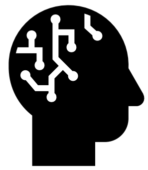
John T Foxworthy Winter 2022

Image by Ariel Skelley of Getty Images

Forecasting Single Family Residential Property Prices with the Long Short – Term Memory Model:

A Practical Application for the Real Estate Industry with Artificial Intelligence

Abstract



The pricing and forecasting of homes are complex and challenging, and we look into the Deep Learning algorithm of LSTM in three different cities of their price history to gain insight. The East Coast city of Boston, the Mid-West city of Chicago, and the West Coast city of Los Angeles are evaluated of their past median sales prices over 35 years for single-family residential properties. A deep dive in Exploratory Data Analysis is done first with Statistical Tests on the nature of the time series followed by three different LSTM experiments for each of the three cities. We find the LSTM did perform strongly but with concerns of overfitting.

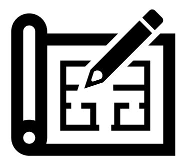
Introduction to Residential Real Estate Market and Modeling

Pricing a single-family residential home and forecasting its price is complicated, and Feature Engineering can appear to be an impossible task. The drivers of home price can vary in quality, quantity, frequency, and more. Making an entire list of features that drive a property price is exhausting and long. Rather than finding a dataset of various columns to explain a separate target of the price of Single-Family Residential Property, we can use Time Series Modeling. Past property sales have features embedded as a standalone one–dimensional array representing a set of real-world observations. Statistically, the univariate time series data will have endogenous predictability. In other words, a single column data series has all the information we need to explain and predict the property prices of single-family residential homes in the U.S. for the past, present, and future.

There are several time series models, but we have selected the Long Short–Term Memory (LSTM)[1]. In the early 1990s, two German computer scientists, Sepp Horchreiter and Jurgen Schmidhuber, created LSTM, which provided a valuable contribution to the subjects of Deep Learning and Artificial Intelligence. Precisely, the LSTM captures the time series behavior by learning the order of dependence in sequence prediction problems, which suits the nature of the residential housing market. *If we define Artificial Intelligence as a non – biological entity behaving as a biological entity, then LSTM is behaving as a pricer and forecaster of Housing*

Prices. For example, it is common for real estate practitioners to refer to the attributes of single-family residential properties sold and resold over the years, quantitative and qualitative. Below is an outline of a few features that LSTM will attempt to capture.

Quantitative Features in Single–Family Residential Property Prices



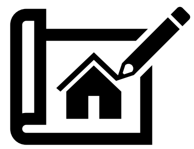
⇒ The economics of a local economy is growing at the zip code level with raising incomes wanting a purchase a larger home, and pushing up property prices.

- Or vice versa, the national or state economy outpacing a particular zip code causes falling prices in a residential area.

⇒ The demography of a positive inflow of workers, such as the growing technology sector of Austin, Texas right now, raises the population level locally.

- Or the negative outflow of workers of the rust belt states in the Mid–West because of the closing of many factories in manufacturing.

Qualitative Features in Single–Family Residential Property Prices



⇒ Amenities like a swimming pool, fireplace, garden, balcony, and terrace are not considered in the pricing of the property and add value to the purchase of a property. The home is sold at a price that is higher than expected.

- The opposite would be the absence of amenities compared to all the other rival properties in a particular residential area, causing a loss in value. The home is sold at a lower price than expected.

⇒ A renovation of an up-to-date kitchen or a property torn down and rebuilt, causing the property's age to be the current year coupled with up-to-date construction materials. Several bids push the price higher than realized.

- An overpriced home for sale has long been ignored and needs substantial renovation or needs to be torn down. The home is sold for a lower price.

Furthermore, the overwhelming majority of residential properties are not for sale because people live in them, so we cannot obtain price observations today or for several years in the past. A data scientist can only use the past sales of properties and not a list of every parcel in the country. We can estimate house prices that are not for sale as a workaround for the lack of observations. Still, the complexity of quantitative and qualitative features that are consistent, intermittent, dormant, ephemeral, and more make price discovery challenging. Given the direct nature of the U.S. Real Estate Sector, then a time series models LSTM is reasonable and justified.

