Exercise 2 : Basic Statistics

Setup

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 Exercise2_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 Exercise2_solution.ipynb

Note : Don't forget to import the Essential Libraries required for solving the Exercise (check the preparation notebooks)

Preparation

M2 BasicStatistics.ipynb	Check how to import the Pokemon data and perform basic Statistics You will need the CSV data file pokemonData.csv to use this code
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

Problems

Problem 1 : Data Preparation

Download the dataset from the following Kaggle Competition (login required) – Go to "Data", and "Download All".

House Prices Competition : https://www.kaggle.com/c/house-prices-advanced-regression-techniques

a) Import the "train.csv" data from the downloaded data folder (has four files) in Jupyter Notebook.
b) What are the data types ("dtypes") – int64/float64/object – of the variables (columns) in the dataset?
c) Extract only the variables (columns) of type Integer (int64), and store as a new Pandas DataFrame.
d) Read "data_description.txt" (from the Kaggle data folder) to identify the actual Numeric variables.
   Note : You have to manually read through the text file, and try to judge the actual variable types.
e) Drop non-Numeric variables from the DataFrame to have a clean DataFrame with Numeric variables.

Problem 2 : Statistical Summary

a) Find the Summary Statistics (Mean, Median, Quartiles etc) of SalePrice from the Numeric DataFrame.
b) Visualize the summary statistics and distribution of SalePrice using standard Box-Plot, Histogram, KDE.
c) Find the Summary Statistics (Mean, Median, Quartiles etc) of LotArea from the Numeric DataFrame.
d) Visualize the summary statistics and distribution of LotArea using standard Box-Plot, Histogram, KDE.
e) Plot SalePrice (y-axis) vs LotArea (x-axis) using jointplot, and check the Correlation between the two.

Important

Try to solve the problems on your own. Take help/hints from the "Preparation" codes and 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.