Top 5 Zip Codes to Invest Real Estate

What this project is about:

We were tasked by a real estate firm to help them decide which zip codes they should invest in.

Summary

We approached the problem with a top-down approach. First we leveraged U.S. Census data in order to narrow down to which state we want to invest in. Then we utilized the given dataset from Zillow to further narrow down our choices. We focused mainly on where the people are going and which zip codes were the largest in the area. Once we had our zip codes, we used Facebook Prophet to run a time series analysis on each zip code. We created a model to forecast what the prices would look like for the next few years. Based on our observations, the top five zip codes we chose showed promise in the future for rising home values.

The Data

I utilized a custom dataset from the U.S. Census where I obtained the population of each zip code in the chosen state in order to order the zip codes by population. Dataset is taken from the U.S. Census API with a **custom link that I made**

Additionally, the Zillow Dataset from our partners at Flatiron School was used. **Link to the Dataset**

Agenda

1. Look at general population trends for every state in the U.S.
2. Narrow down from state to city and then to the most populous zip codes in that city.
3. Run time series analysis using Facebook Prophet and forecast with model.
4. Interpret model results and diagnostics.
5. Give business recommendations from analysis.

Assumptions

1. Focus on where people are moving to. Which state are people leaving and to which state are the most people going to?
2. The investment firm is a smaller firm looking to expand into a new area.
3. The firm will want to have clusters of zip codes nearby for ease of management.
4. Firm will not be outsourcing work to other property managers. Work will be done in-house.
5. We will not be buying apartment buildings but are open to do so in the future.
6. We will look for areas where laws are favorable to landlords as a bonus.
7. Since dataset given has data until April 2018, we will use data on or before that date to simulate real time

Initial Obtain/Scrub/Exploration of Data

Based on this 2017 U.S. Census press release for fastest growing states, the three photos below are sourced from the article.

Top 10 States in Numeric Growth: 2016 to 2017

Rank	Name	2010	2016	2017	Numeric growth
1	Texas	25,146,100	27,904,862	28,304,596	399,734
2	Florida	18,804,594	20,656,589	20,984,400	327,811
3	California	37,254,518	39,296,476	39,536,653	240,177
4	Washington	6,724,545	7,280,934	7,405,743	124,809
5	North Carolina	9,535,721	10,156,689	10,273,419	116,730
6	Georgia	9,688,690	10,313,620	10,429,379	115,759
7	Arizona	6,392,309	6,908,642	7,016,270	107,628
8	Colorado	5,029,325	5,530,105	5,607,154	77,049
9	Tennessee	6,346,295	6,649,404	6,715,984	66,580
10	South Carolina	4,625,381	4,959,822	5,024,369	64,547

Top 10 Most Populous States: 2017

Rank	Name	2010	2016	2017
1	California	37,254,518	39,296,476	39,536,653
2	Texas	25,146,100	27,904,862	28,304,596
3	Florida	18,804,594	20,656,589	20,984,400
4	New York	19,378,110	19,836,286	19,849,399
5	Pennsylvania	12,702,857	12,787,085	12,805,537
6	Illinois	12,831,565	12,835,726	12,802,023
7	Ohio	11,536,730	11,622,554	11,658,609
8	Georgia	9,688,690	10,313,620	10,429,379
9	North Carolina	9,535,721	10,156,689	10,273,419
10	Michigan	9,884,129	9,933,445	9,962,311

Top 10 States in Percentage Growth: 2016 to 2017

Rank	Name	2010	2016	2017	Percent growth
1	Idaho	1,567,650	1,680,026	1,716,943	2.2
2	Nevada	2,700,691	2,939,254	2,998,039	2.0
3	Utah	2,763,889	3,044,321	3,101,833	1.9
4	Washington	6,724,545	7,280,934	7,405,743	1.7
5	Florida	18,804,594	20,656,589	20,984,400	1.6
6	Arizona	6,392,309	6,908,642	7,016,270	1.6
7	Texas	25,146,100	27,904,862	28,304,596	1.4
8	District of Columbia	601,766	684,336	693,972	1.4
9	Colorado	5,029,325	5,530,105	5,607,154	1.4
10	Oregon	3,831,072	4,085,989	4,142,776	1.4

I took this info and color coded which states showed up multiple times. Although Idaho is the main winner, we see that Texas shows up in all 3 categories of Numeric, Populous, and Percentage Growth. Therefore we will go with Texas as our choice.

Most Populous	Numeric Growth	Percentage Growth
California	Texas	Idaho
Texas	Florida	Nevada
Florida	California	Utah
New York	Washington	Washington
Pennsylvania	North Carolina	Florida
Illinois	Georgia	Arizona
Ohio	Arizona	Texas
Georgia	Colorado	District of Columbia
North Carolina	Tennessee	Colorado
Michigan	South Carolina	Oregon

Using Texas as our starting point, we see that based on value_counts() for the Zillow Data, Houston has the highest number of zip codes.

Therefore the city we will look at is Houston.

```
df_texas_census.head()
```

	NAME	S0101_C01_001E
69	ZCTA5 77084	104582
28	ZCTA5 77036	76605
80	ZCTA5 77095	72081
59	ZCTA5 77072	62162
63	ZCTA5 77077	57757

After choosing Houston, I obtained a list of all the zip codes which are in Houston and pulled the population data for each Houston zip code to find the top five most populous zip codes.

Census data can be **found here**

Custom Link can be accessed with **this link**

I did the following with the Census dataset:

1. Removing all columns I don't need. I only want zip code column and population column
2. Fixing the data type to int in order to sort properly
3. Sorting the dataframe by population highest to lowest
4. .head() to find the top 5

Now that we have the top 5 zip codes we do the following:

1. Use pd.melt() method to keep only the columns that we want and turn price data from row of values to column of values
2. I create new .csv files which only contain the zip code and the associated value of homes
3. Change column names to "ds" and "y" in order for the dataset to play nice with Facebook Prophet time series analysis
4. Run quick check for any nulls/Nans for my sanity

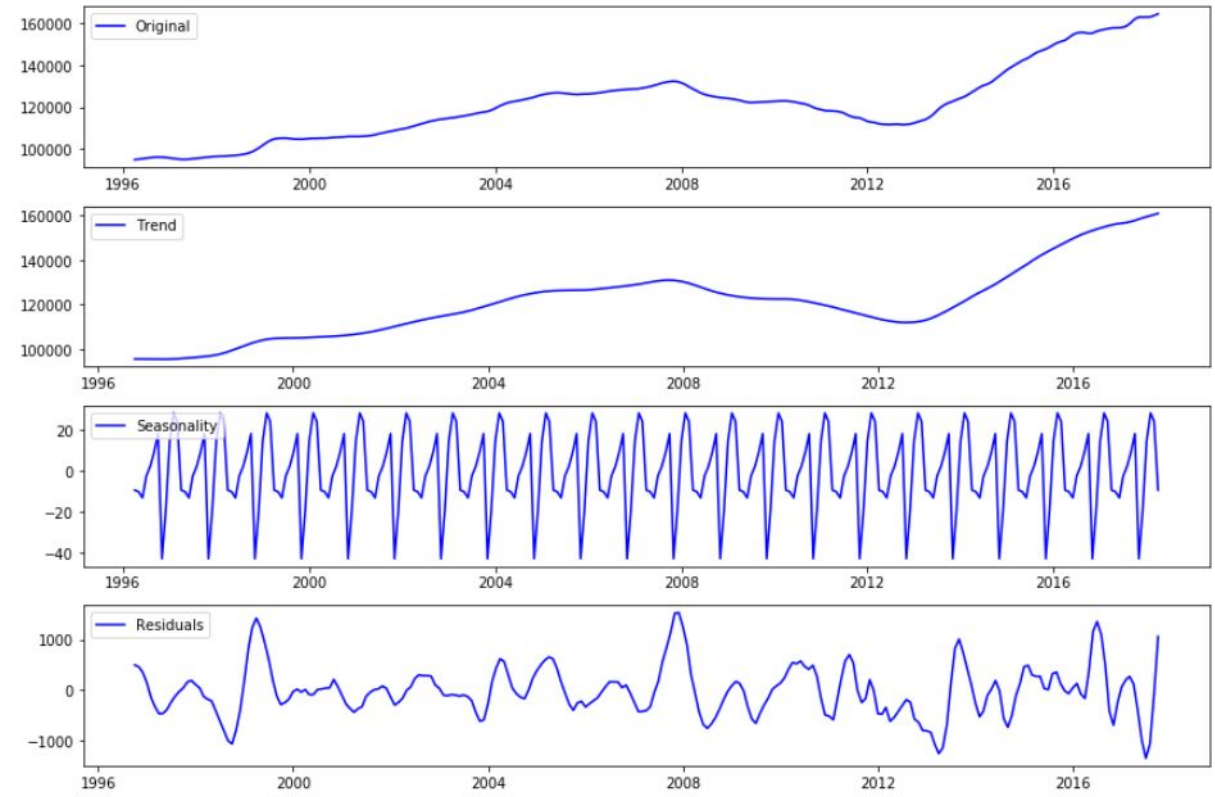
Once the zip codes are chosen, I made a separate .csv file for each zip code to prep it for time series analysis with FBProphet.

The prepped data files can be found **here**

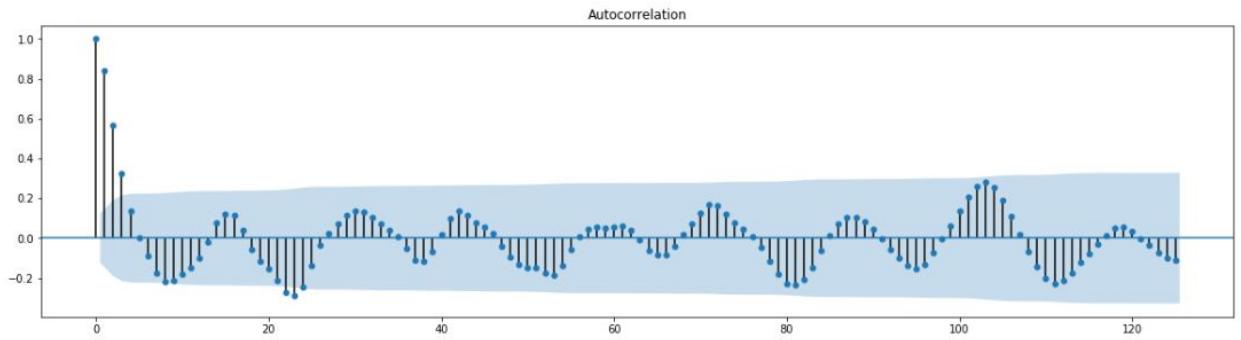
--- Analysis of each zip code: ---

[PICK #1] 77084 - Stationarity and ARMA model

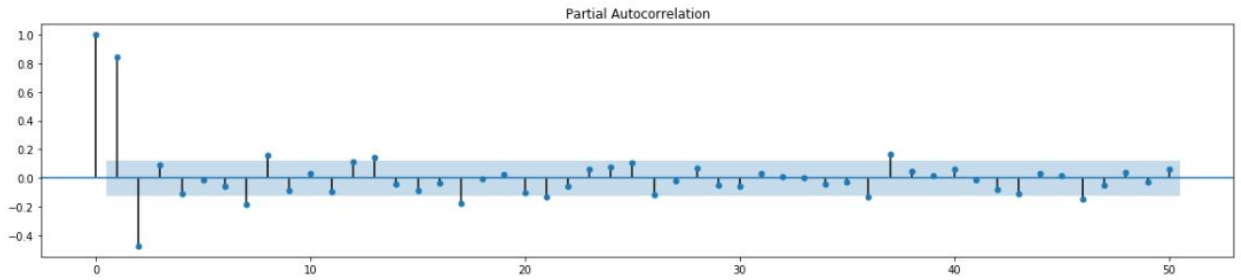
----------Decomposition of Series----------



