Lab 6

Big Data Analytics

Waylon Abernathy

11/9/19

Data

In [7]:
```
1 zillow = pd.read_csv('Zip_Zhvi_SingleFamilyResidence.csv')
2 zillow.shape
```

Out[7]: (15701, 289)

Part 1: Home Sweet Home

In order to get all of the depressed Razorback fans into a single metro dataset, I first needed to see what the listings were for the metro areas. After locating the names, I then took every city for each metro area and ran the means for each month/year in the new dataset. I did this twice - once for all cities in the metro areas and once for only the 4 cities with their corresponding zipcodes.

4 towns with zipcodes

Here is a bottom view of what we were left with when looking at only the 4 towns with their zipcodes. Each zipcode's value was averaged per year.

Out[8]:

	Fayetteville	Little Rock	Hot Springs	Searcy
1996-04	93600.000000	115037.5	73950.0	73800.0
1996-05	93800.000000	115050.0	73950.0	74100.0
1996-06	93766.666667	115025.0	73950.0	74300.0
1996-07	93600.000000	114975.0	73900.0	74600.0
1996-08	93366.666667	114912.5	73850.0	74800.0

All Towns/Zip Codes within Metro Area

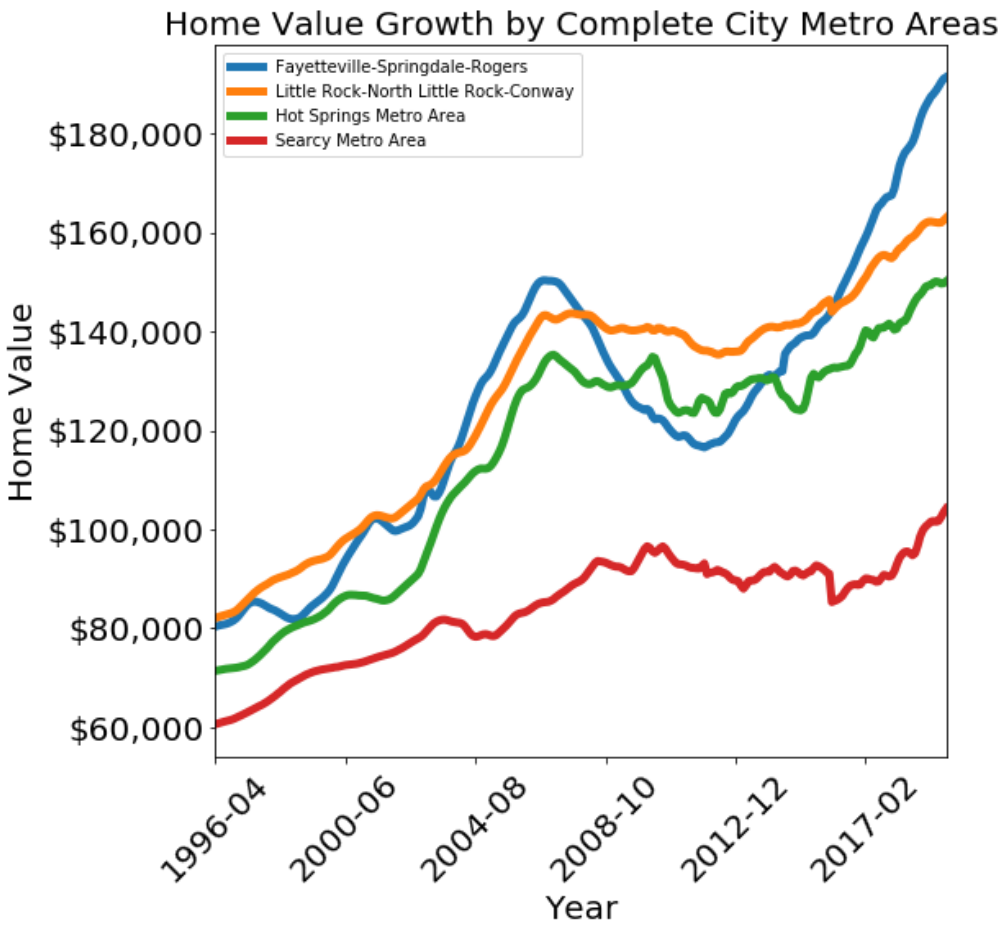
This is bottom view of the four towns including all other towns within the respective metro areas. Searcy (pronounced 'Sir-sea') doesn't receive any favors in home value from the surrounding towns in the Searcy "Metro" Area. Let's see how it all plots together.

