Simple Linear Regression Using Python.

-A MODEL TO PREDICT THE PRICE OF HOUSES IN USA.

CREATED BY VIVEK CHATTOPADHYAY

INTRODUCTION

THIS IS A SCENARIO WHERE A PERSON WANTS TO KNOW THE PRICE OF THE HOUSE THEY ARE BUYING BUT DO NOT HAVE ENOUGH INFORMATION ABOUT HOW THOSE PRICES ARE DETERMINED OR PREDICTED. HENCE, THIS PROJECT IS MADE FOR THE PURPOSE OF PREDICTING THE HOUSE PRICES BASED ON A DATABASE OF USA HOUSE MARKET.

METHODOLOGY

1. TO SEE IF THE DATA IS LINEAR OR NON-LINEAR IN NATURE.
2. TO SEE IF THE DATA HAS ANY OUTLIERS OR ANY MISSING VALUES.
3. CREATING A LINEAR REGRESSION MODEL.
4. REMOVING AUTOCORRELATION.
5. CHECKING IF THE ASSUMPTIONS OF THE LINEAR REGRESSION IS MET OR NOT.

THE USA HOUSING DATABASE - 1:

- The data contains the following columns:
- 'Avg. Area Income': Avg. Income of residents of the city house is located in.
- 'Avg. Area House Age': Avg Age of Houses in same city.
- 'Avg. Area Number of Rooms': Avg Number of Rooms for Houses in same city.
- 'Avg. Area Number of Bedrooms': Avg Number of Bedrooms for Houses in same city.
- 'Area Population': Population of city house is located in.
- 'Price': Price that the house sold at.
- 'Address': Address for the house.

THE USA HOUSING DATABASE - 2:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386

DESCRIPTIVE STATISTICS

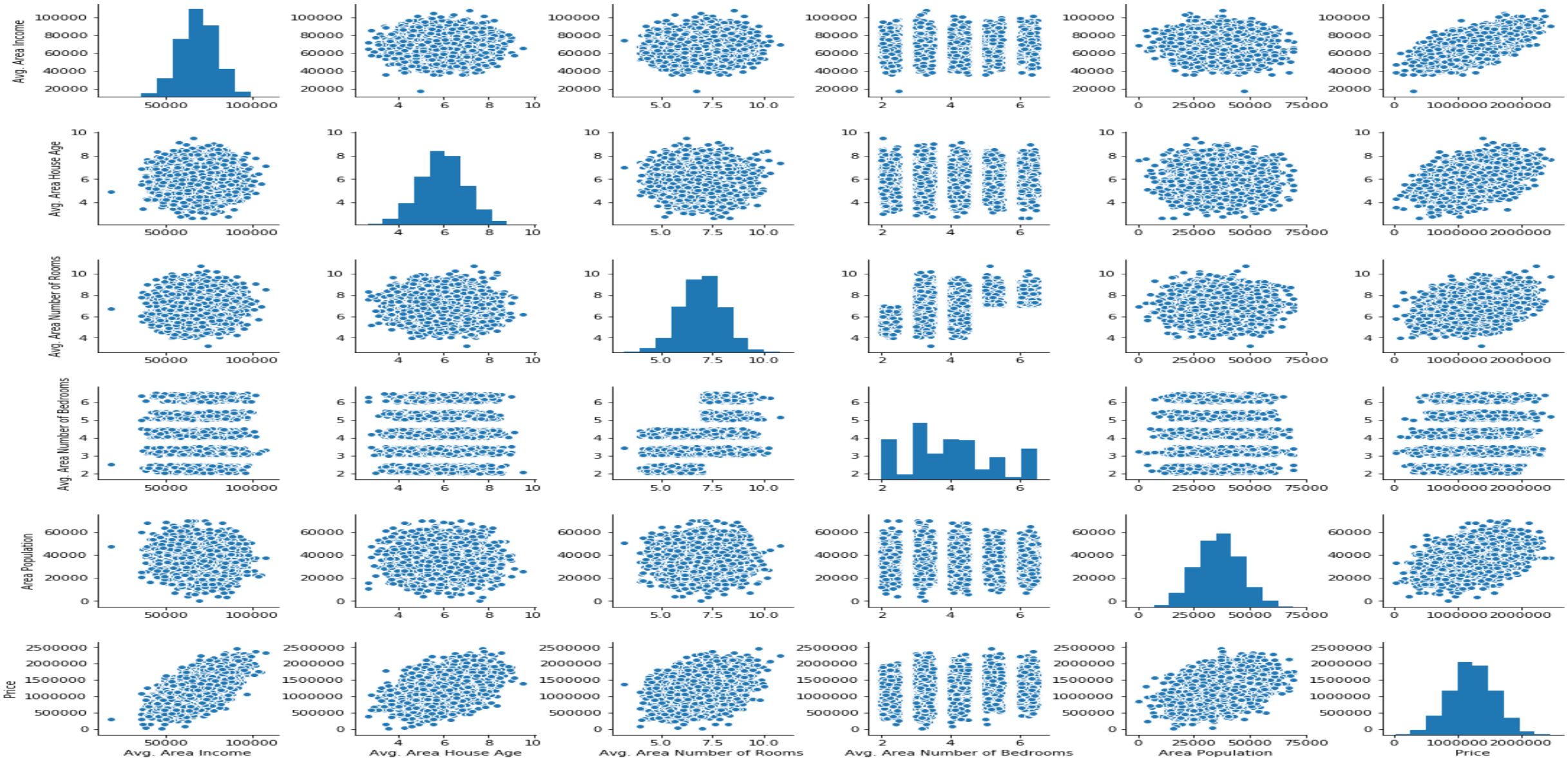
	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

