**EXPLORATORY DATA ANALYSIS PROJECT**

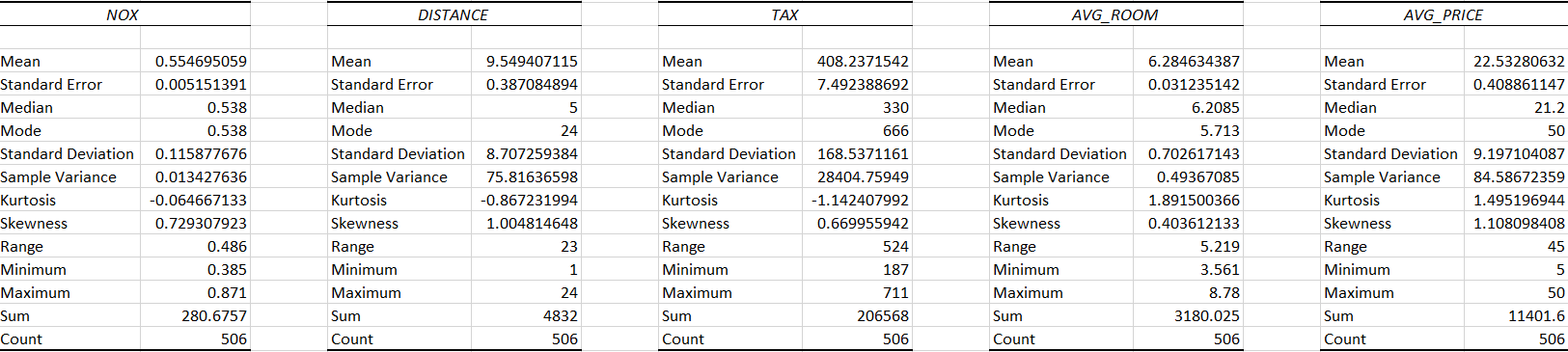
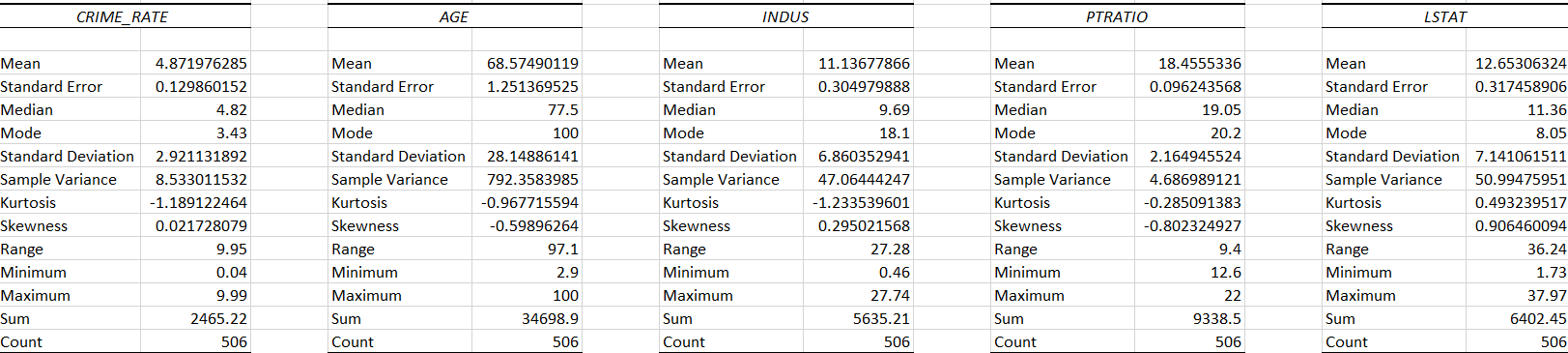
**PROJECT TITLE: TERRO’S REAL ESTATE AGENCY**

**BY**

**MYADAM.SAI SUDHA**

**GLCA-MAR 2023**

1. **Observations:**

****

1. Crime Rate: Out of 506 Observations we have derived that the average crime rate of town is (4.87).
2. Industry: Average of 11.13% of Non-retail business is existing per acre.
3. NOX: In the complete observation we have Minimum of 0.385 and Maximum of 0.871 of NOX Concentration in environment in the town.
4. Average Room: In the 506 observations, we have average of 6 rooms and minimum of 3 rooms per each flat.
5. Age: Average age of observations or buildings that are built after the year 1940 is derived as 68.57%
6. Distance: Distance to the highways from the purchased venture will be the key factor for the customer in terms of purchasing property.

