Sardar Patel Institute of Technology
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058, India
(Autonomous College Affiliated to University of Mumbai)

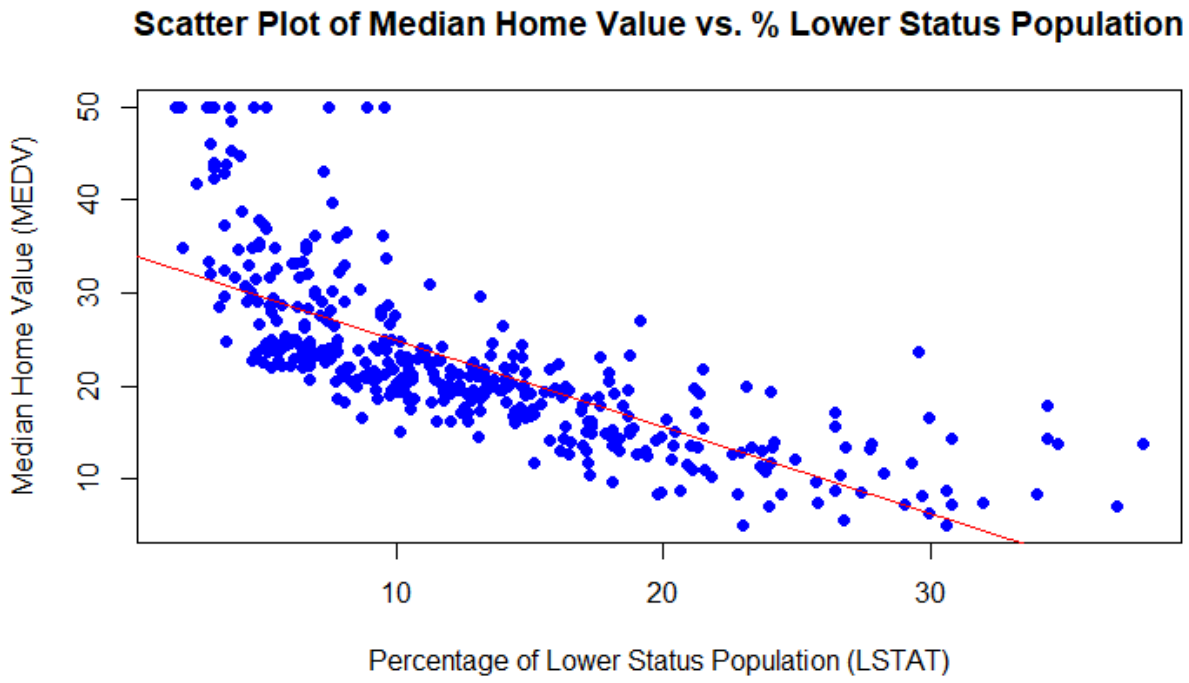
ADVANCED DATA VISUALIZATION

Experiment No.	5
Aim	Perform Linear Regression and Logistic Regression on HOUSING DATASET using R.
Name	Chinmay Jadhav
UID No.	2021300046
Class & Division	BE COMPS A(F)

DATASET: https://www.kaggle.com/datasets/altavish/boston-housing-dataset

1. LINEAR REGRESSION:

SCATTER PLOT:



