

Modeling US Housing Prices



Analysing and forecasting housing time series data



Why predicting and forecasting is important in the real estate market?

- ❖ For investors to determine price trends in different locations to know where to invest.
- ❖ For home buyers to know where and when they can potentially buy a home and plan their finances around forecasted price ranges.
- ❖ For home sellers to determine the value of their home and determine when to sell.

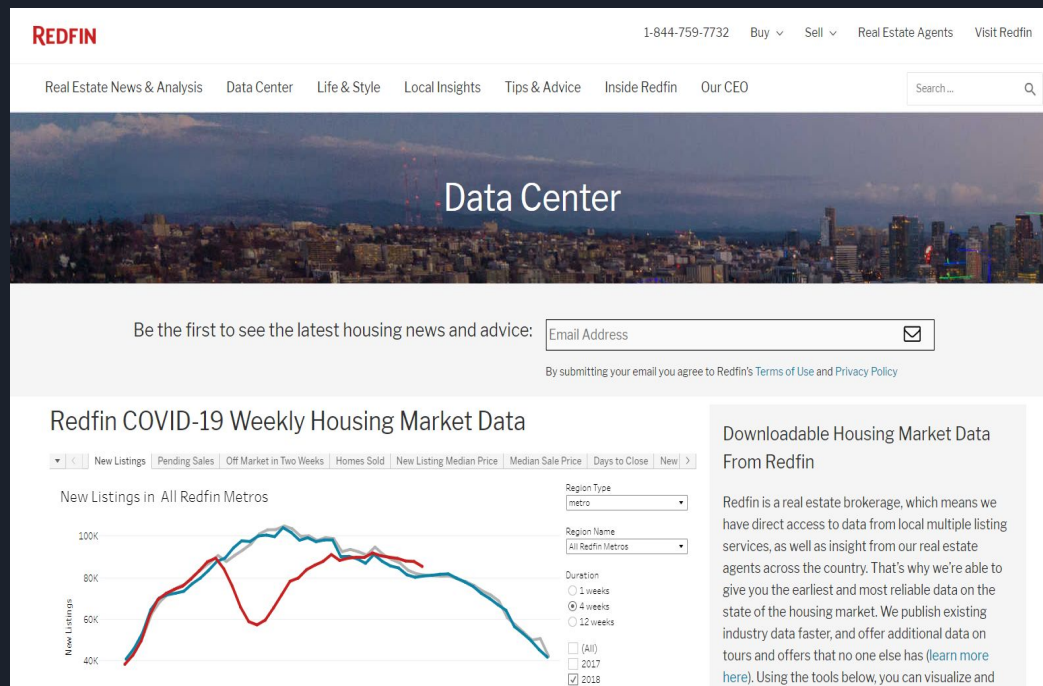


Aim of the project

- ❖ To provide some insight and analysis into some past and current trends in the housing market.
- ❖ To develop and deploy machine learning models that will be used to forecast housing time series data.
- ❖ Determine future housing prices based on the models.

A look at the data source

- ❖ Redfin brokerage database site.
- ❖ They provide real estate news, analysis and historical data as well as property listings across different states in the United States.





Important features from the data

- ❖ Median sales price (*50th percentile price at the end of the month*) ** Target variable
- ❖ Homes sales (*total number of home sales at the end of the month*)
- ❖ New listings (*total number of homes listings with an added date during the month*)
- ❖ Inventory (*total number of listings as of the end of the month*)
- ❖ Median days on the market (*50th percentile of days properties were on the market*)
- ❖ Months of supply (*how many months to have new listings if none is available*)
- ❖ Pending sales (*number of homes that went under contract during the month*)
- ❖ Active listings (*total number of home listings that were active during the month*)
- ❖ The data was collected from 2012 to Aug 2020

