NYC Department of Mental Health and Hygiene Inspections

The following merged datasets contain information on center-based child care/summer camp sites and inspections related to the sites.

In [89]:

```
%matplotlib inline
low_memory = False
import numpy as np
import pandas as pd #allows working with labeled data easy and intuitive
import matplotlib.pyplot as plt #2D ploying library
import seaborn as sns; sns.set() #provides high-level interface for crafting informa
sns.set(style='whitegrid')

dohmhChildFacilities = pd.read_csv('dohmh-childcare-center-inspections.csv', encodi
shortData = pd.read_csv('Childcare_Centers.csv')
```

Below you can find the four different exploratory compenents used to display the data in its simplest form before cleaning:
.describe( )
.shape
.dtypes
display( )

DOHMH Child Inspection Briefing

In [90]:

```
dohmhChildFacilities.shape
```

Out[90]:

```
(60243, 35)
```

In [91]:

```
dohmhChildFacilities.dtypes
```

Out[91]:

```
Center Name                                     object
Legal Name                                      object
Building                                        object
Street                                          object
Borough                                         object
ZipCode                                          int64
Phone                                           object
Permit Number                                  float64
Permit Expiration                               object
Status                                          object
Age Range                                       object
Maximum Capacity                                 int64
Day Care ID                                     object
Facility Type                                   object
Child Care Type                                 object
Building Identification Number                 float64
URL                                             object
Date Permitted                                  object
Actual                                          object
Violation Rate Percent                         float64
Average Violation Rate Percent                 float64
Total Educational Workers                        int64
Average Total Educational Workers              float64
Staff Turnover Rate                            float64
Average Staff Turn Over Rate                   float64
Public Health Hazard Violation Rate            float64
Average Public Health Hazard Violation Rate    float64
Critical Violation Rate                        float64
Average Critical Violation Rate                float64
Inspection Date                                 object
Regulation Summary                              object
Violation Category                              object
Health Code Sub Section                         object
Violation Status                                object
Inspection Summary Result                       object
dtype: object
```

In [92]:

```
dohmhChildFacilities.describe()
```

Out[92]:

	ZipCode	Permit Number	Maximum Capacity	Building Identification Number	Violation Rate Percent	Average Violation Rate Percent	I
count	60243.000000	53700.000000	60243.000000	6.010700e+04	60120.000000	59896.000000	60
mean	10818.706539	31003.218771	73.316667	2.839044e+06	51.093268	40.034620	
std	537.909151	35355.027282	86.886552	1.225077e+06	26.031190	2.691228	
min	10001.000000	375.000000	0.000000	0.000000e+00	0.000000	34.886800	
25%	10451.000000	6599.000000	27.000000	2.011440e+06	33.333300	39.812300	
50%	11204.000000	8101.000000	55.000000	3.072665e+06	50.000000	39.812300	
75%	11235.000000	53140.000000	98.000000	4.031725e+06	66.666700	39.812300	
max	12225.000000	104644.000000	3256.000000	5.155012e+06	100.000000	46.828200	

In [93]:

```
dohmhChildFacilities
```

Out[93]:

	Center Name	Legal Name	Building	Street	Borough	ZipCode	Phone	Permit Number	Ex
0	ALL SEASONS ABC DAY CARE, LLC	ALL SEASONS DAY CARE, LLC	190	E 162ND ST	BRONX	10451	914-490-5231	104386.0	2 26T
1	ALL SEASONS ABC DAY CARE, LLC	ALL SEASONS DAY CARE, LLC	190	E 162ND ST	BRONX	10451	914-490-5231	104386.0	2 26T
2	PRESCHOOL OF AMERICA (USA), INC.	PRESCHOOL OF AMERICA (USA), INC.	44-46	MARKET STREET	MANHATTAN	10002	212-346-9868	8108.0	2 22T

In [94]:

```
shortData.shape
```

Out[94]:

```
(56985, 10)
```

In [95]:

```
shortData.dtypes
```

Out[95]:

```
Center Name             object
Borough                 object
ZipCode                  int64
Phone                   object
Maximum Capacity         int64
Child Care Type         object
URL                     object
Inspection Date         object
Regulation Summary      object
Violation Category      object
dtype: object
```

In [96]:

```
shortData.describe()
```

Out[96]:

	ZipCode	Maximum Capacity
count	56985.000000	56985.000000
mean	10815.237168	66.715662
std	537.126750	57.104154
min	10001.000000	0.000000
25%	10451.000000	26.000000
50%	11203.000000	54.000000
75%	11235.000000	93.000000
max	11697.000000	643.000000

In [97]:

```
shortData
```

Out[97]:

