**PROJECT REPORT**

**WEEK 4**

**Case Study:**

Terro’s real estate agency

(TOPICS COVERED: Descriptive Statistics, Covariance, Correlations, Simple Linear Regression, Multiple Linear Regression)

Correlations, Simple Linear Regression,

Multiple Linear Regression)

You have been hired at a Terro’s Real Estate Agency in the capacity of an Auditor. One of the jobs that

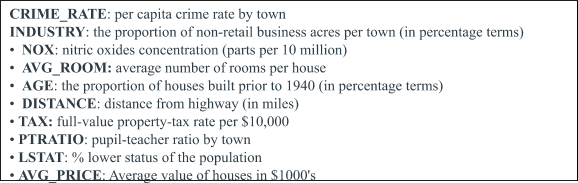
the auditors of this agency do is to map all the relevant features for the properties along with the

information related to the geography around it. The agency wants to understand the relevance of the

parameters that they collect in relation to the value of the house (Avg\_Price).

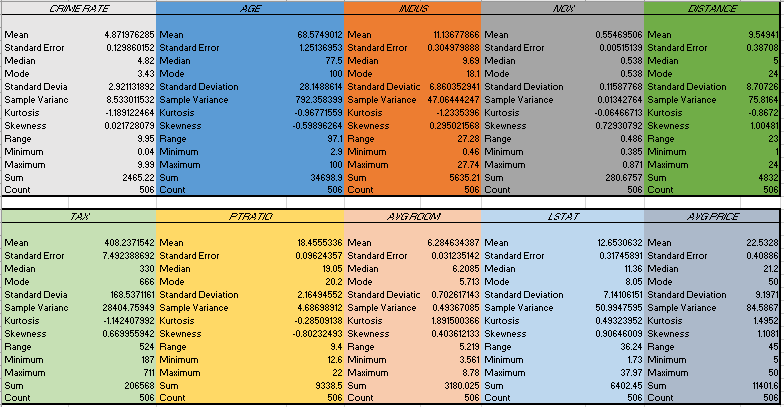
You have been given a dataset of 506 houses in Boston. Please refer to the data dictionary below:

**Data Dictionary:**



Your key job is to analyze the extent and magnitude of each variable relative to the value of the house. For this, you have the following deliverables to execute.

**1)** **The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe? (5 marks)**

****

From the each observation of PRICE & TAX RATIO, AVG ROOM, AVG PRICE I state that,

we can understand and observe that the PTRATIO X PRICE = TAX which depends on the AVERAGE OF ROOM.

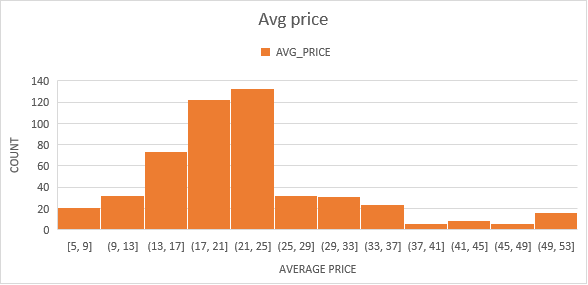
Clearly saying, the price and tax increase as the average of room increases. And comparatively price and tax are more or high for an average room.

Also AVERAGE PRICE is having skewness of 1.1081.

Here AVERAGE PRICE is dependant variable and AVERAGE ROOM is independent.

Also AVG PRICE & AVG ROOM has outliers.

**2)** **Plot the histogram of the Avg\_Price Variable. What do you infer? (5 marks)**



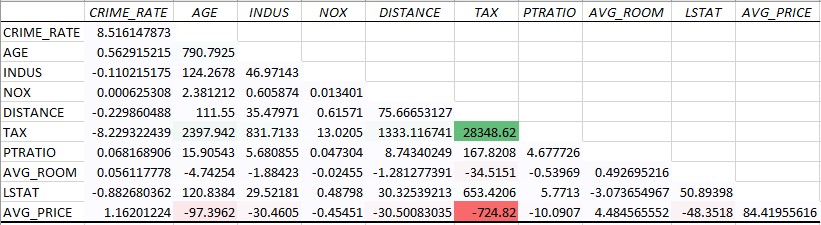
Here Count=COUNT/10

And AVG PRICE is the range of price which lies within the count.