Introduction to the Dataset of Housing Prices Indices

The dataset we are using is the Federal Reserve Economic Data (FRED), a public entity of the U.S. Central Bank that vetted a private data vendor on the Standard and Poor's Case – Shiller Home Price Indices. Robert Shiller of the latter named index is a noble prizing-winning economist to add value to the Data Integrity issues mentioned before since there are no sales of each home in the U.S. every single month. Therefore, I chose major cities with the monthly median sale price data, such as Boston, Chicago, and Los Angeles, from January 1987 to December 2021. There are also quarterly data for other cities because the frequency of sales is less. Overall, the primary objective is to create a baseline model for different parts of the country with three sets of 420 observations. All datasets are indexed to 100 as of January 2000 and are seasonally adjusted.

Literature Review of LSTM and Property Pricing

The residential real estate sector research is flooded in the tradition of classical statistical analysis like linear regression, so finding LSTM research on monthly housing prices is rare. Unfortunately, outside the U.S. is the only solution, like Sweden and Turkey. Finding LSTM articles is not tricky but ignoring the nature of the real estate market mentioned earlier is wrong. Nonetheless, our literature review with real estate datasets is a bachelor's thesis in statistics from Uppsala University in Sweden and an article in the Journal of Business Economics and Management about Turkish housing sales. There is universality in the quantitative features of higher incomes and lower unemployment rates pushing property prices up in countries outside the U.S. coupled with the qualitative features of amenities and a renovation. The methods result further below will demonstrate the global nature of residential real estate.

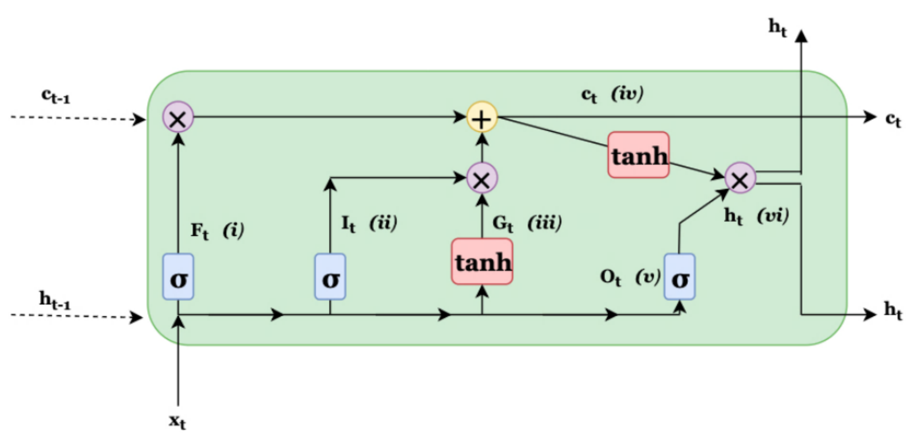


Image by Hansson, Fredrik and Rostami, Jako on a LSTM cell

Fredrick Hansson and Jako Rostami[2] evaluated the performance of different forecasting methodologies from a contemporary, not classical, approach on the monthly house prices in the most densely populated areas of Sweden. First, the modern regression modeling technique of class boundaries of the Support Vector Machine (SVM) and the deep learning of LSTM is compared. After that, there are comparisons of the classical Seasonal Autoregressive Integrated Moving Average (SARIMA) model that outperformance both SVM and LSTM, but there are practical concerns. For example, LSTM makes no assumptions about the monthly dataset over SARIMA, giving the advantages of restudying and reapplication over other models. Also, both Hansson and Rostami admit, more importantly, the data preprocessing of LSTM might have reduced the accuracy, and their choice of hyperparameters is open to question. In other words, LSTM may have fitted differently and given a better result in the accuracy scores and requires further research.

Melek Akgun, Ayse Soy Temur, and Gunay Temur[3] try to predict housing sales, not prices, in the 81 provinces in Turkey. Like the article before, they use a classical model ARIMA and compare the results with LSTM and their Hybrid model. We assume the change in the classical model from Seasonal ARIMA (SARIMA) to ARIMA may be due to the lack of seasonality in the housing datasets in Turkey because of the warmer climate than Sweden. Technically, the research attempts to address normality, linearity, and stationarity issues in housing sales. For example, does the time series go up as much as it goes down symmetrically like a normal distribution? Are there any functions in the time series that require an exponent, so it is non–linear sometimes or not at all? Is the common tendency and variation, i.e., the average and standard deviation, stable over time? The results confirm that the Hybrid model, a mixture of the classical ARIMA and the contemporary LSTM, performed the best. Yet, there is no questioning of the chosen hyperparameters to fit the models, especially LSTM, thereby raising the validity of the results. It is difficult to remove the bias of classical statistical modeling in any real estate sector, regardless of the country.