TECHNICAL INFORMATION

- Data columns (total 7 columns):
- Avg. Area Income 5000 non-null float64
- Avg. Area House Age 5000 non-null float64
- Avg. Area Number of Rooms 5000 non-null float64
- Avg. Area Number of Bedrooms 5000 non-null float64
- Area Population 5000 non-null float64
- Price 5000 non-null float64
- Address 5000 non-null object
- dtypes: float64(6), object(1)

ADDRESS IS A VARIABLE WHICH WILL NOT CONTRIBUTE TO THE MODEL ITSELF AND HENCE WILL BE DROPPED.

IS THE DATA LINEAR OR NON-LINEAR?



ANSWER:

YES, THE DATA IS LINEAR AS SHOWN BY THE PAIRPLOT OF THE PREVIOUS SLIDE WHICH SHOWS THAT ALL THE INDEPENDENT VARIABLE HAVE A LINEAR RELATIONSHIP WITH THE DEPENDENT VARIABLE ‘PRICE’ EXCEPT FOR THE VARIABLE - AVG. AREA NUMBER OF BEDROOMS, WHICH WILL BE DROPPED WHILE CREATING THE REGRESSION MODEL.

MISSING VALUES:

- Avg. Area Income 0
- Avg. Area House Age 0
- Avg. Area Number of Rooms 0
- Avg. Area Number of Bedrooms 0
- Area Population 0
- Price 0
- Address 0

THERE ARE NO MISSING VALUES PRESENT IN THIS DATA.

OUTLIERS- COMPARING MEAN AND MEDIAN

MEAN

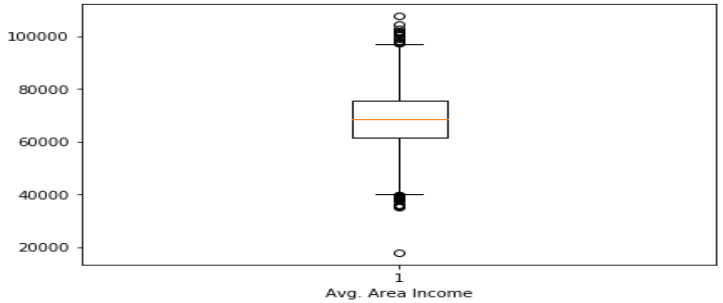
Avg. Area Income	6.858311e+04
Avg. Area House Age	5.977222e+00
Avg. Area Number of Rooms	6.987792e+00
Avg. Area Number of Bedrooms	3.981330e+00
Area Population	3.616352e+04
Price	1.232073e+06

