Assessment

Submitted By:

Lamima Tabassum Ava

Bsc Engg in CSE

lamima7777@gmail.com

Dataset Description: The given dataset has two columns one column illustrate the date of cars monthly and other column depicts the number of registered cars.

```
print("There are {:,} rows and {} columns in the  Registered Car dataset set.".format(df.shape[0], df.shape[1]))
print("The time series starts on {} and ends on {}.\n".format(df.Time.min(), df.Time.max()))
print("There are {:,} rows and {} columns in the Custom Default data set.".format(df1.shape[0], df1.shape[1]))
```

```
There are 324 rows and 2 columns in the  Registered Car dataset set.
The time series starts on 1995M01 and ends on 2021m02.

There are 324 rows and 2 columns in the Custom Default data set.
```

Dataset Info:

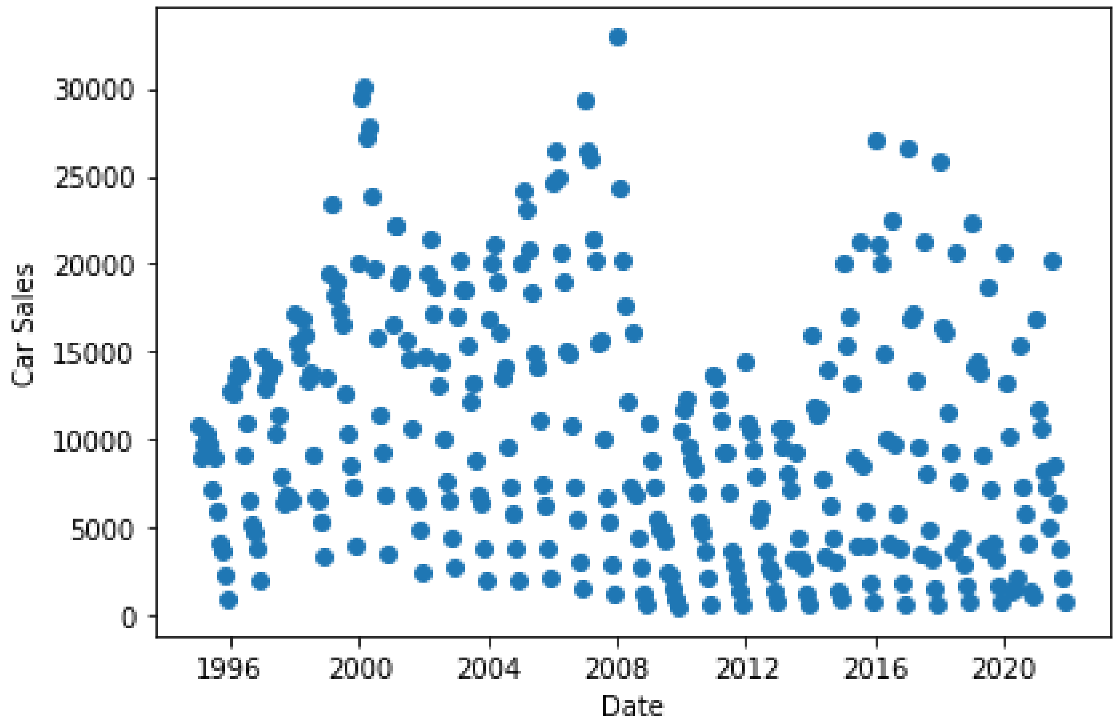
count	324.000000
mean	10494.172840
std	7180.582829
min	474.000000
25%	4313.500000
50%	9332.500000
75%	15369.750000
max	32961.000000

Summary statistics of categorical columns

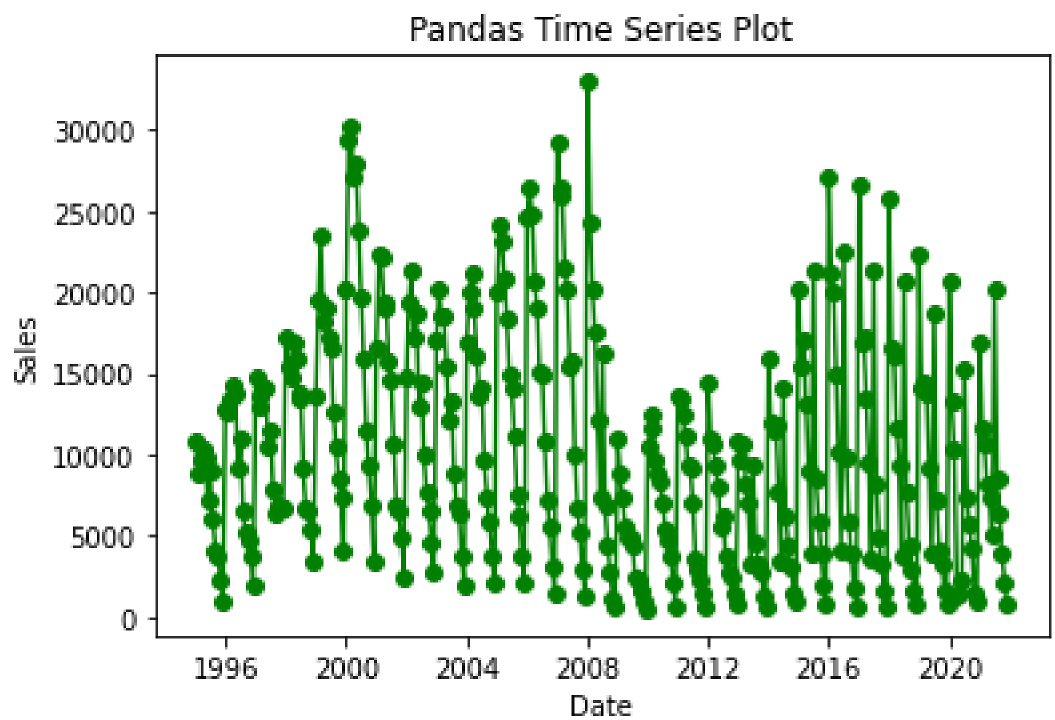
Direction	Number of Observations	Average Congestion
2008-01	1	32961.0
2000-03	1	30125.0
2000-02	1	29419.0
2007-01	1	29281.0
2000-05	1	27832.0
2000-04	1	27147.0
2016-01	1	27106.0
2017-01	1	26668.0
2007-02	1	26495.0
2006-02	1	26384.0
2007-03	1	25974.0
2018-01	1	25813.0
2006-03	1	24858.0

Visualization:

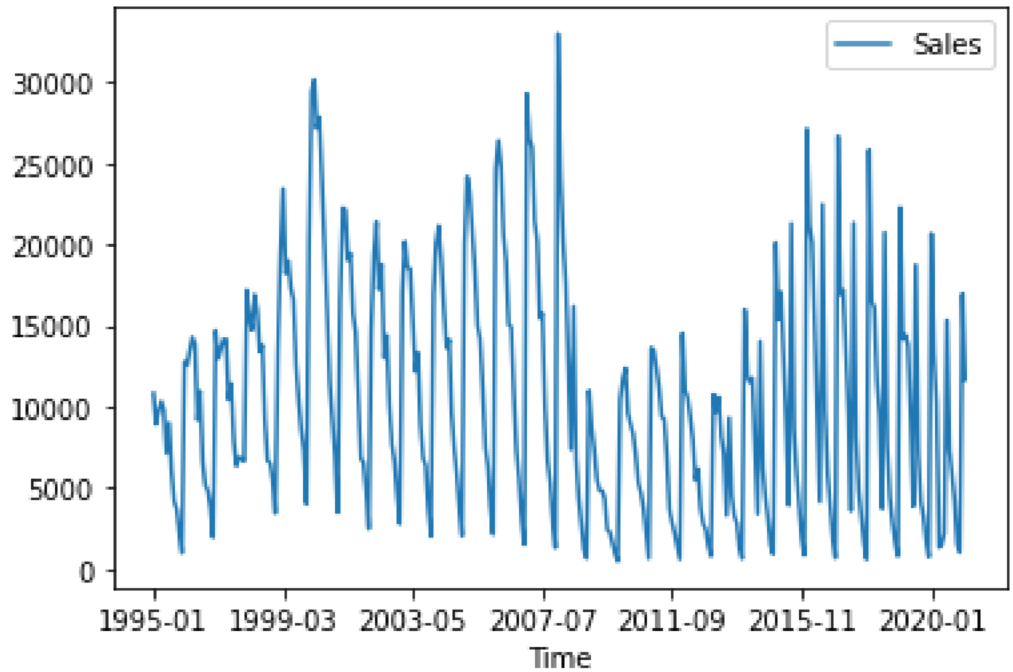
Scatter plot:



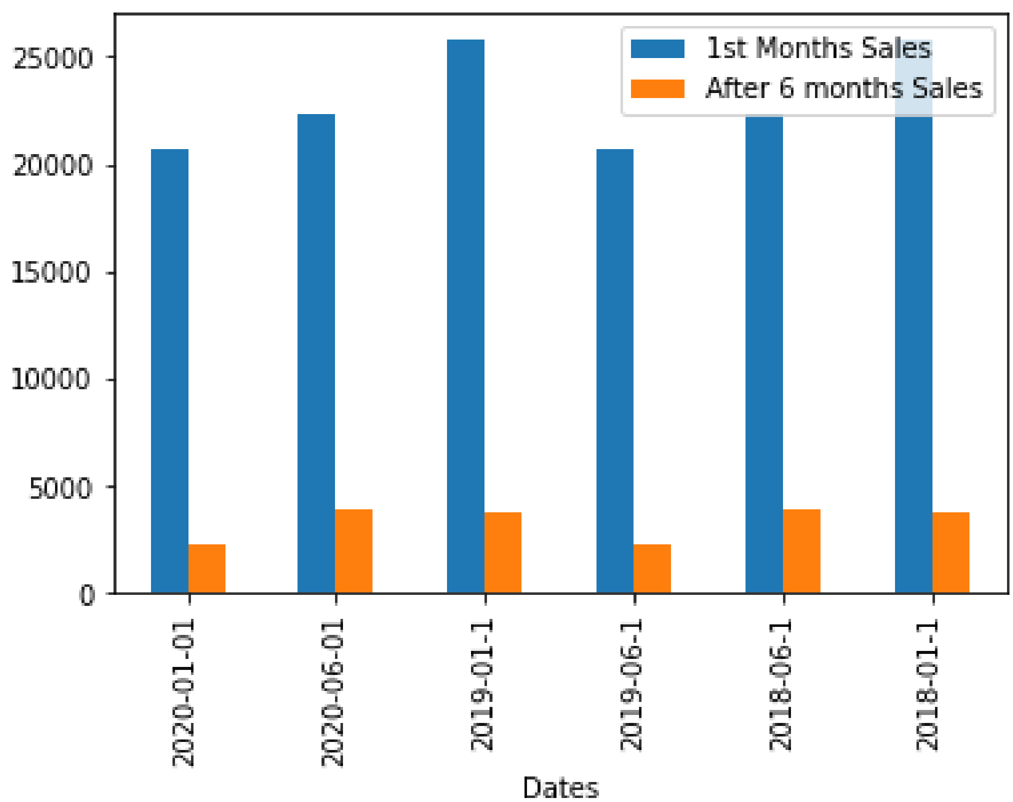
30000
25000
20000
15000
10000
5000
0
Car Sales
1996
2000
2004
2008
2012
2016
2020
Date

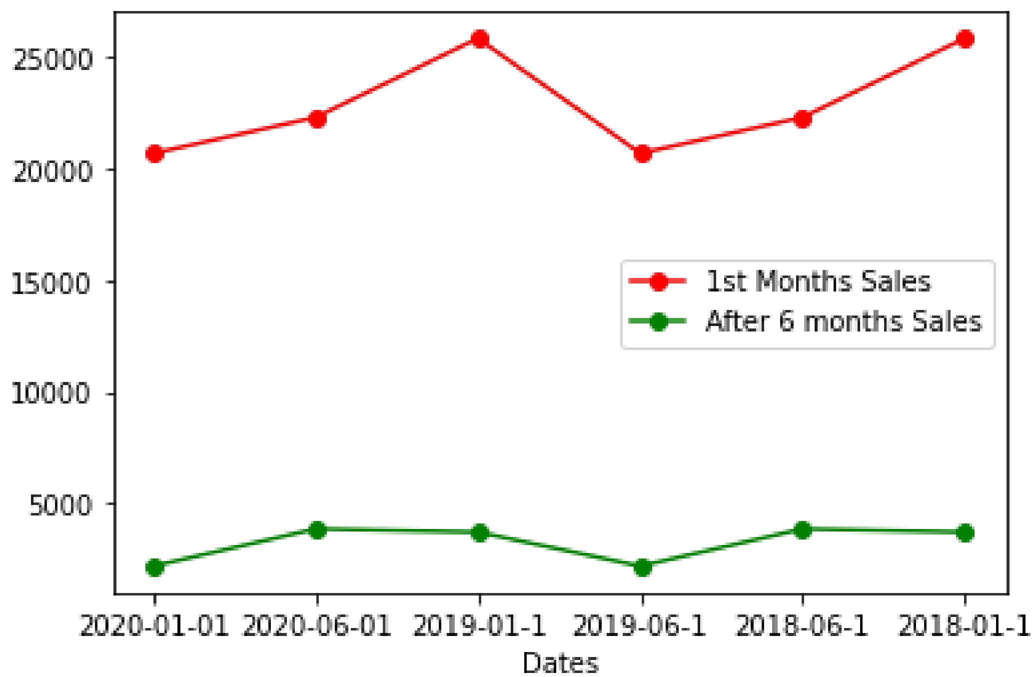


Pandas Time Series Plot
30000
25000
20000
15000
10000
5000
0
Sales
1996
2000
2004
2008
2012
2016
2020
Date



Comparing Sales 6 month Interval:





ETS MODEL:

The ETS models are a family of time series models with an underlying state space model consisting of a level component, a trend component (T), a seasonal component (S), and an error term (E). This state space formulation can be turned into a different formulation, a forecast and a smoothing equation (as can be done with all ETS models).

