

```
In [1]: import pandas as pd
from openpyxl.workbook import Workbook

#reading excel file
df = pd.read_excel('C:/Users/svasudev/OneDrive - IGMFinancial/Documents/Predictive Analyti
engine='openpyxl', sheet_name=0, header=0, index_col=False, keep_default_na=
```

```
In [2]: #Top 5 rows of the Excel file
df.head()
```

Out[2]:

	Date	Ottawa	Toronto West/Ouest	Toronto East/Est	Windsor	London	Peterborough	St. Catharine's	Sudbury	Sault Saint Marie	Thunder Bay	N
0	1990-01-03	55.9	49.1	48.7	45.2	50.1	0.0	0.0	56.4	54.8	56.6	
1	1990-01-10	55.9	47.7	46.8	49.7	47.6	0.0	0.0	56.4	54.9	56.8	
2	1990-01-17	55.9	53.2	53.2	49.6	53.7	0.0	0.0	55.8	54.9	56.8	
3	1990-01-24	55.9	53.2	53.5	49.0	52.1	0.0	0.0	55.7	54.9	56.8	
4	1990-01-31	55.9	51.9	52.6	48.6	49.1	0.0	0.0	55.6	54.8	56.8	

```
In [3]: #Datatype and Null Information about the columns in the Excel file
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9962 entries, 0 to 9961
Data columns (total 20 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Date                                     9962 non-null   datetime64[ns]
1   Ottawa                                 9962 non-null   float64
2   Toronto West/Ouest                     9962 non-null   float64
3   Toronto East/Est                       9962 non-null   float64
4   Windsor                                9962 non-null   float64
5   London                                 9962 non-null   float64
6   Peterborough                           9962 non-null   float64
7   St. Catharine's                        9962 non-null   float64
8   Sudbury                                9962 non-null   float64
9   Sault Saint Marie                      9962 non-null   float64
10  Thunder Bay                             9962 non-null   float64
11  North Bay                              9962 non-null   float64
12  Timmins                                9962 non-null   float64
13  Kenora                                 9962 non-null   float64
14  Parry Sound                            9962 non-null   float64
15  Ontario Average/Moyenne provinciale    9962 non-null   float64
16  Southern Average/Moyenne du sud de l'Ontario  9962 non-null   float64
17  Northern Average/Moyenne du nord de l'Ontario 9962 non-null   float64
```

```
18 Fuel Type          9962 non-null    object
19 Type de carburant   9962 non-null    object
dtypes: datetime64[ns](1), float64(17), object(2)
memory usage: 1.5+ MB
```

```
In [4]: #Creating a new dataframe
df2 = pd.DataFrame()
```

```
In [5]: #Choosing rows from the original dataframe with Fuel Type "Premium Gasoline" and pasting it
df2 = df.loc[df['Fuel Type']=='Premium Gasoline']
```

```
In [6]: #Datatype and Null Information about the columns in the new dataframe
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1675 entries, 6612 to 8286
Data columns (total 20 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Date                                     1675 non-null   datetime64[ns]
1   Ottawa                                 1675 non-null   float64
2   Toronto West/Ouest                    1675 non-null   float64
3   Toronto East/Est                     1675 non-null   float64
4   Windsor                               1675 non-null   float64
5   London                                1675 non-null   float64
6   Peterborough                         1675 non-null   float64
7   St. Catharine's                      1675 non-null   float64
8   Sudbury                              1675 non-null   float64
9   Sault Saint Marie                    1675 non-null   float64
10  Thunder Bay                          1675 non-null   float64
11  North Bay                            1675 non-null   float64
12  Timmins                              1675 non-null   float64
13  Kenora                               1675 non-null   float64
14  Parry Sound                          1675 non-null   float64
15  Ontario Average/Moyenne provinciale  1675 non-null   float64
16  Southern Average/Moyenne du sud de l'Ontario  1675 non-null   float64
17  Northern Average/Moyenne du nord de l'Ontario  1675 non-null   float64
18  Fuel Type                             1675 non-null   object
19  Type de carburant                     1675 non-null   object
dtypes: datetime64[ns](1), float64(17), object(2)
memory usage: 274.8+ KB
```

```
In [7]: df2.head()
```

```
Out[7]:
```

