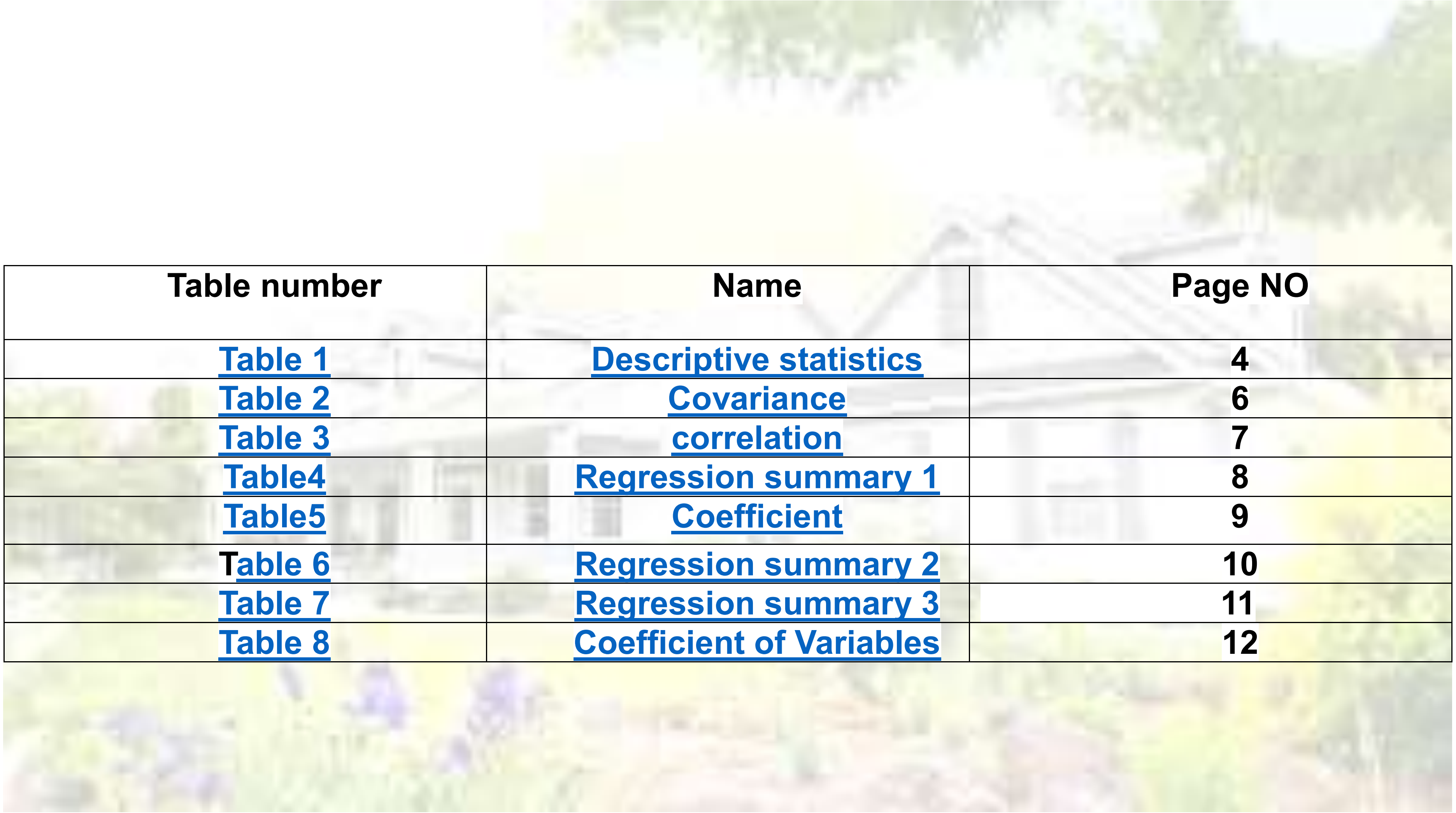
"Finding out the most relevant features for pricing" of a house

Submitted By :	Ankith H Poojary	
Batch	:	GLCA-DA-Online Sep 23
Date	:	30/09/2023

Contents



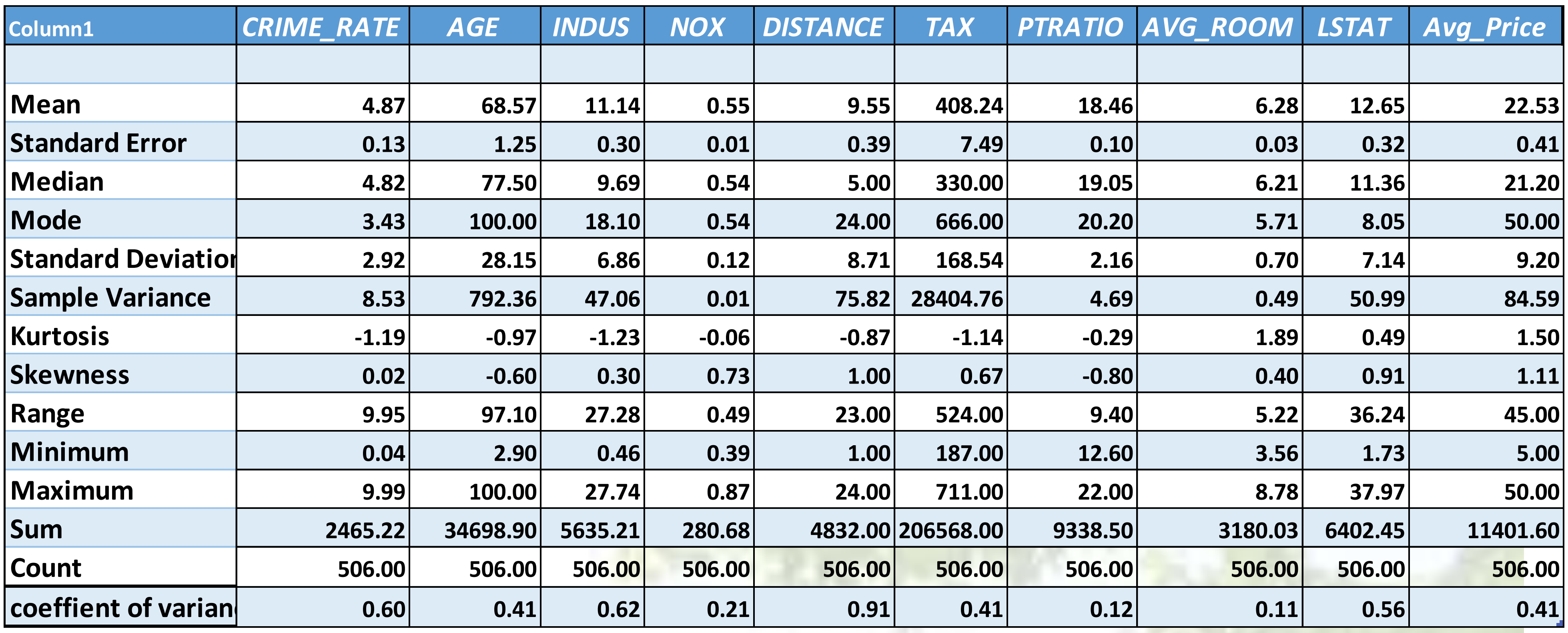
Introduction

The central challenge revolves around understanding what truly drives house prices in specific areas. To tackle this, Terro's has generously shared a dataset featuring 506 houses in Boston, each defined by an array of attributes. These attributes encompass diverse aspects such as crime rates, educational facilities, pollution levels, and more. Terro's seeks to unravel the web of factors influencing property values.

