# shudhi_transform

April 28, 2018

```
In [1]: import pandas as pd
        import numpy as np
        import sklearn
        import matplotlib.pyplot as plt
        import seaborn as sns
        import warnings
        from sklearn.preprocessing import StandardScaler, MaxAbsScaler, MinMaxScaler, RobustSca
        import time
        from scipy.stats import zscore
        from datetime import datetime
        from IPython.display import display, HTML
```

## 1 Import Shudhi Modules

```
In [2]: from shudhi_describe import shudhi_describe
        from shudhi_transform import shudhi_transform
```

### 1.1 Import Datasets: we have a small and a big dataset

```
In [3]: df_four = pd.read_json('foursquare_test.json', orient='records')
```

```
In [4]: df_four.shape
```

```
Out[4]: (400, 11)
```

```
In [5]: df_complaint = pd.read_csv('Top_5_complaints.csv')
```

```
In [6]: df_complaint.shape
```

```
Out[6]: (376062, 36)
```

### 1.2 Run Shudhi Describe: Baseline

```
In [7]: shudhi_describe(df_four, plot=False)
```

```
                                   SUMMARY STATISTICS
```

```
           Feature                Feature Type  count  # Unique  # Missing  \
0          country               String/Object    400         1          0
1               id               String/Object    400       400          0
2         latitude                  Real Value    400       399          0
3         locality               String/Object    400         5          0
4        longitude                  Real Value    400       399          0
5             name               String/Object    400       386          0
6            phone               String/Object    400       199        181
7      postal_code  Real Value saved as string    400        53          0
8           region               String/Object    400         1          0
9   street_address               String/Object    400       310          0
10         website               String/Object    400        69          0

   # Outliers   mean median      min      max
0
1
2           1  40.75  40.74  40.6605  42.3531
3
4           1 -73.97 -73.99 -74.0159 -71.0541
5
6
7
8
9
10
```

```
-----------------------------------------------------------------------------------------------
```

```
In [8]: shudhi_describe(df_complaint, plot=False)
```

```
                                        SUMMARY STATISTICS


                            Feature   Feature Type   count  # Unique  \
0                        Unique Key        Integer  376062    376062
1                      Created Date  String/Object  376062    334662
2                       Closed Date  String/Object  376062    273775
3                            Agency  String/Object  376062         7
4                       Agency Name  String/Object  376062        39
5                    Complaint Type  String/Object  376062         5
6                        Descriptor  String/Object  376062        62
7                     Location Type  String/Object  376062        20
8                       Incident Zip     Real Value  376062       209
9                  Incident Address  String/Object  376062    151287
10                      Street Name  String/Object  376062      8597
11                    Cross Street 1  String/Object  376062      9418
```

```
12                 Cross Street 2  String/Object  376062      9406
13          Intersection Street 1  String/Object  376062      4951
14          Intersection Street 2  String/Object  376062      4860
15                   Address Type  String/Object  376062         5
16                           City  String/Object  376062        90
17                       Landmark  String/Object  376062        42
18                         Status  String/Object  376062         5
19                       Due Date  String/Object  376062    217043
20         Resolution Description  String/Object  376062        98
21  Resolution Action Updated Date  String/Object  376062    274121
22                Community Board  String/Object  376062        76
23                        Borough  String/Object  376062         6
24      X Coordinate (State Plane)     Real Value  376062     69741
25      Y Coordinate (State Plane)     Real Value  376062     85139
26             Park Facility Name  String/Object  376062         1
27                   Park Borough  String/Object  376062         6
28                   Vehicle Type     Real Value  376062         0
29                Garage Lot Name     Real Value  376062         0
30             Ferry Terminal Name     Real Value  376062         0
31                       Latitude     Real Value  376062    164759
32                      Longitude     Real Value  376062    164663
33                       Location  String/Object  376062    165828
34                           year        Integer  376062         8
35                          month        Integer  376062        12

    # Missing # Outliers         mean       median          min          max
0           0          0  3.34056e+07  3.37504e+07  1.95111e+07  3.89483e+07
1           0
2        2239
3           0
4           0
5           0
6           0
7       31986
8        2370          0      10795.9        11203            0        11697
9       34479
10      34479
11     122944
12     123443
13     341669
14     341916
15       1292
16       2356
17     375970
18          0
19     139610
20        327
21       1152
```

```
22          0
23          0
24       2996          0  1.00514e+06   1.0039e+06       913495  1.06717e+06
25       2996          0       207940       204824       121212       271876
26          0
27          0
28     376062          0
29     376062          0
30     376062          0
31       2996          0        40.74        40.73      40.4991      40.9129
32       2996          0       -73.92       -73.93     -74.2544     -73.7008
33       2996
34          0          0      2015.85         2016         2011         2018
35          0          0         6.47            6            1           12


-------------------------------------------------------------------------------------------------
```

