Imports

Loading Data

```
(145063, 551)
```

	Page	2015-07-01	2015-07-02	2015-07-03	2015-07-04	2015-07-05	2015-07-06	2015-07-07	2015-07-08	2015-07-09	...	2
0	2NE1_zh.wikipedia.org_all-access_spider	18.0	11.0	5.0	13.0	14.0	9.0	9.0	22.0	26.0	...	
1	2PM_zh.wikipedia.org_all-access_spider	11.0	14.0	15.0	18.0	11.0	13.0	22.0	11.0	10.0	...	
2	3C_zh.wikipedia.org_all-access_spider	1.0	0.0	1.0	1.0	0.0	4.0	0.0	3.0	4.0	...	
3	4minute_zh.wikipedia.org_all-access_spider	35.0	13.0	10.0	94.0	4.0	26.0	14.0	9.0	11.0	...	
4	52_Hz_I_Love_You_zh.wikipedia.org_all-access_s...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	

5 rows × 551 columns

	2015-07-01	2015-07-02	2015-07-03	2015-07-04	2015-07-05	2015-07-06	2015-07-07	
count	1.243230e+05	1.242470e+05	1.245190e+05	1.244090e+05	1.244040e+05	1.245800e+05	1.243990e+05	1
mean	1.195857e+03	1.204004e+03	1.133676e+03	1.170437e+03	1.217769e+03	1.290273e+03	1.239137e+03	1
std	7.275352e+04	7.421515e+04	6.961022e+04	7.257351e+04	7.379612e+04	8.054448e+04	7.576288e+04	6
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0
25%	1.300000e+01	1.300000e+01	1.200000e+01	1.300000e+01	1.400000e+01	1.100000e+01	1.300000e+01	1
50%	1.090000e+02	1.080000e+02	1.050000e+02	1.050000e+02	1.130000e+02	1.130000e+02	1.150000e+02	1
75%	5.240000e+02	5.190000e+02	5.040000e+02	4.870000e+02	5.400000e+02	5.550000e+02	5.510000e+02	5
max	2.038124e+07	2.075219e+07	1.957397e+07	2.043964e+07	2.077211e+07	2.254467e+07	2.121089e+07	1

8 rows × 550 columns

```
(550, 1)
```

	Exog
0	0
1	0
2	0
3	0
4	0

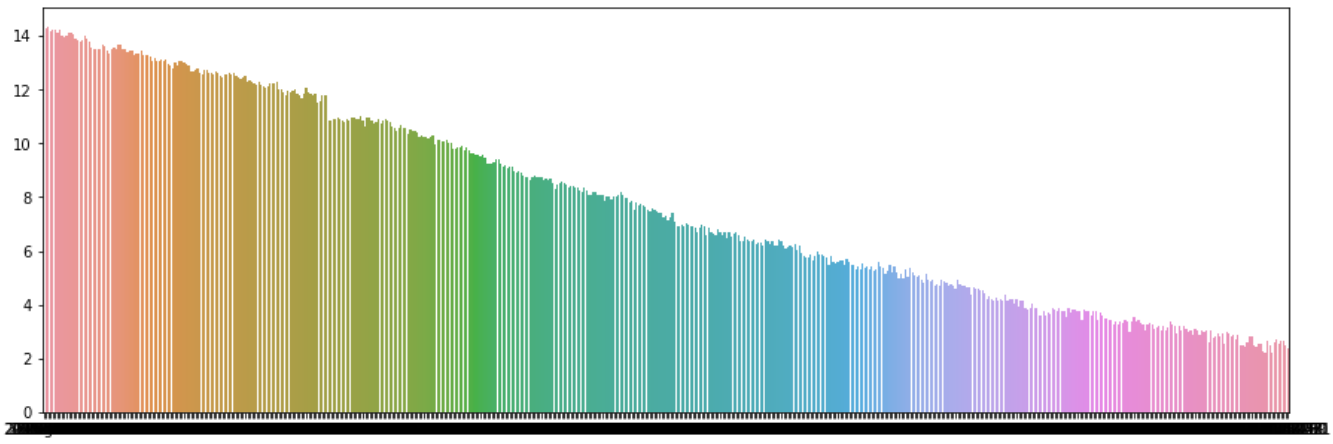
```
Exog
0       496
1        54
dtype: int64
```

EDA

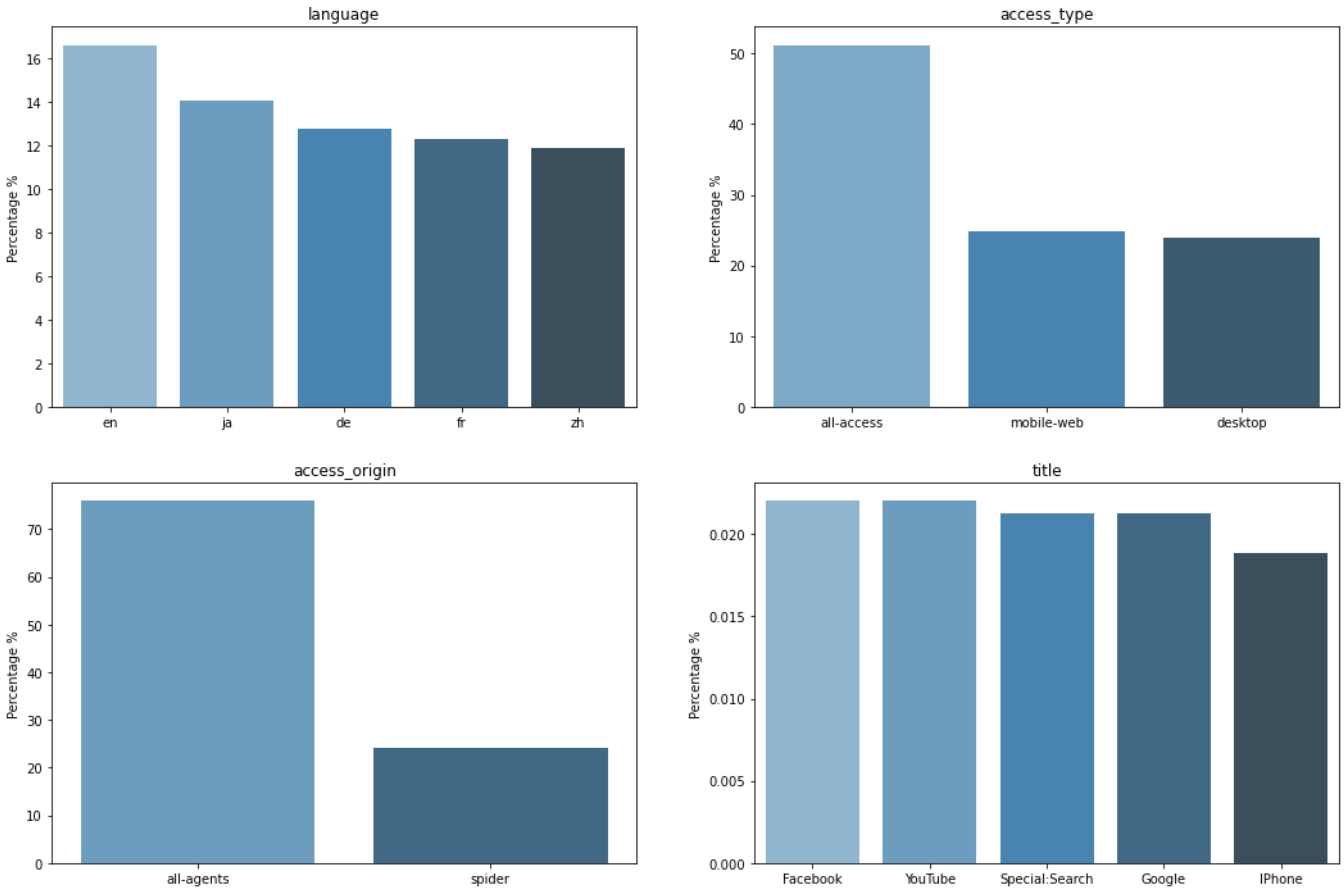
	Page	2015-07-01	2015-07-02	2015-07-03	2015-07-04	2015-07-05	2015-07-06	2015-07-07	2015-07-08	2015-07-09	...	2
0	2NE1_zh.wikipedia.org_all-access_spider	18.0	11.0	5.0	13.0	14.0	9.0	9.0	22.0	26.0	...	
1	2PM_zh.wikipedia.org_all-access_spider	11.0	14.0	15.0	18.0	11.0	13.0	22.0	11.0	10.0	...	
2	3C_zh.wikipedia.org_all-access_spider	1.0	0.0	1.0	1.0	0.0	4.0	0.0	3.0	4.0	...	
3	4minute_zh.wikipedia.org_all-access_spider	35.0	13.0	10.0	94.0	4.0	26.0	14.0	9.0	11.0	...	
4	52_Hz_I_Love_You_zh.wikipedia.org_all-access_s...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	

5 rows × 551 columns

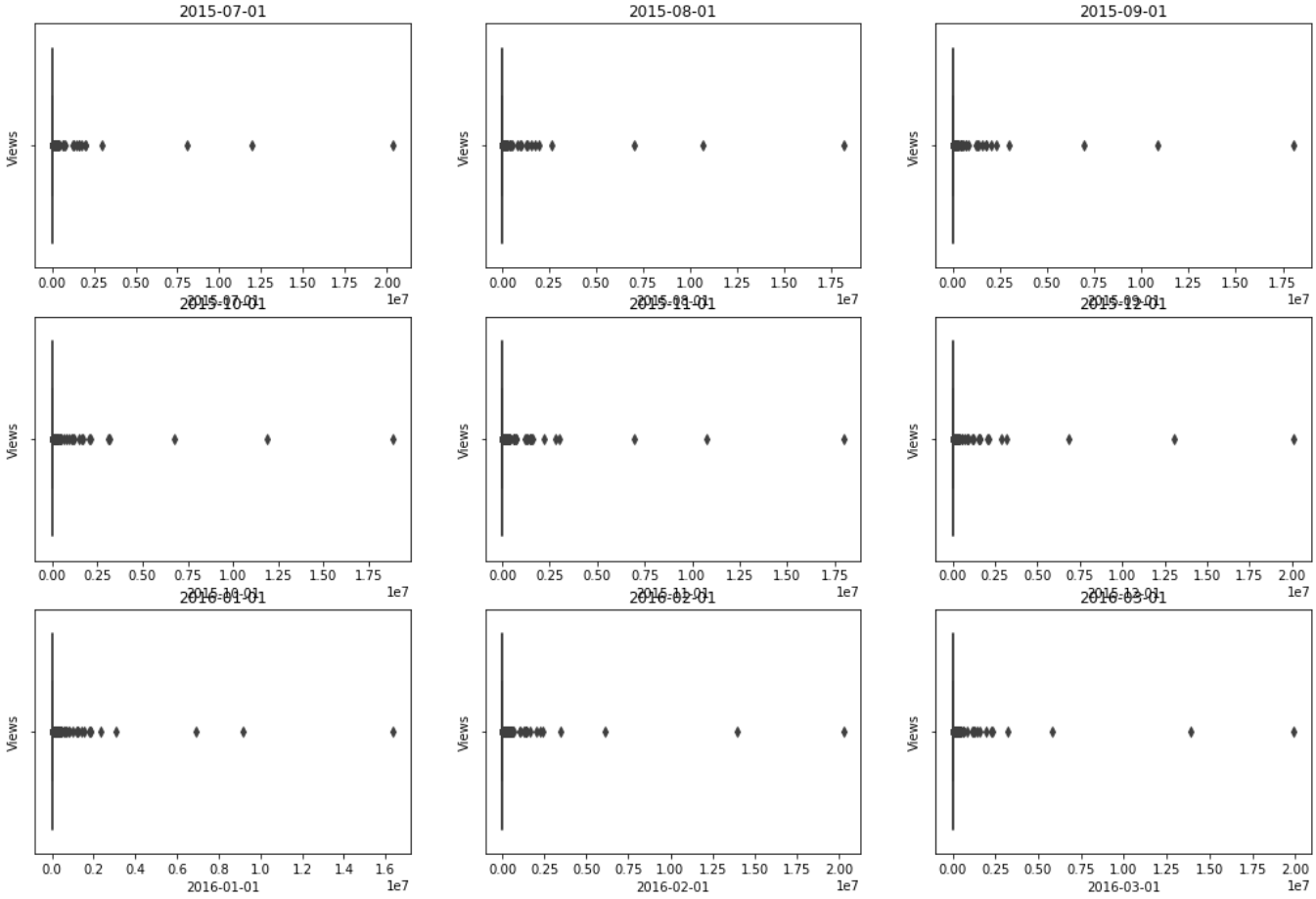
```
<AxesSubplot:>
```



- There is clear linear trend in null rates, decreasing with time
- So most of the sites are being visited in Dec'16 (98%) as compared to Jul'15 (85%)