This report explores Terro's Real Estate Agency's pursuit of refining its house pricing strategy, recognizing the critical importance of precise pricing in the dynamic real estate landscape. The central challenge at hand is unraveling the factors that truly drive house prices in specific locales. To address this, Terro's has provided a dataset comprising 506 houses in Boston, each characterized by attributes including crime rates, educational facilities, pollution levels, and more. These data attributes form the cornerstone of our analysis, and our objective is to uncover the intricate relationships within this dataset, ultimately empowering Terro's Real Estate Agency with actionable insights to enhance their pricing decisions and competitiveness in the real estate market.

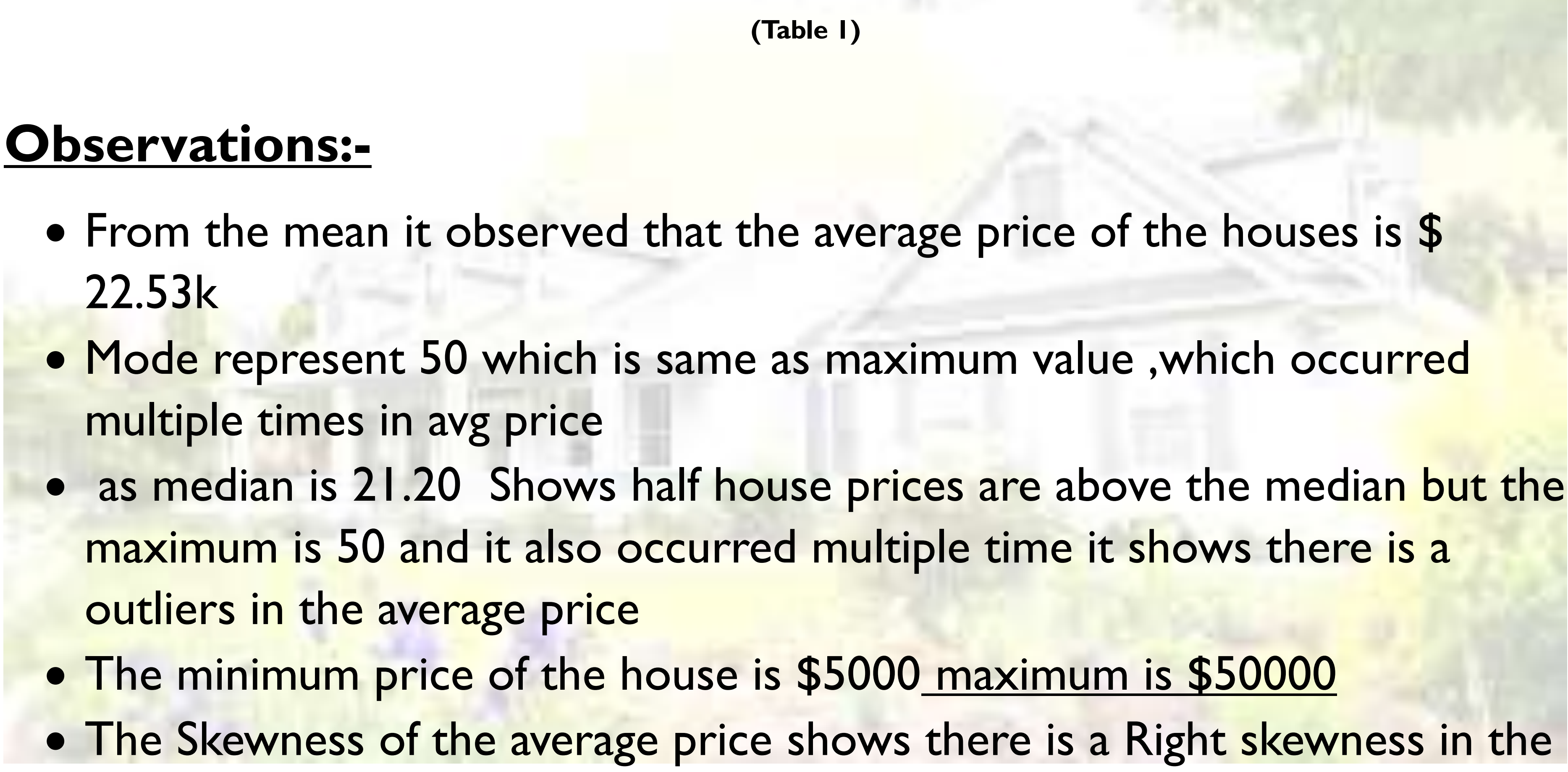
1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

Descriptive statistics



Column1	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	Avg_Price
Mean	4.87	68.57	11.14	0.55	9.55	408.24	18.46	6.28	12.65	22.53
Standard Error	0.13	1.25	0.30	0.01	0.39	7.49	0.10	0.03	0.32	0.41
Median	4.82	77.50	9.69	0.54	5.00	330.00	19.05	6.21	11.36	21.20
Mode	3.43	100.00	18.10	0.54	24.00	666.00	20.20	5.71	8.05	50.00
Standard Deviatior	2.92	28.15	6.86	0.12	8.71	168.54	2.16	0.70	7.14	9.20
Sample Variance	8.53	792.36	47.06	0.01	75.82	28404.76	4.69	0.49	50.99	84.59
Kurtosis	-1.19	-0.97	-1.23	-0.06	-0.87	-1.14	-0.29	1.89	0.49	1.50
Skewness	0.02	-0.60	0.30	0.73	1.00	0.67	-0.80	0.40	0.91	1.11
Range	9.95	97.10	27.28	0.49	23.00	524.00	9.40	5.22	36.24	45.00
Minimum	0.04	2.90	0.46	0.39	1.00	187.00	12.60	3.56	1.73	5.00
Maximum	9.99	100.00	27.74	0.87	24.00	711.00	22.00	8.78	37.97	50.00
Sum	2465.22	34698.90	5635.21	280.68	4832.00	206568.00	9338.50	3180.03	6402.45	11401.60
Count	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00	506.00
coeffient of varian	0.60	0.41	0.62	0.21	0.91	0.41	0.12	0.11	0.56	0.41

