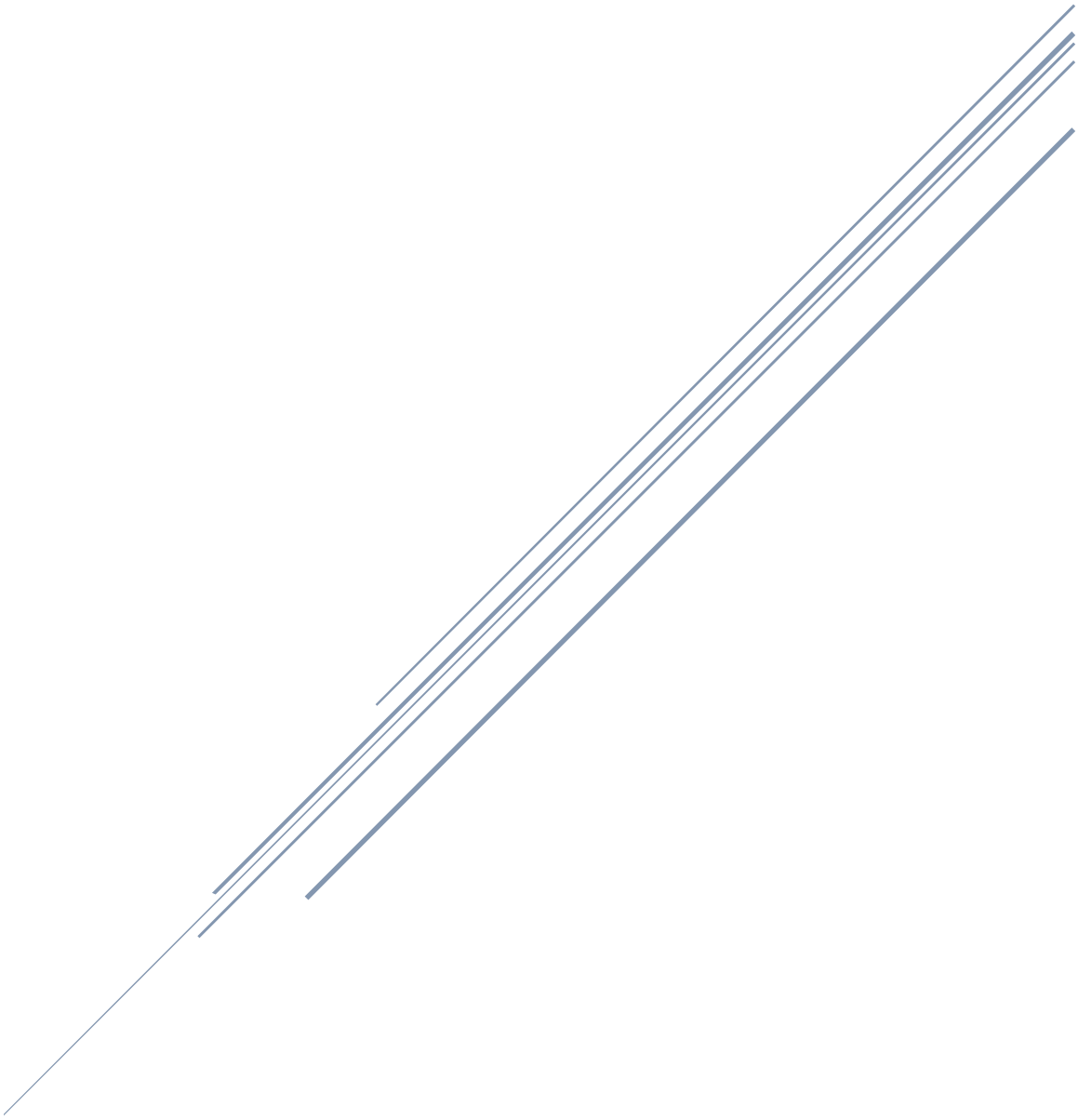
WRANGLE AND ANALYZE DATA

Wrangle Report



Udacity Data Analysis Nanodegree

Introduction

Real-world data is rarely accurate. I will collect data in a variety of forms and sources using Python and its libraries, evaluate its quality and tidiness, and then clean it. We call this "data wrangling." I'll keep track of my data wrangling activities in a Jupyter Notebook and present them using Python analyses and visualizations. The dataset that I will be 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.