	Date	Ottawa	Toronto West/Ouest	Toronto East/Est	Windsor	London	Peterborough	St. Catharine's	Sudbury	Sault Saint Marie	Thunder Bay
6612	1990-01-03	59.5	52.9	52.7	49.2	54.0	0.0	0.0	60.2	58.6	60.4
6613	1990-01-10	59.6	51.6	50.9	53.7	51.5	0.0	0.0	60.2	58.7	60.5
6614	1990-01-17	59.7	57.4	57.4	53.6	57.6	0.0	0.0	59.6	58.6	60.6
6615	1990-01-24	59.8	57.1	57.5	53.1	56.0	0.0	0.0	59.5	58.6	60.5

	Date	Ottawa	Toronto West/Ouest	Toronto East/Est	Windsor	London	Peterborough	St. Catharine's	Sudbury	Sault Saint Marie	Thunder Bay
6616	1990-01-31	59.8	55.9	56.6	52.6	53.1	0.0	0.0	59.5	58.6	60.6

```
In [8]: #Creating a new dataframe
df3 = pd.DataFrame()
```

```
In [9]: #Analyzing 'Premium Gasoline' price for Sudbury Region only and pasting it into new dataframe
df3.insert(0, "Date", df2['Date'], True)
df3.insert(1, "Sudbury", df2['Sudbury'], True)
```

```
In [10]: df3.head()
```

```
Out[10]:
```

	Date	Sudbury
6612	1990-01-03	60.2
6613	1990-01-10	60.2
6614	1990-01-17	59.6
6615	1990-01-24	59.5
6616	1990-01-31	59.5

```
In [11]: #Date column only had weekly values. Sometimes the week started on Wednesday and sometimes
#this inconsistency, converting the date column into daily values and assigning the previous values to
#daily values.
df3.set_index('Date', inplace=True)
df3 = df3.resample('D').ffill().reset_index()
```

```
In [12]: #Last 10 values from the dataset
df3.tail(10)
```

```
Out[12]:
```

	Date	Sudbury
11707	2022-01-22	177.7
11708	2022-01-23	177.7
11709	2022-01-24	180.2
11710	2022-01-25	180.2
11711	2022-01-26	180.2
11712	2022-01-27	180.2
11713	2022-01-28	180.2
11714	2022-01-29	180.2
11715	2022-01-30	180.2
11716	2022-01-31	181.8

```
In [13]: #Checking for null values  
df3.isnull().values.any()
```

```
Out[13]: False
```

```
In [14]: #Splitting the "Date" column into year, month and week to explore trends  
df3['Year']=df3['Date'].dt.year  
df3['Month']=df3['Date'].dt.month  
df3['Week']=df3['Date'].dt.isocalendar().week
```

```
In [15]: df3.head(10)
```

```
Out[15]:
```

