گزارش سری زمانی

AUTHOR
امیرمحمد عبداله پور

خلاصه ای از مسیر

کارهایی که با داده انجام شد

- ساختن یک ستون جدید به عنوان قیمت کل، با ضرب کردن مقدار کمیت با قیمت واحد
- داده ها از روزانه به جمع درآمد در روز، هفته و ماه تبدیل شد
- حذف کشورهای به غیر از انگلیس به دلیل اینکه تقریبا ۹۰ درصد داده ها از انگلیس آمده اند و استفاده از بقیه کشورها منطقی نیست، چون اطلاعات کافی راجع به فروش در آن کشور ها نداریم.
- حذف داده هایی که شماره خرید آنها با حرف انگلیسی C شروع می شود چون این فروش ها کنسل شده اند و درآمد از آنها نباید در فروش کل قرار بگیرد
- این داده ها در دو سری به وجود آمده اند که اولین سری از سال ۲۰۰۹ تا ۲۰۱۰ بود و سری دوم از سال ۲۰۱۰ تا ۲۰۱۱ بود که براساس زمان این ها را تبدیل به یک دیتاست کردم و به بررسی دو سال از فروش این فروشگاه می پردازیم
- در تاریخ نوامبر ۲۰۱۰ تا مارچ ۲۰۱۱ ترند به سمت پایین آمد به دلیل اینکه اکه به تاریخ در آن زمان در انگلستان نگاه کنیم متوجه اعتراضات دانشجویی می شویم
- در این فروشگاه، آخرهفته ها روز فروش نیست و درآمد به صفر میرسه و به این دلیل آخرهفته ها را از داده پاک میکنیم
- در داده یک سری قیمت منفی داشتیم که نشان دهنده، پس دادن کالا توسط مشتری ها بود که برای یکم ساده تر شدن مدل از داده حذف شد

تست هایی که گرفته شد

- **Dicky-fuller (ADF):**
  - برای نشان دادن مانا بودن، اگر کوچک بود نشان دهنده مانایی هست و اگه از ۰.۰۵ بزرگتر باشه مانا نیست
- **Runs-test:**
  - برای سنجش پایانی باقی مانده ها استفاده میشه که تصادفی بودن باقی مانده ها را تایید یا رد میکنه
- **کلموگیروف-اسمینیروف**
  - نرمال بودن یا نبودن بافی مانده را مورد سنجش میزاریم

چرا قیمت‌ها صفر هستند؟

در این تاریخ‌ها بازارهای مالی بریتانیا (UK) تعطیل بوده‌اند و هیچ معامله‌ای انجام نشده، بنابراین قیمت‌ها به صورت 0.0 ثبت شده‌اند.

کریسمس (تعطیلات پایان سال)

بازارهای بریتانیا در دوره کریسمس معمولاً چند روز بسته هستند:

- 2009-12-24 تا 31-12-2009
- 2010-12-24 تا 31-12-2010

روز سال نو

- 2010-01-01
- 2011-01-03 (تعطیلی جایگزین برای New Year’s Day)

تعطیلات عید پاک (Easter Holidays)

این تعطیلات شامل دو روز رسمی است: Good Friday و Easter Monday

- 2010-04-02 (جمعه نیک – Good Friday)
- 2010-04-05 (دوشنبه عید پاک – Easter Monday)
- 2011-04-22 (جمعه نیک)
- 2011-04-25 (دوشنبه عید پاک)

Early May Bank Holiday

یک تعطیلی رسمی که اولین دوشنبه ماه می برگزار می‌شود:

- 2010-05-03
- 2011-05-02

Spring Bank Holiday

یک تعطیلی رسمی که آخرین دوشنبه ماه می برگزار می‌شود:

- 2010-05-31
- 2011-05-30

Summer Bank Holiday

یک تعطیلی رسمی که آخرین دوشنبه ماه آگوست برگزار می‌شود:

