Housing Project

Submitted by:

Suraj Kumar

ACKNOWLEDGMENT

Following online researches have done for data cleaning.

For outliers: **There are some techniques used to deal with outliers.**

1. Deleting observations.
2. Transforming values.
3. Imputation.
4. Separately treating.
5. Deleting observations. Sometimes it's best to completely remove those records from your dataset to stop them from skewing your analysis.

I have dropped the columns that contains huge number of outliers as compare to workable data.

For zscore: taken help from following url to calculate zscore of numerical columns.

https://stackoverflow.com/questions/64254167/remove-outliers-while-preserving-the-timestamps-in-dataframe

INTRODUCTION

- Business Problem Framing

  Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy.

  So, when any company tries to enter or which are already in this field do their research of every aspect which effect a house price so that they can buy or sell houses to make maximum profit.

- Motivation for the Problem Undertaken

  The objective to make this project is to build a model which give an idea about approx price of house under certain conditions. So that the house can be purchased or sell to make best possible profit.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  The data contains object as well as numeric columns.
  Following statistical analysis are done on numeric columns:
  1→ Mean of every numerical data is found.
  2→ Standard deviation
  3→ Minimum and maximum value is found.
  4→ 25^{th}, 50^{th} and 75^{th} percentile is found.

  Following are the analysis of these data:
  1→ Many columns have mean is greater than median.
  2→ Certain columns have large difference between 75th percentile and maximum value.
  3→There are outliers present in certain columns.
  4→ Standard deviation in certain columns are very high.

- Data Pre-processing Done

  Following steps are done for data cleaning:
  - Finding the correlation between target column and other independent variables.
  - Dropped the negatively correlated columns.
  - Finding the outliers using boxplot.
  - Dropped the columns which contains huge outliers as compare to workable data
  - Removed the outliers of other columns.

- ## Data Inputs- Logic- Output Relationships

  The input data contains category as well as numeric type where category data describes about the facility and all, while numeric data gives the information about the area, year, number of rooms etc.

  Certain data affect the output in some way like increase in area increase some price of house.

  Latest built house has certain high prices as compare to old ones.

- ## Hardware and Software Requirements and Tools Used

  Following libraries are used for data analysis:

  - Pandas: - Loading and handling the dataframe.
  - NumPy: - Working with arrays and some mathematical operations.
  - Matplotlib: - For visualisations
  - Seaborn: - For visualisations