**Housing price prediction**

Sohan Thakur

Bellevue University

**Business Problem**

Predict the median price of home in Arizona in next 12 months based on the historic home price and other pointers available for last 10years.

**Background/history**

With the real estate market drifting toward a more normalized pace, many industry experts envision the number of existing homes for sale increasing next year, especially as current homeowners look to move. It appears to be a risky investment now to either buy a new house or go for a rental property next year. According to Fannie Mae, we will still see an almost 50 percent shortage of homes available to meet a normal demand of buyers. The national average interest rate will likely stay somewhere around 3.25% for 2022. There is not enough stability in the market to sustain a large increase in interest rates.

In my opinion home prices to continue to rise, primarily due to limited supply. However, price increases will moderate next year to accommodate household affordability.

Thus, my goal is to predict the median house price in next 3 months with very high accuracy and eventually for 12 months with relatively less accuracy (as there would be more changes in Q1 2022 that will impact housing price beyond Q1 2022).

**Data Explanation (Data Prep/Data Dictionary)**

With the real estate market drifting toward a more normalized pace, many industry experts envision the number of existing homes for sale increasing next year, especially as current homeowners look to move.

Redfin provide earliest and most reliable data on the state of the housing market. They publish existing industry data faster, and offer additional data on tours and offers that no one else has.

Below are the attributes I am getting from Redfin housing data:

|  |  |
| --- | --- |
| **Attribute name** | **Description** |
| period\_begin | Month start date |
| period\_end | Month End date |
| period\_duration | Period duration |
| region\_type | Is region a state or city etc. |
| is\_seasonally\_adjusted | is\_seasonally\_adjusted |
| region | Region name |
| city | City name |
| state | State name |
| state\_code | State code |
| property\_type | Type of Property |
| median\_sale\_price | Median sale Price |
| median\_sale\_price\_mom | Median sale Price month over month |
| median\_sale\_price\_yoy | Median sale Price year over year |
| median\_list\_price | Median list price |
| median\_list\_price\_mom | Median list price Month over month |
| median\_list\_price\_yoy | Median list price year over year |
| median\_ppsf | Median price per squre foot |
| median\_ppsf\_mom | Median price per squre foot month over month |
| median\_ppsf\_yoy | Median price per squre foot year over year |
| median\_list\_ppsf | Medan list price per square foot |
| median\_list\_ppsf\_mom | Medan list price per square foot month over month |
| median\_list\_ppsf\_yoy | Medan list price per square foot year over year |
| homes\_sold | Number of home sold |
| homes\_sold\_mom | Number of home sold month over month |
| homes\_sold\_yoy | Number of home sold year over year |
| pending\_sales | Pending sales |
| pending\_sales\_mom | Pending sales month over month |
| pending\_sales\_yoy | Pending sales year over year |
| new\_listings | number of new listing |
| new\_listings\_mom | number of new listing month over month |
| new\_listings\_yoy | number of new listing year over year |
| inventory | number of inventory |
| inventory\_mom | number of inventory month over month |
| inventory\_yoy | number of inventoryyear over year |
| months\_of\_supply | Month of supply |
| months\_of\_supply\_mom | Change in number of months in supply month over month |
| months\_of\_supply\_yoy | Change in number of months in supply year over year |
| median\_dom | median days of month in market |
| median\_dom\_mom | Median days of month in market month over month |
| median\_dom\_yoy | Median days of month in market year over year |
| avg\_sale\_to\_list | average sales to list price |
| avg\_sale\_to\_list\_mom | Change in average sales price to list proce month over month |
| avg\_sale\_to\_list\_yoy | Change in average sales price to list proce year over year |
| sold\_above\_list | Fraction of houses sold over list price |
| sold\_above\_list\_mom | Change in fraction of house sold over list price month over month |
| sold\_above\_list\_yoy | Change in fraction of house sold over list price year over year |
| price\_drops | Price drop |
| price\_drops\_mom | Change in Price drop month over month |
| price\_drops\_yoy | Change in Price drop year over year |
| off\_market\_in\_two\_weeks | Fraction of house moving off market in two weeks |
| off\_market\_in\_two\_weeks\_mom | Change in fraction of house moving off market in two weeks month over month |
| off\_market\_in\_two\_weeks\_yoy | Change in fraction of house moving off market in two weeks year over year |
| parent\_metro\_region | Parent metro region |
| parent\_metro\_region\_metro\_code | Parent metro region metro code |
| last\_updated | Last update tiemstamp |

**Data Preparation:**

Below are the specific data operation activities that has been carried out on top of Redfin housing data:

1. Removed unwanted columns like surrogate ids.
2. Removed redundant columns like City name and city code. Just kept one out of these two.
3. Checked the quality of data to ensure there are no missing/null values.
4. Checked the data for skewness and biasing for any particular value.
5. The methods that I used in this use case is Facebook prophet and ARIMA(Auto Regressor Integrated Moving Average). We just need the time and the value attribute for these models, hence I removed rest other attributes.
6. For Facebook prophet, created a dataset with date and median house price. I picked the month start date as the date for that month.
7. For ARIMA, we need a series with values. I indexed the date column to keep a track of which value is corresponding to what date.