- 2010-08-30
- 2011-08-29

عروسی سلطنتی

در این روز شاهزاده ویلیام و کاترین میدلتون ازدواج کردند و دولت بریتانیا این روز را تعطیل رسمی ملی اعلام کرد - 29-04-2011

ساختار داده ها

- Country: کشورهای که فروش در آنجا انجام شده به دلیل اینکه انگلیس اکثر داده ما را تشکیل می داد از بقیه کشورها صرف نظر شد و به تحلیل فقط میزان فروش در انگلیس پرداخته شد.
- Customer ID : یک کد مختص به یک مشتری
- **Price**: قیمت هر واحد از کالا
- **Quantity**: تعداد خرید مشتری از یک نوع کالا
- Description: توضیح راجب کالا
- StockCode: کد مختص به سهام کالا
- Invoice: شماره خرید
- **InvoiceDate**: زمان خرید
- **Holiday:** نشان دادن اینکه در آن روز به خصوص آیا تعطیل هست یا خیر

پیش پردازش داده ها

```
<'class 'pandas.core.frame.DataFrame>
DatetimeIndex: 1067371 entries, 2009-12-01 07:45:00 to 2011-12-09 12:50:00
:Data columns (total 7 columns)
 Column       Non-Null Count    Dtype   #
 -----    --------------       ------  ---
 Invoice      1067371 non-null  object   0
 StockCode    1067371 non-null  object   1
 Description  1062989 non-null  object   2
 Quantity     1067371 non-null  int64    3
Price        1067371 non-null  float64   4
Customer ID  824364 non-null   float64   5
 Country      1067371 non-null  object   6
   dtypes: float64(2), int64(1), object(4)
                memory usage: 65.1+ MB
```

	Quantity	Price	Customer ID
count	1.067371e+06	1.067371e+06	824364.000000
mean	9.938898e+00	4.649388e+00	15324.638504
std	1.727058e+02	1.235531e+02	1697.464450
min	8.099500e+04-	5.359436e+04-	12346.000000
25%	1.000000e+00	1.250000e+00	13975.000000
50%	3.000000e+00	2.100000e+00	15255.000000
75%	1.000000e+01	4.150000e+00	16797.000000
max	8.099500e+04	3.897000e+04	18287.000000