And from the above histogram we can observe, there are range/count of people who’s price is similar and also priced with different measures.

And there is a count of 140 at the range of (21,25). Further stating that there is a positive skewness between AVG PRICES & COUNT.

**3)** **Compute the covariance matrix. Share your observations. (5 marks)**

****

Observing the covariance, we can clearly state that there is a negative relation of

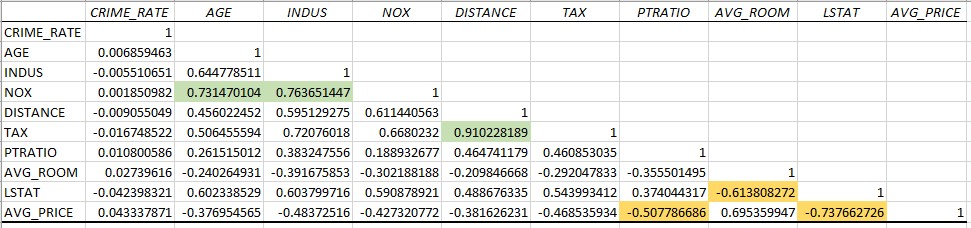
-724.82 between TAX and AVG PRICE. And TAX alone tends towards positive scale with 28348.62. Also the relation between AGE and AVG PRICE tends towards negative scale of

-97.3962

**4)** **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.**

**(5 marks)**

****

The top 3 positive correlation lies between:

TAX and DISTANCE with 0.910228189

NOX and INDUS with 0.763651447

NOX and AGE with 0.731470104

The top 3 negative correlation lies between:

AVG PRICE and PTRATIO with -0.507786686

LSTAT and AVG ROOM with -0.613808272

AVG PRICE and LSTAT with -0.737662726

**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.**

**(8 marks)**

**a. What do you infer from the Regression Summary Output in terms of variance**

**explained, coefficient value, Intercept and the Residual plot?**

**b. 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 | 23243.914 | 601.6178711 | 5.08E-88 |  |  |  |
| Residual | 504 | 19472.38142 | 38.63567742 |  |  |  |  |  |
| 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 | 61.41514552 | 3.7431E-236 | 33.44846 | 35.65922 | 33.44846 | 35.65922 |
| LSTAT | -0.950049354 | 0.038733416 | -24.52789985 | 5.0811E-88 | -1.02615 | -0.87395 | -1.02615 | -0.87395 |

|  |  |  |
| --- | --- | --- |
| RESIDUAL OUTPUT |  |  |
|  |  |  |
| *Observation* | *Predicted AVG\_PRICE* | *Residuals* |
| 1 | 29.8225951 | -5.822595098 |
| 2 | 25.87038979 | -4.270389786 |
| 3 | 30.72514198 | 3.974858016 |
| 4 | 31.76069578 | 1.639304221 |
| 5 | 29.49007782 | 6.709922176 |
| 6 | 29.60408375 | -0.904083746 |
| 7 | 22.74472741 | 0.155272588 |
| 8 | 16.36039575 | 10.73960425 |
| 9 | 6.118863721 | 10.38113628 |
| 10 | 18.30799693 | 0.59200307 |
| 11 | 15.1253316 | -0.125331595 |
| 12 | 21.94668596 | -3.046685955 |
| 13 | 19.62856553 | 2.071434468 |
| 14 | 26.70643322 | -6.306433217 |
|  |  |  |
|  |  |  |
|  |  |  |

a)The observation state that,

The relation between the Coefficient value and Intercept is 34.55384088

And The relation between the Coefficient value and LSTAT is -0.950049354

Also from Residual Plot I infer there is a positive skewness between LSTAT & Residuals

b)

The **significance F** of **5.0811E-88** is verified through hypothesis,

as significance is less than 0.05 resulting in rejecting the alternative hypothesis and giving the null hypothesis. we can observe that LSTAT variable is not significant for analysis.

