I studied data cleaning. Data cleaning to correct errors and inconsistencies in data is a cumbersome but essential process. I learned anew that there are also types of missing value. There are three types of missing value, and the concepts and differences of MCAR (Missing Complete at Random), MAT (Missing at Random), and MNAR (Missing at Random) have been newly learned. Methods of finding missing value include making a heatmap, using the isnull function to make a list of columns with missing value, and visualizing it with histogram. Methods of dealing with missing value include removing the value through the dropna function and filling the value through the fillna function. Among the methods of filling the value, it was interesting that it could be filled with KNN. There is a method of visualizing the box plot, but if the data is large, you can find the outlier using the normal distribution of zscore. Finally, we learned how to eliminate repetitive data and correct capitalization, formatting, and spelling errors so that inconsistent data does not occur.

The following strategies were established to carry out the first exercise.

state: remove value greater than 5, fill missing value with mean

floor: remove value greater than 40, fill missing value with mean

max_floor: remove value greater than 50, fill missing value with mean

build_year: fill missing value with the most frequent value

life_sq: remove value greater than 100, fill missing value with mean

However, after hearing the presentation and the solution explanation, there were many things I didn't think of.

Using the z-score, it was found that methods such as discarding values greater than the threshold, checking and correcting contradictions between floor and max_floor, changing the 0, 1, 3, and 20 values of build_year to 2000, 2001, 2003, 1920, and filling the missing value using Knn are necessary.

I didn't feel well during the time to learn aggregation & pivot table after lunch, so I didn't take classes. I'm going to study and understand by myself.