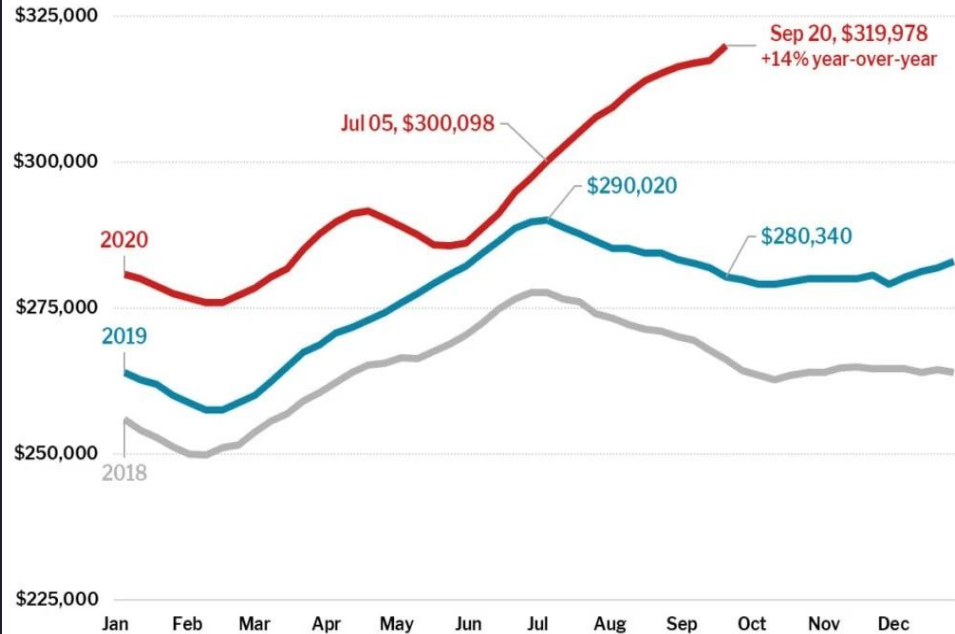
Link to the metadata

https://docs.google.com/spreadsheets/d/1YNT5VfZTwSnUK7nqGAPGZUqOVBbsIC_M1vTgYCCLtVg/edit#gid=635767466

