Conor Parke – Udacity

Data Wrangling Project (WeRateDogs Twitter Data)

For this project I was set with the task of cleaning tweet data belonging to the popular WeRateDogs twitter account. This report describes the data wrangling process undertaken.

Gather – The data for this project was gathered from three sources. The twitter archive, which is a csv file containing the tweet, rating, dog name & dog stage data. An Image prediction file was a TSV file that held the tweet & associated tweet Image details, as well as prediction columns that tell us what dog breed the tweet image contains. The twitter API was accessed to collect retweet and favorite counts for each tweet. A Twitter developer account had to be created at this step-in order to gain access to the Twitter API account.

Assess – The data at this step was assessed visually to start with in order to identify any notable / easily identifiable issues with the data. This involved opening the data sources in excel to begin with, were the layout and structure was reviewed. The data was then programmatically assessed using Pandas, based on quality and tidiness predominantly.

Clean – Each of the items documented during the assessing phase then had to cleaned. The aim here was to deliver a tidier, higher quality dataframe at the end which was easier for a data analyst to work with and interpret.

All three phases were performed in the attached python project file: wrangle_act.ipynb in Jupyter Notebook.

Gather:

To start with, the necessary python libraries had to be imported.

I collected data from three data sources in total. The next step involved reading in the twitter archive csv file using read_csv(). A quick check was performed after to view the structure and layout of the dataframe using .head(). The tweet image predictions file was also downloaded from Udacity servers using the requests library. The tweet data from all the tweets was accessed using Twitter's API and the tweet data was stored in JSON format in a .txt file.

Assess:

Dataframe summary:

- twitter_archive
- tweet_extended_df
- image_predict

Each dataframe was assessed for quality and tidiness at this phase, both visual and programmatic assessment took place here.

Quality & Tidiness issues:

Quality:

- Unnecessary html tags in Source column. Convert into their values.
- Remove columns using drop() that hold no / low amounts of data or data that is not required.
- Identify and remove dogs from analysis that contain more than one dog stage.
- Identify and remove all rows that are not dogs.
- Bad naming of columns.
- No column available to interpret what gender the dog is. Extract from text column.
- Correct invalid name column entries
- Tweet Date values in incorrect data format
- Remove retweet related rows from dataframe.
- Incorrect values displaying in rating numerator columns for values with decimal.

Tidiness:

- Dog stages are spaced among multiple columns (4), create one single column called stage, Drop Dog Stage columns
- Extract dog breed and correlating confidence % from p#, p#_conf and p#_dog columns.
- All three data sources can be combined into one master dataframe.

Clean:

I performed the data cleaning process in three stages: Define, Code & Test. All the data cleaning steps carried out are documented in depth in wrangle_act.ipynb.

Store Data:

After completing the cleaning process, the final dataframe output file was exported out to Jupyter notebook in csv format.

Store Data

```
In [79]: #Exported master df out to a csv file to view data in new columns
         all_df_new.to_csv('MasterDF.csv')
```

Conclusion

I believe the objective was achieved in this project, the data was successfully collected from three separate sources, and it was extracted, assessed and cleaned for use. The data is now more understandable and easier to work with. Observations and visualizations can be gathered from this data now. I have shown some examples in the act_report.pdf file.