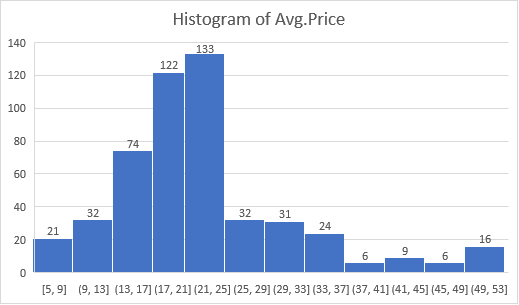
* In terms of these 506 observations, we have the distance to highways are within the range 23 miles and average of 9.54 miles.

1. Tax: The average tax of the properties is evaluated as 408.237 in the town.
2. PT-Ratio: PT-Ratio is nothing but people to teacher ratio which basically states the availability of teachers to the total existing students.

* As per the observations made, we had evaluated that average of 18.45 teachers are available to the total students.

1. LSTAT: On an average 12% of population has lower status.
2. Average-Price: The average price of the properties has been observed 22,530$, where minimum and maximum values are 5000$ and 50,000$

**2) Plot a histogram of the Average Price variable. What do you infer?**



1)The average price range is started from $5.

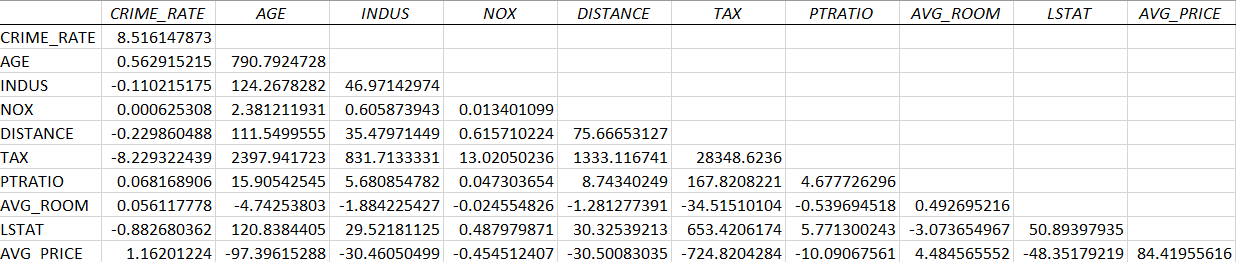
2)The most average price range falls ($21,$25)

3)The average price range started from $5 value 21 after that it rapidly increased to $25 value 133.But after 133 value suddenly decreasing.

4)The distribution appears to be slightly right skewed with a relatively longer tail on the right side.

Note: The values shown on x-axis of the histogram has to take in term’s of 1000’s $ as mentioned in the attachment.

**3) Compute the covariance matrix. Share your observations?**



**Covariance:**

1)In terms of statistics covariance is nothing but finding relationship between two or more variables.

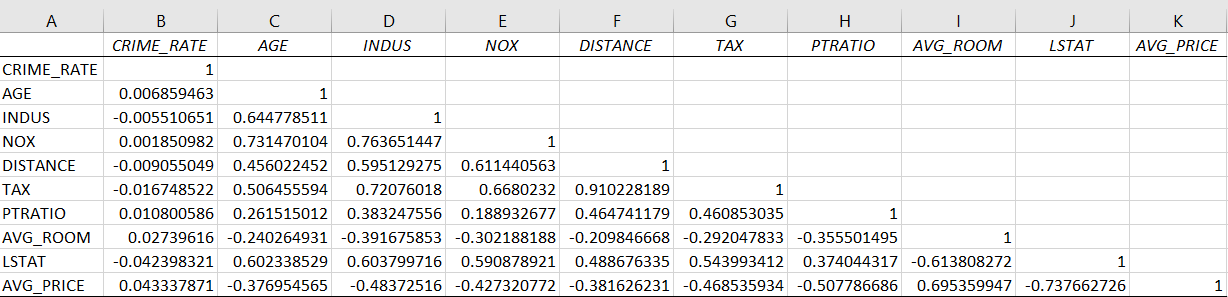
2)If the output value of 2 variables is positive then it is said to be both the variables are travelling in same direction, which means the variables are exhibiting similar behavior which is said to be positive covariance.