Current market trends and analysis

Home Sale Prices Up 14% to Another New High

4-week rolling average of the median sale price of homes sold

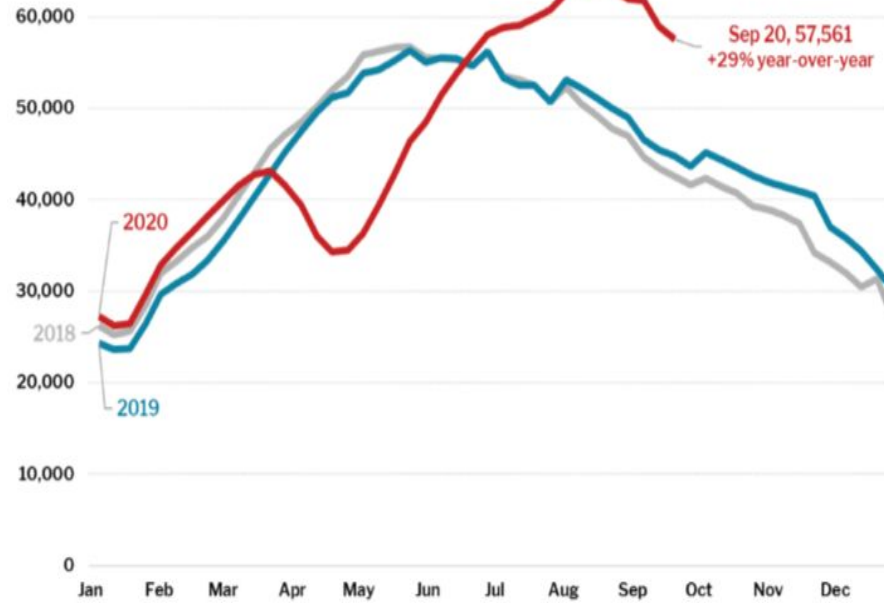


REDFIN

Source: Redfin analysis of MLS data

Pending Sales Up 29% From a Year Earlier

4-week rolling average of weekly pending sales

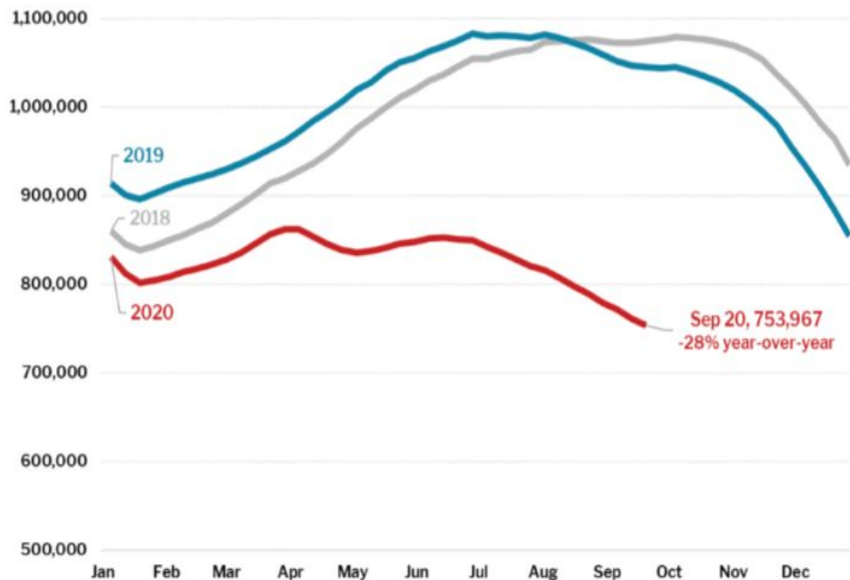


REDFIN

Source: Redfin analysis of MLS data

Active Listings of Homes For Sale Down 28% From 2019

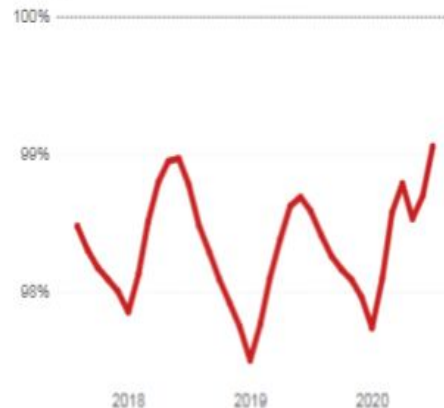
4-week rolling average of weekly active listings of homes for sale



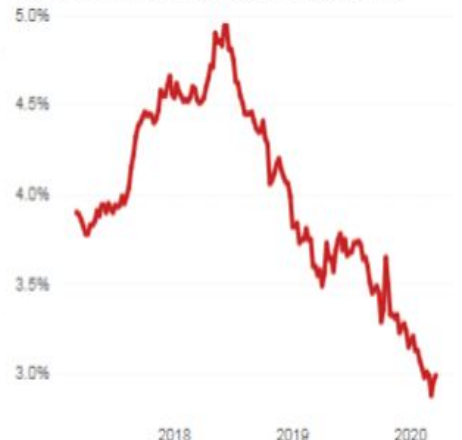
REDFIN



Average Sale-to-List Percentage
Shows how closely the final sale price matches the original asking price.



Mortgage Rates
Shows how mortgage rates have changed over time.



*Source: MLS
Information provided may be incomplete and should be independently verified.
Details are subject to change; amounts provided are estimates. Visit redfin.com/disclosure for full terms and conditions.

REDFIN

Data analysis from [Tableau](https://www.tableau.com/)

Period End	Geo Region	Property Type	State Code	parent_metro_region	active_listings	Median Dom	Avg Sale To List	Homes Sold	New Listings	Inventory	pending_sales	months_of_supply	Median Sale Price
2012-02-01	South	All Residential	FL	Orlando, FL	21554.0	120.0	0.963734	2478.0	3744.0	18765.0	1151.0	7.6	\$115K
2012-02-01	South	All Residential	TX	San Antonio, TX	10658.0	90.0	0.958310	1301.0	2232.0	8625.0	1005.0	6.6	\$150K
2012-02-01	Midwest	All Residential	MI	Sturgis, MI	355.0	102.0	0.897743	26.0	50.0	333.0	13.0	12.8	\$29K
2012-02-01	West	All Residential	CA	Vallejo, CA	1745.0	44.0	0.987707	485.0	566.0	1092.0	454.0	2.3	\$170K
2012-02-01	West	All Residential	CA	San Diego, CA	13044.0	62.0	0.970295	2548.0	3699.0	9099.0	2561.0	3.6	\$302K
...
2020-07-01	West	All Residential	OR	Portland, OR	9487.0	14.0	1.006187	4315.0	4694.0	4778.0	4082.0	1.1	\$440K
2020-07-01	South	All Residential	VA	Richmond, VA	4165.0	13.0	1.001111	1952.0	1958.0	1877.0	1944.0	1.0	\$289K
2020-07-01	Midwest	All Residential	IA	Pella, IA	111.0	16.0	0.973929	49.0	57.0	64.0	43.0	1.3	\$170K
2020-07-01	Midwest	All Residential	NE	Fremont, NE	101.0	17.0	0.976522	54.0	51.0	43.0	50.0	0.8	\$177K
2020-07-01	Northeast	All Residential	ME	Lewiston, ME	145.0	11.0	1.007384	154.0	65.0	117.0	22.0	0.8	\$210K