Out[10]:

	Fayetteville-Springdale-Rogers	Little Rock-North Little Rock-Conway	Hot Springs Metro Area	Searcy Metro Area
1996-04	80373.684211	82064.516129	71425.0	60640.0
1996-05	80547.368421	82277.419355	71525.0	60800.0
1996-06	80663.157895	82454.838710	71650.0	60940.0
1996-07	80757.894737	82622.580645	71725.0	61100.0
1996-08	80878.947368	82758.064516	71825.0	61240.0

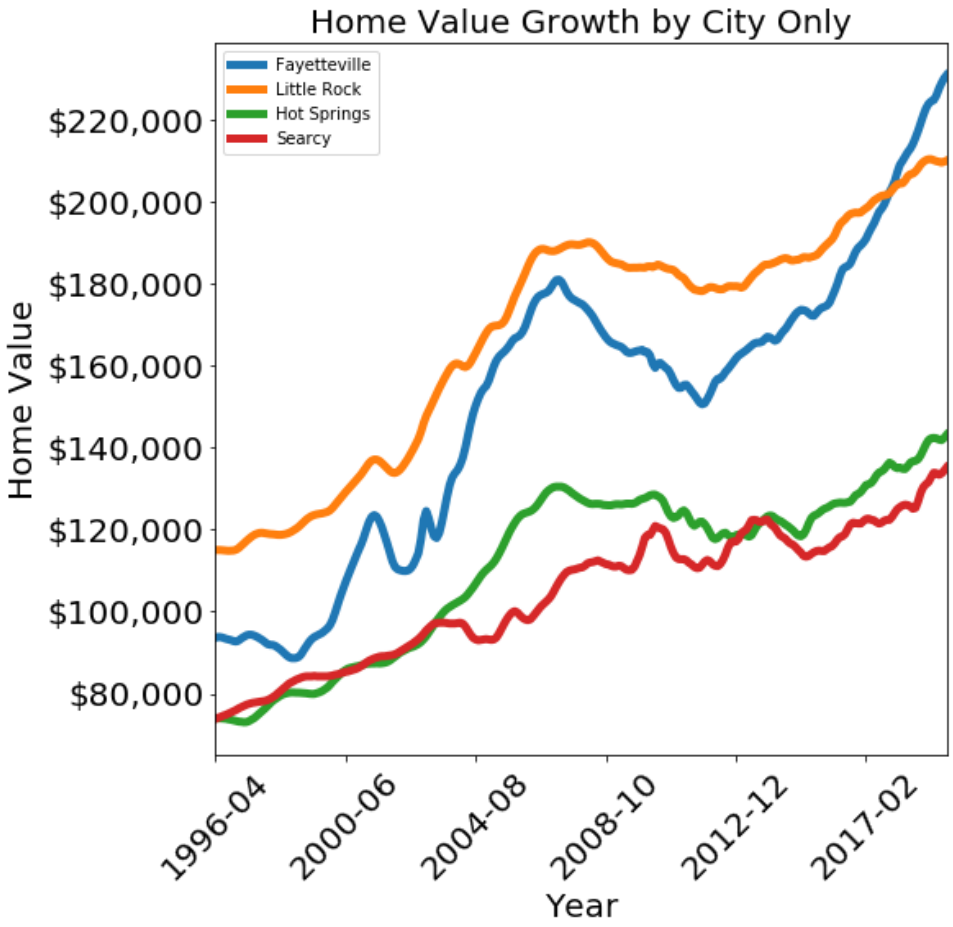
Metro Area Plot

Sidenote: Finally learned how to get '$' and ',' on the y-axis ticks... worth it.



City Plot:

This is why brands with names like 'Fayettechill' can survive.



Part 2

Finding Zipcodes for the Prophet

With the help of some code online (see references below) I was able to wrangle the percent change over time for the zipcodes as well as the coefficient of variance, which will make it helpful in narrowing my scope. I needed to decrease the data size due to issues with prophet + colab (also tried spark) running the full data. I believe I can make a decent choice of zip codes to use with this information and the extra data used below. Using subsetting to help choose zipcodes isn't as impressive as running all the data through a predictor model, but it will have to do for now.

The timeframe for the calculations was set for December 2017 as requested by the assignment parameters.

Fun Fact: The water crisis in Flint, MI made a significant impact on home values

Out[20]:

	City	State	std	mean	pchange	CV
4537	Flint	MI	19843.125877	41073.563218	-124.519231	0.483112
4339	Flint	MI	19520.459813	40047.892720	-105.429864	0.487428
1069	Indianapolis	IN	16182.102223	72282.758621	-56.626506	0.223872
14597	Capron	IL	21904.616560	131182.375479	-43.393276	0.166978
7034	Rock Falls	IL	10659.313198	79201.915709	-40.472879	0.134584

Dont Worry, Be Happy...dataset.

Here, I pulled data from wallethub's listing of best places to live in the US based on data already analyzed by the websites data science team. This dataset is based on multiple metrics such as job growth, crime rate, income-growth rate, poverty, and wellness/life expectancy. Merging this data with the Zillow data helped narrow the scope further.