- English is the most popular language ~16%
- The most common type of access is all-access ~ 50%
- all-agents (~75%) is more common access origin than spider
- Top 5 web pages are : FB, YT, Special Search, Google, Iphone



- The data has a lot of outliers. It would be better to take the median when grouping by language to tackle this issue

Formatting Data for Model

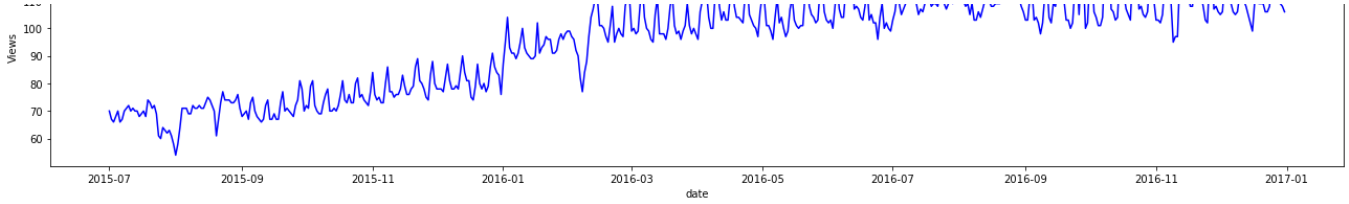
```
(145063, 554)
```

```
(144411, 554)
```

```
(550, 8)
```

	date	de	en	es	fr	ja	ru	zh
0	2015-07-01	87.5	165.0	287.0	85.0	167.0	157.0	70.0
1	2015-07-02	86.0	161.0	284.0	86.0	162.0	159.0	67.0
2	2015-07-03	85.0	151.0	255.0	85.0	171.0	158.0	66.0
3	2015-07-04	85.0	149.0	215.0	84.0	184.0	155.0	68.0
4	2015-07-05	97.5	157.0	254.0	93.0	183.0	164.0	70.0

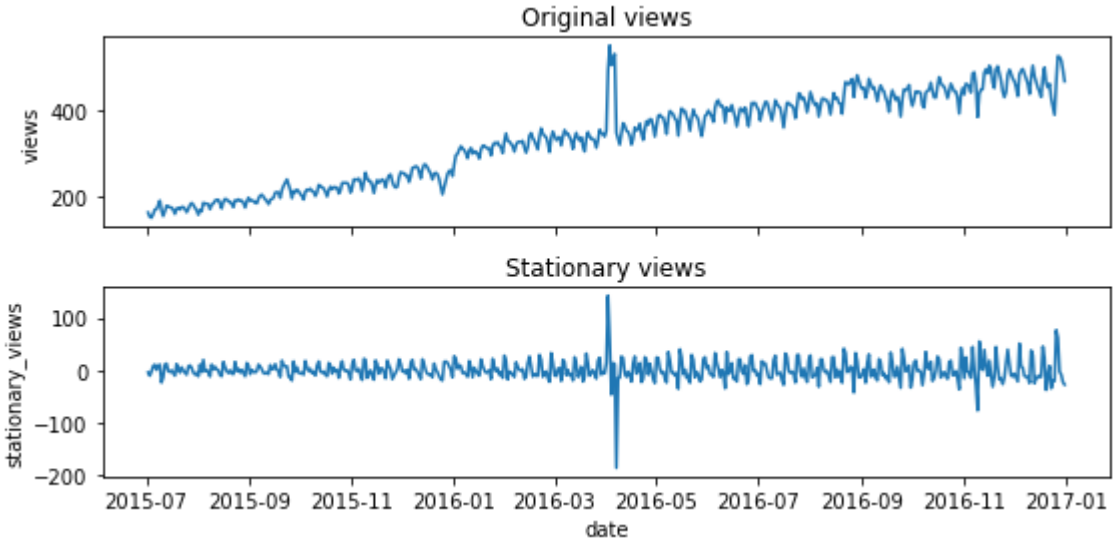

de
Views
date
en
Views
date
es
Views
date
fr
Views
date
ja
Views
date
ru
Views
date
zh



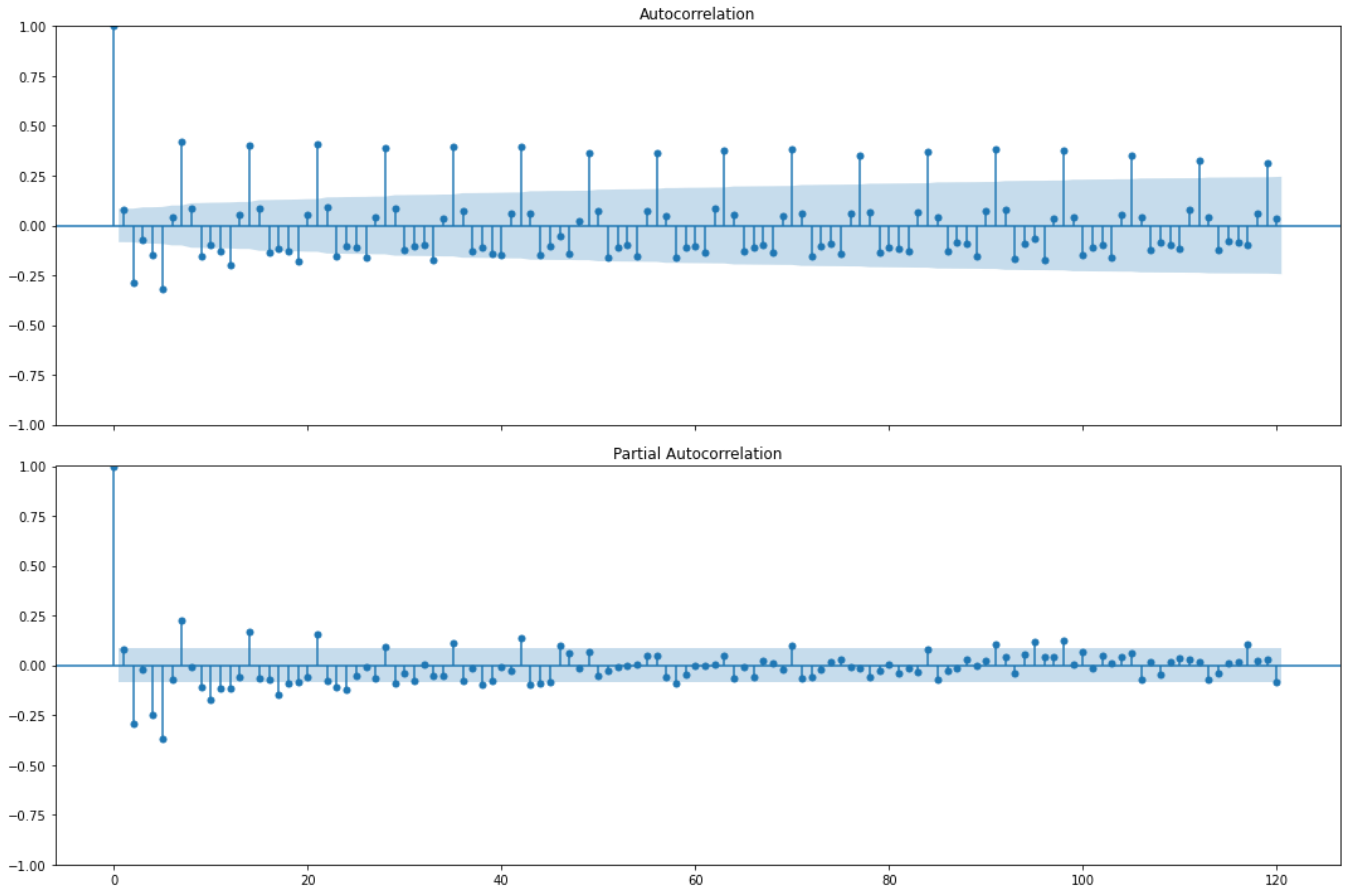
Time Series Analysis for English

	date	views	stationary_views	exog
0	2015-07-01	165.0	165.0	0
1	2015-07-02	161.0	161.0	0
2	2015-07-03	151.0	151.0	0
3	2015-07-04	149.0	149.0	0
4	2015-07-05	157.0	157.0	0

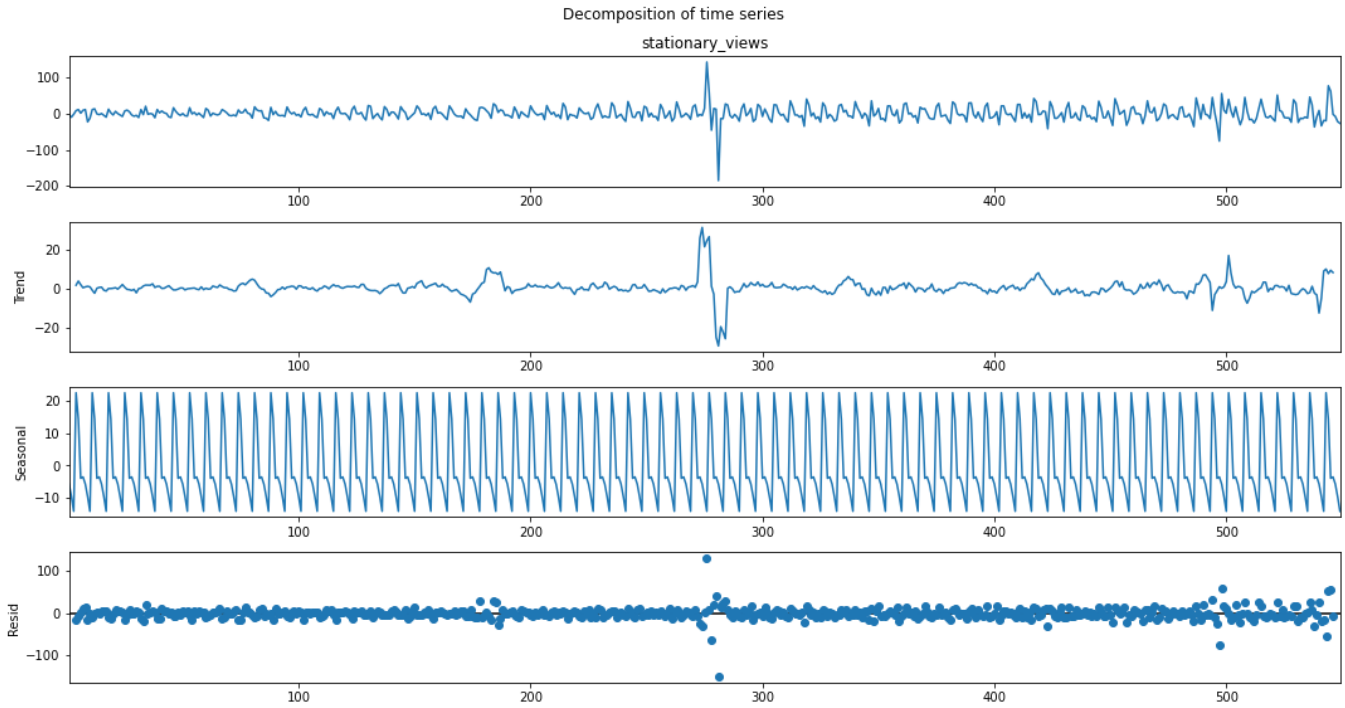
********** Stationary views obtained after 1 differencing! **********



	date	views	stationary_views	exog
1	2015-07-02	161.0	-4.0	0
2	2015-07-03	151.0	-10.0	0
3	2015-07-04	149.0	-2.0	0
4	2015-07-05	157.0	8.0	0
5	2015-07-06	168.0	11.0	0



- From acf plot we can see that there is weekly seasonality



Modelling

```
((411, 4), (138, 4))
```

```
(4, 8, 2)
```

- Searching for best parameters for SARIMAX model...

