Exercise 2 : Basic Statistics

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 Exercise2_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 Exercise2_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 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 **train.csv** and the associated text file **data_description.txt** posted with this Exercise. The dataset and description are collected from Kaggle. You may also want to download the files directly from the Kaggle Competition (Login > Go to “Data” > “Download All”). Either way, read the competition description to get an idea about the task.

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

a) Import the “train.csv” data you downloaded (either from NTU Learn or Kaggle) in Jupyter Notebook.
b) What are the data types (“dtypes”) – int64/float64/object – of the variables (columns) in the dataset?
c) Create a new Pandas DataFrame consisting of only the variables (columns) of type Integer (int64).
d) Open the “data_description.txt” file you downloaded (either from NTU Learn or Kaggle) in Wordpad.
Read the description for each variable carefully and try to identify the “actual” Numeric variables.
Categorical variables are often “encoded” as Numeric variables for easy representation. Spot them.
e) Drop non-Numeric variables from the DataFrame to have a clean DataFrame with Numeric variables.

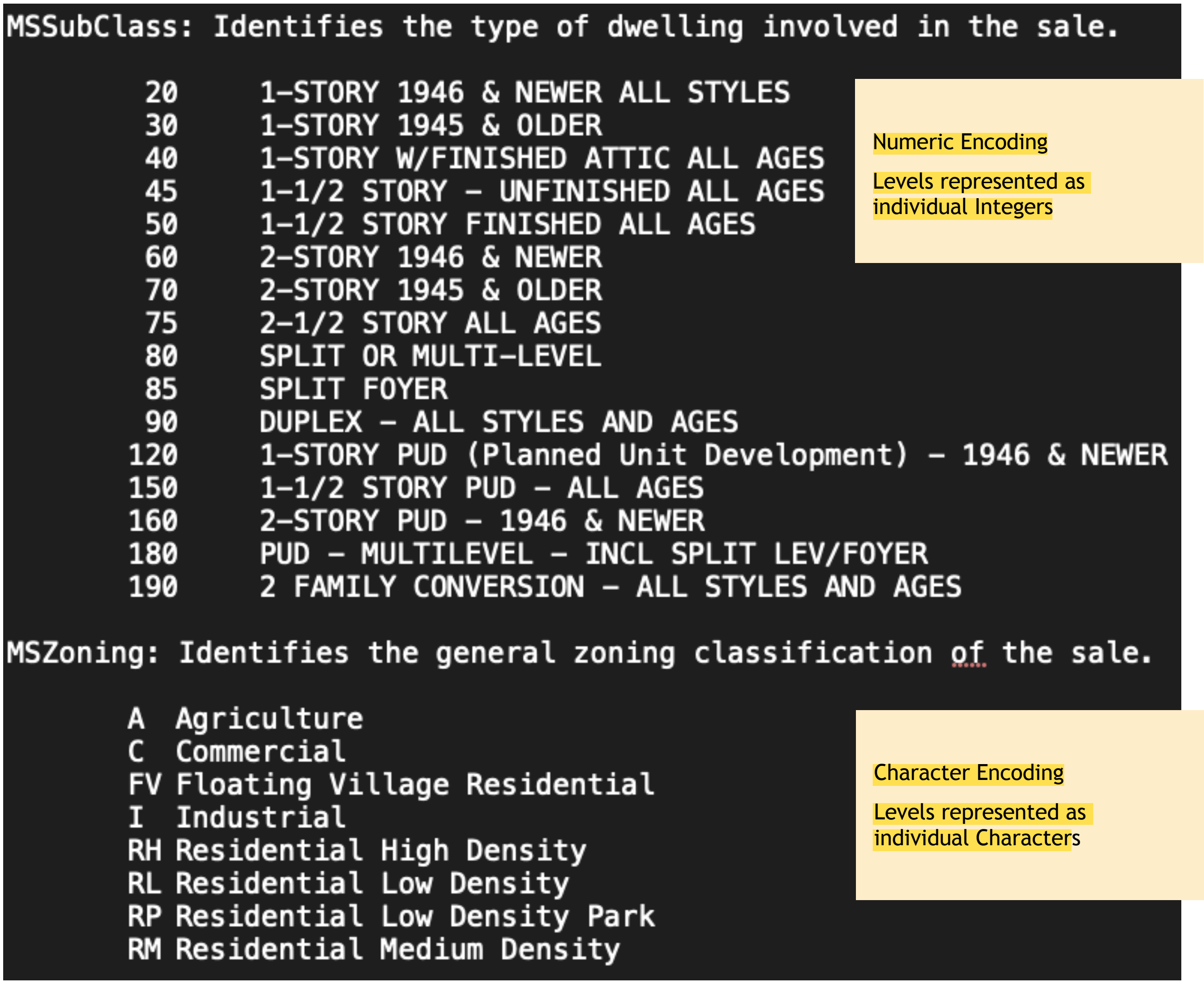
Problem 2 : Statistical Summary

Now that you have a “clean” DataFrame with only Numeric variables, we can safely perform standard statistics.

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 find the Correlation between the two.

Data Description

Note carefully that Categorical variables can be “encoded” in either of two ways, as follows. Even if a categorical variable is “encoded” as numbers, interpreting it as a numeric variable is wrong. Thus, one should be careful in reading the given data description file and identifying the “actual” numeric variables from the dataset to perform statistical exploration.



```
MSSubClass: Identifies the type of dwelling involved in the sale.

        20      1-STORY 1946 & NEWER ALL STYLES
        30      1-STORY 1945 & OLDER
        40      1-STORY W/FINISHED ATTIC ALL AGES
        45      1-1/2 STORY - UNFINISHED ALL AGES
        50      1-1/2 STORY FINISHED ALL AGES
        60      2-STORY 1946 & NEWER
        70      2-STORY 1945 & OLDER
        75      2-1/2 STORY ALL AGES
        80      SPLIT OR MULTI-LEVEL
        85      SPLIT FOYER
        90      DUPLEX - ALL STYLES AND AGES
       120      1-STORY PUD (Planned Unit Development) - 1946 & NEWER
       150      1-1/2 STORY PUD - ALL AGES
       160      2-STORY PUD - 1946 & NEWER
       180      PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
       190      2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

       A   Agriculture
       C   Commercial
       FV  Floating Village Residential
       I   Industrial
       RH  Residential High Density
       RL  Residential Low Density
       RP  Residential Low Density Park
       RM  Residential Medium Density
```

Numeric Encoding

Levels represented as individual Integers

Character Encoding

Levels represented as individual Characters

Bonus Problem

Create a new Pandas DataFrame consisting of all variables (columns) of type Integer (int64) or Float (float64). Read the description for each variable carefully and try to identify the “actual” Numeric variables in the data.

Drop non-Numeric variables from the DataFrame to have a clean DataFrame with only the Numeric variables. Plot SalePrice vs each of the Numeric variables you identified to understand their correlation or dependence.