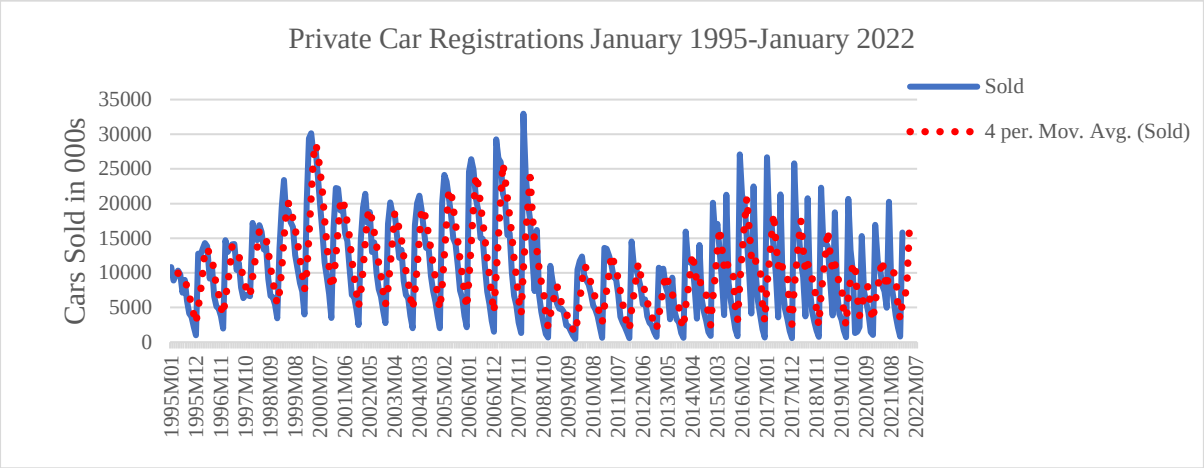
Higher Diploma in Science in Data Analytics (HDSDA_JAN22) Statistics 2

Kieran O'Hanlon

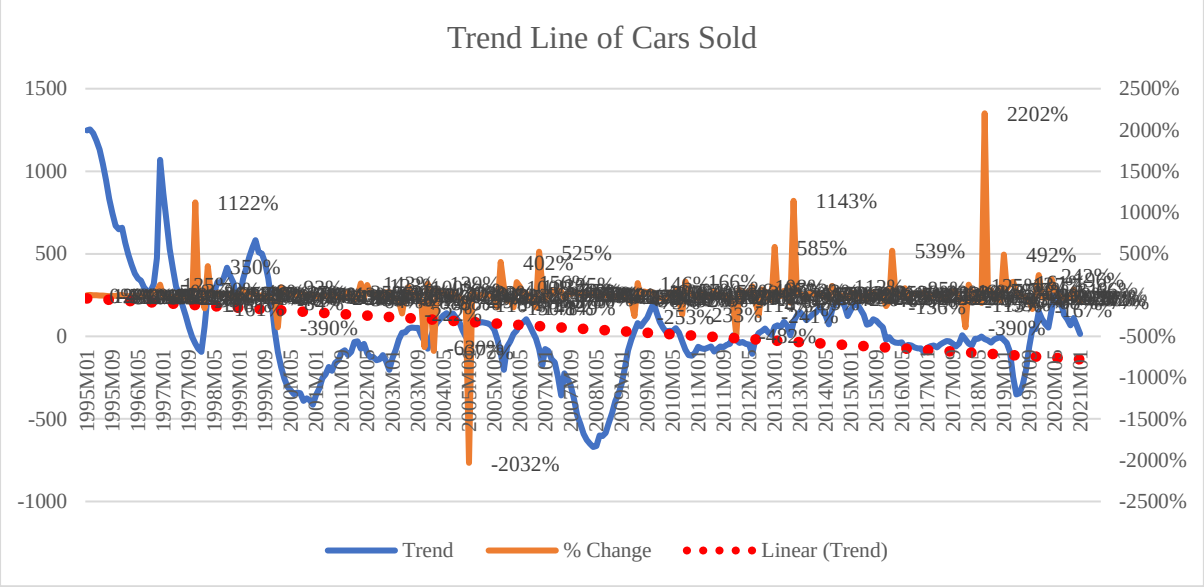
National College of Ireland

x21190020@student.ncirl.ie

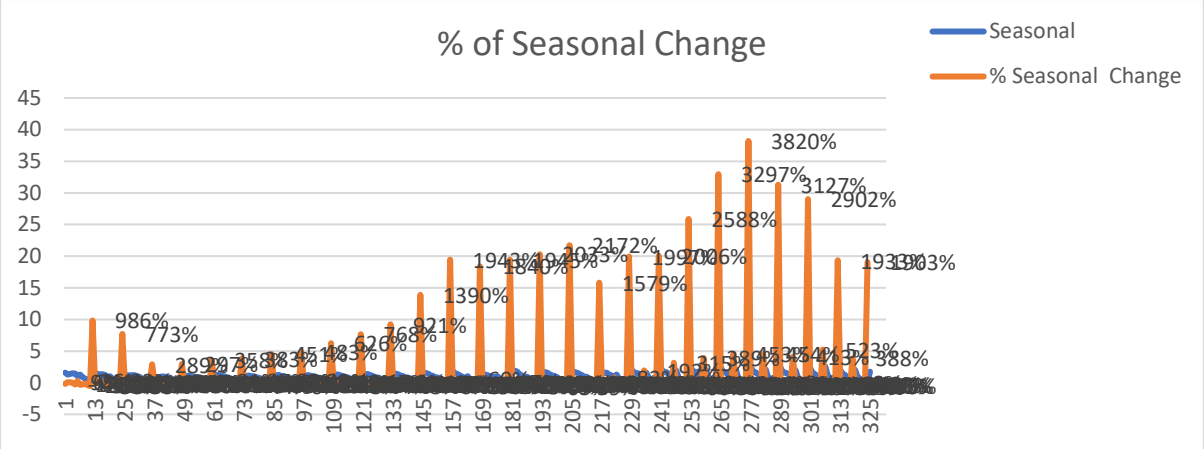
An initial observation from the data concludes there is a direct correlation between an increase in new cars registered in the years first month. In turn, there is a significant decrease in new cars registered in the last month of the year. What is also evident from the graph is the end of the Celtic Tiger era where 32,000 cars were registered in January 2008 to just under 11,000 the following year. This period signalled the start of the Great Recession.