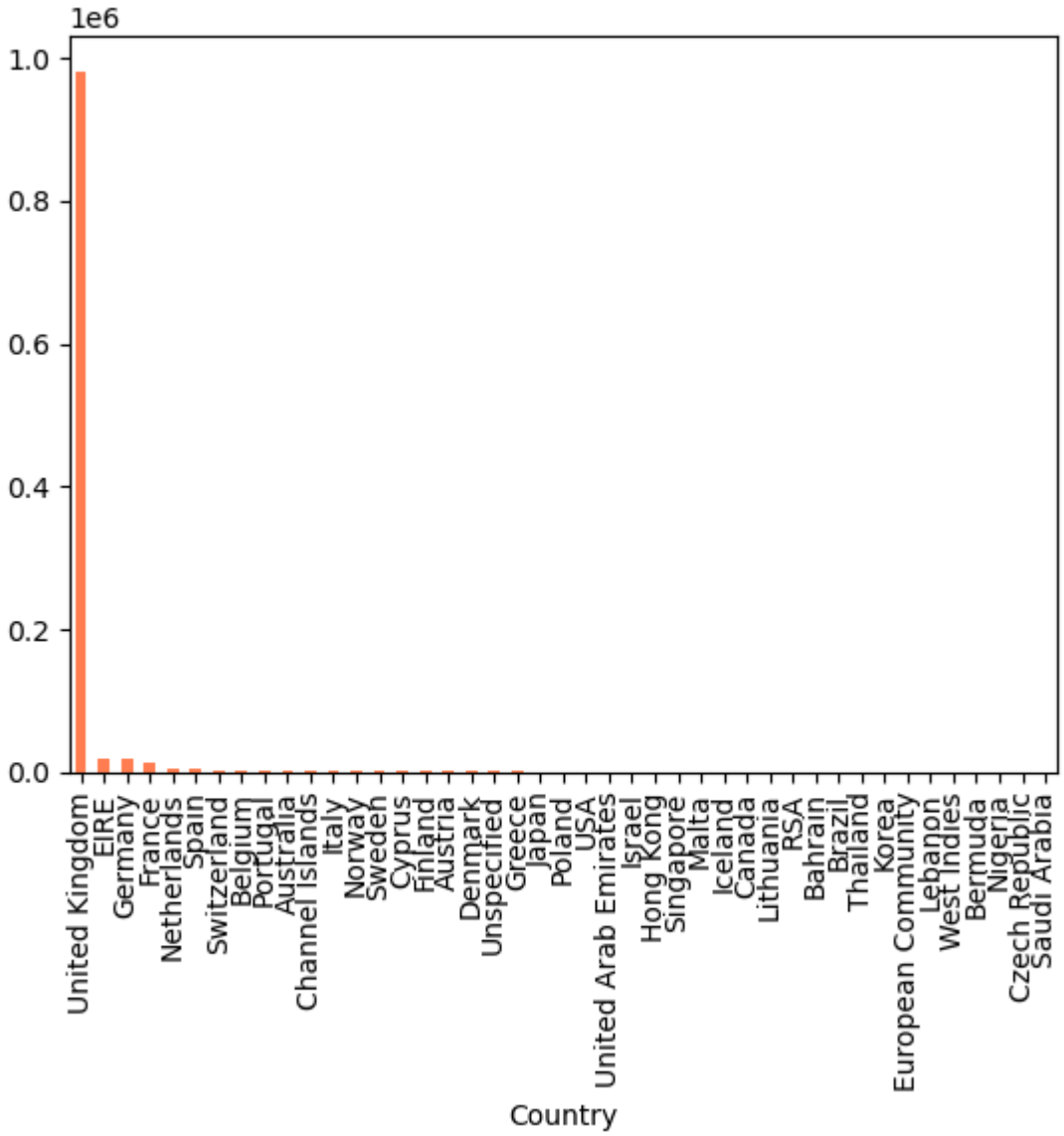
تعداد داده های ناموجود در هر ستون

	na_count	na_proportion
Customer ID	243007	22.77
Description	4382	0.41

	na_count	na_proportion
Invoice	0	0.00
StockCode	0	0.00
Quantity	0	0.00
Price	0	0.00
Country	0	0.00

حذف کشورها

همینطور که در نمودار زیر می بینیم اکثر داده های ما از کشور انگلیس میاد و کار درست این هست که فقط داده های فروش در این کشور مورد بررسی قرار بگیره



Country	Customer ID	Price	Quantity	Description	StockCode	Invoice	InvoiceDate
United Kingdom	13085.0	6.95	12	15CM CHRISTMAS GLASS BALL 20 LIGHTS	85048	489434	**2009-12-01 07:45:00**

InvoiceDate	Invoice	StockCode	Description	Quantity	Price	Customer ID	Country
2009-12-01 07:45:00	489434	79323P	PINK CHERRY LIGHTS	12	6.75	13085.0	United Kingdom
2009-12-01 07:45:00	489434	79323W	WHITE CHERRY LIGHTS	12	6.75	13085.0	United Kingdom
2009-12-01 07:45:00	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2.10	13085.0	United Kingdom
2009-12-01 07:45:00	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	1.25	13085.0	United Kingdom

