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 type of columns
5 White spaces in column names
6 Bad Data Problem
7 A New Column

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: Incompatible types of columns: The variables of 'ZipCode' column were not compatible with the column type. Therefore, I added an argument to the importing the data code to change it from 'float' to 'string.'

```
data=pd.read_csv('DOHMH_Childcare_Center_Inspections.csv',
                 dtype={'ZipCode': 'str'})
```

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(' ', '')
```

Bad Data Problem: There was a bad data problem in ‘Facility Type’ column. The Camp variable used to exist as ‘Camp’ and ‘CAMP’. I solved the issue with the following code:

```
data_c.FacilityType=data_c.FacilityType.str.lower()
```

A New Column: I created 'Operated_time' column through subtracting two columns.

```
data_c['Operated_time']=data_c['PermitExpiration']-data_c['DatePermitted']
```

Before the subtraction, I changed the type of these columns from object to datetime.

```
data_c['PermitExpiration']=data_c['PermitExpiration'].apply(pd.to_datetime)
```

I also changed the 'Inspection Date' columns' type object to datetime.