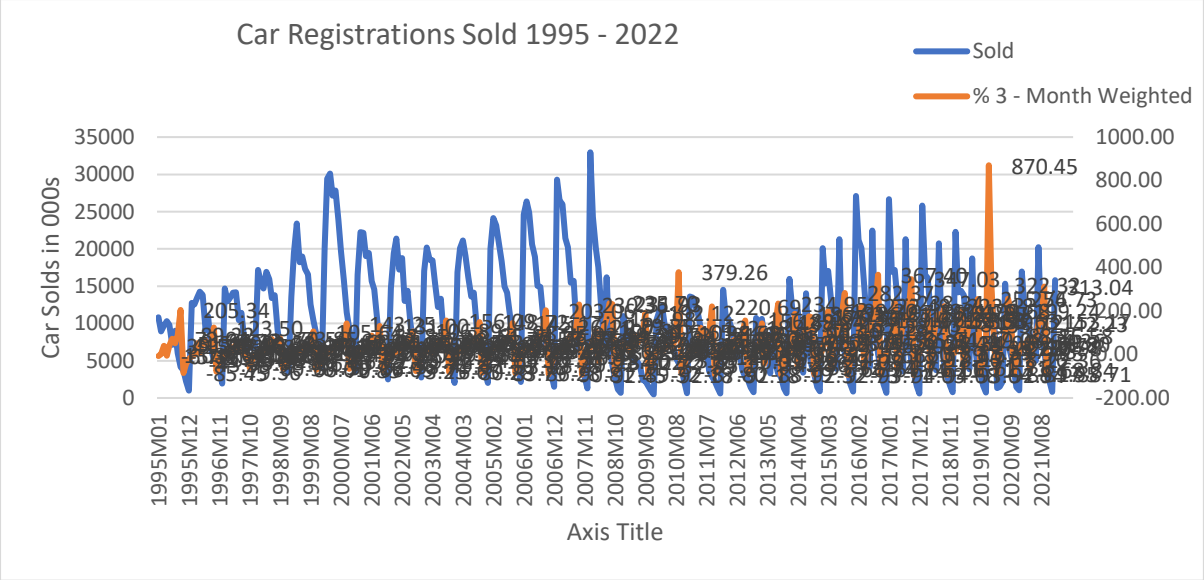
The trendline below further illustrates this with a 2000% decrease in 2005 and although car sales do recover most notably around 2019 the linear trend is downward.



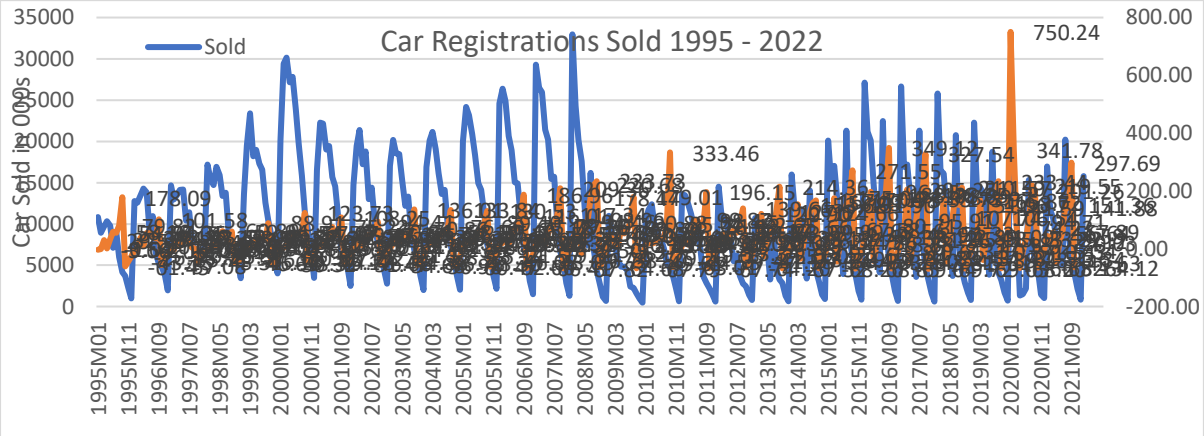
The seasonal chart illustrates the regular and predictable pattern that occurs each year from peaks in January to troughs in December. Companies that understand this pattern can plan and implement strategies accordingly for the best needs of the business. However, data must be adjusted accordingly, for example, in the car sales industry which has a significant pattern comparing January sales to the previous year's average sales companies in the industry may get a false impression.