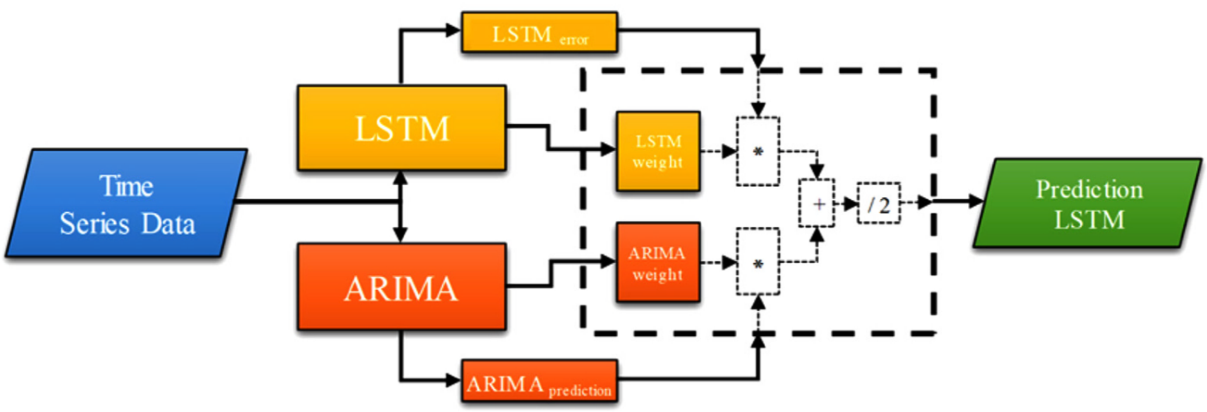
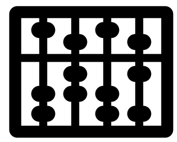


Image by Temur, Ayse Soy, and Temur, Gunay, and Akgun, Melek on the Hybrid Model

The critical takeaway from the literature is both historical and structural. For most of human history, residential real estate was linear and stable as an illiquid asset with little activity in sales. In the past two decades, a regime switch in the time series of housing has become less liquid, more sales, more investing options, and non–normal, non–linear, and non – stationary datasets. A relaxing in the laws of ownership so individuals and institutions can purchase more and securitize residential loans into financial instruments like Mortgage-Backed Securities (MBS) in the U.S. plays a role in the behavior of the times series of housing prices. In addition, the frequency of house sales has increased throughout the years. More importantly, hyperparameter tuning plays a critical role in the accuracy of the LSTM results on the pricing of single-family residential properties. Also, both research papers do not even have half the amount of monthly data in our dataset below.