	Center Name	Borough	ZipCode	Phone	Maximum Capacity	Child Care Type	UF
0	BEDROCK PRESCHOOL (P/S)	BRONX	10463	718-884-0020	91	Child Care - Pre School	Na
1	GOOD SAMARITAN FULTON DAY CARE CENTER-PRESCHOOL	BROOKLYN	11221	718-443-6463	30	Child Care - Pre School	Na
2	GATEWAY CATHEDRAL, INC.	STATEN ISLAND	10309	718-966-8695	30	Child Care - Infants/Toddlers	GATEWAYACADEMYNY.OF
	WELLSPRING			718-		School Based	

Data Cleaning

In [98]:

```
#data merge, removal of duplicate/unnecessary columns
cols_to_use = dohmhChildFacilities.columns.difference(shortData.columns)
dfNew = pd.merge(shortData, dohmhChildFacilities[cols_to_use], left_index=True, righ
dfNew = dfNew.drop(columns = ['URL'])
dfNew.shape
```

Out[98]:

```
(60243, 34)
```

In [99]:

```
newViolationCat= dfNew.groupby(['Violation Category'])
newViolationCat.size().sort_values()
```

Out[99]:

```
Violation Category
PUBLIC HEALTH HAZARD     7007
CRITICAL                15152
GENERAL                 21288
dtype: int64
```

In [100]:

```
dfNew[dfNew['Violation Category']==""] = np.NaN
dfNew['Violation Category'] = dfNew['Violation Category'].fillna('OTHER')
newViolationCat= dfNew.groupby(['Violation Category'])
newViolationCat.size().sort_values()
```

Out[100]:

```
Violation Category
PUBLIC HEALTH HAZARD      7007
CRITICAL                 15152
OTHER                    16796
GENERAL                  21288
dtype: int64
```

In [102]:

```
dfNew = dfNew.dropna(how='any',axis=0)
dfNew.shape
```

Out[102]:

```
(32890, 34)
```

#NAN values were removed for now. Given that under 60% of the data (25040/60243) had missing values, predictions of those missing values can be made. Linear regression and replacemnt of NAN values with respective averaged will be used in the near future to determine the values of the following categories:

- Violation Rate Percent
- Average Violation Rate Percent
- Average Staff Turn Over Rate
- Staff Turnover Rate

In [103]:

```
#display of new cleaned columns
dfNew.sample(5)
```

Out[103]:

	Center Name	Borough	ZipCode	Phone	Maximum Capacity	Child Care Type	Inspection Date	Regul Sum
49473	KIDS CIRCLE DAY CARE INC	QUEENS	11432.0	718-380-1280	30.0	Child Care - Pre School	04/23/2018	There n viol obs at t
45229	PRESCHOOL OF AMERICA (USA) INC	QUEENS	11366.0	718-380-0032	44.0	Child Care - Infants/Toddlers	07/18/2017	tea hav rec trai
34113	MAGIC MOMENTS ACADEMY PRESCHOOL & CHILD CARE	BROOKLYN	11238.0	917-806-7777	24.0	Child Care - Pre School	07/26/2018	There n viol obs at t
29574	SHIRA ASSOCIATION, INC.	BROOKLYN	11219.0	718-435-7700	58.0	Child Care - Pre School	12/07/2017	There n viol obs at t
40110	NORTHEAST BRONX DAY CARE CENTER, INC.	BRONX	10466.0	718-547-0501	199.0	Child Care - Pre School	12/09/2016	tea hav rec trai

5 rows × 34 columns

In [104]:

```
dfNew.dtypes
```

Out[104]:

```
Center Name                                     object
Borough                                         object
ZipCode                                        float64
Phone                                           object
Maximum Capacity                               float64
Child Care Type                                 object
Inspection Date                                 object
Regulation Summary                              object
Violation Category                              object
Actual                                          object
Age Range                                       object
Average Critical Violation Rate                float64
Average Public Health Hazard Violation Rate    float64
Average Staff Turn Over Rate                   float64
Average Total Educational Workers              float64
Average Violation Rate Percent                 float64
Building                                        object
Building Identification Number                 float64
Critical Violation Rate                        float64
Date Permitted                                  object
Day Care ID                                     object
Facility Type                                   object
Health Code Sub Section                         object
Inspection Summary Result                       object
Legal Name                                      object
Permit Expiration                               object
Permit Number                                  float64
Public Health Hazard Violation Rate            float64
Staff Turnover Rate                            float64
Status                                          object
Street                                          object
Total Educational Workers                      float64
Violation Rate Percent                         float64
Violation Status                                object
dtype: object
```