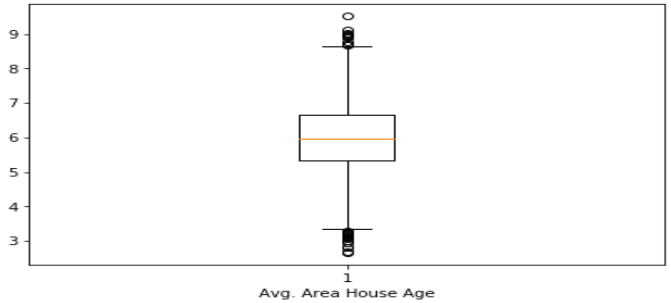
MEDIAN

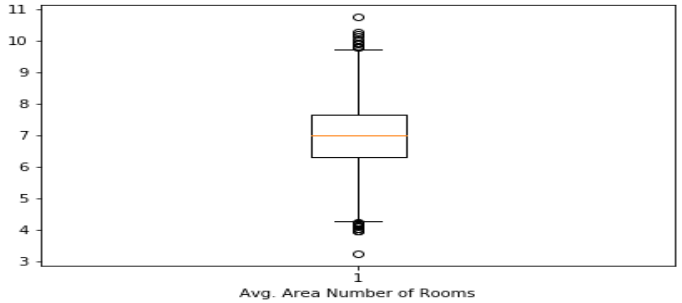
Avg. Area Income	6.880429e+04
Avg. Area House Age	5.970429e+00
Avg. Area Number of Rooms	7.002902e+00
Avg. Area Number of Bedrooms	4.050000e+00
Area Population	3.619941e+04
Price	1.232669e+06

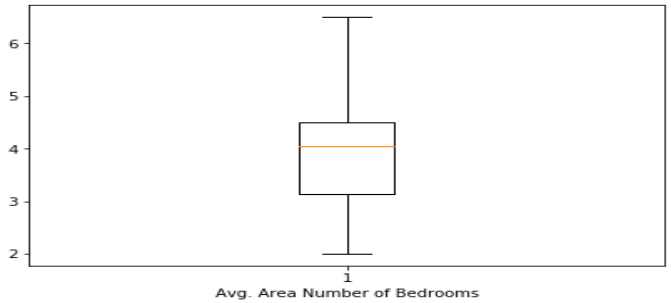
NOTE -- Avg. Area Number of Bedrooms and Avg. Area Number of Rooms may have Outliers