**Methods**

To begin with I will be performing data cleansing and variable transformation. Post that, below methods will be used:

1. Facebook prophet
2. ARIMA

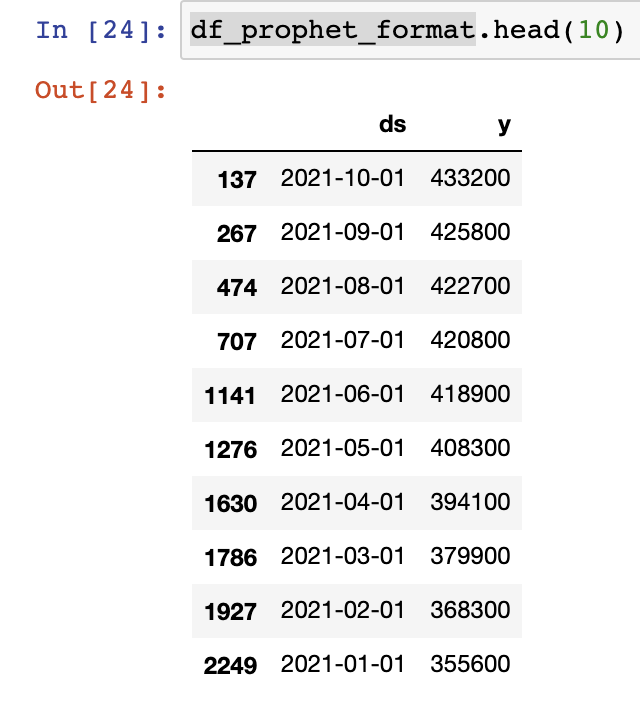
Let’s see the result under each one of the methods:

**Facebook prophet**

Prophet follows the sklearn model API. It creates an instance of the Prophet class and then call its fit and predict methods.

The input to Prophet is always a dataframe with two columns: ds and y. The ds (datestamp) column should be of a format expected by Pandas, ideally YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp. The y column must be numeric, and represents the measurement we wish to forecast.

