Problem:

A real estate economist collects information on 1000 house price sales from two similar neighborhoods, one called "University Town" bordering a large state university, and another a neighborhood about three miles from the university. He specifies the following regression equation:

$$y = \beta_0 + \beta_1 x_1 + \delta_2 D_2 + \gamma(D_2 * x_1) + \beta_3 x_3 + \delta_4 D_4 + \delta_5 D_5 + \varepsilon$$

where

y = house prices in \$1000

x_1 = the number of hundreds of square feet of living area

D_2 ={

1 house near university

0 otherwise

x_3 = age of the house (in years)

D_4 ={

1 house has pool

0 otherwise

D_5 ={

1 if fireplace is present

0 otherwise

Discuss the effect of these variables on house prices

Solution

Read the dataset###########################

df <- read.table("C:/Users/Asus/Downloads/housing.txt", header = TRUE)

head(df, n=10)

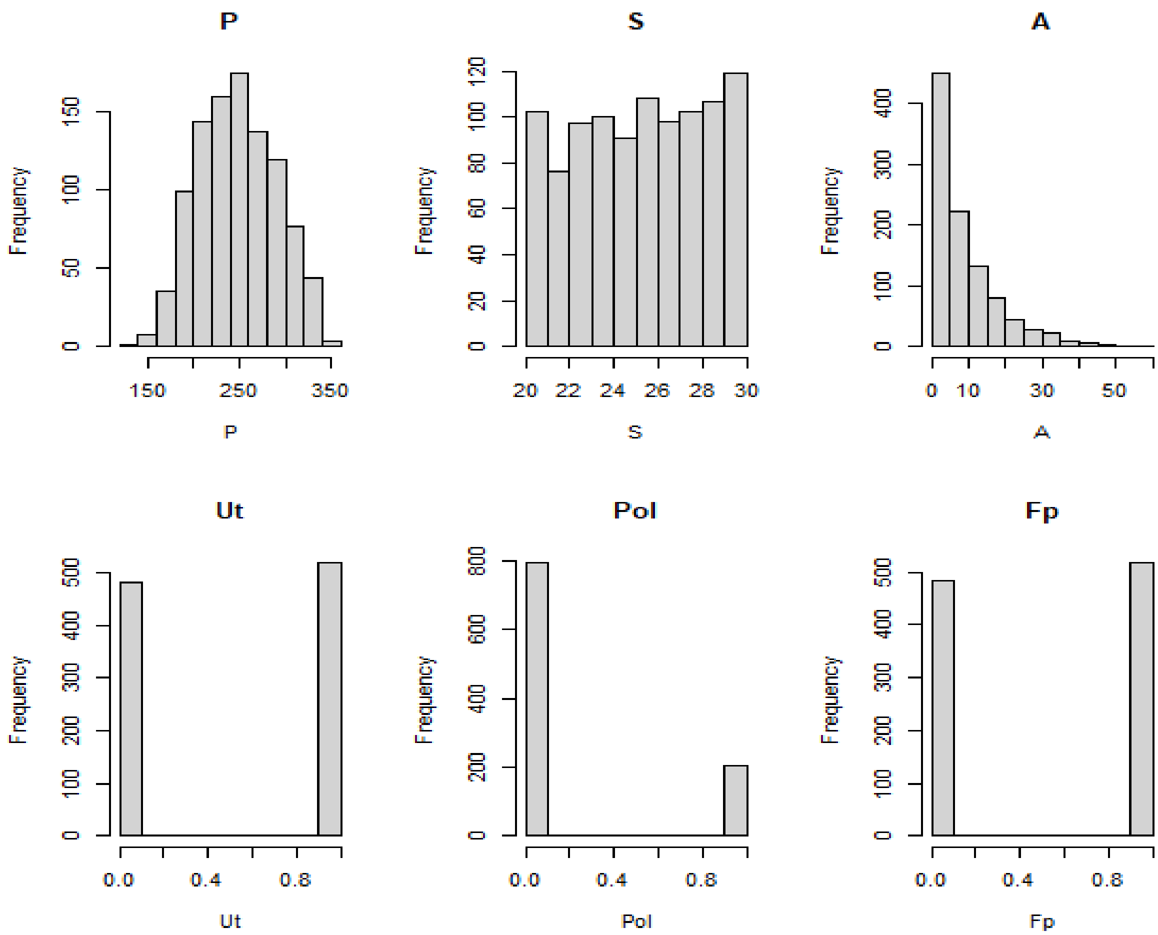
```
          P     S A Ut Pol Fp
1  205.452 23.46 6  0   0  1
2  185.328 20.03 5  0   0  1
3  248.422 27.77 6  0   0  0
4  154.690 20.17 1  0   0  0
5  221.801 26.45 0  0   0  1
6  199.119 21.56 6  0   0  1
7  272.134 29.91 9  0   0  1
8  250.631 27.98 0  0   0  1
9  197.240 24.80 0  0   1  0
10 235.755 27.50 0  0   0  0
```