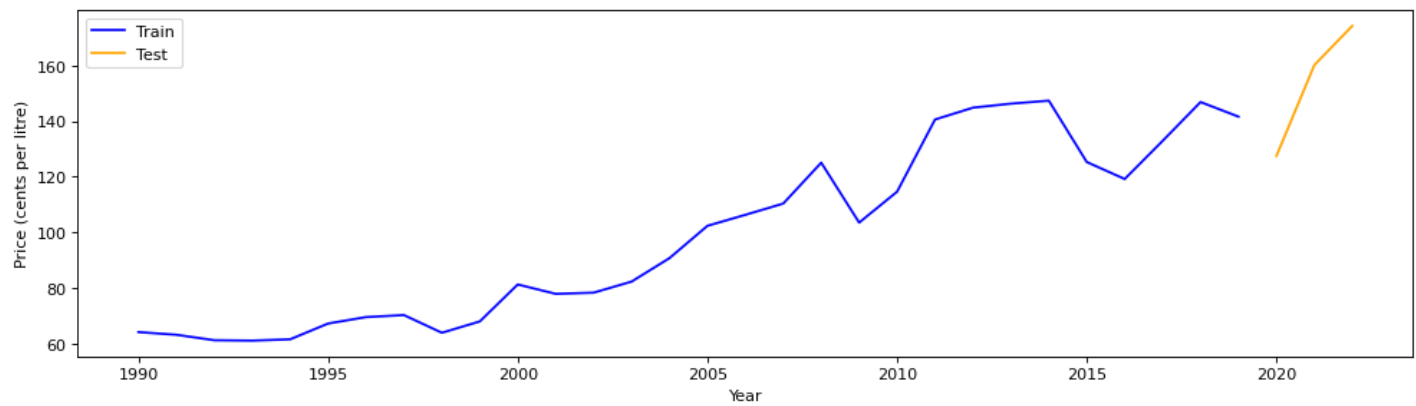
	Date	Sudbury	Year	Month	Week
0	1990-01-03	60.2	1990	1	1
1	1990-01-04	60.2	1990	1	1
2	1990-01-05	60.2	1990	1	1
3	1990-01-06	60.2	1990	1	1
4	1990-01-07	60.2	1990	1	1
5	1990-01-08	60.2	1990	1	2
6	1990-01-09	60.2	1990	1	2
7	1990-01-10	60.2	1990	1	2
8	1990-01-11	60.2	1990	1	2
9	1990-01-12	60.2	1990	1	2

```
In [16]: #Splitting the dataset in Train and Test  
#Train from Year 1990 to Year 2019  
#Test from Year 2020  
  
train = df3[(df3['Date'] > '1990-01-01') & (df3['Date'] <= '2019-12-31')]  
test = df3[df3['Date'] >= '2020-01-01']
```

Yearly Price Visualization on Train and Test Dataset

```
In [17]: import matplotlib.pyplot as plt  
from matplotlib.pyplot import figure  
  
yearly_train_Price = train.groupby(['Year'])['Sudbury'].mean()  
yearly_test_Price = test.groupby(['Year'])['Sudbury'].mean()  
  
figure(figsize=(15, 4), dpi=80)  
plt.plot(yearly_train_Price, label='Train',c='blue')  
plt.plot(yearly_test_Price, label='Test',c='orange')  
plt.legend(loc='best')  
plt.suptitle('Premium Gasoline Prices for Sudbury', fontsize=20)  
plt.xlabel('Year')  
plt.ylabel('Price (cents per litre)')  
plt.show()
```

Premium Gasoline Prices for Sudbury



DataPrep for Time Series

```
In [18]: train.index = pd.DatetimeIndex(train['Date'])
          #Changing the frequency of the index to Daily
          train.index = train.asfreq('d').index

          test.index = pd.DatetimeIndex(test['Date'])
          #Changing the frequency of the index to Daily
          test.index = test.asfreq('d').index
```

Train and Time Series Dataset

```
In [19]: train_time_series = pd.DataFrame()
          train_time_series.index = train.index
          train_time_series.insert(0,"Sudbury Gas Price Train",train['Sudbury'],True)

          test_time_series = pd.DataFrame()
          test_time_series.index = test.index
          test_time_series.insert(0,"Sudbury Gas Price Test",test['Sudbury'],True)

          print(train_time_series.tail())
          print(test_time_series.tail())
```

```
                Sudbury Gas Price Train
Date
2019-12-27                142.5
2019-12-28                142.5
2019-12-29                142.5
2019-12-30                138.5
2019-12-31                138.5
                Sudbury Gas Price Test
Date
2022-01-27                180.2
2022-01-28                180.2
2022-01-29                180.2
2022-01-30                180.2
2022-01-31                181.8
```

ARIMA Model

To predict time series with ARIMA, we need to set the values of three parameters (p,d,q):

p: The order of the auto-regressive (AR) model (i.e., the number of lag observations)

d: The degree of differencing.

q: The order of the moving average (MA) model.