40679 rows × 13 columns

A snapshot of the full datatable used extracted from the Redfin database

Data table for modeling

Period End	Geo Region	Property Type	State Code	parent_metro_region	active_listings	Median Dom	Avg Sale To List	Homes Sold	New Listings	Inventory	pending_sales	months_of_supply	Median Sale Price
2012-02-01	West	All Residential	CA	Vallejo, CA	1745.0	44.0	0.987707	485.0	566.0	1092.0	454.0	2.3	170000
2012-02-01	West	All Residential	CA	San Diego, CA	13044.0	62.0	0.970295	2548.0	3699.0	9099.0	2561.0	3.6	302000
2012-02-01	Northeast	All Residential	DC	Washington, DC	21201.0	63.0	0.976828	3663.0	5998.0	15629.0	3660.0	4.3	278000
2012-02-01	West	All Residential	NM	Santa Fe, NM	761.0	171.0	0.985821	99.0	94.0	682.0	21.0	6.9	295000
2012-02-01	West	All Residential	CA	Clearlake, CA	446.0	100.0	0.962962	80.0	82.0	361.0	41.0	4.5	105000
...
2020-07-01	West	All Residential	CA	Santa Maria, CA	1237.0	52.0	0.985959	374.0	395.0	824.0	307.0	2.2	715000
2020-07-01	Northeast	All Residential	MA	Boston, MA	14710.0	20.0	1.005931	5340.0	6319.0	8215.0	5701.0	1.5	555000
2020-07-01	West	All Residential	CA	San Rafael, CA	906.0	28.0	1.009217	383.0	331.0	483.0	380.0	1.3	1350000
2020-07-01	West	All Residential	NM	Santa Fe, NM	932.0	50.0	1.000883	229.0	272.0	694.0	173.0	3.0	450000
2020-07-01	West	All Residential	CA	Napa, CA	597.0	50.0	0.988584	179.0	161.0	393.0	146.0	2.2	755000

3762 rows × 13 columns

A snapshot of the datatable of the top 5 ranked states based on sales price 2012 -2020

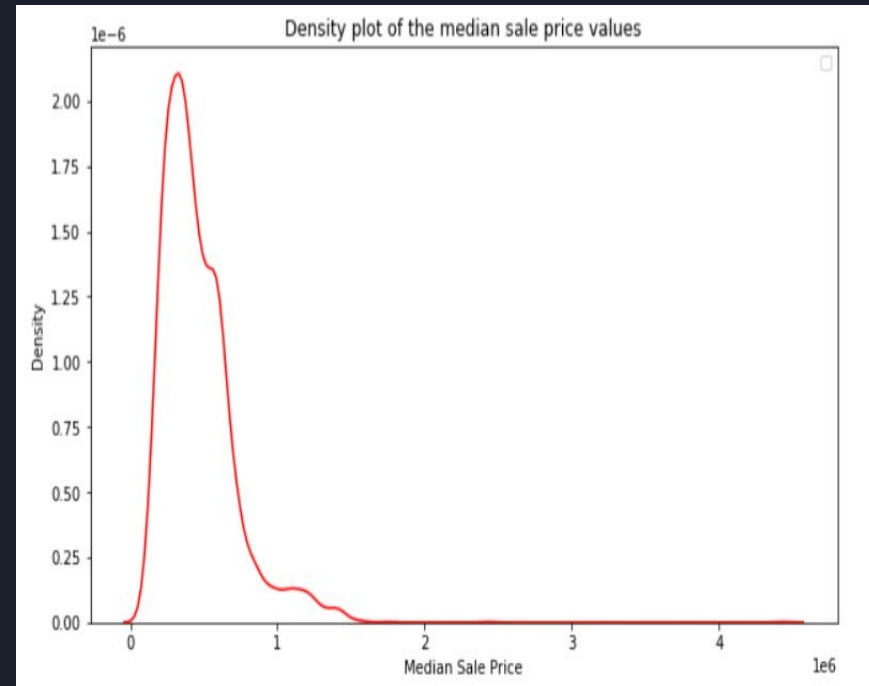
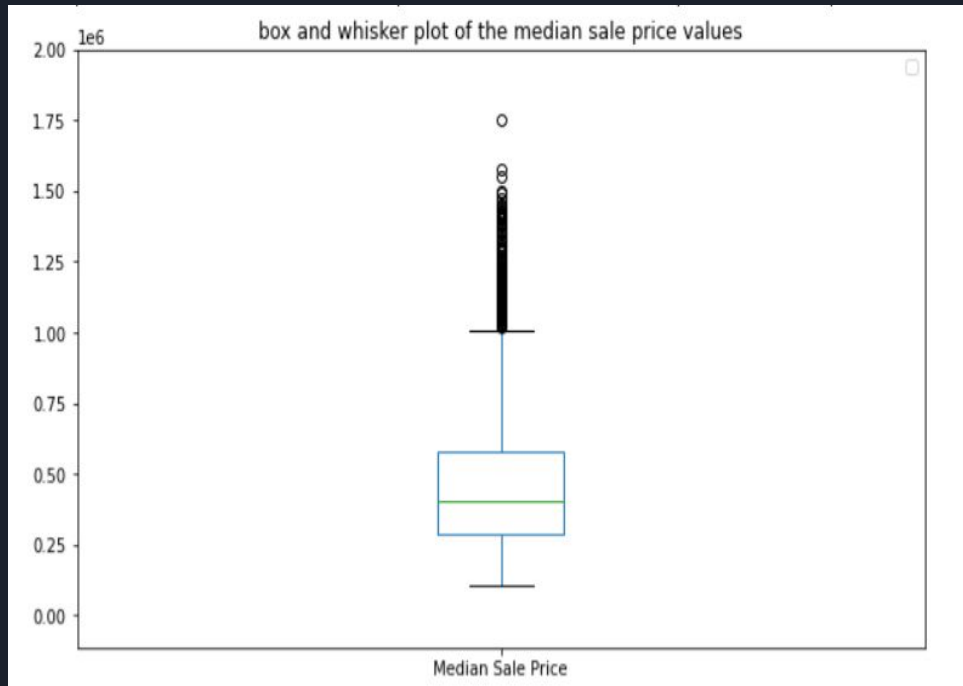
Data summary statistics