Methods

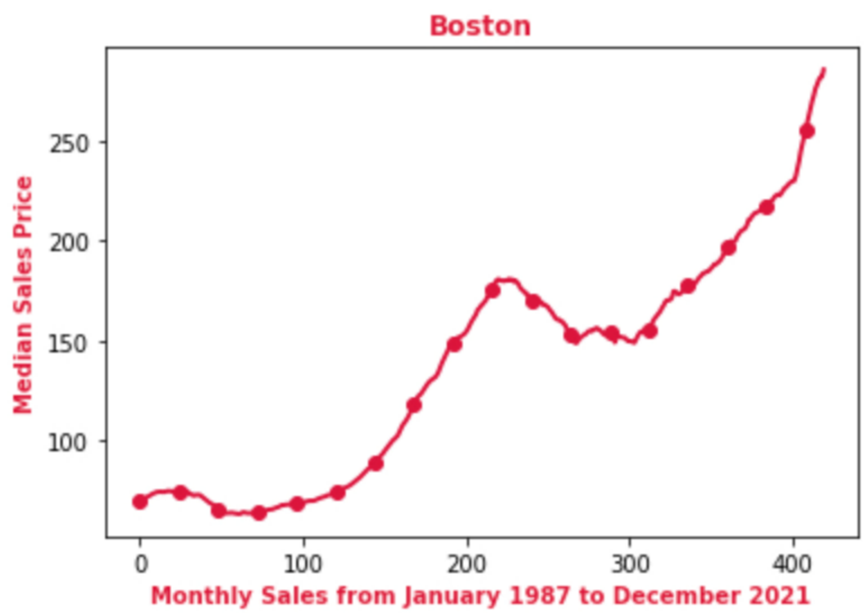


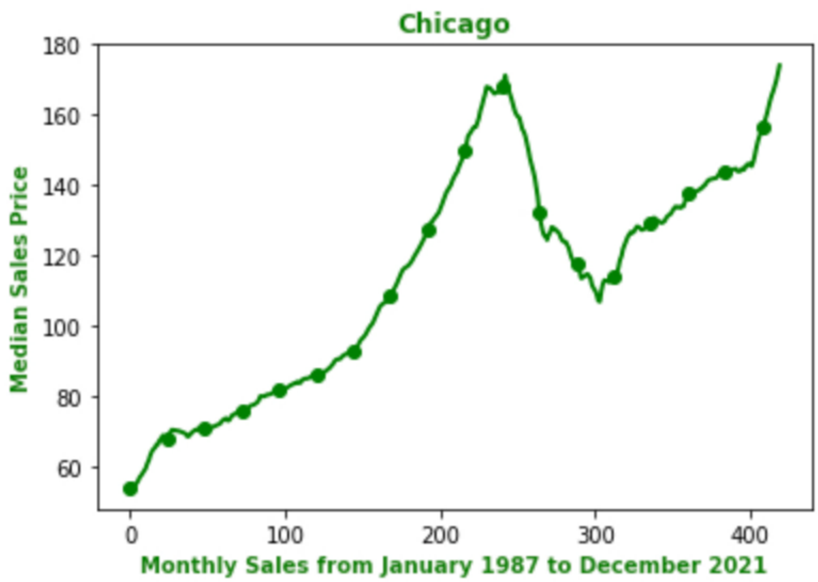
We used the Google Colab with the high processing power of a Hardware Accelerator set to TPU and a Runtime Shape of High – RAM. The first configuration with files beginning with 01A04 has 4 LSTM cells with one input shape feature and the input shape time step set to a look back of 1, coupled with one dense unit layer. In addition, the activation function was a rectified linear unit, 100 training iterations in the epochs, a single batch size, and the optimizer function is the adaptative moment estimation[4].

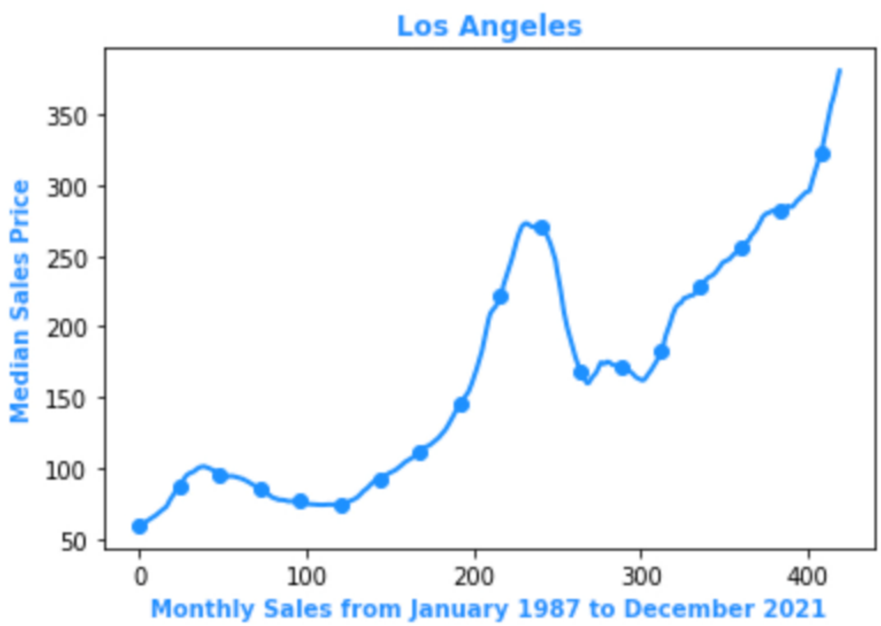
The second configuration is the reduction of the epochs from 100 to 5 to see how training impacts accuracy. The third configuration is the reduction of LSTM cells from 2 to 4 to see how the number of LSTM cell units contributes to accuracy.   Finally, we have a series of assessment metrics with a timer to evaluate the results   The file names beginning with 02A04 and 03A04 correspond to the second and third configurations. Altogether, all experiments have a 70% of the test set split into the training set and a Min-Max Scaler from 0 to 1 transformation.