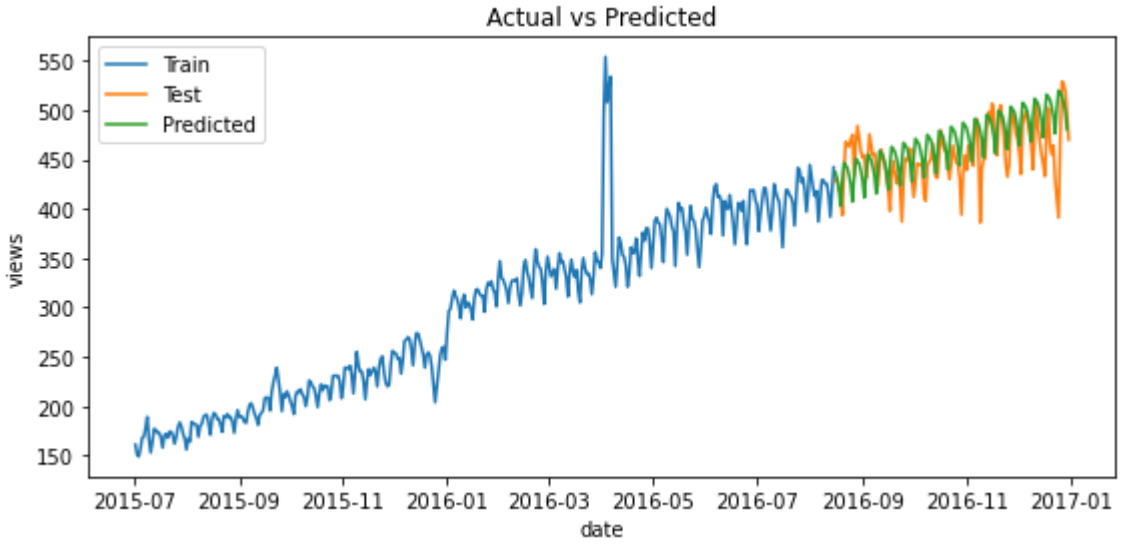
```
100%|██████████| 4/4 [00:14<00:00,  3.73s/it]
**************** Best model AIC is 3330.1270838111077 found with order (1, 1, 1) and seasonal order (0, 1, 1, 7) and exog False******************
```

```
((1, 1, 1), (0, 1, 1, 7), False)
```

Building best SARIMAX model

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                             views   No. Observations:                  411
Model:             SARIMAX(1, 1, 1)x(0, 1, 1, 7)   Log Likelihood               -1661.064
Date:                           Wed, 05 Apr 2023   AIC                           3330.127
Time:                                   23:47:33   BIC                           3346.123
Sample:                                        0   HQIC                          3336.460
                                             - 411
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5186      0.219     -2.373      0.018      -0.947      -0.090
ma.L1          0.6542      0.191      3.416      0.001       0.279       1.029
ma.S.L7       -0.9304      0.025    -36.514      0.000      -0.980      -0.880
sigma2       215.0422      2.812     76.479      0.000     209.531     220.553
===================================================================================
Ljung-Box (L1) (Q):                   2.36   Jarque-Bera (JB):            115458.97
Prob(Q):                              0.12   Prob(JB):                         0.00
Heteroskedasticity (H):              11.24   Skew:                            -2.03
Prob(H) (two-sided):                  0.00   Kurtosis:                        85.82
===================================================================================

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

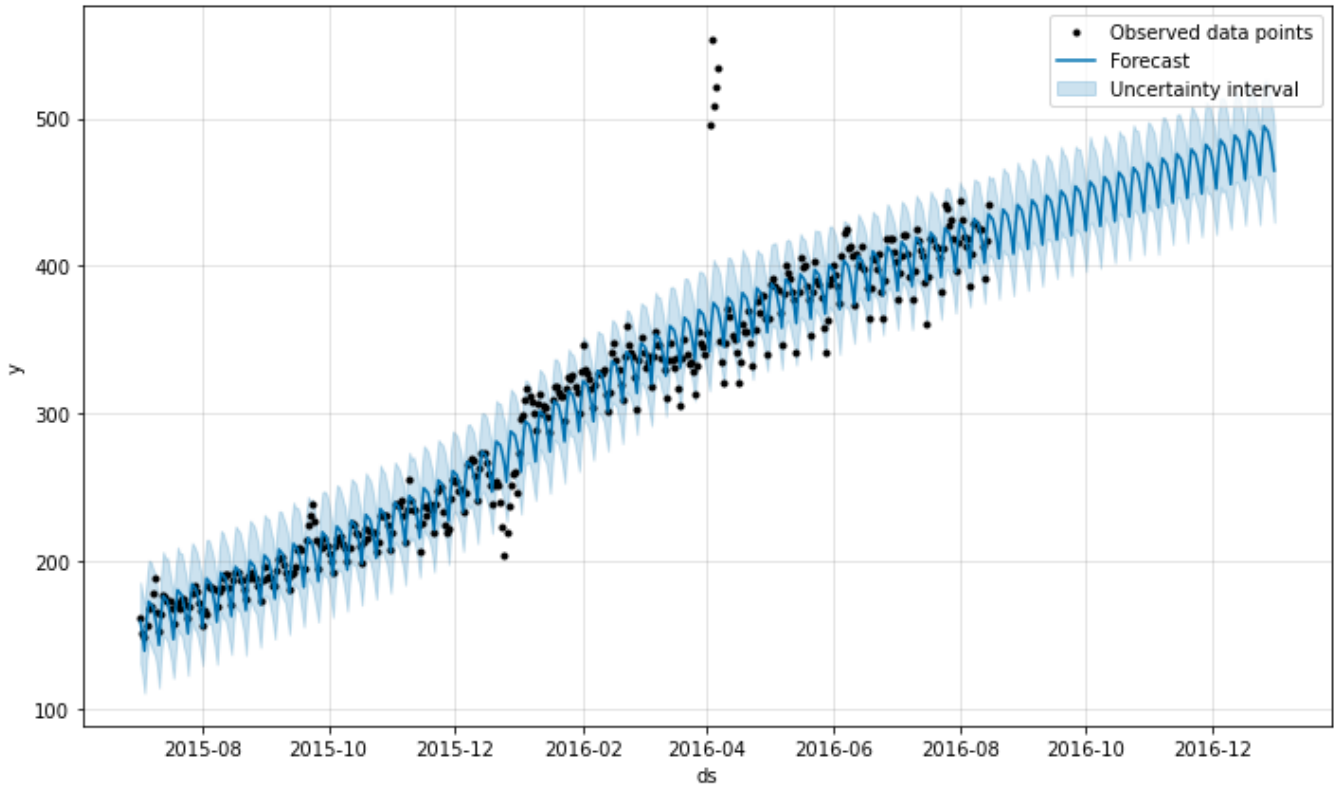
```
**************** Performance Metrics ****************

RMSE : 26.73
MAPE: 4.66 %
```



Training using prophet

```
23:47:37 - cmdstanpy - INFO - Chain [1] start processing
23:47:37 - cmdstanpy - INFO - Chain [1] done processing
```

```
<prophet.forecaster.Prophet at 0x1f8c34efa60>
```