	count	mean	std	min	25%	50%	75%	max
active_listings	3762.0	4443.235779	6402.451265	23.000000	817.00000	1651.000000	4111.000000	3.290500e+04
Median Dom	3762.0	59.278841	60.017219	7.000000	26.00000	42.000000	68.000000	1.079000e+03
Avg Sale To List	3762.0	0.986702	0.025828	0.828685	0.97525	0.986109	0.996225	1.121664e+00
Homes Sold	3762.0	1152.489102	1648.661126	2.000000	190.00000	427.500000	1119.500000	8.639000e+03
New Listings	3762.0	1376.784689	2031.706513	1.000000	210.00000	476.500000	1319.500000	1.086500e+04
Inventory	3762.0	3087.669856	4529.396669	11.000000	560.25000	1229.500000	3018.500000	2.557000e+04
pending_sales	3762.0	1028.308878	1523.828145	1.000000	119.00000	337.500000	1095.000000	8.032000e+03
months_of_supply	3762.0	3.668740	3.348194	0.300000	2.10000	2.900000	4.100000	5.700000e+01
Median Sale Price	3762.0	462801.701223	252067.176018	104000.000000	287000.00000	405000.000000	580000.000000	4.426000e+06

Summary statistics for the data distribution for the data feature columns for the top 5 states

The table shows the data count, mean, standard deviation, minimum, maximum as well as the 25th percentile and 75th percentile values

Median sales price data distribution plots



Modeling the target variable will require a robust model to capture the outliers

Data resampling


To train time series models the data frequency had to be resampled by month

```
df_top5 = df_top5.resample('MS').mean()
```

```
Df_top5
```

	active_listings	Median Dom	Avg Sale To List	Homes Sold	New Listings	Inventory	pending_sales	months_of_supply	Median Sale Price
Period End									
2012-02-01	5557.805556	100.138889	0.966955	921.555556	1406.055556	4176.861111	845.722222	5.433333	286972.222222
2012-03-01	5702.027778	98.388889	0.972665	1182.583333	1629.138889	4110.000000	1064.916667	4.350000	294083.333333
2012-04-01	5541.722222	88.305556	0.973483	1190.277778	1524.888889	4063.138889	1044.138889	4.119444	308111.111111
2012-05-01	5518.000000	79.416667	0.977199	1355.277778	1545.444444	3948.916667	1125.250000	4.836111	318055.555556
2012-06-01	5304.888889	74.416667	0.980372	1388.638889	1436.916667	3808.527778	1049.027778	3.647222	319583.333333
...
2020-03-01	3709.081081	48.135135	0.992018	1061.567568	1350.621622	2457.216216	990.162162	2.821622	589000.000000
2020-04-01	3387.729730	44.945946	0.987829	866.081081	960.000000	2462.729730	754.054054	3.456757	598189.189189
2020-05-01	3754.945946	45.702703	0.981610	784.513514	1321.081081	2576.675676	967.945946	4.697297	557729.729730
2020-06-01	4021.216216	53.324324	0.986052	1148.756757	1471.486486	2463.540541	1238.972973	2.967568	598864.864865
2020-07-01	4040.675676	40.621622	0.992940	1430.567568	1603.459459	2321.945946	1399.324324	2.151351	621567.567568
102 rows × 9 columns									

The top five states data table has to be aggregated and resampled daily, weekly or monthly before model training
The data table represents the top five states as a whole from 2012 to 2020



Quick overview of VAR models, Facebook prophet and LSTM

- ❖ VAR Models: Extension of univariate time series models used to develop multivariate time series models. Belongs to the class of VARMA (Vector Autoregression moving average) models
- ❖ Facebook's prophet: Open source python library used for time series forecasting. Additive model that fits non-linear time series data to daily, weekly, monthly or yearly seasonality.
- ❖ LSTM model: It is a form of RNN a class of deep learning network that has both feedforward and feedback features. It is suitable for time series prediction because it handles the vanishing gradient problem.



