**ASSIGNMENT-2**

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

|  |  |
| --- | --- |
| **CRIME\_RATE** | |
|  |  |
| **Mean** | 4.871976285 |
| **Standard Error** | 0.129860152 |
| **Median** | 4.82 |
| **Mode** | 3.43 |
| **Standard Deviation** | 2.921131892 |
| **Sample Variance** | 8.533011532 |
| **Kurtosis** | -1.189122464 |
| **Skewness** | 0.021728079 |
| **Range** | 9.95 |
| **Minimum** | 0.04 |
| **Maximum** | 9.99 |
| **Sum** | 2465.22 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(4.8719) and Median(4.82) values are almost near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.599) ,it shows that the spread is high
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(0.0217),it describes that there are almost equal values both left and right of the mean as the value of skewness value is almost equal to 0
* From the value of Kurtosis(-1.1891),it shows that the peak is flat as the value of kurtosis is in Negative value(Platykurtic)

|  |  |
| --- | --- |
| **AGE** | |
|  |  |
| **Mean** | 68.57490119 |
| **Standard Error** | 1.251369525 |
| **Median** | 77.5 |
| **Mode** | 100 |
| **Standard Deviation** | 28.14886141 |
| **Sample Variance** | 792.3583985 |
| **Kurtosis** | -0.967715594 |
| **Skewness** | -0.59896264 |
| **Range** | 97.1 |
| **Minimum** | 2.9 |
| **Maximum** | 100 |
| **Sum** | 34698.9 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(68.574) and Median(77.5) values are some what near near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.410) ,it shows that the spread is Normal
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(-0.598),it describes that more values are in right of the mean as the skewness value is in Negative
* From the value of Kurtosis(-0.9677),it shows that the peak is flat as the value of kurtosis is in Negative value(Platykurtic)

|  |  |
| --- | --- |
| **INDUS** | |
|  |  |
| **Mean** | 11.13677866 |
| **Standard Error** | 0.304979888 |
| **Median** | 9.69 |
| **Mode** | 18.1 |
| **Standard Deviation** | 6.860352941 |
| **Sample Variance** | 47.06444247 |
| **Kurtosis** | -1.233539601 |
| **Skewness** | 0.295021568 |
| **Range** | 27.28 |
| **Minimum** | 0.46 |
| **Maximum** | 27.74 |
| **Sum** | 5635.21 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(11.13) and Median(9.69) values are some what near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.616) ,it shows that the spread is High
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(0.295),it describes that it is left leaning as the value is greater than 0
* From the value of Kurtosis(-1.233),it shows that the peak is flat as the value of kurtosis is in Negative value(Platykurtic)

|  |  |
| --- | --- |
| **NOX** | |
|  |  |
| **Mean** | 0.554695059 |
| **Standard Error** | 0.005151391 |
| **Median** | 0.538 |
| **Mode** | 0.538 |
| **Standard Deviation** | 0.115877676 |
| **Sample Variance** | 0.013427636 |
| **Kurtosis** | -0.064667133 |
| **Skewness** | 0.729307923 |
| **Range** | 0.486 |
| **Minimum** | 0.385 |
| **Maximum** | 0.871 |
| **Sum** | 280.6757 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(0.554) and Median(0.538) values are almost near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.208) ,it shows that the spread is Normal
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(0.729),it describes that it is left leaning as the value is greater than 0
* From the value of Kurtosis(-0.064),it shows that the peak is flat as the value of kurtosis is in Negative value(Platykurtic)