حذف بازگشت های عجیب و غریب

در اینجا با عدد های خیلی زیادی برخورد می کنیم که نشان دهنده درستی از بازگشت کالا توسط مشتری نیست! به همین دلیل میایم و بازگشت های بالاتر از ۵۰ تا را حذف میکنیم

```
,array([   -3,    -2,    -6,    -1,   -12,    -4,   -24,   -81
,246-   ,252-   ,600-   ,504-   ,9-     ,8-     ,5-     ,48-
,16-    ,50-    ,96-    ,18-    ,60-    ,32-    ,10-    ,36-
,20-    ,25-    ,13-    ,23-    ,7-     ,168-   ,64-    ,120-
,14-    ,30-    ,72-    ,190-   ,144-   ,17-    ,100-   ,11-
,21-    ,300-   ,22-    ,26-    ,150-   ,27-    ,132-   ,44-
,80-    ,28-    ,108-   ,432-   ,15-    ,240-   ,192-   ,40-
,288-   ,34-    ,1440-  ,38-    ,408-   ,200-   ,42-    ,74-
,140-   ,280-   ,500-   ,400-   ,19-    ,309-   ,160-   ,204-
,49-    ,51-    ,312-   ,69-    ,1200-  ,156-   ,250-   ,128-
,576-   ,61-    ,29-    ,99-    ,98-    ,210-   ,830-   ,41-
,118-   ,180-   ,55-    ,90-    ,85-    ,720-   ,648-   ,320-
,56-    ,384-   ,1152-  ,700-   ,59-    ,31-    ,35-    ,324-
,213-   ,212-   ,256-   ,248-   ,2400-  ,243-   ,270-   ,148-
,117-   ,129-   ,65-    ,37-    ,45-    ,68-    ,82-    ,110-
,46-    ,102-   ,76-    ,94-    ,74215- ,47-    ,9360-  ,33-
,1350-  ,960-   ,3114-  ,2000-  ,1930-  ,1300-  ,670-   ,58-
,121-   ,75-    ,360-   ,164-   ,480-   ,420-   ,1515-  ,52-
,756-   ,113-   ,126-   ,318-   ,334-   ,701-   ,828-   ,39-
,220-   ,130-   ,70-    ,53-    ,186-   ,43-    ,79-    ,152-
,276-   ,67-    ,162-   ,112-   ,66-    ,468-   ,1296-  ,840-
                                ([80995- ,234-   ,244-   ,184-
```

حذف فروش هایی که کنسل شده اند که با حرف C اول ستون شماره خرید قرار گرفته شده.

دو ستون کمیت و قیمت دو ستون مهم برای ما هست تا با این دو ستون یک ستون جدید به نام قیمت کل بسازیم.