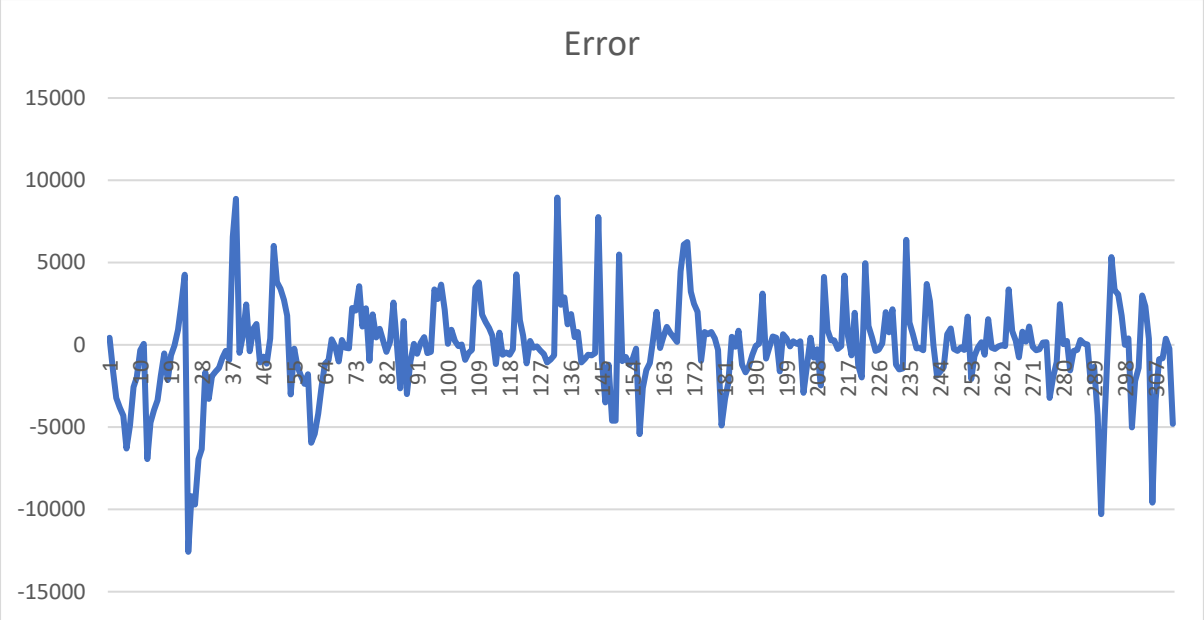
The 3-month weighted moving average percent is considerably less than the actual sales. For example, in January 1996 12773 cars were registered whereas the four month moving average predicted just under 2,000 cars would be registered. This is a difference of just over 85%.



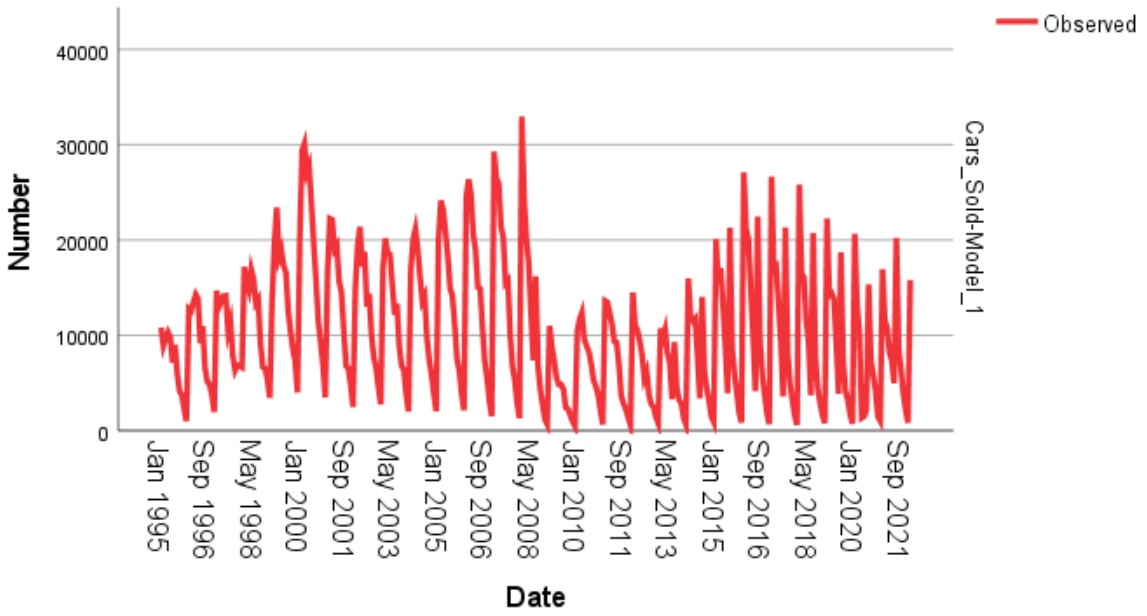
There is an improvement in the four-month moving average illustrated below compared to the three-month weighted average. Using the same month for illustrative purposes the four-month moving average predicted just under 5,000 cars would be a registered, a difference of 61%.



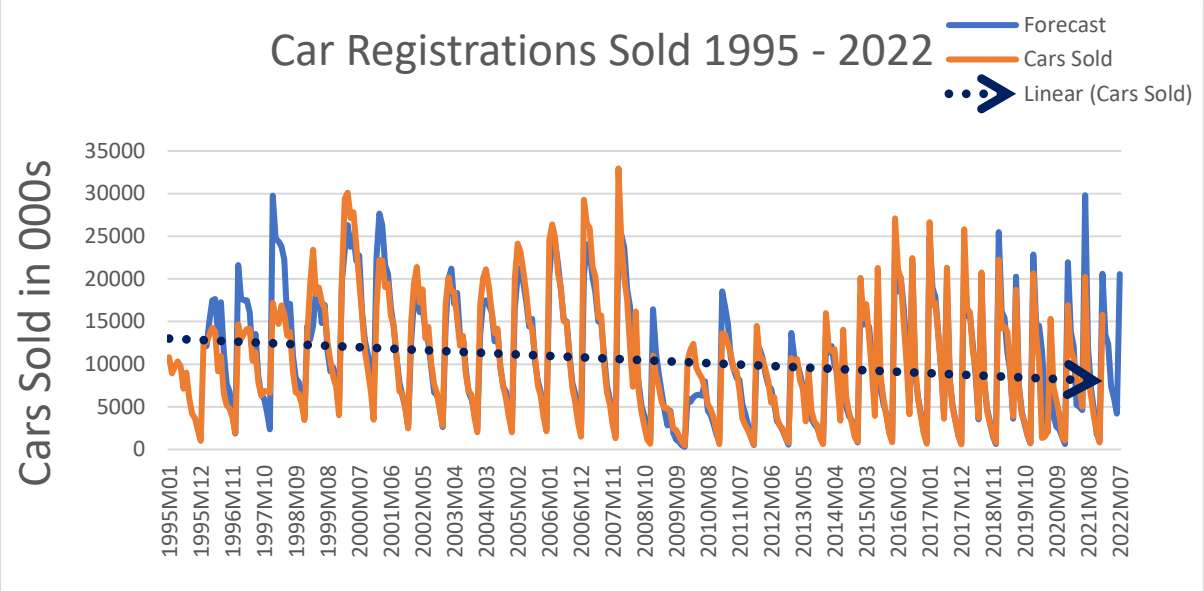
The error term is the difference between actual sales and forecasted sales. In general, though the line is centered around zero and with a MAPE of 20% it is reasonable to assume the forecast figure is accurate



