Data Wrangling Process of Capstone Project

In this project, I work on the DOHMH Childcare Center Inspections data. Before starting to data wrangling, I explored my data. To be able to do that I applied shape, columns, dtypes, info attributes, and describe, isnull, duplicated methods of pandas to the data set.

After getting familiar with data, I find out five types of problems. These problems are:

1 Unnecessary information columns
2 Duplicated rows
3 Missing data
4 Incompatible types of columns
5 White spaces in column names

There is no outlier issue in this data set.

Unnecessary information columns: At this stage, I filtered out unnecessary columns and dropped them.

```
df=df(columns=['column_name'])
```

Duplicated rows: Duplicates are data points which are repeated rows in the dataset. I removed the duplicated rows by use of 'drop_duplicates()' method.

Missing data: I handled missing data in two steps. First of all, I kept only the rows with at least 23 non-NA values.

```
df=df.dropna(thresh=23)
```

Then, I used 'fillna' method to replace the missing values with zero.

```
df=df.fillna(0)
```

Incompatible types of columns: There are two columns that variables were not compatible with the column type. The first one is 'ZipCode' column. Its' type was used to be 'float'. I changed 'ZipCode' columns' type to 'integer'.The second column is the 'Maximum Capacity' column. I also changed its' type as 'integer'.

```
df['Maximum Capacity'].astype('int32')
df['ZipCode'].astype('int32')
```

White spaces in column names: There were some spaces between the word of column names. I removed these white spaces.

```
df.columns = df.columns.str.replace(' ', '')
```