A Project Synopsis on

“House Price Prediction Using Regression”.

Submitted to **Manipal University Jaipur** towards the partial fulfilment for the award of the degree of

Bachelors Of Technology(*B.Tech*) In Information Technology(*IT*)

2022 – 2026

*By Lakshya Pawar(**229302177**), Ayonija(**229302173**), Navodit Kapoor(**229302204**) and Sidhali Singh(**229302191**)*

MANIPAL UNIVERSITY
JAIPUR
(University under Section 2(f) of the UGC Act)

Under the guidance of ***Mrs. Nandani Sharma***

Department of Information Technology

School of Computing and Information Technology

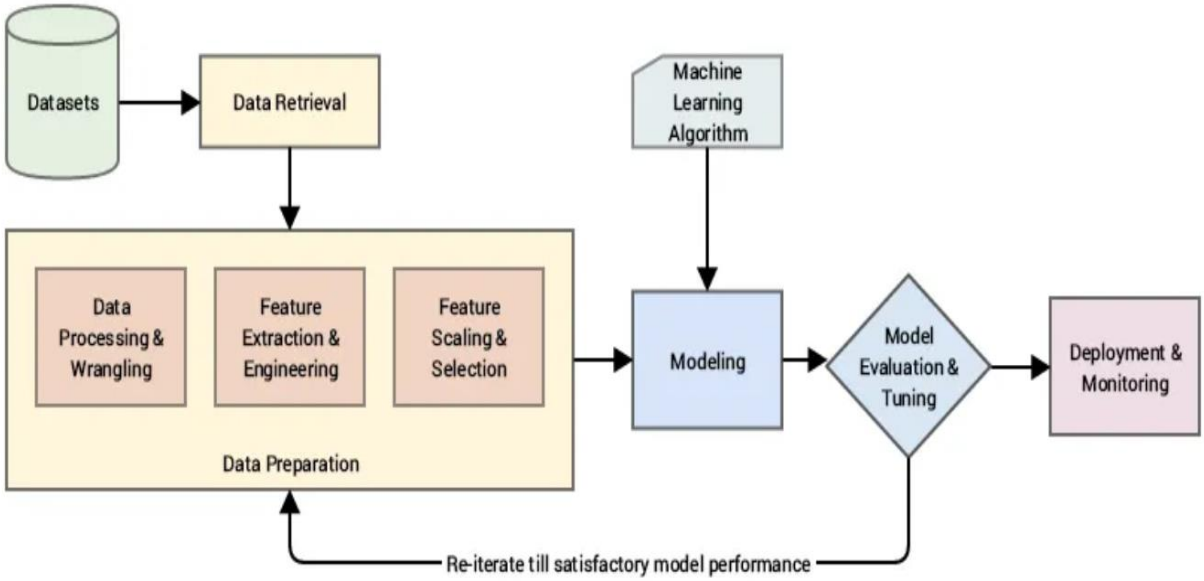
Manipal University Jaipur

Jaipur, Rajasthan

1. Introduction:

House Price Prediction using Regression ***(Focus: Supervised Learning (Linear Regression))*** aims to predict house prices based on various features of the houses. The objective is to build a regression model to estimate house prices.

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview



2. Motivation:

The real estate market plays a crucial role in the global economy, where accurate house price predictions can significantly benefit buyers, sellers, investors, and financial institutions. Traditionally, estimating house prices relied on intuition and basic heuristics, which are often subjective and inaccurate. This inspired me to leverage data science to develop a more accurate and data-driven approach.

The primary motivation for this project is to explore how machine learning techniques, particularly regression models, can predict house prices based on various features such as location, size, number of rooms, and neighborhood amenities.