However, before we begin, there is a series of Exploratory Data Analysis files beginning with 00_EDA to evaluate any insights that may help us evaluate the results. First, we conduct a hypothesis test on the presence or absence of stationarity with the Augmented Dickey-Fuller

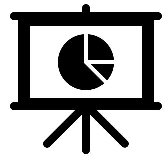
Test, and similarly on a normal distribution with the Jarque – Bera Test. The findings will help us with the historical changes in the time series, if any, with our model evaluation. The Literature Review raised some questions on the real estate market changes over the past decades and if that impacts our dataset and model.




Chicago
180
160
140
120
100
80
60
Median Sales Price
0
100
200
300
400
Monthly Sales from January 1987 to December 2021


Los Angeles
350
300
250
200
150
100
50
Median Sales Price
0
100
200
300
400
Monthly Sales from January 1987 to December 2021

Results



The Exploratory Data Analysis points to the historical changes in the real estate market as the first quarter of Boston, the first sixth of Chicago, and the first third of Los Angeles datasets have a stationary time series. Los Angeles begins with a normal distribution as median sale prices symmetrically rise as they fall, but both Boston and Chicago have weak non–normality in the same periods. The implication is that cities in the U.S. begin with stable and normally distributed home prices, then break out to non – stationarity and non–normality. Of course, we would need more tests on other cities for verification. Nonetheless, the first periods of the cities explain the tradition of relying on linear models, but the real estate market has changed, as mentioned earlier in the literature review.

	Augmented Dickey-Fuller Test			Jarque - Bera Test		
	ADF Statistic	p - value		JB Statistic	p - value	
Boston	1.0647	0.9949	**Fail to Reject Null**	17.6250	0.0000	**Reject Null**
Chicago	-0.6248	0.8653	**Hypothesis so non**	24.6670	0.0000	**Hypothesis so non -**
Los Angeles	0.5562	0.9865	**- Stationary**	31.763	0.0000	**Normal**

	Augmented Dickey-Fuller Test			Jarque - Bera Test		
	ADF Statistic	p - value		JB Statistic	p - value	
Boston (First Quarter)	-3.1713	0.0217	**Reject Null**	9.0590	0.0110	**Non - Normal**
Chicago (First Sixth)	-3.8176	0.0027	**Hypothesis so**	15.5200	0.0000	**Non - Normal**
Los Angeles (First Third)	-3.2149	0.0191	**Stationary**	4.3960	0.1110	**Normal Distribution**

Overall, the Long Short – Term Memory model performed strongly. Unlike the literature reviews from both Sweden and Turkey that did not vary hyperparameters and left their configuration an open question, our performance summary in the appendix outlines some critical findings. First, the goodness of fit measure, the R – Squared, was north of 0.90 with all the experiments except for Chicago's 0.69 when the epochs were reduced from 200 to 5. If we averaged the R – Squared in the three sets of experiments, we have 0.9669, 0.8814, and 0.9653, which is impressive. Second, the Mean Squared Error (MSE) is close to zero for the average squared differences between the estimated actual values. Thirdly, the Mean Absolute Percentage Error (MAPE) for all experiments that measures accuracy is okay and in an acceptable range of 19 and 33. Fourth, all the experiments took about 2 minutes or less to process given the Google Colab environment and high processing power settings.

The concern of the results is overfitting as the Root Mean Square Deviation (RMSE) has small training scores and higher testing scores. Boston, in particular, keeps repeating this behavior in all of the three experiments, but Chicago does a better job with much more minor RMSE differences. Boston RMSE Training scores are 0.95, 1.70, and 1.14 compared to the more significant RMSE Testing scores of 9.56, 5.55, and 9.25, respectively. Chicago's differences in RMSE Training and Test scores are just 0.06, 1.28, and 0.09 for all the experiments. Los Angeles was in between both cities but closer to Boston.