|  |  |
| --- | --- |
| **DISTANCE** | |
|  |  |
| **Mean** | 9.549407115 |
| **Standard Error** | 0.387084894 |
| **Median** | 5 |
| **Mode** | 24 |
| **Standard Deviation** | 8.707259384 |
| **Sample Variance** | 75.81636598 |
| **Kurtosis** | -0.867231994 |
| **Skewness** | 1.004814648 |
| **Range** | 23 |
| **Minimum** | 1 |
| **Maximum** | 24 |
| **Sum** | 4832 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(9.549) and Median(5) values are Far away, so we can take mean as the Center as it is high
* **Measure of Dispersion**:
* From the coefficient of variance(0.911) ,it shows that the spread is High
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(1.004),it describes that the data in left of the mean is high as value is greater than 0
* From the value of Kurtosis(-0.867),it shows that the peak is flat as the value of kurtosis is in Negative value(Platykurtic)

|  |  |
| --- | --- |
| **TAX** | |
|  |  |
| **Mean** | 408.2371542 |
| **Standard Error** | 7.492388692 |
| **Median** | 330 |
| **Mode** | 666 |
| **Standard Deviation** | 168.5371161 |
| **Sample Variance** | 28404.75949 |
| **Kurtosis** | -1.142407992 |
| **Skewness** | 0.669955942 |
| **Range** | 524 |
| **Minimum** | 187 |
| **Maximum** | 711 |
| **Sum** | 206568 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(408.23) and Median(330) values are some what near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.412) ,it shows that the spread is Normal
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(0.66),it describes that the data in left of the mean is high as value is greater than 0
* From the value of Kurtosis(-1.142),it shows that the peak is flat as the value of kurtosis is in Negative value(Platykurtic)

|  |  |
| --- | --- |
| **PTRATIO** | |
|  |  |
| **Mean** | 18.4555336 |
| **Standard Error** | 0.096243568 |
| **Median** | 19.05 |
| **Mode** | 20.2 |
| **Standard Deviation** | 2.164945524 |
| **Sample Variance** | 4.686989121 |
| **Kurtosis** | -0.285091383 |
| **Skewness** | -0.802324927 |
| **Range** | 9.4 |
| **Minimum** | 12.6 |
| **Maximum** | 22 |
| **Sum** | 9338.5 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(18.45) and Median(19.05) values are almost near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.117) ,it shows that the spread is Low
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(-0.802),it describes that the data in right of the mean is high as value is less than 0
* From the value of Kurtosis(-0.285),it shows that the peak is flat as the value of kurtosis is in Negative value(Platykurtic)

|  |  |
| --- | --- |
| **AVG\_ROOM** | |
|  |  |
| **Mean** | 6.284634387 |
| **Standard Error** | 0.031235142 |
| **Median** | 6.2085 |
| **Mode** | 5.713 |
| **Standard Deviation** | 0.702617143 |
| **Sample Variance** | 0.49367085 |
| **Kurtosis** | 1.891500366 |
| **Skewness** | 0.403612133 |
| **Range** | 5.219 |
| **Minimum** | 3.561 |
| **Maximum** | 8.78 |
| **Sum** | 3180.025 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(6.284) and Median(6.208) values are almost near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.111) ,it shows that the spread is Low
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(0.403),it describes that the data is left learning which have more data on left of mean
* From the value of Kurtosis(1.891),it shows that the peak is sharp as the value of kurtosis is Positive(Leptokurtic)

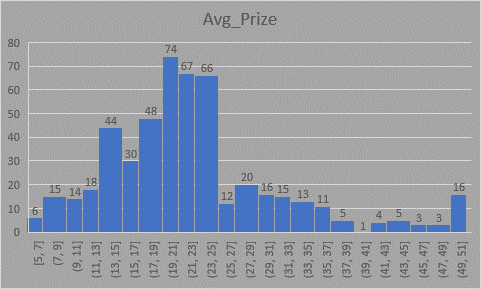
|  |  |
| --- | --- |
| **LSTAT** | |
|  |  |
| **Mean** | 12.65306324 |
| **Standard Error** | 0.317458906 |
| **Median** | 11.36 |
| **Mode** | 8.05 |
| **Standard Deviation** | 7.141061511 |
| **Sample Variance** | 50.99475951 |
| **Kurtosis** | 0.493239517 |
| **Skewness** | 0.906460094 |
| **Range** | 36.24 |
| **Minimum** | 1.73 |
| **Maximum** | 37.97 |
| **Sum** | 6402.45 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(12.65) and Median(11.36) values are some what near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.564) ,it shows that the spread is High
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(0.90),it describes that the data is left learning which have more data on left of mean
* From the value of Kurtosis(0.493),it shows that the peak is sharp as the value of kurtosis is Positive(Leptokurtic)