Models development

underlying assumptions and preprocessing

- ❖ **VAR model:** The features used in the model building process for each time series should exhibit codependency .
- ❖ **Facebook prophet library:** Modeling is easy and it handles the preprocessing automatically as well as choice of parameters for your model.
- ❖ **LSTM model:** LSTM works best with maximum between 200 and 400 time steps. Each time series can be predicted using others or one can be predicted by the others.

VAR Model

Lagged values up to 10 were tested to choose the best max lag value with the least AIC value

```
model = VAR(diff train2[diff train2.columns])  
for i in [1,2,3,4,5,6,7,8,9,10]:  
    result = model.fit(i)
```

A VAR model with a maximum of 8 lags was fit to the data

```
fitted_model = model.fit(maxlags=8)  
fitted_model.summary()
```

```
Summary of Regression Results  
=====
```

Model:	VAR		
Method:	OLS		
Date:	Mon, 28, Sep, 2020		
Time:	11:13:54		

```
-----
```

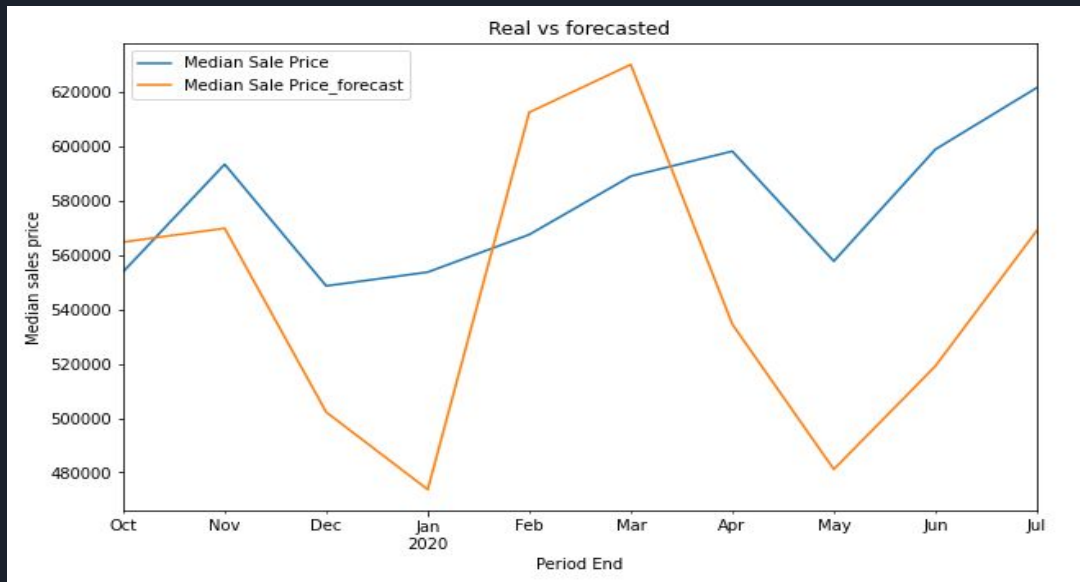
No. of Equations:	9.00000	BIC:	31.5289
Nobs:	82.0000	HQIC:	19.9877
Log likelihood:	-892.256	FPE:	3.04645e+09
AIC:	12.2458	Det(Omega_mle):	9.88856e+06

```
-----
```

Results for equation active_listings

Model performance

The model isn't the best forecasting and the RMSE for the model was with consideration given to the variance of the sales price values



RMSE for the overall forecast: 193388.94187380048

Facebook prophet library

The date column (period end) and target column (median sales price) were fitted into a facebook prophet object with the other time series as added regressors. They are represented by ds and y in the table below