Although the ARIMA model follows a similar cycle to the other models the Root Mean Square Error (RMSE) of over 7,000 and the MAPE figure of 165 are both obvious limitations of this model.



The forecasted data is very similar to the actual amount of cars registered over that period and for the following six months of 2022, the forecast is following the same seasonal pattern. The mean absolute percentage error (MAPE) suggests the forecasted figures are on average 21% away from the actual average. Because of the pattern and the very obvious seasonal element to the data, the Holt-Winters is the best model to work with. This method allows us to include seasonality while still being able to make the prediction along with the trend. The smoothing technique takes into account the average along with the trend and in doing this it can also make a time series prediction. One drawback of the model is the Root Mean Squared Error (RMSE) which is particularly high (2,690). The RMSE is an absolute error that squares the deviations to keep positive and negative deviations from offsetting each other.



Question 2 Part1

Introduction

The intent of this research is to determine how Principal Component Analysis (PCA) operates for data sets with a number of variables illustrating a trend. PCA is an algorithm that is used to reduce the number of dimensions and intricacies of a dataset. Although it reduces the number of variables through a reduction technique it still wants to retain as much intelligence as possible with regard to the reduced dimensions. This is possible as the intelligence is protected in the variance that the PCA explains. This research is to determine how real estate is valued as several factors can determine whether those factors are tangible or intangible (Sisman & Sisman, 2016). For the objective of this assignment, I will use the Boston housing dataset which I obtained from Github.

Background

By reducing the factor numbers more applicable factors are established as an element using PCA. In other words, PCA is a method that helps find the appropriate components that lie beneath different variables. This method extrapolates the greatest common variance from all the variables and puts them into a generic score. Like most statistical analyses there are several assumptions. For example, there is no multicollinearity, the data is linear, relevant variables are used and there is a true correlation between factors and variables.

The housing market contributes enormously to all major economies worldwide leading to a plethora of research in this field. Gaining insight into predicting house prices is essential for consumers who are looking to buy or sell their homes (Sarkar et al., 2021). Furthermore, a system for predicting the value of houses helps consumers who are buying and selling their homes to be on an equal level with investors, insurance companies, and real estate companies. One such way is through Hedonic price models which will help determine the relationship between prices and the features of the home (Rosen, 1974).

The Boston housing dataset was initially compiled by David Harrison Jr. of Harvard University and Daniel L. Rubinfeld National Bureau of Economic Research. The study wanted to consider the methodological difficulties of how much consumers were willing to pay for clean air with the use of the housing market data in Boston. However, the study is not without controversy with the variable black proportion of the population (B) with the authors reasoning that an increase in B will have a negative effect on housing value if Black people are considered undesirable neighbours by whites.

Data Description

The data set has 506 which is relatively small and has 14 variables (13 features and the median value (MEDV as the target variable) which are described in the appendices. From the descriptive statistics, it was determined

there are no missing values. I omitted two variables black proportion of the population and the dummy variable Charles River (CHAS). The median value per home is $22,000 with a standard deviation of $9,000

Results

Correlation Matrix[a]

		medv	crim	zn	indus	nox	rm	age	dis	rad	tax	lstat	ptratio
Correlation	medv	1.000	-.388	.360	-.484	-.427	.695	-.377	.250	-.382	-.469	-.738	-.508
	crim	-.388	1.000	-.200	.407	.421	-.219	.353	-.380	.626	.583	.456	.290
	zn	.360	-.200	1.000	-.534	-.517	.312	-.570	.664	-.312	-.315	-.413	-.392
	indus	-.484	.407	-.534	1.000	.764	-.392	.645	-.708	.595	.721	.604	.383
	nox	-.427	.421	-.517	.764	1.000	-.302	.731	-.769	.611	.668	.591	.189
	rm	.695	-.219	.312	-.392	-.302	1.000	-.240	.205	-.210	-.292	-.614	-.356
	age	-.377	.353	-.570	.645	.731	-.240	1.000	-.748	.456	.506	.602	.262
	dis	.250	-.380	.664	-.708	-.769	.205	-.748	1.000	-.495	-.534	-.497	-.232
	rad	-.382	.626	-.312	.595	.611	-.210	.456	-.495	1.000	.910	.489	.465
	tax	-.469	.583	-.315	.721	.668	-.292	.506	-.534	.910	1.000	.544	.461
	lstat	-.738	.456	-.413	.604	.591	-.614	.602	-.497	.489	.544	1.000	.374
	ptratio	-.508	.290	-.392	.383	.189	-.356	.262	-.232	.465	.461	.374	1.000
Sig. (1-tailed)	medv		<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
	crim	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	zn	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000
	indus	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000
	nox	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000
	rm	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000
	age	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000
	dis	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000
	rad	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000
	tax	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000
	lstat	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000
	ptratio	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	