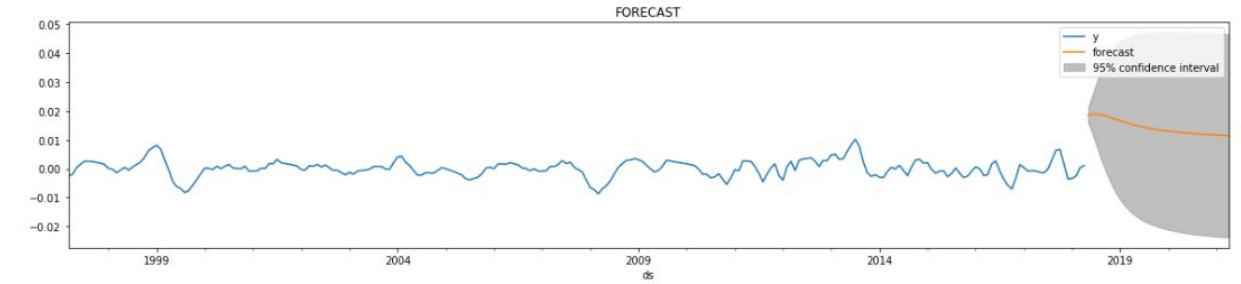
Auto-Correlation and Partial-Auto-Correlation (After De-trended and Transformed Series to Stationary)




```
#PACF plot
rcParams['figure.figsize'] = 20, 4
plot_pacf(logged_df_diff_roll_mean1, lags=50, alpha=0.05);
```



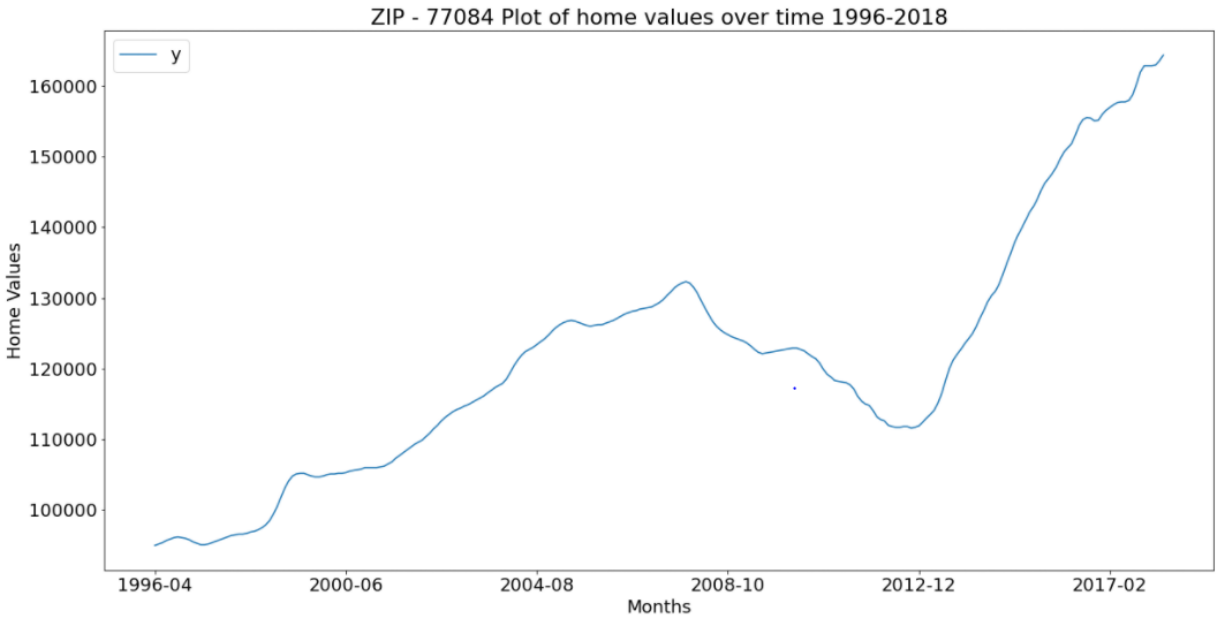
-------- ARMA Model with 3 year forecast ---------



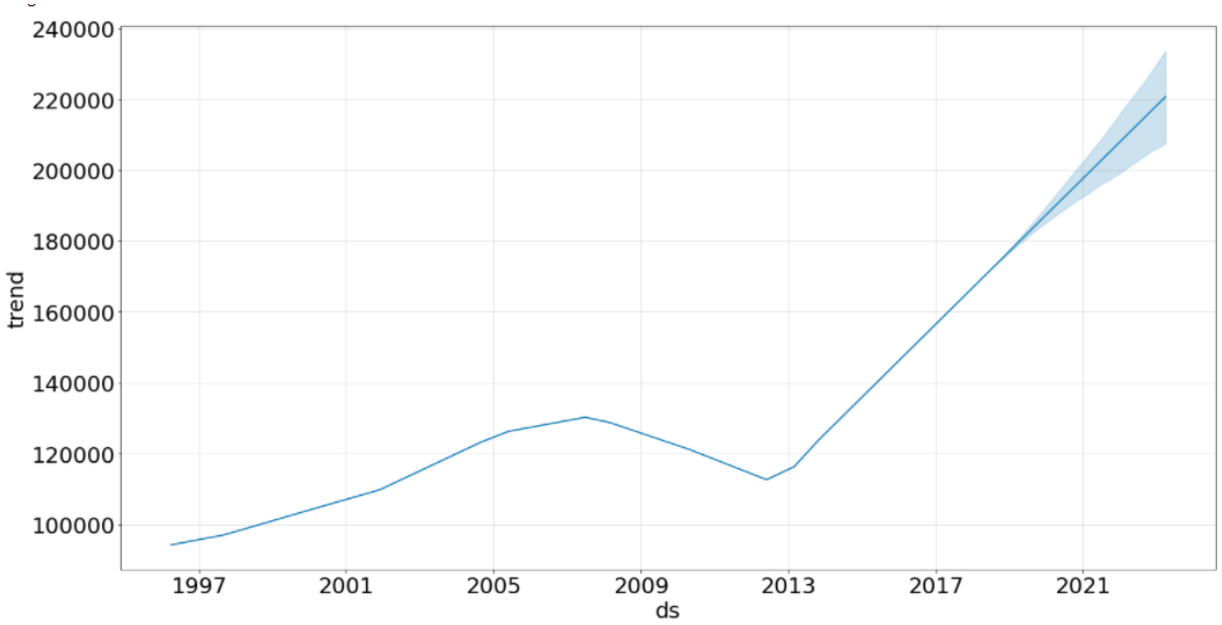
We see prices are predicted to be higher. Could be a good investment zip code.

[PICK #1] 77084 - FACEBOOK PROPHET MODEL

----------Plot of Data----------



----------Forecast----------

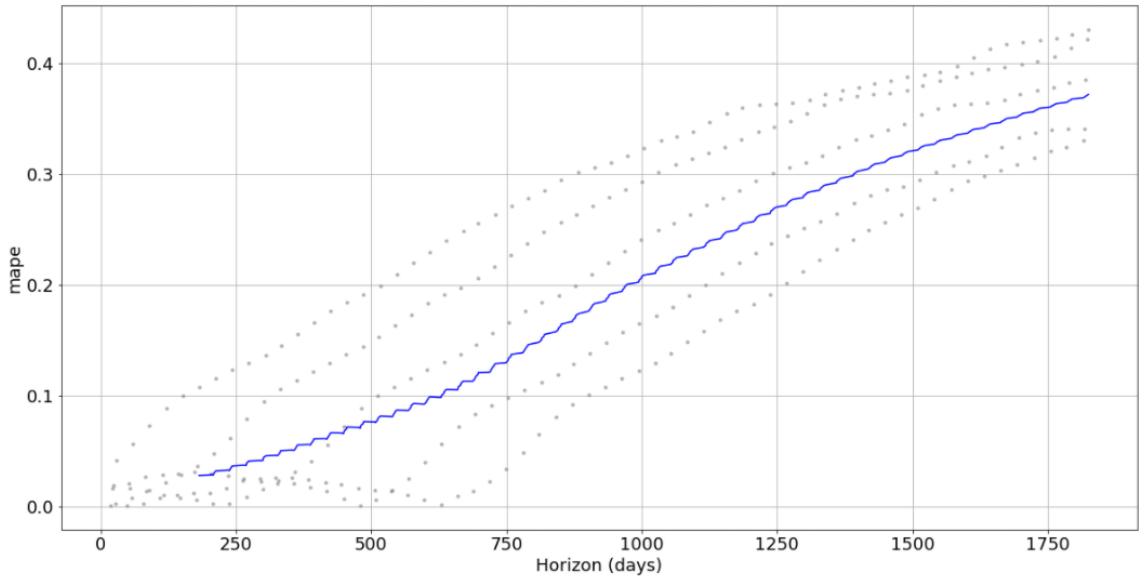


----------Analysis of MAPE (Mean Average Percent Error)----------

MAPE (Mean Average Percent Error) - Observation:

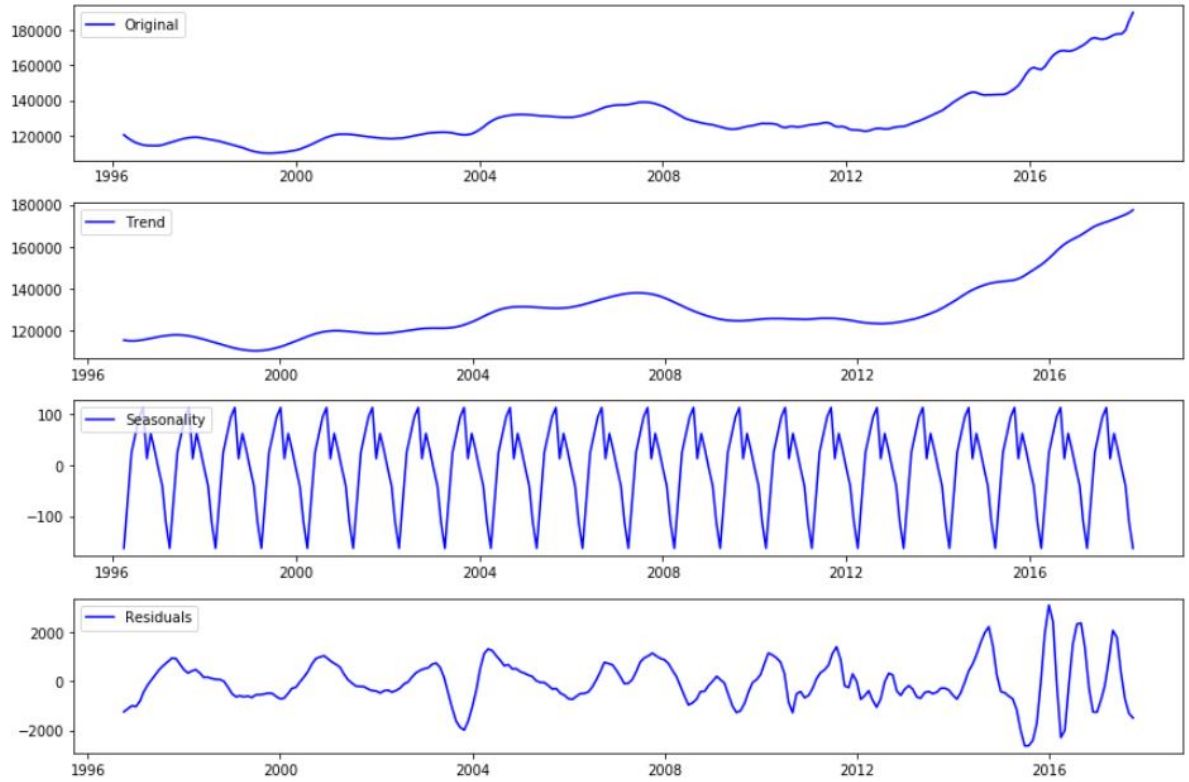
1. We see that MAPE increases over time
2. I am willing to tolerate MAPE of 0.1 to 0.2
    - This gets exceeded after about 1100 days
3. We will focus on MAPE as our main diagnostic metric.
    - Shows the model was about 80% accurate at 1000 days
    - Bullish prediction for the next 2-3 years
    - Supports the high upward trend we saw in the graph of all the data points for the zip code