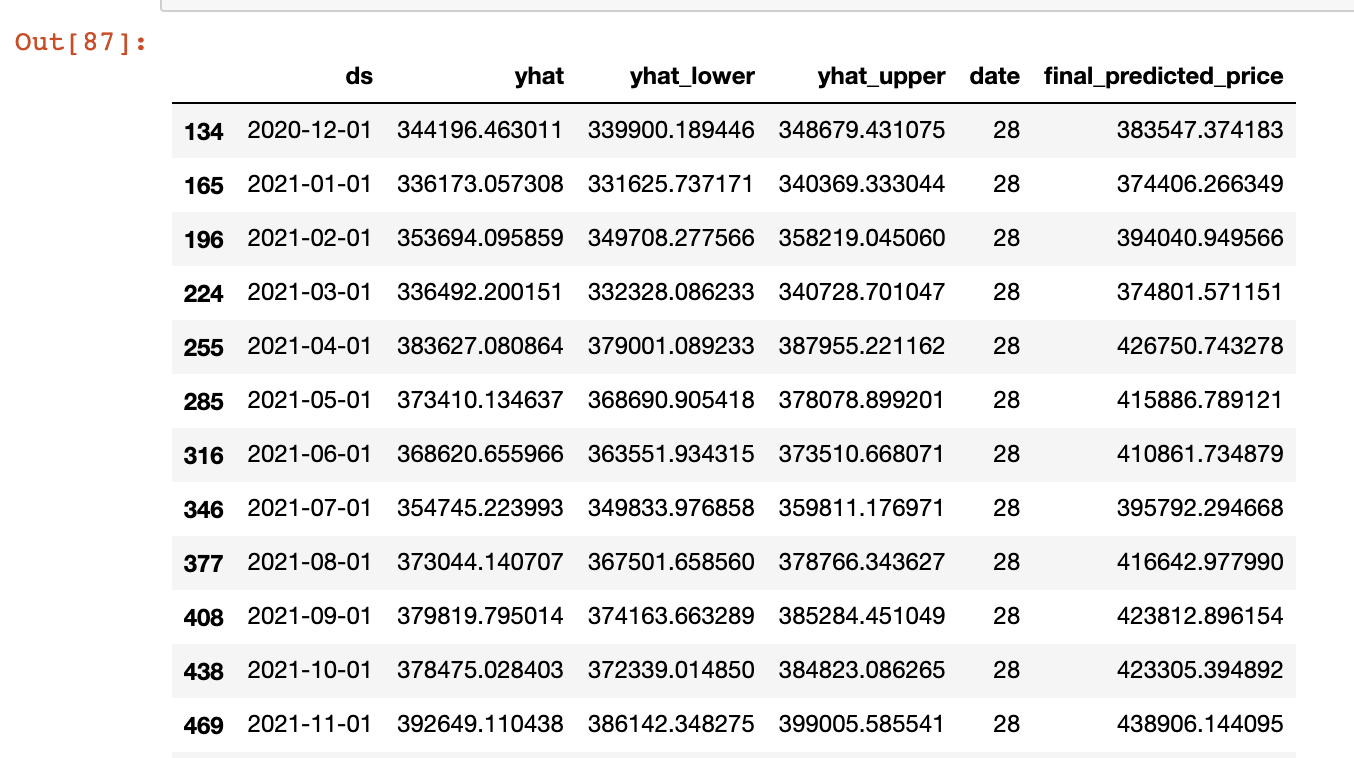
**Source data sample:**



**Prediction adjustments:**

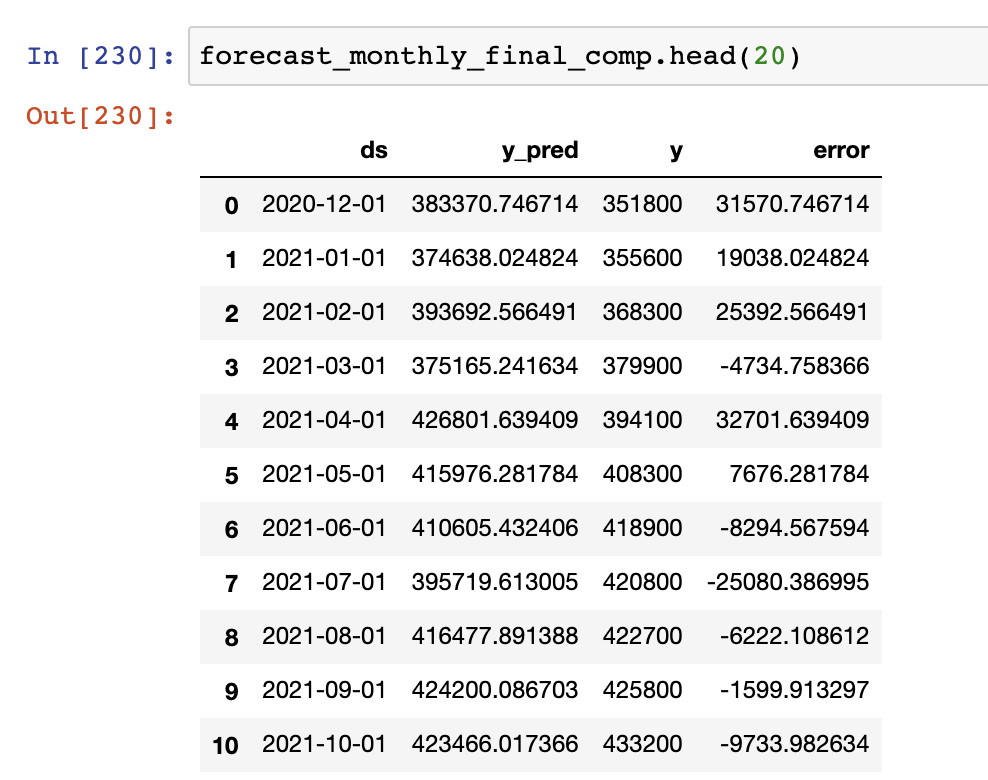
When I performed prediction, I found that the prediction is fine based on historic data. However there has been a unexpected increase in the house price in 2021, thus it was falling short mostly for all months consistently.

In order to absorbed the impact of this jump, I decided to use a multiplier of 1.1 to the predicted price.



