Advanced data visualization

Experiment-5

Name	Parth Gandhi
UID	2021300033
Batch	Batch H
ṇDepartment	COMPS A

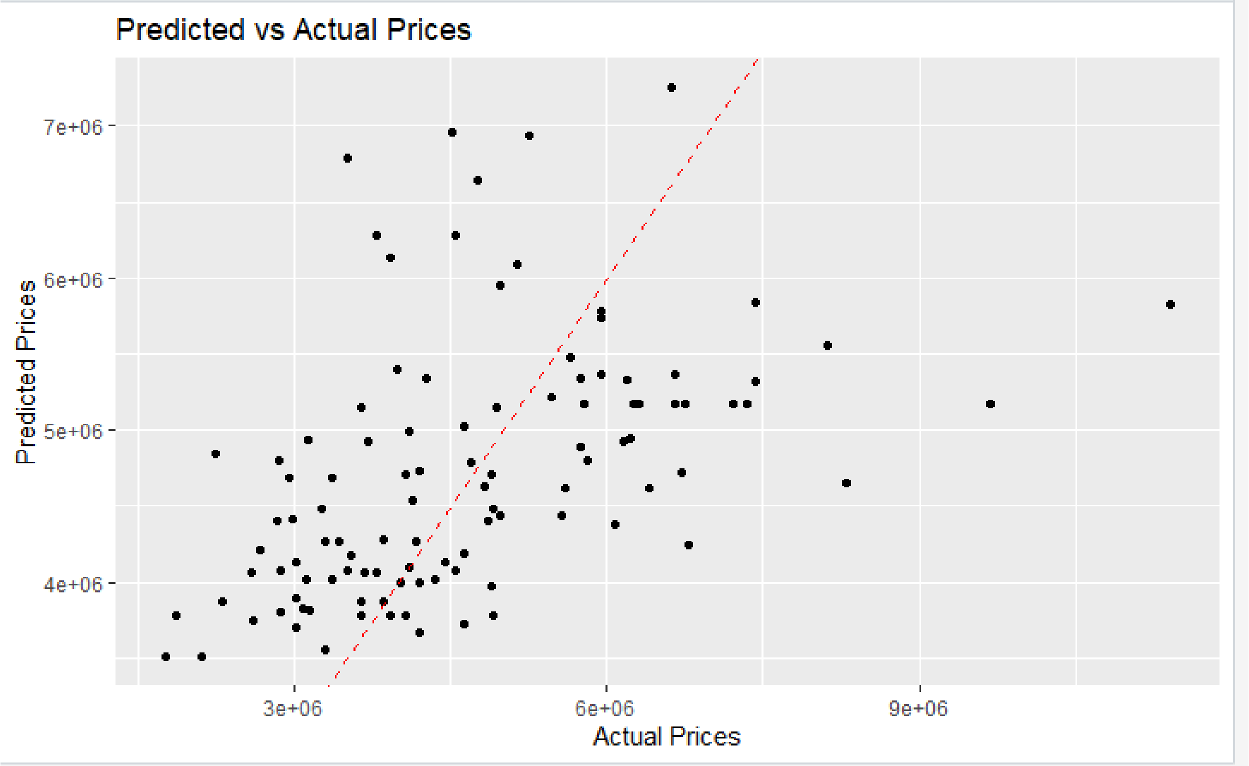
Aim: Create advance chart charts (linear regression and logistic regression) using R on dataset Housing data set

Linear regression

Dataset description: The dataset used is a simple housing dataset based on certain factors like house area, bedrooms, furnished, nearness to mainroad, etc. Description of each attribute is given below:

1. **price**: The price of the house in currency units.
2. **area**: The total area of the house in square feet.
3. **bedrooms**: Number of bedrooms in the house.
4. **bathrooms**: Number of bathrooms in the house.
5. **stories**: Number of floors (stories) in the house.
6. **mainroad**: Whether the house is located near a main road (yes/no).
7. **guestroom**: Whether the house has a guest room (yes/no).
8. **basement**: Whether the house has a basement (yes/no).
9. **hotwaterheating**: Whether the house has hot water heating (yes/no).
10. **airconditioning**: Whether the house has air conditioning (yes/no).
11. **parking**: Number of parking spaces available.
12. **prefarea**: Whether the house is located in a preferred area (yes/no).
13. **furnishingstatus**: The furnishing status of the house, i.e., furnished, semi-furnished, or unfurnished.