[12]:
```
fig = plot_cross_validation_metric(cv_results, metric='mape', figsize=(20,10))
```

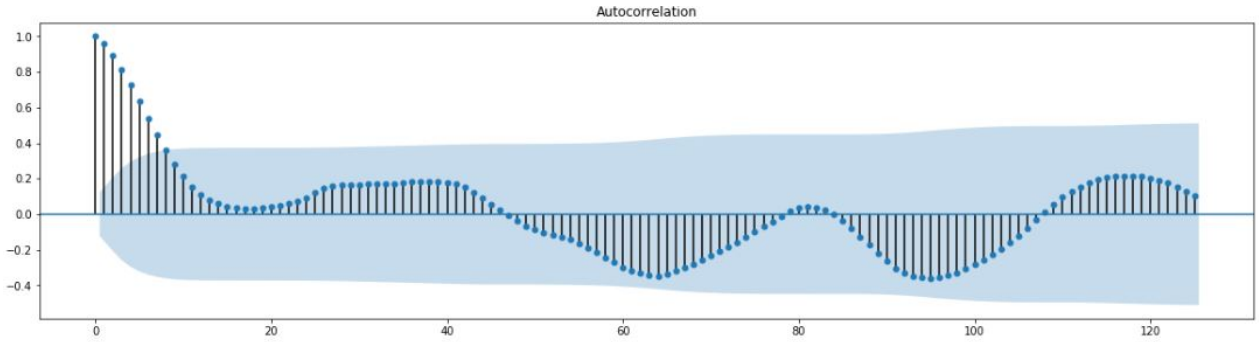


[PICK #2] 77036 - Stationarity and ARMA model

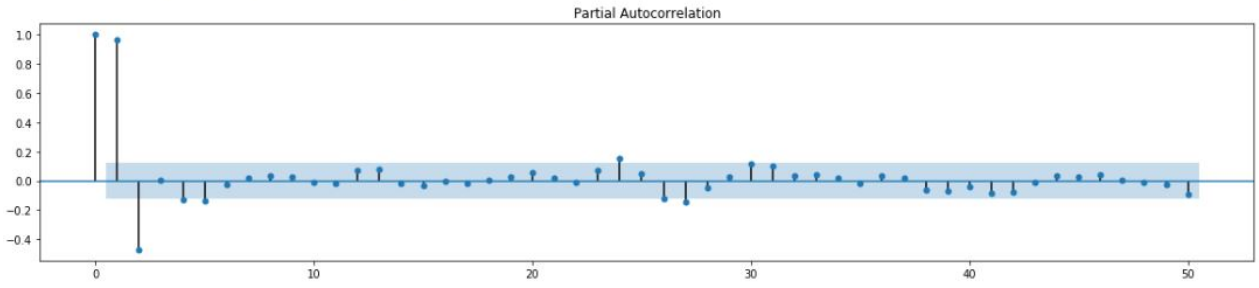
----------Decomposition of Series----------



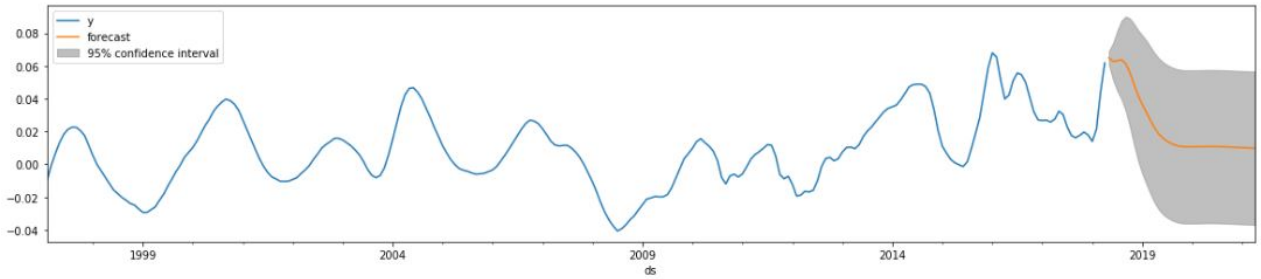
Auto-Correlation and Partial-Auto-Correlation (After De-trended and Transformed Series to Stationary)




```
#PACF plot
acParams['figure.figsize'] = 20, 4
plot_pacf(logged_df_minus_roll_mean1, lags=50, alpha=0.05);
```