Checking if data is stationary - We can see that it is not based on the P-value - Augmented Dickey Fuller Test

In [20]:

```
from statsmodels.tsa.stattools import adfuller
results = adfuller(train_time_series['Sudbury Gas Price Train'])
print('ADF Statistic: ', results[0])
print('p-value: ', results[1])
print('Critical Values', results[4])
```

ADF Statistic: -1.77598586093174

p-value: 0.392447522946673

Critical Values {'1%': -3.4309490326940666, '5%': -2.8618047392710215, '10%': -2.5669109155836605}

Taking First difference - P value is < 0.05. We can stop at the First Difference; d = 1

In [21]:

```
train_time_series_stationary1 = train_time_series.diff().dropna()
results1 = adfuller(train_time_series_stationary1['Sudbury Gas Price Train'])
print('ADF Statistic: ', results1[0])
print('p-value: ', results1[1])
print('Critical Values', results1[4])
```

ADF Statistic: -17.71190628042678

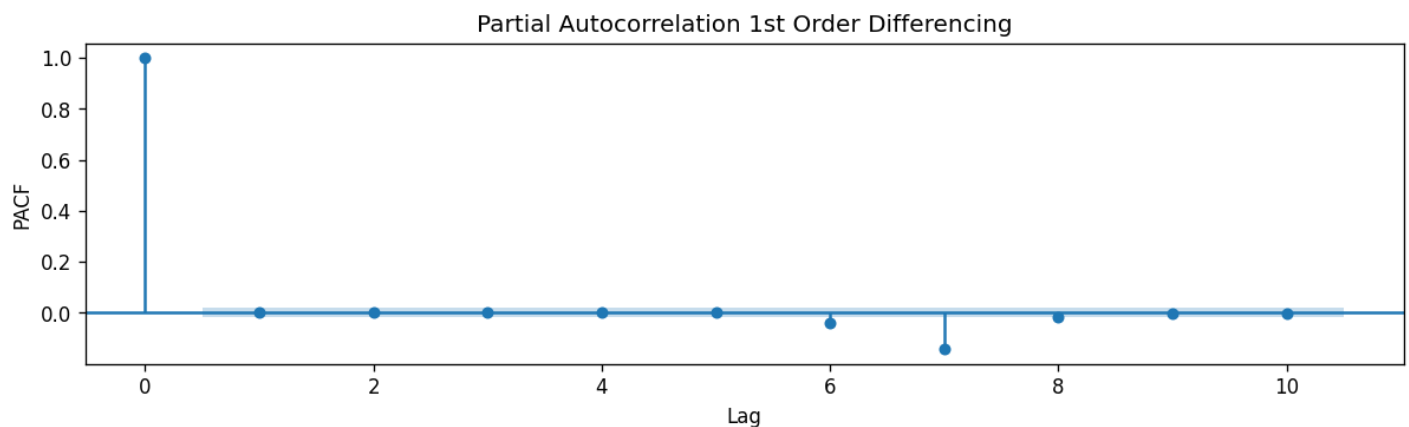
p-value: 3.4992234549468696e-30

Critical Values {'1%': -3.4309490326940666, '5%': -2.8618047392710215, '10%': -2.5669109155836605}

The Order of Autoregressive Term p; p = 0

In [22]:

