**Case Study: Terro’s Real Estate Agency**

1. **The first step to any project is understanding the data. So for this step, generate the summary**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
|  |  |  |  |  |  |  |  |  |  |  |
| Mean | 4.871976285 | 68.575 | 11.1368 | 0.5547 | 9.549407 | 408.2 | 18.4555 | 6.28463439 | 12.6531 | 22.53280632 |
| Standard Error | 0.129860152 | 1.2514 | 0.30498 | 0.00515 | 0.387085 | 7.492 | 0.09624 | 0.03123514 | 0.31746 | 0.408861147 |
| Median | 4.82 | 77.5 | 9.69 | 0.538 | 5 | 330 | 19.05 | 6.2085 | 11.36 | 21.2 |
| Mode | 3.43 | 100 | 18.1 | 0.538 | 24 | 666 | 20.2 | 5.713 | 8.05 | 50 |
| Standard Deviation | 2.921131892 | 28.149 | 6.86035 | 0.11588 | 8.707259 | 168.5 | 2.16495 | 0.70261714 | 7.14106 | 9.197104087 |
| Sample Variance | 8.533011532 | 792.36 | 47.0644 | 0.01343 | 75.81637 | 28405 | 4.68699 | 0.49367085 | 50.9948 | 84.58672359 |
| Kurtosis | -1.18912246 | -0.9677 | -1.2335 | -0.0647 | -0.867232 | -1.142 | -0.28509 | 1.89150037 | 0.49324 | 1.495196944 |
| Skewness | 0.021728079 | -0.599 | 0.29502 | 0.72931 | 1.004815 | 0.67 | -0.80232 | 0.40361213 | 0.90646 | 1.108098408 |
| Range | 9.95 | 97.1 | 27.28 | 0.486 | 23 | 524 | 9.4 | 5.219 | 36.24 | 45 |
| Minimum | 0.04 | 2.9 | 0.46 | 0.385 | 1 | 187 | 12.6 | 3.561 | 1.73 | 5 |
| Maximum | 9.99 | 100 | 27.74 | 0.871 | 24 | 711 | 22 | 8.78 | 37.97 | 50 |
| Sum | 2465.22 | 34699 | 5635.21 | 280.676 | 4832 | 2E+05 | 9338.5 | 3180.025 | 6402.45 | 11401.6 |
| Count | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 | 506 |

The average value of owner occupied houses (AVG\_Price) in the data set is 22.53 (‘000 USD), and the median value is 21.20.The range of values is between 5 to 50. The data has a bit of skewness.The average AGE is 68.5 and the median age is 77.5 suggesting negative skewness in this variable. Mean of AVG\_Rooms is 6.28 and median is 6.2 suggesting that this variable could be normally distributed (more analysis would be required to know the exact picture). Most frequent value of AVG\_ROOMS is 5.7

1. **Plot the histogram of the Avg\_Price Variable. What do you infer?**

The histogram shows that the variable, AVG\_PRICE Is positively skewed as there is a tail towards the right, indicating the presence of some very high values in the data. Most of the houses are priced in the 17.88 to 24.49 bracket (prices are in ‘000 USD)

1. **Compute the covariance matrix. Share your observations.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
| CRIME\_RATE | 8.516147873 |  |  |  |  |  |  |  |  |  |
| AGE | 0.562915215 | 790.792 |  |  |  |  |  |  |  |  |
| INDUS | -0.11021518 | 124.268 | 46.9714 |  |  |  |  |  |  |  |
| NOX | 0.000625308 | 2.38121 | 0.60587 | 0.0134 |  |  |  |  |  |  |
| DISTANCE | -0.22986049 | 111.55 | 35.4797 | 0.6157 | 75.666531 |  |  |  |  |  |
| TAX | -8.22932244 | 2397.94 | 831.713 | 13.021 | 1333.1167 | 28349 |  |  |  |  |
| PTRATIO | 0.068168906 | 15.9054 | 5.68085 | 0.0473 | 8.7434025 | 167.82 | 4.677726 |  |  |  |
| AVG\_ROOM | 0.056117778 | -4.7425 | -1.88423 | -0.025 | -1.281277 | -34.52 | -0.53969 | 0.492695216 |  |  |
| LSTAT | -0.88268036 | 120.838 | 29.5218 | 0.488 | 30.325392 | 653.42 | 5.7713 | -3.073654967 | 50.894 |  |
| AVG\_PRICE | 1.16201224 | -97.396 | -30.4605 | -0.455 | -30.50083 | -724.8 | -10.0907 | 4.484565552 | -48.352 | 84.4195562 |

