# assess_data

August 3, 2021

0.1 # assess_data

Written in the Python 3.7.9 Environment

By Nicole Lund

This Jupyter Notebook reviews the header row for all of the downloaded csv files and looks for inconsistencies

[1]:
```
#Import Dependencies
import os
import pandas as pd
```

[2]:
```
# Get list of files in the folder
file_list = os.listdir()
```

[3]:
```
# Initialize empty dataframes
df_1 = pd.DataFrame()
df_2 = pd.DataFrame()
df_3 = pd.DataFrame()
```

[4]:
```
# Collect header and first row of data for all csv files
for file in file_list:
    if file[-3:] == "csv":
        csv_df = pd.read_csv(file,nrows=1)
        csv_df['filename'] = file
        if (df_1.size == 0):
            df_1 = csv_df
        elif (set(df_1.columns) == set(csv_df)):
            df_1 = df_1.append(csv_df)
        else:
            if (df_2.size == 0):
                df_2 = csv_df
            elif (set(df_2.columns) == set(csv_df)):
                df_2 = df_2.append(csv_df)
            else:
                if (df_3.size == 0):
                    df_3 = csv_df
                elif (set(df_3.columns) == set(csv_df)):
                    df_3 = df_3.append(csv_df)
                else:
                    print('Not enough dataframes')
    else:
        print(file + " not a csv")
```

```
assess_data.ipynb not a csv
retrieve_data.ipynb not a csv
__MACOSX not a csv
```

[5]:

```
df_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 131 entries, 0 to 0
Data columns (total 16 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   tripduration             131 non-null    int64
 1   starttime                131 non-null    object
 2   stoptime                 131 non-null    object
 3   start station id         131 non-null    int64
 4   start station name       131 non-null    object
 5   start station latitude   131 non-null    float64
 6   start station longitude  131 non-null    float64
 7   end station id           131 non-null    int64
 8   end station name         131 non-null    object
 9   end station latitude     131 non-null    float64
 10  end station longitude    131 non-null    float64
 11  bikeid                   131 non-null    int64
 12  usertype                 131 non-null    object
 13  birth year               128 non-null    object
 14  gender                   131 non-null    int64
 15  filename                 131 non-null    object
dtypes: float64(4), int64(5), object(7)
memory usage: 17.4+ KB
```

[6]:

```
df_2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 25 entries, 0 to 0
Data columns (total 16 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Trip Duration            25 non-null     int64
 1   Start Time               25 non-null     object
 2   Stop Time                25 non-null     object
 3   Start Station ID         25 non-null     int64
 4   Start Station Name       25 non-null     object
 5   Start Station Latitude   25 non-null     float64
 6   Start Station Longitude  25 non-null     float64
 7   End Station ID           25 non-null     int64
 8   End Station Name         25 non-null     object
 9   End Station Latitude     25 non-null     float64
 10  End Station Longitude    25 non-null     float64
 11  Bike ID                  25 non-null     int64
 12  User Type                25 non-null     object
 13  Birth Year               25 non-null     int64
 14  Gender                   25 non-null     int64
 15  filename                 25 non-null     object
dtypes: float64(4), int64(6), object(6)
memory usage: 3.3+ KB
```

[7]:
```
df_3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10 entries, 0 to 0
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ride_id             10 non-null     object
 1   rideable_type       10 non-null     object
 2   started_at          10 non-null     object
 3   ended_at            10 non-null     object
 4   start_station_name  10 non-null     object
 5   start_station_id    10 non-null     object
 6   end_station_name    10 non-null     object
 7   end_station_id      10 non-null     object
 8   start_lat           10 non-null     float64
 9   start_lng           10 non-null     float64
 10  end_lat             10 non-null     float64
 11  end_lng             10 non-null     float64
 12  member_casual       10 non-null     object
 13  filename            10 non-null     object
dtypes: float64(4), object(10)
memory usage: 1.2+ KB
```

[9]:
```
df_2.filename
```

[9]:
```
0       201610-citibike-tripdata.csv
0       201611-citibike-tripdata.csv
0       201612-citibike-tripdata.csv
0       201701-citibike-tripdata.csv
0       201702-citibike-tripdata.csv
0       201703-citibike-tripdata.csv
0    JC-201509-citibike-tripdata.csv
0    JC-201510-citibike-tripdata.csv
0    JC-201511-citibike-tripdata.csv
```

```
0    JC-201512-citibike-tripdata.csv
0    JC-201601-citibike-tripdata.csv
0    JC-201602-citibike-tripdata.csv
0    JC-201603-citibike-tripdata.csv
0    JC-201604-citibike-tripdata.csv
0    JC-201605-citibike-tripdata.csv
0    JC-201606-citibike-tripdata.csv
0    JC-201607-citibike-tripdata.csv
0    JC-201608-citibike-tripdata.csv
0    JC-201609-citibike-tripdata.csv
0    JC-201610-citibike-tripdata.csv
0    JC-201611-citibike-tripdata.csv
0    JC-201612-citibike-tripdata.csv
0    JC-201701-citibike-tripdata.csv
0    JC-201702-citibike-tripdata.csv
0    JC-201703-citibike-tripdata.csv
Name: filename, dtype: object
```

[10]:

```
df_3.filename
```

[10]:

```
0       202102-citibike-tripdata.csv
0       202103-citibike-tripdata.csv
0       202104-citibike-tripdata.csv
0       202105-citibike-tripdata.csv
0       202106-citibike-tripdata.csv
0    JC-202102-citibike-tripdata.csv
0    JC-202103-citibike-tripdata.csv
0    JC-202104-citibike-tripdata.csv
0    JC-202105-citibike-tripdata.csv
0    JC-202106-citibike-tripdata.csv
Name: filename, dtype: object
```