In [167]:
```
1 happy = pd.read_csv('Happy.csv')
2 happy.shape
```

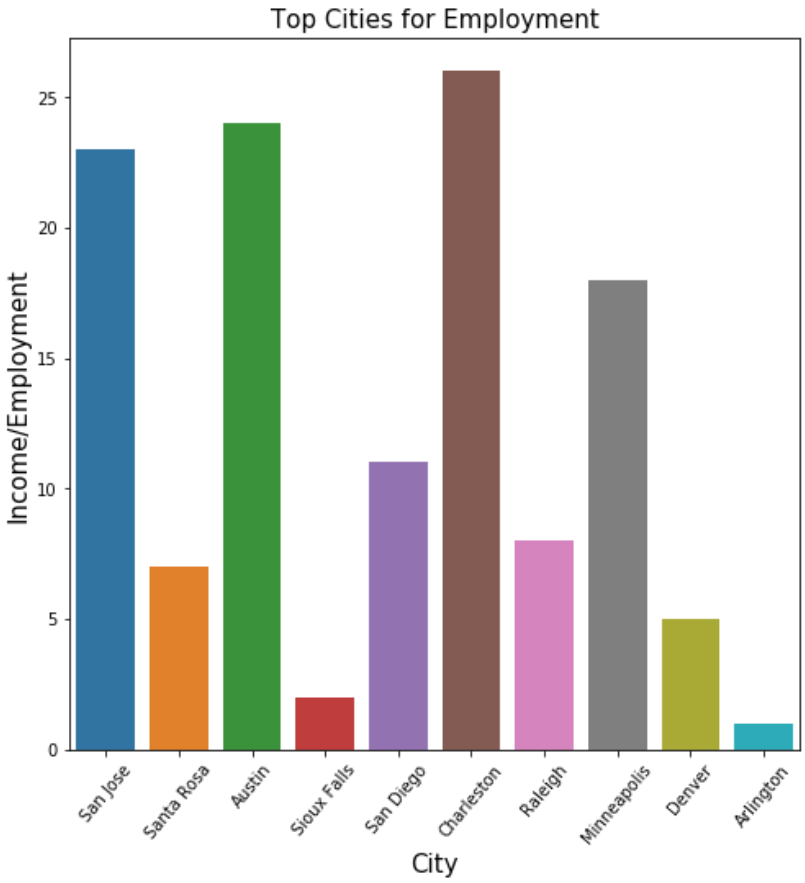
Out[167]: (182, 7)

Zips with highest Income/Employment opportunity

This plot represents the zipcodes based on the wallethub data variable: Income/Employment. Overall ranking, happiness, and physical/emotional wellness are all important parts of the happiness dataset, but we are looking for investments to make money. Using the assumption that an increase in job availability creates population/housing growth and demand is the reason I focused on this particular variable. San Jose, Austin, and Charleston all scored in the top 20 out of the 181 cities that made the list in this area, so they could possibly be options for investment

After localizing the 10 cities seen below based on Income/Employment, I used parameters of Percent Change and Coefficient of Variance (CV) to subset the data further. I used only cities with a percentage of positive change greater than 70% with a 0.3 or less CV. This helped narrow down to 5 possible zipcodes.

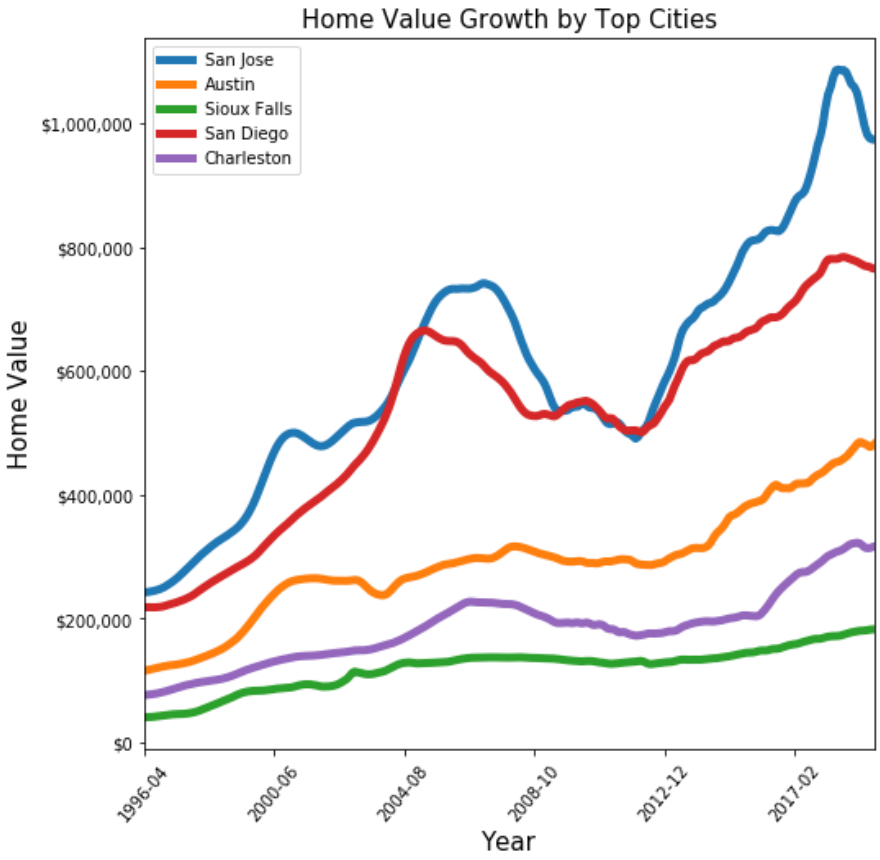
I worry that such a low variance may lead to overfitting later in the process with Prophet.



