**Cagri Aslanbas – Data Wrangling Project Report**

I applied data wrangling on three datasets ("twitter\_archive\_enhanced.csv”, “image\_predictions.tsv” and “tweet\_json.txt”), only one of which was provided ("twitter\_archive\_enhanced.csv”).

* **Gathering:**

I have used the following methods to gather these 3 datasets:

1. **twitter\_archive\_enhanced.csv:** This was already provided. I’ve used pandas’ *read\_csv* method to load this file in a dataframe named “tw”.
2. **image\_predictions.tsv:** This file resides in the internet. I’ve used Requests library’s *get* method to to obtain it and used pandas’ *read\_csv* method to load this file in a dataframe named “im”.
3. **tweet\_json.txt:** This file was completely created by myself using Tweepy (Python Twitter API package). Then I’ve used pandas’ *read\_csv* method to load this file in a dataframe named “rt”.

* **Assessing:**

By using pandas’ *head, tail, info, describe, sample, value\_counts, duplicated* functions, I have found out some quality issues in *twitter\_archive\_enhanced* and *image\_predictions* dataframes. Furthermore I’ve also seen some tidiness issues in *twitter\_archive\_enhanced* and *tweet\_json* dataframes. The issues I’ve found and its types are as follows:

Quality Issues:

* “twitter-archive-enhanced” table:
  + timestamp is string not datetime **(Validity)**
  + name is entered as "a" sometimes instead of None **(Accuracy)**
  + 181 entries are retweets not original ratings **(Accuracy)**
  + 59 entries do not have images **(Accuracy)**
  + inaccurate rating\_denominator values (different than 10) **(Accuracy)**
  + duplicated expanded\_url values **(Accuracy)**
* “image-predictions” table:
  + All lowercase p1, p2, p3 sometimes, first letter is uppercase other times. **(Consistency)**
  + p1, p2, p3 containing breaks use underscore ("\_") sometimes, dash ("-") other times. (**Consistency)**

Tidiness Issues:

* One variable in four columns in “twitter-archive-enhanced” table (doggo, floofer, pupper, puppo) **(Each variable forms a column)**
* “rt” table, which contains retweet\_count and favorite\_count, should be part of the “twitter-archive-enhanced” table . **(Each type of observational unit forms a table)**
* **Cleaning:**

I’ve created copies of all the three dataframes before cleaning. I have cleaned the tidiness issues in first place and quality issues in second. I have used various functions I’ve found from different sources (sources are provided in the end of the *wrangle\_act.ipynb* file) to do cleaning.

In the end of cleaning process, I had two files that are ready for explanatory data analyses: *twitter\_archive\_master.csv* and *image\_predictions\_master.tsv.* As I have merged my two dataframes, tw and rt, there is no need for the tweet\_json.txt *file anymore.*

My cleaning process consisted of defining, coding and testing each 10 issues separately. I’ve made the data more tidy and fixed non-accurate, non-consistent and non-valid data.