The steps I will follow for data wrangling are:

- Gathering Data.
- Assessing Data.
- Cleaning Data.
- Storing Data.

Gathering Data

This is the first step for data wrangling, I will gather all three pieces of data which are the WeRateDogs Twitter archive data, The tweet image predictions and Additional data from the Twitter API for this project and load them in the notebook.

The three pieces of data which are:

1. The WeRateDogs Twitter archive:

I downloaded directly the WeRateDogs Twitter archive data by download (twitter-archive-enhanced.csv) manually then uploaded it and read the data into a pandas dataframe by using pd.read_csv() method.

2. The tweet image predictions:

I downloaded the tweet image prediction data (image_predictions.tsv) programmatically by using the Requests library.

3. Additional data from the Twitter API:

I downloaded tweet-json.text file manually then I read it to extract necessary data which is tweet id, retweet count and favorite count, after that I store all of this data with each other in dataframe.

As a summary of the gathering data step, we gathered three dataframes in different methods which is:

1. twitter_archive_df.
2. image_predictions_df.
3. tweet_json_df.

Assessing Data

This is the second step for data wrangling, in this step I will assessing the three dataframes that we obtained through the previous gathering data step to identify data quality issues and data tidiness issues in two ways as following:

- Visual Assessment : This assessment is very simple it is done by scrolling to check the data visually to take a general overview of the dataframes.
- Programmatic Assessment: This assessment is done by coding utilizing various methods and functions to give information about data quality issues and data tidiness issues. In this assessment I used a lot of functions such as: sample(), info(), describe(), value_counts(), isnull() and sum().

After we finished both visual and programmatic assessment, we got the following quality and tidiness issues for all three dataframes:

1- Quality issues:

Data that has problems with its quality (data content), such as missing data, invalid data, incorrect dates, or inconsistent data. For our data we got the following quality issues:

twitter_archive_df

1. We only want original ratings (no retweets) that have images, so we should delete retweets and replies.
2. Missing values in the following columns:in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls.
3. Wrong datatype for tweet_id and timestamp. tweet_id it is a string (object) datatype not integer and timestamp it is a datetime datatype not string (object)
4. (None) value instead of (NaN) value in name column.
5. Invalid values of rating_denominator such as: 120 and 170, because ratings almost always have a denominator of 10.

image_predictions_df

6. Wrong datatype for tweet_id. tweet_id it is a string (object) datatype not integer.
7. There are 66 duplicated jpg_url.
8. Inconsistency in the data in each of the following columns:p1, p2 and p3 some of the data are written in uppercase letters, some in lowercase letters, and some contain a underscore sign (-).
9. Delete wrong predictions.
10. Delete unused column.

tweet_json_df

11. Wrong datatype for tweet_id. tweet_id it is a string (object) datatype not integer.

2- Tidiness issues:

Data with special structural problems (columns, rows or table) that impede down cleaning, analyzing, and displaying. For our data we got the following tidiness issues:

twitter_archive_df

1. The last four columns:doggo, floofer, pupper and puppo should be in one column called dog_stage.

image_predictions_df

2. The following three columns:p1, p2 and p3 should be in one column called breeds_of_dogs.
3. The following three columns:p1_conf, p2_conf and p3_conf should be in one column called dog_prediction_confidence.
4. Combine all three dataframes:twitter_archive_df_copy, image_predictions_df_copy and tweet_json_df_copy into one dataframe based on tweet_id column.

Cleaning Data

This is the third step for data, in this step first ,we make copies of original pieces of data to avoid losing or changing data in case of error by using copy() method, then we clean all quality and tidiness issues we got from the previous assessing data step. I used the define-code-test framework for each issue as following:

- Define: In words wrangling, describe how you intend to solve the issue.
- Code: Transform your define step to programming code.
- Test: Test the code you wrote to make sure it works properly.

After we solved all quality and tidiness issues in the pieces of data, we combine all three dataframes:twitter_archive_df_copy, image_predictions_df_copy and tweet_json_df_copy into one dataframe based on tweet_id column.

Storing Data

In this step, I will save the gathered, assessed, and cleaned master dataset to a CSV file named "twitter archive master.csv" . Thus, the data became ready for the step of data analysis and visualization.