3)If the output value of two variables is negative then it is said to be both the variables are travelling in the opposite direction and exhibiting different behavior which is said to be negative covariance.

4)In the above generated covariance matrix TAX variable has positively strong covariance with all variables except for CRIME\_RATE variable which has negative covariance.

5)Tax with Age variable has high variance in positive side.

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

****

**Correlation: It is a statistical measure which expresses the extension to which the variables are linearly inclined to each other and the value lies between -1 to +1.**

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

* From the above generated correlation matrix output we have 3 strongly correlated variables as mentioned below.

1)Distance – Tax

2)NOX – Age

3)NOX – Indus

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

* We have 3 negatively correlated pairs as mentioned below

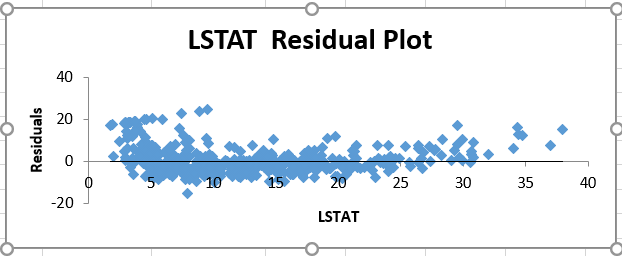
1)LSTAT – Avg-Room

2)Avg-Price – PTRATIO

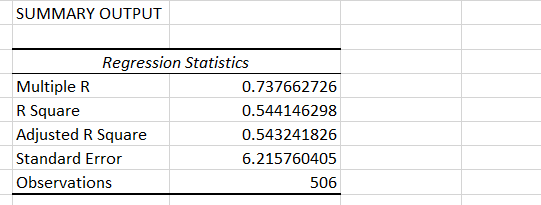
3)Avg-Price – LSTAT

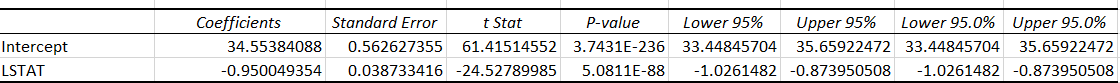
**5) Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.**

**Generated Output:**

****

1. The R-squared value is 0.544.
2. which means approximately 54.4% of the variance in AVG\_PRICE is explained by the independent variable (LSTAT).
3. This indicates a moderate level of variance explained.
4. Adjusted R-squared: The adjusted R-squared value is 0.544,
5. which is slightly lower than the R-squared value.
6. The model's goodness of fit.
7. **What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?**

****

****

1. R-squared (Multiple R-squared): The R-squared value is 0.544.
2. which means approximately 54.4% of the variance in AVG\_PRICE is explained by the independent variable (LSTAT).
3. This indicates a moderate level of variance explained.
4. Adjusted R-squared: The adjusted R-squared value is 0.544,

which is slightly lower than the R-squared value.

1. The model's goodness of fit.
2. From the above generated output, we see that there is 54% of the variation in the average price can be explained by the LSTAT.
3. The coefficient of LSTAT for this model is -0.950049354.
4. Intercept of LSTAT for this model is 34.55384088.

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

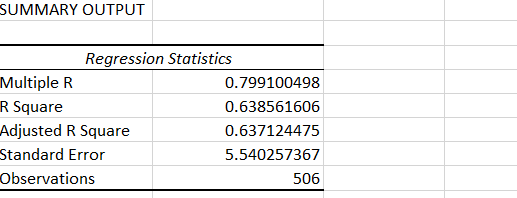
1)Yes, LSTAT is the significant variable of the avg-price for this model.

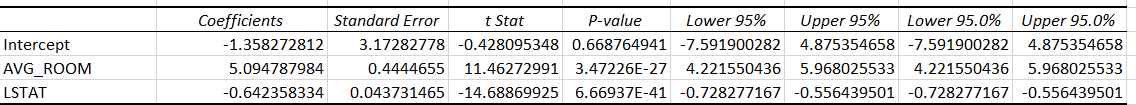
2)As the p-value (5.08E-88) we obtained from this model is very less than 0.05.

and we already know the p-value is less than 0.05 it was significant to target variable.

