**Business Situation:**

Predicting price of real estate is challenging as it depends on a normal of variable like pollution, square ft, neighborhood. We have to predict housing process in baston area.

**Task:**

Understand characteristics and attributes of population & neighbor, so that we can predict the prices of a property.

**Action:**

**Data:** CRIME\_RATE, AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG\_ROOM, LSTAT, AVG\_PRICE

**Descriptive Analysis:**

**1)**

**Crime Rate:**

The standard deviation of 2.92 indicates that the data is quite spread out and there is a significant variation in the crime rates of different areas in the locality.

Minimum - 0.04, Maximum - 9.99, Mode - 3.43

The skewness value of 0.02 indicates that the data is approximately symmetric, with a slight right skew.

The kurtosis value of -1.19 indicates that it has thinner tails and a lower peak than a normal distribution.

The mode value of 3.43 indicates that the most common crime rate in the locality is around 3.43.

**Age:**

The age variable is slightly skewed to the left, meaning that there are more data points on the right side of the distribution.

The kurtosis value of -0.97 suggests that it has fewer outliers than a normal distribution.

Mean - 68.57, Median - 77.5, Mode - 100, Standard Deviation - 28.15(High spread out)

**Indus:**

About 11% of the land in each town is used for non-retail business.

The standard deviation of INDUS is 6.86, which means that there is a relatively large amount of variability in the proportion of non-retail business acres per town.

The minimum value of INDUS is 0.46, indicating that there are some towns where almost no land is used for non-retail businesses.

**Nox:**

The distribution of NOX values is slightly skewed to the right, as indicated by the positive skewness value of 0.7293.

Most of the areas as 0.538 ppm

**Distance:**

The distribution of distance values is heavily skewed to the right, as indicated by the positive skewness value of 1.0048.

The mean distance value is 9.5494, with a standard deviation of 8.7073.

**Tax:**

The median tax rate is 330, which is significantly lower than the mean, suggesting that the distribution of tax rates is skewed to the right.

Overall, the tax data appears to be highly variable and skewed, with some suburbs having very high tax rates and a large number of suburbs having tax rates below the mean.

**PTRATIO:**

The median pupil-teacher ratio is 19.05, which is close to the mean, suggesting that the distribution of ratios is approximately symmetrical.

Overall, the pupil-teacher ratio data appears to be less variable and less skewed than the tax data, with most suburbs having ratios relatively close to the mean value.

**Avg\_Room:**

The mean number of rooms per dwelling is 6.28, with a standard deviation of 0.70.

The minimum number of rooms is 3.56 and the maximum is 8.78.

The data is slightly positively skewed (skewness=0.40), and has a moderately high peak (kurtosis=1.89).

**Lstat:**

The mean percentage of lower status of the population is 12.65, with a standard deviation of 7.14.

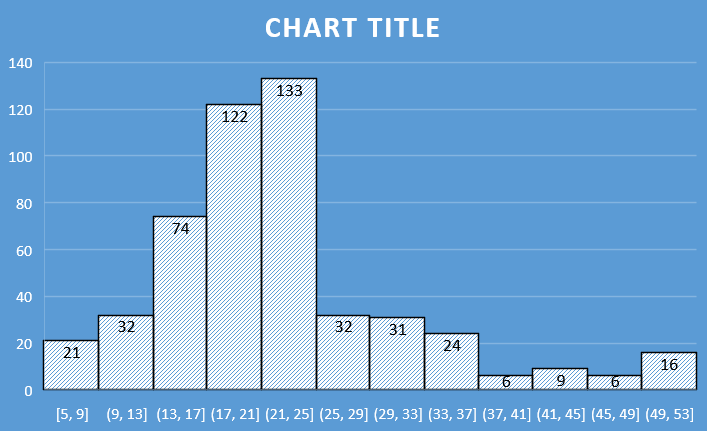
The minimum percentage is 1.73 and the maximum is 37.97.

The data is positively skewed (skewness=0.91), and has a moderate peak (kurtosis=0.49).

**Avg\_Price:**

Avg\_price of the 506 houses is $22K, with highest price of %50K and lowest with $5K

**2)**

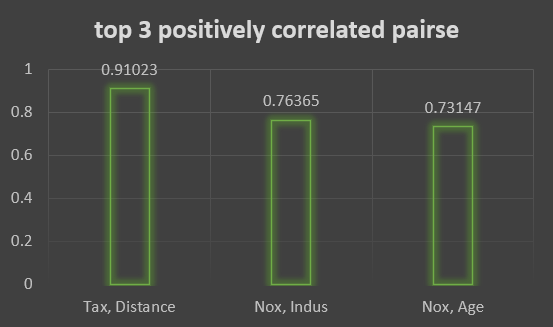


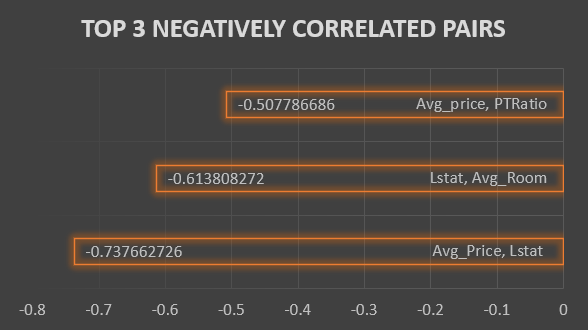
* The distribution of prices is right-skewed, with a long tail to the right indicating a few very high-priced homes.
* The majority of homes are priced between $10k and $30k, with a peak in the frequency around $20k.
* There are relatively few homes priced above $40k, indicating that these are likely luxury homes or outliers.

