Milestone Report

Capstone Project 1

Table of Contents

Problem Statement

- Buying a house is a very important juncture for the majority of individuals, both emotionally and financially and most probably a one time investment.
- Hence, this decision goes through a lot of consideration with endless comparison of features a house has to offer vs the Selling price of the house.
- Often few of the critical factors are overlooked and one ends up paying a premium sum for an unworthy house, effectively making the whole exercise a painful experience.
- Through this project, We would try to address this pain point of the house buyers i.e feature wise comparison of house vs. the price offered for the house.
- We are aiming to build a model, which would compare all the major/ minor features of a house (e.g Total plot area, Garage availability etc.) and based on the analysis try predicting a reasonable price for the house i.e. Selling price of the house.
- In short, the end product of this Project would be a reasonable selling price of a house.

Dataset Description

- To train and test this model, we are using the Ames Housing dataset, a dataset of 79 explanatory variables which describes in detail about the various aspects of the residential homes in Ames, Iowa.
- The dataset is sourced from kaggle, made available as part of on-going Kaggle competition

Data Cleaning and Preparation

The below steps were taken to clean up the data.

- First of all, we addressed the columns with missing values. All the columns were identified which consisted of even a single NULL/ NaN value.
- If the number of missing values exceed 50% of the total entries for a particular column, then we drop the column altogether.
- For columns containing categorical values, We replace the missing values with the mode value of the column.
- For columns with numerical variables, replace the missing values with the median value of the column.
- We also plotted a box plot for all the numerical variables, and in turn observed how the data is distributed across the range values.
- From the box plot above, We also got an understanding of the outliers present for each numerical column, which would be addressed during the EDA part.

Initial Findings from the Exploratory Data Analysis