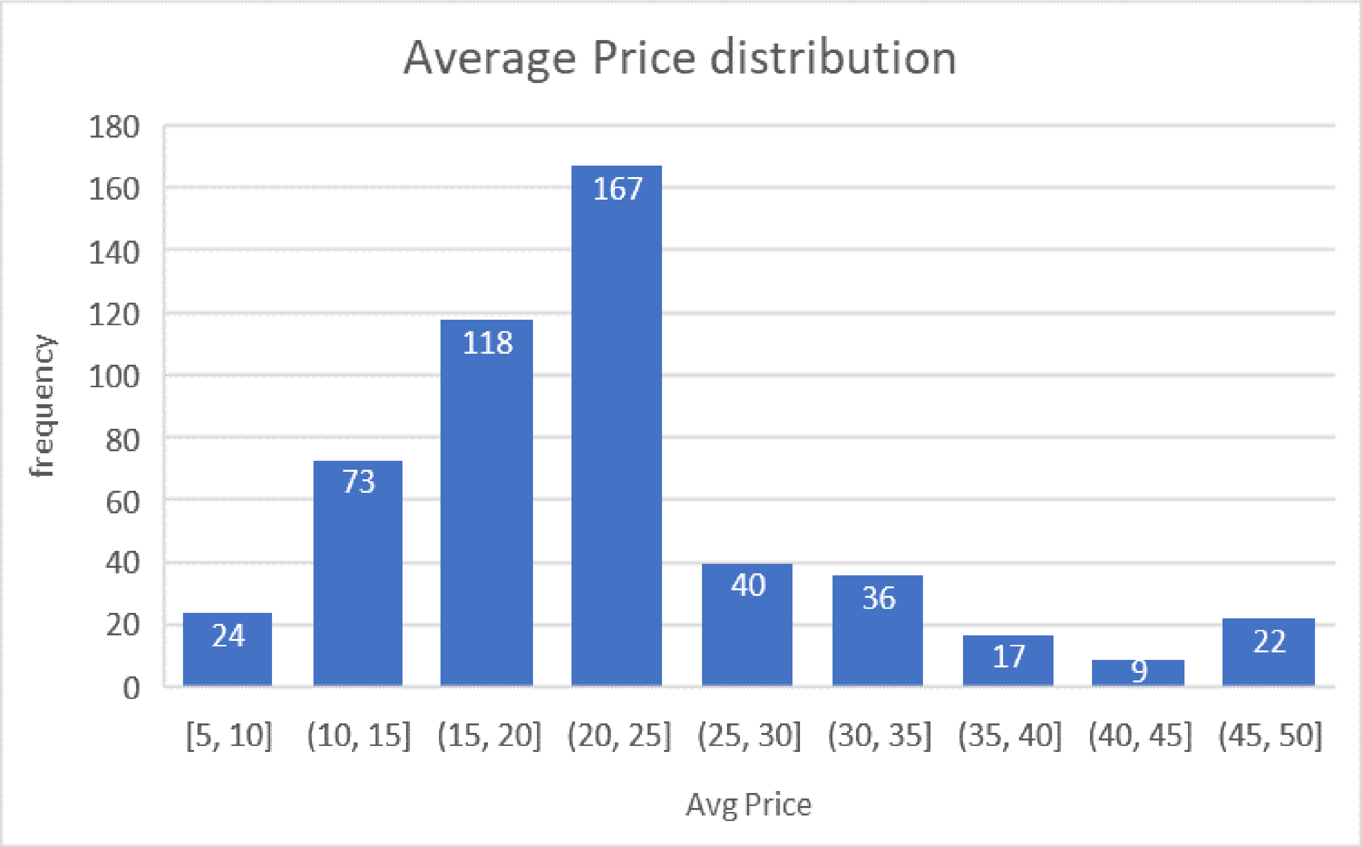
(Table 1)



Observations:-

- From the mean it observed that the average price of the houses is $ 22.53k
- Mode represent 50 which is same as maximum value ,which occurred multiple times in avg price
- as median is 21.20 Shows half house prices are above the median but the maximum is 50 and it also occurred multiple time it shows there is a outliers in the average price
- The minimum price of the house is $5000 maximum is $50000
- The Skewness of the average price shows there is a Right skewness in the data. That means the number of houses is more below the price of average and there is an outlier to the right-side.
- The age of the house is on an average of 69 % of houses built prior to 1940, but the skewness shows there is negative skewness that shows there are outliers.
- The co-effient of variance of avg price shows there is 41% variation in the avg price.

2) Plot a histogram of the Avg_Price variable. What do you infer?



(F.1)

Approach:-Selected Average price range and inserted histogram for this range

It is a right skewed graph.

In the above chart is shows the distribution of average prices of the houses, most of the houses are between the prices of $20000-$25000 and there many houses below that range so the most houses are available below $25000,

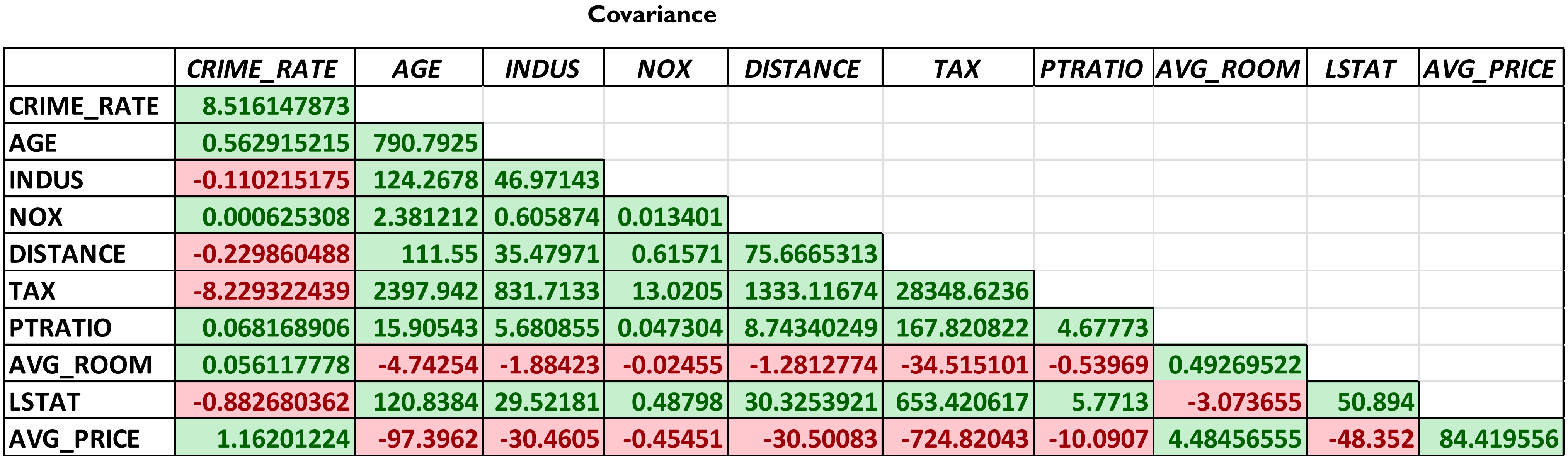
It showing right skewness the number of the houses above $25000 is less

