**ASSIGNMENT – TERRO’S REAL ESTATE AGENCY**

Reported by

Meenatchi

**Problem Statement (Situation): “Finding out the most relevant features for pricing of a house”**

Terro’s real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an “Auditor”, who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property.

**Objective (Task):**

**Q1. Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation**.

>>to generate the summary statistics for each variable in table, I used data analysis tool pack.

>>In tool pack we should select the descriptive statistics and select the arrays which we want to find the summary statistics.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *CRIME\_RATE* |  | *AGE* |  | *INDUS* |  | *NOX* |  |
|  |  |  |  |  |  |  |  |
| Mean | 4.871976 | Mean | 68.5749 | Mean | 11.13678 | Mean | 0.554695 |
| Standard Error | 0.12986 | Standard Error | 1.25137 | Standard Error | 0.30498 | Standard Error | 0.005151 |
| Median | 4.82 | Median | 77.5 | Median | 9.69 | Median | 0.538 |
| Mode | 3.43 | Mode | 100 | Mode | 18.1 | Mode | 0.538 |
| Standard Deviation | 2.921132 | Standard Deviation | 28.14886 | Standard Deviation | 6.860353 | Standard Deviation | 0.115878 |
| Sample Variance | 8.533012 | Sample Variance | 792.3584 | Sample Variance | 47.06444 | Sample Variance | 0.013428 |
| Kurtosis | -1.18912 | Kurtosis | -0.96772 | Kurtosis | -1.23354 | Kurtosis | -0.06467 |
| Skewness | 0.021728 | Skewness | -0.59896 | Skewness | 0.295022 | Skewness | 0.729308 |
| Range | 9.95 | Range | 97.1 | Range | 27.28 | Range | 0.486 |
| Minimum | 0.04 | Minimum | 2.9 | Minimum | 0.46 | Minimum | 0.385 |
| Maximum | 9.99 | Maximum | 100 | Maximum | 27.74 | Maximum | 0.871 |
| Sum | 2465.22 | Sum | 34698.9 | Sum | 5635.21 | Sum | 280.6757 |
| Count | 506 | Count | 506 | Count | 506 | Count | 506 |

**Insights**

* The age of the house represent how old is that house and if that house in good condition. By the seeing statistics we can say that most of the houses are old. The mode is 100 so there are many houses which are 100 years old.
* The average industry acres per town is 11. The percentage of industry in town can increase the house price in that town.

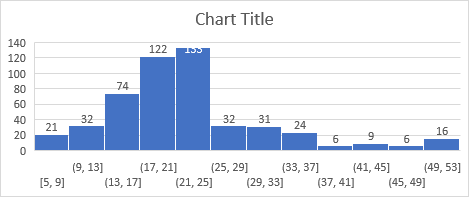
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *DISTANCE* |  | *TAX* |  | *PTRATIO* |  |
|  |  |  |  |  |  |
| Mean | 9.549407115 | Mean | 408.2371542 | Mean | 18.4555336 |
| Standard Error | 0.387084894 | Standard Error | 7.492388692 | Standard Error | 0.096243568 |
| Median | 5 | Median | 330 | Median | 19.05 |
| Mode | 24 | Mode | 666 | Mode | 20.2 |
| Standard Deviation | 8.707259384 | Standard Deviation | 168.5371161 | Standard Deviation | 2.164945524 |
| Sample Variance | 75.81636598 | Sample Variance | 28404.75949 | Sample Variance | 4.686989121 |
| Kurtosis | -0.867231994 | Kurtosis | -1.142407992 | Kurtosis | -0.285091383 |
| Skewness | 1.004814648 | Skewness | 0.669955942 | Skewness | -0.802324927 |
| Range | 23 | Range | 524 | Range | 9.4 |
| Minimum | 1 | Minimum | 187 | Minimum | 12.6 |
| Maximum | 24 | Maximum | 711 | Maximum | 22 |
| Sum | 4832 | Sum | 206568 | Sum | 9338.5 |
| Count | 506 | Count | 506 | Count | 506 |