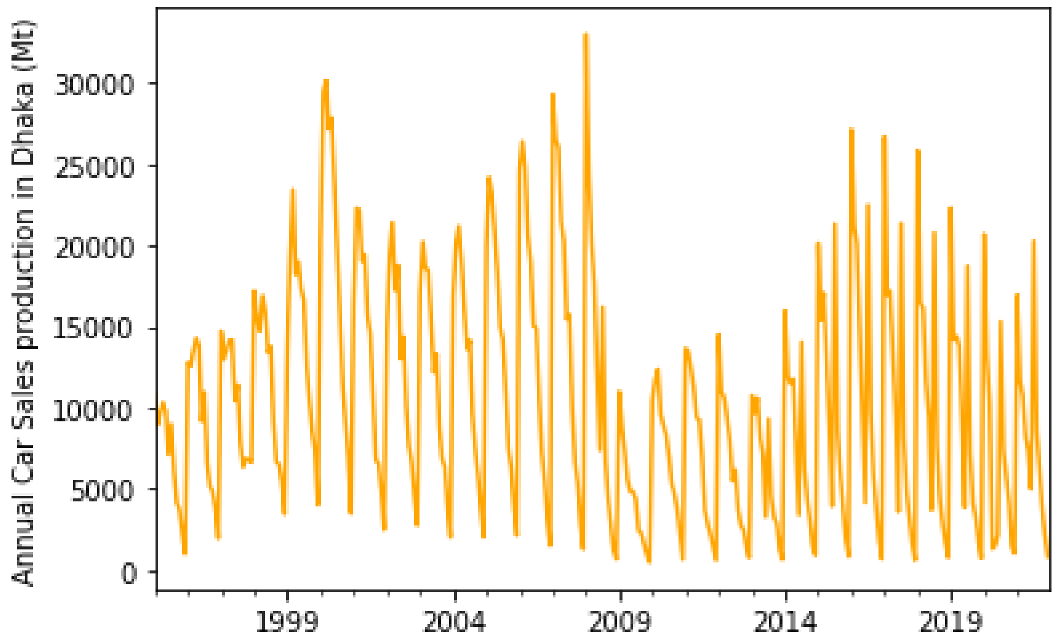


Figure: Visualization

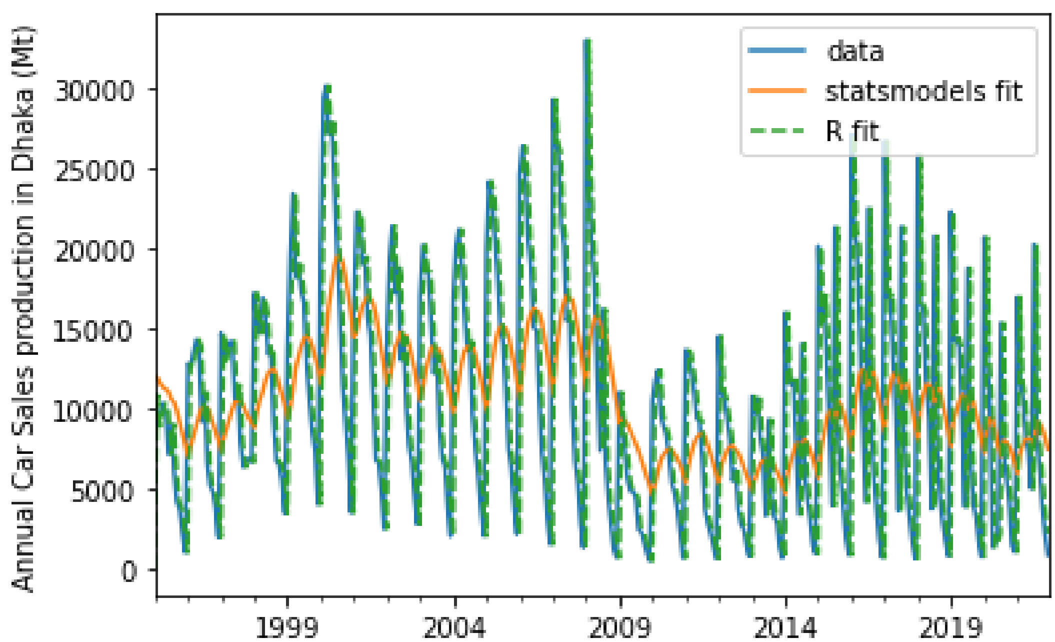


Figure: Comparison Between Statsmodel fit and R fit.

The model summary of ETS model:

```
                              ETS Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  324
Model:                       ETS(ANN)   Log Likelihood               -3318.222
Date:                Mon, 23 May 2022   AIC                           6642.444
Time:                        15:17:11   BIC                           6653.786
Sample:                    01-01-1995   HQIC                          6646.971
                         - 12-01-2021   Scale                     46037829.854
Covariance Type:               approx
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
smoothing_level     0.1193      0.055      2.186      0.029       0.012       0.226
initial_level    1.203e+04   3519.165      3.419      0.001    5135.963    1.89e+04
===================================================================================
Ljung-Box (Q):                       80.70   Jarque-Bera (JB):                22.88
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.35   Skew:                             0.64
Prob(H) (two-sided):                  0.12   Kurtosis:                         2.82
===================================================================================
```

Heuristic Model:

A heuristic, or heuristic technique, is any approach to problem-solving that uses a practical method or various shortcuts in order to produce solutions that may not be optimal but are sufficient given a limited timeframe or deadline.