We have many houses nearly 167 of the price around $25000,And followed by the how price between $15000-$20000

The Graphs Shows there is an outliers to the rightside the number of the houses above the average is low but the houses of high value between $45000-$50000 is more it affect the average of price

>>**The Regression model may affected by the outliers in the price**

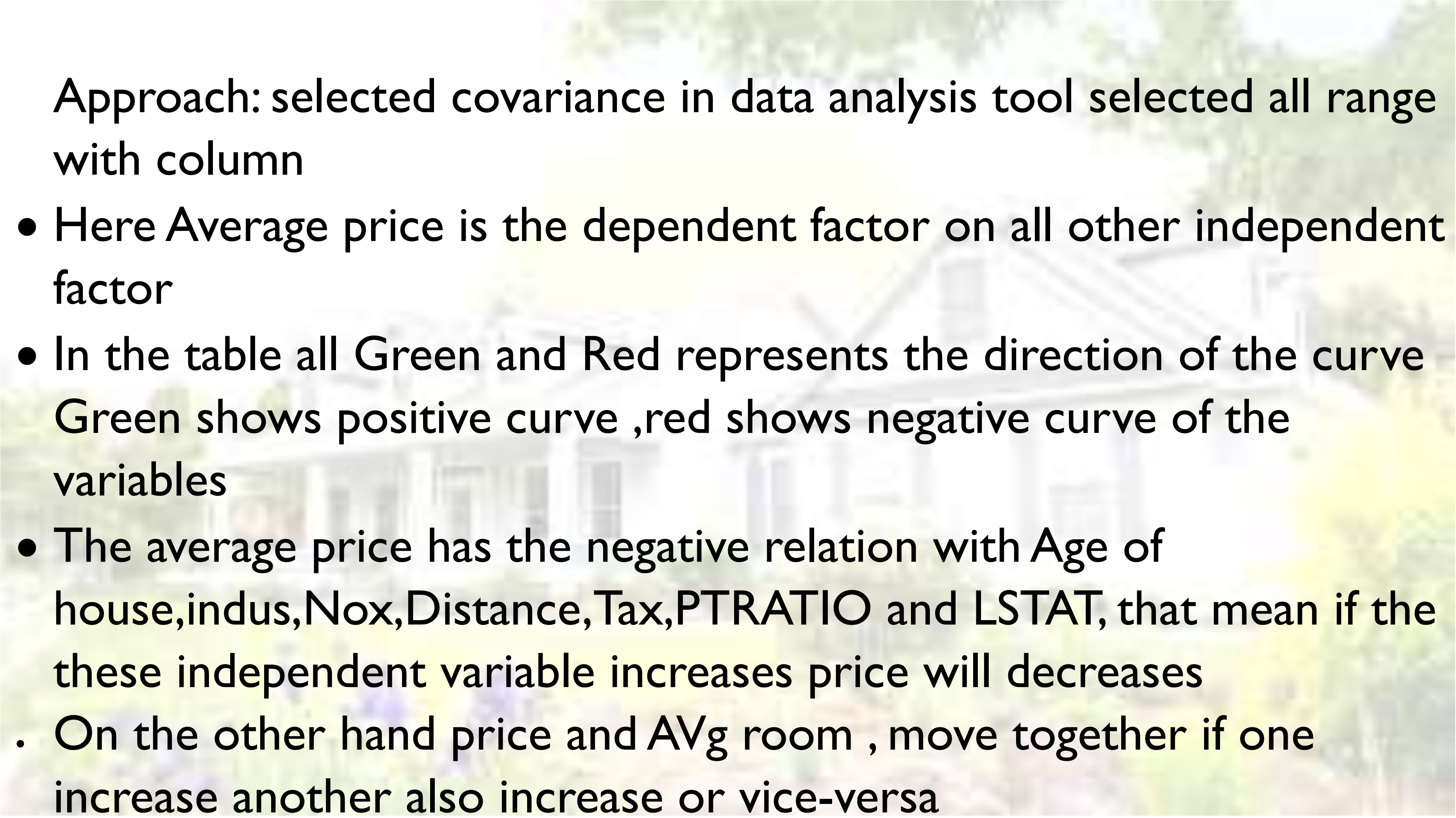
3. Compute the covariance matrix. Share your observations.



Covariance

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.6665313					
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.11674	28348.6236				
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.74340249	167.820822	4.67773			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.2812774	-34.515101	-0.53969	0.49269522		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.3253921	653.420617	5.7713	-3.073655	50.894	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.50083	-724.82043	-10.0907	4.48456555	-48.352	84.419556

Table :2



Approach: selected covariance in data analysis tool selected all range with column

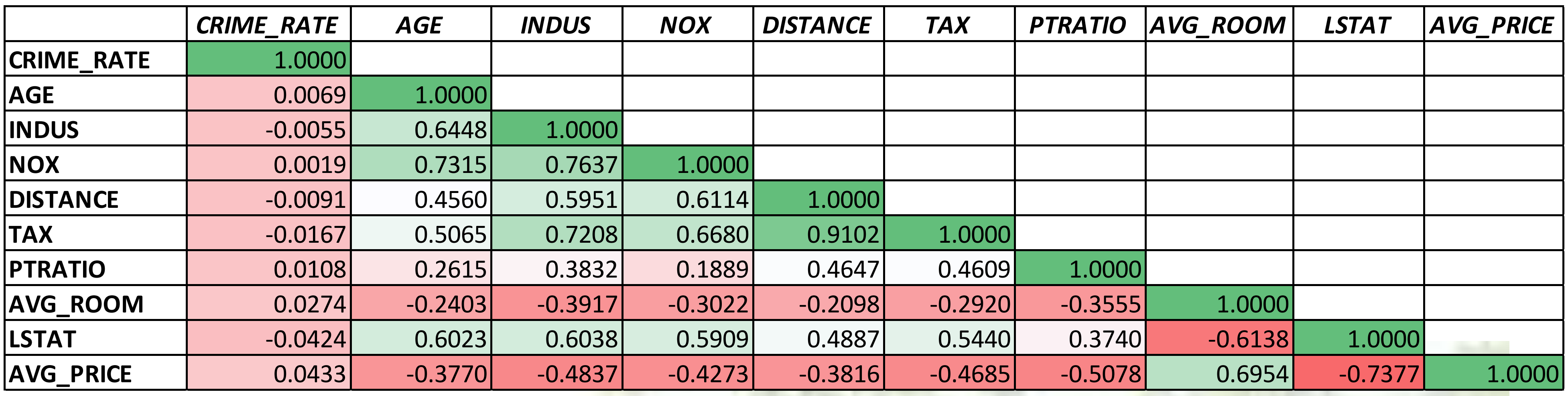
- Here Average price is the dependent factor on all other independent factor
- In the table all Green and Red represents the direction of the curve Green shows positive curve ,red shows negative curve of the variables
- The average price has the negative relation with Age of house,indus,Nox,Distance,Tax,PTRATIO and LSTAT, that mean if the these independent variable increases price will decreases
- On the other hand price and AVg room , move together if one increase another also increase or vice-versa
- **There is a inter-relationship between the independent variables it leads to Multi-colliniarity**
- All the variables holds a significant relationship with avg price

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

a) Which are the top 3 positively correlated pairs.

b) Which are the top 3 negatively correlated pairs.