InvoiceDate	Quantity	Price	TotalPrice
07:45:00 2009-12-01	12	6.95	83.4

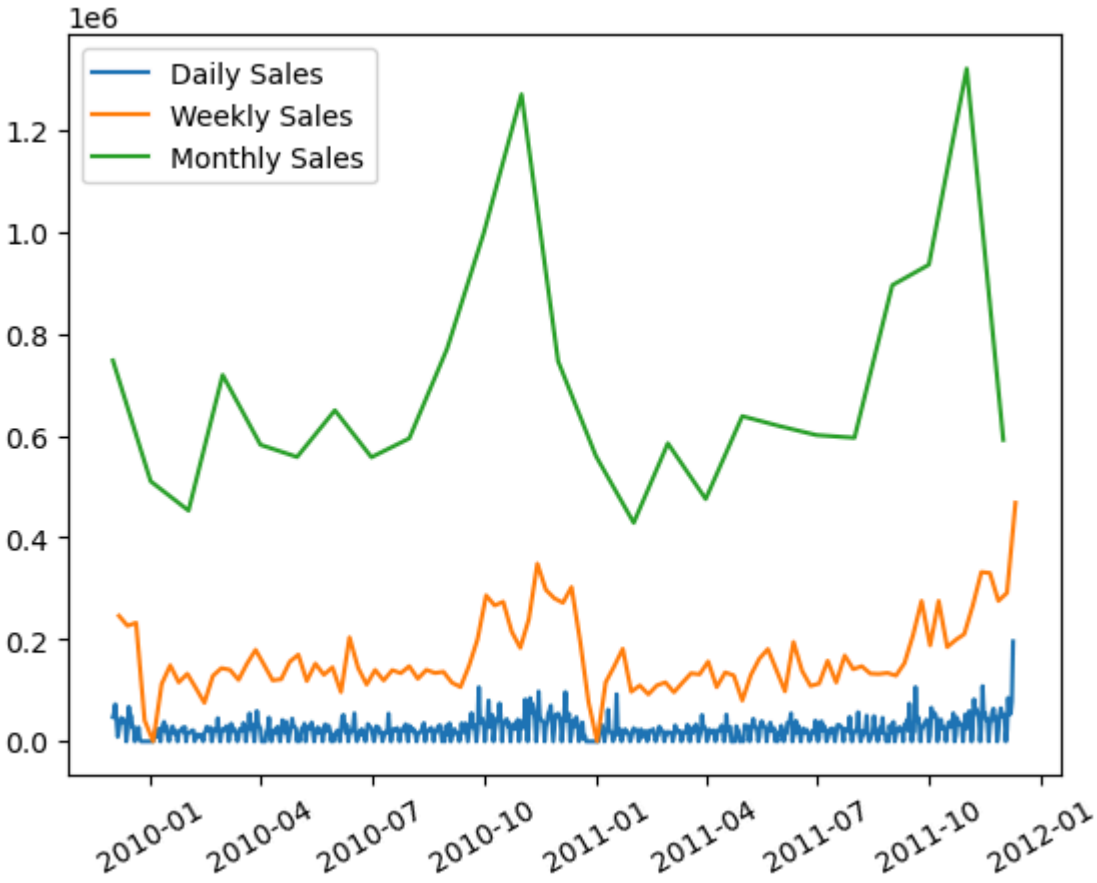
	TotalPrice	Price	Quantity
InvoiceDate			
07:45:00 2009-12-01	81.0	6.75	12
07:45:00 2009-12-01	81.0	6.75	12
07:45:00 2009-12-01	100.8	2.10	48
07:45:00 2009-12-01	30.0	1.25	24
...	...	...	...
12:31:00 2011-12-09	23.4	1.95	12
12:49:00 2011-12-09	23.6	2.95	8
12:49:00 2011-12-09	30.0	1.25	24
12:49:00 2011-12-09	214.8	8.95	24
12:49:00 2011-12-09	70.8	7.08	10

rows × 3 columns 930836

ستون قیمت کل که به عنوان متغیر مورد بررسی در این سری زمانی انتخاب می کنیم.

تغییر زمان داده ها

داده ها به صورت روزانه ، هفتگی و ماهانه در نمودار زیر قرار گرفته اند.