a. Determinant = 5.209E-5

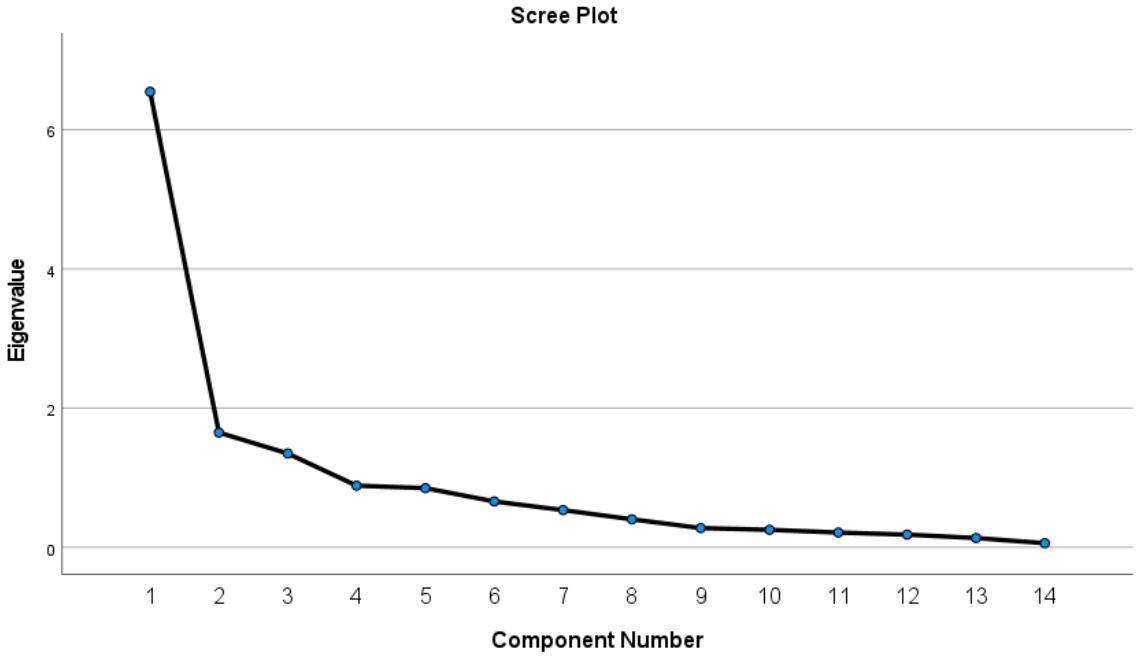
Although no variable coefficient is close to 1.000 in terms of the median house value all the variables are significantly correlated to each other. This is proven with the Bartletts test of sphericity as the P-value is < .05 and therefore we can reject the null hypothesis as the variables are significantly correlated. Per capita crime rate by town (CRIM) does imply that as crime rises house prices decrease. Average number of rooms per dwelling (RM) has the highest coefficient and again this is pretty intuitive that the more bedrooms the higher the price of the house. The lower status (LSTAT) is defined as an adult with less than a ninth-grade level of education or a male that works as a labourer (Harrison & Rubinfeld, 1978). With a significantly low coefficient (-.738) the study proves at a statistically significant level respondents with less education are less likely to own a home in the median value range. This is also true of the tax coefficient, the more property tax that is paid the less likely the respondents will have a home in the median value range. However, tax and access to radial highways are highly correlated (.910)

The total variance and scree plot deal with the factor extraction method. The Eigenvalue is an index of the strength of the component, in other words, the amount of variance it accounts for. By using an Eigenvalue of greater than one I have reduced the number of components down to two and a maximum of three. This is further illustrated in the scree plot with a significant drop initially and the variance of 47% explains this. However, the other two components have a very low variance with a small drop in the scree plot before it eventually tapers off. The cumulative variation of the three components accounts for 68.1% of the variation in the data.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.323	52.694	52.694	6.323	52.694	52.694	3.628	30.235	30.235
2	1.522	12.684	65.379	1.522	12.684	65.379	2.850	23.751	53.986
3	1.249	10.412	75.791	1.249	10.412	75.791	2.617	21.805	75.791
4	.822	6.852	82.643						
5	.539	4.493	87.136						
6	.405	3.371	90.507						
7	.281	2.342	92.849						
8	.263	2.194	95.043						
9	.214	1.779	96.823						
10	.184	1.534	98.357						
11	.136	1.135	99.492						
12	.061	.508	100.000						

Extraction Method: Principal Component Analysis.



Communalities

	Initial	Extraction
medv	1.000	.828
crim	1.000	.602
zn	1.000	.640
indus	1.000	.753
chas	1.000	.300
nox	1.000	.802
rm	1.000	.748
age	1.000	.743
dis	1.000	.828
rad	1.000	.840
tax	1.000	.846
ptratio	1.000	.446
b	1.000	.430
lstat	1.000	.737

Extraction Method: Principal Component Analysis.

A communality is the sum of the squared loadings and represents the amount of variance in that variable accounted for by all the components. Many of the variables do account for a high variance with several accounting for at least 70% of the variation. Two or more components were extracted meaning the components are uncorrelated. By running a bivariate correlation (see appendices) and with a P-value > .05 we fail to reject the null hypothesis that the factors are uncorrelated. This is also known as orthogonal rotation and why the varimax method is used for the purpose of this discussion. Furthermore, the component matrix is consistent with the total variance explained model. The loaded factors from column one are all > .5 with seven of the variables close to 80% or above. This further illustrates the high variance from the scree plot and the initial Eigenvalue percent of variance at close to 53%. From the two-component matrices (see below) the original variables are on the x-axis and the loaded factors are in the columns. I determined >.3 was sufficient for the absolute value which makes both figures easier to comprehend.