Final Five

Out[95]:

	San Jose	Austin	Sioux Falls	San Diego	Charleston
1996-04	242500.0	115900.0	41000.0	218700.0	76800.0
1996-05	242900.0	116900.0	41100.0	218600.0	77200.0
1996-06	243500.0	118000.0	41300.0	218500.0	77600.0
1996-07	244300.0	119100.0	41600.0	218400.0	78200.0
1996-08	245200.0	120200.0	42000.0	218500.0	78800.0



Take aways from this plot:

Recessions hurt. San Jose and San Diego both took large dips in home value from the 2008 recession, which could be an advantage for us depending on if you're a "glass-half-full" type. San Jose took a massive drop; however, it recovered very nicely until what looks like another dip approaching in 2018-2019.

San Jose and San Diego might not be great options if we are investing tomorrow, but they may present opportunities in the future if/when the housing market becomes cheaper to buy into.

Our safest options may be in Austin, Sioux Falls, and Charleston as the home value hit was not as significant in 2008.

Furthermore, I am interested in seeing how Prophet handles San Jose since we are training on dates up to 2017. We might be able to see how the low variance in the zipcodes should show us how Prophet handles overfitting, which would help us finalize on 3 zipcodes to invest.

Prophet

San Diego 92128 Predictions

```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.


WARNING:fbprophet:Seasonality has period of 365.25 days which is larger than initi
al window. Consider increasing initial.
INFO:fbprophet:Making 42 forecasts with cutoffs between 1997-06-01 06:52:12 and 20
17-11-30 18:10:48
INFO:fbprophet:n_changepoints greater than number of observations.Using 11.
INFO:fbprophet:n_changepoints greater than number of observations.Using 15.
INFO:fbprophet:n_changepoints greater than number of observations.Using 20.
```

San Diego 2018 Predictions vs 2018 Actual

San Diego's predictions appear right on the money when based off of the training set of 2017 and prior. We can see that the predictions of y (actual) vs yhat (predicted) are within $10-20K.

Out[235]:

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
500	2018-08-01	769071.405088	756065.619106	783144.271088	783300.0	2017-11-30 18:10:48
501	2018-09-01	772002.310700	757452.384646	785881.871826	783900.0	2017-11-30 18:10:48
502	2018-10-01	774789.845144	759807.344205	791571.268343	783000.0	2017-11-30 18:10:48
503	2018-11-01	777566.638926	760884.354570	794661.996351	781400.0	2017-11-30 18:10:48
504	2018-12-01	779802.157637	761333.768473	800047.819522	779900.0	2017-11-30 18:10:48

San Diego Model Performance

Out[236]:

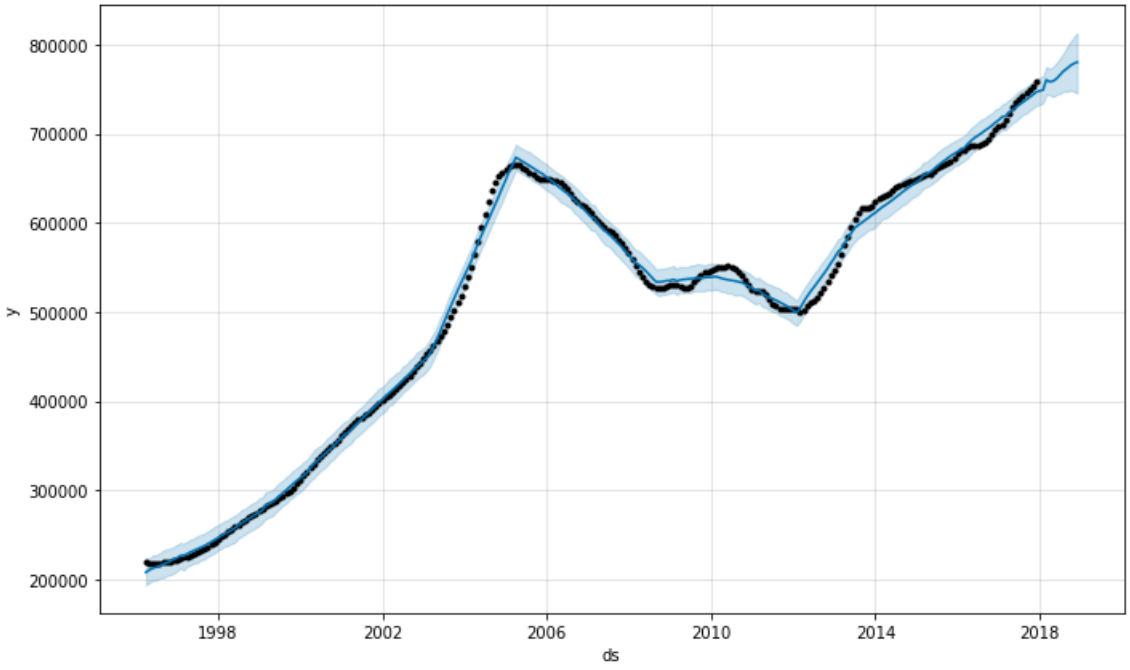
	horizon	mse	rmse	mae	mape	coverage
451	365 days 03:39:36	4.109021e+09	64101.646769	43535.127156	0.078257	0.38
452	365 days 04:22:48	4.189979e+09	64730.048478	44528.712604	0.079762	0.38
453	365 days 05:06:00	4.236854e+09	65091.119226	44849.581448	0.080022	0.38
454	365 days 05:29:24	4.215933e+09	64930.214373	44468.864128	0.079725	0.38
455	365 days 05:49:12	4.205018e+09	64846.112585	44003.603413	0.078805	0.40