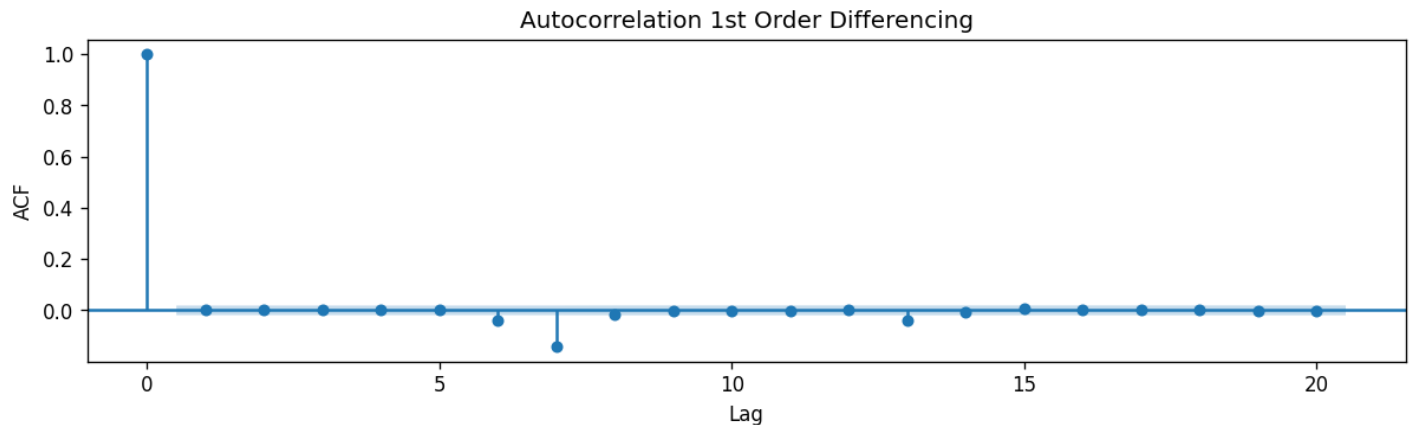
```
plt.rcParams.update({'figure.figsize': (12, 3), 'figure.dpi': 120})
from statsmodels.graphics.tsaplots import plot_pacf
plot_pacf(train_time_series_stationary1, lags=10, title="Partial Autocorrelation 1st Order Differencing")
plt.xlabel('Lag')
plt.ylabel('PACF')
plt.show()
```



The order of the Moving Average term q; q = 0

In [23]:

```
plt.rcParams.update({'figure.figsize':(12,3), 'figure.dpi':120})
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(train_time_series_stationary1, lags=20, title="Autocorrelation 1st Order Differenced")
plt.xlabel('Lag')
plt.ylabel('ACF')
plt.show()
```



Check the above p,d,q parameters with Auto Arima - Best Model ARIMA (0,1,0)

In [24]:

```
import pmdarima as pm
from pmdarima.model_selection import train_test_split
import numpy as np

model1 = pm.auto_arima(train_time_series, trace=True, error_action='ignore', suppress_warnings=True)
model1.fit(train_time_series)

forecast1 = model1.predict(n_periods=len(test_time_series))
forecast1 = pd.DataFrame(forecast1, index = test_time_series.index, columns=['Prediction'])
```

Performing stepwise search to minimize aic

```
ARIMA(2,1,2) (0,0,0) [0] intercept : AIC=37271.611, Time=1.49 sec
ARIMA(0,1,0) (0,0,0) [0] intercept : AIC=37263.611, Time=0.25 sec
ARIMA(1,1,0) (0,0,0) [0] intercept : AIC=37265.611, Time=0.64 sec
ARIMA(0,1,1) (0,0,0) [0] intercept : AIC=37265.611, Time=0.91 sec
ARIMA(0,1,0) (0,0,0) [0]           : AIC=37261.929, Time=0.12 sec
ARIMA(1,1,1) (0,0,0) [0] intercept : AIC=37267.611, Time=0.99 sec
```

Best model: ARIMA(0,1,0)(0,0,0)[0]

Total fit time: 4.442 seconds

Model Summary

In [25]:

```
import statsmodels.api as sm
model = sm.tsa.arima.ARIMA(train_time_series, order=(0,1,0))
model_result = model.fit()
print(model_result.summary())
```

SARIMAX Results

Dep. Variable:	Sudbury Gas Price Train	No. Observations:	10955
Model:	ARIMA(0, 1, 0)	Log Likelihood	-18629.965
Date:	Sat, 05 Mar 2022	AIC	37261.929
Time:	15:21:24	BIC	37269.231
Sample:	01-03-1990	HQIC	37264.389
	- 12-31-2019		
Covariance Type:	opg		

	coef	std err	z	P> z	[0.025	0.975]
sigma2	1.7571	0.004	452.332	0.000	1.749	1.765
=====						
Ljung-Box (L1) (Q):			0.00	Jarque-Bera (JB):		2423976.05
Prob(Q):			1.00	Prob(JB):		0.00
Heteroskedasticity (H):			2.55	Skew:		2.78
Prob(H) (two-sided):			0.00	Kurtosis:		75.66
=====						