Furthermore, the reduction in the training passes in the second experiment in files that begin with 02A04 shows a decrease of the goodness of fit measure in R – squared with a significant drop in Chicago to 0.69, but not much change in the other accuracy metrics. The third experiment of reducing LSTM cells shows the slight importance of the number of parameters. When you reduce the parameters for LSTM and a single Dense unit from 96 to 32 and from 5 to 3, respectively, there is a slight reduction in the accuracy. The metrics for the third experiment are not much different from the first experiment.

Conclusion

The Long Short – Term Memory model is a strong candidate to price and forecast single family residential homes in the U.S. There were no issues with recency in the time series because the pandemic did influence housing prices in 2020 and 2021, plus the evolution of the U.S. real estate market from 1987 to 2021 can cause difficulty, especially for linear models. It is very difficult to price homes in each city in the U.S., but LSTM provides the flexibility of hyperparameter tuning and its cell state. The three experiments in three different cities exposed the variation in accuracy with different configurations, and of course more can be done to keep the high level of accuracy and reduce overfitting. The processing object of the LSTM cell is like a conveyor belt and the LSTM gates control what information is added or removed, providing better process flow than past models relying on the assumption of linearity, normality, and stationarity.

More research and more experiments are required, of course, because there are many other cities in the U.S., but this is a promising start. An open question is the overfitting as the data series can be autoregressive, so that cross-validation may be a problem. In other words, the monthly inputs of data depend on the order of events as the sequence of months matters in signing a property deal and closing a property deal months later. Therefore, you cannot remove months like January and February and pretend nothing has happened because that is not how the residential real estate market works. Realistically, an accepted bid on a property sale can be agreed today, but money does not change hands until a few days later or even six months later when the sale price is reported in a database. As long as the agreed date of a sale and the date the sale is recorded can vary in months, cross-validation is an open debate. Nonetheless, more data on cities across the U.S. with more resources for processing power and more configurations of the LSTM model may lead to better pricing and forecasting as the real estate industry changes even more in the future.

Sources

(1) Horchreiter, Sepp and Schmidhuber, Jurgen. *"Long Short–Term Memory."* November 15, 1997. Neural Computation. Volume 9, Issue 8. https://doi.org/10.1162/neco.1997.9.8.1735

(2) Hansson, Fredrik and Rostami, Jako. "*Time Series Forecasting of Housing Prices: An evaluation of a Support Vector Machine and a Recurrent Neural Network with LSTM cells.*" May 24, 2019. Uppsala University. https://www.diva-portal.org/smash/get/diva2:1325965/FULLTEXT01.pdf

(3) Temur, Ayse Soy, and Temur, Gunay, and Akgun, Melek. "Predicting Housing Sales in Turkey using ARIMA, LSTM and Hybrid Models." July 2019. Volume 20 and Issue 5. Journal of Business Economics and Management.

(4) Ba, Jimmy and Kingma, Diederik P. "Adam: A Method for Stochastic Optimization." Cornel University. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Appendix

Performance Summary In Google Colab with TPU Hardware Accelerator and High - RAM Runtime Shape

		Mean Squared Error	Mean Absolute Percentage Error	Root Mean Square Deviation			
	R squared	MSE	MAPE	RMSE Train	RMSE Test	Total Time	File Name 01A04
Boston	0.924451	0.000021763	27.08	0.95	9.56	1 min 25s	
Chicago	0.992938	0.000010274	21.20	1.17	1.23	56.5 s	
Los Angeles	0.983186	0.000053078	32.73	2.63	6.73	1 min 39s	
Averages	0.9669	0.000028372	27.00	1.58333333	5.84	1 min 27s	

Number of parameters	Number of LSTM cells	Input Shape Features	Input Shape Time Steps	Dense Units	Activation	Epochs	Optimizer	Batch Size
96, 5	4	1	look back =1	1	relu	100	adam	1
96, 5	4	1	look back =1	1	relu	100	adam	1
96, 5	4	1	look back =1	1	relu	100	adam	1