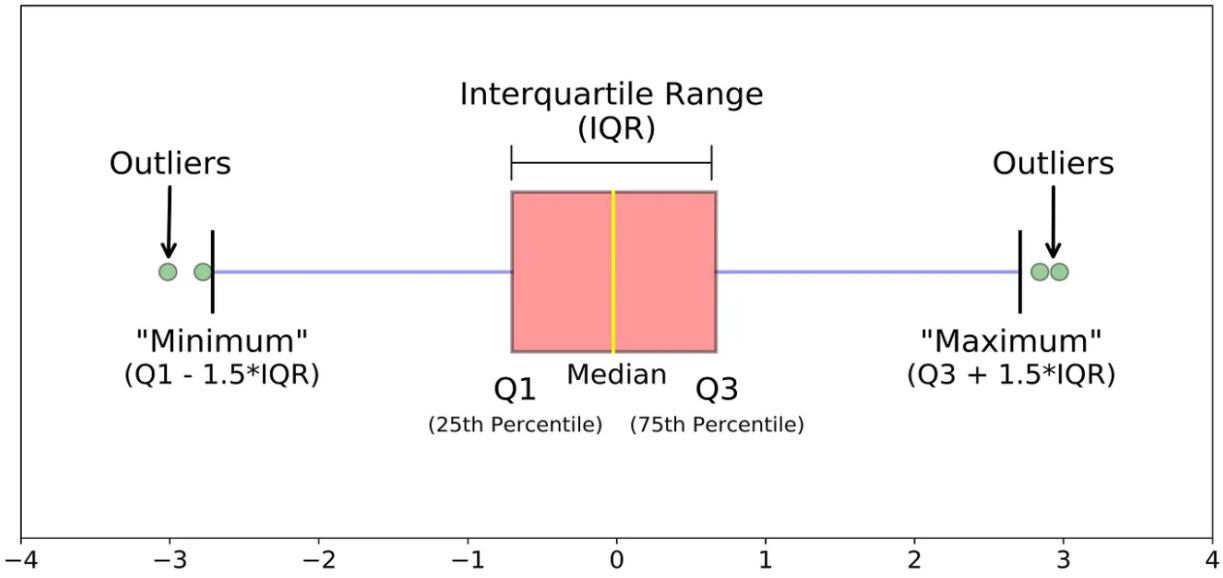
Furthermore, this project offers a hands-on opportunity to apply data science concepts, enhance our skills in data analysis, feature engineering, and machine learning, and gain experience in working with real-world datasets. The ultimate goal is to combine theoretical knowledge with practical application to solve a relevant and impactful problem in the real estate domain.

3. Project Objectives:

The primary objective of the House Price Prediction project is to develop a machine learning model that accurately predicts the selling price of houses based on various features, such as location, property size, number of bedrooms, and other relevant attributes. This involves leveraging regression techniques to build a model that can analyze historical housing data, identify patterns, and provide price estimates for new listings.

To achieve this, the project will place a significant focus on exploratory data analysis (EDA) and data visualization to uncover hidden trends and correlations within the dataset. Furthermore, the project aims to utilize data visualization tools such as Matplotlib, Seaborn to generate meaningful insights from the data.

Visualizations like heatmaps, histograms, and box plots will be used to explore the distribution of house prices, identify outliers, and examine the impact of various factors on property values.



Box plot

4. Methodology:

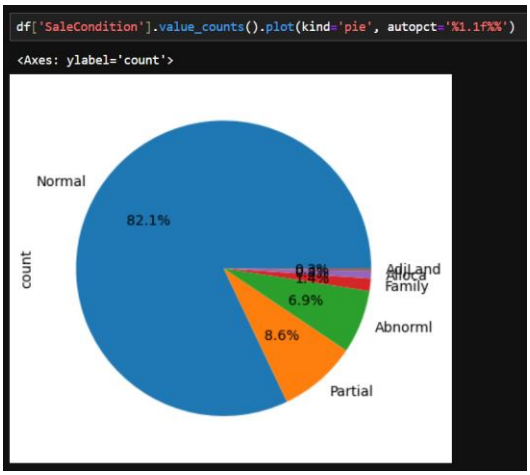
The methodology consists of several key stages, including data exploration, data cleaning, feature engineering, data preprocessing, and model building.

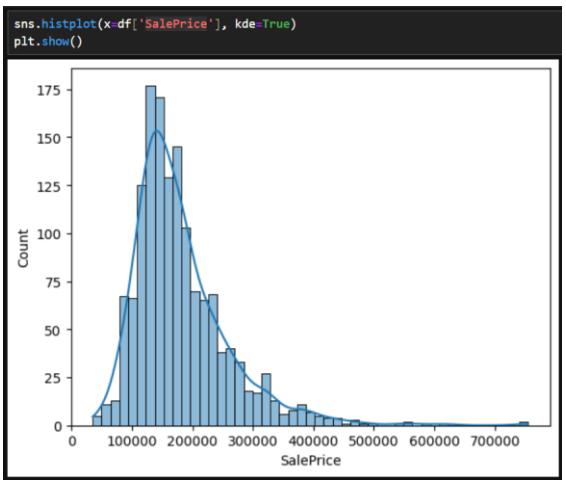
https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

*i. **Data Exploration**:*

- The first step involves conducting **exploratory data analysis (EDA)** to gain an understanding of the dataset's structure, distribution, and key characteristics. Using data visualization techniques like histograms, pie charts, and box plots, we explored relationships between features.