<Figure size 1000x1500 with 0 Axes>

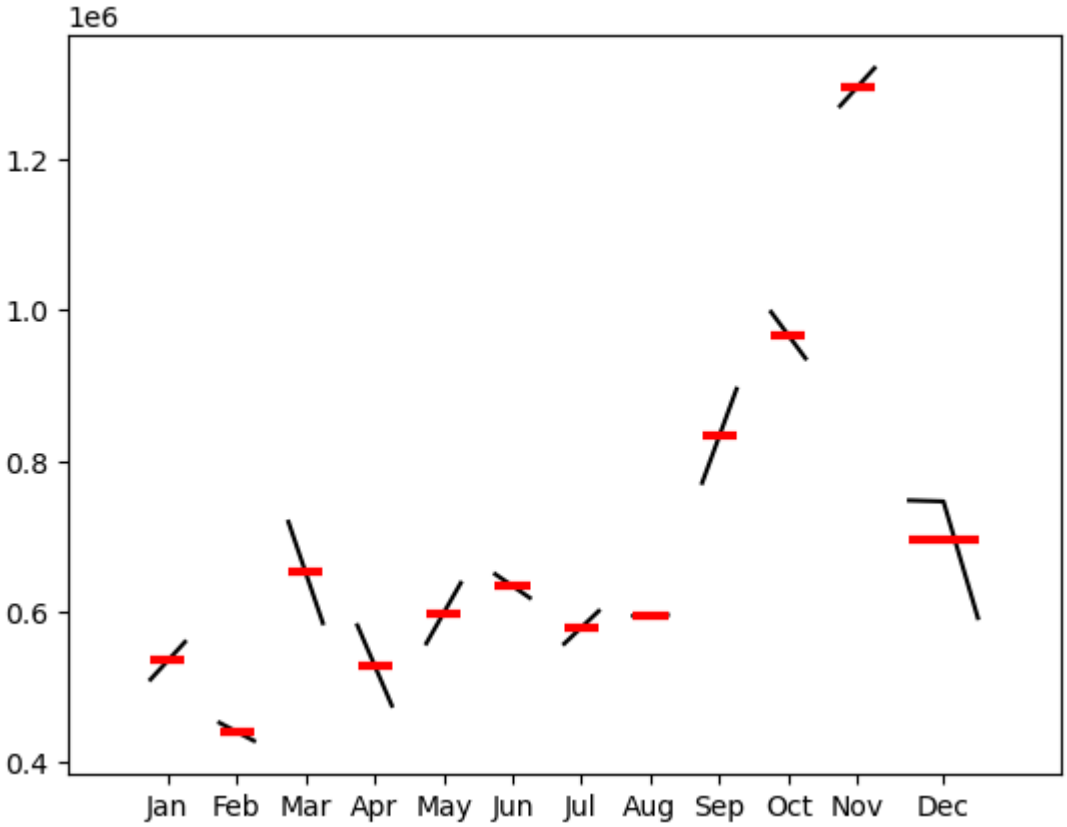
	InvoiceDate	TotalPrice
0	2009-12-01	47668.86
1	2009-12-02	55875.26
2	2009-12-03	72820.90
3	2009-12-04	37966.35
4	2009-12-05	9042.36
...	...	...
734	2011-12-05	85331.87
735	2011-12-06	52484.19
736	2011-12-07	55214.81
737	2011-12-08	78999.51
738	2011-12-09	196114.48