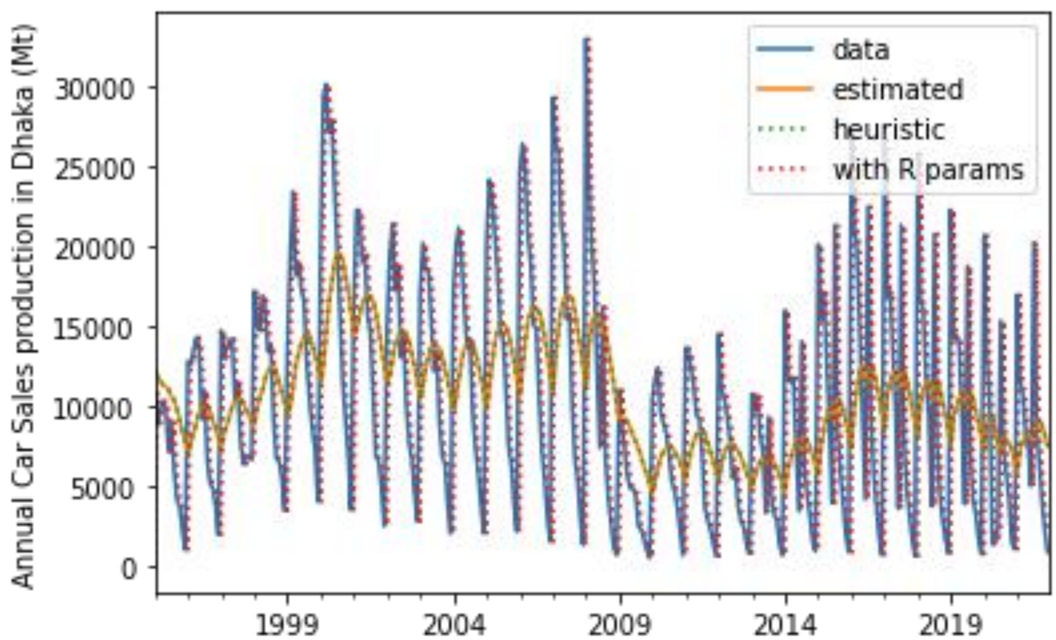


Figure: The comparison Between Heuristic and R fit

Result and Summary of the model:

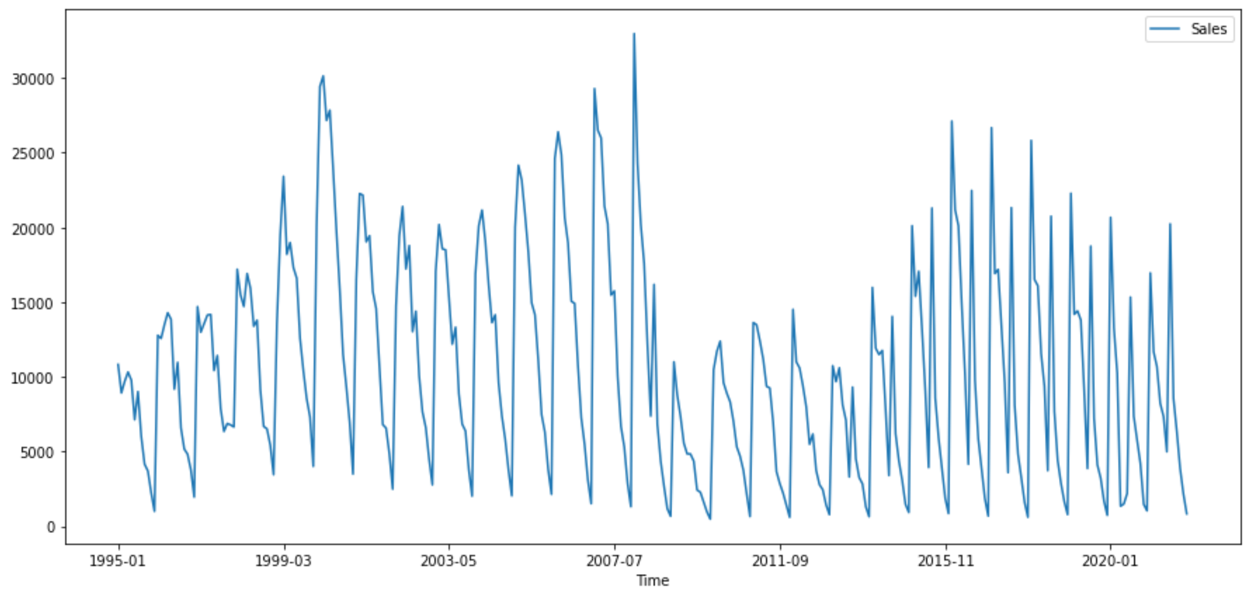
```
                                 ETS Results
==============================================================================
Dep. Variable:                        y   No. Observations:                  324
Model:                         ETS(ANN)   Log Likelihood               -3318.222
Date:                  Mon, 23 May 2022   AIC                           6640.444
Time:                          15:17:14   BIC                           6648.005
Sample:                      01-01-1995   HQIC                          6643.462
                           - 12-01-2021   Scale                     46037829.855
Covariance Type:                 approx
==================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
smoothing_level     0.1193      0.050      2.386      0.017       0.021       0.217
==============================================
              initialization method: heuristic
----------------------------------------------
initial_level                        1.203e+04
==================================================================================
Ljung-Box (Q):                       80.70   Jarque-Bera (JB):                22.88
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.35   Skew:                             0.64
Prob(H) (two-sided):                  0.12   Kurtosis:                         2.82
==================================================================================

Warnings:
[1] Covariance matrix calculated using numerical (complex-step) differentiation.
```

ARIMA/SARIMA model:

ARIMA and SARIMA are both algorithms for forecasting. **ARIMA takes into account the past values (autoregressive, moving average) and predicts future values based on that. SARIMA similarly uses past values but also takes into account any seasonality patterns**.

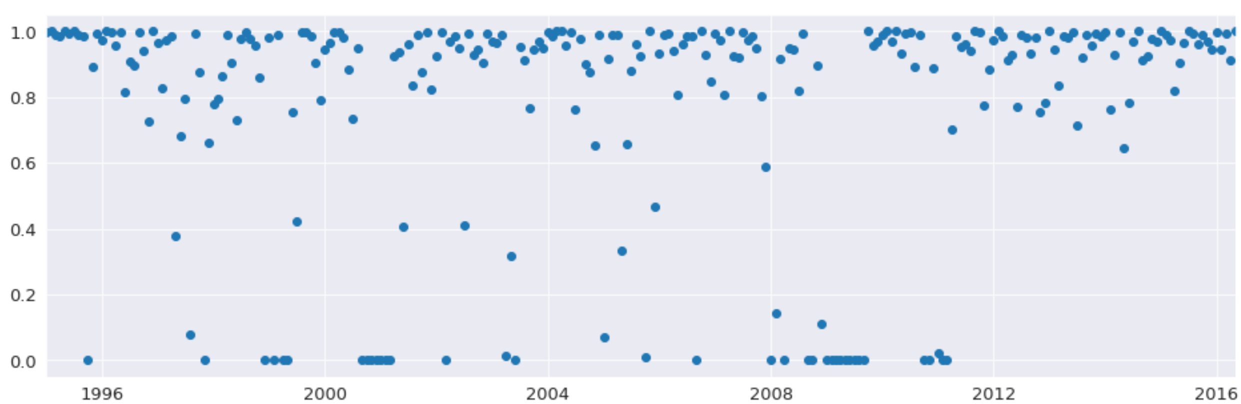
In this assessment I used ARIMA model because the given dataset is not seasonal. To try to improve the Moving Average forecast, I will add a second-order Autoregressive term, creating a seasonal Autoregressive Integrated Moving Average model in the order of ARIMA $(2,1,1)$(2,1,1) $(0,1,0)$(0,1,0) .