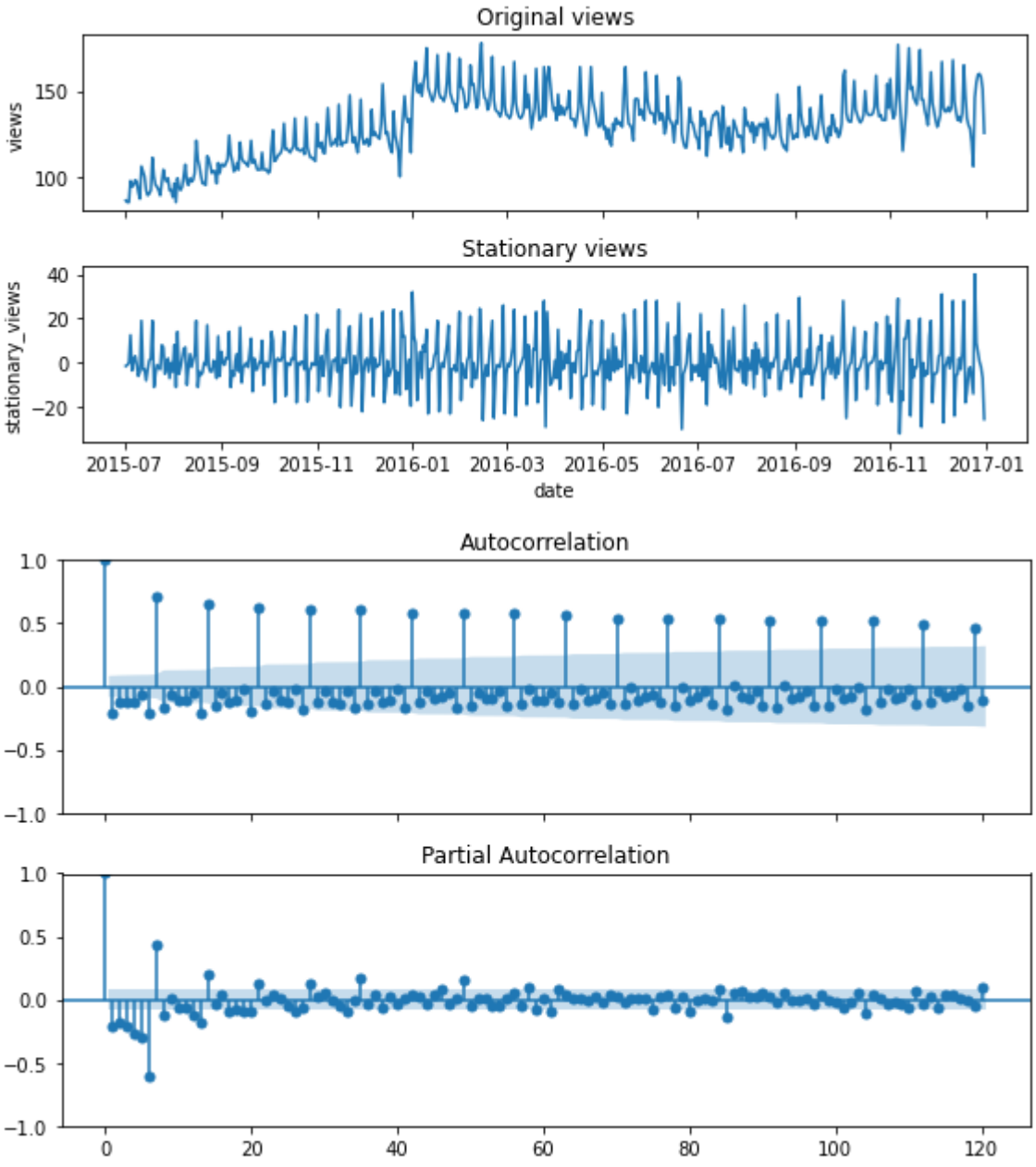


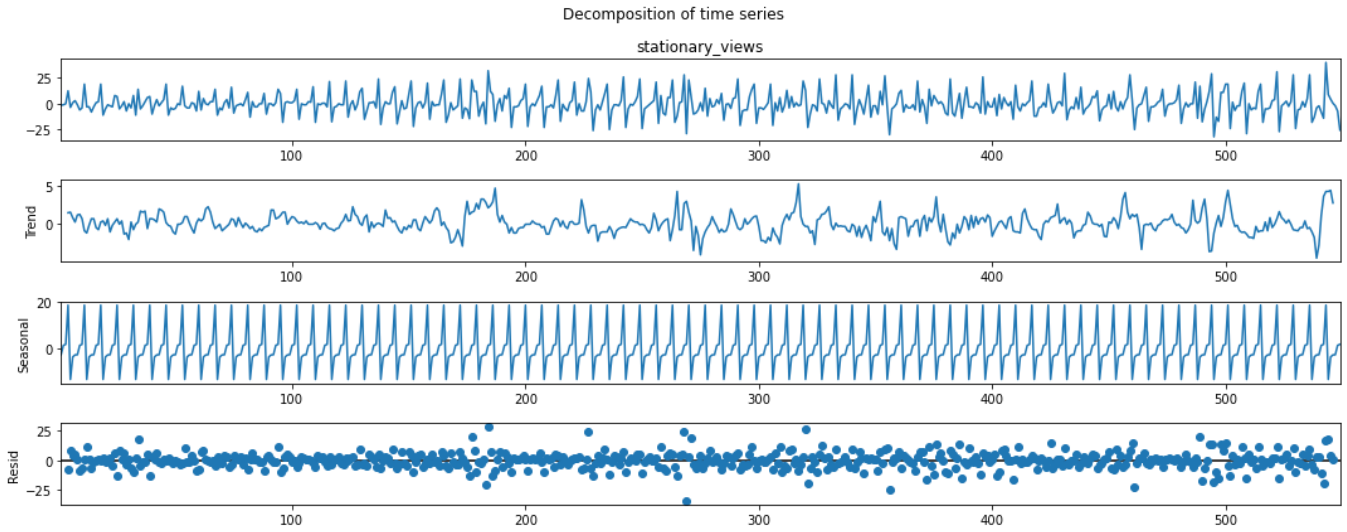

```
********** Performance Metrics **********

RMSE : 21.99
MAPE: 3.67 %
```

Running for other languages

********** Language : de **********

********** Stationary views obtained after 1 differencing! **********






```
100%|██████████| 4/4 [00:02<00:00,  1.76it/s]
```

********** Best model AIC is 2647.0356865121207 found with order (1, 1, 1) and seasonal order (1, 0, 1, 7) and exog False**********

SARIMAX Results

Dep. Variable:	views	No. Observations:	411
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 7)	Log Likelihood	-1318.518
Date:	Wed, 05 Apr 2023	AIC	2647.036
Time:	23:47:48	BIC	2667.116
Sample:	0	HQIC	2654.980
	- 411		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5005	0.048	10.401	0.000	0.406	0.595
ma.L1	-0.8952	0.034	-26.213	0.000	-0.962	-0.828
ar.S.L7	0.9877	0.007	138.213	0.000	0.974	1.002
ma.S.L7	-0.8229	0.037	-22.461	0.000	-0.895	-0.751
sigma2	35.4220	1.515	23.379	0.000	32.452	38.392

Ljung-Box (L1) (Q):	0.31	Jarque-Bera (JB):	288.04
Prob(Q):	0.58	Prob(JB):	0.00
Heteroskedasticity (H):	2.68	Skew:	0.74
Prob(H) (two-sided):	0.00	Kurtosis:	6.83

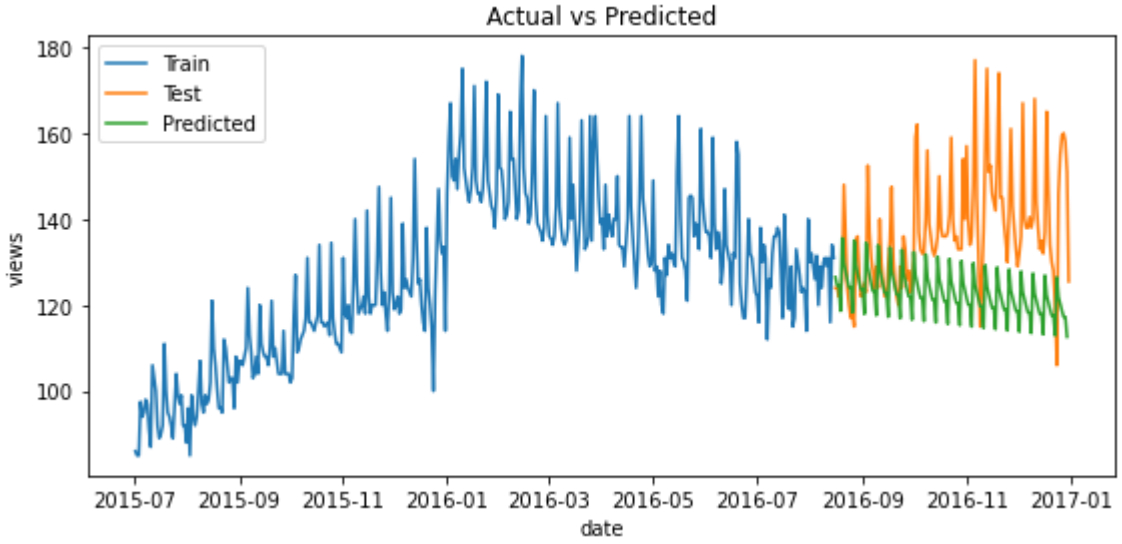
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

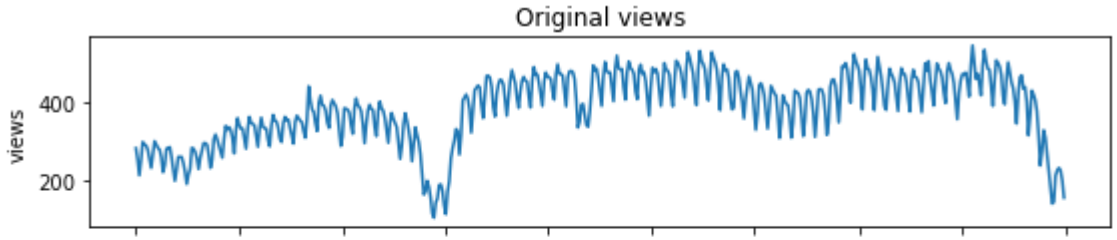
********** Performance Metrics **********

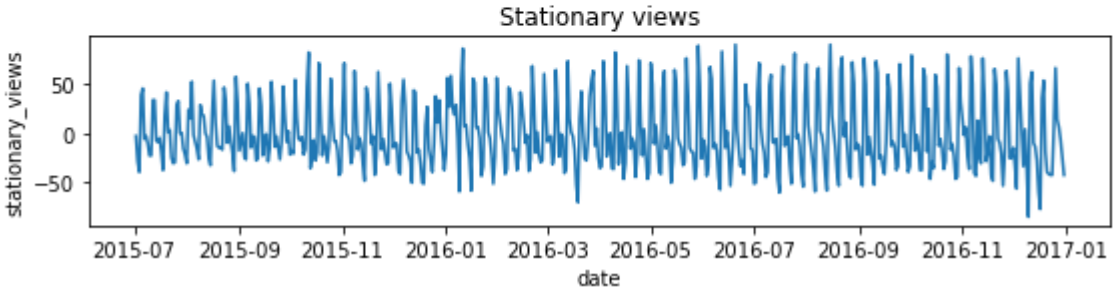
RMSE : 19.24
MAPE: 10.62 %

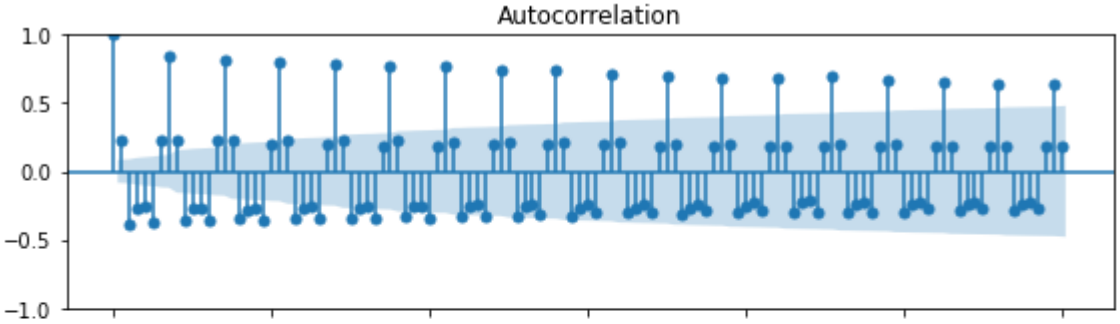
********** Language : es **********

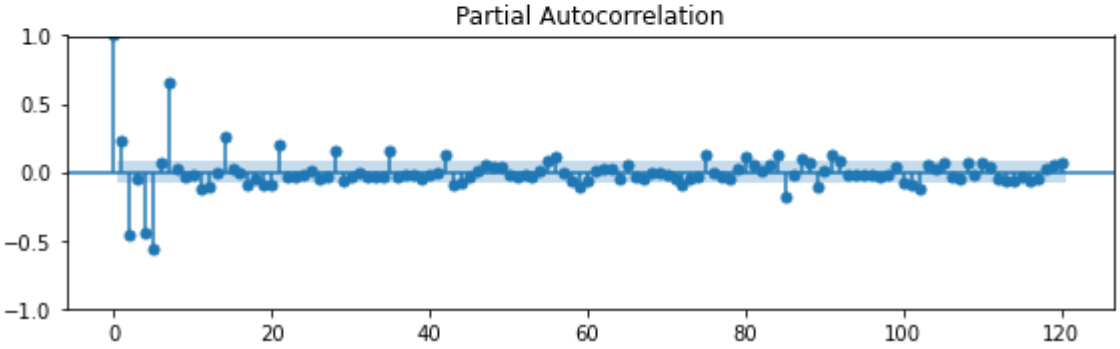
********** Stationary views obtained after 1 differencing! **********