3)

* There is a positive relationship between Age with Tax, Indus, Lstat, and distance.
* As industries increase in areas, the tax also increases. So, there is a positive relationship with a covariance of 831.71.
* The least relationship is between Avg\_Price & Tax, Avg\_price & Age, Avg\_price & Lstat.
* As the Avg\_price increases, the Tax, Age, and Lstat decrease.
* Looking at the off-diagonal elements, we can see that there are both positive and negative covariances between different pairs of variables.
* Covariance by itself doesn't tell us about the strength or direction of the relationship between variables. To understand the strength and direction of the relationship, we need to look at the correlation coefficient, which takes into account the standard deviations of the variables.

**4)**

****

****

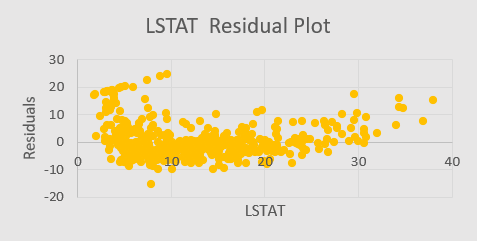
**Top 3 Positive:**

* Tax & Distance with 0.91023
* Nox & Indus with 0.76365
* Nox & Age with 0.73147

**Top 3 Negative:**

* Avg\_price & Lstat with -0.737662726
* Lstat & Toom with -0.613808272
* Avg\_price & PtRation with -0.507786686

**5)**

****

**A.**

Here, we have performed linear regression between Lstat and Avg\_price. The P-value obtained is 5.0811E-88, which is less than 0.05. Hence, it provides evidence to reject the null hypothesis and accept the alternative hypothesis.

However, upon examining the R Square value, it indicates that the model only fits around 54% of the data, making it an imperfect fit. Additionally, when we look at the average Residual, it is -2.7365E-14, indicating that most of the predicted values are approximately 2.73 less than the actual data points.

**B.**

The result is somewhat significant, but not highly significant or insignificant. While the p-value is below 0.05, indicating evidence to reject the null hypothesis, the linear regression only explains around 54% of the variance in the data (as measured by R squared).

**6)**

**A.**

The formula Y = MX + C can be used to derive the regression equation.

With the given values of Avg\_room = 7 and Lstat = 20, and coefficients of Avg\_room = 5.094787984, Lstat = -0.642358334, and Intercept = -1.358272812,

The regression equation can be calculated as follows:

**Regression Equation = (Coefficient of Avg\_room \* Avg\_Room) + (Coefficient of Lstat \* Lstat) + Intercept**

= (5.094787984 \* 7) + (-0.642358334 \* 20) + (-1.358272812)

= 21.45807639

Therefore, based on the regression equation, the predicted value of the house price is $21.46K, which is less than the quoted value of $30K by the company. This suggests that the company is overcharging

**B.**

Here, the p-values are less than 0.05, indicating a statistically significant relationship between the independent variables and the dependent variable. The variance explained by the model is represented by the R-squared value in the regression summary output.

The R-squared value of 0.6386 suggests that the independent variables can explain around 63.86% of the variation in the dependent variable (AVG\_PRICE), which is an improvement from the previous model with only 54%. However, the remaining 36.14% of the variation is still unexplained by the model, likely due to other factors not included in the analysis.

**7)**

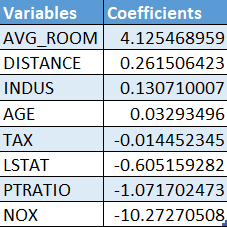
In this regression model, multiple variables were considered. We found that all variables, except Crime\_rate, have a p-value less than 0.05. The R-Square value is 69.38%, which is better than the previous two models.

However, the actual fitting percentage is 68.82% from adjusted R-Square, which takes into account only the effective variables that the dependent variable mostly depends on.

The coefficients for the independent variables NOX, TAX, PTRATIO, and LSTAT are negative, indicating that an increase in any of these variables will result in a decrease in the predicted average house price.

On the other hand, the coefficients for DISTANCE, AVG ROOM, and CRIME RATE are positive, indicating that an increase in any of these variables will result in an increase in the predicted average house price.

**8)**

****

**B.**

In this model, we have included all the variables except for crime\_rate. Since the p-value for crime\_rate in the previous model was greater than 0.05, we can say that the crime\_rate variable is insignificant. The adjusted R-Square for this model is 0.688683682, which is slightly higher than the previous model's adjusted R-Square of 0.688298647. Therefore, we can conclude that this model performs better than the previous one, based on the adjusted R-Square.

**C.**

We can see that the variable NOX has the lowest negative coefficient, which means that if the level of NOX is high in a locality of this town, then it will lead to a drop or reduction in the average price.

**D.**

The regression equation for this model can be expressed as a combination of the coefficients and variables, which is given by:

= (0.03293496 \* C1) + (0.130710007 \* C2) + (-10.27270508 \* C3) + (0.261506423 \* C4) + (-0.014452345 \* C5) + (-1.071702473 \* C6) + (4.125468959 \* C7) + (-0.605159282 \* C8) + 29.42847349