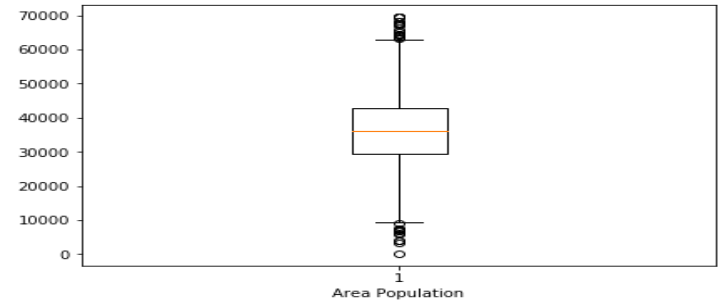
OUTLIERS - BOXPLOTS

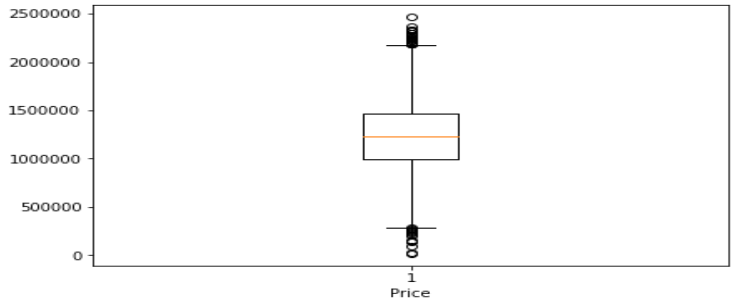








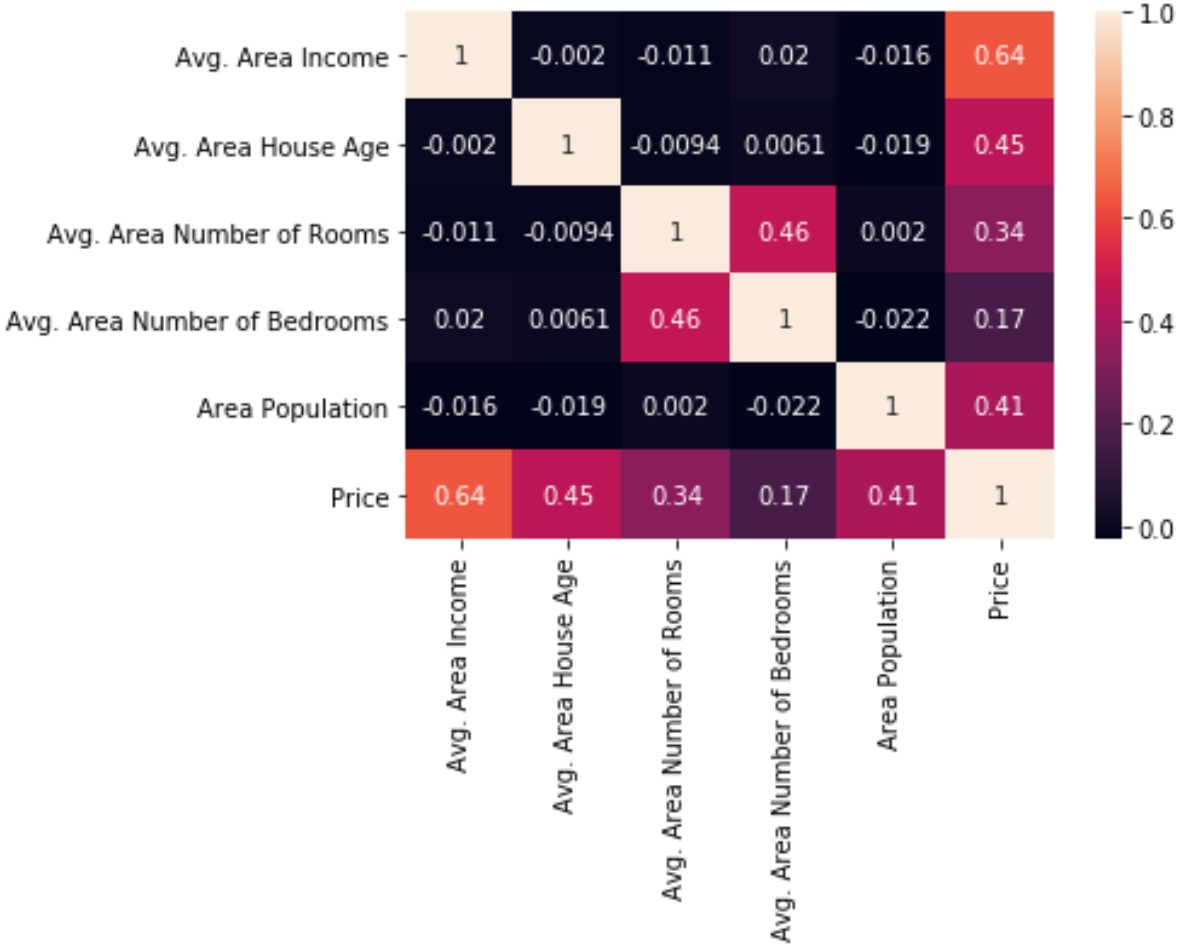




THE BOXPLOTS SHOWS THAT THERE ARE OUTLIERS IN EACH VAIRABLE. THUS THE VALUES OUTSIDE OF THE WHISKERS ARE CAPPED TO THE VALUE OF THE WHISKER ITSELF.

THE VALUE OF THE WHISKERS IN THE BOXPLOT IS CALCULATED AS:
INTER-QUARTILE RANGE +/- 1.5

THE CORRELATION HEATMAP



FINDINGS:

THERE IS A STRONG RELATIONSHIP BETWEEN AVG. AREA NUMBER OF ROOMS AND AVG. AREA NUMBER OF BEDROOMS.

THIS RELATIONSHIP CAN BE EXPLAINED THAT BEDROOM ITSELF IS A SUBSET OF ROOM HENCE THE COUNT ROOMS AND THE COUNT OF BEDROOMS WILL INCREASE WITH INCREASE IN ONE OF IT'S VALUES BY 1 UNIT. THIS CREATES MULTI-COLLINEARITY IN THE DATA SET.

ONE OF THEM HAS TO BE DROPPED AND THE NO. OF BEDROOMS HAS SHOWN ENOUGH EVIDENCE FROM THE PAST SLIDES AS TO WHY IT SHOULD BE DROPPED. THUS THE VARIABLE AVG. AREA NUMBER OF BEDROOMS WILL BE DROPPED.