**6) 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.**

**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?**

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 | 13638.49311 | 444.3308922 | 7.0085E-112 |  |  |  |
| Residual | 503 | 15439.3092 | 30.69445169 |  |  |  |  |  |
| Total | 505 | 42716.29542 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -1.358272812 | 3.17282778 | -0.428095348 | 0.668764941 | -7.591900282 | 4.875354658 | -7.591900282 | 4.875354658 |
| AVG\_ROOM | 5.094787984 | 0.4444655 | 11.46272991 | 3.47226E-27 | 4.221550436 | 5.968025533 | 4.221550436 | 5.968025533 |
| LSTAT | -0.642358334 | 0.043731465 | -14.68869925 | 6.66937E-41 | -0.728277167 | -0.556439501 | -0.728277167 | -0.556439501 |

a)From the observations based on regression I can clearly state that there is a 63.85% of variance in LSTAT can be accounted for AVG ROOM AND AVG PRICE.

To calculate the average price I have used the equation:

y=mx+b

|  |  |
| --- | --- |
| Here  Y=m1x1+m1x2+c |  |
| |  |  | | --- | --- | | average price | 21458.07639 | |  |  | | 30000 | 71.52692131 | | difference percentage | 28.47307869 | |  |
|  |  |
|  |  |
|  |  |
|  |  |

After calculating the AVG PRICE there is a quote of price around **21458 USD** which is **28.47%** more compared to other company quoting the value of **30000 USD**.

And it is overcharging.

b)

R square is basically variance between dependant and independent variable.

When that (variance) R square compared with AVG PRICE VS LSTAT

And LSTAT VS AVG PRICE & AVG ROOM. We can observer that,

AVG PRICE VS LSTAT has variance of **0.544146 or 54.41%**

LSTAT VS AVG PRICE & AVG ROOM has variance of **0.638561606 or 63.85%**

From the observation we can say the 6th question model performs better compared to 5th. The betterment of table depends on the changes in R value.

**7)** **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. (8 marks)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ***Column1*** | ***Coefficients*** | ***Standard Error*** | ***t Stat*** | ***P-value*** | ***Lower 95%*** | ***Upper 95%*** | ***Lower 95.0%*** | ***Upper 95.0%*** |
| AGE | 0.032770689 | 0.013097814 | 2.501997 | 0.0126704 | 0.00703665 | 0.058504728 | 0.00703665 | 0.058504728 |
| AVG\_ROOM | 4.125409152 | 0.442758999 | 9.317505 | 3.893E-19 | 3.255494742 | 4.995323561 | 3.255494742 | 4.995323561 |
| CRIME\_RATE | 0.048725141 | 0.078418647 | 0.621346 | 0.5346572 | -0.105348544 | 0.202798827 | -0.105348544 | 0.202798827 |
| DISTANCE | 0.261093575 | 0.067947067 | 3.842603 | 0.0001375 | 0.127594012 | 0.394593138 | 0.127594012 | 0.394593138 |
| INDUS | 0.130551399 | 0.063117334 | 2.068392 | 0.0391209 | 0.006541094 | 0.254561704 | 0.006541094 | 0.254561704 |
| Intercept | 29.24131526 | 4.817125596 | 6.070283 | 2.54E-09 | 19.77682784 | 38.70580267 | 19.77682784 | 38.70580267 |
| LSTAT | -0.603486589 | 0.053081161 | -11.3691 | 8.911E-27 | -0.70777824 | -0.499194938 | -0.70777824 | -0.499194938 |
| NOX | -10.3211828 | 3.894036256 | -2.65051 | 0.0082939 | -17.97202279 | -2.670342809 | -17.97202279 | -2.670342809 |
| PTRATIO | -1.074305348 | 0.133601722 | -8.0411 | 6.586E-15 | -1.336800438 | -0.811810259 | -1.336800438 | -0.811810259 |
| TAX | -0.01440119 | 0.003905158 | -3.68774 | 0.0002512 | -0.022073881 | -0.0067285 | -0.022073881 | -0.0067285 |

a)

From the above observation R squared value is 0.6883. And

Basically from the regression model we can understand that adding a significant value increases the Adjusted R and adding insignificant values tends to decrease the

Adjusted R. Also here “Adjusted R” value is the variant/different version of “R Square”

Which helps In numerous predictors/predictions.