R squared

Out[237]: `0.8828589661172318`

San Diego Prophet Prediction Plot

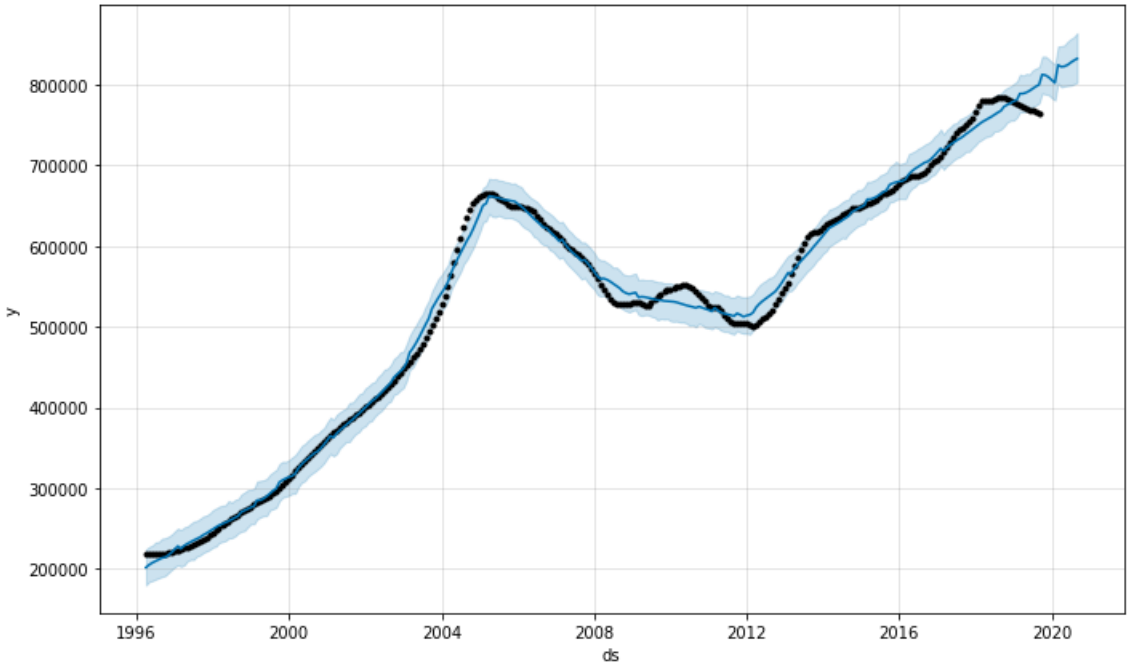
```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.
```



San Diego Prediction with full time series

As I suspected, the model appears to be overfit and may miss big drops in home values. This lead me to believe that we should select the zipcodes with the least amount of movement in the time series data: Austin, Charleston, and Sioux Falls. These zipcodes are likely better insulated from large shifts in the economy, which would lead to better prediction and safer investment. We'll continue forward with analysis of the three zipcodes, but will drop San Jose as the timeseries revealed a big dip in the house market in that area.

```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.
```



Charleston 29407 Predictions

```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.
```

```
WARNING:fbprophet:Seasonality has period of 365.25 days which is larger than initi
al window. Consider increasing initial.
INFO:fbprophet:Making 42 forecasts with cutoffs between 1997-06-01 06:52:12 and 20
17-11-30 18:10:48
INFO:fbprophet:n_changepoints greater than number of observations.Using 11.
INFO:fbprophet:n_changepoints greater than number of observations.Using 15.
INFO:fbprophet:n_changepoints greater than number of observations.Using 20.
```

Charleston 2018 Predictions vs 2018 Actual

It looks like prophet is short changing us a little on the prediction here at a 95% confidence level. The prediction expects about 10-30K lower than the actual home value.