CRIME\_RATE and AVG\_PRICE, AVG\_ROOM and AVG\_PRICE are positively related, rest all variables are negatively related with AVG\_PRICE

1. **Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |
| CRIME\_RATE | 1 |  |  |  |  |  |  |  |  |  |
| AGE | 0.006859463 | 1 |  |  |  |  |  |  |  |  |
| INDUS | -0.005510651 | 0.64478 | 1 |  |  |  |  |  |  |  |
| NOX | 0.001850982 | 0.73147 | 0.76365 | 1 |  |  |  |  |  |  |
| DISTANCE | -0.009055049 | 0.45602 | 0.59513 | 0.61144 | 1 |  |  |  |  |  |
| TAX | -0.016748522 | 0.50646 | 0.72076 | 0.66802 | 0.910228 | 1 |  |  |  |  |
| PTRATIO | 0.010800586 | 0.26152 | 0.38325 | 0.18893 | 0.464741 | 0.46085 | 1 |  |  |  |
| AVG\_ROOM | 0.02739616 | -0.24026 | -0.39168 | -0.3022 | -0.209847 | -0.29205 | -0.355501 | 1 |  |  |
| LSTAT | -0.042398321 | 0.60234 | 0.6038 | 0.59088 | 0.488676 | 0.54399 | 0.3740443 | -0.61380827 | 1 |  |
| AVG\_PRICE | 0.043337871 | -0.37695 | -0.48373 | -0.4273 | -0.381626 | -0.46854 | -0.507787 | 0.695359947 | -0.73766 | 1 |

The red cells suggest high positive correlations. Top 3 positively correlated pairs –TAX and Distance (0.891), NOX and INDUS (0.76) and NOX and AGE (0.73). Top 3 negative correlations –LSTAT and AVG\_Price (-0.74), LSTAT and AVG\_ROOM (-0.61) and PTRATIO and AVG\_PRICE (-0.51)

5. **Build an initial regression model with AVG\_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too.**

1. *What do you infer from the Regression Summary Output in terms of variance explained, coefficient value, Intercept and the Residual plot?*
2. *Is LSTAT variable significant for the analysis based on your model?*

|  |  |
| --- | --- |
| SUMMARY OUTPUT |  |
|  |  |
| *Regression Statistics* | |
| Multiple R | 0.737662726 |
| R Square | 0.544146298 |
| Adjusted R Square | 0.543241826 |
| Standard Error | 6.215760405 |
| Observations | 506 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 1 | 23243.914 | 23243.91 | 601.6179 | 5.0811E-88 |
| Residual | 504 | 19472.38142 | 38.63568 |  |  |
| Total | 505 | 42716.29542 |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 34.55384088 | 0.562627355 | 61.41515 | 3.7E-236 | 33.44845704 | 35.65922472 | 33.44845704 | 35.65922472 |
| LSTAT | -0.95004935 | 0.038733416 | -24.5279 | 5.08E-88 | -1.0261482 | -0.873950508 | -1.0261482 | -0.873950508 |

1. *What do you infer from the Regression Summary Output in terms of variance explained, coefficient value, Intercept and the Residual plot?*

The model has an R-squared value of . The intercept is 34.55 which suggests that even if LSTAT is 0, the value of AVG\_PRICE will be positive, i.e. 34.55. Looking at the residual plot, we see more concentration of points towards the lower values of LSTAT, visually it suggests that there might be a pattern here, so we should explore more models we could get a better model.

1. *Is LSTAT variable significant for the analysis based on your model?*

LSTAT has a significance value very close to 0, but it cannot be absolute 0. Since it is less than the significance level of 0.05, this variable LSTAT is significant and should be retained in our analysis

1. **Build another instance of the Regression model but this time including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as the dependent variable.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | | | | | |
|  | | |  | | |
| *Regression Statistics* | | | | | |
| Multiple R | | | 0.799100498 | | |
| R Square | | | 0.638561606 | | |
| Adjusted R Square | | | 0.637124475 | | |
| Standard Error | | | 5.540257367 | | |
| Observations | | | 506 | | |
| ANOVA | | | |  |  | |  |  |  | | |
|  | | | | *df* | *SS* | | *MS* | *F* | *Significance F* | | |
| Regression | | | | 2 | 27276.98621 | | 13638.49 | 444.3309 | 7.0085E-112 | | |
| Residual | | | | 503 | 15439.3092 | | 30.69445 |  |  | | |
| Total | | | | 505 | 42716.29542 | |  |  |  | | |
|  | *Coefficients* | *Standard Error* | | | *t Stat* | | *P-value* | *Lower 95%* | | | *Upper 95%* | | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -1.35827281 | 3.17282778 | | | -0.4281 | | 0.668765 | -7.591900282 | | 4.875354658 | | | -7.59190028 | 4.875354658 |
| AVG\_ROOM | 5.094787984 | 0.4444655 | | | 11.46273 | | 3.47E-27 | 4.221550436 | | 5.968025533 | | | 4.22155044 | 5.968025533 |
| LSTAT | -0.64235833 | 0.043731465 | | | -14.6887 | | 6.67E-41 | -0.728277167 | | -0.5564395 | | | -0.72827717 | -0.5564395 |

