WRANGLE REPORT

INTRODUCTION

This report contains the data wrangling work for Udacity Data Analyst Nanodegree project project 4:Wrangle and analyze data.

The data wrangling in this project included the following processes:

1. Gather data

From at least the three (3) different sources in at least the three (3) different file formats on the Project Details page. Each piece of data is imported into a separate pandas DataFrame at first.

2.Assessment

Visual assessment and Programmatic assessment. it is required to detect atleast eight (8) data quality issues and two (2) tidiness issues, and include the issues to clean to satisfy the Project Motivation. Each issue is documented in one to a few sentences each.

3.Cleaning data

by applying the steps define, code, and test

Save the master dataset at the end of cleaning process

4.Analysis and findings

Analyze the data to give insights

The master dataset is analyzed using pandas or SQL in the Jupyter Notebook and at least three (3) separate insights are produced and create atleast one (1) labeled visualization is produced in the Jupyter Notebook using Python's plotting libraries or in Tableau.

ISSUE No.1 'Empty Columns wholly filled with NaN'
Dropping columns with NaN

ISSUE NO.2 Timestamp column data type as Object
Changing Timestamp to date type

ISSUE NO.3 Empty Columns with NaN
Dropping columns media_url,media_url_https,favourites_count with NaN

ISSUE NO.4 Mismatched key column in tweet_json table
Renaming column in tweet_json table

ISSUE.NO.6. Missing values after matching datasets
Using isna().sum() functions the dataset contained the following NaN values.

```
expanded_urls          59
```

```
jpg_url              281
img_num              281
p1                   281
p1_conf              281
p2                   281
p2_conf              281
p3                   281
p3_conf              281
full text              2
```

ISSUE NO.7. rating_numerator column with values greater than 13

ISSUE NO.8. rating_denominator column with values greater than 10

SAVING TO MASTER DATASET

The clean dataset saved as twitter_archive_master.csv