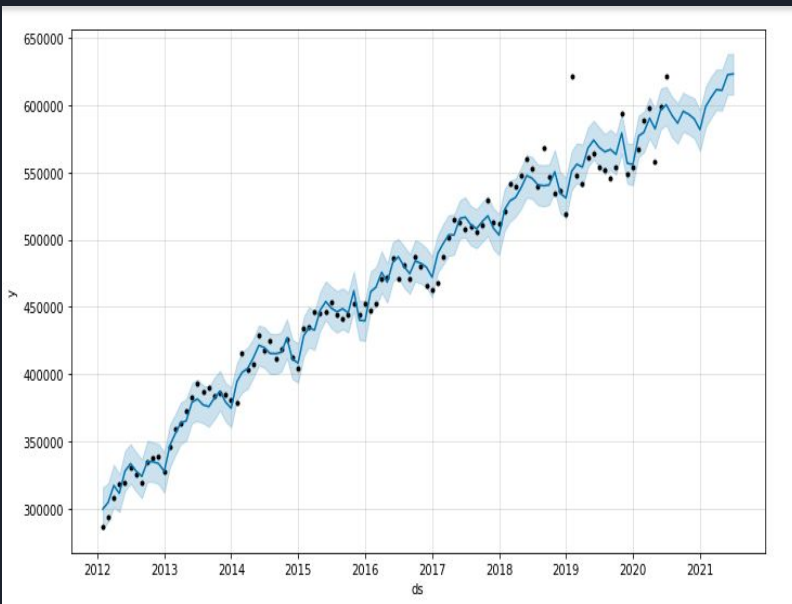
```
m = Prophet(seasonality_mode='additive')
for i in df3.columns[2:]:
    m.add_regressor(i)
m.fit(train)
```

ds	y	active_listings	Median Dom	Avg Sale To List	Homes Sold	New Listings	Inventory	pending_sales	months_of_supply
2012-02-01	286972.222222	5557.805556	100.138889	0.966955	921.555556	1406.055556	4176.861111	845.722222	5.433333
2012-03-01	294083.333333	5702.027778	98.388889	0.972665	1182.583333	1629.138889	4110.000000	1064.916667	4.350000
2012-04-01	308111.111111	5541.722222	88.305556	0.973483	1190.277778	1524.888889	4063.138889	1044.138889	4.119444
2012-05-01	318055.555556	5518.000000	79.416667	0.977199	1355.277778	1545.444444	3948.916667	1125.250000	4.836111
2012-06-01	319583.333333	5304.888889	74.416667	0.980372	1388.638889	1436.916667	3808.527778	1049.027778	3.647222
...
2020-03-01	589000.000000	3709.081081	48.135135	0.992018	1061.567568	1350.621622	2457.216216	990.162162	2.821622
2020-04-01	598189.189189	3387.729730	44.945946	0.987829	866.081081	960.000000	2462.729730	754.054054	3.456757
2020-05-01	557729.729730	3754.945946	45.702703	0.981610	784.513514	1321.081081	2576.675676	967.945946	4.697297
2020-06-01	598864.864865	4021.216216	53.324324	0.986052	1148.756757	1471.486486	2463.540541	1238.972973	2.967568
2020-07-01	621567.567568	4040.675676	40.621622	0.992940	1430.567568	1603.459459	2321.945946	1399.324324	2.151351

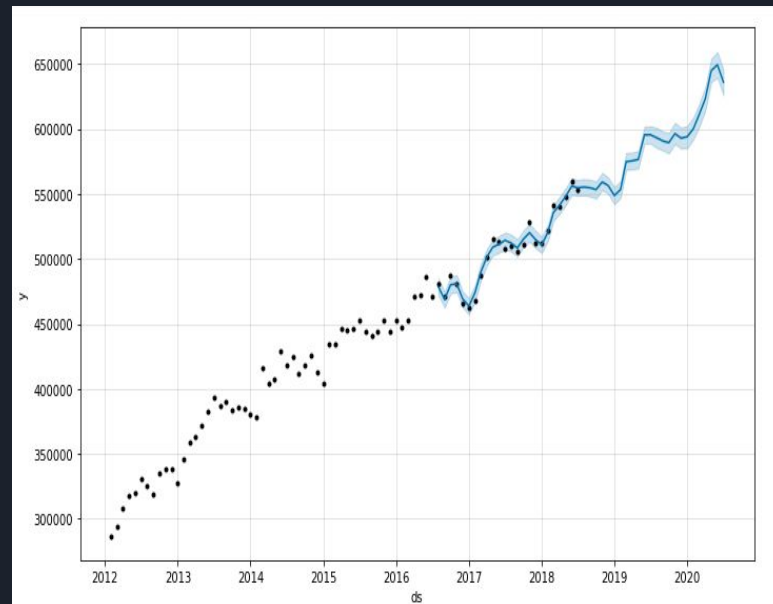
Model performance

The prophet model performed better than the VAR model with capturing the variations with a lower RMSE and better predictions

RMSE for the forecast: 26432.238681781066



Median sales forecast 12 months from July 2020



Model predictions ending July 2020



LSTM model

The LSTM model had 1 LSTM layer with 60 nodes and a dense layer. The model was fitted for 30 epochs and a batch size of 7.

```
Model: "sequential"
```