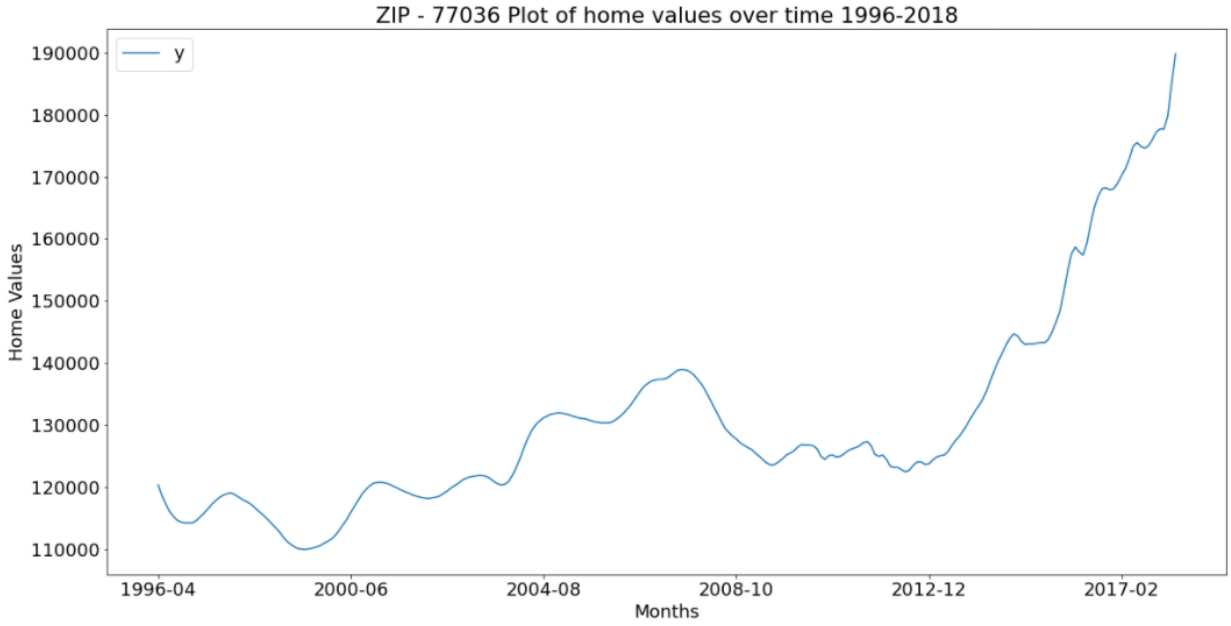


-------- ARMA Model with 3 year forecast ---------

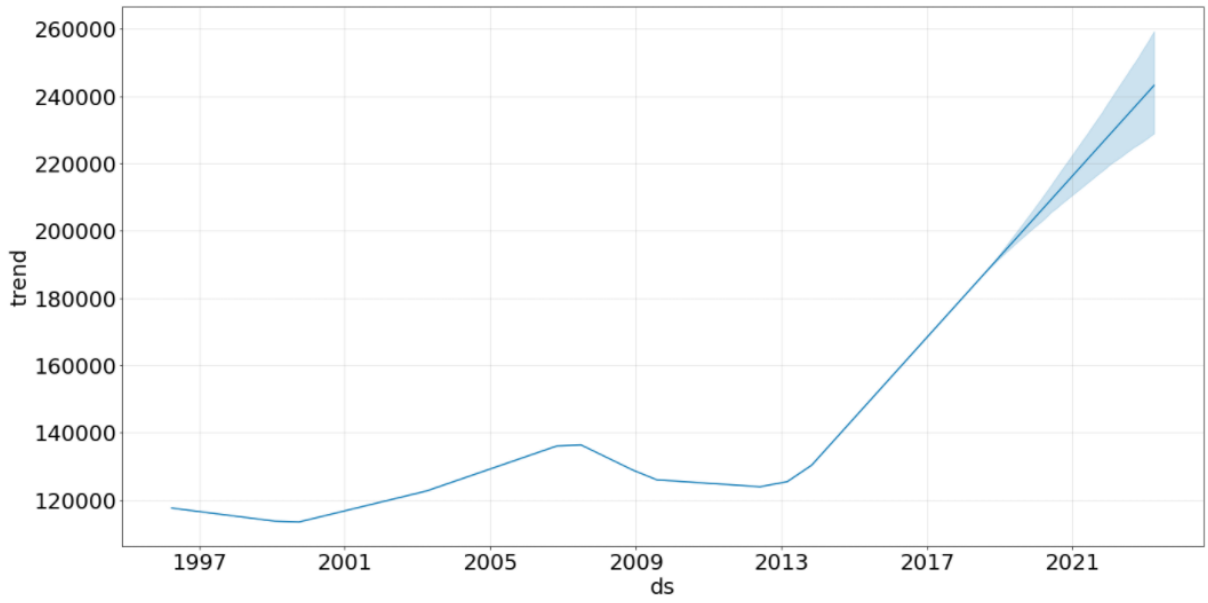


[PICK #2] 77036 - FACEBOOK PROPHET MODEL

----------Plot of Data----------



----------Forecast----------

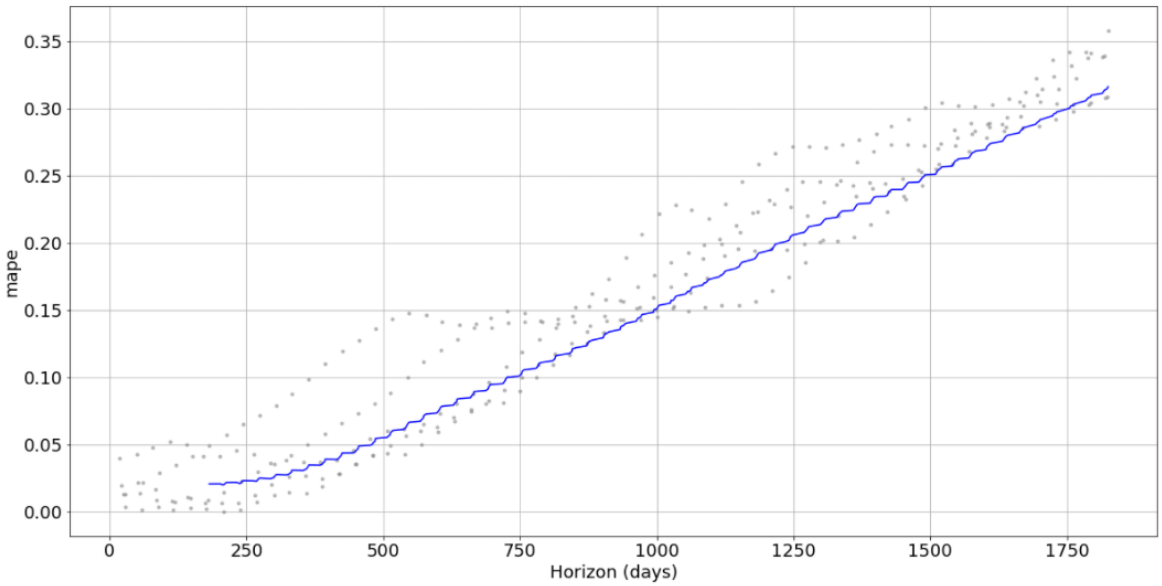


----------Analysis of MAPE (Mean Average Percent Error)----------

MAPE (Mean Average Percent Error) - Observation:

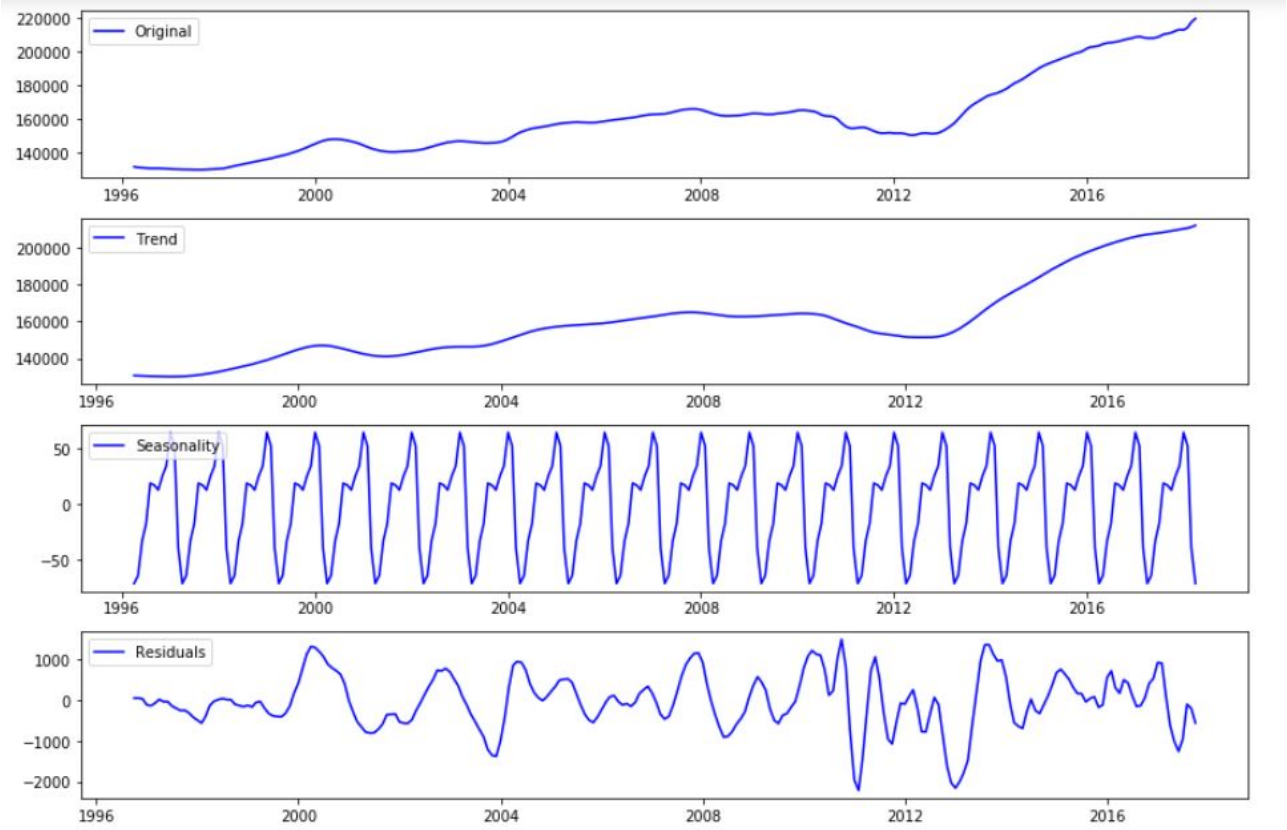
1. We see that MAPE increases over time
2. I am willing to tolerate MAPE of 0.1 to 0.2
    - This gets exceeded after about 1100 days
3. We will focus on MAPE as our main diagnostic metric.
    - Shows the model was about 85% accurate at 1000 days
    - Bullish prediction for the next 2-3 years
    - Supports the high upward trend we saw in the graph of all the data points for the zip code

n [12]:
```
fig = plot_cross_validation_metric(cv_results, metric='mape', figsize=(20,10))
```

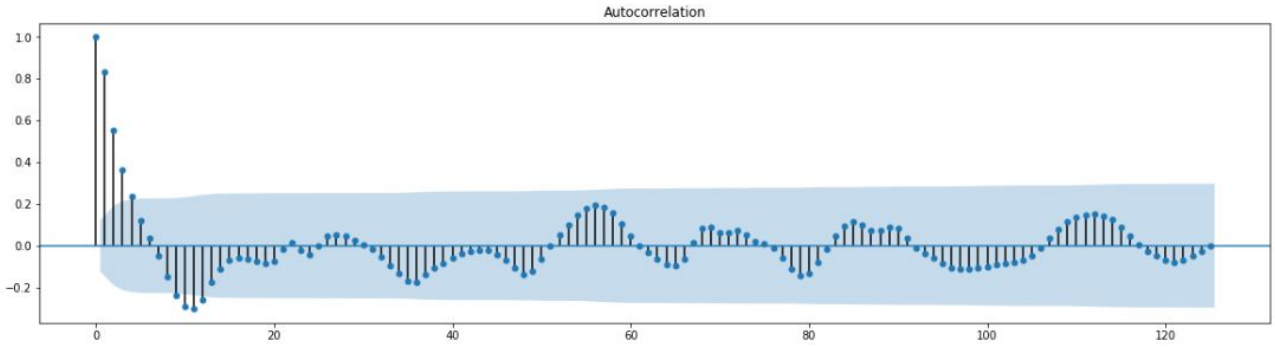


[PICK #3] 77095 - Stationarity and ARMA model

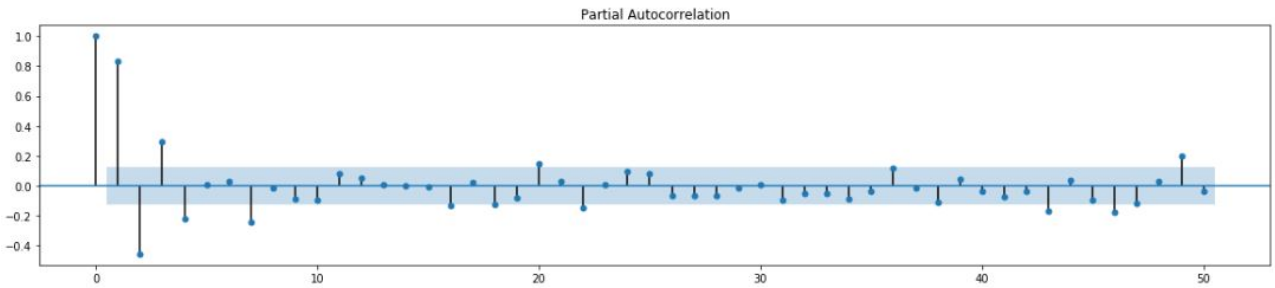
----------Decomposition of Series----------



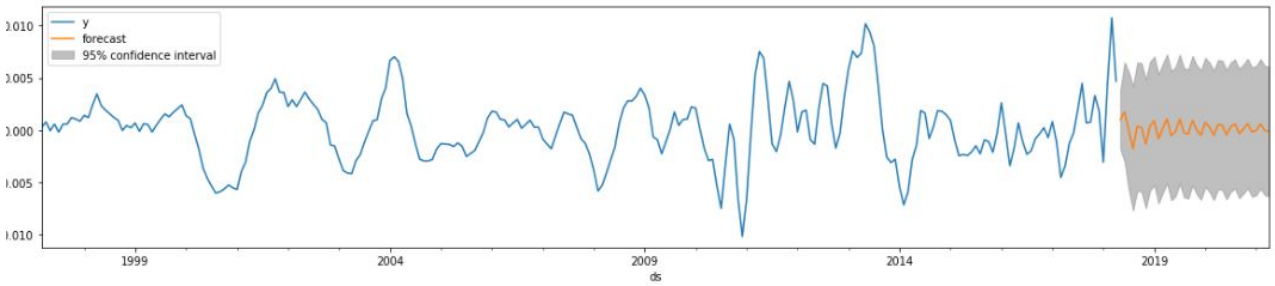
Auto-Correlation and Partial-Auto-Correlation (After De-trended and Transformed Series to Stationary)




```
#PACF plot
rcParams['figure.figsize'] = 20, 4
plot_pacf(logged_df_diff_roll_mean1, lags=50, alpha=0.05);
```