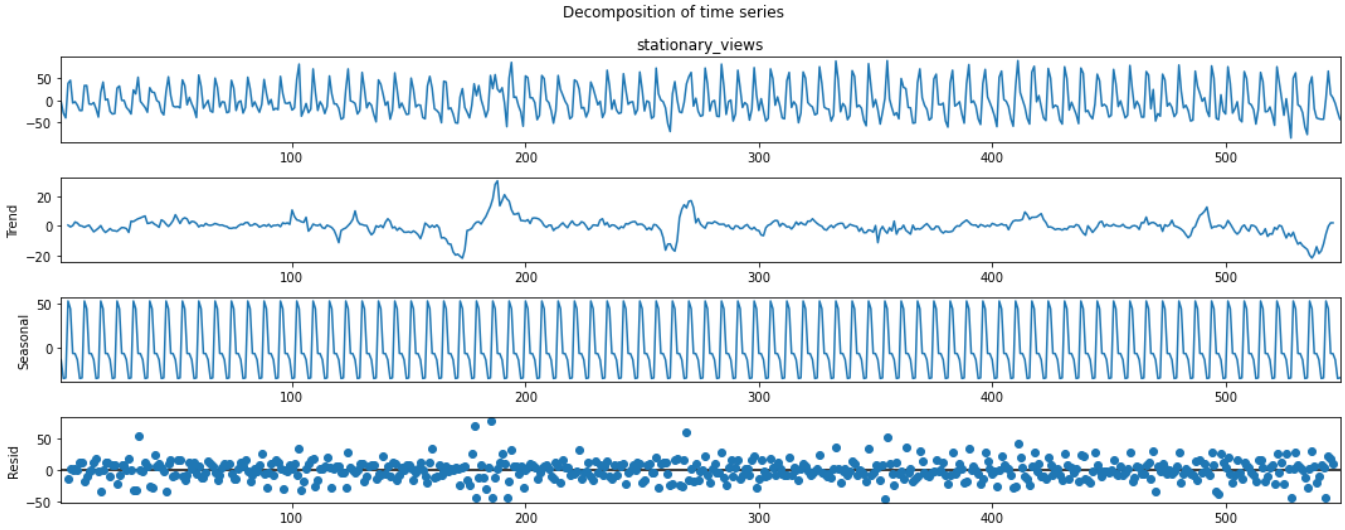

Actual vs Predicted
Train
Test
Predicted
views
date


Original views
views


Stationary views
stationary_views
date


Autocorrelation


Partial Autocorrelation




```
100%|██████████| 4/4 [00:02<00:00,  1.85it/s]
```

********** Best model AIC is 3476.0251259879037 found with order (0, 1, 0) and seasonal order (1, 0, 1, 7) and exog False**********

SARIMAX Results

Dep. Variable:	views	No. Observations:	411
Model:	SARIMAX(0, 1, 0)x(1, 0, [1], 7)	Log Likelihood	-1735.013
Date:	Wed, 05 Apr 2023	AIC	3476.025
Time:	23:47:53	BIC	3488.074
Sample:	0	HQIC	3480.792
	- 411		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L7	0.9956	0.004	256.476	0.000	0.988	1.003
ma.S.L7	-0.8107	0.027	-30.472	0.000	-0.863	-0.759
sigma2	265.7709	11.409	23.295	0.000	243.409	288.132

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	378.76
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	1.14	Skew:	0.90
Prob(H) (two-sided):	0.45	Kurtosis:	7.35

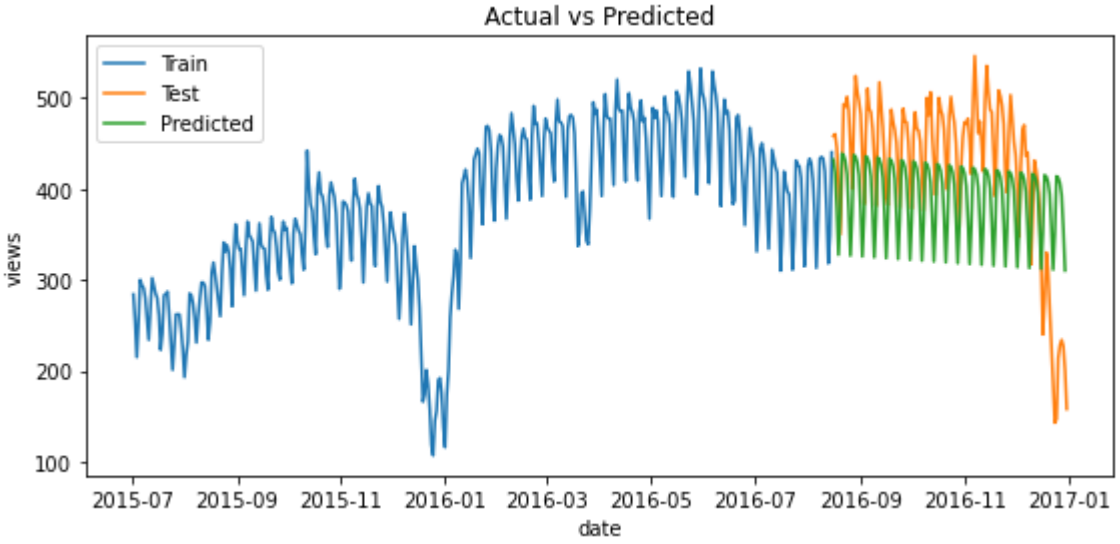
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

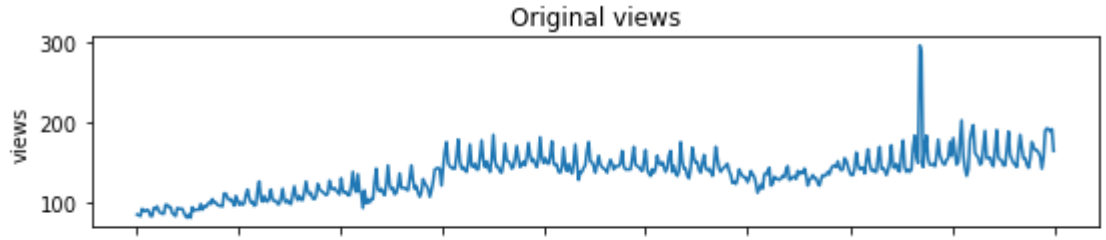
********** Performance Metrics **********

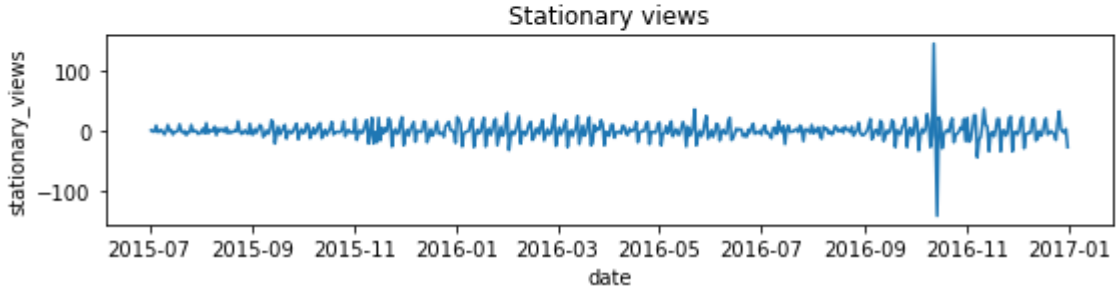
RMSE : 80.35
MAPE: 19.74 %

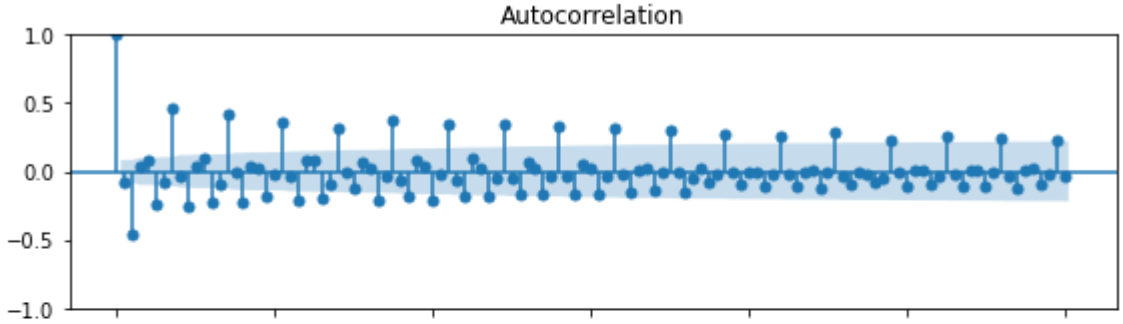
********** Language : fr **********

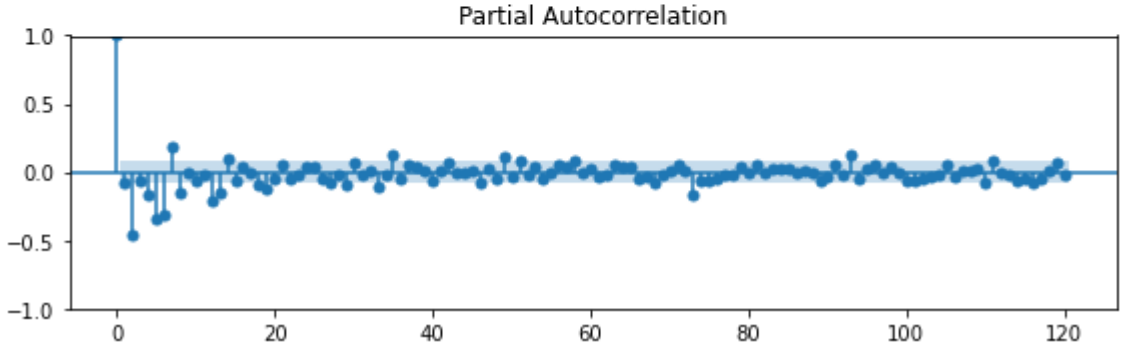
********** Stationary views obtained after 1 differencing! **********

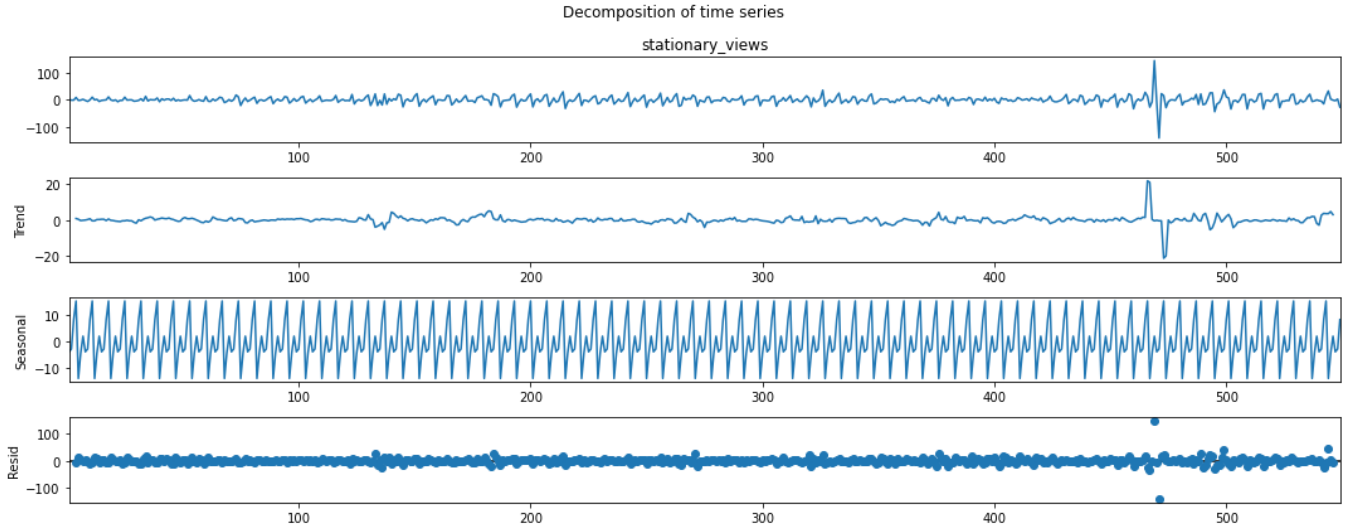

Actual vs Predicted
Train
Test
Predicted
views
500
400
300
200
100
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01
date


Original views
views
300
200
100


Stationary views
stationary_views
100
0
−100
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01
date


Autocorrelation
1.0
0.5
0.0
−0.5
−1.0


Partial Autocorrelation
1.0
0.5
0.0
−0.5
−1.0
0
20
40
60
80
100
120




```
100%|██████████| 4/4 [00:01<00:00,  2.68it/s]
```