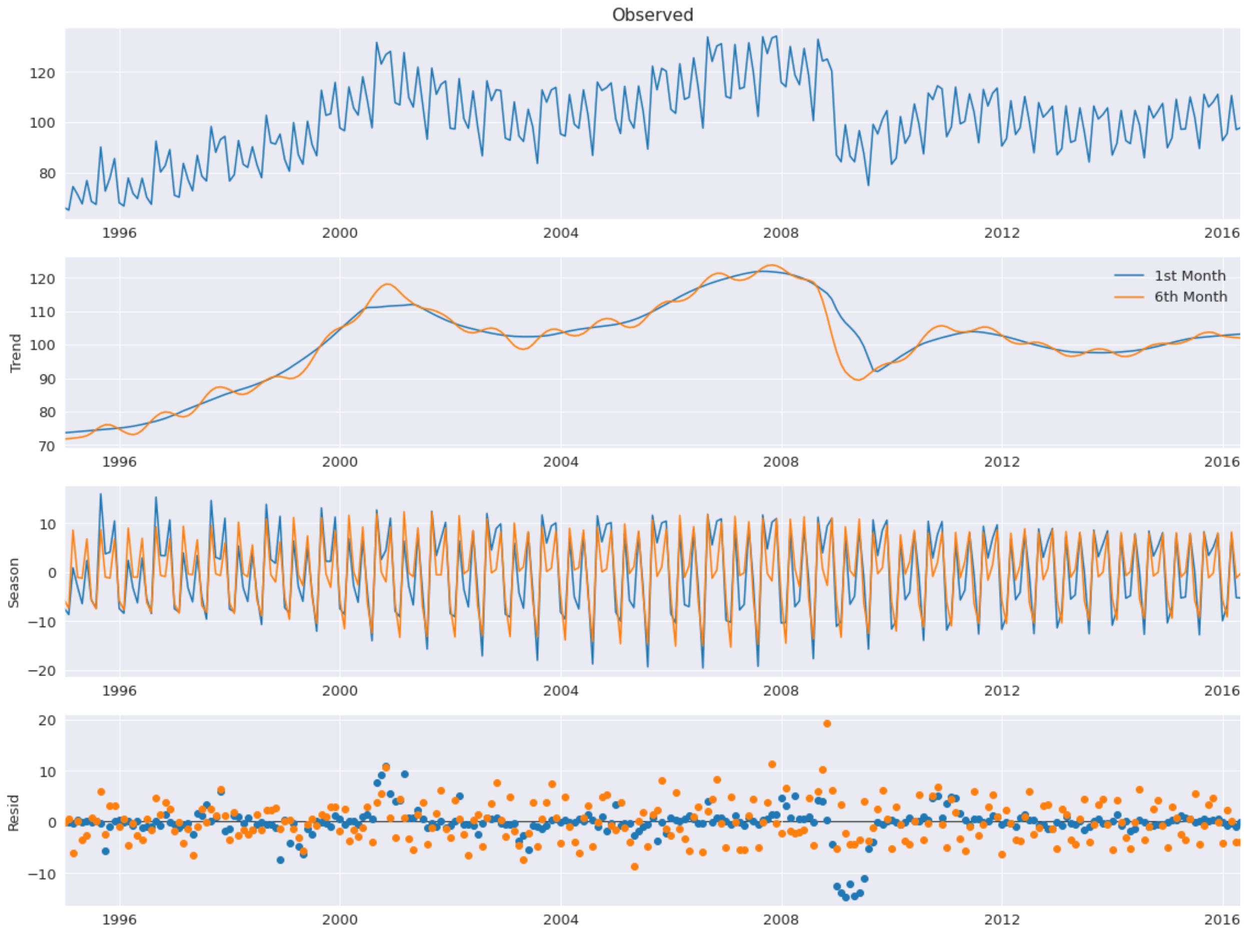
	coef	std err	z	P>|z|	[0.025	0.975]
intercept	1.9849	0.085	23.484	0.0	1.819	2.151
x1	3.0231	0.011	277.150	0.0	3.002	3.044
ar.L1	0.7969	0.009	93.735	0.0	0.780	0.814
sigma2	0.9886	0.020	49.941	0.0	0.950	1.027