Out[246]:

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
500	2018-08-01	287737.602526	279143.022810	296212.871143	310600.0	2017-11-30 18:10:48
501	2018-09-01	289801.500786	281067.084783	298605.600620	312800.0	2017-11-30 18:10:48
502	2018-10-01	291708.265310	282470.289483	300742.971064	316000.0	2017-11-30 18:10:48
503	2018-11-01	293522.318932	284831.128955	302606.344997	318900.0	2017-11-30 18:10:48
504	2018-12-01	294467.464724	285485.184266	304156.298938	321000.0	2017-11-30 18:10:48

Charleston Model Performance

Out[247]:

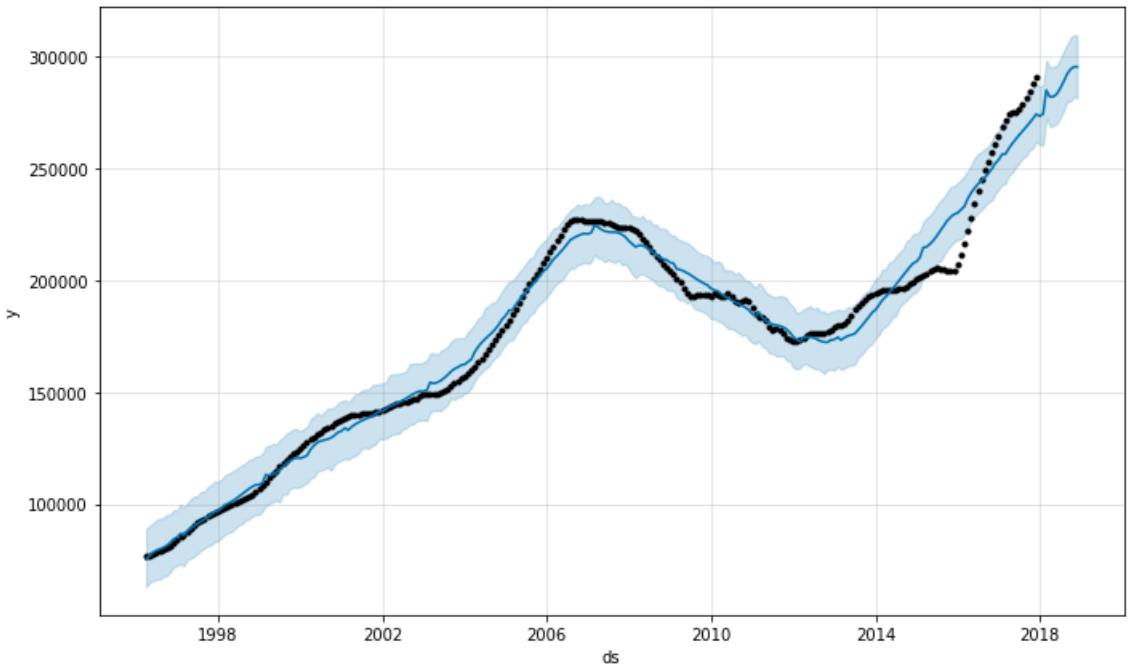
	horizon	mse	rmse	mae	mape	coverage
451	365 days 03:39:36	4.737118e+08	21764.920796	16393.801164	0.082580	0.32
452	365 days 04:22:48	4.770274e+08	21840.956269	16639.538337	0.083840	0.30
453	365 days 05:06:00	4.890219e+08	22113.840280	17004.208088	0.085504	0.30
454	365 days 05:29:24	4.553391e+08	21338.675634	16204.923071	0.081918	0.32
455	365 days 05:49:12	4.685729e+08	21646.545113	16605.522168	0.082829	0.32

R squared

Out[248]: 0.864926055956807

Charleston Prediction Plot

```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.
```



Austin 78751

```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.


WARNING:fbprophet:Seasonality has period of 365.25 days which is larger than initi
al window. Consider increasing initial.
INFO:fbprophet:Making 42 forecasts with cutoffs between 1997-06-01 06:52:12 and 20
17-11-30 18:10:48
INFO:fbprophet:n_changepoints greater than number of observations.Using 11.
INFO:fbprophet:n_changepoints greater than number of observations.Using 15.
INFO:fbprophet:n_changepoints greater than number of observations.Using 20.
```

Austin 2018 Predictions vs 2018 Actual

Predictions here are within 10K of the actual value. This model appears to be pretty accurate in forecasting this market, but could be too overfit to anticipate rapid change (positive or negative).

Out[253]:

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
500	2018-08-01	466224.568008	458061.838382	473957.722889	455600.0	2017-11-30 18:10:48
501	2018-09-01	468739.110691	460766.211784	477001.138177	458400.0	2017-11-30 18:10:48
502	2018-10-01	471169.780381	463050.863651	479788.633652	461500.0	2017-11-30 18:10:48
503	2018-11-01	473746.242824	464148.720122	483312.884454	465900.0	2017-11-30 18:10:48
504	2018-12-01	476643.740549	467238.947148	486857.929680	471200.0	2017-11-30 18:10:48

Austin Model Performance

Out[254]:

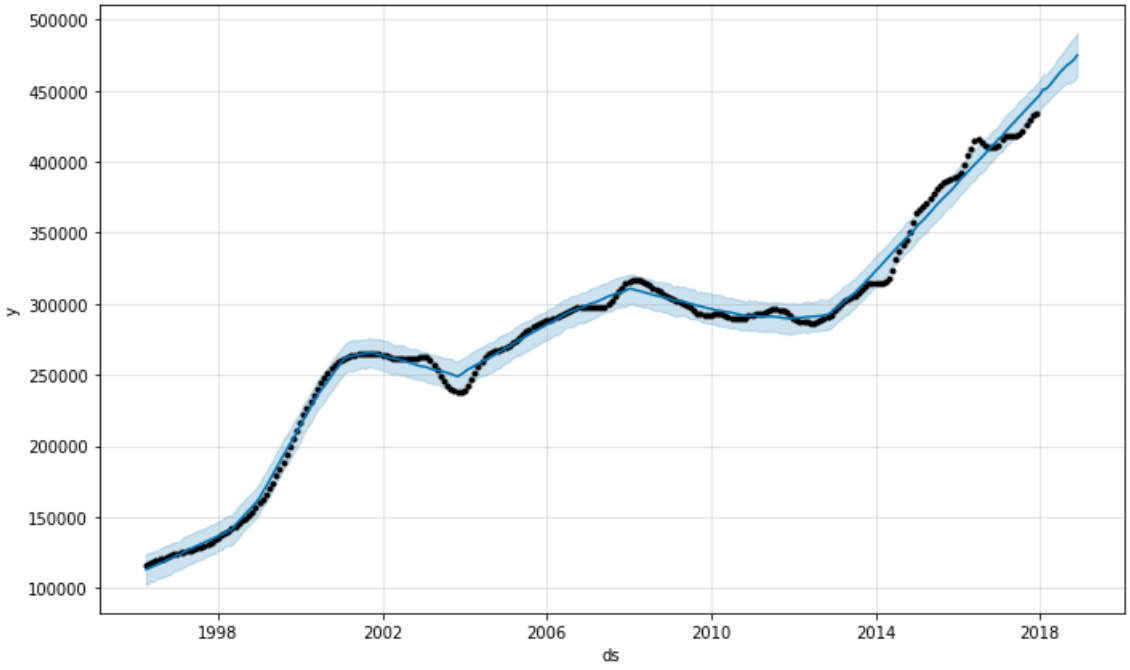
	horizon	mse	rmse	mae	mape	coverage
451	365 days 03:39:36	1.057846e+09	32524.540717	24226.737425	0.093834	0.26
452	365 days 04:22:48	1.052750e+09	32446.108613	24074.116788	0.092978	0.26
453	365 days 05:06:00	1.111790e+09	33343.518953	25090.187187	0.095741	0.24
454	365 days 05:29:24	1.120508e+09	33473.997107	25476.919546	0.097383	0.22
455	365 days 05:49:12	1.119881e+09	33464.625980	25429.590867	0.097069	0.22

R squared

Out[255]: 0.725068791140455

Austin Prediction Plot

```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.
```



Sioux Falls 57103

```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.
```

Sioux Falls 2018 Predictions vs 2018 Actual

Sioux Falls is a cheap place to live! This model is predicting within $5,000 accurately. Should maybe buy a plan ticket.

```
WARNING:fbprophet:Seasonality has period of 365.25 days which is larger than initi
al window. Consider increasing initial.
INFO:fbprophet:Making 42 forecasts with cutoffs between 1997-06-01 06:52:12 and 20
17-11-30 18:10:48
INFO:fbprophet:n_changepoints greater than number of observations.Using 11.
INFO:fbprophet:n_changepoints greater than number of observations.Using 15.
INFO:fbprophet:n_changepoints greater than number of observations.Using 20.
```

Out[260]:

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
500	2018-08-01	170083.761514	167025.395272	172990.251769	173200.0	2017-11-30 18:10:48
501	2018-09-01	170658.147905	167381.067822	173739.642532	174500.0	2017-11-30 18:10:48
502	2018-10-01	171252.406638	168203.346025	174223.856632	175900.0	2017-11-30 18:10:48
503	2018-11-01	171884.554493	168484.658158	175053.791697	177200.0	2017-11-30 18:10:48
504	2018-12-01	172474.321423	169174.910111	175837.102770	178100.0	2017-11-30 18:10:48

Sioux Falls Model Performance

Out[261]:

	horizon	mse	rmse	mae	mape	coverage
451	365 days 03:39:36	9.852152e+07	9925.800973	7740.731538	0.071863	0.28
452	365 days 04:22:48	9.510743e+07	9752.303860	7565.048253	0.069402	0.28
453	365 days 05:06:00	9.710217e+07	9854.043326	7752.794172	0.070745	0.26
454	365 days 05:29:24	1.011640e+08	10058.031070	7949.788293	0.073481	0.26
455	365 days 05:49:12	1.015681e+08	10078.100878	7994.650271	0.073588	0.26