LINEAR REGRESSION MODEL:

OLS Regression Results

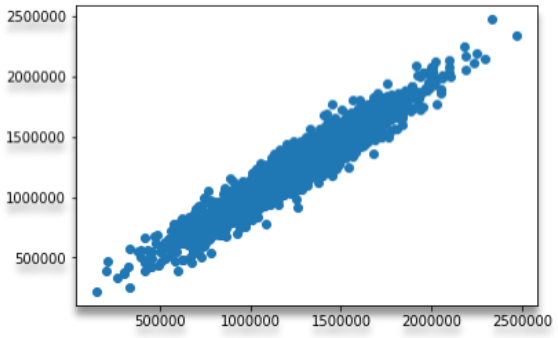
Dep. Variable:	Price	R-squared:	0.964
Model:	OLS	Adj. R-squared:	0.964
Method:	Least Squares	F-statistic:	2.217e+04
Date:	Wed, 02 May 2018	Prob (F-statistic):	0.00
Time:	10:42:03	Log-Likelihood:	-46312.
No. Observations:	3350	AIC:	9.263e+04
Df Residuals:	3346	BIC:	9.266e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Avg. Area Income	10.2380	0.330	31.069	0.000	9.592	10.884
Avg. Area House Age	5.105e+04	3653.293	13.974	0.000	4.39e+04	5.82e+04
Avg. Area Number of Rooms	-8270.5799	3380.972	-2.446	0.014	-1.49e+04	-1641.598
Area Population	8.2908	0.403	20.598	0.000	7.502	9.080

Omnibus:	3.034	Durbin-Watson:	1.999
Prob(Omnibus):	0.219	Jarque-Bera (JB):	2.981
Skew:	-0.048	Prob(JB):	0.225
Kurtosis:	2.889	Cond. No.	7.78e+04

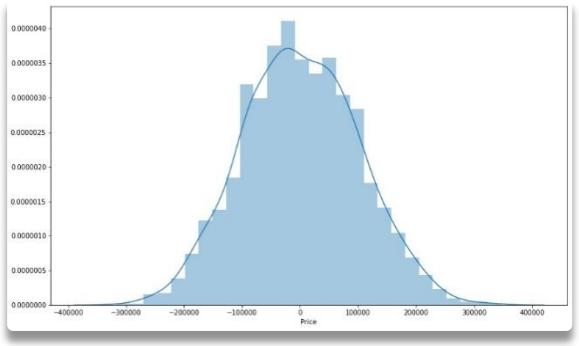
Mean Absolute Percentage Error: 7.405043178356214

ASSUMPTIONS OF LINEAR REGRESSION



HOMOSCEDASTICITY

THE ABOVE GRAPH SHOWS THAT THE DATA IS PERFECTLY HOMOSCEDASTIC.



NORMAL RESIDUALS

THE ABOVE GRAPH SHOWS THAT THE RESIDUALS(DIFFERENCE BETWEEN ACTUAL AND PREDICTED PRICES) IS NORMALLY DISTRIBUTED.

1.999

DURBIN – WATSON TEST

THE DURBIN WATSON STATISTIC LIES BETWEEN 1 AND 2 WHICH CLEARLY STATES THAT THERE EXISTS NO AUTO-CORRELATION.

BESIDES THESE THE ASSUMPTIONS OF LINEARITY AND MULTI-CORRELATION IS MET AT SLIDE 9 AND 13.

CONCLUSION

METRICS:

THE LINEAR REGRESSION SUMMARY SHOWS ADJUSTED R^2 IS 0.964 WHICH MEANS 96% OF THE INDEPENDENT VARIABLE IS CORRELATED WITH THE DEPENDENT VARIABLE. THE MODEL HAS A 7.4 % ERROR RATE. THESE TWO FEATURES INDICATE THAT THE MODEL IS GOOD ENOUGH TO PREDICT THE PRICE OF THE HOUSES.

ASSUMPTIONS:

ALL THE ASSUMPTIONS ARE MET AND HENCE IT CAN BE SAID THAT THE MODEL IS A GENERALIZED MODEL. WHICH MEANS THAT IT WILL GIVE AN IN-GENERAL PRICE OF THE HOUSE RATHER THAN THE ACTUAL. THIS IS SO BECAUSE OF THE PRESENCE OF THE ERROR-RATE. IT IS TO BE NOTED THAT NO MODEL IS WITHOUT ANY ERROR AND HENCE THE GENERALIZATION APPROACH IS CHOSEN AS THE BEST TOOL TO ANSWER ANALYTICAL QUESTIONS.

THANK YOU!

THE CODE FOR THIS PROJECT CAN BE VIEWED BY SENDING ME A MAIL AT: vivekc0395@gmail.com