END 305 E Assignment 3

Due Date: 06/02/2020

Task:

In this assignment, you will use K-Means algorithm to conduct a clustering analysis on a real estate data collected from the City of Daegu (a city in South Korea). The data includes sales price of apartments together with detailed characteristics of the apartment itself. Please download "Daegu_Real_Estate_data.csv" file from Ninova as the data file.

Steps:

1. Conduct necessary preprocessing of data:
    a. Encoding of categorical data (use pandas.get_dummies)
    b. Scale the data (MinMaxScaler)
2. Try **k** values from 2 to 30 and select the best **k** value. (Check both inertia and silhouette score.). Explain why you selected that particular k in your code/notebook
3. Rerun clustering algorithm with best k, print out the cluster centers.
4. Identify 2 clusters:
    a. The cluster consisting of the houses with **highest sales value**s
    b. The cluster consisting of the houses with **lowest sales value**s
5. Explain the characteristics defining these clusters by looking at the center values. Which features/columns are definitive and different from rest of the clusters.

Deliverables:

Deliver a single Jupyter notebook (preferred) or python code file that includes both your code and comments.