FinalProject_MS2_KanaparthiVenkata

March 3, 2022

Perform at least 5 data transformation and/or cleansing steps to your flat file data. For example: . Replace Headers . Format data into a more readable format . Identify outliers and bad data . Find duplicates . Fix casing or inconsistent values . Conduct Fuzzy Matching

[1]:
```
import numpy as np
import pandas as pd
```

[2]:
```
df=pd.read_csv("california.csv")
df.head(5)
```

[2]:

	10-barrel-brewing-co-san-diego	10 Barrel Brewing Co	large	1501 E St	Unnamed: 4	Unnamed: 5	San Diego	California	Unnamed: 8	92101-6618	http://10barrel.com	6195782311	2018-07-24	2018-08-23
0	101-north-brewing-company-petaluma	101 North Brewing Company	micro	1304 Scott St Ste D	NaN	NaN	Petaluma	California	NaN	94954-7100	http://www.101northbeer.com	7.077535e+09	2018-07-24	2018-08-11
1	14-cannons-brewing-company-westlake-village	14 Cannons Brewing Company	micro	31125 Via Colinas Ste 907	NaN	NaN	Westlake Village	California	NaN	91362-3974	http://14cannons.com	8.186996e+09	2018-07-24	2018-08-23
2	1850-brewing-company-mariposa	1850 Brewing Company	micro	NaN	NaN	NaN	Mariposa	California	NaN	95338	http://www.1850restaurant.com	NaN	2018-07-24	2018-08-23
3	2-tread-brewing-co-santa-rosa	2 Tread Brewing Co	brewpub	1018 Santa Rosa Plz	NaN	NaN	Santa Rosa	California	NaN	95401-6399	http://www.2treadbrewing.com	4.152331e+09	2018-07-24	2018-08-23
4	21st-amendment-brewery-cafe-san-francisco	21st Amendment Brewery Cafe	brewpub	563 2nd St	NaN	NaN	San Francisco	California	NaN	94107-1411	http://www.21st-amendment.com	4.153691e+09	2018-07-24	2018-08-23

```
   United States  -117.129593  32.714813  Unnamed: 17
0  United States          NaN        NaN          NaN
1  United States  -118.802397  34.153340          NaN
2  United States  -119.903659  37.570148          NaN
3  United States  -122.716773  38.438777          NaN
4  United States  -122.392577  37.782448          NaN
```

0.0.1 1. Replace the Headers

[3]:
```
names = []
with open('california_names.txt', 'r') as f:
    for line in f:
        f.readline()
        var = line.split(":")[0]
        names.append(var)
names
```

[3]:
```
['obdb_id',
 'brewery_name',
 'brewery_type',
 'street_name',
 'addressline_2',
 'addressline_3',
 'city_name',
 'state_name',
 'county_province',
 'postal_code',
 'website_url',
 'phone',
 'created_at',
 'updated_at',
 'country',
 'longitude',
 'latitude',
 'tags']
```

[4]:
```
df = pd.read_csv("california.csv", names = names)
df.head(5)
```

[4]:
```
                                    obdb_id                brewery_name  \
0              10-barrel-brewing-co-san-diego        10 Barrel Brewing Co
1          101-north-brewing-company-petaluma   101 North Brewing Company
2  14-cannons-brewing-company-westlake-village  14 Cannons Brewing Company
3               1850-brewing-company-mariposa        1850 Brewing Company
4               2-tread-brewing-co-santa-rosa          2 Tread Brewing Co

  brewery_type               street_name  addressline_2  addressline_3  \
```

```
0        large                  1501 E St            NaN            NaN
1        micro        1304 Scott St Ste D            NaN            NaN
2        micro  31125 Via Colinas Ste 907            NaN            NaN
3        micro                        NaN            NaN            NaN
4      brewpub        1018 Santa Rosa Plz            NaN            NaN

          city_name  state_name  county_province postal_code  \
0         San Diego  California              NaN  92101-6618
1          Petaluma  California              NaN  94954-7100
2  Westlake Village  California              NaN  91362-3974
3          Mariposa  California              NaN       95338
4       Santa Rosa  California              NaN  95401-6399

                     website_url         phone  created_at  updated_at  \
0           http://10barrel.com  6.195782e+09  2018-07-24  2018-08-23
1    http://www.101northbeer.com  7.077535e+09  2018-07-24  2018-08-11
2          http://14cannons.com  8.186996e+09  2018-07-24  2018-08-23
3  http://www.1850restaurant.com           NaN  2018-07-24  2018-08-23
4   http://www.2treadbrewing.com  4.152331e+09  2018-07-24  2018-08-23

         country   longitude   latitude  tags
0  United States -117.129593  32.714813   NaN
1  United States         NaN        NaN   NaN
2  United States -118.802397  34.153340   NaN
3  United States -119.903659  37.570148   NaN
4  United States -122.716773  38.438777   NaN
```

0.0.2 2. Create a data set with required columns

[5]:
```
df1=df[['brewery_name','brewery_type','street_name','addressline_2','addressline_3','city_name
df1.head(5)
```

[5]:
```
                brewery_name brewery_type               street_name  \
0       10 Barrel Brewing Co        large                 1501 E St
1   101 North Brewing Company        micro        1304 Scott St Ste D
2  14 Cannons Brewing Company        micro  31125 Via Colinas Ste 907
3       1850 Brewing Company        micro                        NaN
4         2 Tread Brewing Co      brewpub        1018 Santa Rosa Plz

   addressline_2  addressline_3         city_name  state_name        country  \
0            NaN            NaN         San Diego  California  United States
1            NaN            NaN          Petaluma  California  United States
2            NaN            NaN  Westlake Village  California  United States
3            NaN            NaN          Mariposa  California  United States
4            NaN            NaN        Santa Rosa  California  United States

  postal_code
```

```
0  92101-6618
1  94954-7100
2  91362-3974
3       95338
4  95401-6399
```

0.0.3 3. Find duplicates

[6]:
```
print("Postal Code is duplictaed - {}".format(any(df.postal_code.duplicated())))
```

```
Postal Code is duplictaed - True
```

0.0.4 4. Find Null values

[7]:
```
print("The column street name contains NaN - %r " % df.street_name.isnull().
 ↪values.any())
```

```
The column street name contains NaN - True
```

0.0.5 5. Identify outliers and bad data

[8]:
```
size_prev = df.shape
print(df['longitude'])
df = df[np.isfinite(df['longitude'])]
size_after = df.shape
print("The size of previous data was - {prev[0]} rows and the size of the new␣
 ↪one is - {after[0]} rows".
      format(prev=size_prev, after=size_after))
```

```
0     -117.129593
1             NaN
2     -118.802397
3     -119.903659
4     -122.716773
          …
891   -122.688361
892   -121.541958
893   -118.316351
894   -119.690657
895   -118.352467
Name: longitude, Length: 896, dtype: float64
The size of previous data was - 896 rows and the size of the new one is - 630
rows
```