Ford-GoBike-Data-Wrangling

June 18, 2022

1 Part 0 - Ford GoBike Data Wrangling

1.1 by Dane

1.2 Introduction

This data set includes information about individual rides made in a bike-sharing system covering the greater San Francisco Bay area.

1.3 Preliminary Wrangling

This data was fairly clean and only needed some minor wranling.

[1]:
```
# import all packages and set plots to be embedded inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb

%matplotlib inline
```

1.3.1 Data Gathering & Exploration

[2]:
```
# Read in file & print first few statements
bikes = pd.read_csv('201902-fordgobike-tripdata.csv')
bikes.head()
```

[2]:
```
   duration_sec                start_time                  end_time  \
0         52185  2019-02-28 17:32:10.1450  2019-03-01 08:01:55.9750
1         42521  2019-02-28 18:53:21.7890  2019-03-01 06:42:03.0560
2         61854  2019-02-28 12:13:13.2180  2019-03-01 05:24:08.1460
3         36490  2019-02-28 17:54:26.0100  2019-03-01 04:02:36.8420
4          1585  2019-02-28 23:54:18.5490  2019-03-01 00:20:44.0740

   start_station_id                                start_station_name  \
0              21.0  Montgomery St BART Station (Market St at 2nd St)
1              23.0                     The Embarcadero at Steuart St
2              86.0                           Market St at Dolores St
3             375.0                           Grove St at Masonic Ave
```

```
4               7.0                                Frank H Ogawa Plaza

   start_station_latitude  start_station_longitude  end_station_id  \
0               37.789625              -122.400811            13.0   
1               37.791464              -122.391034            81.0   
2               37.769305              -122.426826             3.0   
3               37.774836              -122.446546            70.0   
4               37.804562              -122.271738           222.0   

                               end_station_name  end_station_latitude  \
0                Commercial St at Montgomery St             37.794231   
1                            Berry St at 4th St             37.775880   
2  Powell St BART Station (Market St at 4th St)             37.786375   
3                        Central Ave at Fell St             37.773311   
4                         10th Ave at E 15th St             37.792714   

   end_station_longitude  bike_id   user_type  member_birth_year  \
0            -122.402923     4902    Customer             1984.0   
1            -122.393170     2535    Customer                NaN   
2            -122.404904     5905    Customer             1972.0   
3            -122.444293     6638  Subscriber             1989.0   
4            -122.248780     4898  Subscriber             1974.0   

  member_gender bike_share_for_all_trip  
0          Male                      No  
1           NaN                      No  
2          Male                      No  
3         Other                      No  
4          Male                     Yes  
```

Initial data overview: data is pretty clean except for some missing data for gender and birthyear.

[3]:
```
# Print overview of data types
bikes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype  
---  ------                   --------------   -----  
 0   duration_sec             183412 non-null  int64  
 1   start_time               183412 non-null  object 
 2   end_time                 183412 non-null  object 
 3   start_station_id         183215 non-null  float64
 4   start_station_name       183215 non-null  object 
 5   start_station_latitude   183412 non-null  float64
 6   start_station_longitude  183412 non-null  float64
 7   end_station_id           183215 non-null  float64
 8   end_station_name         183215 non-null  object
 9   end_station_latitude     183412 non-null  float64
 10  end_station_longitude    183412 non-null  float64
 11  bike_id                  183412 non-null  int64
 12  user_type                183412 non-null  object
 13  member_birth_year        175147 non-null  float64
 14  member_gender            175147 non-null  object
 15  bike_share_for_all_trip  183412 non-null  object
dtypes: float64(7), int64(2), object(7)
memory usage: 22.4+ MB
```

It looks like the data is pretty clean except for a few data type issues and some missing data.

1.3.2 Quality issues

1. start_time & end_time are strings and should be datetime.
2. start_station_id & end_station_id are floats and should be strings (also drop .0)
3. bike_id: int to string
4. member_gender: string to category
5. user_type: string to category
6. birth_year: float to int

1.3.3 Tidiness issues

1. Separate station information into a separate DF & file: (station)id, latitude, longitude, & name. In main DF, keep start_station_id & end_station_id in the column where the ID refers to the station DF.