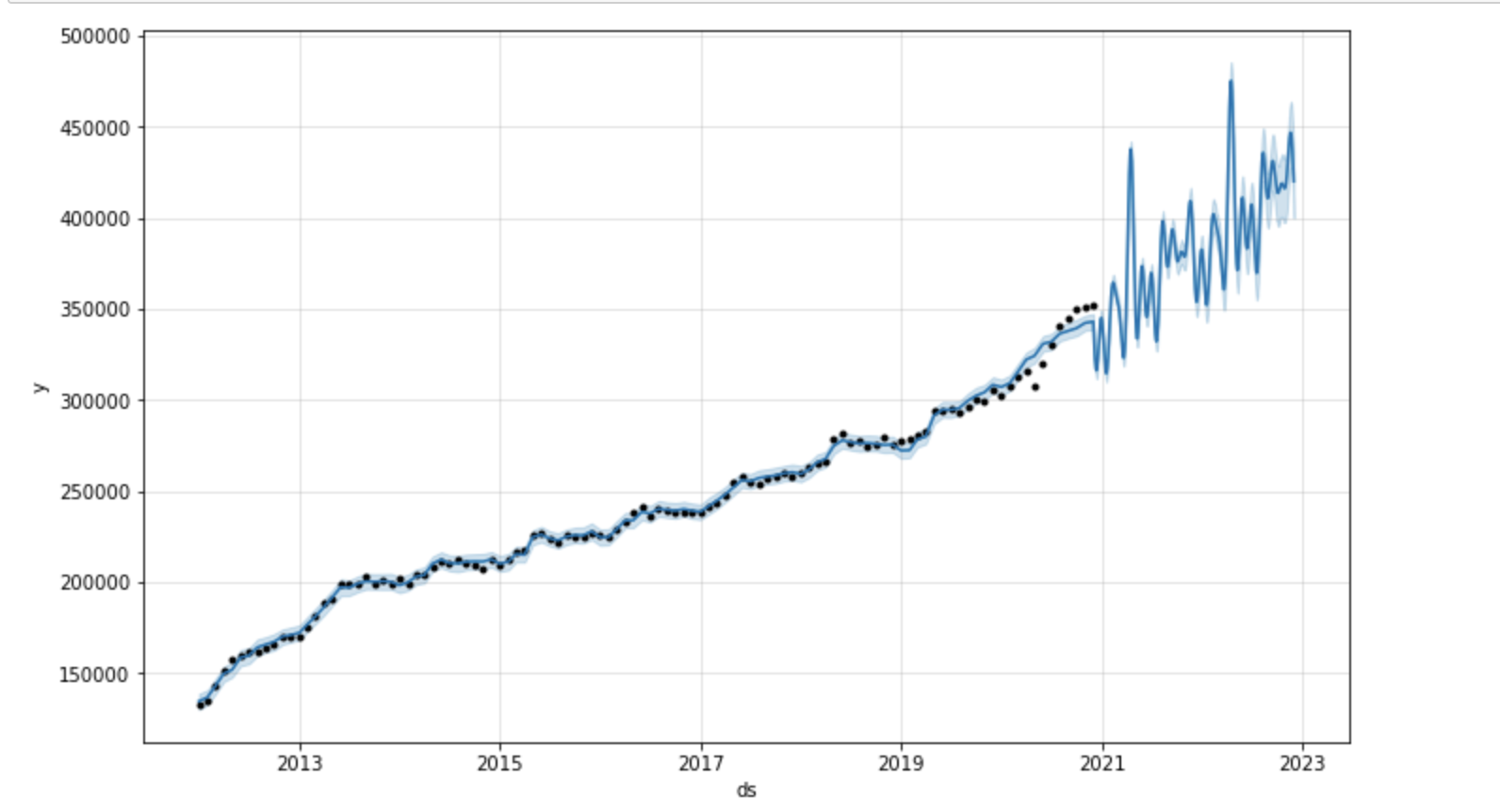
**Model Accuracy:**

I used data before 2021 as the test data and predicted 2021 all month’s median home price to check the model accuracy. Below are the stats for standard error:

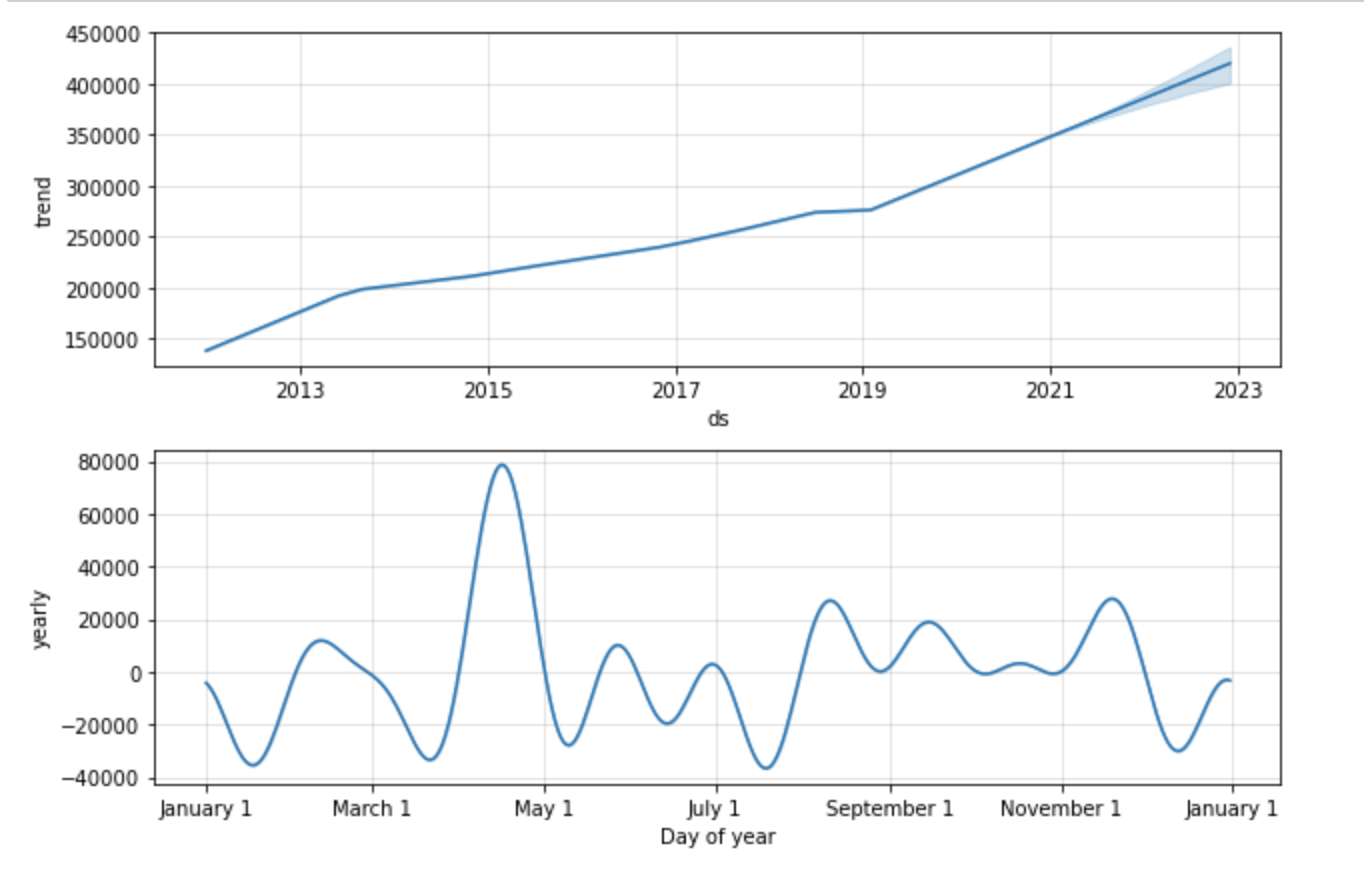


**Actual Vs the forecast data:**

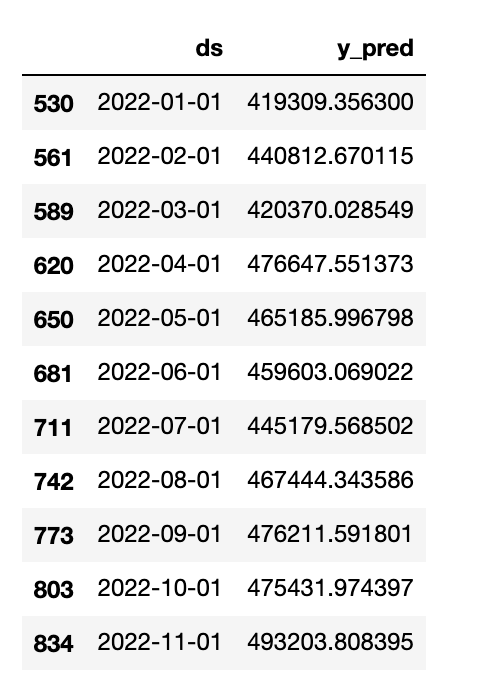
**Overall forecast value:**



**Components of forecast(trend and seasonality):**

****

**FINAL PREDICTION FOR 2022 HOUSING PRICE USING FACEBOOK PROPHET**

****

**ARIMA**

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

The parameters of the ARIMA model are defined as follows:

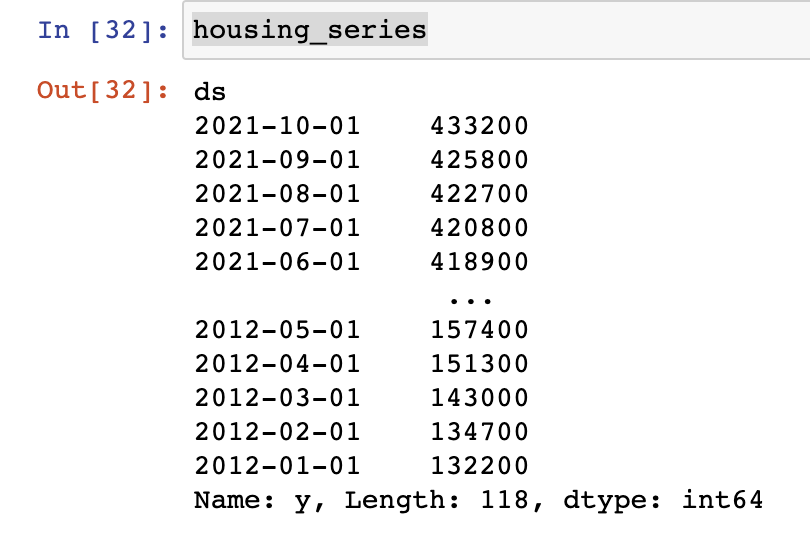
p: The number of lag observations included in the model, also called the lag order.

d: The number of times that the raw observations are differenced, also called the degree of differencing.