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Model Prediction

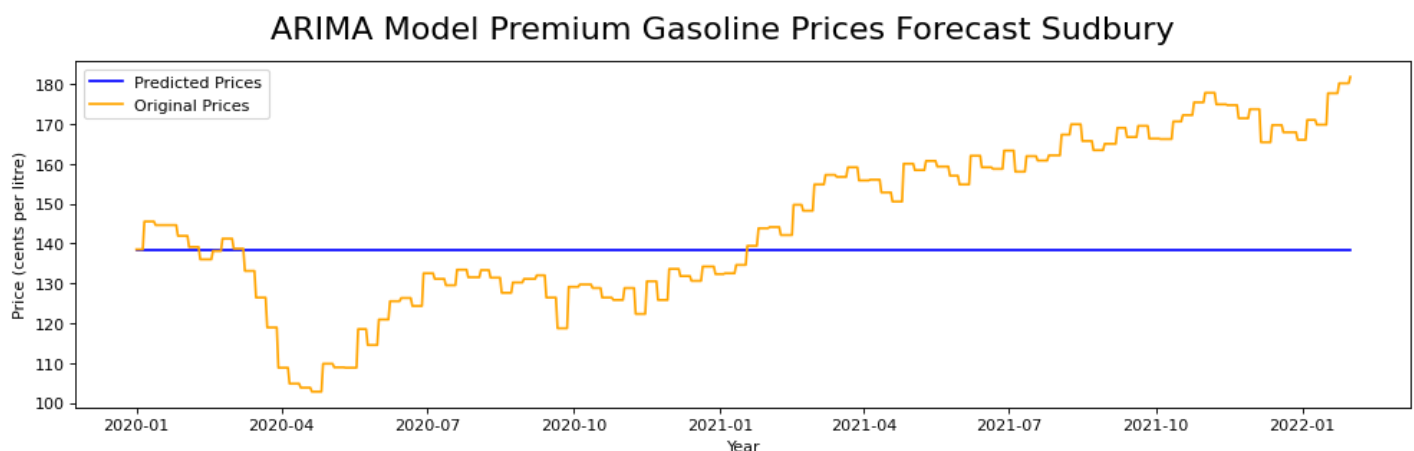
```
In [26]: import warnings
warnings.filterwarnings('ignore')
ARIMA_Predict = model_result.predict(start='1/1/2020', end='1/31/2022')
ARIMA_Predict_df = pd.DataFrame(ARIMA_Predict)
```

```
In [27]: ARIMA_Predict_df.tail()
```

```
Out[27]:
```

	predicted_mean
2022-01-27	138.5
2022-01-28	138.5
2022-01-29	138.5
2022-01-30	138.5
2022-01-31	138.5

```
In [28]: figure(figsize=(15, 4), dpi=80)
plt.plot(ARIMA_Predict_df, label='Predicted Prices',c='blue')
plt.plot(test_time_series, label='Original Prices',c='orange')
plt.legend(loc='best')
plt.suptitle('ARIMA Model Premium Gasoline Prices Forecast Sudbury', fontsize=20)
plt.xlabel('Year')
plt.ylabel('Price (cents per litre)')
plt.show()
```



Evaluation of the Model

Mean Absolute Error (MAE) ARIMA

```
In [29]: from sklearn.metrics import mean_absolute_error
maeARIMA=mean_absolute_error(test_time_series['Sudbury Gas Price Test'],ARIMA_Predict)
print('Mean Absolute Error ARIMA = {}'.format(round(maeARIMA, 2)))
```

Mean Absolute Error ARIMA = 17.82

Mean squared error (MSE) ARIMA

```
In [30]: from sklearn.metrics import mean_squared_error
mseARIMA=mean_squared_error(test_time_series['Sudbury Gas Price Test'],ARIMA_Predict)
print('The Mean Squared Error ARIMA = {}'.format(round(mseARIMA, 2)))
```

The Mean Squared Error ARIMA = 441.82

Root mean squared error (RMSE) ARIMA

```
In [31]: from numpy import sqrt
rmseARIMA = sqrt(mseARIMA)
print('The Root Mean Squared Error ARIMA = {}'.format(round(rmseARIMA, 2)))
```