	R squared	MSE	MAPE	RMSE Train	RMSE Test	Total Time	File Name 02A04
Boston	0.974508	0.003200000	28.71	1.70	5.55	5.27 s	
Chicago	0.695659	0.006200000	19.38	9.37	8.09	4.71 s	
Los Angeles	0.974086	0.000322730	32.36	3.26	8.35	6.31 s	
Averages	0.8814	0.003240910	26.82	4.78	7.33	5.43 s	

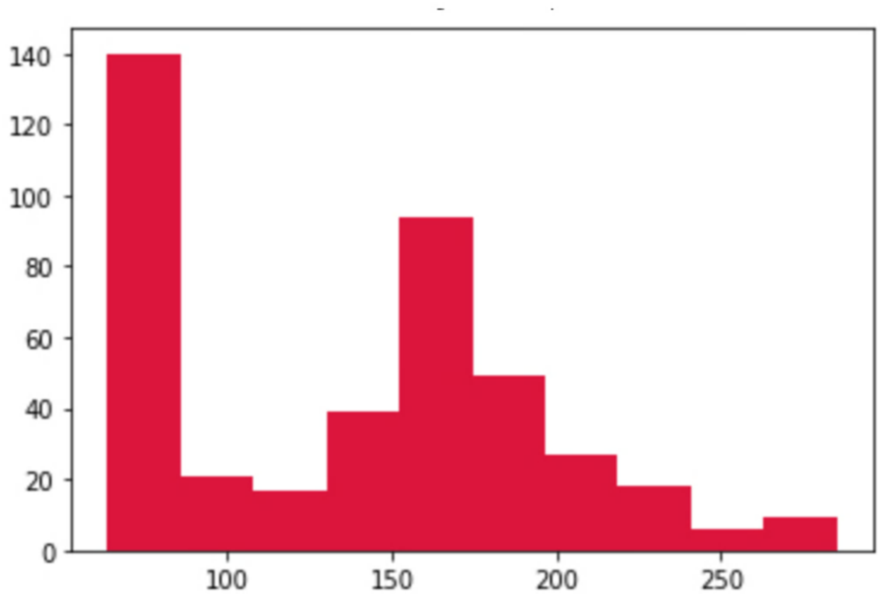
Number of parameters	Number of LSTM cells	Input Shape Features	Input Shape Time Steps	Dense Units	Activation	Epochs	Optimizer	Batch Size
96, 5	4	1	look back = 1	1	relu	5	adam	1
96, 5	4	1	look back = 1	1	relu	5	adam	1
96, 5	4	1	look back = 1	1	relu	5	adam	1

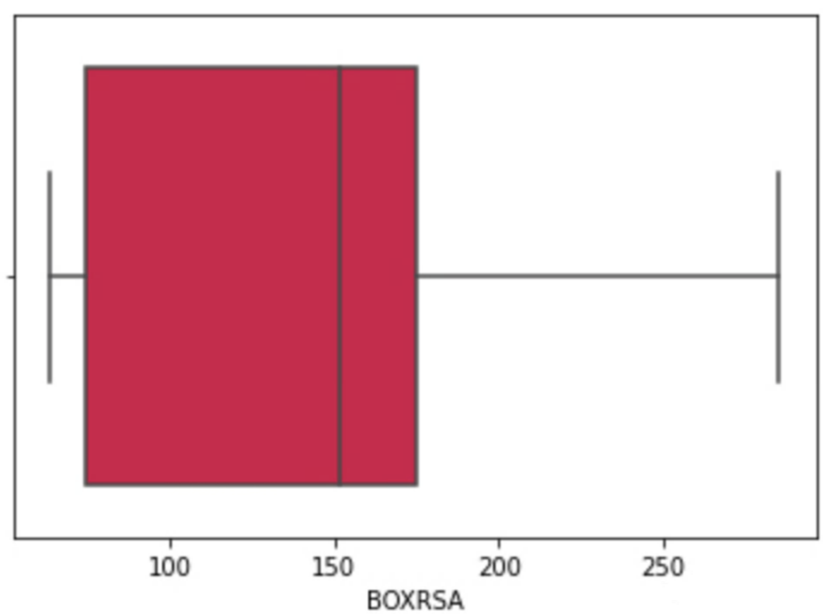
	R squared	MSE	MAPE	RMSE Train	RMSE Test	Total Time	File Name 03A04
Boston	0.929144	0.000020500	27.13	1.14	9.25	1 min 24s	
Chicago	0.994778	0.000102940	21.05	1.15	1.06	55.9 s	
Los Angeles	0.97201	0.000062624	32.12	2.39	8.68	1 min 7s	
Averages	0.9653	0.000062021	26.77	1.56	6.33	1 min 9s	