1. *Write the Regression equation. If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then what will be the value of AVG\_PRICE? How does it compare to the company quoting a value of 30000 USD for this locality? Is the company overcharging/ undercharging?*

Regression Equation

y = B1X1+ B2X2+ c

y = (5.09 \* X1) – (0.64 \* X2) -1.36

X1= 7 (AVG\_ROOM)

X2= 20 (LSTAT)

Y = (5.09 \* 7) – (0.64 \* 20) –1.36

Y = 35.63 –12.8 –1.36 = 21.47

The company is quoting a value of 30 against a prediction of 21.47, which suggests that the company is overcharging

1. *Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square. Explain.*

The R-squared value here is 0.64 as compared to 0.54 of the previous model, this means by adding AVG\_ROOMS to our existing model, we are able to capture additional 10% of the variance in AVG\_Price, because of which this is a better model than the previous one

1. **Now, build a Regression model with all variables. AVG\_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG\_price. Explain.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  | | |
|  | |  | | |
| *Regression Statistics* | | | | |
| Multiple R | | 0.832978824 | | |
| R Square | | 0.69385372 | | |
| Adjusted R Square | | 0.688298647 | | |
| Standard Error | | 5.1347635 | | |
| Observations | | 506 | | |
| ANOVA | |  | |  | | |  | |  | |  | |
|  | | *df* | | *SS* | | | *MS* | | *F* | | *Significance F* | |
| Regression | | 9 | | 29638.8605 | | | 3293.207 | | 124.9045 | | 1.9328E-121 | |
| Residual | | 496 | | 13077.43492 | | | 26.3658 | |  | |  | |
| Total | | 505 | | 42716.29542 | | |  | |  | |  | |
|  | *Coefficients* | | *Standard Error* | | | *t Stat* | | *P-value* | | *Lower 95%* | | *Upper 95%* | | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 29.24131526 | | 4.817125596 | | | 6.0702829 | | 2.54E-09 | | 19.77682784 | | 38.70580267 | | 19.77682784 | 38.70580267 |
| CRIME\_RATE | 0.048725141 | | 0.078418647 | | | 0.6213464 | | 0.5346572 | | -0.105348544 | | 0.202798827 | | -0.105348544 | 0.202798827 |
| AGE | 0.032770689 | | 0.013097814 | | | 2.5019968 | | 0.0126704 | | 0.00703665 | | 0.058504728 | | 0.00703665 | 0.058504728 |
| INDUS | 0.130551399 | | 0.063117334 | | | 2.0683922 | | 0.0391209 | | 0.006541094 | | 0.254561704 | | 0.006541094 | 0.254561704 |
| NOX | -10.3211828 | | 3.894036256 | | | -2.6505102 | | 0.0082939 | | -17.97202279 | | -2.670342809 | | -17.97202279 | -2.670342809 |
| DISTANCE | 0.261093575 | | 0.067947067 | | | 3.8426026 | | 0.0001375 | | 0.127594012 | | 0.394593138 | | 0.127594012 | 0.394593138 |
| TAX | -0.01440119 | | 0.003905158 | | | -3.6877361 | | 0.0002512 | | -0.022073881 | | -0.0067285 | | -0.022073881 | -0.0067285 |
| PTRATIO | -1.074305348 | | 0.133601722 | | | -8.0411041 | | 6.586E-15 | | -1.336800438 | | -0.811810259 | | -1.336800438 | -0.811810259 |
| AVG\_ROOM | 4.125409152 | | 0.442758999 | | | 9.3175049 | | 3.893E-19 | | 3.255494742 | | 4.995323561 | | 3.255494742 | 4.995323561 |
| LSTAT | -0.603486589 | | 0.053081161 | | | -11.369129 | | 8.911E-27 | | -0.70777824 | | -0.499194938 | | -0.70777824 | -0.499194938 |

This particular model has an R-squared value of 0.6939 against a R-squared value of 0.64 in the previous model (with LSTAT and AVG\_ROOM), this model captures more variance as compared to the previous model. Also, here the adjusted R-square value is 0.6883 suggesting that the significant variables are contributing to 68.83% of the variance. The intercept value is 29.24, suggesting that even if all the independent variables were zero, the AVG\_PRICE would be 29.24. Looking at the p-values, CRIM\_RATE should be dropped as its p-value is more than 0.05. Rest all variables are significant.