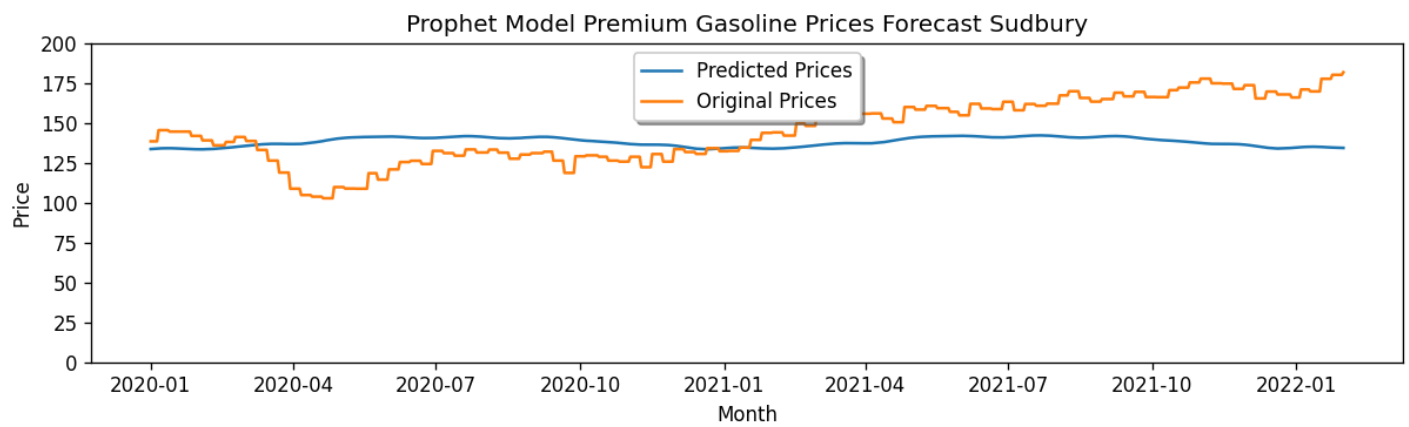
The Root Mean Squared Error ARIMA = 21.02

Prophet Model

```
In [32]: from fbprophet import Prophet
d={'ds':train['Date'],'y':train['Sudbury']}
df_pred=pd.DataFrame(data=d)
model_prophet = Prophet(daily_seasonality=False)
model_prophet_result = model_prophet.fit(df_pred)
```

```
In [33]: future = model_prophet.make_future_dataframe(periods=765)
forecast = model_prophet.predict(future)
forecast = forecast[(forecast['ds'] >= '2020-01-01') & (forecast['ds'] <= '2022-01-31')]
```

```
In [34]: fig, ax = plt.subplots()
ax.plot(forecast['ds'], forecast['yhat'], label='Predicted Prices')
ax.plot(test['Date'], test['Sudbury'], label='Original Prices')
plt.ylim([0,200])
legend = ax.legend(loc='upper center', shadow=True)
plt.title('Prophet Model Premium Gasoline Prices Forecast Sudbury')
plt.xlabel('Month')
plt.ylabel('Price')
plt.show()
```



Mean Absolute Error (MAE) Prophet

```
In [35]: maeProphet=mean_absolute_error(test['Sudbury'],forecast['yhat'])  
print('Mean Absolute Error Prophet = {}'.format(round(maeProphet, 2)))
```

Mean Absolute Error Prophet = 18.27

Mean squared error (MSE) Prophet

```
In [36]: mseProphet = mean_squared_error(test['Sudbury'],forecast['yhat'])  
print('The Mean Squared Error Prophet = {}'.format(round(mseProphet, 2)))
```

The Mean Squared Error Prophet = 458.82

Root mean squared error (RMSE) Prophet

```
In [37]: rmseProphet = sqrt(mseProphet)  
print('The Root Mean Squared Error Prophet = {}'.format(round(rmseProphet, 2)))
```

The Root Mean Squared Error Prophet = 21.42