1.4 Cleaning Data

1.4.1 Issue #1: start_time & end_time

Convert start_time & end_time columns to datetime (from strings)

[4]:

```
bikes['start_time'] = pd.to_datetime(bikes.start_time)
bikes['end_time'] = pd.to_datetime(bikes.end_time)
```

[5]:

```
bikes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   duration_sec             183412 non-null  int64
 1   start_time               183412 non-null  datetime64[ns]
 2   end_time                 183412 non-null  datetime64[ns]
 3   start_station_id         183215 non-null  float64
 4   start_station_name       183215 non-null  object
 5   start_station_latitude   183412 non-null  float64
 6   start_station_longitude  183412 non-null  float64
 7   end_station_id           183215 non-null  float64
 8   end_station_name         183215 non-null  object
 9   end_station_latitude     183412 non-null  float64
 10  end_station_longitude    183412 non-null  float64
 11  bike_id                  183412 non-null  int64
 12  user_type                183412 non-null  object
 13  member_birth_year        175147 non-null  float64
 14  member_gender            175147 non-null  object
 15  bike_share_for_all_trip  183412 non-null  object
dtypes: datetime64[ns](2), float64(7), int64(2), object(5)
memory usage: 22.4+ MB
```

[6]:

```
bikes.head()
```

[6]:

```
   duration_sec              start_time                end_time  \
0         52185 2019-02-28 17:32:10.145 2019-03-01 08:01:55.975
1         42521 2019-02-28 18:53:21.789 2019-03-01 06:42:03.056
2         61854 2019-02-28 12:13:13.218 2019-03-01 05:24:08.146
3         36490 2019-02-28 17:54:26.010 2019-03-01 04:02:36.842
4          1585 2019-02-28 23:54:18.549 2019-03-01 00:20:44.074

   start_station_id                                start_station_name  \
0              21.0  Montgomery St BART Station (Market St at 2nd St)
1              23.0                     The Embarcadero at Steuart St
2              86.0                           Market St at Dolores St
3             375.0                           Grove St at Masonic Ave
4               7.0                               Frank H Ogawa Plaza

   start_station_latitude  start_station_longitude  end_station_id  \
0               37.789625              -122.400811            13.0
1               37.791464              -122.391034            81.0
2               37.769305              -122.426826             3.0
3               37.774836              -122.446546            70.0
4               37.804562              -122.271738           222.0

                               end_station_name  end_station_latitude  \
0                Commercial St at Montgomery St             37.794231
1                            Berry St at 4th St             37.775880
2  Powell St BART Station (Market St at 4th St)             37.786375
3                        Central Ave at Fell St             37.773311
4                         10th Ave at E 15th St             37.792714

   end_station_longitude  bike_id   user_type  member_birth_year  \
0            -122.402923     4902    Customer             1984.0
1            -122.393170     2535    Customer                NaN
2            -122.404904     5905    Customer             1972.0
```

```
3             -122.444293     6638  Subscriber              1989.0
4             -122.248780     4898  Subscriber              1974.0

  member_gender bike_share_for_all_trip
0          Male                      No
1           NaN                      No
2          Male                      No
3         Other                      No
4          Male                     Yes
```

1.5 Issue #2: start_station_id & end_station_id

Convert stat_station_id & end_station_id to strings from floats and replace ‘.0’ with ’ ’ from string.

