Exercise 3 : Exploratory Analysis

Workflow

1. Create a folder on your Desktop and name it Cx1015_[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 (already installed on the Lab computer) 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 (under Course Content) 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
7. 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 the 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

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

Download the dataset **train.csv** and the associated text file **data_description.txt** posted with this Exercise.

Problem 1 : Analysis of Numeric Variables

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) Comment if the distributions look like "Normal Distribution", or different. Use the .skew() method to find the "skewness" of each of the five distributions. Which of the variables has the maximum number of outliers?
c) Check the relationship amongst the variables using mutual correlation and the correlation heatmap. Comment which of the variables has the strongest correlation with "SalePrice". Is this useful in predicting "SalePrice"?
d) Check the relationship amongst the variables using mutual jointplots and an overall pairplot. Comment which of the variables has the strongest linear relation with "SalePrice". Is this useful in predicting "SalePrice"?

Problem 2 : Analysis of Categorical Variables

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) One may check the relation amongst two categorical variables through the bi-variate joint heatmap of counts. Use groupby() command to generate joint heatmap of counts for "OverallQual" against the other three variables. Comment if this is useful in identifying the relation between "OverallQual" with the other variables.
d) Draw boxplots of "SalePrice" against each of these categorical variables. Notice the patterns in these boxplots. Comment on which of these variables has the most influence in predicting "SalePrice".

You may read more about the functions you will need to use in this exercise in the following references.

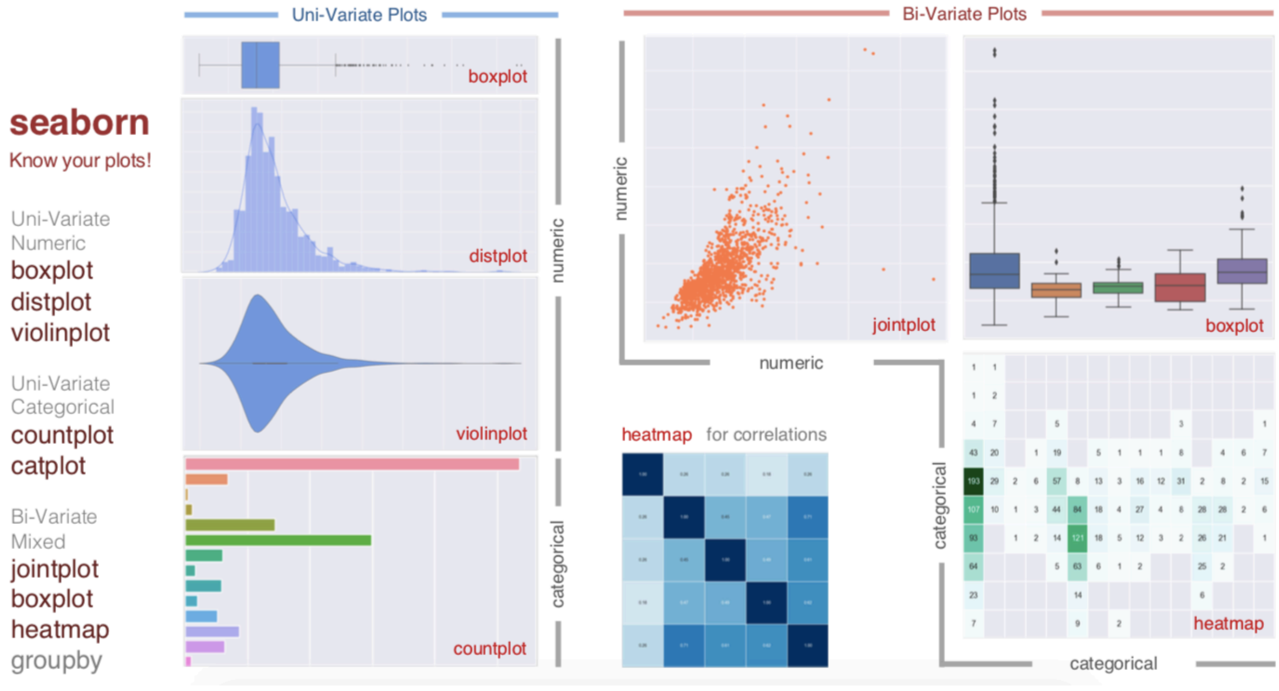
Part (b) : catplot : https://seaborn.pydata.org/generated/seaborn.catplot.html
Part (c) : groupby : https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.groupby.html
Part (d) : boxplot : https://seaborn.pydata.org/generated/seaborn.boxplot.html

Bonus Problem

Perform a similar analysis on every other variable in the dataset, against "SalePrice". This will let you gain more insight about the data, and find out which variables are actually useful in predicting "SalePrice". Warning : It is a painful process!

Seaborn : Know your plots!

Your plots in this exercise will be from the following common plots. Seek help from the Instructor if you face problems.


seaborn
Know your plots!
Uni-Variate Numeric
boxplot
distplot
violinplot
Uni-Variate Categorical
countplot
catplot
Bi-Variate Mixed
jointplot
boxplot
heatmap
groupby
Uni-Variate Plots
boxplot
distplot
violinplot
countplot
numeric
categorical
Bi-Variate Plots
numeric
numeric
jointplot
boxplot
heatmap for correlations
categorical
categorical
heatmap