Output and analysis:



The scatter plot titled "Predicted vs Actual Prices" is a visualization used to evaluate the performance of a linear regression model for predicting house prices from a housing dataset. The plot compares the predicted prices (on the y-axis) against the actual prices (on the x-axis) of houses.

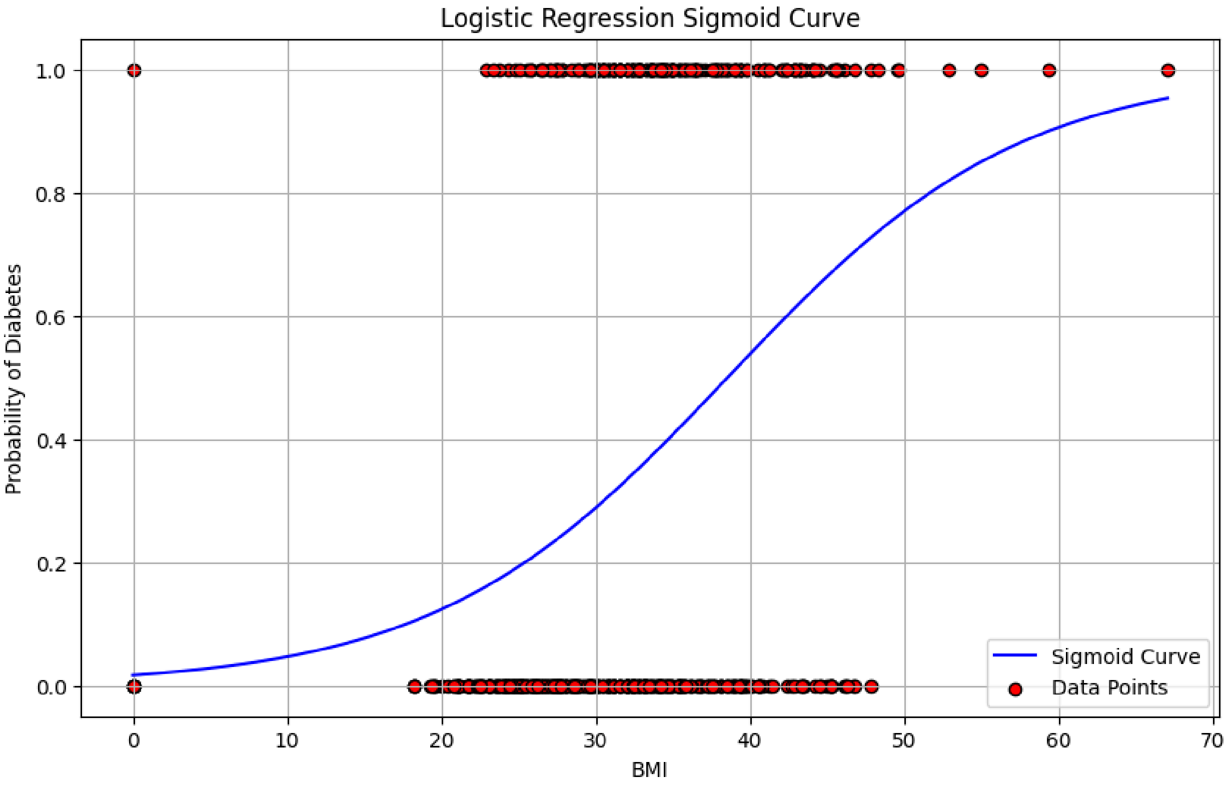
1. **Line of Perfect Prediction**: The dashed red line represents the line of perfect prediction, where predicted prices would equal actual prices. Points along this line indicate perfect predictions.
2. **Data Distribution**:
   - The concentration of data points primarily around the lower and middle part of the price range suggests that the model performs relatively well for lower to moderately priced homes.
   - There's a noticeable spread in points as the actual prices increase, indicating variability in the accuracy of predictions for higher-priced houses.
3. **Outliers and Predictive Accuracy**:
   - There are a few points far from the line of perfect prediction, especially at the higher end of the price spectrum. These represent cases where the model significantly underpredicted or overpredicted the actual values.
   - The model appears to underperform when predicting the highest prices, as many points at higher actual prices lie below the line of perfect prediction.

