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

**Ans**

## calculate summary statistics using the data analysis tool pack.

### Open the data analysis tool pack by clicking on the "Data" tab in the top menu bar, then selecting "Data Analysis" in the "Analysis" section.

### Select the "Descriptive Statistics" for summary statistics.

### Enter the input range of data.

### Choose where you want to output your results, either in a new worksheet or in a new range of cells. So I select the new worksheet.

### Select the summary statistics you want to calculate, such as mean, median, standard deviation, Standard Error, sum, count, Max, Min, Range, skewness, kurtosis, sample Variance.

### Click "OK" to generate the summary statistics.

1. **The mean and median crime\_rate are relatively close, indicating that the data is not skewed significantly.**
2. **Also The mean and median indus & nox are relatively close, indicating that the data is not skewed significantly.**
3. The standard deviation of Avg\_Price 9.19 indicates moderate variability in the data.
4. The age data ranges from 2 to 100

**Q.2) Plot a histogram of the Avg\_Price variable. What do you infer?**

**Ans**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| here, Maximum AVG\_PRICE of house is (21,25) Of 133 house | | | |  |  |
| total Count Of House is 509 |  |  |  |  |  |
| hear, minimum AVG\_PRICE of house is (37,41) of 6 home & (45,49) of 6 home. | | | | | |
| hear second Maximum AVG\_PRICE of house is (17,21) of 122 house | | | | | |

Select the Avg\_Price variable column in your dataset.

Click on the "Insert" tab in the top ribbon.

Click on "Histogram" in the "Charts" section.

Choose the appropriate bin range and bin width for your histogram.

Click "OK".

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

**Ans**

If the covariance between two variables is zero, it suggests that there is no linear relationship between them. So hear some variable that are covariance between two variables is zero

Crime\_rate & Avg\_room , nox & crime rate , LSTAT & Nox and some other.

**If the covariance is negative, it suggests that the variables tend to move in opposite directions.**

**Avg\_room & Age , avg\_price & age , indus & avg\_price , tax & Crime \_rate**

**If the covariance between two variables is very large or very small, it suggests that there may be a strong linear relationship between them.**

Q.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

**Ans.**

A) positively correlated pairs B) negatively correlated pairs

1) Distance & Tax 1) LSTAT & Avg\_price

2) Indus & Tax 2) Avg\_room & Lstat

3) Nox & Tax 3) Ptratio & Avg Price

Look at the off-diagonal entries in the covariance matrix.

The correlation coefficient ranges from -1 to 1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation.

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

**Ans**

Regression Summary output provides valuable information about the relationship between the dependent variable and the independent variables in a linear regression model.

**Variance explained:** The R-squared value indicates the proportion of variance in the dependent variable that is explained by the independent variables. A higher R-squared value indicates that the model is a good fit for the data and that the independent variables are able to explain a larger proportion of the variation in the dependent variable.

In my initial regression model R-squared value of 0.544146298 indicates that your initial regression model explains 54.4% of the variation in the dependent variable, which is moderate.

***Coefficient values:*** *The coefficients indicate the strength and direction of the relationship between each independent variable and the dependent variable. A positive coefficient value suggests a positive relationship, whereas a negative coefficient value suggests a negative relationship. In my*  initial regression model coefficient values between Intercept & coefficient is 34.55384088 is *positive relationship and coefficient & LSTAT is* -0.950049354 is *negative relationship*

**Intercept**: The intercept in my regression model is 34.55384088. This represents the estimated value of the dependent variable (Y) when all the independent variables (X) in the model are equal to zero.

**B) Is LSTAT variable significant for the analysis based on your model?**

In your my summary output, the variable "LSTAT" has a coefficient estimate of -0.950049354, a standard error of 0.038733416, a t-value of -24.52789985, and a very small p-value (5.0811E-88). This indicates that the variable is statistically significant and has a strong negative relationship with the dependent variable.

**Q.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?**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **y = a + bX** | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | average price= -1.35 + 5.095\*(avg\_room)-0.64\*(LSTAT) | | | | | **21.515** |  |  |
|  |  |  |  |  |  |  |  |  |
|  | putting the value of avg room & LSTAT in to the equation | | | | |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | 21.515/1000 = 0.021515 | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **so as par the Regression model company quoting a value is 30000 USD for locality 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**  Comparing the two models, we can see that the current model has a higher R-squared value of 0.639 compared to the previous model's R-squared value of 0.544. This indicates that the current model explains more variability in the dependent variable than the previous model. The current model also has a lower standard error of 5.540 compared to the previous model's standard error of 6.216, indicating that the current model has a better fit to the data.  ANOVA table of the current model shows that the F-statistic is higher, with a much lower p-value compared to the previous model. This suggests that the current model is a better fit for the data.  Overall, based on these metrics, we can say that the current model performs better than the previous model in explaining the variability in the dependent variable using the independent variables.  **Q. 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. (8 marks)**  **Ans**  The multiple regression model has an adjusted R square of 0.688, which means that 68.8% of the variation in the dependent variable (AVG\_PRICE) can be explained by the independent variables in the model.  The Intercept value of 29.24 indicates that when all the independent variables are 0, the average price of a house in the area is $29,240.  CRIME\_RATE: The coefficient of 0.049 implies that a one-unit increase in crime rate has a positive but insignificant effect on the average price of the house.  AGE: The coefficient of 0.033 indicates that a one-unit increase in the proportion of houses built before 1940 leads to an average price increase of $33.  INDUS: The coefficient of 0.131 shows that a one-unit increase in the proportion of non-retail business acres per town leads to an average price increase of $131.  NOX: The coefficient of -10.321 suggests that a one-unit increase in nitric oxide concentration leads to an average price decrease of $10,321.  DISTANCE: The coefficient of 0.261 indicates that a one-unit increase in the weighted distances to five Boston employment centers leads to an average price increase of $261.  TAX: The coefficient of -0.014 implies that a one-unit increase in the full-value property-tax rate per $10,000 leads to an average price decrease of $14.  PTRATIO: The coefficient of -1.074 shows that a one-unit increase in the pupil-teacher ratio by town leads to an average price decrease of $1,074.  AVG\_ROOM: The coefficient of 4.125 suggests that a one-unit increase in the average number of rooms per dwelling leads to an average price increase of $4,125.  LSTAT: The coefficient of -0.603 indicates that a one-unit increase in the percentage of lower status of the population leads to an average price decrease of $603. | | | | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  |  | |  |
|  |  | |  |  |  |
|  |  | |  |  |  |
|  |  |  |  |  |  |

### 