********** Best model AIC is 2640.2367273049176 found with order (1, 1, 1) and seasonal order (1, 0, 1, 7) and exog False**********

SARIMAX Results

Dep. Variable:	views	No. Observations:	411
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 7)	Log Likelihood	-1315.118
Date:	Wed, 05 Apr 2023	AIC	2640.237
Time:	23:47:58	BIC	2660.318
Sample:	0	HQIC	2648.181
	- 411		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5575	0.047	11.853	0.000	0.465	0.650
ma.L1	-0.9235	0.024	-38.041	0.000	-0.971	-0.876
ar.S.L7	0.9695	0.011	89.820	0.000	0.948	0.991
ma.S.L7	-0.7181	0.041	-17.695	0.000	-0.798	-0.639
sigma2	35.0540	1.504	23.305	0.000	32.106	38.002

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	258.46
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	1.66	Skew:	0.64
Prob(H) (two-sided):	0.00	Kurtosis:	6.67

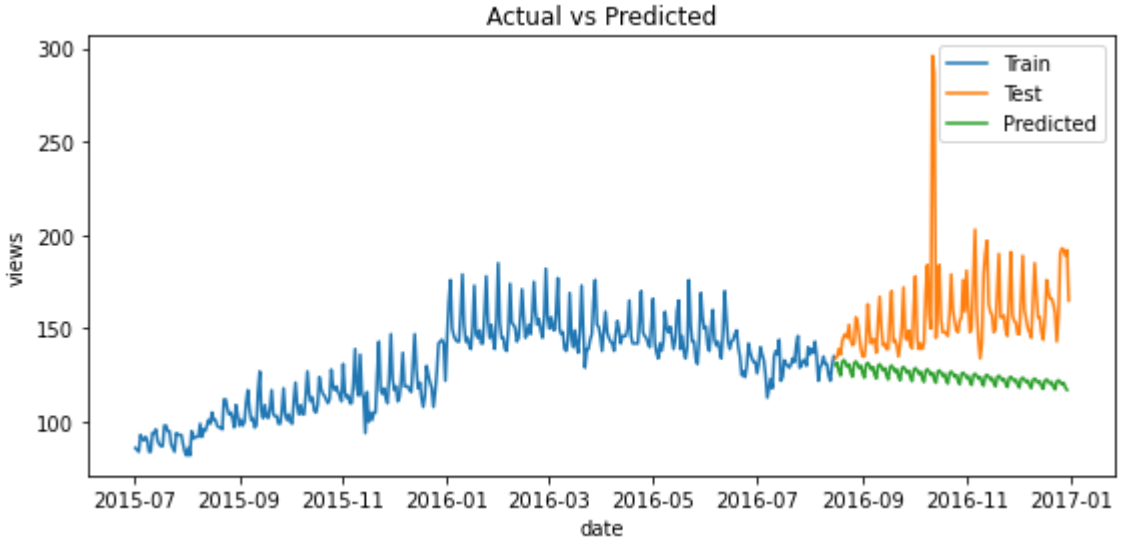
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

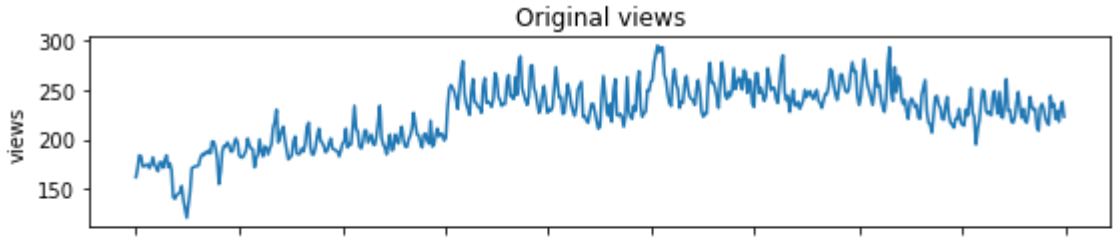
********** Performance Metrics **********

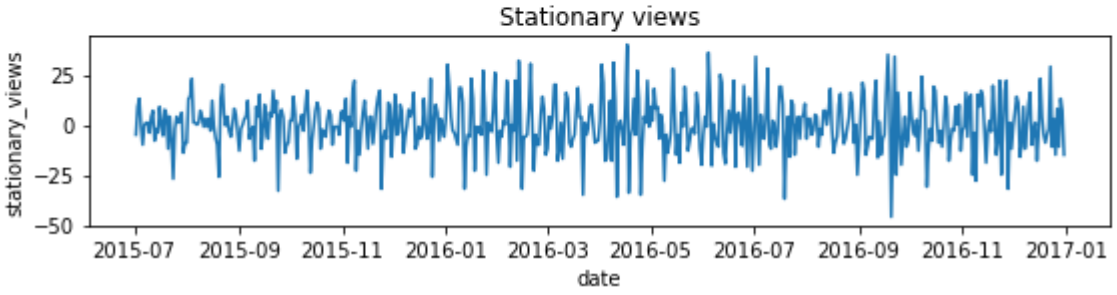
RMSE : 40.98
MAPE: 19.81 %

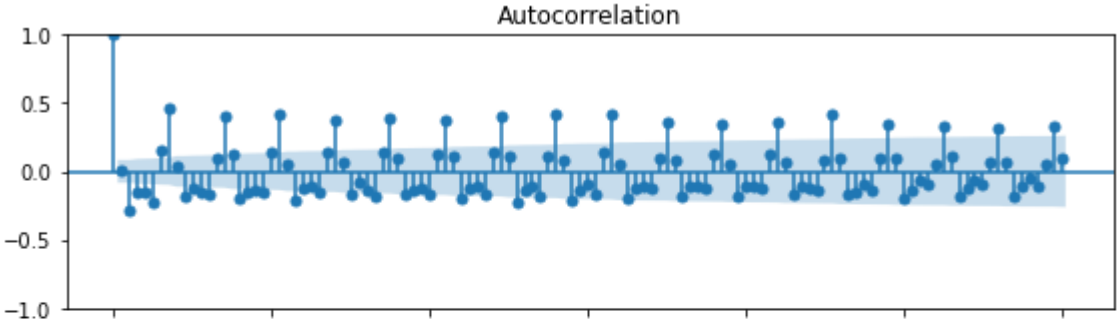
********** Language : ja **********

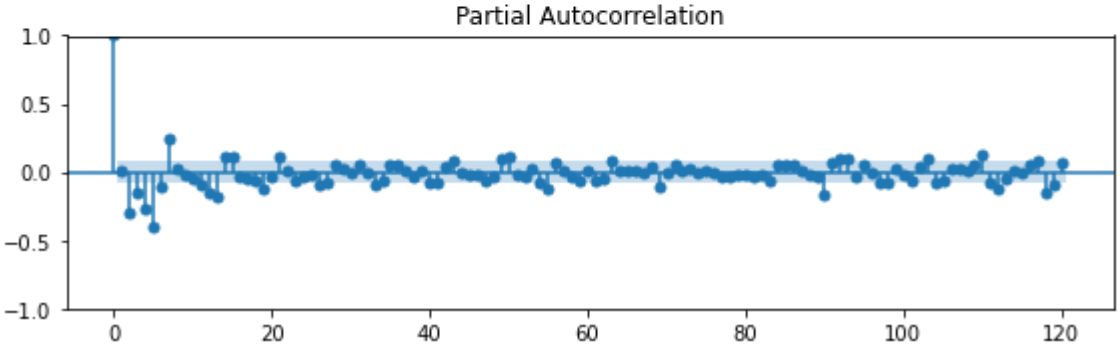
********** Stationary views obtained after 1 differencing! **********

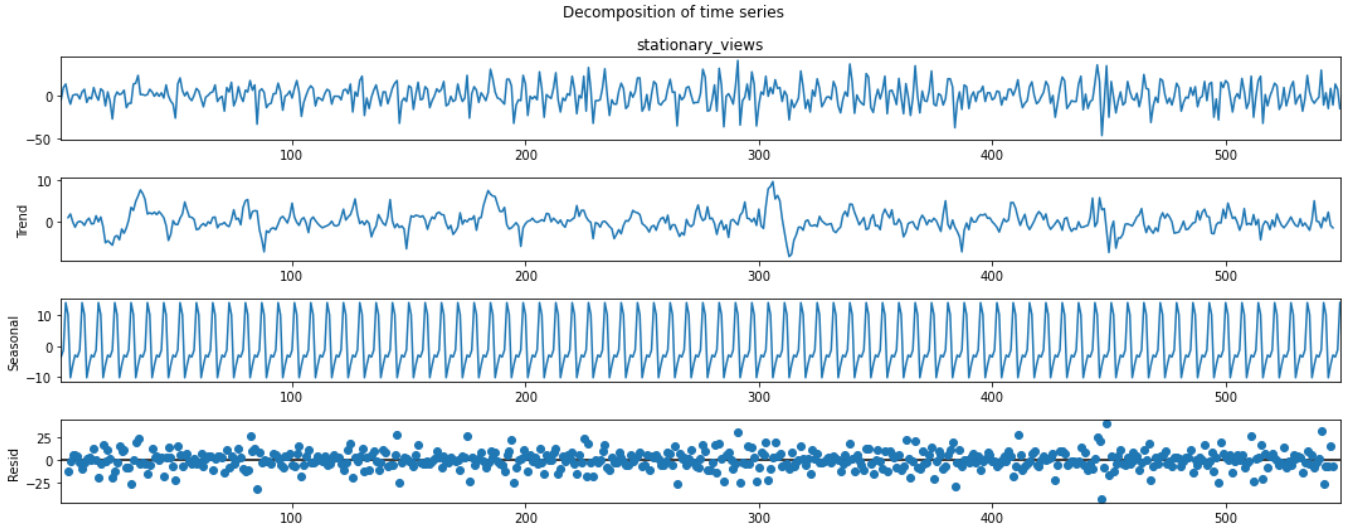

Actual vs Predicted
Train
Test
Predicted
views
date
300
250
200
150
100
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01


Original views
views
300
250
200
150


Stationary views
stationary_views
date
25
0
−25
−50
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01


Autocorrelation
1.0
0.5
0.0
−0.5
−1.0


Partial Autocorrelation
1.0
0.5
0.0
−0.5
−1.0
0
20
40
60
80
100
120




```
100%|██████████| 4/4 [00:03<00:00,  1.25it/s]
```

********** Best model AIC is 3025.5207187681995 found with order (1, 1, 1) and seasonal order (1, 0, 1, 7) and exog False**********

SARIMAX Results

Dep. Variable:	views	No. Observations:	411
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 7)	Log Likelihood	-1507.760
Date:	Wed, 05 Apr 2023	AIC	3025.521
Time:	23:48:04	BIC	3045.602
Sample:	0	HQIC	3033.465
	- 411		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7890	0.034	23.193	0.000	0.722	0.856
ma.L1	-0.9814	0.012	-80.410	0.000	-1.005	-0.957
ar.S.L7	0.9822	0.012	83.235	0.000	0.959	1.005
ma.S.L7	-0.8539	0.039	-22.128	0.000	-0.930	-0.778
sigma2	89.9767	5.435	16.555	0.000	79.324	100.629

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	10.04
Prob(Q):	0.98	Prob(JB):	0.01
Heteroskedasticity (H):	1.46	Skew:	0.08
Prob(H) (two-sided):	0.03	Kurtosis:	3.75