|  |  |
| --- | --- |
| **AVG\_PRICE** | |
|  |  |
| **Mean** | 22.53280632 |
| **Standard Error** | 0.408861147 |
| **Median** | 21.2 |
| **Mode** | 50 |
| **Standard Deviation** | 9.197104087 |
| **Sample Variance** | 84.58672359 |
| **Kurtosis** | 1.495196944 |
| **Skewness** | 1.108098408 |
| **Range** | 45 |
| **Minimum** | 5 |
| **Maximum** | 50 |
| **Sum** | 11401.6 |
| **Count** | 506 |

* **Measure of Central Tendancy:**
* The Mean(22.53) and Median(21.2) values are almost near, so we can take mean as the Center
* **Measure of Dispersion**:
* From the coefficient of variance(0.408) ,it shows that the spread is Normal
* **Measure of Symmentry/Peakedness**:
* From the Skewness value(1.108),it describes that the data is left learning which have more data on left of mean
* From the value of Kurtosis(1.495),it shows that the peak is sharp as the value of kurtosis is Positive(Leptokurtic)

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



**From the Histogram,it describes that,**

* The Mean(22.53) and Median(21.2) values are almost near, so we can take mean as the Center
* From the coefficient of variance(0.408) ,it shows that the spread is Normal
* From the Skewness value(1.108),it describes that there are more value on the left side of center(Mean) and there is a Tail on the right
* From the value of Kurtosis(1.495),it shows that the peak is sharp as the value of kurtosis is Positive(Leptokurtic)

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

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Column1*** | ***CRIME\_RATE*** | ***AGE*** | ***INDUS*** | ***NOX*** | ***DISTANCE*** | ***TAX*** | ***PTRATIO*** | ***AVG\_ROOM*** | ***LSTAT*** | ***AVG\_PRICE*** |
| CRIME\_RATE | 8.516147873 |  |  |  |  |  |  |  |  |  |
| AGE | 0.562915215 | 790.7924728 |  |  |  |  |  |  |  |  |
| INDUS | -0.110215175 | 124.2678282 | 46.97142974 |  |  |  |  |  |  |  |
| NOX | 0.000625308 | 2.381211931 | 0.605873943 | 0.013401099 |  |  |  |  |  |  |
| DISTANCE | -0.229860488 | 111.5499555 | 35.47971449 | 0.615710224 | 75.66653127 |  |  |  |  |  |
| TAX | -8.229322439 | 2397.941723 | 831.7133331 | 13.02050236 | 1333.116741 | 28348.6236 |  |  |  |  |
| PTRATIO | 0.068168906 | 15.90542545 | 5.680854782 | 0.047303654 | 8.74340249 | 167.8208221 | 4.677726296 |  |  |  |
| AVG\_ROOM | 0.056117778 | -4.74253803 | -1.884225427 | -0.024554826 | -1.281277391 | -34.51510104 | -0.539694518 | 0.492695216 |  |  |
| LSTAT | -0.882680362 | 120.8384405 | 29.52181125 | 0.487979871 | 30.32539213 | 653.4206174 | 5.771300243 | -3.073654967 | 50.89397935 |  |
| AVG\_PRICE | 1.16201224 | -97.39615288 | -30.46050499 | -0.454512407 | -30.50083035 | -724.8204284 | -10.09067561 | 4.484565552 | -48.35179219 | 84.41955616 |