[19]:

```
bikes['start_station_id'] = bikes.start_station_id.astype(str).str.replace('.
 ↪0', '', regex = False)
bikes['end_station_id'] = bikes.end_station_id.astype(str).str.replace('.0',␣
 ↪'', regex = False)
bikes.head()
```

[19]:

```
   duration_sec               start_time                 end_time  \
0         52185 2019-02-28 17:32:10.145 2019-03-01 08:01:55.975
1         42521 2019-02-28 18:53:21.789 2019-03-01 06:42:03.056
2         61854 2019-02-28 12:13:13.218 2019-03-01 05:24:08.146
3         36490 2019-02-28 17:54:26.010 2019-03-01 04:02:36.842
4          1585 2019-02-28 23:54:18.549 2019-03-01 00:20:44.074

  start_station_id                                start_station_name  \
0               21  Montgomery St BART Station (Market St at 2nd St)
1               23                     The Embarcadero at Steuart St
2               86                           Market St at Dolores St
3              375                           Grove St at Masonic Ave
4                7                               Frank H Ogawa Plaza

   start_station_latitude  start_station_longitude end_station_id  \
0               37.789625              -122.400811             13
1               37.791464              -122.391034             81
2               37.769305              -122.426826              3
3               37.774836              -122.446546             70
4               37.804562              -122.271738            222

                               end_station_name  end_station_latitude  \
0                Commercial St at Montgomery St             37.794231
1                            Berry St at 4th St             37.775880
2  Powell St BART Station (Market St at 4th St)             37.786375
3                        Central Ave at Fell St             37.773311
```

```
4                          10th Ave at E 15th St             37.792714

   end_station_longitude  bike_id   user_type  member_birth_year  \
0            -122.402923     4902    Customer             1984.0
1            -122.393170     2535    Customer                NaN
2            -122.404904     5905    Customer             1972.0
3            -122.444293     6638  Subscriber             1989.0
4            -122.248780     4898  Subscriber             1974.0

  member_gender bike_share_for_all_trip
0          Male                      No
1           NaN                      No
2          Male                      No
3         Other                      No
4          Male                     Yes
```

[20]:

```python
bikes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   duration_sec             183412 non-null  int64
 1   start_time               183412 non-null  datetime64[ns]
 2   end_time                 183412 non-null  datetime64[ns]
 3   start_station_id         183412 non-null  object
 4   start_station_name       183215 non-null  object
 5   start_station_latitude   183412 non-null  float64
 6   start_station_longitude  183412 non-null  float64
 7   end_station_id           183412 non-null  object
 8   end_station_name         183215 non-null  object
 9   end_station_latitude     183412 non-null  float64
 10  end_station_longitude    183412 non-null  float64
 11  bike_id                  183412 non-null  int64
 12  user_type                183412 non-null  object
 13  member_birth_year        175147 non-null  float64
 14  member_gender            175147 non-null  object
 15  bike_share_for_all_trip  183412 non-null  object
dtypes: datetime64[ns](2), float64(5), int64(2), object(7)
memory usage: 22.4+ MB
```

1.6 Issue #3: bike_id

Convert bike_id from int to string.

[21]:

```python
bikes['bike_id'] = bikes.bike_id.astype(str)
bikes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   duration_sec             183412 non-null  int64
 1   start_time               183412 non-null  datetime64[ns]
 2   end_time                 183412 non-null  datetime64[ns]
 3   start_station_id         183412 non-null  object
 4   start_station_name       183215 non-null  object
 5   start_station_latitude   183412 non-null  float64
 6   start_station_longitude  183412 non-null  float64
 7   end_station_id           183412 non-null  object
 8   end_station_name         183215 non-null  object
 9   end_station_latitude     183412 non-null  float64
 10  end_station_longitude    183412 non-null  float64
 11  bike_id                  183412 non-null  object
 12  user_type                183412 non-null  object
 13  member_birth_year        175147 non-null  float64
 14  member_gender            175147 non-null  object
 15  bike_share_for_all_trip  183412 non-null  object
dtypes: datetime64[ns](2), float64(5), int64(1), object(8)
memory usage: 22.4+ MB
```

1.7 Issue #4: member_gender

Convert member_gende to pandas category without order.

