Data Wrangling Report

Generally, Data wrangling goes through three phases that are:

1- Gathering data
2- Assessing data
3- Cleaning data

I will talk briefly about each phase of them in my project.

First Phase: Gathering data

To accomplish this, I used two methods that are loading file in hands and Programmatically Downloading files.

1- **Loading file in hands (**Files available to be manually downloaded**)**
   - "twitter-archive-enhanced.csv".
   - "tweet-json.txt"
2- **Programmatically Downloading (**Files available to be downloaded using a known URL**)**
   - "image-predictions.tsv"

For files.csv and files.tsv, they are easily read as dataframes using pd.read_csv(file) method.

For json files I needed to iterate over its content to extract important data tweet_id, favorite_count, retweet_count into dictionary, then save extracted data into a dataframe, then finally to csv file called "tweet_info.csv".

Finally, we gathered three dataframes that are:

1 - twitter_archive

2- image_predictions

3- tweet_info

Second Phase: Assessing data

With basics of assessing I managed accomplished this phase.

Data Quality Dimensions

- Completeness: Do we have missing records or not? Are there specific rows, columns, or cells missing?
- Validity: Do we have invalid values for our columns?
- Accuracy: Do we have wrong valid values in our columns?
- Consistency: Do we have multiple columns refer to same thing?

Data Tidiness

- Every column is a variable.
- Every row is an observation.
- Every cell is a single value.

Using Visual Assessing

Using methods df.sample() and explore data In Ms. Excel I figured out some data quality and tidiness issues that needed to be cleaned for further analysis.

Using Programmatic Assessing

Same task as in visual assessing but using more methods like

- Df.describe()
- Df.info()
- Df.column.value_counts()
- Df.duplicated().sum()

To get more descriptive information about data quality and tidiness.

After all I came out with these issues:

Quality Issues

- **twitter_archive Dataframe**

1- 181 non original ratings (retweets) should be removed.

2- Completeness: missing values in columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp,name, expanded_urls)

3- Validity: invalid values in "name" columns like(a,none)

4- Validity: invalid timestamp type (str) instead of (datetime)

5- Accuracy: 440 rating_numerator values are less than 10 also some values are very high [outliers].

6- Accuracy: 23 rating_denominator values are not equal to 10

- **image_predictions Dataframe**

7- Completeness: missing id values in columns 2075 id compared to 2365 in twitter_archive Dataframe so missing images.

8- inconsistency: some values in columns (p1, p2, p3) start with small letter while others start with capital one.

- **tweet_info Dataframe**

9- Completeness: misssing id values in columns 2354 id compared to 2365 in twitter_archive Dataframe so missing data about favorite_count,retweet_count for some ids.

Tidiness Issues

- **twitter_archive Dataframe**

1- Every cell is a single value: timestamp column has an observation that can be splited into separate columns [hour, month, day] for analysis.

2- Every column is a variable: (doggo, floofer, pupper, puppo) columns can be merged in one column.

- **General**

3- Tweet_info Dataframe and image_predictions Dataframe can be merged to twitter_archive Dataframe to form one detailed dataframe

Third Phase: Cleaning data

I used the three steps rule **(Define - Code -Test)** to solve critical issues that are listed above which can affect accuracy of our analysis.