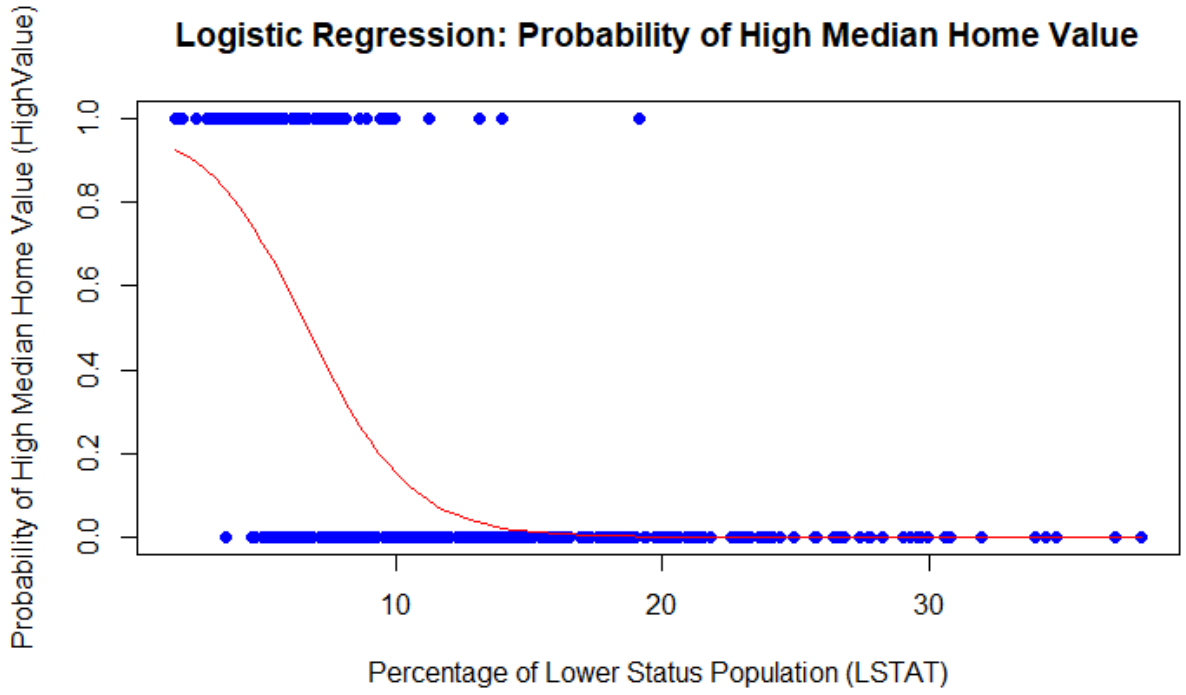
CONCLUSION: The scatter plot helps us understand how the percentage of lower-status population affects median home values. For instance, if there's a clear downward trend, we can say that areas with a higher percentage of lower-status populations tend to have lower median home values.

```
> data= read.csv("C:\Users\students\Desktop\HousingData.csv")
Error: '\U' used without hex digits in character string (<input>:1:20)
> data= read.csv("C:/Users/students/Desktop/HousingData.csv")
> summary(data)
      CRIM                ZN              INDUS            CHAS
NOX
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.
:0.3850
 1st Qu.: 0.08190   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st
Qu.:0.4490
 Median : 0.25372   Median :  0.00   Median : 9.69   Median :0.00000
Median :0.5380
 Mean   : 3.61187   Mean   : 11.21   Mean   :11.08   Mean   :0.06996   Mean
:0.5547
 3rd Qu.: 3.56026   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd
Qu.:0.6240
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.
:0.8710
 NA's   :20         NA's   :20       NA's   :20      NA's   :20
       RM              AGE              DIS              RAD              TAX
 Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.
:187.0
 1st Qu.:5.886   1st Qu.: 45.17   1st Qu.: 2.100   1st Qu.: 4.000   1st
Qu.:279.0
 Median :6.208   Median : 76.80   Median : 3.207   Median : 5.000   Median
:330.0
 Mean   :6.285   Mean   : 68.52   Mean   : 3.795   Mean   : 9.549   Mean
:408.2
 3rd Qu.:6.623   3rd Qu.: 93.97   3rd Qu.: 5.188   3rd Qu.:24.000   3rd
Qu.:666.0
 Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.
:711.0
                 NA's   :20
    PTRATIO             B              LSTAT             MEDV
 Min.   :12.60   Min.   :  0.32   Min.   : 1.730   Min.   : 5.00
 1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 7.125   1st Qu.:17.02
 Median :19.05   Median :391.44   Median :11.430   Median :21.20
 Mean   :18.46   Mean   :356.67   Mean   :12.715   Mean   :22.53
 3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.955   3rd Qu.:25.00
 Max.   :22.00   Max.   :396.90   Max.   :37.970   Max.   :50.00
                                  NA's   :20
> str(data)
'data.frame':     506 obs. of  14 variables:
 $ CRIM   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ ZN     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ INDUS  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ CHAS   : int  0 0 0 0 0 0 NA 0 0 NA ...
 $ NOX    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524
```

```
...
 $ RM      : num  6.58 6.42 7.18 7 7.15 ...
 $ AGE     : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ DIS     : num  4.09 4.97 4.97 6.06 6.06 ...
 $ RAD     : int  1 2 2 3 3 3 5 5 5 5 ...
 $ TAX     : int  296 242 242 222 222 222 311 311 311 311 ...
 $ PTRATIO: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ B       : num  397 397 393 395 397 ...
 $ LSTAT   : num  4.98 9.14 4.03 2.94 NA ...
 $ MEDV    : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
> sum(is.na(data))
[1] 120
> data <- na.omit(data)
> model <- lm(medv ~ lstat, data = data)
Error in eval(predvars, data, env) : object 'medv' not found
> colnames(data)  # List all column names
 [1] "CRIM"    "ZN"      "INDUS"   "CHAS"    "NOX"     "RM"      "AGE"
"DIS"     "RAD"
[10] "TAX"     "PTRATIO" "B"       "LSTAT"   "MEDV"
> model <- lm(MEDV ~ LSTAT, data = data)
> summary(model)

Call:
lm(formula = MEDV ~ LSTAT, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-9.833 -3.944 -1.334  2.094 24.628

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.23580    0.62154   55.08   <2e-16 ***
LSTAT       -0.93007    0.04226  -22.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.123 on 392 degrees of freedom
Multiple R-squared:  0.5527,  Adjusted R-squared:  0.5516
F-statistic: 484.4 on 1 and 392 DF,  p-value: < 2.2e-16

> plot(data$LSTAT, data$MEDV,
+     main = "Scatter Plot of Median Home Value vs. % Lower Status
Population",
+     xlab = "Percentage of Lower Status Population (LSTAT)",
+     ylab = "Median Home Value (MEDV)",
+     pch = 19,  # Type of point
+     col = "blue")  # Color of points
> abline(model, col = "red")
```