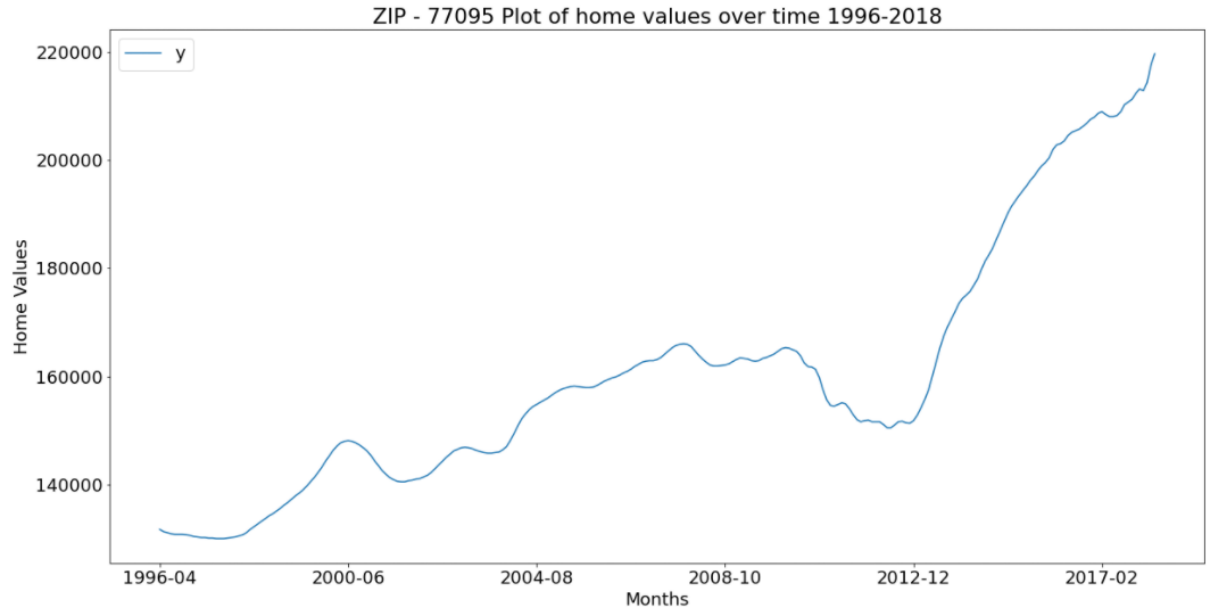


-------- ARMA Model with 3 year forecast ---------

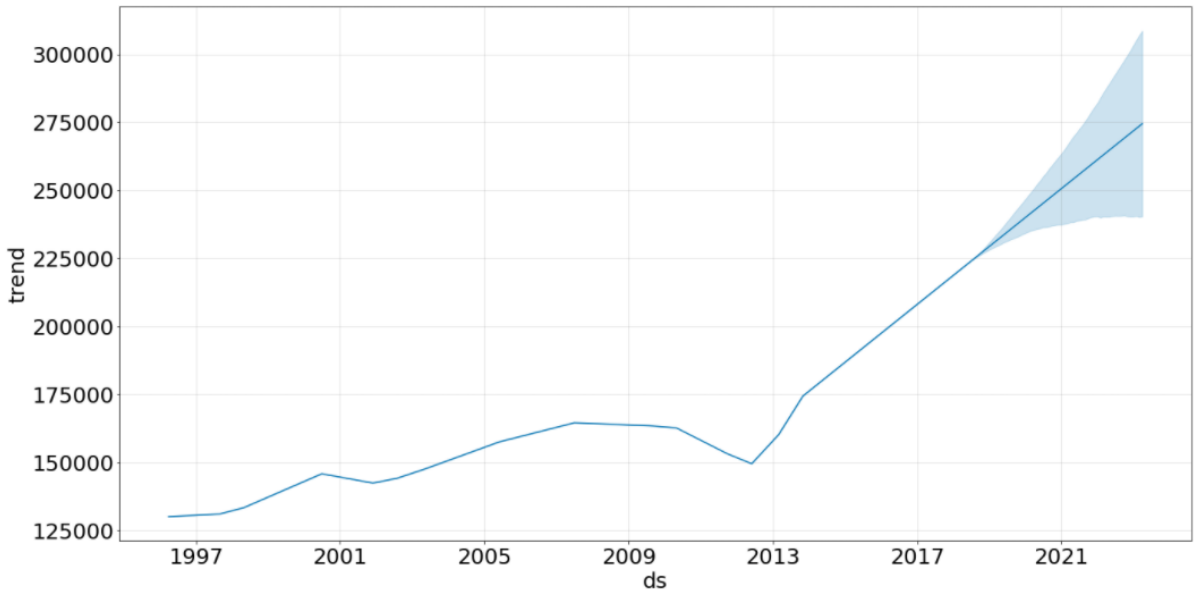


[PICK #3] 77095 - FACEBOOK PROPHET MODEL

----------Plot of Data----------



----------Forecast----------

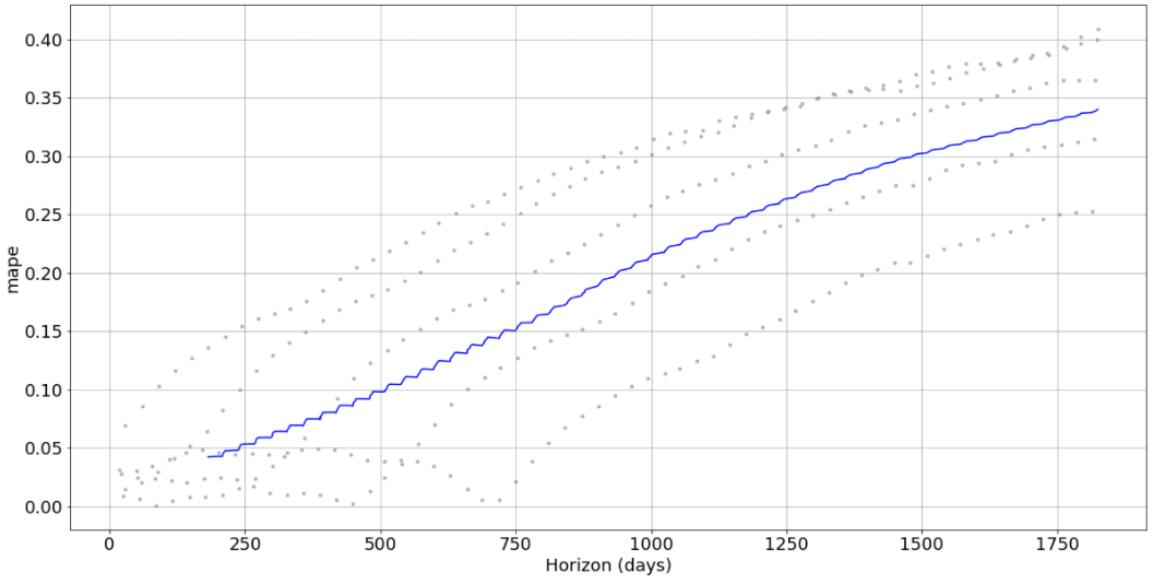


----------Analysis of MAPE (Mean Average Percent Error)----------

MAPE (Mean Average Percent Error) - Observation:

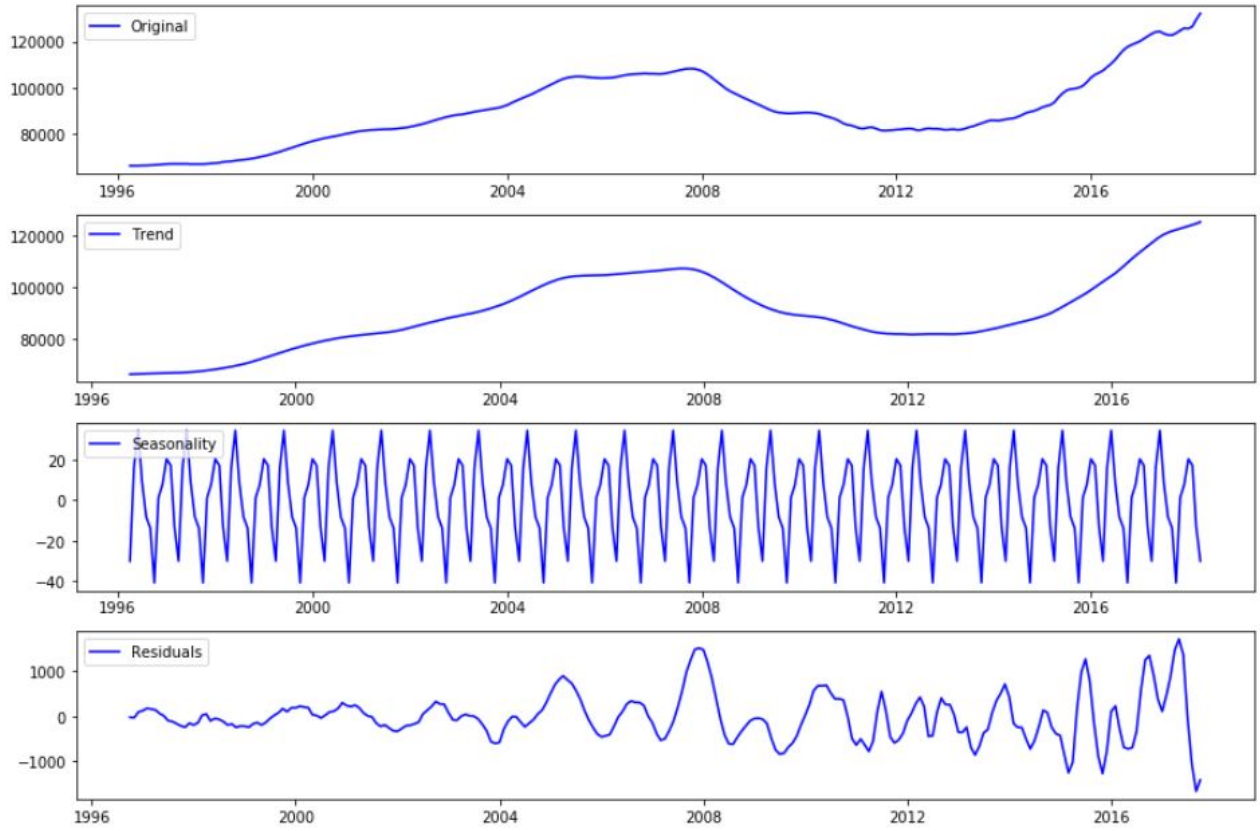
1. We see that MAPE increases over time
2. I am willing to tolerate MAPE of 0.1 to 0.2
    - This gets exceeded after about 800 days
3. We will focus on MAPE as our main diagnostic metric.
    - Shows the model was about 80% accurate at 800 days
    - Bullish prediction for the next 2-3 years
    - Supports the high upward trend we saw in the graph of all the data points for the zip code

In [12]:
```
fig = plot_cross_validation_metric(cv_results, metric='mape', figsize=(20,10))
```