summary(df)

```
       P               S               A               Ut       
 Min.   :134.3   Min.   :20.03   Min.   : 0.000   Min.   :0.000  
 1st Qu.:215.6   1st Qu.:22.83   1st Qu.: 3.000   1st Qu.:0.000  
 Median :245.8   Median :25.36   Median : 6.000   Median :1.000  
 Mean   :247.7   Mean   :25.21   Mean   : 9.392   Mean   :0.519  
 3rd Qu.:278.3   3rd Qu.:27.75   3rd Qu.:13.000   3rd Qu.:1.000  
 Max.   :345.2   Max.   :30.00   Max.   :60.000   Max.   :1.000  

      Pol              Fp       
 Min.   :0.000   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:0.000  
 Median :0.000   Median :1.000  
 Mean   :0.204   Mean   :0.518  
 3rd Qu.:0.000   3rd Qu.:1.000  
 Max.   :1.000   Max.   :1.000  
```

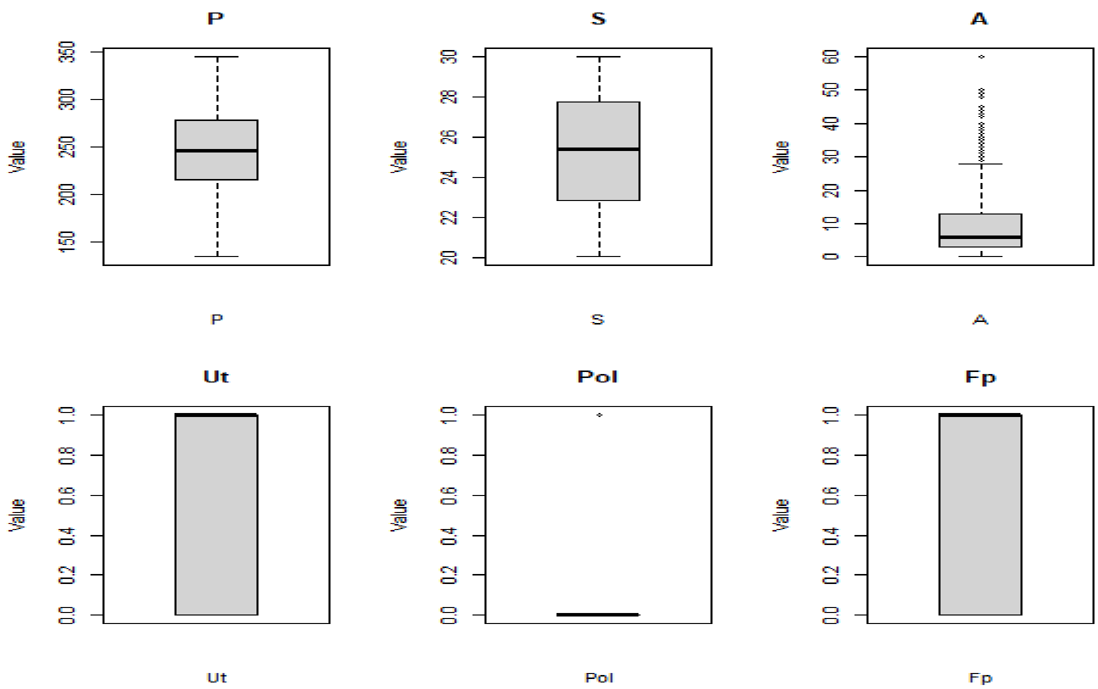
Interpretation: Here from this summery, it is noticeable that for the feature Age the max value is significantly larger than all the measure of central tendency.

```
###################################### Histogram ######################

par(mfrow = c(2, 3))

for (col in colnames(df)) {

  hist(df[[col]], main = col, xlab = col, ylab = "Frequency")

}
```