Layer (type)	Output Shape	Param #
=====		
lstm (LSTM)	(None, 60)	14880
=====		
dense (Dense)	(None, 1)	61
=====		

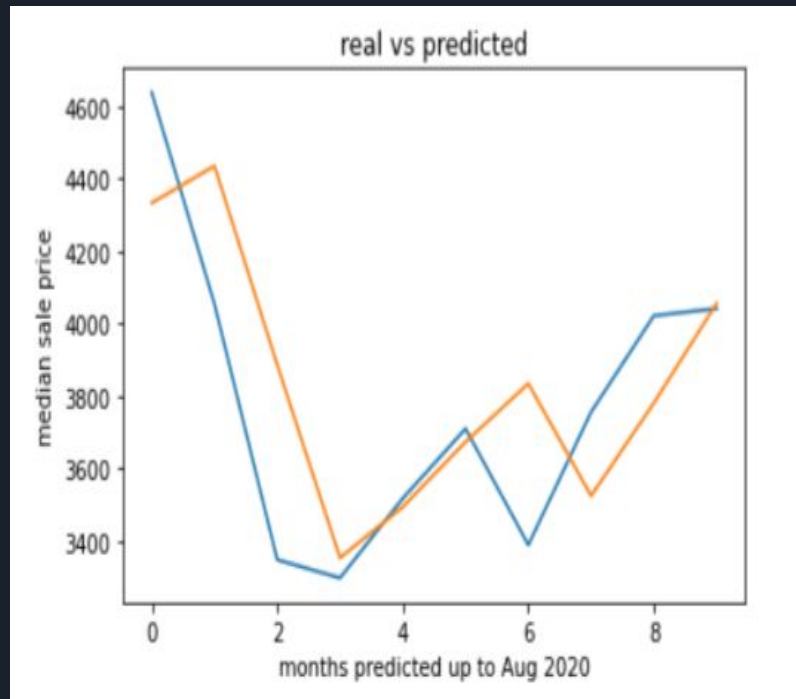
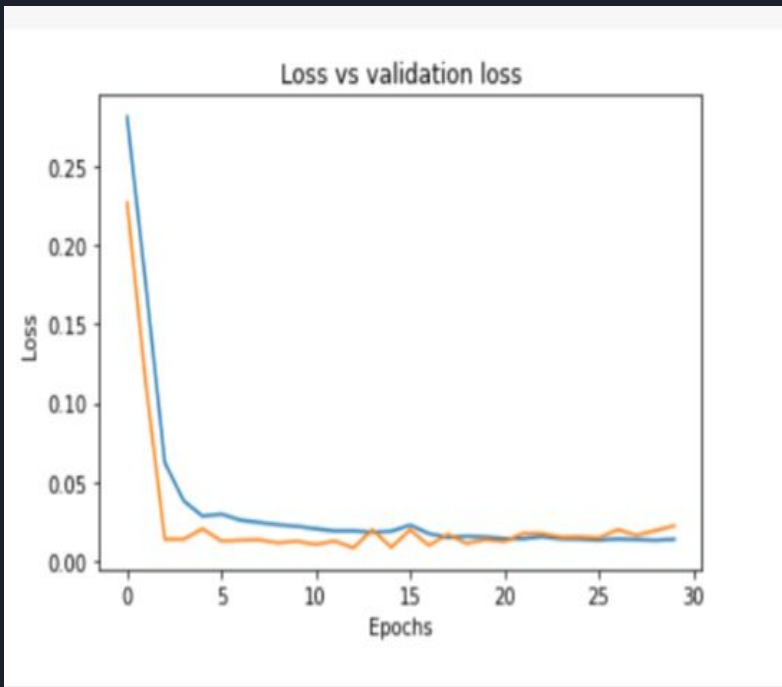
```
Total params: 14,941
```

```
Trainable params: 14,941
```

```
Non-trainable params: 0
```

Model performance

The loss plots show no overfitting and the RMSE is lower than the VAR model but not better than facebook prophet.





Conclusions

- ❖ Facebook prophet library is more robust in making predictions.
- ❖ The current housing market appears to be a seller's market especially in the top 5 states where prices are rising.
- ❖ In the next 12 months in the top 5 ranked states for prices, they are expected to be about \$462,000 with its maximum at \$492,000 and minimum of \$477,000.



Challenges and future work

- ❖ Find and pay for data with more information
- ❖ Use different combinations of deep learning methods like a combination of LSTMs and CNNs
- ❖ Explore using exogenous data for modeling using VAR models



THANKS FOR YOUR ATTENTION