Number of parameters	Number of LSTM cells	Input Shape Features	Input Shape Time Steps	Dense Units	Activation	Epochs	Optimizer	Batch Size
32, 3	2	1	look back = 1	1	relu	100	adam	1
32, 3	2	1	look back = 1	1	relu	100	adam	1
32, 3	2	1	look back = 1	1	relu	100	adam	1

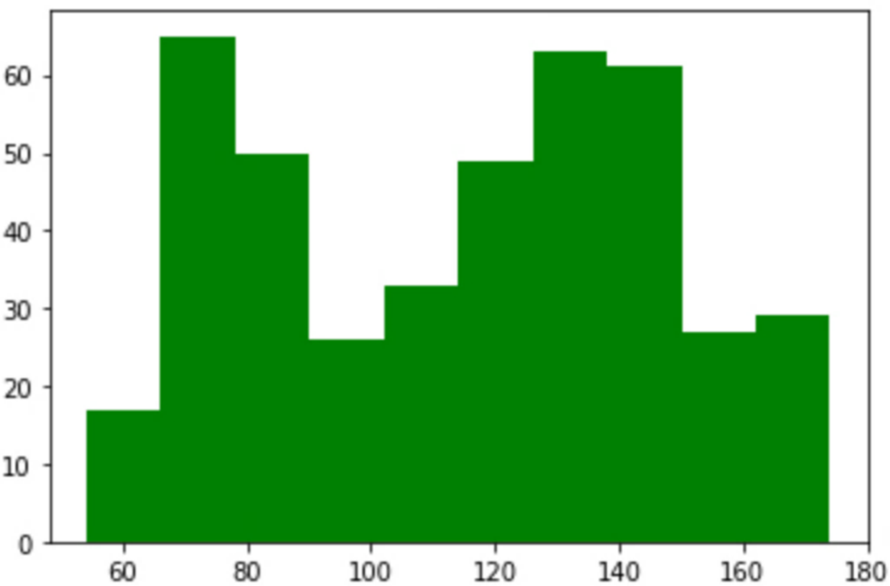
Please note that the number of parameters is the LSTM parameters followed by the Dense parameters

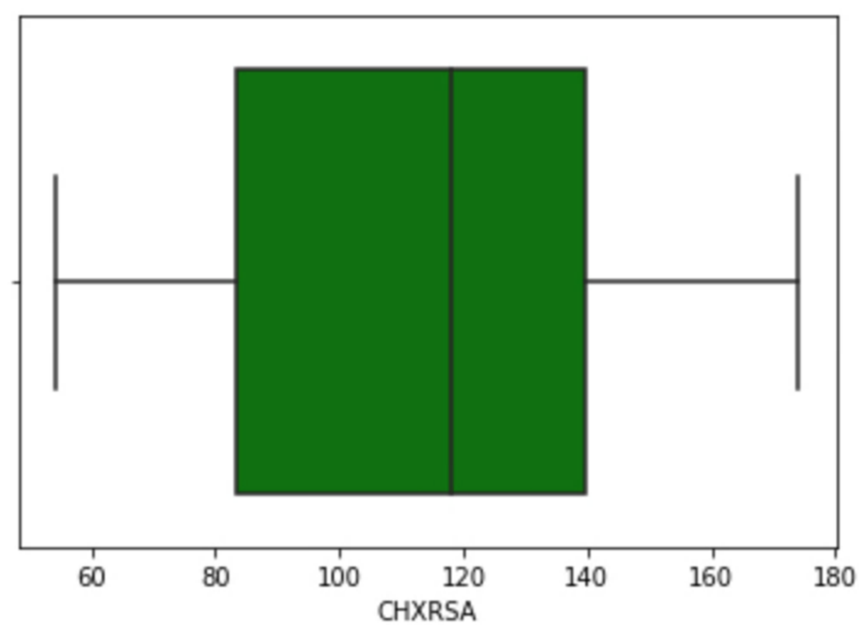
Boston from January 1987 to December 2021





Chicago from January 1987 to December 2021





Los Angeles from January 1987 to December 2021