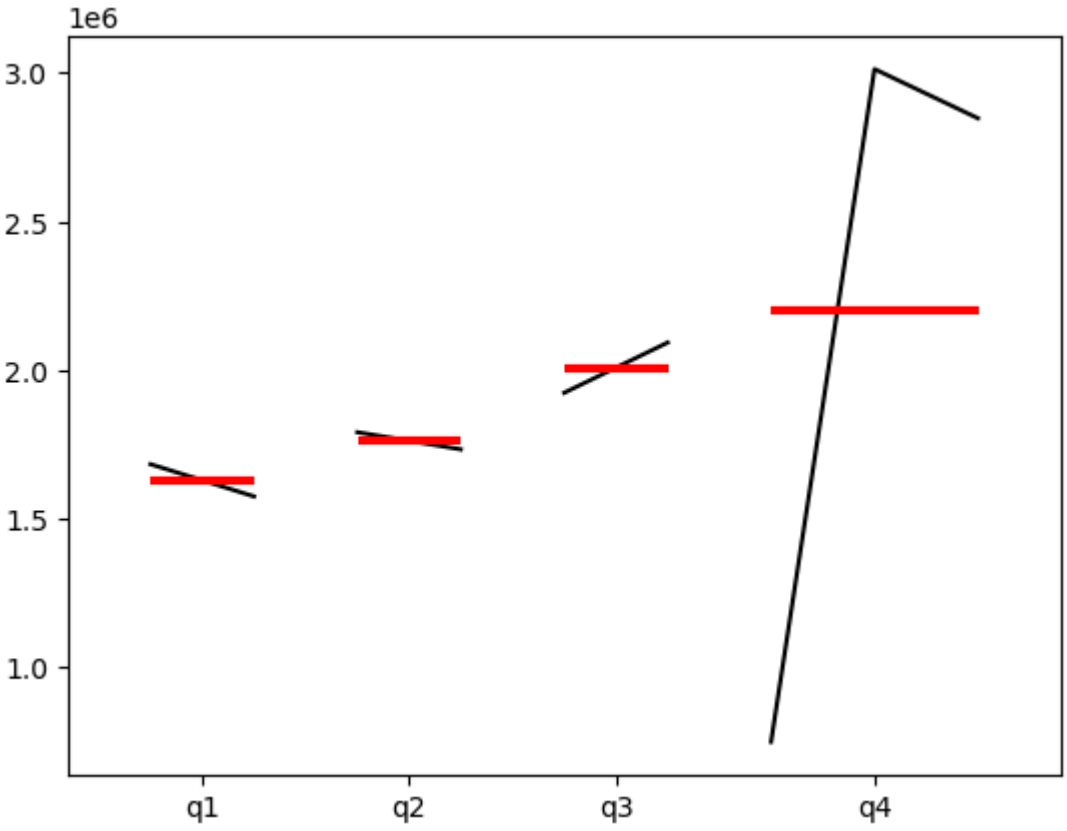
739 rows × 2 columns

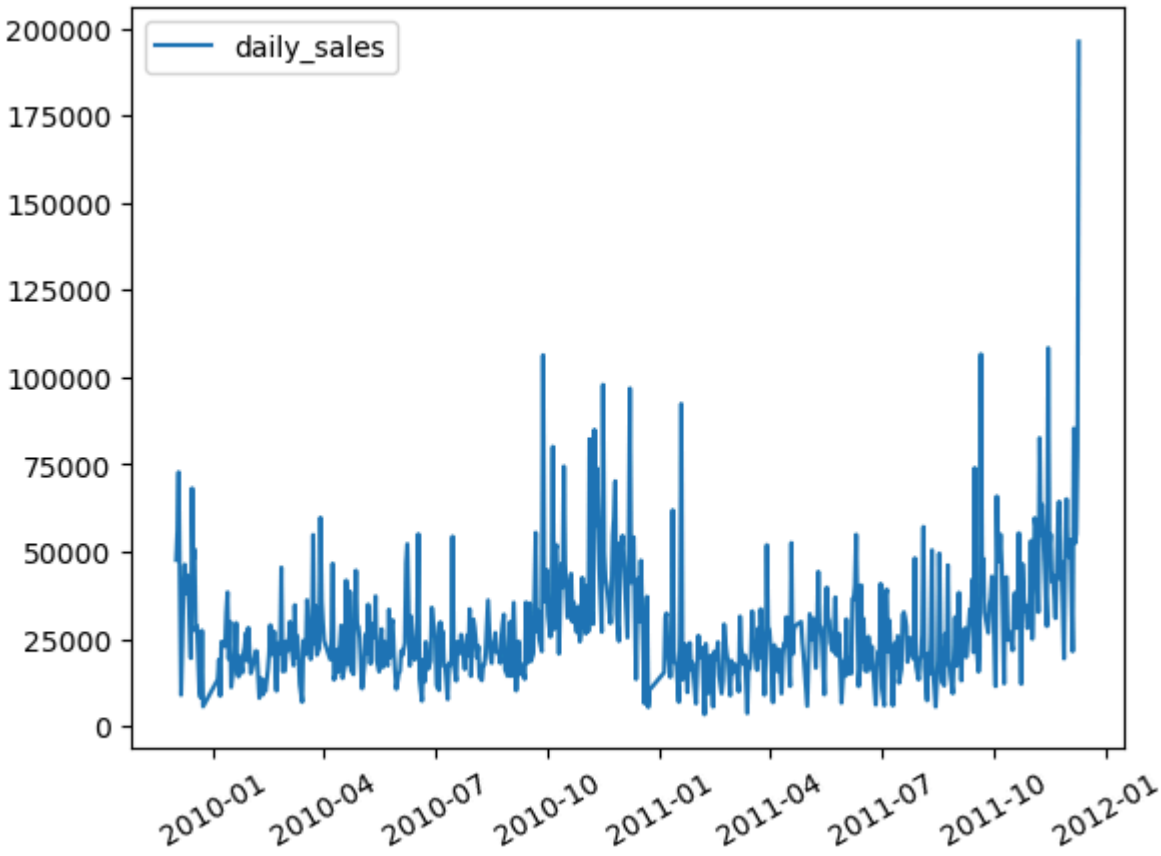
	InvoiceDate	TotalPrice
0	2009-12-01	47668.86
1	2009-12-02	55875.26
2	2009-12-03	72820.90
3	2009-12-04	37966.35
4	2009-12-05	9042.36
...	...	...
734	2011-12-05	85331.87
735	2011-12-06	52484.19
736	2011-12-07	55214.81
737	2011-12-08	78999.51
738	2011-12-09	196114.48