Correlation



	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1.0000									
AGE	0.0069	1.0000								
INDUS	-0.0055	0.6448	1.0000							
NOX	0.0019	0.7315	0.7637	1.0000						
DISTANCE	-0.0091	0.4560	0.5951	0.6114	1.0000					
TAX	-0.0167	0.5065	0.7208	0.6680	0.9102	1.0000				
PTRATIO	0.0108	0.2615	0.3832	0.1889	0.4647	0.4609	1.0000			
AVG_ROOM	0.0274	-0.2403	-0.3917	-0.3022	-0.2098	-0.2920	-0.3555	1.0000		
LSTAT	-0.0424	0.6023	0.6038	0.5909	0.4887	0.5440	0.3740	-0.6138	1.0000	
AVG_PRICE	0.0433	-0.3770	-0.4837	-0.4273	-0.3816	-0.4685	-0.5078	0.6954	-0.7377	1.0000

Table :3

A) Top 3 positive correlated pairs:-

TAX and DISTANCE

NOX and INDUS

NOX and AGE

B) Top 3 negative correlated Pairs:-

Avg Price and LSTAT

LSTAT and AVG Room

AVG price and PTRATIO

5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

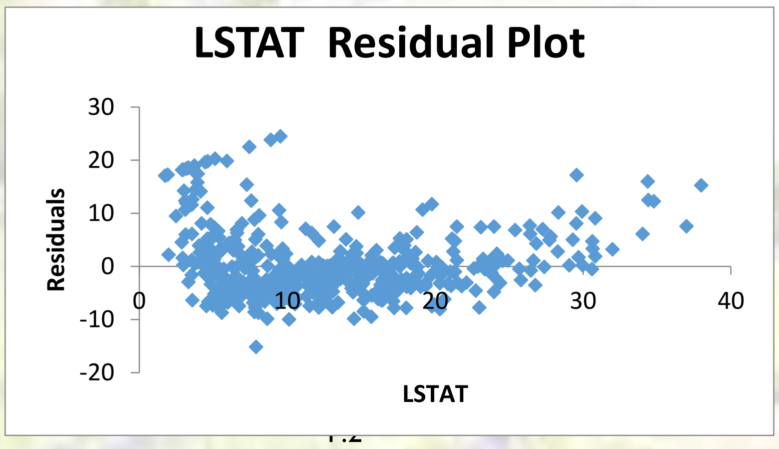
a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and Residual plot?

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

Regression summary 1

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.73766273							
R Square	0.5441463							
Adjusted R Square	0.54324183							
Standard Error	6.21576041							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	23243.914	23243.914	601.61787	5.0811E-88			
Residual	504	19472.38142	38.6356774					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.5538409	0.562627355	61.4151455	3.74E-236	33.44845704	35.6592247	33.448457	35.6592247
LSTAT	-0.9500494	0.038733416	-24.5279	5.081E-88	-1.0261482	-0.8739505	-1.0261482	-0.8739505

Table 4



a)

- The R Square is less so not satisfy the first condition of good model
- The coefficient of LSTAT is -0.95 that shows it have significant impact on the AVG price of the house. A unit change in LSTAT will result in huge change in AVG price in negative direction
- The p value of the LSTAT Is significant as it is less than 5%
- The intercept is shows the Dependent factor when all independent Factors are 0

The Residual plot:-

It representing the relationship between residuals and the LSTAT It not showing any trends that means residual is independent of Variable LSTAT so it satisfies the condition of Residuals

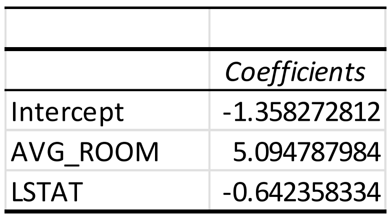
b)

Yes, LSTAT is a significant factor as it have good coefficient and P value is also less than 5%.

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?**

The coefficient Table



	Coefficients
Intercept	-1.358272812
AVG_ROOM	5.094787984
LSTAT	-0.642358334

Table:5

Formula : Y=Mx+c

Y= Dependent

M=Estimated slope (coefficient of Variable)

C=Intercept

Predicted value

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

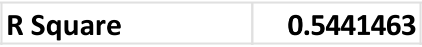
21.45807639

So The predicted value is **$21458**,

If the company charge 30000 USD it is nearly 9000 USD more the predicted value, so it is Overcharging

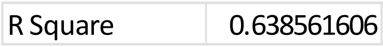
b) **Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain**

Previous model



R Square	**0.5441463**

This Model



R Square	0.638561606

Yes ,there is a significant change in the R square Between 2 models In previous model R square is only 54% which didn't satisfied the condition to consider as a good model,

But in next model R square id 63.9% Which is on significant side this R square of the model satisfies the first condition of a good predictive model.

Conclude:- The R square is Good in model second in compare to model I

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 R Square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.832979
R Square	0.693854
Adjusted R Square	0.688299
Standard Error	5.134764
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.86	3293.207	124.9045	1.9E-121
Residual	496	13077.43	26.3658		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24132	4.817126	6.070283	2.54E-09	19.77683	38.7058	19.77683	38.7058
CRIME_RATE	0.048725	0.078419	0.621346	0.534657	-0.10535	0.202799	-0.10535	0.202799
AGE	0.032771	0.013098	2.501997	0.01267	0.007037	0.058505	0.007037	0.058505
INDUS	0.130551	0.063117	2.068392	0.039121	0.006541	0.254562	0.006541	0.254562
NOX	-10.3212	3.894036	-2.65051	0.008294	-17.972	-2.67034	-17.972	-2.67034
DISTANCE	0.261094	0.067947	3.842603	0.000138	0.127594	0.394593	0.127594	0.394593
TAX	-0.0144	0.003905	-3.68774	0.000251	-0.02207	-0.00673	-0.02207	-0.00673
PTRATIO	-1.07431	0.133602	-8.0411	6.59E-15	-1.3368	-0.81181	-1.3368	-0.81181
AVG_ROOM	4.125409	0.442759	9.317505	3.89E-19	3.255495	4.995324	3.255495	4.995324
LSTAT	-0.60349	0.053081	-11.3691	8.91E-27	-0.70778	-0.49919	-0.70778	-0.49919

Table:6

Approach used: go to data analysis tool –selected regression analysis-selected range of avg price as a y (dependent) and all other variable as x (independent variable).