2. LOGISTIC REGRESSION

PLOT:



CONCLUSION: The plot tells us that homes in areas with a greater proportion of lower-income or less-privileged residents are generally less likely to have high median values. The curve helps predict this probability smoothly for different levels of the `LSTAT` variable.

```
data$HighValue <- ifelse(data$MEDV > 25, 1, 0) # Here, we'll
create a binary variable 'HighValue' which is 1 if MEDV > 25,
and 0 otherwise
> logistic_model <- glm(HighValue ~ LSTAT, data = data,
family = "binomial")
>
> summary(logistic_model)

Call:
glm(formula = HighValue ~ LSTAT, family = "binomial", data =
data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.4267     0.4783   7.164 7.83e-13 ***
LSTAT        -0.5148     0.0613  -8.398  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 432.96  on 393  degrees of freedom
Residual deviance: 243.71  on 392  degrees of freedom
AIC: 247.71

Number of Fisher Scoring iterations: 7

> data$predicted_probabilities <- predict(logistic_model,
type = "response")
>
> plot(data$LSTAT, data$HighValue,
+     main = "Logistic Regression: Probability of High Median
Home Value",
+     xlab = "Percentage of Lower Status Population (LSTAT)",
+     ylab = "Probability of High Median Home Value
(HighValue)",
+     pch = 19,  # Type of point
+     col = "blue")
> curve(predict(logistic_model, data.frame(LSTAT = x), type =
"response"), add = TRUE, col = "red")
```