Exercise 6: Clusters and Anomalies

Setup of Working Environment

1. Create a folder on your Desktop and name it EE0005_[LabGroup], where [LabGroup] is the name of your Group.
2. Download the .ipynb files and data files posted corresponding to this exercise and store in the aforesaid folder.
3. Open Jupyter Notebook and navigate to the aforesaid folder on Desktop.
4. Open and explore the .ipynb files (notebooks) that you downloaded, and go through "Preparation", as follows.
5. The walk-through videos posted on NTU Learn may help you with this "Preparation" too.
6. Create a new Jupyter Notebook, name it Exercise6_solution.ipynb, and save it in the same folder on the Desktop

Preparation for the Exercises

M5 ClusteringPatterns.ipynb	Check how to perform basic Clustering on the Pokemon data (pokemonData.csv)
M5 DetectingAnomalies.ipynb	Check how to perform basic Anomaly Detection on the Pokemon dataset

Objective

Let us assume that the houses in our dataset vary by their Living Area (GrLivArea) and Garages (GarageArea) in general. In this exercise, we will try to find patterns in the data by clustering the house as per their Living Area and Garage Area. We will also try to identify major anomalies in the dataset, once again, in terms of their Living Area and Garage Area.

Problems

Problem 1 : Clustering using GrLivArea and GarageArea

Download the Kaggle dataset "train.csv" from NTU Learn, posted corresponding to this Example Class.
Import the complete dataset "train.csv" in Jupyter, as `houseData = pd.read_csv('train.csv')`

a) Extract the two variables in consideration from the dataset

```
X = pd.DataFrame(houseData[['GrLivArea','GarageArea']])
```

b) Visualize the 2D distribution of the two variables extracted above, using a standard scatter plot.

c) Import k-Means Clustering model from Scikit-Learn : `from sklearn.cluster import KMeans`

d) Guess the number of clusters from the 2D scatterplot, and perform k-Means clustering with that.

e) Print the cluster centers, view their countplot, and visualize the clusters on the 2D scatterplot.

Problem 2 : Anomaly Detection with the same Variables

a) Import Anomaly model from Scikit-Learn : `from sklearn.neighbors import LocalOutlierFactor`

b) Guess the parameters from the 2D scatterplot, and perform Anomaly Detection with those parameters.

c) View their countplot of Anomalies vs Normal Data, and visualize the anomalies on the 2D scatterplot.

Bonus : Are you happy with the Clusters and the Anomalies? If not, what would you do to make it better?