**From the table, we Infer the following:**

* The Age and Tax has high Positive covariance value(2397.941)--🡪Both the values are located in the positive Quadrant(Positive relationship)
* The Avg\_Prize and Tax has high Negative covariance value(-724.82)🡪Both the values are located in the Negative Quadrant(Negative relationship)
* In the table ,some of the Covariance values are related opposite quadrants(2,4),which defines that there is less/No relation between the catagories

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

**a) Which are the top 3 positively correlated pairs and**

**b) Which are the 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 | 0.644778511 | 1 |  |  |  |  |  |  |  |
| **NOX** | 0.001850982 | 0.731470104 | 0.763651447 | 1 |  |  |  |  |  |  |
| **DISTANCE** | -0.009055049 | 0.456022452 | 0.595129275 | 0.611440563 | 1 |  |  |  |  |  |
| **TAX** | -0.016748522 | 0.506455594 | 0.72076018 | 0.6680232 | 0.910228189 | 1 |  |  |  |  |
| **PTRATIO** | 0.010800586 | 0.261515012 | 0.383247556 | 0.188932677 | 0.464741179 | 0.460853035 | 1 |  |  |  |
| **AVG\_ROOM** | 0.02739616 | -0.240264931 | -0.391675853 | -0.30218819 | -0.209846668 | -0.292047833 | -0.355501495 | 1 |  |  |
| **LSTAT** | -0.042398321 | 0.602338529 | 0.603799716 | 0.590878921 | 0.488676335 | 0.543993412 | 0.374044317 | -0.613808272 | 1 |  |
| **AVG\_PRICE** | 0.043337871 | -0.376954565 | -0.48372516 | -0.42732077 | -0.381626231 | -0.468535934 | -0.507786686 | 0.695359947 | -0.737662726 | 1 |

**The Top Positive Correlated Pairs:**

* Tax and Distance-🡪(0.910228188533182)
* NOX and INDUS🡪(0.763651446920914)
* INDUS and TAX🡪( 0.720760179951544)

**The Top Negative Correlated Pairs**:

* AVG\_PRIZE And LSTAT🡪( -0.737662726174014)
* LSTAT and AVG\_ROOM🡪( -0.613808271866396)
* AVG\_PRIZE and PTRAITO🡪( -0.507786685537561)

**5) Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as**

**Independent Variable. Generate the residual plot.**

**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?**

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | 34.55384088 |
| LSTAT | -0.95004935 |

|  |
| --- |
| *P-value* |
| 3.74E-236 |
| 5.0811E-88 |

|  |  |
| --- | --- |
| Adjusted R Square | 0.543241826 |

A)From the Regression model it describes that:

* The P-value is 5.0811E-88
* The Adjusted R Square value is 0.5432
* The Intercept (Value of y, when x=0) is 34.5538
* There is no pattern in Residual Plot
* The Co-efficient (Change of y with increase in X) value of LSTAT is -0.9500

B)The Regression model is suitable for the Prediction, By following below condition:

* The P-value is less than 0.05
* The Adjusted R square is some what near one
* There is no Pattern in the Residual Plot

6)

**build a new Regression model including LSTAT and AVG\_ROOM together as Independent variables**

**and AVG\_PRICE as 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 and explain**.

|  |
| --- |
| *P-value* |
| 0.668764941 |
| 3.47226E-27 |
| 6.66937E-41 |

|  |  |
| --- | --- |
| Adjusted R Square | 0.637124475 |

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | -1.358272812 |
| AVG\_ROOM | 5.094787984 |
| LSTAT | -0.642358334 |

A)Regression Equation=Intercept+Coefficient\*AVG\_ROOM+Coefficient\*LSTAT

AVG\_PRIZE= -1.3582+5.094\*7+ -0.6423\*x20

AVG\_PRIZE= 21.46 USD

From the AVG\_PRIZE value, we can come to an conclusion that the company is Overcharging