[26]:

```
bikes.member_gender.value_counts()
```

[26]:

```
Male      130651
Female     40844
Other       3652
Name: member_gender, dtype: int64
```

[29]:

```
gender_list = ['Male', 'Female', 'Other']
gender_cat_type = pd.api.types.CategoricalDtype(categories = gender_list, ordered = False)
bikes['member_gender'] = bikes.member_gender.astype(gender_cat_type)
bikes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   duration_sec             183412 non-null  int64
 1   start_time               183412 non-null  datetime64[ns]
 2   end_time                 183412 non-null  datetime64[ns]
 3   start_station_id         183412 non-null  object
 4   start_station_name       183215 non-null  object
 5   start_station_latitude   183412 non-null  float64
 6   start_station_longitude  183412 non-null  float64
 7   end_station_id           183412 non-null  object
 8   end_station_name         183215 non-null  object
 9   end_station_latitude     183412 non-null  float64
 10  end_station_longitude    183412 non-null  float64
 11  bike_id                  183412 non-null  object
 12  user_type                183412 non-null  object
 13  member_birth_year        175147 non-null  float64
 14  member_gender            175147 non-null  category
 15  bike_share_for_all_trip  183412 non-null  object
dtypes: category(1), datetime64[ns](2), float64(5), int64(1), object(7)
memory usage: 21.2+ MB
```

[30]:

```
bikes.member_gender.value_counts()
```

[30]:

```
Male      130651
Female     40844
Other       3652
Name: member_gender, dtype: int64
```

1.8 Issue #5: user_type

Convert user_type from string to unordered category

[32]:

```
bikes.user_type.value_counts()
```

[32]:

```
Subscriber    163544
Customer       19868
Name: user_type, dtype: int64
```

[33]:

```
user_list = ['Subscriber', 'Customer']
user_cat_type = pd.api.types.CategoricalDtype(categories = user_list, ordered =␣
 ↪False)
bikes['user_type'] = bikes.user_type.astype(user_cat_type)
bikes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   duration_sec             183412 non-null  int64
 1   start_time               183412 non-null  datetime64[ns]
 2   end_time                 183412 non-null  datetime64[ns]
 3   start_station_id         183412 non-null  object
 4   start_station_name       183215 non-null  object
 5   start_station_latitude   183412 non-null  float64
 6   start_station_longitude  183412 non-null  float64
 7   end_station_id           183412 non-null  object
 8   end_station_name         183215 non-null  object
 9   end_station_latitude     183412 non-null  float64
 10  end_station_longitude    183412 non-null  float64
 11  bike_id                  183412 non-null  object
 12  user_type                183412 non-null  category
 13  member_birth_year        175147 non-null  float64
 14  member_gender            175147 non-null  category
 15  bike_share_for_all_trip  183412 non-null  object
dtypes: category(2), datetime64[ns](2), float64(5), int64(1), object(6)
memory usage: 19.9+ MB
```

1.9 Issue #6: member_birth_year

Convert member_birth_year from float to int64.

[35]:
```
bikes.member_birth_year.value_counts()
```

[35]:
```
1988.0    10236
1993.0     9325
1989.0     8972
1990.0     8658
1991.0     8498
          …
1928.0        1
1878.0        1
1930.0        1
1910.0        1
1927.0        1
Name: member_birth_year, Length: 75, dtype: int64
```

[38]:
```
bikes.member_birth_year.isna().value_counts()
```

[38]:
```
False    175147
True       8265
Name: member_birth_year, dtype: int64
```

[48]:
```
bikes['member_birth_year'] = bikes.member_birth_year.astype('Int64')
bikes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype
```

```
---  ------                   --------------   -----
 0   duration_sec             183412 non-null  int64
 1   start_time               183412 non-null  datetime64[ns]
 2   end_time                 183412 non-null  datetime64[ns]
 3   start_station_id         183412 non-null  object
 4   start_station_name       183215 non-null  object
 5   start_station_latitude   183412 non-null  float64
 6   start_station_longitude  183412 non-null  float64
 7   end_station_id           183412 non-null  object
 8   end_station_name         183215 non-null  object
 9   end_station_latitude     183412 non-null  float64
 10  end_station_longitude    183412 non-null  float64
 11  bike_id                  183412 non-null  object
 12  user_type                183412 non-null  category
 13  member_birth_year        175147 non-null  Int64
 14  member_gender            175147 non-null  category
 15  bike_share_for_all_trip  183412 non-null  object
dtypes: Int64(1), category(2), datetime64[ns](2), float64(4), int64(1),
object(6)
memory usage: 20.1+ MB
```

[49]:

```
bikes.member_birth_year.isna().value_counts()
```

[49]:

```
False    175147
True       8265
Name: member_birth_year, dtype: int64
```

1.10 Issue #7: data tidiness

Separate station information into a separate station DF: id, name, latitude, & longitude.