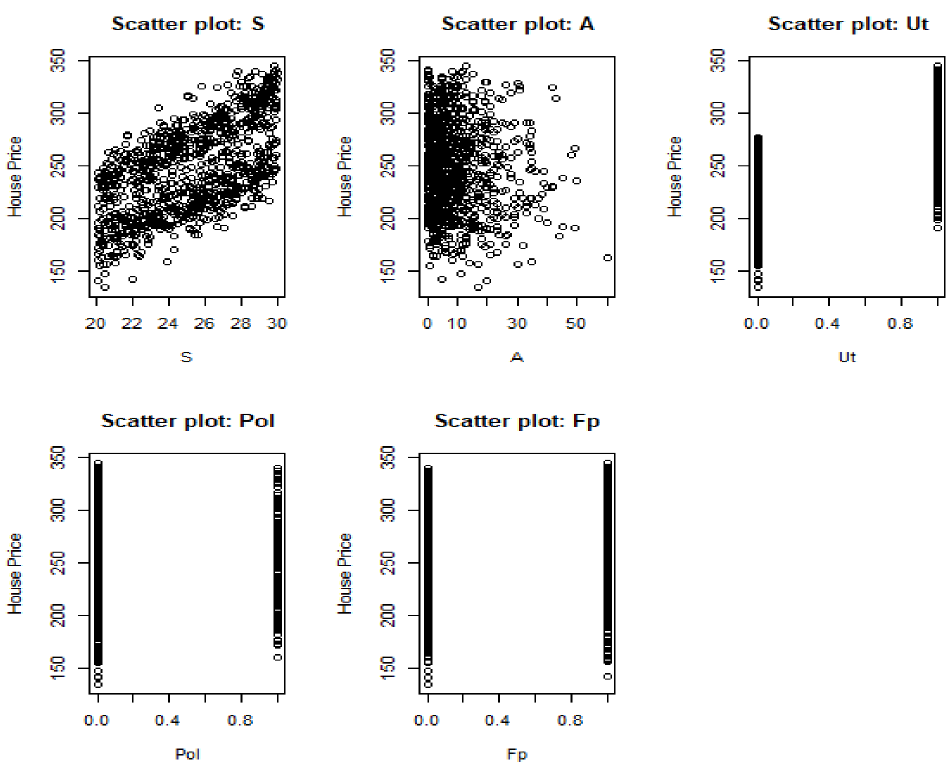
Firstly, the histogram for the feature price is in normal/gausian distribution.it depicts the average pricing for houses is 250.secondly the size of those houses are almost same.thirdly , the shape of the feature Age is positively skewed.as it gives us the information that only a few houses are very old which is nearly 60 years. It also shows that the urbanization started in that town within 10 years.

```
##################################### Boxplot ##########################

par(mfrow = c(2, 3))  # Adjust the numbers to arrange the boxplots in the desired layout

for (col in colnames(df)) {

  boxplot(df[[col]], main = col, xlab = col, ylab = "Value")

}
```



From this boxplot we see that for the feature Age the value 60 probably be detected as outlier.but all other features are free from outlier.

```
################################### correlation matrix ##################

cor_matrix <- cor(df)

print(cor_matrix)

################################### relation between price and all other variables #########

par(mfrow = c(2, 3))  # Adjust the numbers to arrange the plots in the desired layout

for (col in colnames(df)[-which(colnames(df) == "P")]) {

  plot(df[[col]], df$P, main = paste("Scatter plot:", col), xlab = col, ylab = "House Price")

}
```



From this plot we may interpret that there is a linear relationship between House price and the area of houses.on the other hand, the price is high for those houses which are built within nearly 10 years which has the intention to decrese with house maturity.for the most matured house (60 years) the price is noticeably low.

It is also seen that the price is high if the house is near the university .in contrast for other two feature pool and fireplace price is constant, that means price doesn’t matter wheather there is a pool and fireplace or not.