B)From the Regression Model,

* The Adjusted R square value in 5th question is 0.5432
* The Adjusted R square value in 6th question is 0.6371

By comparing the both Adjusted R square value ,it shows that MLR-2 (AVG\_ROOM,LSTAT vs AVG\_value) 🡪The MLR-2 is more better than SLR (LSTAT vs AVG\_PRIZE) Model

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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 29.24131526 | 4.817126 | 6.070283 | 2.53978E-09 |
| CRIME\_RATE | 0.048725141 | 0.078419 | 0.621346 | 0.534657201 |
| AGE | 0.032770689 | 0.013098 | 2.501997 | 0.012670437 |
| INDUS | 0.130551399 | 0.063117 | 2.068392 | 0.03912086 |
| NOX | -10.3211828 | 3.894036 | -2.65051 | 0.008293859 |
| DISTANCE | 0.261093575 | 0.067947 | 3.842603 | 0.000137546 |
| TAX | -0.01440119 | 0.003905 | -3.68774 | 0.000251247 |
| PTRATIO | -1.074305348 | 0.133602 | -8.0411 | 6.58642E-15 |
| AVG\_ROOM | 4.125409152 | 0.442759 | 9.317505 | 3.89287E-19 |
| LSTAT | -0.603486589 | 0.053081 | -11.3691 | 8.91071E-27 |

|  |  |
| --- | --- |
| Adjusted R Square | 0.688298647 |

From the Regression model, it describes the following:

* Intercept value🡪29.24131526
* Adjusted R square🡪0.68829864
* The P-value of the CRIME\_RATE is greater than 0.05 (0.53465) ,so we can not use this model for the Prediction

**8) Regression model using only the significant variables you just picked and answer the questions below:**

1. **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.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 29.42847349 | 4.804729 | 6.124898 | 1.84597E-09 |
| AGE | 0.03293496 | 0.013087 | 2.516606 | 0.012162875 |
| INDUS | 0.130710007 | 0.063078 | 2.072202 | 0.038761669 |
| NOX | -10.27270508 | 3.890849 | -2.64022 | 0.008545718 |
| DISTANCE | 0.261506423 | 0.067902 | 3.851242 | 0.000132887 |
| TAX | -0.014452345 | 0.003902 | -3.70395 | 0.000236072 |
| PTRATIO | -1.071702473 | 0.133454 | -8.03053 | 7.08251E-15 |
| AVG\_ROOM | 4.125468959 | 0.442485 | 9.3234 | 3.68969E-19 |
| LSTAT | -0.605159282 | 0.05298 | -11.4224 | 5.41844E-27 |

|  |  |
| --- | --- |
| Adjusted R Square | 0.688683682 |

* Intercept value🡪29.42847349
* Adjusted R square🡪0.688683682
* All P-values are less than 0.05,so we can use this model for Prediction
* By comparing the Adjusted R square of (7 & 8th Question),it describes that not much change in the value and also from the (Adjusted square🡪0.688683682) it satisfies the second condition (Adjusted R square is some what near 1)
* There is no Pattern in the Residual Plot

|  |  |
| --- | --- |
| ***Column1*** | ***Coefficients*** |
| NOX | -10.27270508 |
| PTRATIO | -1.071702473 |
| LSTAT | -0.605159282 |
| TAX | -0.014452345 |
| AGE | 0.03293496 |
| INDUS | 0.130710007 |
| DISTANCE | 0.261506423 |
| AVG\_ROOM | 4.125468959 |
| Intercept | 29.42847349 |

* When there is increase in NOX value in the locality ,the AVG\_PRIZE may be decrease with respect to NOX
* The Regression Equation=29.42+0.032\*AGE+0.130\*INDUS+(-10.272)\*NOX+0.261\*DISTANCE+(-0.014)\*TAX+(-1.071)\*PTRATIO+4.125\*AVG\_ROOM+(-0.605)\*LSTAT