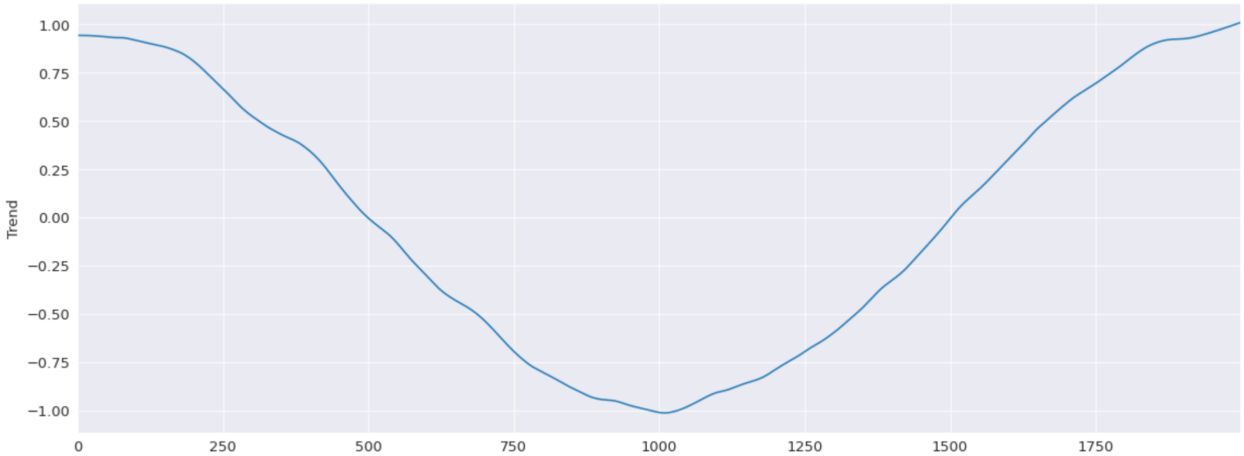
Visualization

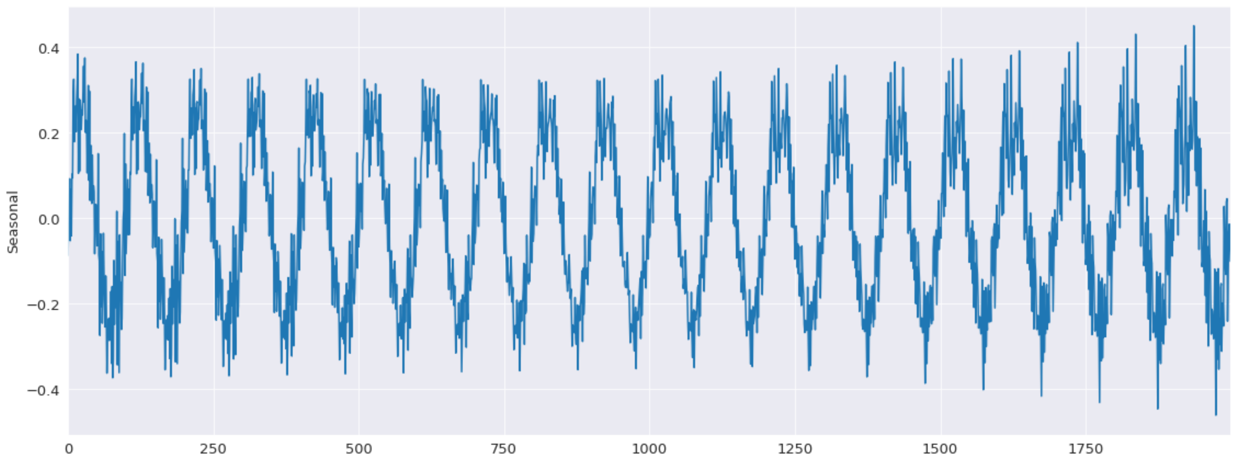
Scatter plot of ARIMA:



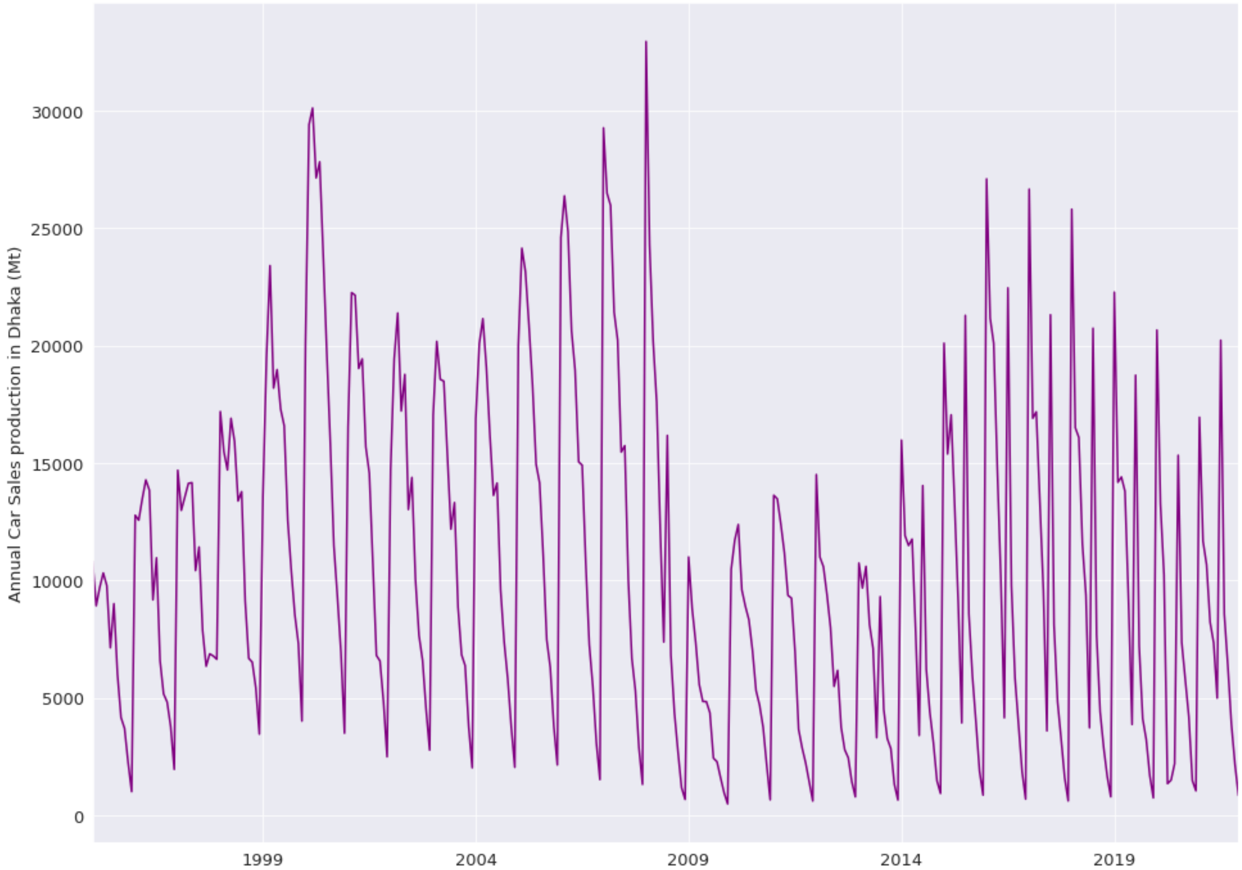


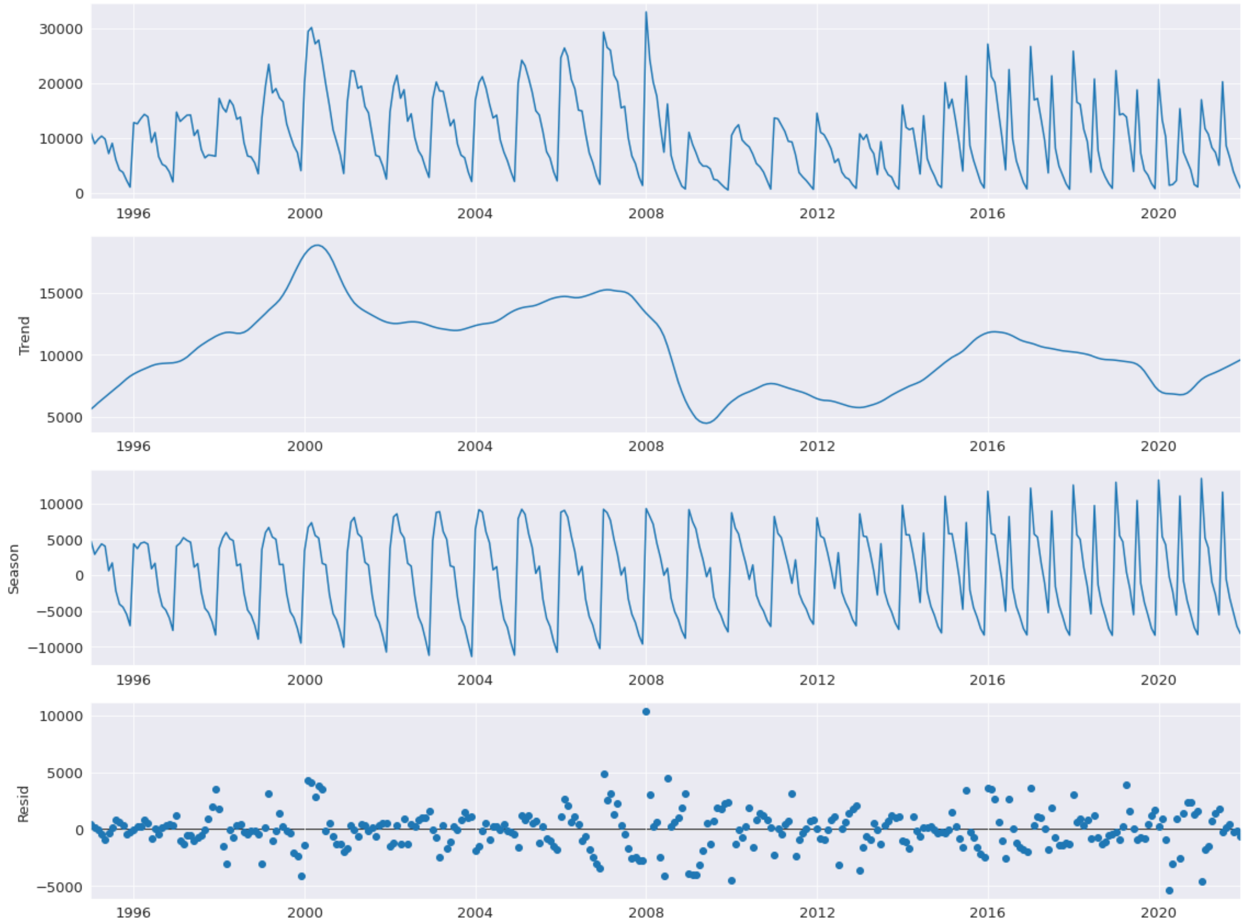
Observed
120
100
80
1996
2000
2004
2008
2012
2016
Trend
120
110
100
90
80
70
1st Month
6th Month
Season
10
0
−10
−20
Resid
20
10
0
−10

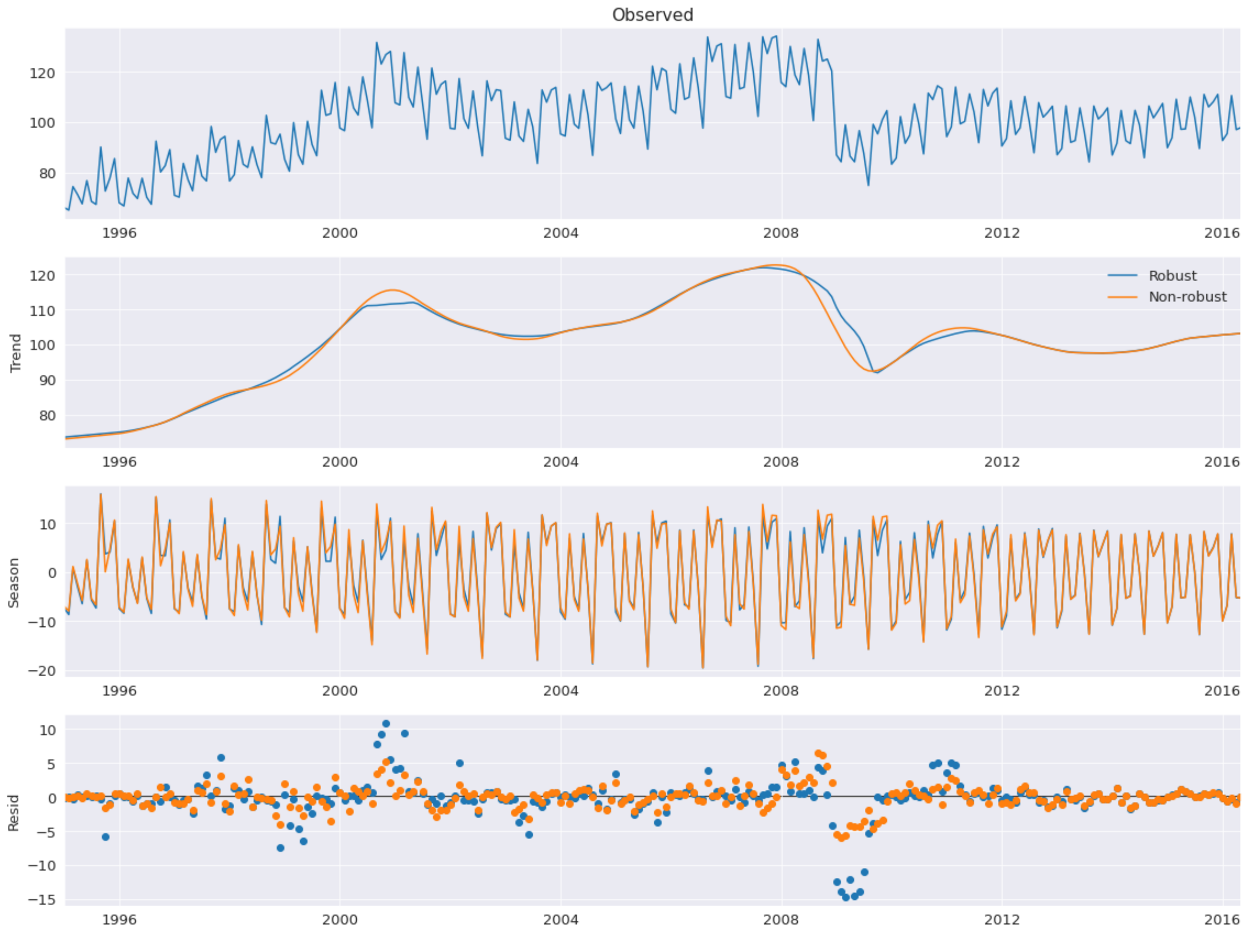




Using SLT FORECAST:


Annual Car Sales production in Dhaka (Mt)
0
5000
10000
15000
20000
25000
30000
1999
2004
2009
2014
2019


30000
20000
10000
0
1996
2000
2004
2008
2012
2016
2020
Trend
15000
10000
5000
Season
10000
5000
0
−5000
−10000
Resid
10000
5000
0
−5000


Observed
Trend
Season
Resid
Robust
Non-robust
1996
2000
2004
2008
2012
2016

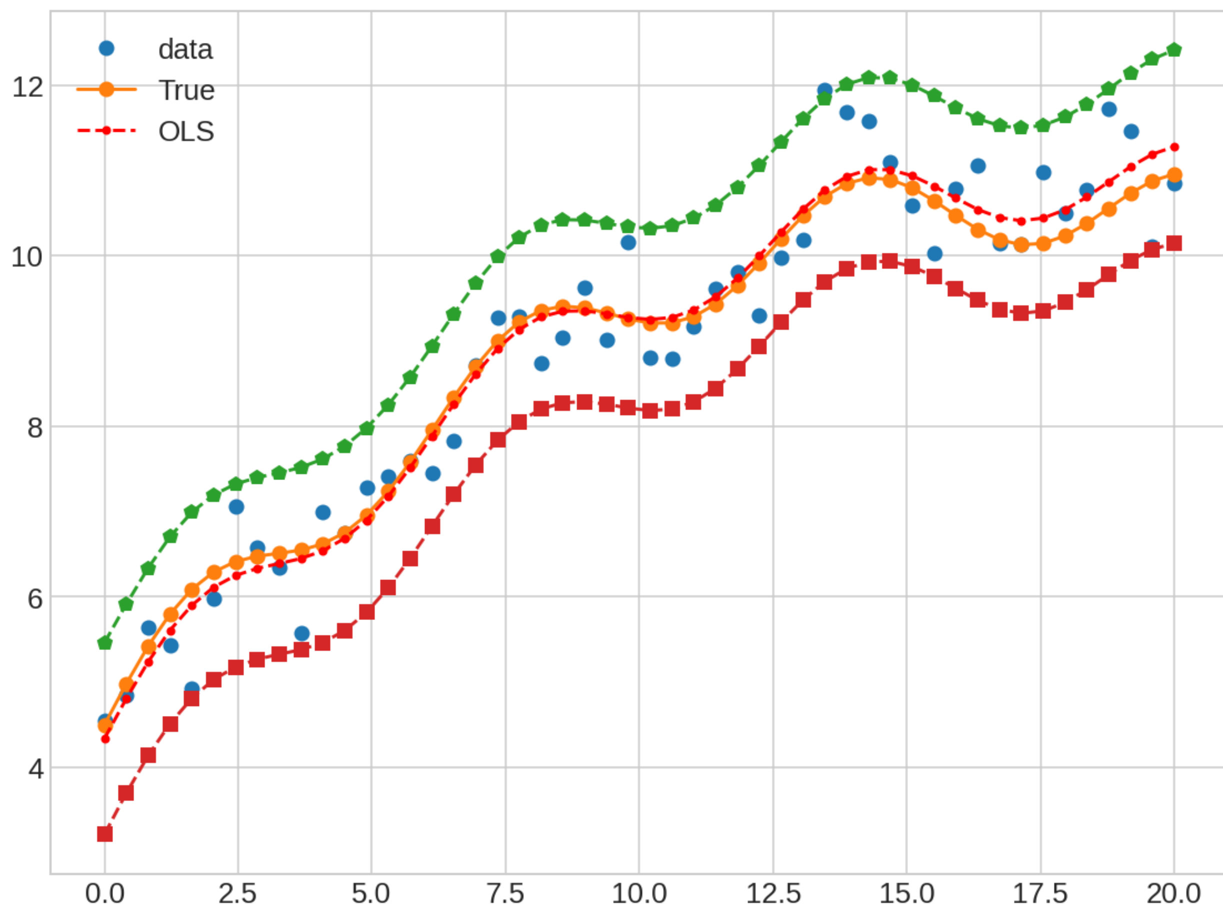
Result:

```
                    STL Decomposition and SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  257
Model:                 ARIMA(1, 1, 0)   Log Likelihood                -522.434
Date:                Mon, 23 May 2022   AIC                           1050.868
Time:                        16:34:37   BIC                           1061.504
Sample:                    01-01-1995   HQIC                          1055.146
                         - 05-01-2016
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1171      0.118      0.995      0.320      -0.113       0.348
ar.L1         -0.0435      0.049     -0.880      0.379      -0.140       0.053
sigma2         3.4682      0.188     18.406      0.000       3.099       3.837
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):               223.01
Prob(Q):                              0.92   Prob(JB):                         0.00
Heteroskedasticity (H):               0.33   Skew:                            -0.26
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.54
                               STL Configuration
================================================================================
Period:                            12       Trend Length:                     23
Seasonal:                           7       Trend deg:                         1
Seasonal deg:                       1       Trend jump:                        1
Seasonal jump:                      1       Low pass:                         13
Robust:                         False       Low pass deg:                      1
--------------------------------------------------------------------------------
```

Linear Models

Ordinary Least Squares

Ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. It is **a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable** (simple or multiple linear regression).



Result:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 4.051e+06
Date:                Mon, 23 May 2022   Prob (F-statistic):          1.97e-239
Time:                        17:25:42   Log-Likelihood:                -146.21
No. Observations:                 100   AIC:                             298.4
Df Residuals:                      97   BIC:                             306.2
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5575      0.312      1.788      0.077      -0.061       1.176
x1             0.1556      0.144      1.080      0.283      -0.130       0.442
x2             9.9988      0.014    717.112      0.000       9.971      10.026
==============================================================================
Omnibus:                        0.008   Durbin-Watson:                   1.760
Prob(Omnibus):                  0.996   Jarque-Bera (JB):                0.089
Skew:                          -0.021   Prob(JB):                        0.956
Kurtosis:                       2.859   Cond. No.                         144.
==============================================================================
```

Comparison:

In this assessment using ARIMA gives me best performance. Here is the comparison table.

Models	Coef	Std Error	z	p>|z|
ETS	0.11	0.55	2.18	0.02
Heuristic	0.11	0.55	2,38	0.17
ARIMA	0.11	0.11	0.99	0.32
OLS	0.55	0.31	1.78	0.07