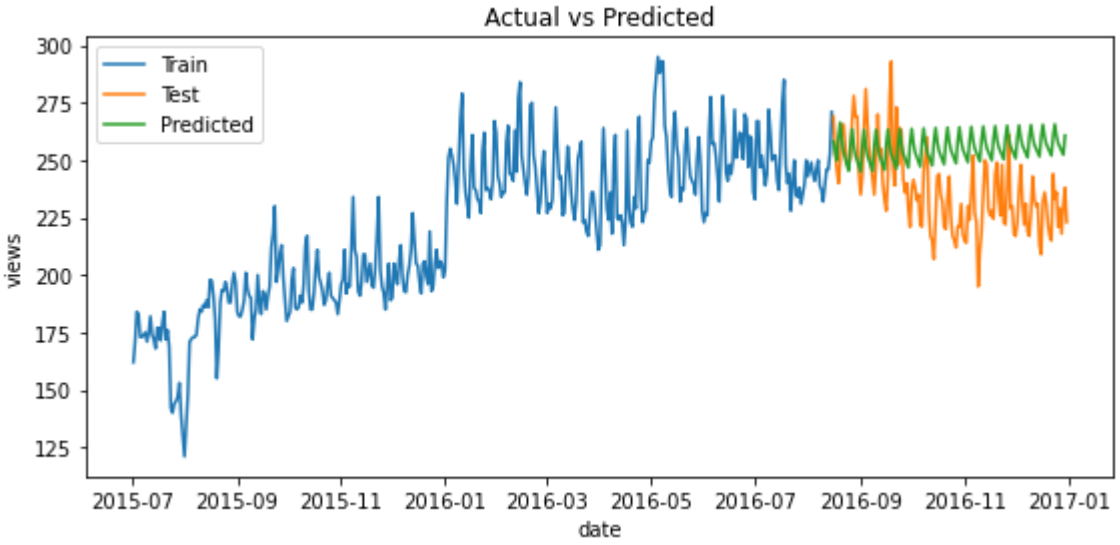
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

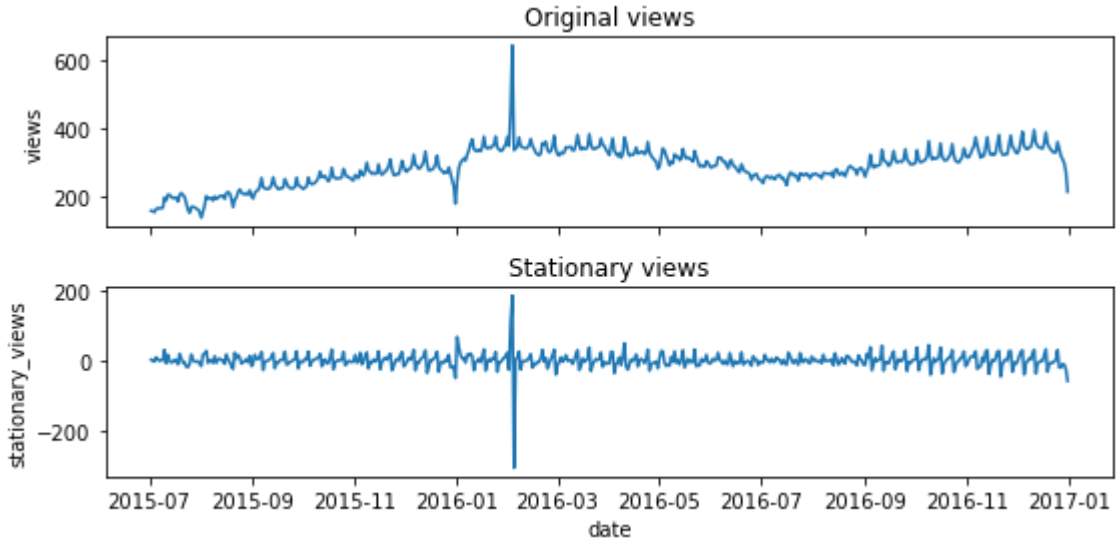
********** Performance Metrics **********

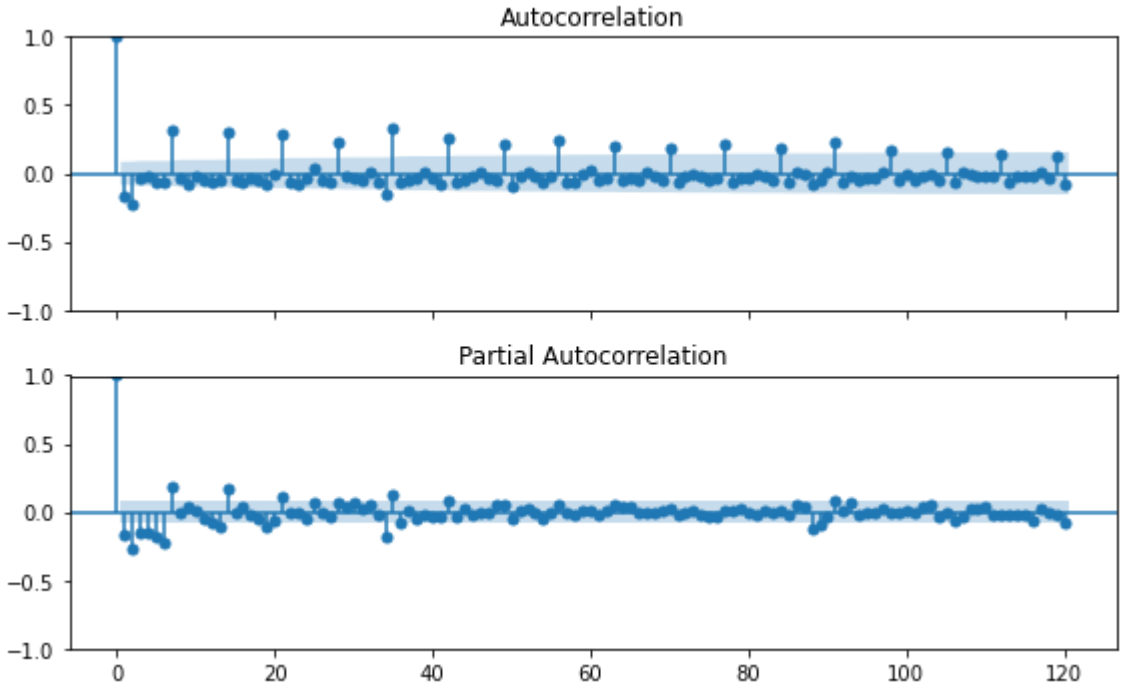
RMSE : 24.54
MAPE: 9.35 %

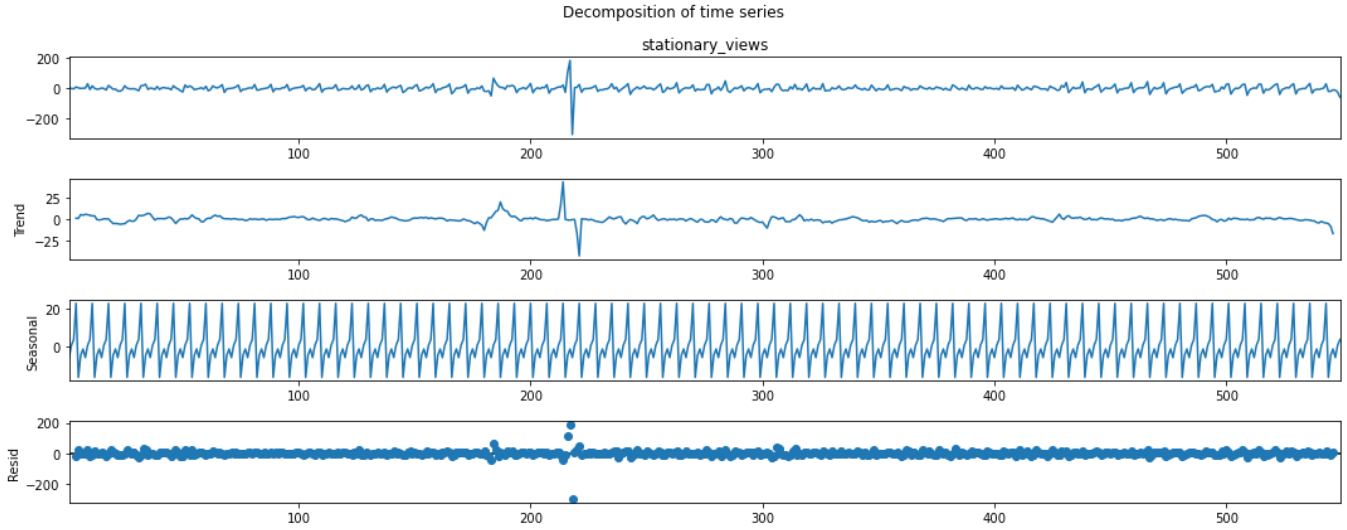
********** Language : ru **********

********** Stationary views obtained after 1 differencing! **********


Actual vs Predicted
Train
Test
Predicted
views
date


Original views
views
Stationary views
stationary_views
date


Autocorrelation
Partial Autocorrelation




```
100%|██████████| 4/4 [00:03<00:00,  1.13it/s]
```

********** Best model AIC is 3616.6536667150995 found with order (1, 1, 1) and seasonal order (1, 0, 1, 7) and exog False**********

SARIMAX Results

Dep. Variable:	views	No. Observations:	411
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 7)	Log Likelihood	-1803.327
Date:	Wed, 05 Apr 2023	AIC	3616.654
Time:	23:48:10	BIC	3636.734
Sample:	0	HQIC	3624.598
	- 411		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4848	0.049	9.818	0.000	0.388	0.582
ma.L1	-0.8852	0.041	-21.615	0.000	-0.965	-0.805
ar.S.L7	0.9693	0.032	29.958	0.000	0.906	1.033
ma.S.L7	-0.8836	0.057	-15.631	0.000	-0.994	-0.773
sigma2	383.7163	4.608	83.270	0.000	374.685	392.748

Ljung-Box (L1) (Q):	0.53	Jarque-Bera (JB):	80704.78
Prob(Q):	0.47	Prob(JB):	0.00
Heteroskedasticity (H):	0.99	Skew:	2.04
Prob(H) (two-sided):	0.96	Kurtosis:	71.61

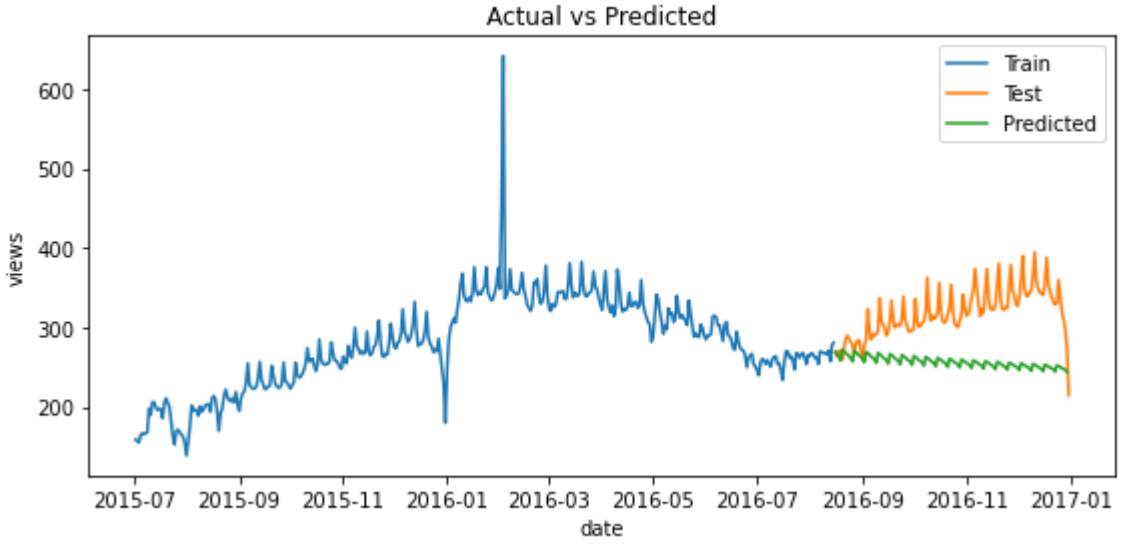
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