## 2 Run Shudhi Transform

```
In [9]: # df_four has no missing in continuous variables, hence, only scaling with std scaler
        # Scale for latitude and longitude; One hot for locality

        df_new= shudhi_transform(df_train=df_four, cols=['latitude', 'longitude', 'locality'],
                                 scale_strategy='std', one_hot=True )
```

```
Warning: Entered inconsistent column types. Only Continuous features will be scaled.
Warning: Entered inconsistent column types. Only Categorical features will be one hot encoded
```

```
In [10]: # The warning above is because we have both categorical and continuous features in co
```

```
In [11]: # Fill in missing values with mean and then scale with "max_abs" scaler

         df_c_new= shudhi_transform(df_train=df_complaint, cols=['Latitude', 'Longitude', 'Inci
                                    missing_strategy='mean',  scale_strategy='max_abs')
```

```
Warning: If a column has >10% missing values, it will not be acted upon unless "override=True"
['Latitude', 'Longitude', 'Incident Zip']
```

### 2.1 Check with describe if this has worked: Ofcourse it has!

```
In [13]: shudhi_describe(df_new, plot=False)
```

```
                                       SUMMARY STATISTICS
```

```
                          Feature                Feature Type  count  \
0                         country               String/Object    400
1                              id               String/Object    400
2                        latitude                  Real Value    400
3                       longitude                  Real Value    400
4                            name               String/Object    400
5                           phone               String/Object    400
6                     postal_code  Real Value saved as string    400
7                          region               String/Object    400
8                  street_address               String/Object    400
9                         website               String/Object    400
10  locality_822 Lexington Avenue                     Integer    400
11              locality_Brooklyn                     Integer    400
12              locality_Elmhurst                     Integer    400
13              locality_New York                     Integer    400
14             locality_Sunnyside                     Integer    400

    # Unique  # Missing # Outliers  mean median       min      max
0          1          0
1        400          0
2        399          0          0    -0   -0.1  -1.07619  18.5324
3        399          0          0     0  -0.09 -0.288801  19.6258
4        386          0
5        199        181
6         53          0
7          1          0
8        310          0
9         69          0
10         2          0          0     0      0         0        1
11         2          0          0     0      0         0        1
12         2          0          0     0      0         0        1
13         2          0          0  0.99      1         0        1
14         2          0          0     0      0         0        1


-----------------------------------------------------------------------------------------------
```

```
In [16]: shudhi_describe(df_c_new, cols= ['Latitude', 'Longitude', 'Incident Zip'], plot=False
```

```
                                       SUMMARY STATISTICS


        Feature Feature Type   count  # Unique  # Missing  # Outliers  mean  \
0      Latitude   Real Value  376062    164760          0           0  1.00
1     Longitude   Real Value  376062    164664          0           0 -1.00
2  Incident Zip   Real Value  376062       210          0           0  0.92
```

```
   median       min       max
0    1.00  0.989887  1.000000
1   -1.00 -1.000000 -0.992544
2    0.96  0.000000  1.000000


-------------------------------------------------------------------------------------------
```