* The average distance (away from the highway) is 9.5 miles. The mode is 24, so there are many houses which are away from highway. The nearest house should have high price.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *AVG\_ROOM* |  | *LSTAT* |  | *AVG\_PRICE* |  |
|  |  |  |  |  |  |
| Mean | 6.284634387 | Mean | 12.65306324 | Mean | 22.53280632 |
| Standard Error | 0.031235142 | Standard Error | 0.317458906 | Standard Error | 0.408861147 |
| Median | 6.2085 | Median | 11.36 | Median | 21.2 |
| Mode | 5.713 | Mode | 8.05 | Mode | 50 |
| Standard Deviation | 0.702617143 | Standard Deviation | 7.141061511 | Standard Deviation | 9.197104087 |
| Sample Variance | 0.49367085 | Sample Variance | 50.99475951 | Sample Variance | 84.58672359 |
| Kurtosis | 1.891500366 | Kurtosis | 0.493239517 | Kurtosis | 1.495196944 |
| Skewness | 0.403612133 | Skewness | 0.906460094 | Skewness | 1.108098408 |
| Range | 5.219 | Range | 36.24 | Range | 45 |
| Minimum | 3.561 | Minimum | 1.73 | Minimum | 5 |
| Maximum | 8.78 | Maximum | 37.97 | Maximum | 50 |
| Sum | 3180.025 | Sum | 6402.45 | Sum | 11401.6 |
| Count | 506 | Count | 506 | Count | 506 |

* The town has average of 6 rooms per house. The house has minimum 3 rooms per house.
* Average price increases when the house has more rooms.

**Q2. Plot a histogram of the Avg-Price variable. What do you infer?**



* More than 250 houses have average price between 17 to 25.
* Most of the houses average price below 25. There are only few houses have high average price.

**Q3. Compute the covariance matrix. Share your observations.**

**>>** To compute the covariance matrix, I used data analysis tool pack.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_RATE* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_ROOM* | *LSTAT* | *AVG\_PRICE* |  |
| CRIME\_RATE | 8.516148 |  |  |  |  |  |  |  |  |  |  |
| AGE | 0.562915 | 790.7925 |  |  |  |  |  |  |  |  |  |
| INDUS | -0.11022 | 124.2678 | 46.97143 |  |  |  |  |  |  |  |  |
| NOX | 0.000625 | 2.381212 | 0.605874 | 0.013401 |  |  |  |  |  |  |  |
| DISTANCE | -0.22986 | 111.55 | 35.47971 | 0.61571 | 75.66653 |  |  |  |  |  |  |
| TAX | -8.22932 | 2397.942 | 831.7133 | 13.0205 | 1333.117 | 28348.62 |  |  |  |  |  |
| PTRATIO | 0.068169 | 15.90543 | 5.680855 | 0.047304 | 8.743402 | 167.8208 | 4.677726 |  |  |  |  |
| AVG\_ROOM | 0.056118 | -4.74254 | -1.88423 | -0.02455 | -1.28128 | -34.5151 | -0.53969 | 0.492695 |  |  |  |
| LSTAT | -0.88268 | 120.8384 | 29.52181 | 0.48798 | 30.32539 | 653.4206 | 5.7713 | -3.07365 | 50.89398 |  |  |
| AVG\_PRICE | 1.162012 | -97.3962 | -30.4605 | -0.45451 | -30.5008 | -724.82 | -10.0907 | 4.484566 | -48.3518 | 84.41956 |  |

* Average price has positive relationship with average room per house.
* Average price increases when the house has more rooms.
* Other than Avg\_room and crime\_rate, all variables have negative relationship with Avg\_price.
* If the house is old and the distance from highway is high, then the price of the house has to be low.

**Q4.** **Create a correlation matrix of all the variables (Use Data analysis tool pack).**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *CRIME\_* | *AGE* | *INDUS* | *NOX* | *DISTANCE* | *TAX* | *PTRATIO* | *AVG\_R* | *LSTAT* | *AVG\_P* |
| CRIME\_RATE | 1 |  |  |  |  |  |  |  |  |  |
| AGE | 0.006859 | 1 |  |  |  |  |  |  |  |  |
| INDUS | -0.00551 | 0.644779 | 1 |  |  |  |  |  |  |  |
| NOX | 0.001851 | 0.73147 | 0.763651 | 1 |  |  |  |  |  |  |
| DISTANCE | -0.00906 | 0.456022 | 0.595129 | 0.611441 | 1 |  |  |  |  |  |
| TAX | -0.01675 | 0.506456 | 0.72076 | 0.668023 | 0.910228 | 1 |  |  |  |  |
| PTRATIO | 0.010801 | 0.261515 | 0.383248 | 0.188933 | 0.464741 | 0.460853 | 1 |  |  |  |
| AVG\_ROOM | 0.027396 | -0.24026 | -0.39168 | -0.30219 | -0.20985 | -0.29205 | -0.3555 | 1 |  |  |
| LSTAT | -0.0424 | 0.602339 | 0.6038 | 0.590879 | 0.488676 | 0.543993 | 0.374044 | -0.6138 | 1 |  |
| AVG\_PRICE | 0.043338 | -0.37695 | -0.48373 | -0.42732 | -0.38163 | -0.46854 | -0.50779 | 0.69536 | -0.7376 | 1 |

1. **Which are the top 3 positively correlated pairs?**

* The Tax and distance have highest positive corelation with 0.91 corelation.
* The second highest corelation is Industry and Nitric oxides concentration with o.76 corelation.
* The third highest corelated pair is Age and NOX with 0.73.