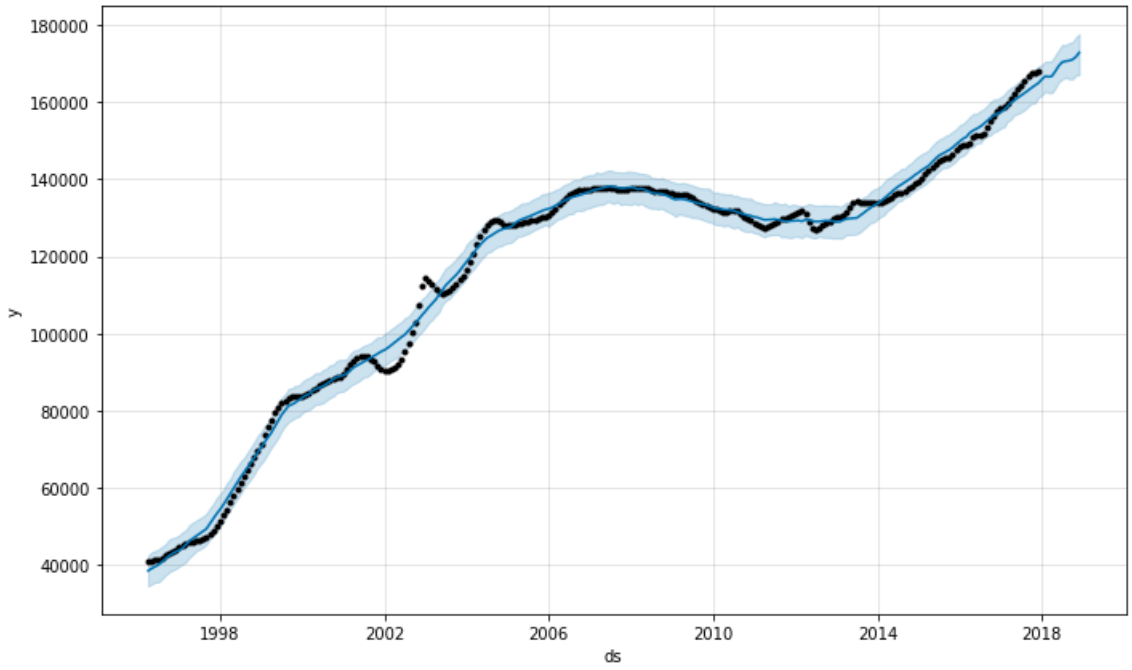
R squared

Out[262]: 0.9011630207852886

```
INFO:fbprophet:Disabling weekly seasonality. Run prophet with weekly_seasonality=T
rue to override this.
INFO:fbprophet:Disabling daily seasonality. Run prophet with daily_seasonality=Tru
e to override this.
```

	ds	yhat	yhat_lower	yhat_upper
268	2018-07-31	170501.611003	166324.672765	174729.579497
269	2018-08-31	170636.816636	165903.349804	175225.073213
270	2018-09-30	170909.308944	165857.545228	175439.313643
271	2018-10-31	171609.040946	166810.453017	176732.969799
272	2018-11-30	172735.023023	166978.264733	177579.724616

```
Prediction for city Sioux Falls for 2018 172735.023023174
```



Conclusions

Three Zipcodes:

I'm going to go with: Austin, 78751; Sioux Falls, 57103; and Charleston, 29407

The average growth for all zipcodes analyzed were above 70%; however, the future predictons can lead us astray if new data comes out that the home values have taken a sudden downfall due to the overfitting of Prophet.

In the analysis with Prophet we went for a 95% confidence interval. We could see that the predictions within this interval in 2018 (yhat lower and yhat upper) were accurate and the predicted value was typically within ~10K of the actual value -- pretty good when considering some values like Austin were a half million. Zipcodes in smaller communities like Sioux Falls may not bring the most %yield, but they are likely safer bets if another recession is looming.

Further analysis that might be beneficial could include a deeper dive into the selected cities such as:

- What industries in the zipcodes have been/could be affected by the China Trade War?
- What are the updated government census statistics?
- Are the number of Walmarts in a location an indicator for population growth? (just kidding, kind of)

References

Bowne-Anderson, Hugo. "Time Series Analsis Tutorial with Python". DataCamp, 1 Jan. 2018, https://www.datacamp.com/community/tutorials/time-series-analysis-tutorial (https://www.datacamp.com/community/tutorials/time-series-analysis-tutorial)

Fernando, Aguilar. "Time Series Analysis on Zillow's Housing Data". Medium, 15 Jul, 2019, https://medium.com/@feraguilari/time-series-analysis-modfinalproyect-b9fb23c28309 (https://medium.com/@feraguilari/time-series-analysis-modfinalproyect-b9fb23c28309)

Unknown. "How to get cross validation and performance metrics on monthly data in Python". GitHub, 29 Apr, 2019, https://github.com/facebook/prophet/issues/949 (https://github.com/facebook/prophet/issues/949)

Brown, Eric. "Forcasting Time Series data with Prophet - Part 4". Python Data, 1 Jan, 2018, https://pythondata.com/forecasting-time-series-data-prophet-part-4/ (https://pythondata.com/forecasting-time-series-data-prophet-part-4/)