Component Matrix[a]

	Component 1	2	3
indus	.853		
nox	.833	.324	
tax	.823		.452
lstat	.798		
dis	-.770	-.464	
rad	.768		.539
age	.768	.320	
medv	-.681	.605	
zn	-.640		.470
crim	.610		.499
ptratio	.537	-.405	
rm	-.529	.648	

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

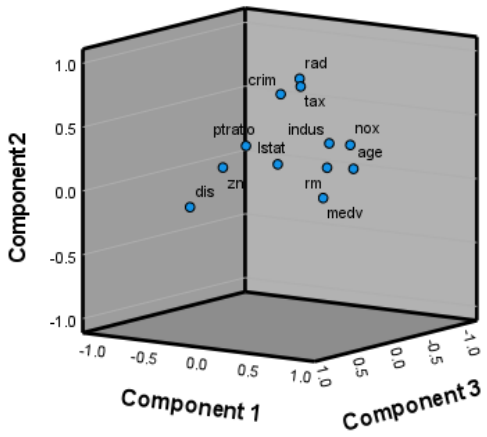
Rotated Component Matrix[a]

	Component 1	2	3
dis	-.889		
age	.825		
nox	.783	.414	
zn	-.753		
indus	.686	.438	.302
rad		.883	
tax	.360	.840	
crim		.752	
rm			-.872
medv			-.864
lstat	.451	.304	.672
ptratio		.393	.561

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Component Plot in Rotated Space



Discussion and Analysis of Results

The first component of quality of life included distance to employment centres (DIS), residential land zones (ZN), non-retail business acres (INDUS), a proxy variable for pollution (NOX), and proportion of owner-occupied dwellings (AGE). Clearly the closer a residential home is to employment centres or industry the higher the pollution in that area. The owner-occupied homes are most likely to be in the suburbs and generally would have further travel to the employment centres. Furthermore, this fits in with non-retail business acres as there would be more amenities, for example, parks, walkways et cetera away from industrial businesses.

In the second component per capita crime rate per town, accessibility to radial highway (RAD), and property tax are strongly correlated. This would imply the homes in this area have more value, pay more property tax because of this, and have access to the highway but on the other hand, easier access to the highway increases the likelihood of more crimes being committed.

The third component is status where the lower status of the population, the pupil-teacher ratio by town, the median value of homes, and the average number of rooms per dwelling are highly correlated Intuitively this stands to reason, for example, the more or fewer rooms in a dwelling will have an effect on the median value with the same reasoning being applied to the lower status of the population, the higher educated will be the above the median and conversely the less educated will be below the median. Furthermore, the pupil-teacher ratio will be higher in more affluent areas whilst fewer teachers to pupils are most likely to be in less affluent areas and those dwellings will again be below the median value.

One area for further analysis is the outliers in the dataset as this is one limitation of PCA analysis that the method is not robust against outliers. This was proven by running a Shapiro-Wilks test against all variables the P-values are <.01 meaning the results are statistically significant and therefore I can reject the null hypothesis that the variables are not normally distributed.

References

Das Sarkar, S.S., Ali, M.E., Yuan-Fang, L., Yong-Bin, K. and Timos, S., (2021). Boosting house price predictions using geo-spatial network embedding. *Data Mining and Knowledge Discovery*, **35**(6), pp. 2221-2250.

Harrison, D. and Rubinfeld, D.L. (1978) 'Hedonic housing prices and the demand for clean air', Journal of Environmental Economics and Management, 5(1), pp. 81–102.

Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. J Polit Econ 82(1):34–55

Sisman, Yasemin & Sisman, Aziz. (2016). Principal Component Analysis Approach in the Determination of House Value. PONTE International Scientific Research Journal. 72.

Appendices

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town
- B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.853
Bartlett's Test of Sphericity	Approx. Chi-Square	5132.764
	df	91
	Sig.	.000

Correlations

		Cost_Value	Area_Value	House_Value
Cost_Value	Pearson Correlation	1	.000	.000
	Sig. (2-tailed)		1.000	1.000
	N	506	506	506
Area_Value	Pearson Correlation	.000	1	.000
	Sig. (2-tailed)	1.000		1.000
	N	506	506	506
House_Value	Pearson Correlation	.000	.000	1
	Sig. (2-tailed)	1.000	1.000	
	N	506	506	506

Question 2 Part 2

Introduction & Background

After reducing the initial dataset to three variables I decided to run those three variables against the median house value to determine if they had an effect on the median house value. Multiple Linear Regression uses at least two independent variables which will help predict the outcome of the dependent variable. A simple example in this context would be the location of the property, what amenities are nearby, and the actual size of the dwelling (Limsombunc et al., 2004). All of these variables will have a direct effect on a person's utility. Construction and renovation play an enormous part in any economy. The industry also affects the demand for ancillary industries such as the supply of construction goods as well as durables for the household. (Li et al., 2011).

Multicollinearity

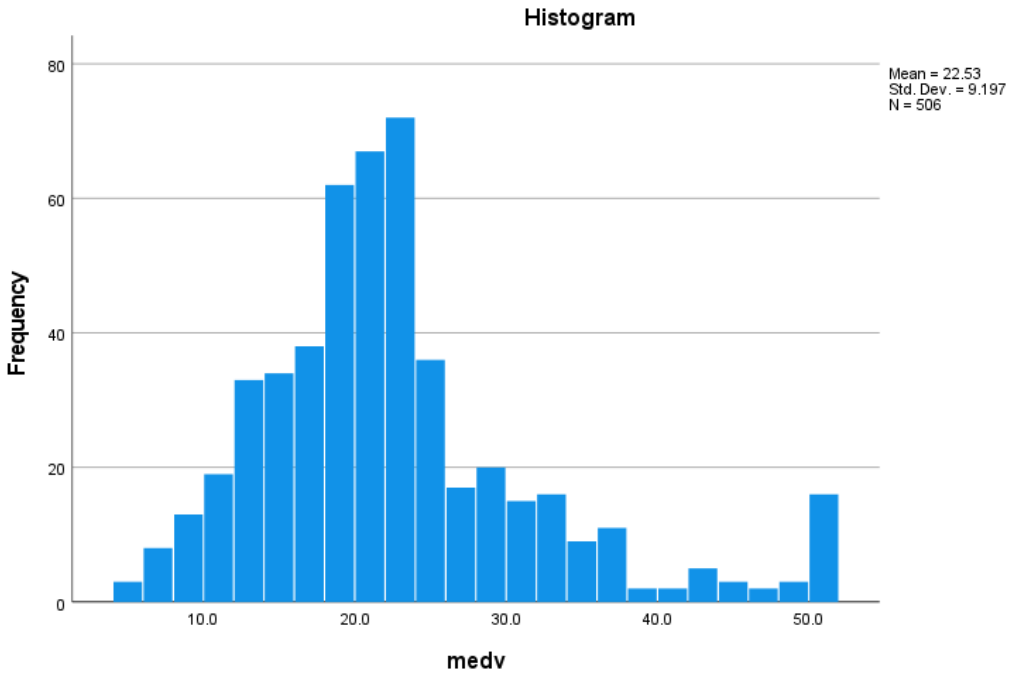
Multicollinearity can cause several problems in regression analysis. This occurs when one or more of the independent variables are highly correlated as it will make the inferences less dependable. An example of this would be a person's height and weight would be highly correlated and most likely it would be better to use the person's body mass index (BMI). From the correlation table (below) and with a P-value > .05, we fail to reject the null hypothesis that the factors are correlated. The general rule is that the correlations are < .8 for the purpose of linear regression. Unfortunately from the same output, it is also quite clear two of the independent variables are not correlated with the dependent variable as neither are > .03. The crime variable is highly correlated with a coefficient of -.864.