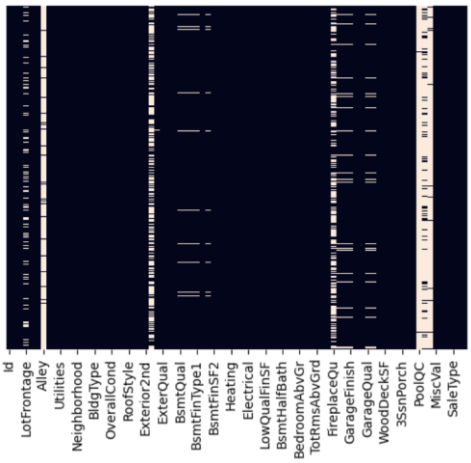
- This analysis helped identify trends, correlations, and patterns in the data, allowing for a deeper understanding of how different variables impact house prices.

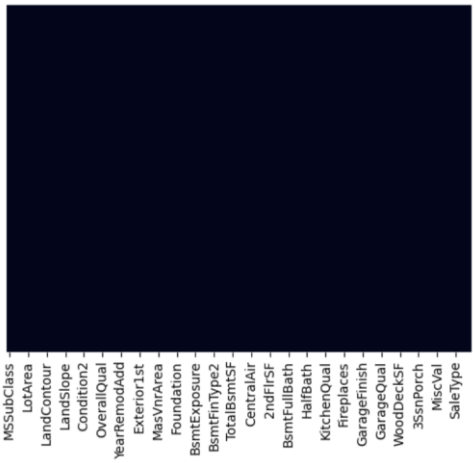




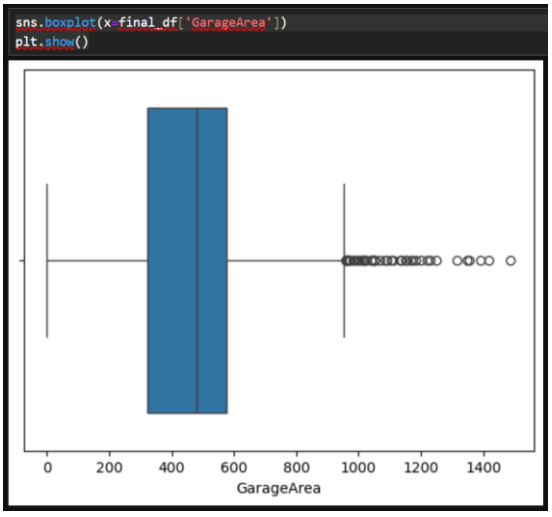
ii. ***Data Cleaning****:*

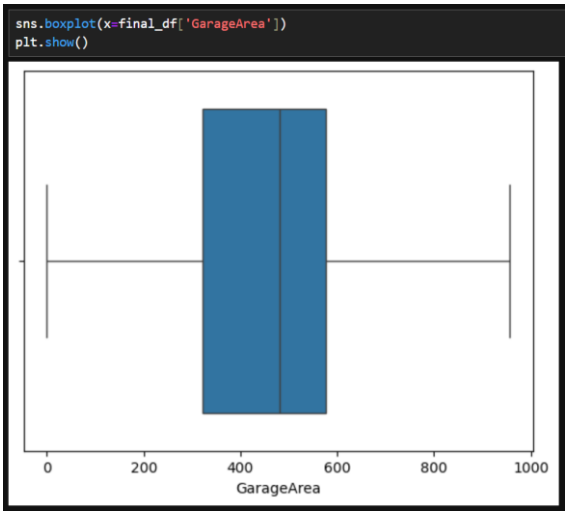
- The next step focused on **data cleaning** to ensure the dataset's quality and integrity. This involved handling missing values, which were either imputed with appropriate statistics (mean, median, or mode) or removed if necessary.





- Outliers that could negatively affect model performance were identified using visualizations like box plots and handled accordingly, either by replacing them with threshold values or removing them.





iii. ***Feature Engineering****:*

- In this phase, feature engineering was performed to create new features that could enhance the model's predictive power. For example, features such as the property area, and cost per square foot could be derived from existing data.