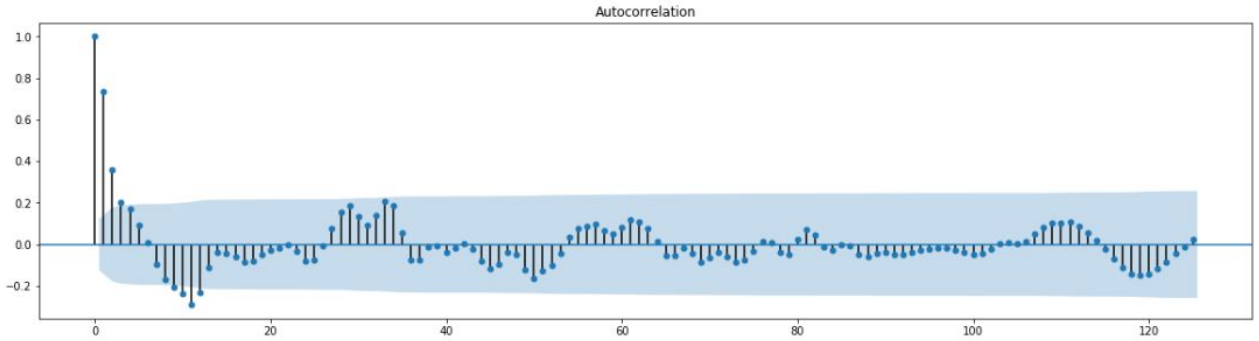


[PICK #4] 77072 - Stationarity and ARMA model

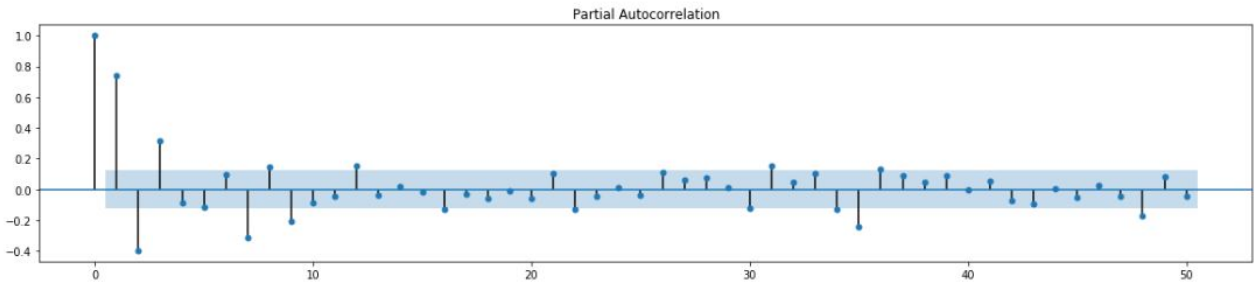
----------Decomposition of Series----------



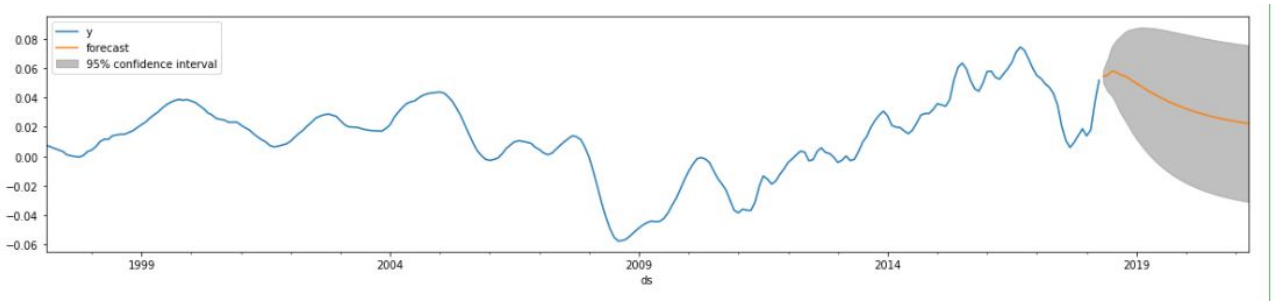
Auto-Correlation and Partial-Auto-Correlation (After De-trended and Transformed Series to Stationary)




```
#PACF plot
rcParams['figure.figsize'] = 20, 4
plot_pacf(logged_df_diff_roll_mean1, lags=50, alpha=0.05);
```

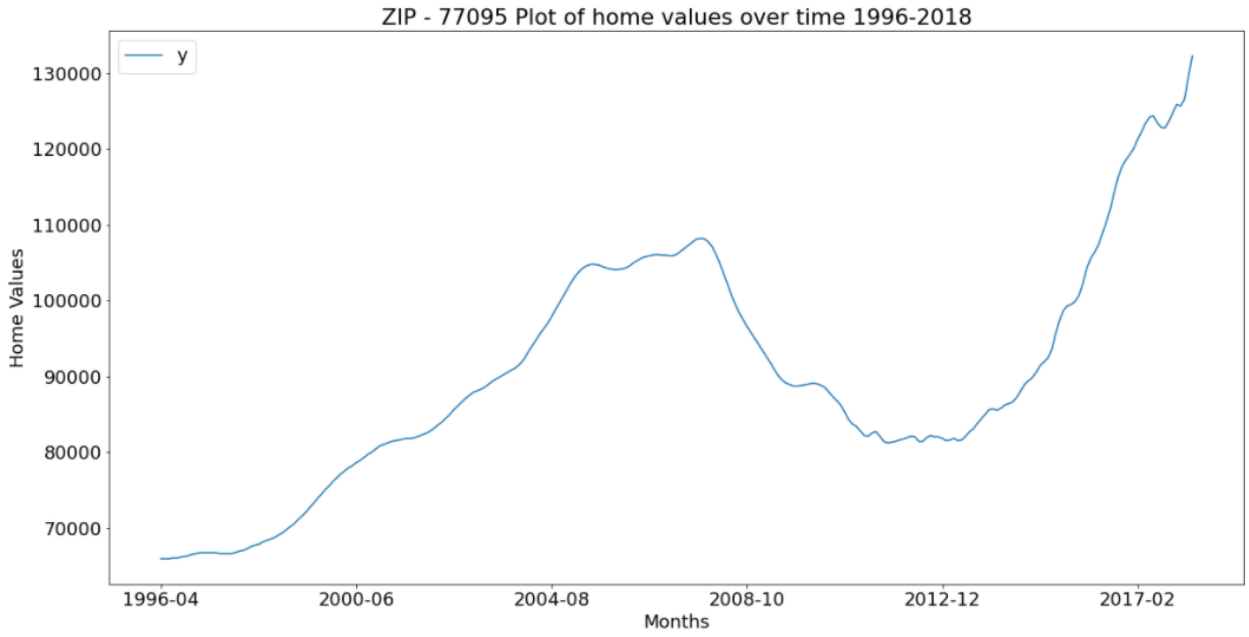


-------- ARMA Model with 3 year forecast ---------

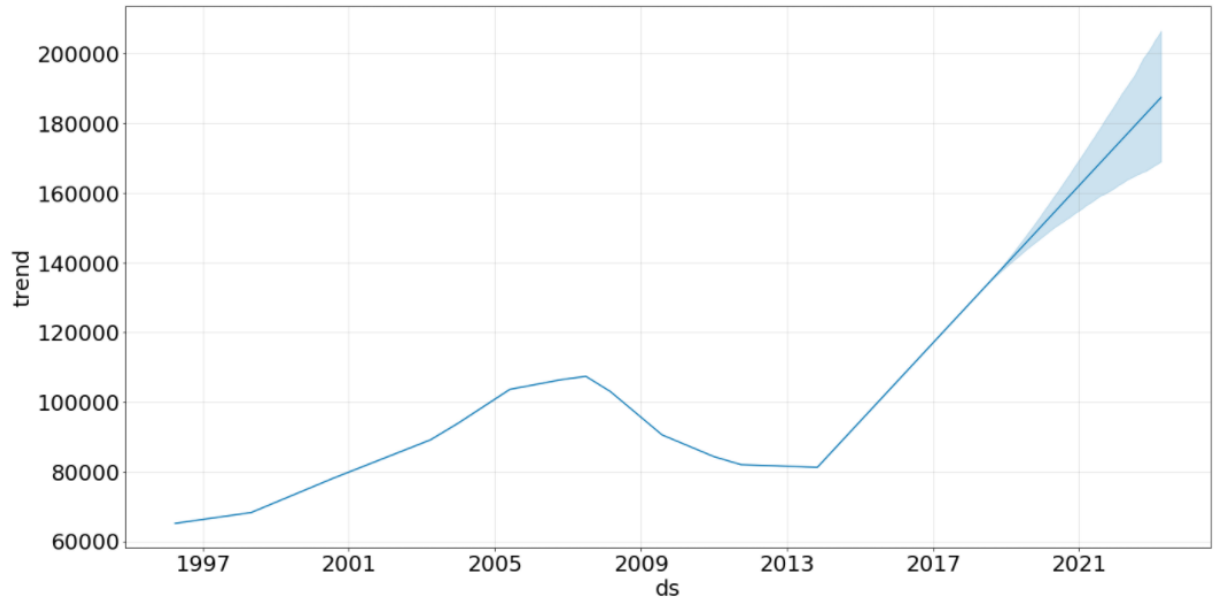


[PICK #4] 77072 - FACEBOOK PROPHET MODEL

----------Plot of Data----------



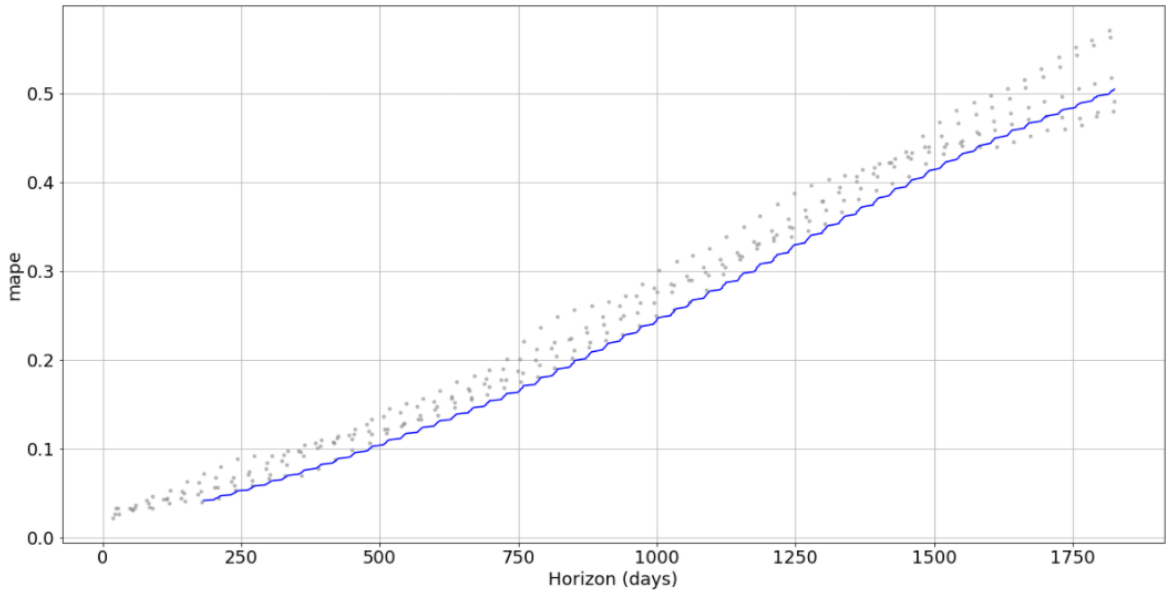
----------Analysis of MAPE (Mean Average Percent Error)----------



MAPE (Mean Average Percent Error) - Observation:

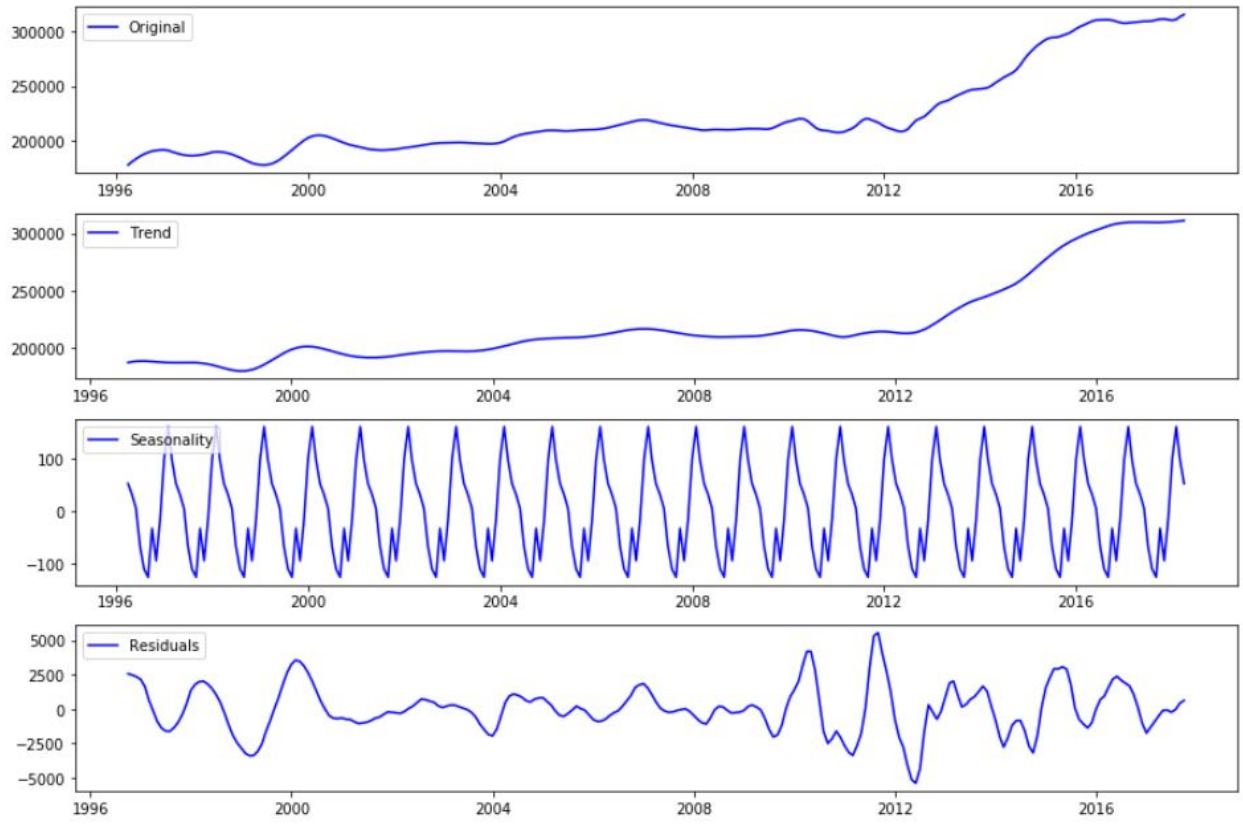
1. We see that MAPE increases over time
2. I am willing to tolerate MAPE of 0.1 to 0.2
   - This gets exceeded after about 800 days
3. We will focus on MAPE as our main diagnostic metric.
   - Shows the model was about 80% accurate at 800 days
   - Bullish prediction for the next 2-3 years
   - Supports the high upward trend we saw in the graph of all the data points for the zip code

In [14]:
```
fig = plot_cross_validation_metric(cv_results, metric='mape', figsize=(20,10))
```

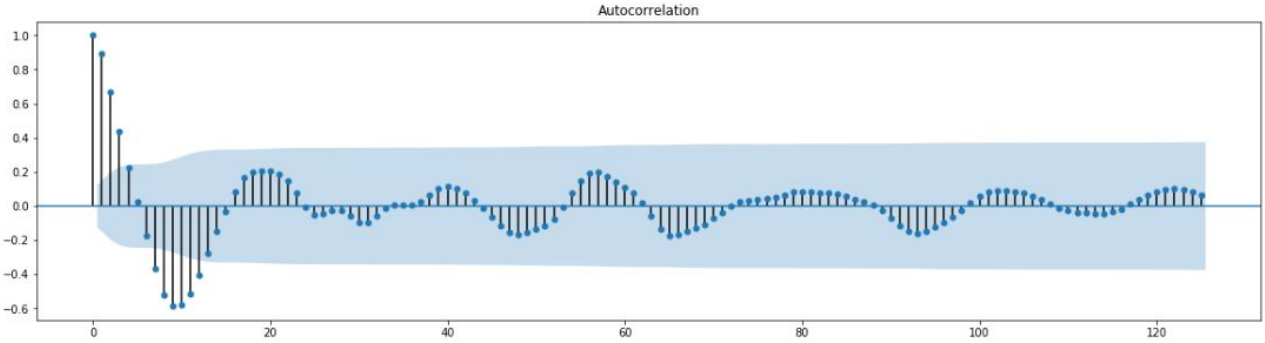


[PICK #5] 77077 - Stationarity and ARMA model

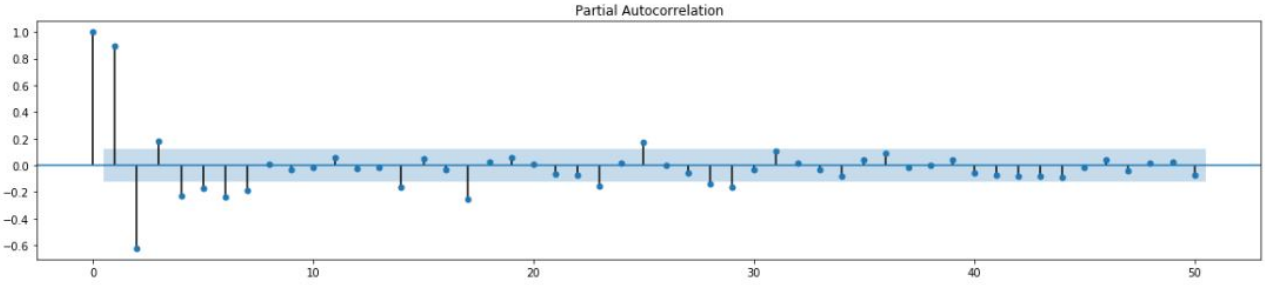
----------Decomposition of Series----------



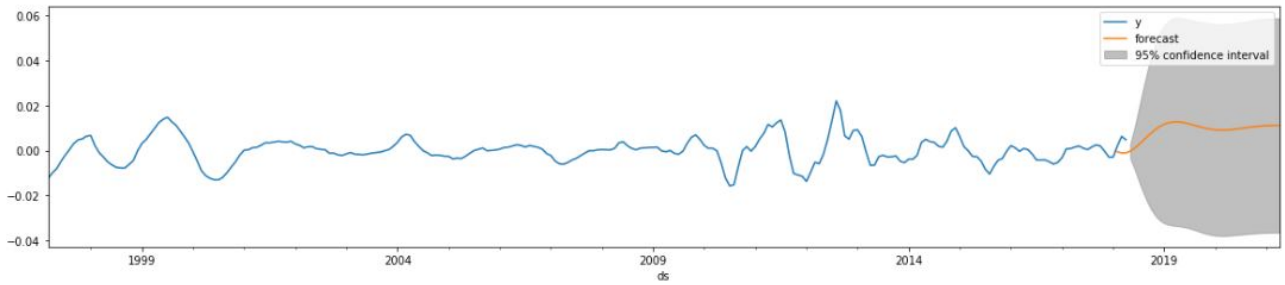
Auto-Correlation and Partial-Auto-Correlation (After De-trended and Transformed Series to Stationary)




```
#PACF plot
rcParams['figure.figsize'] = 20, 4
plot_pacf(logged_df_diff_roll_mean1, lags=50, alpha=0.05);
```

