Weekly Report for CSE 6940 (Graduate Research Methods in ComputerScience)  
By: Nikitha Kondreddy Venkatramireddy  
Date: September 16, 2024

**House Price Prediction using Machine Learning in Python**

**Objective:**

The goal of this project is to predict house prices using various machine learning techniques based on available housing data. This week focused on project setup, data acquisition, cleaning, and exploratory data analysis (EDA).

**Project Setup:**

* **Libraries Installed:** Installed necessary Python libraries using the following command

**A black rectangular object with a black stripe

Description automatically generatedA black and grey rectangular object

Description automatically generatedA black and grey rectangular object

Description automatically generatedA black and grey rectangular object

Description automatically generatedA black rectangular object with a black stripe

Description automatically generated**

These libraries are essential for data manipulation, visualization, and machine learning model building.

* **Environment Setup:**Set up a Jupyter Notebook environment for interactive coding and organized code into structured cells for better readability.
* **Version Control:**Initialized a Git repository to track the progress of the project. The code and project updates are being pushed to a GitHub repository.

<https://github.com/Nikitha130731/House-Price-Prediction-using-ML>

**Dataset Acquisition:**

* **Dataset Source:**

Acquired the dataset from Kaggle, which contains detailed information about houses in Ames, Iowa. This dataset includes 1,460 rows and 81 columns, with various features like house area, quality, and neighborhood.

<https://www.kaggle.com/code/gusthema/house-prices-prediction-using-tfdf/input>

* **Dataset Loaded:**

Loaded the dataset into a pandas DataFrame using the following code

A screenshot of a computer

Description automatically generated

**Data Cleaning and Preprocessing:**

* **Missing Data Handling**

Checked for missing values in the dataset

A screen shot of a computer code

Description automatically generated

Some features like PoolQC, MiscFeature, and Alley had a large number of missing values. These columns were dropped as part of data cleaning, while features like LotFrontage were targeted for imputation.

* **Imputation Strategy:**

I plan to use the median or mean for numerical features and the most frequent category for categorical features during the next phase of data preprocessing.

**Feature Selection Strategy:**

* + **OverallQual** (Overall Quality of the house)
  + **GrLivArea** (Above-ground living area in square feet)
  + **GarageCars** (Size of garage in car capacity)
  + **YearBuilt** (Year the house was built)

I will explore additional features and refine the list in future steps.

**Challenges Faced:**

* **Missing Data:**  
  Several columns had a large percentage of missing values (e.g., Alley, PoolQC), requiring careful handling. I plan to continue imputing missing values in the upcoming week.
* **Data Size:**  
  With 81 columns, careful feature selection is necessary to avoid overfitting or slowing down the model-building process.

**Next Week's Plan:**

* Complete data preprocessing by imputing missing values and encoding categorical features.
* Begin building and testing a basic **Linear Regression model** to predict house prices.
* Analyze model performance using metrics like Mean Squared Error (MSE)

**References:**

* <https://www.kaggle.com/competitions/home-data-for-ml-course/data>
* <https://pandas.pydata.org/docs/user_guide/index.html>
* <https://www.w3schools.com/python/python_ml_getting_started.asp>