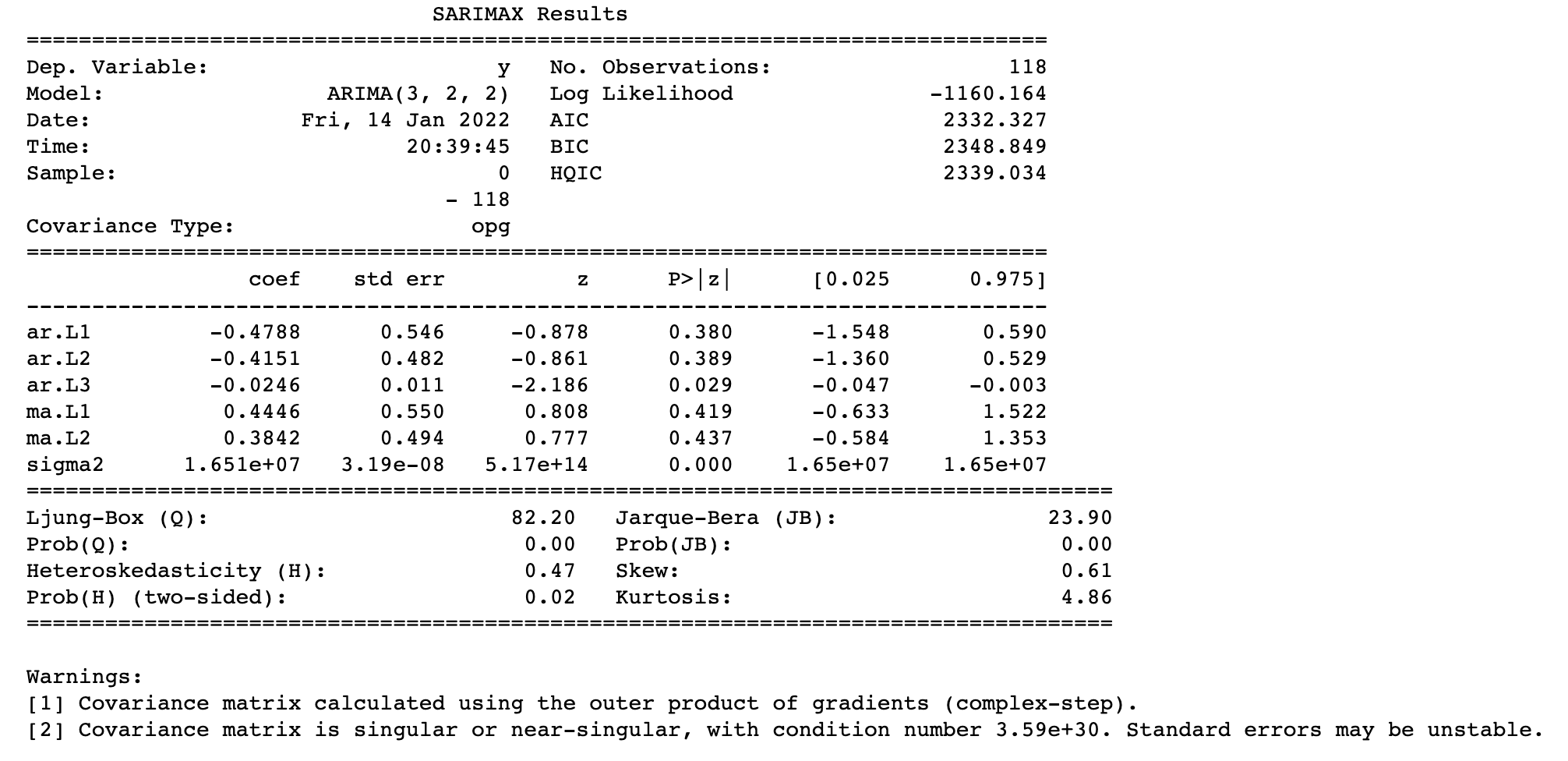
q: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model.