Correlations

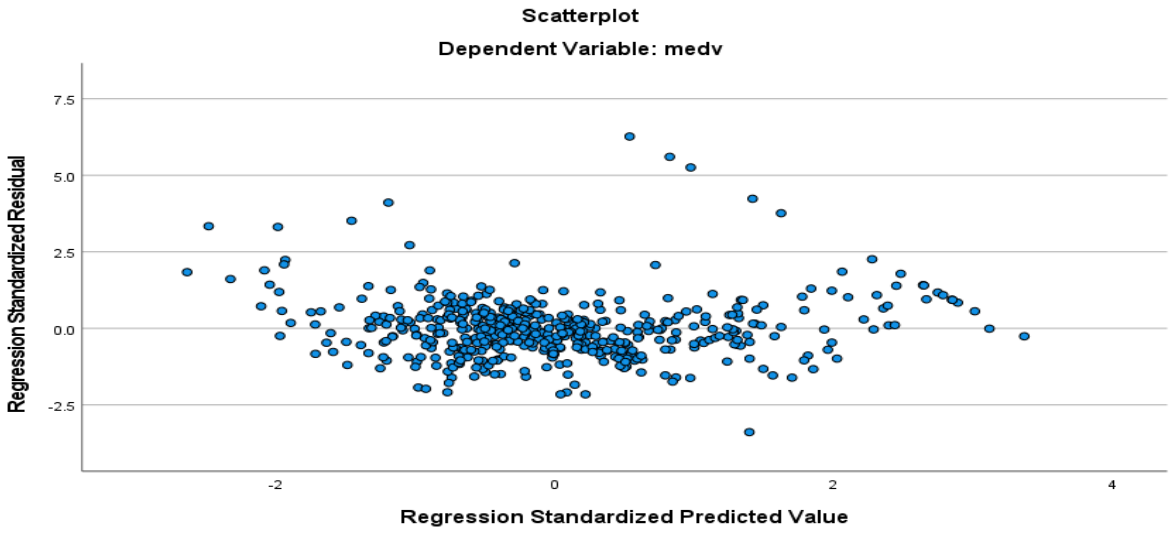
		medv	Quality_Life	Social_Status	Crime
Pearson Correlation	medv	1.000	-.189	-.247	-.864
	Quality_Life	-.189	1.000	.000	.000
	Social_Status	-.247	.000	1.000	.000
	Crime	-.864	.000	.000	1.000
Sig. (1-tailed)	medv	.	<.001	<.001	<.001
	Quality_Life	.000	.	.500	.500
	Social_Status	.000	.500	.	.500
	Crime	.000	.500	.500	.
N	medv	506	506	506	506
	Quality_Life	506	506	506	506
	Social_Status	506	506	506	506
	Crime	506	506	506	506

Results

One of the assumptions for Multiple Linear Regression (MLR) is that there are at least 20 observations for each independent variable but only if the dependent variable is normally distributed. Another assumption of MLR is that there is a linear relationship between the independent variables and the dependent variable. There must be an absence of outliers in all the variables and the variance of the error term is constant. This is known as homoscedasticity. By looking at the scatterplot and the residual statistics this is evident that this assumption is not met as the standard residual ranges from -3.3 to 6.2 meaning there is an unequal scatter which is also known as heteroscedasticity. By looking at the histogram of the dependent variable median value I can clearly see the distribution is not normal. This was also confirmed by the Q-Q plot and the Shapiro-Wilks (see appendices) test as the P-value is < .001 and therefore statistically significant and I can reject the null hypothesis that the dependent variable is not normally distributed.



By looking at the scatterplot in figure (see below) the illustration is showing us that the plots are not uniformly distributed as they range from -2.5 to 7.5 and there is not a linear relationship between the independent and dependent variables.



The adjusted R-square explains 84.2% of the variance in the dependent variable which is statistically significant as the F-stat is < .001 and this is a statistically significant finding. The variation in the dependent variable is explained by the R-square and ideally, this estimate needs to be as close to the population value as possible.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics			
					R Square Change	F Change	df1	df2	Sig. F Change
1	.918[a]	.843	.842	3.6599	.843	895.650	3	502	<.001

a. Predictors: (Constant), Lifestyle, Cost, Area_Value

b. Dependent Variable: medv

The ANOVA analysis checks the null hypothesis that the slope of the line is zero. This is also statistically significant as the P-value <.001 which is significant at an alpha level of .05 with three degrees of freedom and therefore I can reject the null hypothesis that at least one of the variables is statistically different from the others.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35991.956	3	11997.319	895.650	<.001[b]
	Residual	6724.339	502	13.395		
	Total	42716.295	505			

a. Dependent Variable: medv

b. Predictors: (Constant), Lifestyle, Cost, Area_Value

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon$$

y = 22.533 - 1.73 * quality life - 2.27 * social status – 7.94 * crime

The null hypothesis at an alpha level of .05 in this scenario will look like the following:

μ: the independent variables have no effect on the dependent variable

μ1: the independent variables do have an effect on the dependent variable

From the results with one three degrees of freedom, the independent variables are statistically significant at the 95% level and therefore we can reject the null hypothesis and state that the independent variables have an effect on the dependent variable. The crime variable indicates that for every one-unit increase in crime, the median house value will decrease by 7.94. This is an intuitive assumption, especially as access to highways is correlated with this variable which would make it easier for criminals to escape. The quality of life variable also has a negative coefficient implying that for every one-unit increase in pollution, the less likely residential homes will be located here and their value will decrease by 1.73 units. Moreover, the social status coefficient is indicating that the fewer bedrooms in a dwelling, the fewer teachers per student, and the lower the education of the respondents the median house value decrease by 2.27 units on average. By looking at the part correlations crime has the biggest impact on the dependent variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	22.533	.163		138.490	.000			
	Quality_Life	-1.735	.163	-.189	-10.651	<.001	-.189	-.429	-.189
	Social_Status	-2.271	.163	-.247	-13.946	<.001	-.247	-.528	-.247
	Crime	-7.944	.163	-.864	-48.775	<.001	-.864	-.909	-.864

a. Dependent Variable: medv

Discussion and Analysis of Results

Although the P-values of the independent variables are statistically significant and the model has a high R-square of 84% some of the basic assumptions of MLR have not been met. There is no linear relationship between the dependent and independent variables and the dependent variable is not normally distributed. For some reason, the Census Service capped the price at $50,000 for houses in the Boston area. For further research, there is certainly a need to try and improve these factors but from an initial point of view, the data does explain the effects of house prices in Boston. Intuitively as crime rises in a neighbourhood it would be reasonable to predict the price of houses in that area will decrease. In a general sense, many variables such as the neighbourhood, how many bedrooms and bathrooms as well as the distance to schools and local transport can give an accurate prediction of house prices. The housing market is a large part of any economy and with all this data being so transparent this can only help both buyers and sellers in this market. In terms of machine learning for this analysis the linear method is not the best choice but other algorithms such as the generalised additive model (GAM) is possibly a better fit. This model allows for nonlinear relationships and most likely will predict the coefficients with a higher degree of accuracy.

An area for further research is the high R-square as two of the independent variables are not correlated with the dependent

References

Harrison, D. and Rubinfeld, D.L. (1978) 'Hedonic housing prices and the demand for clean air', Journal of Environmental Economics and Management, 5(1), pp. 81–102.

Li, Y., Leatham, D. J., et al. (2011). Forecasting Housing Prices: Dynamic Factor Model versus LBVAR Model.

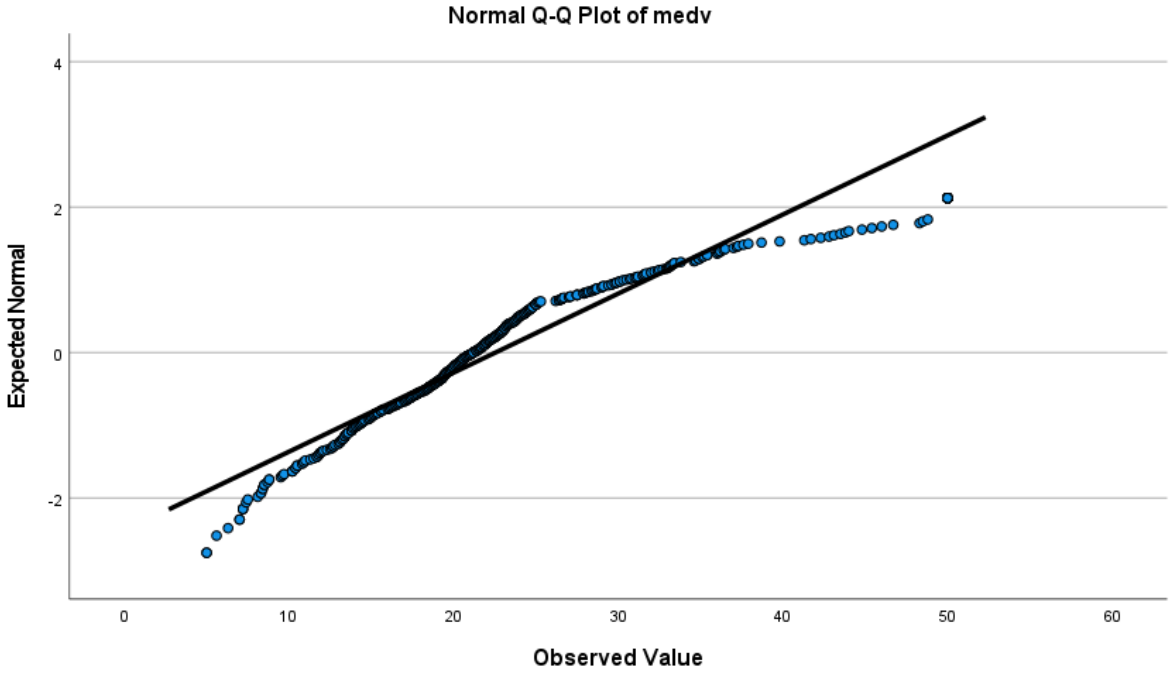
Limsombunc, V., Gan, C., and Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. American Journal of Applied Sciences, 1(3):193–201.

Appendices

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
medv	.149	506	<.001	.917	506	<.001

a. Lilliefors Significance Correction



Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.277	50.944	22.533	8.4422	506
Residual	-12.4039	22.9417	.0000	3.6490	506
Std. Predicted Value	-2.636	3.365	.000	1.000	506
Std. Residual	-3.389	6.268	.000	.997	506

a. Dependent Variable: medv