Also it determines if the correlation is reliable and how much it is determined by the addition of the independent variables.

b)

Also the coefficient of certain variables are positive meaning directly proportional to variable y AVG PRICE. And some have negative coefficient which is inversely proportionate to AVG PRICE.

Hence from the above observation the crime rate is not significant variable in predicting the AVG 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. (8**

**marks)**

**(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)**

**Answer the questions below:**

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

**b. 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?**

**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?**

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

**a)Here Adjusted r square is 0.688298647.**

**To make the multiple linear regression model perform better we should analyse the significant and insignificant variables in the model from the p value. And removing crime rate which is insignificant gives chance to the model perform better, further creating another model excepting Crime rate.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.832836 |  |  |  |  |  |  |  |
| R Square | 0.693615 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.688684 |  |  |  |  |  |  |  |
| Standard Error | 5.131591 |  |  |  |  |  |  |  |
| Observations | 506 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 8 | 29628.68 | 3703.585 | 140.643 | 1.9E-122 |  |  |  |
| Residual | 497 | 13087.61 | 26.33323 |  |  |  |  |  |
| Total | 505 | 42716.3 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 29.42847 | 4.804729 | 6.124898 | 1.85E-09 | 19.98839 | 38.86856 | 19.98839 | 38.86856 |
| AGE | 0.032935 | 0.013087 | 2.516606 | 0.012163 | 0.007222 | 0.058648 | 0.007222 | 0.058648 |
| INDUS | 0.13071 | 0.063078 | 2.072202 | 0.038762 | 0.006778 | 0.254642 | 0.006778 | 0.254642 |
| NOX | -10.2727 | 3.890849 | -2.64022 | 0.008546 | -17.9172 | -2.62816 | -17.9172 | -2.62816 |
| DISTANCE | 0.261506 | 0.067902 | 3.851242 | 0.000133 | 0.128096 | 0.394916 | 0.128096 | 0.394916 |
| TAX | -0.01445 | 0.003902 | -3.70395 | 0.000236 | -0.02212 | -0.00679 | -0.02212 | -0.00679 |
| PTRATIO | -1.0717 | 0.133454 | -8.03053 | 7.08E-15 | -1.33391 | -0.8095 | -1.33391 | -0.8095 |
| AVG\_ROOM | 4.125469 | 0.442485 | 9.3234 | 3.69E-19 | 3.256096 | 4.994842 | 3.256096 | 4.994842 |
| LSTAT | -0.60516 | 0.05298 | -11.4224 | 5.42E-27 | -0.70925 | -0.50107 | -0.70925 | -0.50107 |

The better performance of a model is observed after removing insignificant value

**Crime rate**, and that shows effect on adjusted R too by slightly increasing compared to the previous model. And also an observation which shows the whole model is also good as the value is lower than **0.05.**

b)Adjusted r squared for 7th  question is 83.29%

Adjusted r squared for 7th  question is 83.28%

When compared, both are similar and equal Adjusted R squared values.

c)

|  |  |
| --- | --- |
| ***Column1*** | ***Coefficients*** |
| NOX | -10.3211828 |
| PTRATIO | -1.074305348 |
| LSTAT | -0.603486589 |
| TAX | -0.01440119 |
| AGE | 0.032770689 |
| CRIME\_RATE | 0.048725141 |
| INDUS | 0.130551399 |
| DISTANCE | 0.261093575 |
| AVG\_ROOM | 4.125409152 |
| Intercept | 29.24131526 |

d)

Predicted avg\_price = 0.03\*age + 0.13\*Indus – 10.27\*NOX + 0.26\*distance – 0.014\*tax – 1.07\*ptratio + 4.125\*avg\_room – 0.605\*LSTAT + 29.42