Insights:-R value is greater than 62% and also coefficient also Good , p Value is less than 5% for all variable except Crime rate , so we have to remove the crime rate from our model

Interpretation:

- Here Adjusted R square is within 1 % deffer from R square so it is good sign.(so there are less penalised value)
- Coefficient shows how much the independent variable can contribute to the prediction of Dependent variable
  Coefficient shows how much the dependent variable will vary for a unit change in independent variable.
  In the model for predicting AVG price crime rate can contribute 4%
  Indus contribute 13%
  NOX will negatively contribute
  Avg room hold large positive coefficient to predict the Avg price
  (so the coefficient explain how much strength the independent variable hold to predict the dependent variable)

- Intercept is showing the value of the Avg price when all independent variable kept to ZERO.

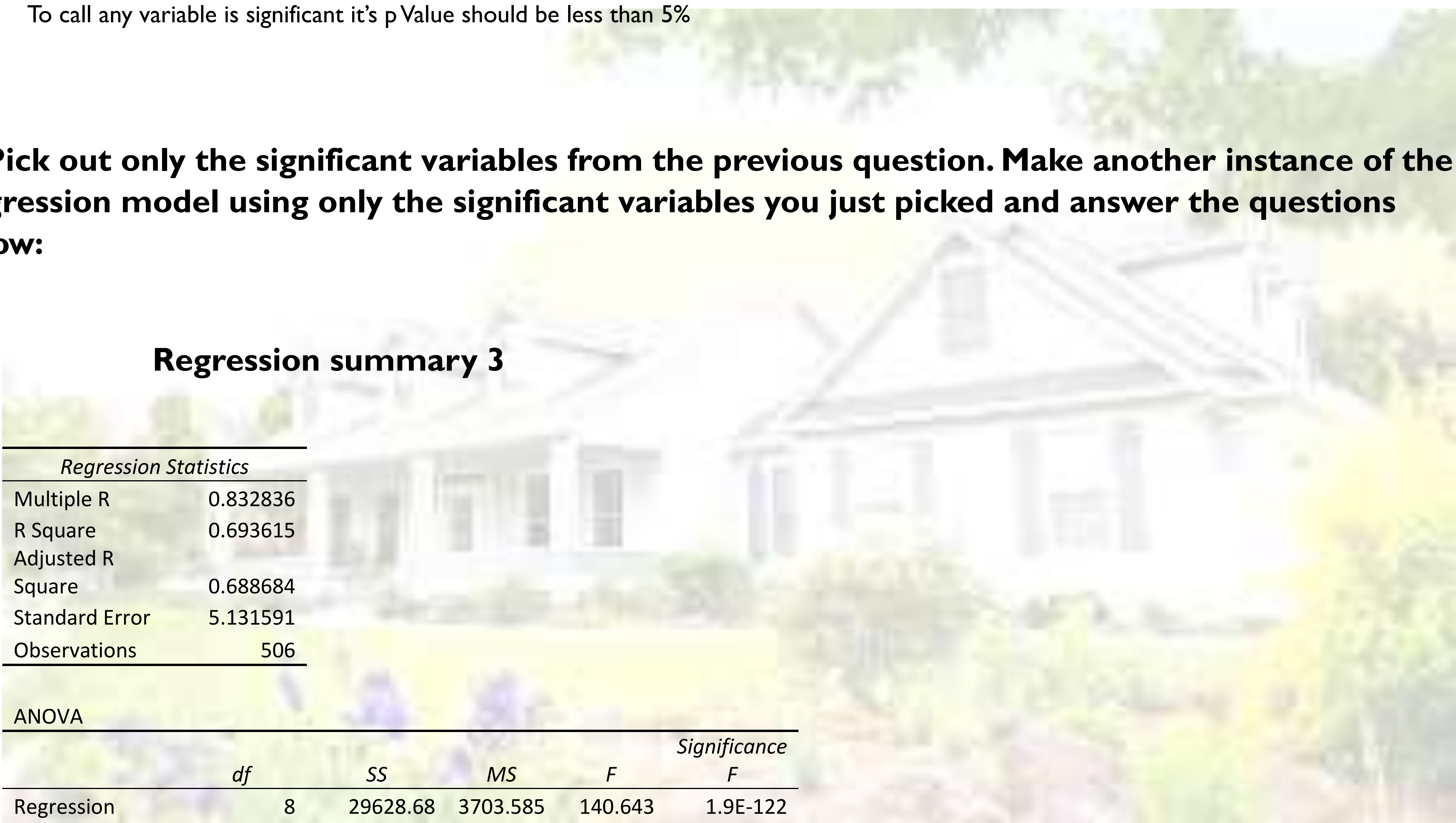
- **Significance of Independent variable:-**

1)Crime rate is insignificant as its p value more than 5% which is(0.53) → Insignificant

2).Age is significant factor as it's p value is less than 5%
3) **INDU**S is significant with p value 0.039
4)**NOX** is a significant factor in predicting AVG price
5)DISTANCE is a significant factor as its p value less than 5%
6)TAX is also a significant factor as its p value less than 5%
7)PTRATIO is a significant factor it have great coefficient and less p value
8)AVG ROOM is a significant factor in predicting Price as its coefficient is good and p value also less than 5%
9)LSTAT is significant in terms of coefficient and p value also less

significant

To call any variable is significant it's p Value should be less than 5%



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:

