Exercise 3 : Exploratory Analysis

Workflow

1. Download the .ipynb files and data files posted with this exercise and store them all in a folder on your Desktop.
2. Open Jupyter Notebook (already installed on the Lab computers) and navigate to the aforesaid folder on Desktop.
3. Open and explore the .ipynb files (notebooks) that you downloaded, and go through "Preparation", as follows.
4. The walk-through videos posted on NTU Learn (under Course Content) may help you with this "Preparation" too.
5. Create a new Jupyter Notebook, name it Exercise3_solution.ipynb, and save it in the same folder on the Desktop.
6. Solve the "Problems" posted below by writing code, and corresponding comments, in Exercise3_solution.ipynb.

Try to solve the problems on your own. Take help and hints from the "Preparation" codes and the walk-through videos. **If you are still stuck, talk to your friends in the Lab to get help/hints.** If that fails too, approach your Lab Instructor.

Note : Don't forget to import the Essential Python Libraries required for solving the Exercise. Write code in the usual "Code" cells, and notes/comments in "Markdown" cells of the Notebook. Check the preparation notebooks for guidance.

Preparation

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

In this Lab Exercise, our main goal is to analyze the most relevant numeric and categorical variables in the given dataset, which may affect the sale price of a house, and hence, will be most relevant in predicting "SalePrice". You will extract some variables, perform statistical exploration and visualization, and try to analyze their relationship with "SalePrice". In addition, you will also try to answer a few specific questions on the dataset using basics of Exploratory Data Analysis.

Disclaimer: There may be several ways to solve these problems and there is no single correct answer. Try to explore on your own, talk to your friends and the Lab Instructor, and make sure you are happy with your own justifications. You will get marks for your solutions as long as your justifications make sense, and you can explain those clearly.

Marks distribution

4 points for Problem 1	2 points for (a) + 1 point for (b) + 1 point for (c)	
3 points for Problem 2	1 point for (a) + 2 points for (b)	
3 points for Problem 3	3 points for (a) OR 3 points for (b)	Choose any one between (a) and (b)

Know your plots!

Your plots in this exercise will be from the common seaborn plots. Seek help from Lab Instructors if you face problems. You will also find more data visualization suggestions with standard statistical plots at https://www.data-to-viz.com

Problems

Problem 1 : Analysis of Numeric Variables

In this problem, your job is to analyze the following numeric variables in the dataset and their relationship with SalePrice.

```
['LotArea', 'GrLivArea', 'TotalBsmtSF', 'GarageArea']
```

a) Which of these variables has the maximum number of outliers as per box-plot? How many outliers does it have?
b) Which of these variables is the most skewed from a regular normal distribution? Is the skew positive or negative?
c) Choose the top two variables that you think will help us the most in predicting 'SalePrice' of houses in this data.

Hints and Pointers

- In case of box-plot, outliers are the datapoints outside the whiskers, which are at Q1 – 1.5 IQR and Q3 + 1.5 IQR.
- Pandas has a bunch of statistical measures built in as methods/functions; .median() and .mean(), for example.
- Predicting one numeric variable with another numeric variable is easiest when they have a strong relationship.

Problem 2 : Analysis of Categorical Variables

In this problem, your job is to analyze the following categorical variables in the data and their relationship with SalePrice.

```
['MSSubClass', 'Neighborhood', 'BldgType', 'OverallQual']
```

a) Which of these variables has the highest number of levels? Which of the levels has the highest number of houses?
b) Choose the top two variables that you think will help us the most in predicting 'SalePrice' of houses in this data.

Hints and Pointers

- Levels for a categorical variable means the number of unique values. For example, gender has 3 levels, F, M, O.
- Each level of a categorical variable may contain a number of datapoints. For example, 14 M found in a dataset.
- When you want to find relationship between a numeric variable and a categorical one, you can't do Correlation.
- Check box-plot function in seaborn carefully – there is a way to plot a numeric vs a categorical in the two axes.
- Think: If there was a relationship between Salary and Gender, what would box-plot of Salary be across F, M, O?

Problem 3 : Interesting Questions for EDA

Choose **any ONE** of the following questions to answer. You may always try out the other one at home if you want. ☺

a) Does the SalePrice of a house get affected by whether it has a Garage or not? Justify your answer using EDA.
b) Does the SalePrice of a house get affected by how recently it got Remodeled? Justify your answer using EDA.

Hints and Pointers

- Check the data_description file very carefully and pick the variables you want to work with for these problems.
- In some cases, it is alright to create new variables out of the original ones, especially if they help you analyze.
- Do keep an eye out for missing values in the variables that you tackle and see what you can make out of those.

Hints are not meant to tell you exactly what to do for the problems; use these as pointers to search online. Take a close look at **Pandas DataFrame documentation** and read the **data_description** file carefully to solve most of these problems.