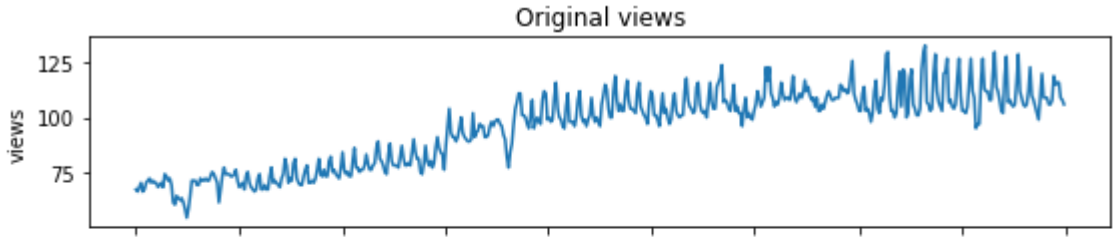
********** Performance Metrics **********

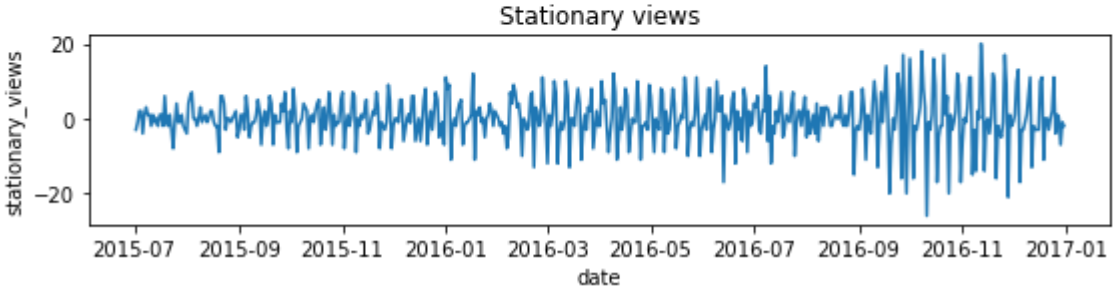
RMSE : 70.01
MAPE: 18.75 %

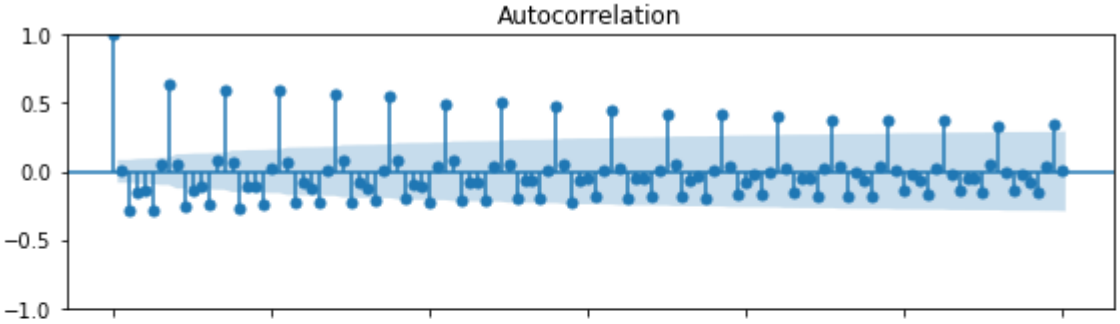
********** Language : zh **********

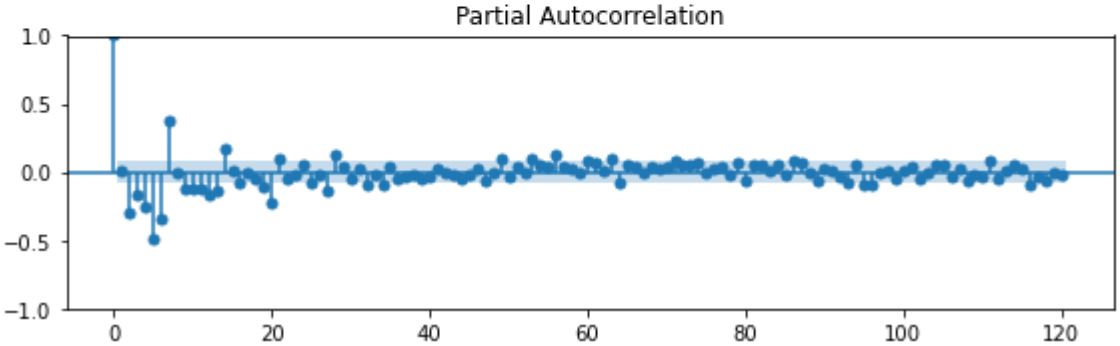
********** Stationary views obtained after 1 differencing! **********

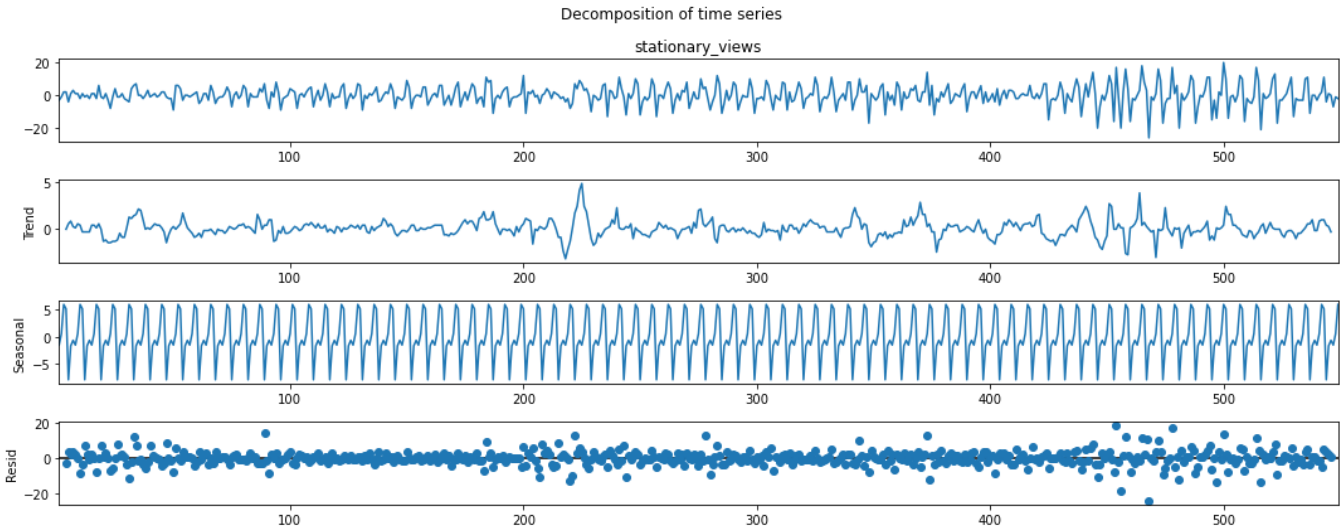

Actual vs Predicted
Train
Test
Predicted
views
600
500
400
300
200
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01
date


Original views
views
125
100
75


Stationary views
stationary_views
20
0
−20
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01
date


Autocorrelation
1.0
0.5
0.0
−0.5
−1.0


Partial Autocorrelation
1.0
0.5
0.0
−0.5
−1.0
0
20
40
60
80
100
120



100%|██████████| 4/4 [00:01<00:00, 2.58it/s]

********** Best model AIC is 2159.698775215993 found with order (1, 1, 1) and seasonal order (1, 0, 1, 7) and exog False**********

SARIMAX Results

Dep. Variable:	views	No. Observations:	411
Model:	SARIMAX(1, 1, 1)x(1, 0, 1, 7)	Log Likelihood	-1074.849
Date:	Wed, 05 Apr 2023	AIC	2159.699
Time:	23:48:14	BIC	2179.780
Sample:	0 - 411	HQIC	2167.643
Covariance Type:	opg		

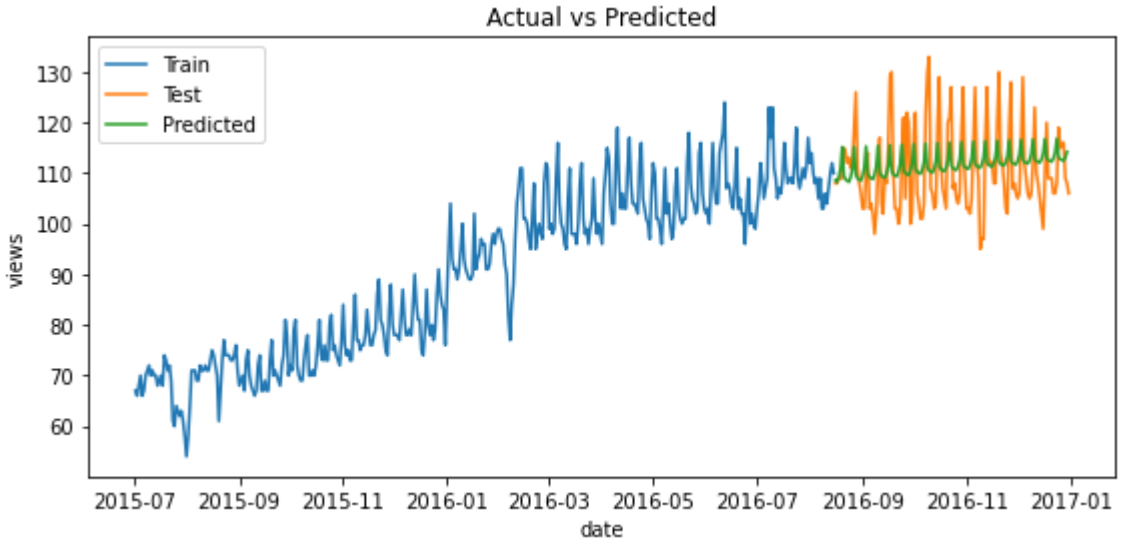
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7194	0.029	25.178	0.000	0.663	0.775
ma.L1	-0.9757	0.012	-84.338	0.000	-0.998	-0.953
ar.S.L7	0.9756	0.012	80.032	0.000	0.952	0.999
ma.S.L7	-0.7845	0.036	-21.837	0.000	-0.855	-0.714
sigma2	10.8751	0.535	20.344	0.000	9.827	11.923

Ljung-Box (L1) (Q):	1.32	Jarque-Bera (JB):	116.21
Prob(Q):	0.25	Prob(JB):	0.00
Heteroskedasticity (H):	1.52	Skew:	0.21
Prob(H) (two-sided):	0.01	Kurtosis:	5.57

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

********** Performance Metrics **********

RMSE : 7.19
MAPE: 5.35 %



Questionnare

Defining the problem statements and where can this and modifications of this be used?

- We have to predict the number of views for a particular language for a particular day.
- We can modify the frequency of the data to be daily or weekly or monthly.

Write 3 inferences you made from the data visualizations

- English is the most popular language ~16%
- The most common type of access is all-access ~ 50%
- all-agents (~75%) is more common access origin than spider
- Top 5 web pages are : FB, YT, Special Search, Google, Iphone

What does the decomposition of series do?

- Decomposition breaks a time series into trend, seasonality and noise

What level of differencing gave you a stationary series?

- Difference -> 1

Difference between arima, sarima & sarimax

- Sarima includes seasonality for AR and MA methods
- Sarimax can include exogenous variables

Compare the number of views in different languages

```
de    129.270909
en    336.082915
es    387.381480
fr    135.845036
ja    224.789091
ru    288.667829
zh     95.338874
dtype: float64
```

What other methods other than grid search would be suitable to get the model for all languages?

- We can use grid search or use auto arima