1. **Which are the top 3 negatively correlated pairs?**

* The LSTAT and AVG\_PRICE have lowest negatived corelation with -0.73.
* The second lowest corelated is LSTA and average room with -0.61.
* The third lowest corelated is AVG\_PRICE and PRATIO with -0.5.

**Q5. Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.737663 |  |  |  |  |  |  |  |
| R Square | 0.544146 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.543242 |  |  |  |  |  |  |  |
| Standard Error | 6.21576 |  |  |  |  |  |  |  |
| Observations | 506 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 23243.91 | 23243.91 | 601.6179 | 5.08E-88 |  |  |  |
| Residual | 504 | 19472.38 | 38.63568 |  |  |  |  |  |
| Total | 505 | 42716.3 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 34.55384 | 0.562627 | 61.41515 | 3.7E-236 | 33.44846 | 35.65922 | 33.44846 | 35.65922 |
| LSTAT | -0.95005 | 0.038733 | -24.5279 | 5.08E-88 | -1.02615 | -0.87395 | -1.02615 | -0.87395 |

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

* The p-value is very close to zero (5.08E-88), indicating that the regression model is statistically significant. We could typically reject the null hypothesis that all regression coefficients are zero.
* The coefficient value is 34.55 so it suggests that for each unit increase in the LSTAT variable, the AVG\_PRICE variable is expected to increase by 34 units.
* The intercept of 3.7E-236 is extremely small number. The combination of a large coefficient and an extremely small intercept should be scrutinized. It might indicate issues such as collinearity, outliers, or other problems in the data.
* The data have randomly separated residual plot, It represents the average contribution to unexplained error.

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

* Yes. The p-value associated with the F-statistic. In this case, the p-value is very close to zero (5.08E-88), indicating that the regression model is statistically significant.

**Q6. Build a new Regression model including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as dependent variable.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.973885 |  |  |  |  |  |  |  |
| R Square | 0.948453 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.946366 |  |  |  |  |  |  |  |
| Standard Error | 5.535767 |  |  |  |  |  |  |  |
| Observations | 506 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 2 | 284181.4 | 142090.7 | 4636.712 | 0 |  |  |  |
| Residual | 504 | 15444.93 | 30.64471 |  |  |  |  |  |
| Total | 506 | 299626.3 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0 | #N/A | #N/A | #N/A | #N/A | #N/A | #N/A | #N/A |
| AVG\_ROOM | 4.906906 | 0.070193 | 69.90558 | 1.6E-261 | 4.768998 | 5.044814 | 4.768998 | 5.044814 |
| LSTAT | -0.65574 | 0.030559 | -21.4585 | 4.81E-73 | -0.71578 | -0.5957 | -0.71578 | -0.5957 |
|  |  |  |  |  |  |  |  |  |

**a) 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: AVG PRICE=coefficient of Avg\_room \* ROOM + coefficient of LSTAT \* LSTAT
* AVG\_PRICE = 4.9\*7-0.65\*20 = 21.3
* The average price of the house which has 7 rooms and has a value of 20 for L\_STAT is 21300 USD. But the company quoting a value of 30000 USD which is **Overcharging.**

**b) Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain.**

* The performance of this model better than the previous model. The adjusted R-square of previous model was 0.54 and this model’s adjusted R-square is 0.94. This model’s R-square is higher than the previous one.

**Q7. Build another Regression model with all variables where AVG\_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted R-square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.832979 |  |  |  |  |  |  |  |
| R Square | 0.693854 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.688299 |  |  |  |  |  |  |  |
| Standard Error | 5.134764 |  |  |  |  |  |  |  |
| Observations | 506 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 9 | 29638.86 | 3293.207 | 124.9045049 | 1.9328E-121 |  |  |  |
| Residual | 496 | 13077.43 | 26.3658 |  |  |  |  |  |
| Total | 505 | 42716.3 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 29.24132 | 4.817126 | 6.070283 | 2.53978E-09 | 19.77682784 | 38.7058 | 19.77683 | 38.7058 |
| CRIME\_RATE | 0.048725 | 0.078419 | 0.621346 | 0.534657201 | -0.105348544 | 0.202799 | -0.10535 | 0.202799 |
| AGE | 0.032771 | 0.013098 | 2.501997 | 0.012670437 | 0.00703665 | 0.058505 | 0.007037 | 0.058505 |
| INDUS | 0.130551 | 0.063117 | 2.068392 | 0.03912086 | 0.006541094 | 0.254562 | 0.006541 | 0.254562 |
| NOX | -10.3212 | 3.894036 | -2.65051 | 0.008293859 | -17.97202279 | -2.67034 | -17.972 | -2.67034 |
| DISTANCE | 0.261094 | 0.067947 | 3.842603 | 0.000137546 | 0.127594012 | 0.394593 | 0.127594 | 0.394593 |
| TAX | -0.0144 | 0.003905 | -3.68774 | 0.000251247 | -0.022073881 | -0.00673 | -0.02207 | -0.00673 |
| PTRATIO | -1.07431 | 0.133602 | -8.0411 | 6.58642E-15 | -1.336800438 | -0.81181 | -1.3368 | -0.81181 |
| AVG\_ROOM | 4.125409 | 0.442759 | 9.317505 | 3.89287E-19 | 3.255494742 | 4.995324 | 3.255495 | 4.995324 |
| LSTAT | -0.60349 | 0.053081 | -11.3691 | 8.91071E-27 | -0.70777824 | -0.49919 | -0.70778 | -0.49919 |