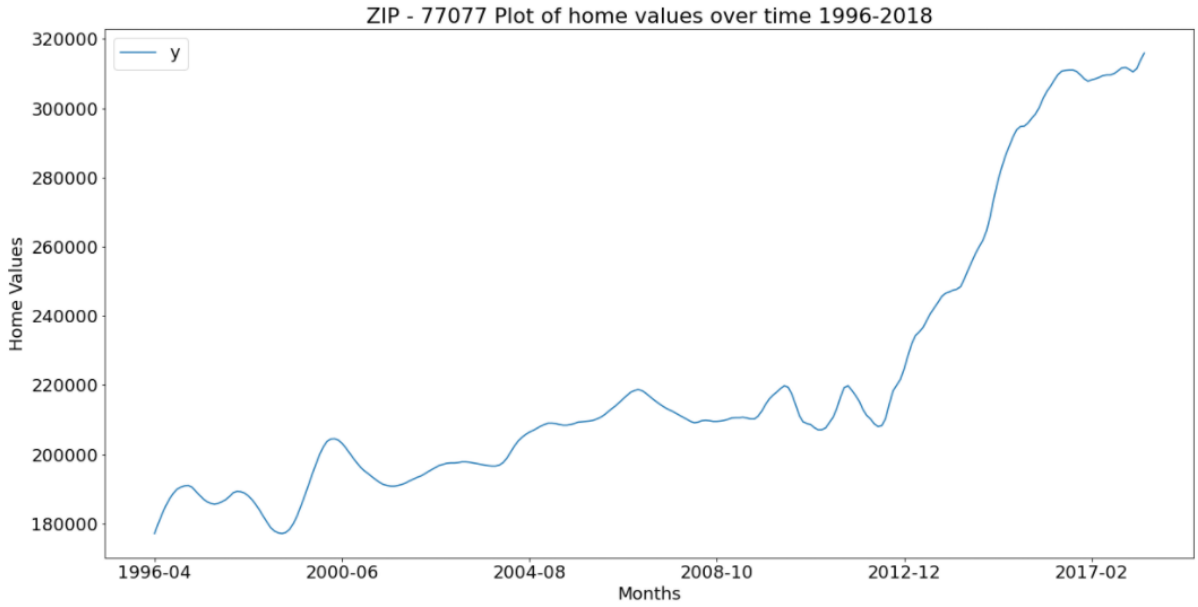


-------- ARMA Model with 3 year forecast ---------

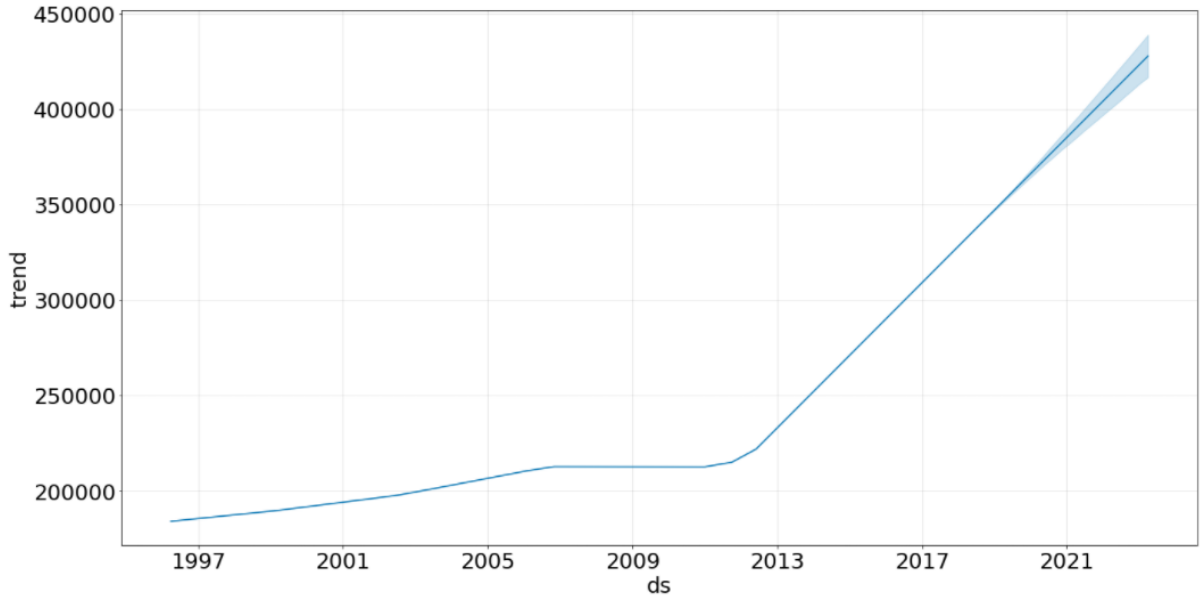


[PICK #5] 77077 - FACEBOOK PROPHET MODEL

----------Plot of Data----------



----------Forecast----------

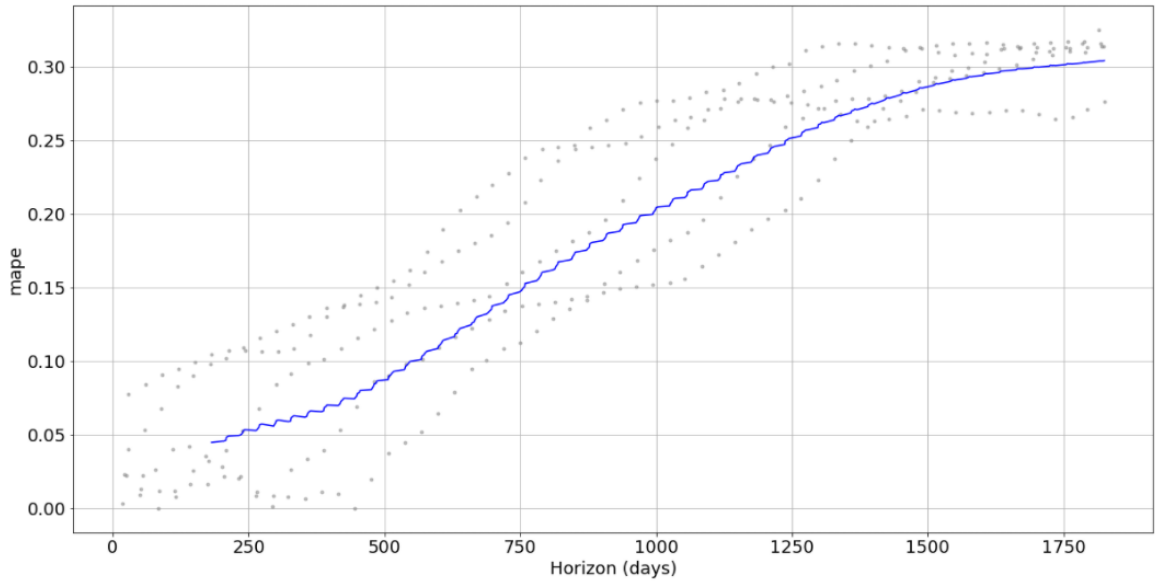


----------Analysis of MAPE (Mean Average Percent Error)----------

MAPE (Mean Average Percent Error) - Observation:

1. We see that MAPE increases over time
2. I am willing to tolerate MAPE of 0.1 to 0.2
    - This gets exceeded after about 1000 days
3. We will focus on MAPE as our main diagnostic metric.
    - Shows the model was about 80% accurate at 1000 days
    - Bullish prediction for the next 2-3 years
    - Supports the high upward trend we saw in the graph of all the data points for the zip code

```
In [15]: fig = plot_cross_validation_metric(cv_results, metric='mape', figsize=(20,10))
```



Overall Observations and Recommendations:

I believe that these five zip codes will be a great starting point. The forecasts of the top 5 zip codes show that there is a strong upward trend in property values. The U.S. Census data also shows that people are moving in and the state as a whole is still growing in numbers. We can supply that demand

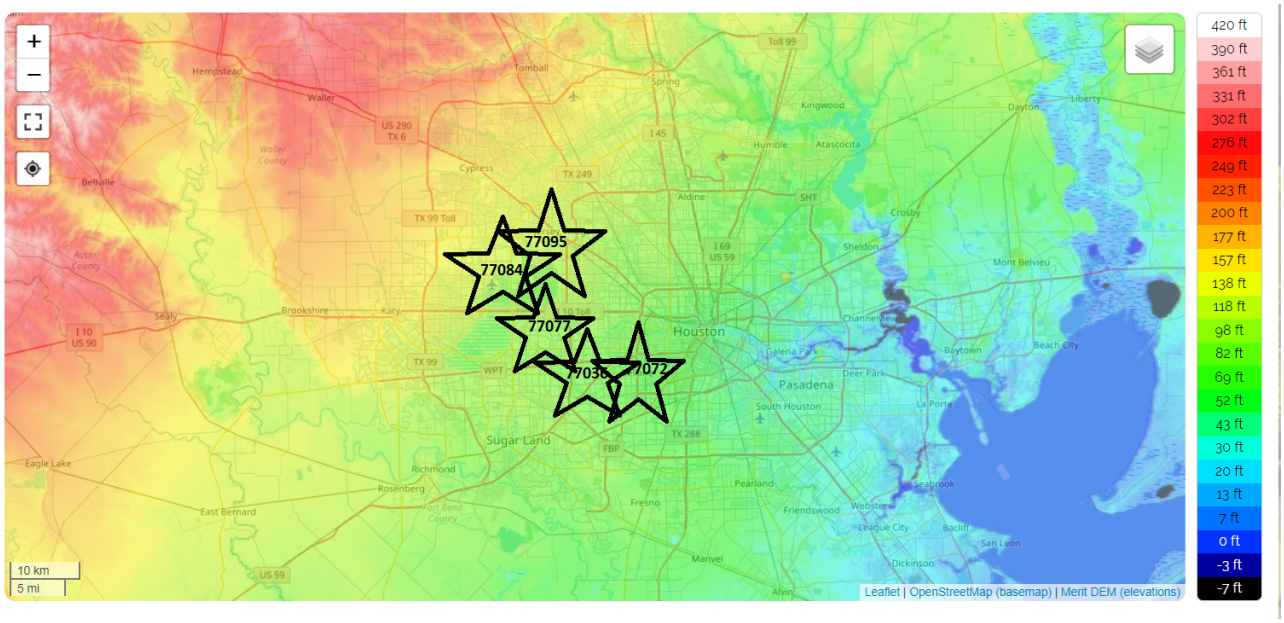
Additional observations

1. Houston has flood zones. However, the choices that we have are in slightly higher elevation areas. I've provided the areas below on a topographic map.

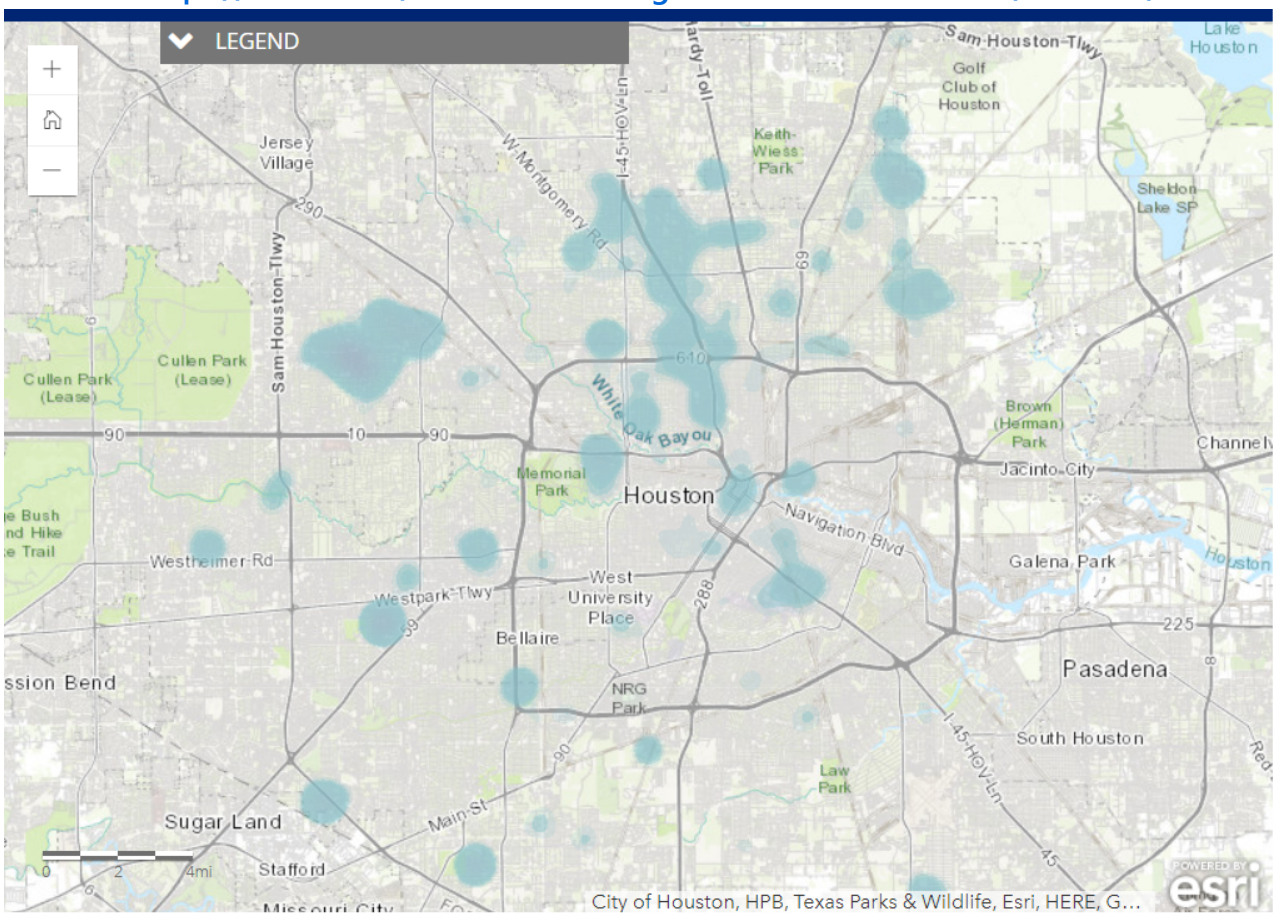
2. There is also a flood map of 311 calls from ABC news which talks about where specifically people were calling from saying their homes are flooded. These can help further narrow our choices down to the house/block level.

3. Texas shows up on lists as a land-lord friendly state. Example

SOURCE: https://en-gb.topographic-map.com/maps/fbcl/Houston/



SOURCE: https://abc13.com/houston-flooding-where-does-it-flood-in/5683641/



Future Work

1. Explore/Use Crime Data from the federal government **Link Here**
2. Zillow Word Cloud on the MLS database **Link Here**
    - This can be useful for finding house patterns like how many bedrooms and bathrooms
3. Utilize different time series tools other than Fbprophet
    - Limitations with tool for monthly data
4. Obtain Daily instead of monthly dataset of home values from same time range and re-apply analysis

Releases

No releases published
Create a new release

Packages

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%