1 Part 1 - Building up a basic predictive model

Load the dataset `Manhattan12.csv` into a pandas dataframe and carry out the following tasks.

Organise your code bearing in mind robustness and maintainability:

1. **Data cleaning and transformation:**

   If you have a closer look at the dataset, you will see that there are lots of missing values. They need be treated appropriately but in the first instance, we will take an aggressive approach to dealing with them.

   - Show the shape of the dataset
   - Rename incorrectly formatted column names (e.g. SALE\nPRICE)
   - Create list of categorical variables and another for the numerical variables
   - For each numerical column, remove the ',' the '$' for the sale price, and then convert them to numeric.
   - Convert the 'SALE DATE' to datetime.
   - For each categorical variable, remove the spaces, and then replace the empty string '' by NaN.
   - Replace the zeros in Prices, Land squares, etc. by NaN
   - Show a summary of all missing values as well as the summary statistics
   - Drop the columns 'BOROUGH', 'EASE-MENT', 'APARTMENT NUMBER'
   - Drop duplicates if any
   - Drop rows with NaN values
   - Identify and remove outliers if any
   - Show the shape of the resulting dataframe.
   - Consider the log of the prices and normalise the data.

2. **Data Exploration**. Consider the resulting dataframe. This first aggressive cleaning should give a smaller dataset, which you can start by exploring relationships between the various features of the dataset.

   - Visualise the prices across neighborhood
   - Visualise the prices over time
   - Show the scatter matrix plot and the correlation matrix
   - Any further plots, which demonstrate your understanding of the data

3. **Model building**. Consider the resulting dataframe.

   - Select the predictors that would have impact in predicting house prices.
   - Build up a first linear model with appropriate predictors and evaluate it. Split the data into a training and test sets; build up the model; and then show a histogram of the residuals. Evaluate your model by using a cross-validation procedure.

2 Part 2 - Improved model

This is an open-ended question and you are free to push your problem-solving skills in order to build up a useful model with higher performance.

1. Consider the entire datasets given in this assignment. Develop an improved predictive model that predicts the sales prices of houses. Make sure to validate your model. You should aim for a model with a higher performance while using a maximum of data points. This implies treating missing values differently for example through imputation rather than dropping them.
2. Use the K-Means algorithm to cluster your cleansed dataset and compare the obtained clusters with the distribution found in the data. Justify your clustering and visualise your clusters as appropriate.
3. Build up local regressors based on your clustering and discuss how this clusters-based regression compares to your regression model obtained in Part 2. 1.