Exercise 3 : Exploratory Analysis

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 Exercise3_solution.ipynb, and save it in the same folder on the Desktop

Preparation for the Exercises

M2 ExploratoryAnalysis.ipynb	Check how to import the Pokemon data and perform Exploratory Analysis You will need the CSV data file pokemonData.csv to use this code

Objective

Our final target is to predict "SalePrice" of a house, based on the other variables given in the Housing Data from Kaggle.

In this Example Class, our main goal is to analyze the most relevant numeric and categorical variables in this dataset, which may affect the sale price of a house, and hence, will be most relevant in predicting "SalePrice". We will extract some variables, perform basic statistical exploration and visualization, and try to gauge their relation with "SalePrice".

Problems

Problem 1 : Analysis of Numeric Variables

Download the Kaggle dataset "train.csv" from NTU Learn, posted corresponding to this Example Class.

Extract the following Numeric variables from the dataset, and store as a new Pandas DataFrame.

```
houseNumData = pd.DataFrame(houseData[['LotArea', 'GrLivArea', 'TotalBsmtSF', 'GarageArea', 'SalePrice']])
```

a) Check the individual statistical description and visualize the statistical distributions of each of these variables.
b) Discuss with your friends if the distributions look like "Normal Distribution". Which one has maximum outliers?
c) Check the relationship amongst the variables using mutual correlation and the correlation heatmap. Discuss with your friends and determine which of the variables has the strongest correlation with "SalePrice". Is it useful?
d) Check the relationship amongst the variables using mutual jointplots and an overall pairplot. Discuss with your friends and determine which of the variables has the strongest linear relation with "SalePrice". Is it useful?

Problem 2 : Statistical Summary

Extract the following Categorical variables from the dataset, and store as a new Pandas DataFrame.

```
houseCatData = pd.DataFrame(houseData[['MSSubClass', 'Neighborhood', 'BldgType', 'OverallQual']])
```

a) Convert each of these variables into "category" data type (note that some are "int64", and some are "object").
b) Check the individual statistical description and visualize the distributions (catplot) of each of these variables.
c) Check the relationship amongst the variables using bi-variate heatmap of counts. Discuss with your friends and see if you can figure any intuitive relationship of "OverallQual" with the other three variables? Is this useful?
d) Draw boxplots of "SalePrice" against each of these categorical variables. Discuss with your friends and find out if you see any pattern in these boxplots. Which of these variables has the strongest relationship with "SalePrice"?

Bonus : Can you perform a similar analysis on other variables in the dataset, and gain more insight about the data?