MODEL

library(car)
ml1=lm(P~S+Ut+I(S*Ut)+A+Pol+Fp+I(Pol*Fp),data=df)
ml1

```
Call:

lm(formula = P ~ S + Ut + I(S * Ut) + A + Pol + Fp + I(Pol *
    Fp), data = df)

Coefficients:
(Intercept)            S           Ut    I(S * Ut)            A
    24.2414       7.6163      27.4700       1.2985      -0.1894

        Pol           Fp  I(Pol * Fp)
     5.0623       1.9341      -1.4093
```

library(car)
vif_values<-vif(ml1)
vif_values

```
          S          Ut   I(S * Ut)           A         Pol
Fp
   2.208352   76.397898   78.041544    1.004610    1.952888    1
.270127

I(Pol * Fp)
   2.167511
```

Based on the provided **Variable Infletion Factor(VIF)** values for the variable , the VIF value of the Ut is 76.397898 which means the high level of correlation with other variables in the model.it potentially indicating multicollinearity.

Similarly, interection between living area and proximity to university the VIF value is 78.041544. It also indicate the high level of multicollinearity with other variable in the model.

plot(ml1)
ssrf=sum(resid(ml1)^2)
ssrf

```
230104.2
```

This represents the sum squared residual of the model is 230104.2

ml2=lm(P~S+A,data=df)
summary(ml2)

```
lm(formula = P ~ S + A, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-79.017 -30.271   3.693  30.292  72.867 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  34.2309     9.4052    3.64 0.000287 ***
S             8.5723     0.3671   23.35  < 2e-16 ***
A            -0.2853     0.1136   -2.51 0.012228 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.85 on 997 degrees of freedom
Multiple R-squared:  0.3577,	Adjusted R-squared:  0.3564 
F-statistic: 277.6 on 2 and 997 DF,  p-value: < 2.2e-16
```

The intercept 34.2309 represents the estimated house price when both the area and Age are 0.

Coeff of **S (8.8523**) indicates that for every additional unit increase in the living area the estimated house price increase by 8.5723.so an increase in the area is associated with higher estimated house price.

In contrast, Coeff of **A (-0.2853)** indicates that for every additional year of age of the house, the estimat ed house price decrease by 0.2853.assumming all other variables in the model is constant. This means th at older houses tend to have lower estimated price.

all coefficients (Intercept, S, and A) have p-values less than 0.05, indicating that they are statistically sign ificant at a 5% significance level.
The Intercept and S has extremely low p-values (< 0.001), indicating highly significant relationships with house prices. The coefficient for A also has a relatively low p-value (0.012), suggesting a statistically sign ificant relationship, although it is less significant compared to the Intercept and S.
The F-statistic of 277.6 with a very low p-value (< 2.2e-16) indicates that the overall model is statistically significant

ssrf2=sum(resid(ml2)^2)
ssrf2

```
[1] 1142293
```

ml3=lm(P~S+Ut+I(S*Ut)+A+Pol+Fp,data=df)
summary(ml3)

```
Call:

lm(formula = P ~ S + Ut + I(S * Ut) + A + Pol + Fp, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.289 -10.141   0.148  10.565  44.783 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  24.5000     6.1917   3.957 8.13e-05 ***
S             7.6122     0.2452  31.048  < 2e-16 ***
Ut           27.4530     8.4226   3.259 0.001154 ** 
I(S * Ut)     1.2994     0.3321   3.913 9.72e-05 ***
A            -0.1901     0.0512  -3.712 0.000217 ***
Pol           4.3772     1.1967   3.658 0.000268 ***
Fp            1.6492     0.9720   1.697 0.090056 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.23 on 993 degrees of freedom
Multiple R-squared:  0.8706,	Adjusted R-squared:  0.8698 
F-statistic:  1113 on 6 and 993 DF,  p-value: < 2.2e-16
```

The multiple R-squared value of 0.8706 indicates that approximately 87.06% of the variability in house p rices can be explained by the independent variables.
The F-statistic of 1113 with a very low p-value (< 2.2e-16) suggests that the overall model is statistically s ignificant

The residual standard error (15.23) represents the estimate of the standard deviation of the errors or re siduals

Conclusion:

Overall, this project suggests that factors such as the square footage of living area, proximity to the univ ersity, age of the house, and the presence of a pool are important determinants of house prices.from th ese variables only the age(A) of he house has negative significance. Understanding these variables can h elp buyers, sellers, and real estate professionals make informed decisions regarding pricing and investm ent strategies