- Transformations were applied to skewed numerical features to make them more normally distributed, which can enhance the performance of models sensitive to feature distributions.

iv. ***Data Preprocessing****:*

- The data preprocessing stage focused on preparing the data for model training. This involved **encoding categorical variables** using techniques like one-hot encoding for nominal features and label encoding for ordinal features.

- Features were then **scaled** using methods like StandardScaler or MinMaxScaler to ensure all numerical variables had comparable scales, which is crucial for models sensitive to feature magnitude.

- The dataset was subsequently **split into training and testing sets** (typically 80:20 split) to evaluate model performance on unseen data.

In [46]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
```

In [47]:
```
X = df_Train.drop(['SalePrice'], axis=1)
y = df_Train['SalePrice']
```

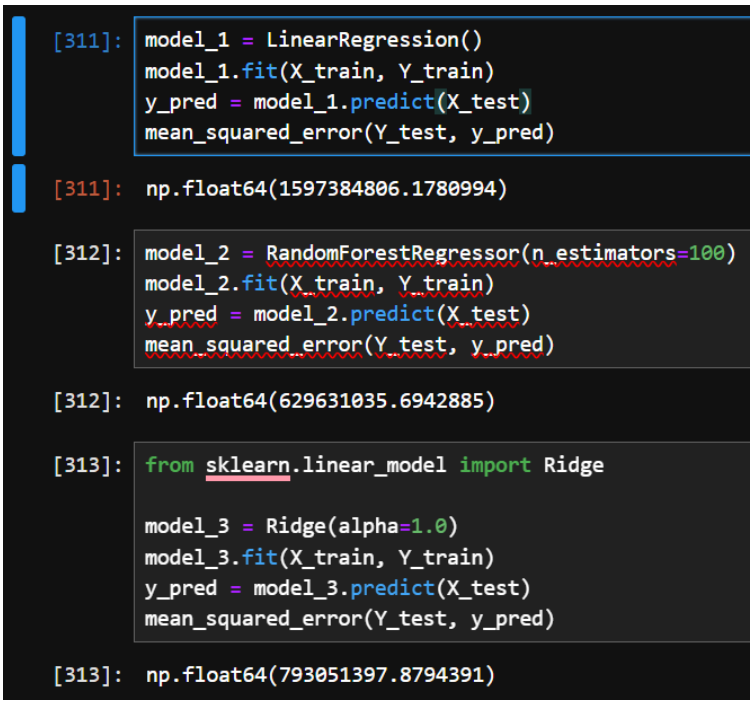
In [48]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state=42)
```

*v. **Model Building**:*

- The final step involved **model building and evaluation**. Various regression models were trained, starting with simple models like **Linear Regression** and advancing to more complex models like **Random Forest Regression** and other ensemble methods.

- The models were evaluated based on metrics such as **Mean Absolute Error (MAE)**, and **Root Mean Squared Error (RMSE)**, to determine their accuracy and robustness.

- Finally, the best-performing model with the least MAE was selected and further analyzed to interpret the importance of various features, providing insights into what factors most significantly influence house prices.



5. Result:

Model	MSE	RMSE
Linear Regression	1597384806.18	39967.30
Random Forest Regressor	629631035.69	25092.09
Ridge Regression	793051397.88	28164.38

The Random Forest Regressor had the lowest RMSE value, so we made the final predictions using Model 2 (Random Forest Regressor).

6. Resources:

To implement the project, we will need a system with these specifications.

- *Python 3.11-3.12*
- *Jupyter Notebook*
- *Pandas*
- *Numpy*
- *Scikit-learn*
- *Matplotlib*
- *Seaborn*

7. Bibliography:

The following resources will be utilized to deepen understanding and expand knowledge on the topic:

- *https://www.w3schools.com/python/matplotlib_intro.asp*
- *https://medium.com/@denizgunay/feature-engineering-data-preprocessing-d6bc219b6b93*
- *https://medium.com/@spinjosovsky/normalize-data-before-or-after-split-of-training-and-testing-data-7b8005f81e26*
- *https://www.javatpoint.com/regression-analysis-in-machine-learning*
- *https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd*

8. Timeline:

- ***August –*** *Data Exploration, Data Cleaning, EDA*
- ***September –*** *Feature Engineering, Data Pre-Processing*
- ***October –*** *Model Building*
- ***November –*** *Documentation*

-x-

Thank You.