1. **Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked. (HINT: Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  | | |
|  | |  | | |
| *Regression Statistics* | | | | |
| Multiple R | | 0.832835773 | | |
| R Square | | 0.693615426 | | |
| Adjusted R Square | | 0.688683682 | | |
| Standard Error | | 5.131591113 | | |
| Observations | | 506 | | |
| ANOVA | |  | |  | | |  | |  | |  | |
|  | | *df* | | *SS* | | | *MS* | | *F* | | *Significance F* | |
| Regression | | 8 | | 29628.68142 | | | 3703.585 | | 140.643 | | 1.911E-122 | |
| Residual | | 497 | | 13087.61399 | | | 26.33323 | |  | |  | |
| Total | | 505 | | 42716.29542 | | |  | |  | |  | |
|  | *Coefficients* | | *Standard Error* | | | *t Stat* | | *P-value* | | *Lower 95%* | | *Upper 95%* | | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 29.42847349 | | 4.804728624 | | | 6.1248982 | | 1.846E-09 | | 19.98838959 | | 38.8685574 | | 19.98838959 | 38.8685574 |
| AGE | 0.03293496 | | 0.013087055 | | | 2.516606 | | 0.0121629 | | 0.007222187 | | 0.058647734 | | 0.007222187 | 0.058647734 |
| INDUS | 0.130710007 | | 0.063077823 | | | 2.0722023 | | 0.0387617 | | 0.006777942 | | 0.254642071 | | 0.006777942 | 0.254642071 |
| NOX | -10.27270508 | | 3.890849222 | | | -2.6402218 | | 0.0085457 | | -17.9172457 | | -2.628164466 | | -17.9172457 | -2.628164466 |
| DISTANCE | 0.261506423 | | 0.067901841 | | | 3.851242 | | 0.0001329 | | 0.128096375 | | 0.394916471 | | 0.128096375 | 0.394916471 |
| TAX | -0.014452345 | | 0.003901877 | | | -3.7039464 | | 0.0002361 | | -0.022118553 | | -0.006786137 | | -0.022118553 | -0.006786137 |
| PTRATIO | -1.071702473 | | 0.133453529 | | | -8.0305293 | | 7.083E-15 | | -1.333905109 | | -0.809499836 | | -1.333905109 | -0.809499836 |
| AVG\_ROOM | 4.125468959 | | 0.44248544 | | | 9.3234005 | | 3.69E-19 | | 3.256096304 | | 4.994841615 | | 3.256096304 | 4.994841615 |
| LSTAT | -0.605159282 | | 0.0529801 | | | -11.422388 | | 5.418E-27 | | -0.70925186 | | -0.501066704 | | -0.70925186 | -0.501066704 |

**Answer the questions below:**

1. *Interpret the output of this model.*

This model explains 69.36% of the variance in AVG\_PRICE. The intercept value is 29.42 suggesting that if all independent variables are 0, then the value of the house would be 29.42. All variables are significant here. This model is acceptable as it has a decent R-square and all variables are significant.

1. *Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?*

Adjusted R-square for this model is 0.6887 vs. 0.6883 in the previous model. Although adjusted R-square value is not up drastically, but we have all significant variables here, so is we consider these two factors together, then this model is a better model than the previous one

1. *Sort the values of the Coefficients in ascending order. What will happen to the average price if the value of NOX is more in a locality in this town?*

|  |  |  |
| --- | --- | --- |
| Ascending order of coefficients | | |
|  |  |  |
|  | *Coefficients* |  |
| NOX | -10.2727051 |  |
| PTRATIO | -1.07170247 |  |
| LSTAT | -0.60515928 |  |
| TAX | -0.01445235 |  |
| AGE | 0.03293496 |  |
| INDUS | 0.130710007 |  |
| DISTANCE | 0.261506423 |  |
| AVG\_ROOM | 4.125468959 |  |
| Intercept | 29.42847349 |  |

NOX and AVG\_Price are negatively related. If the value of NOX increases then value of AVG\_PRICE falls, more specifically every 1-unitincrease in the value of NOX decreased the value of AVG\_PRICE by 10.27

1. *Write the regression equation from this model.*

Y = 29.4285 + 0.0329 \* X1+ 0.1307 \* X2+ -10.2727 \* X3+ 0.2615 \* X4–0.0145 \* X5–1.0717 \* X6+ 4.1255 \* X7–0.6052 \* X8