3)By this way we can say that LSTAT is a significant variable according to this model.

**6) Build a new Regression model including LSTAT and AVG\_ROOM together as independent variables and AVG\_PRICE as dependent variable.**

****

****

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?**

**Calculation:**

A) The regression equation Y= A+BX

Avg-price= -3.37+5.094\*avg\_room-0.642\*LSTAT

* 7 rooms and LSTAT value of 20 substitute value into the equation.
* AVG-Price = -3.37+5.09\*7-0.642\*20=21.472
* The predicted value of avg \_price for this house is 19.42.
* Which is lesser then the company’s quoted value of 30000.

1. **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 model with Avg-room, LSTAT is better than the previous model of only LSTAT.
* Because we clearly observed and understand that r square value of this model was 0.6385 it is greater than previous model with r square value of 0.5441.
* so, this model is more significant and have more variability.

**7) 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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE?**





* From the above generated regression statistics model, we observed that 69% of the variation in the average price is explained by the all the independent variables in table.
* The intercept of this model is derived as 29.24131.
* The coefficient of CRIME\_RATE is 0.048725141 this means that much times the dependent variable changes while changing it.
* The coefficient of variables is derived as below mentioned values from regression model.
* AGE is 0.032770689
* INDUS is 0.130551399
* NOX is -10.3211828
* DISTANCE is 0.261093575
* TAX is -0.01440119
* PTRATIO is -1.074305348
* AVG\_ROOM is 4.125409152
* LSTAT is -0.603486589
* How the independent variable Significance to dependent variable is decided by p value. From this model the p value od every single variable must be less than 0.05 then only we decide that it is significant to the y variable(avg-price).

**Variable p-value significant or non-significant**

CRIME\_RATE 0.534657201 non-significant variable

AGE 0.012670437 significant variable

INDUS 0.03912086 significant variable

NOX 0.008293859 significant variable

DISTANCE 0.000137546 significant variable

TAX 0.000251247 significant variable

PTRATIO 6.58642E-15 significant variable

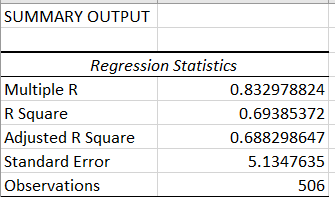
AVG\_ROOM 3.89287E-19 significant variable

LSTAT 8.91071E-27 significant variable

* Finally, I concluded that except crime-rate all the independent variable are significant to the average price.

**8) 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:**

**a) Interpret the output of this model.**

****

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |
|  |  |
| |  |  | | --- | --- | |  |  | |  |  |  * From this regression statistics we observed that 69% of the variation in the average price is explained by the all the independent variables in table. * The intercept of this model is 29.24131. |  |
|  |  |

**8)** **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?**

* By the value of adjusted R-square value little more compared to previous question 7. So, current model performs better in these two models.

**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?**



* Average price will decrease if the value of NOX is increased in a locality of this town.

**d)Write the regression equation from this model.**

* Average Price = 29.4285 + 0.0329 \* AGE + 0.1307 \* INDUS - 10.2727 \* NOX + 0.2615 \* DISTANCE - 0.0145 \* TAX - 1.0717 \* PTRATIO + 4.1255 \* AVG\_ROOM - 0.6052 \* LSTAT.
* 29.42847349+0.03293496(65.2)+ 0.130710007(2.31) -10.27270508(0.538)+ 0.261506423(1) -0.014452345(296) -1.071702473(15.3)+ 4.125468959(6.575) -0.605159282(4.98)

=21.4581

**Conclusion Statement:**

* Terro’s real-estate is an agency provided a table containing 10 different variables. Were the average price was the targeted variable and all other variables are independent variables.
* Using 506 observations, we can state that the average-price of houses was most frequently occurred in the price range of 21,000$ to 25,000$, most of the clients are more interested in purchasing the house falling in the medium price range.
* All the variables are significant to Average-Price except Crime-Rate.