20.04.2020

WRANGLE REPORT

1. INTORDUCTION

The wrangle report is done to explain what I learned from Udacity's Nano Degree program, subsection Data Wrangle. The data which is wrangled here is a tweet archive of the twitter account @dogrates, which is also known as WeRateDogs. This report consists of general steps involved in the wrangling process for this twitter archive.

Project Section

i. Gathering Data
ii. Assessing Data
iii. Cleaning Data

2. GATHERING DATA

The data file we used for this wrangle were obtained from three different sources.

i. **Twitter Archive File:** This file was provided by Udacity
ii. **Tweet Image Prediction**: Hosted on Udacity's server, I downloaded this file programmatically using the library Request.
iii. **Twitter API and Json**: Using python's tweepy library, I gathered information which was used to create the retweet count and favorite count.

3. ASSESSING DATA

The data was assessed based on the Quality and Tidiness.

The basic function used to check the data was done using both the following commands in Jupyter notebook.

i. .head()
ii. .sample()
iii. .info()
iv. .value_counts()

The csv file was manually assessed using the Excel file to see if there was any issues with the data through visual inspection.

The issues we found were separated as mentioned above and answered in the Juypter notebook. When working with cleaning the data, I had a general Idea on

what the end result should be and what are the columns and data I require to create a visual representation. Based on these requirements the data was Assessed.

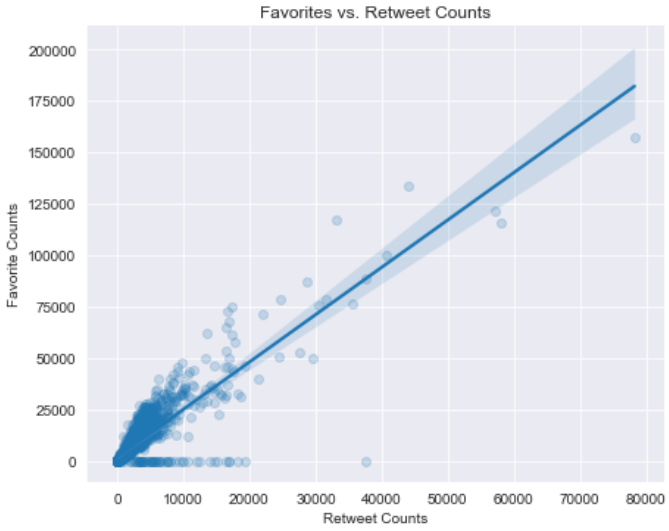
4. CLEANING

Based on the Assessment of the data and issues identified, I did the cleaning of the data and this was separated in three different sections.

I. Define
II. Code
III. Test

These sections are available for each issue I raised and resolved. The initial step was to create the copies of the data frames so that when I make a mistake by accidentally chaining the data I can always revert back. Not just that this helped to toy with the idea of trying different verities of Ideas and if it doesn't execute the way I want the data to be, I could go back and start from first.

The results were finally represented visually. Some of the common visualizations were the favorites vs retweets counts.



I also did a couple of different analysis and that is available in the attached Juypter notebook.

5. CONCLUSION

In real world, especially when you scrap data from internet, it would never be tidy. In some rare instances there is a probability we might get one but for the majority we need to do it. This project's emphasis is to make sure that I get trained in getting used to working with untidy data and the most general issues I would face in real world data and performing analysis on them.