[101]:

```
# Seperate start and end stations
start_stations = bikes[['start_station_id','start_station_name',␣
 ↪'start_station_longitude', 'start_station_latitude']].copy()
end_stations = bikes[['end_station_id', 'end_station_name',␣
 ↪'end_station_longitude', 'end_station_latitude']].copy()
```

[102]:

```
# Rename columns
start_stations.rename(columns = {'start_station_id': 'id',
                                 'start_station_name': 'name',
                                 'start_station_longitude': 'longitude',
                                 'start_station_latitude': 'latitude'},
                      inplace = True)
end_stations.rename(columns = {'end_station_id': 'id',
                               'end_station_name': 'name',
                               'end_station_longitude': 'longitude',
                               'end_station_latitude': 'latitude'},
                    inplace = True)
```

[105]:
```
start_stations.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 4 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   id         183412 non-null  object
 1   name       183215 non-null  object
 2   longitude  183412 non-null  float64
 3   latitude   183412 non-null  float64
dtypes: float64(2), object(2)
memory usage: 5.6+ MB
```

[112]:
```
# Drop any duplicates
start_stations.drop_duplicates(subset = 'id', inplace = True)
end_stations.drop_duplicates(subset = 'id', inplace = True)
```

[108]:
```
# Merge two DFs together to form one for stations
stations = pd.merge(start_stations, end_stations, on = ['id', 'name',␣
 ↪'longitude', 'latitude'], how = 'outer')
```

[109]:
```
stations.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 331 entries, 0 to 330
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         331 non-null    object
 1   name       329 non-null    object
 2   longitude  331 non-null    float64
 3   latitude   331 non-null    float64
dtypes: float64(2), object(2)
memory usage: 12.9+ KB
```

[110]:
```
stations.drop_duplicates(subset = 'id', inplace = True)
```

[111]:
```
stations.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 330 entries, 0 to 329
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         330 non-null    object
 1   name       329 non-null    object 
 2   longitude  330 non-null    float64
 3   latitude   330 non-null    float64
dtypes: float64(2), object(2)
memory usage: 12.9+ KB
```

[114]:
```
bikes.drop(['start_station_name', 'start_station_longitude',␣
 ↪'start_station_latitude',
           'end_station_name', 'end_station_longitude', 'end_station_latitude'],
          axis = 1,
          inplace = True)
```

[115]:
```
bikes.head()
```

[115]:
```
   duration_sec               start_time                end_time  \
0         52185 2019-02-28 17:32:10.145 2019-03-01 08:01:55.975   
1         42521 2019-02-28 18:53:21.789 2019-03-01 06:42:03.056   
2         61854 2019-02-28 12:13:13.218 2019-03-01 05:24:08.146   
3         36490 2019-02-28 17:54:26.010 2019-03-01 04:02:36.842   
4          1585 2019-02-28 23:54:18.549 2019-03-01 00:20:44.074   

  start_station_id end_station_id bike_id   user_type  member_birth_year  \
0               21             13    4902    Customer               1984   
1               23             81    2535    Customer               <NA>   
2               86              3    5905    Customer               1972   
3              375             70    6638  Subscriber               1989   
4                7            222    4898  Subscriber               1974   

  member_gender bike_share_for_all_trip  
0          Male                      No  
1           NaN                      No  
2          Male                      No  
3         Other                      No  
4          Male                     Yes  
```

[121]:
```
bikes.to_csv('gobike-rides-master.csv')
stations.to_csv('gobike-stations-master.csv')
```