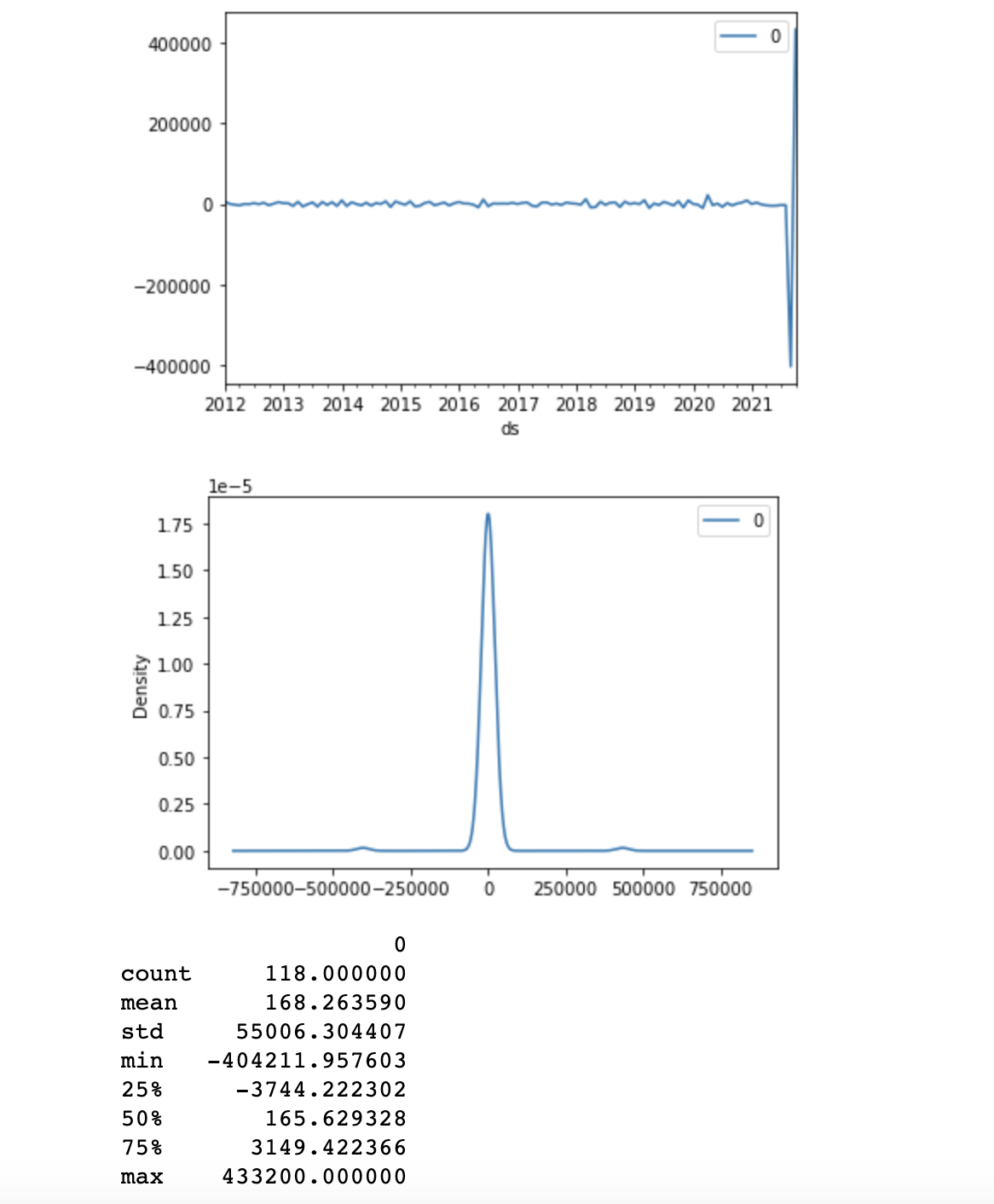
**Source data (converted into a series):**