* Regression of all variables as independent and avg\_price as dependent variable has high coefficient and low p-value.
* The p-value is very close to zero, indicating that the regression model is statistically significant. We could typically reject the null hypothesis that all regression coefficients are zero.
* The adjusted R-square of this model is 0.68. which is considered as low.
* All variables except Crime\_rate have low p-value, which makes overall regression model is significant.
* By that LSTAT, PTRatio and AVG\_ROOM variables have high significance.
* AGE,INDUS,NOX,Distance and Tax have low significance.
* They are all reject the Null hypothesis.

**Q8. 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 and answer the questions below:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.976229 |  |  |  |  |  |  |  |
| R Square | 0.953023 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.950355 |  |  |  |  |  |  |  |
| Standard Error | 5.316393 |  |  |  |  |  |  |  |
| Observations | 506 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 8 | 285550.9 | 35693.86 | 1262.872 | 0 |  |  |  |
| Residual | 498 | 14075.49 | 28.26403 |  |  |  |  |  |
| Total | 506 | 299626.3 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0 | #N/A | #N/A | #N/A | #N/A | #N/A | #N/A | #N/A |
| AGE | 0.011474 | 0.013063 | 0.878344 | 0.380181 | -0.01419 | 0.03714 | -0.01419 | 0.03714 |
| INDUS | 0.049527 | 0.06389 | 0.77519 | 0.438595 | -0.076 | 0.175055 | -0.076 | 0.175055 |
| NOX | 1.282534 | 3.525404 | 0.363798 | 0.716163 | -5.64396 | 8.209033 | -5.64396 | 8.209033 |
| DISTANCE | 0.094786 | 0.064447 | 1.470754 | 0.141989 | -0.03184 | 0.221409 | -0.03184 | 0.221409 |
| TAX | -0.01016 | 0.003977 | -2.55407 | 0.010944 | -0.01797 | -0.00234 | -0.01797 | -0.00234 |
| PTRATIO | -0.49054 | 0.097223 | -5.0455 | 6.35E-07 | -0.68156 | -0.29952 | -0.68156 | -0.29952 |
| AVG\_ROOM | 6.206434 | 0.293687 | 21.13282 | 3.08E-71 | 5.629416 | 6.783452 | 5.629416 | 6.783452 |
| LSTAT | -0.49745 | 0.051776 | -9.60773 | 3.65E-20 | -0.59918 | -0.39573 | -0.59918 | -0.39573 |

* 1. **Interpret the output of this model.**
* The multiple regression model is highly significant (p-value < 0.05), suggesting that at least one of the independent variables is associated with the dependent variable.
* The model explains a substantial portion of the variance in the dependent variable .
* The adjusted R-squared is close to the R-squared value, indicating that the inclusion of multiple predictors does not significantly decrease the goodness of fit.
  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?**
* This model’s R-square is 0.95, which is considered as a good R-Square.
* By selecting significant variables, we have good model which has high significance and R-square.
* This model performs better according to the value of adjusted R-square.

**c) 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?**

|  |  |
| --- | --- |
| **Variable** | **Coefficient** |
| AVG\_ROOM | 6.2064338 |
| NOX | 1.282534479 |
| DISTANCE | 0.094786238 |
| INDUS | 0.049527169 |
| AGE | 0.011474131 |
| TAX | -0.01015638 |
| PTRATIO | -0.4905384 |
| LSTAT | -0.4974544 |

* If the value of NOX is more in a locality in this town, the average price will be more too. Because NOX have positive relationship with a AVG\_Price.
  1. **Write the regression equation from this model.**
* Regression equation

= (6.2 \* avg\_room) +( 1.28\*NOX) + (0.09\*Distance) + (0.04\*INDUS) + (0.01\*Age)

+ (-0.01\*TAX) + (-0.49\*PTRATIO) + (-0.49\*LSTAT)