Regression summary 3

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			

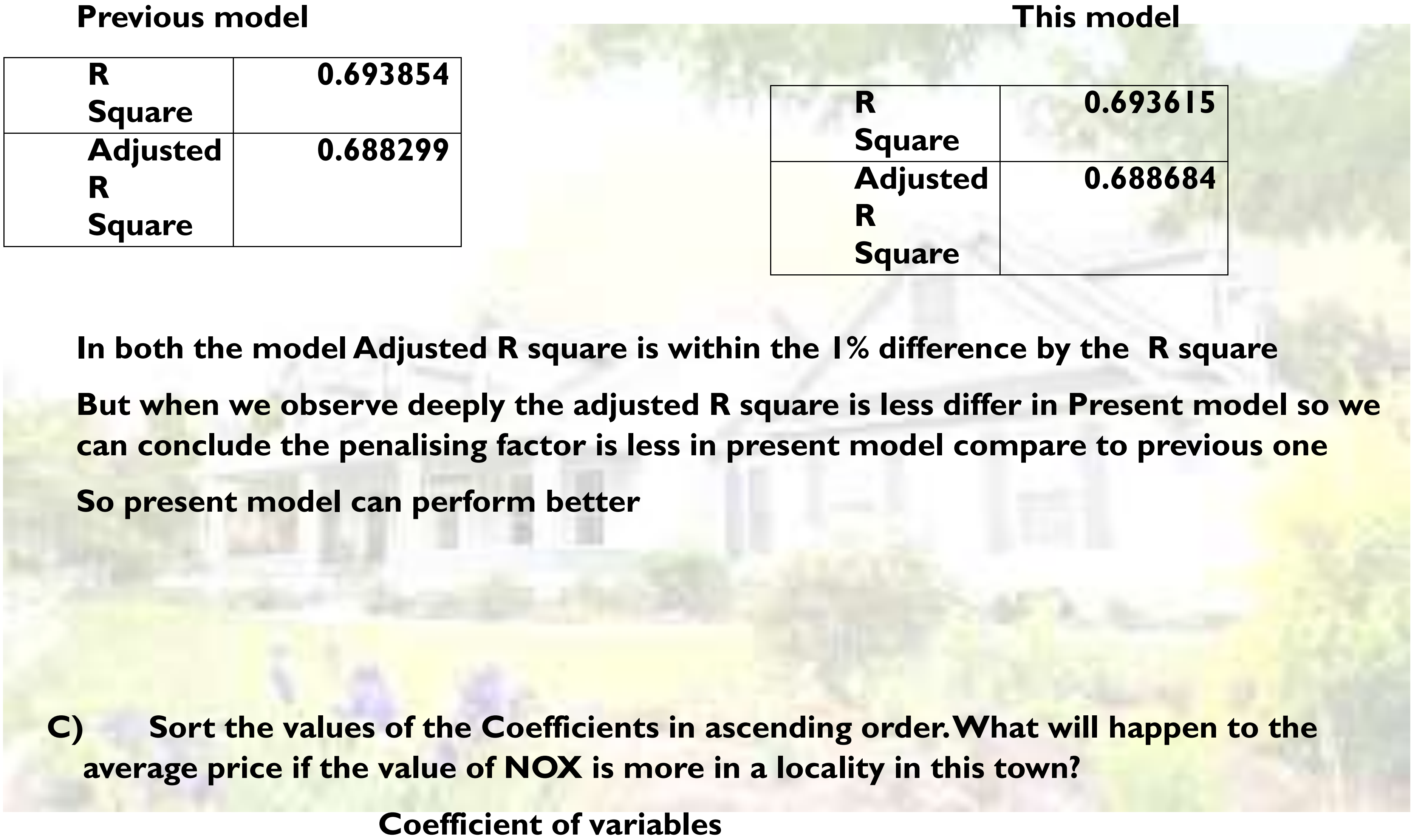
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847	4.804729	6.124898	1.85E-09	19.98839	38.86856	19.98839	38.86856
AGE	0.032935	0.013087	2.516606	0.012163	0.007222	0.058648	0.007222	0.058648
INDUS	0.13071	0.063078	2.072202	0.038762	0.006778	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

Table:7

a) Interpret the output of this model.

- **In this model R Square is Above 62% and adjusted R square is also within 1 % differ from R square .it represent good sign of the variable**
- **The coefficient is also Good it shows all variable has a significant role in predicting Avg price**
- **The P value of all variable is below 5% that confirms all variable here is a significant factor to predict the avg price of the house**

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?



Previous model

R Square	0.693854
Adjusted R Square	0.688299

This model

R Square	0.693615
Adjusted R Square	0.688684

In both the model Adjusted R square is within the 1% difference by the R square

But when we observe deeply the adjusted R square is less differ in Present model so we can conclude the penalising factor is less in present model compare to previous one

So present model can perform better

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?

Coefficient of variables

	Coefficients
NOX	-10.272705
PTRATIO	-1.0717025
LSTAT	-0.6051593
TAX	-0.0144523
AGE	0.03293496
INDUS	0.13071001
DISTANCE	0.26150642
AVG_ROOM	4.12546896
Intercept	29.4284735

Table:8

The coefficient of NOX is -10.272705 it Shows that negative trend .

If the Value of NOX is more the AVG price will reduce in that town (As NOX have a significant negative coefficient

d).Write the regression equation from this model.