****

**Model Summary:**

****

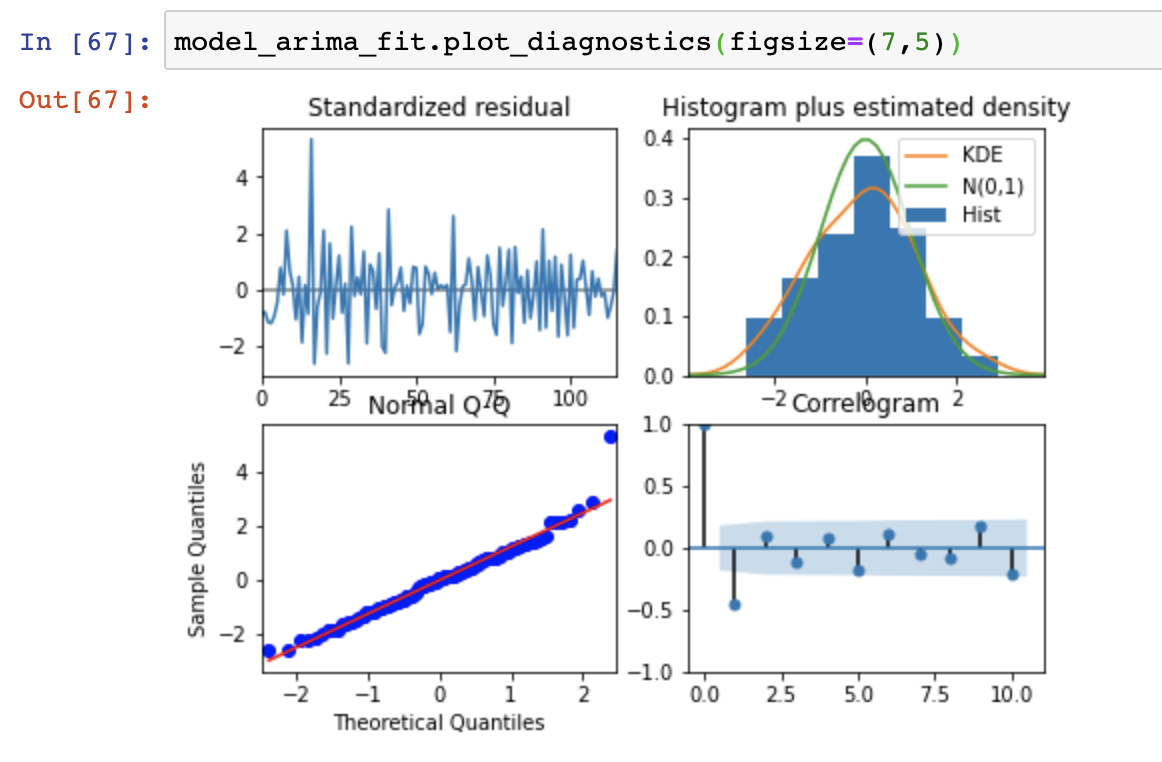
**Residual Error:**

****

This is huge



**Model diagnostic:**

****

**Observations:**

**Standardized residual** : The residual errors seem to fluctuate around a mean of zero and have a SIGNIFICANT VARIATION.

**Estimated density:** The density plot suggest normal distribution with mean zero.

**Normal Q Q plot**: All the dots should fall perfectly in line with the red line. But there are significant deviations hence the prediction is highly skewed.

**Correlogram**: The ACF plot shows the residual errors are autocorrelated which means that there is some pattern in the residual errors which are not explained in the model. So you will need to look for more X’s (predictors) to the model.

Given the unexplained residual error and variance in the prediction , **I have decided not to move ahead with this ARIMA.**

**Conclusion:**

In progress, nothing substantial to share.

**Assumptions:**

Below are the assumptions I am making:

1. Feds wouldn’t change the rates in next quarter as that may have huge impact on demand/supply paradigm.
2. Covid 19 impact will not vanish all of a sudden. It will keep impacting corporate work requirement.

**Limitations:**

Below are the limitations:

1. Covid 19 has impacted the data adversely in last 2 years. Thus, there would be a different trend in last 8years vs just last 2 years. This will impact the model tuning and is inevitable.
2. Fed regulations and rate cut etc. related data is not being used however they may have significant impact on house price.
3. Market trend and supply chain shortages are not being accounted in current analysis.

**Challenges:**

Below are the immediate challenges:

1. New variations of COVID 19 may impact Q1 house price and can totally make the model unusable based on what we know so far.
2. Geo-political events between US-Russia , US-China may cause wide spread economic impact and hence house price may have a solid effect but we don’t have enough data in current setup to account for these events.

**Future uses/Additional Application:**

This model will be a basic and baseline application for predicting median house price in Arizona. This model can be used for:

1. Predicting house price for any state/city of US.
2. Can be tuned to predict the price for subsequent years post 2022.
3. Additional data points can be accumulated and we can tune this model to increase accuracy.

**Recommendations:**

**Based on my work on this use case, I can recommend following:**

1. Facebook prophet worked well for only one state, i.e. Arizona and one house type, i.e. Single family house. When applied for other state and other house type it may not work as well as it did for this use case.
2. Recent surge in house price has challenged the regular model prediction as a result I came up with an additional multiplier. This multiplier has been chooses based upon my experience with sample data. We need to perform further analysis and decide on a more accurate multiplier.
3. ARIMA didn’t perform well for Arizona and Single family home. But that doesn’t mean it can’t work well for full housing data, we need to test that before ruling ARIMA out.

**Ethical Assessment:**

Below are the ethical safeguards I am imposing on my work for this model:

1. No race/gender/color of skin related data would be used to ensure the model is generic and doesn’t produce results based on pre-assumed notions.
2. Data used must originate from the source who adopt fair housing practices as described by state and federal government’s housing policies.

**References:**

<https://www.redfin.com/news/data-center/>

<https://www.redfin.com/news/housing-market-news/>

<https://medium.com/@josemarcialportilla/using-python-and-auto-arima-to-forecast-seasonal-time-series-90877adff03c>