Logistic regression

Dataset description: The dataset used is a diabetes dataset containing information such as BMI, age, glucose level etc. to predict outcome of having diabetes. Below is a description of each attribute:

1. **Pregnancies**: Number of times the patient has been pregnant.
2. **Glucose**: Plasma glucose concentration after a 2-hour oral glucose tolerance test.
3. **BloodPressure**: Diastolic blood pressure (mm Hg).
4. **SkinThickness**: Thickness of the triceps skin fold (mm), a measure related to body fat.
5. **Insulin**: Serum insulin levels (mu U/ml).
6. **BMI**: Body Mass Index, calculated as weight (kg) / (height (m))².
7. **DiabetesPedigreeFunction**: A function that scores the likelihood of diabetes based on family history.
8. **Age**: Age of the patient in years.
9. **Outcome**: Class label (0 or 1), where 1 indicates the presence of diabetes, and 0 indicates its absence.

Output and analysis:



The chart titled "Logistic Regression Sigmoid Curve" represents a logistic regression model's probability estimates for the occurrence of diabetes based on Body Mass Index (BMI) values.

1. **Sigmoid Curve**:
    - The blue line represents the logistic function, which models the probability that an individual has diabetes based on their BMI.
    - The curve is sigmoidal, typical for logistic regression, where the output is bounded between 0 and 1, representing probabilities.
2. **Data Points**:
    - The red dots indicate individual observations from the dataset, where the y-value is binary (0 or 1), representing whether the individual does not have (0) or does have (1) diabetes.
    - Points are clustered at the 0 and 1 y-values across various BMI levels, indicating the observed outcomes.

- **Probability Transition**:
    - For lower BMI values (roughly below 30), the probability of having diabetes is close to zero, which is consistent with medical understanding that lower BMI is generally associated with lower risk of type 2 diabetes.
    - As BMI increases, particularly beyond a threshold around 30, the probability of diabetes increases significantly, sharply rising past BMI 40 and plateauing near 1 as BMI approaches 60 and beyond. This suggests high BMI is a strong indicator of potential diabetes risk.
- **Model Fit**:
    - The sigmoid curve seems to fit the data well, particularly in capturing the critical transition where the probability of diabetes escalates as BMI increases.
    - There are some outliers, particularly at very low and very high BMI values, where the model and data points do not align perfectly, indicating potential for model refinement or the influence of other factors not accounted for by BMI alone.

Conclusion: In my experiment, I applied both linear and logistic regression models to two different datasets. The first chart represents the linear regression model I built to predict housing prices from the housing dataset. While the model captures the general trend of the data, as seen in the alignment of predicted and actual prices, there is noticeable deviation from the ideal fit line. This suggests that the model could be further refined to better capture certain patterns in the data or handle variability more effectively.

The second chart illustrates my logistic regression model applied to the diabetes dataset, using BMI to predict the probability of diabetes. The sigmoid curve effectively demonstrates the non-linear relationship between BMI and the risk of diabetes, with the data points showing a clear separation between the likelihood of having diabetes (1) and not having it (0). This shows that the logistic regression model performs well in predicting binary outcomes. Overall, my experiment highlights the strengths of both regression models for prediction and classification tasks across different types of datasets.