Y = β0 + β1 * X1 + β2 * X2 + … .βn * Xn

β0=Intecept

OR

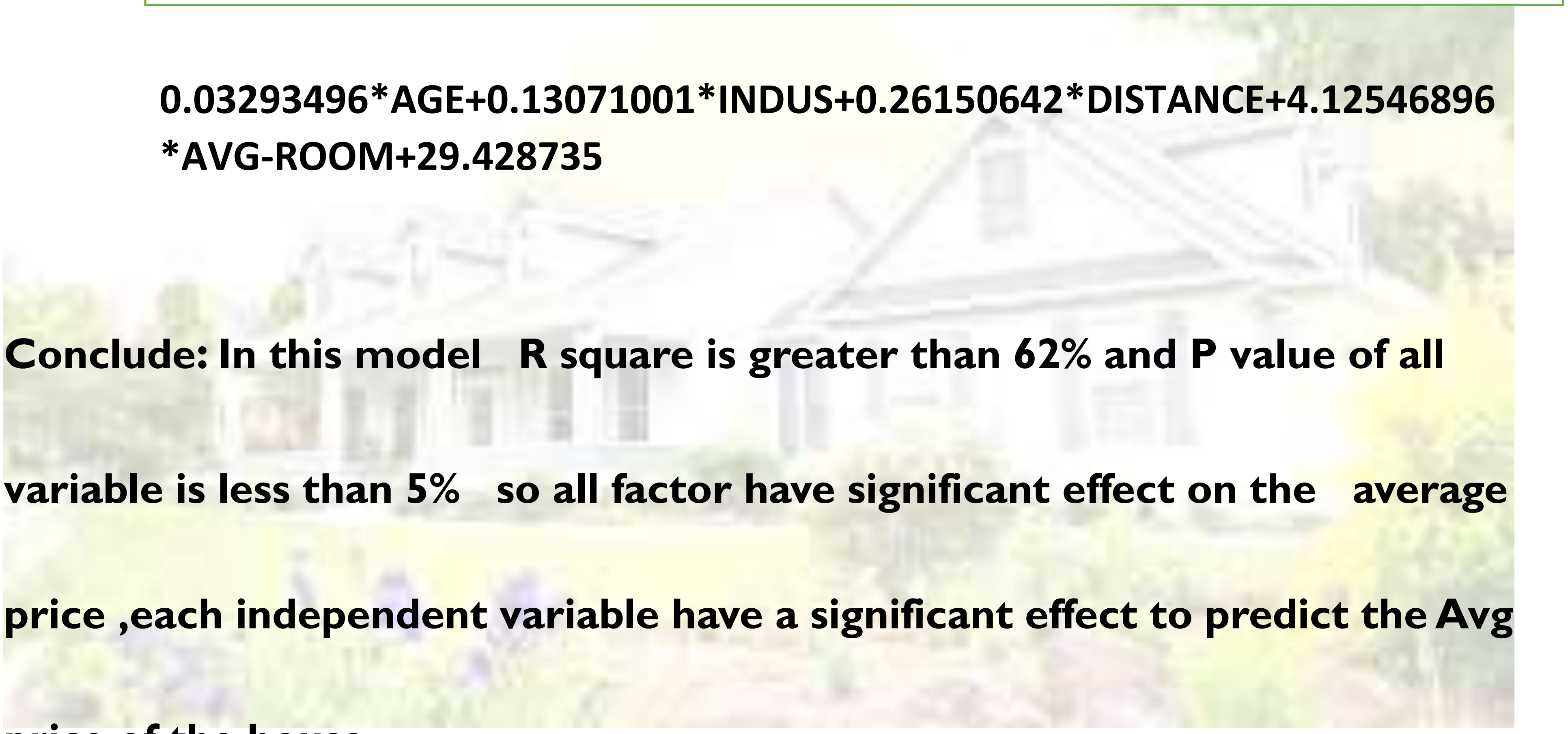
B1 , β2 , … .βn =coefficients of X

Y=MX+C

X=independent variables

Y= -10.272705*NOX+(-1.0717025*PTRATIO)+(-0.6051593*LSTAT)+(-0.0144523*TAX+

0.03293496*AGE+0.13071001*INDUS+0.26150642*DISTANCE+4.12546896*AVG-ROOM+29.428735

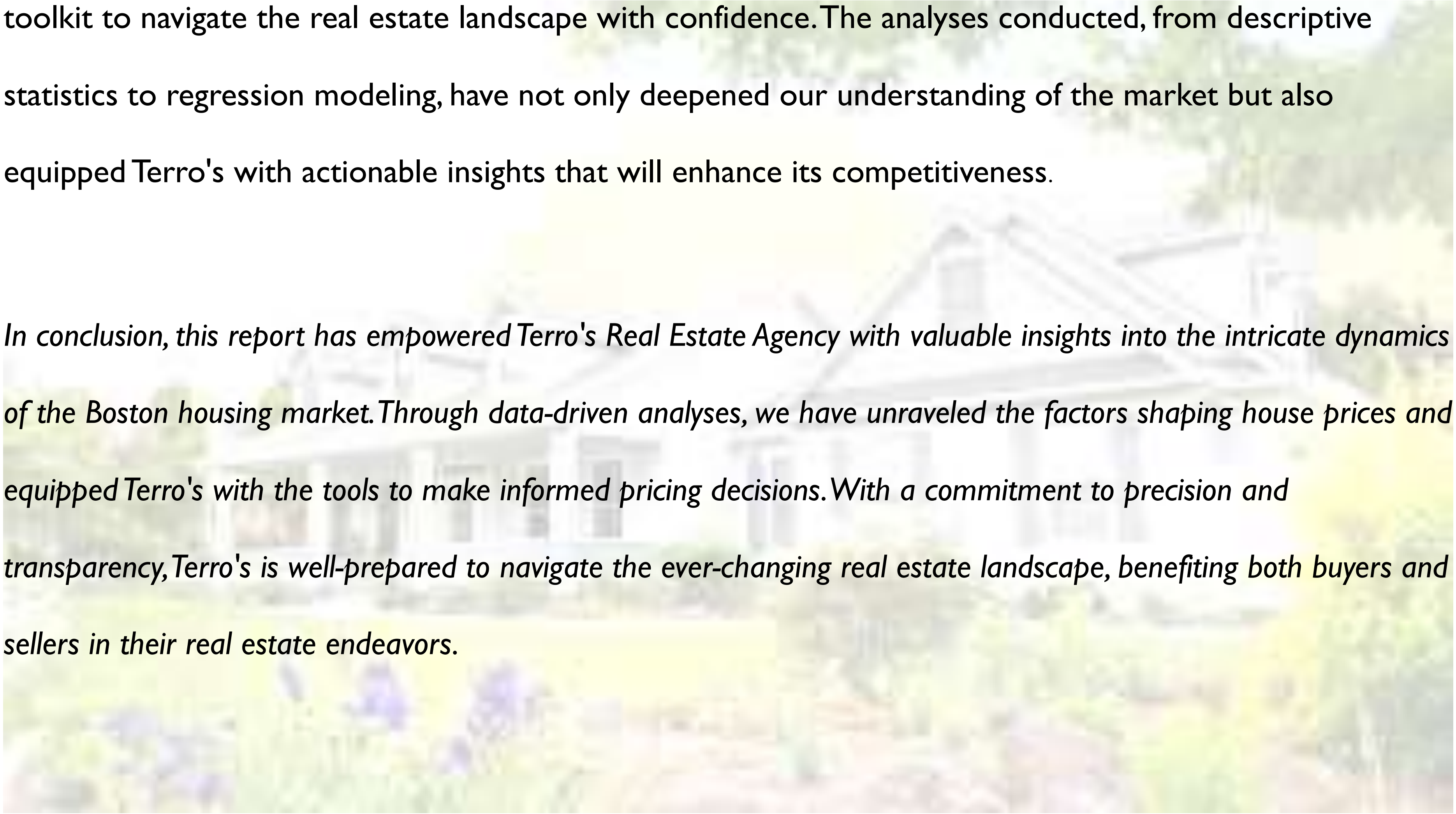


Conclude: In this model R square is greater than 62% and P value of all variable is less than 5% so all factor have significant effect on the average price ,each independent variable have a significant effect to predict the Avg price of the house

Conclusion:

In closing, this report encapsulates Terro's Real Estate Agency's dedicated efforts to refine its house pricing strategy. Our journey through the intricacies of the Boston housing market, guided by a dataset of 506 houses and a multitude of attributes, has illuminated key insights that will empower Terro's to make more informed pricing decisions.

By unraveling the complex web of factors that influence property values, we have provided Terro's with a toolkit to navigate the real estate landscape with confidence. The analyses conducted, from descriptive statistics to regression modeling, have not only deepened our understanding of the market but also equipped Terro's with actionable insights that will enhance its competitiveness.



In conclusion, this report has empowered Terro's Real Estate Agency with valuable insights into the intricate dynamics of the Boston housing market. Through data-driven analyses, we have unraveled the factors shaping house prices and equipped Terro's with the tools to make informed pricing decisions. With a commitment to precision and transparency, Terro's is well-prepared to navigate the ever-changing real estate landscape, benefiting both buyers and sellers in their real estate endeavors.

THANK YOU