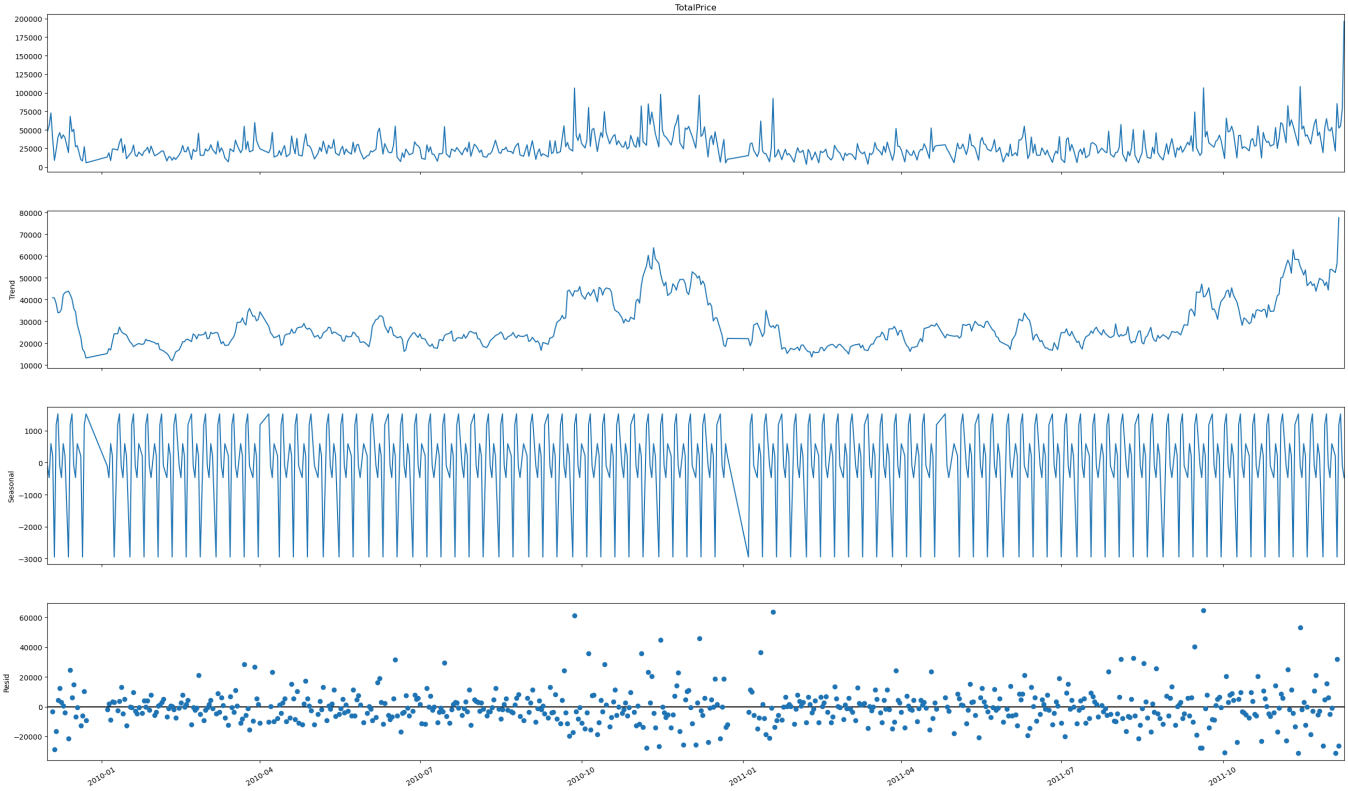
604 rows × 2 columns

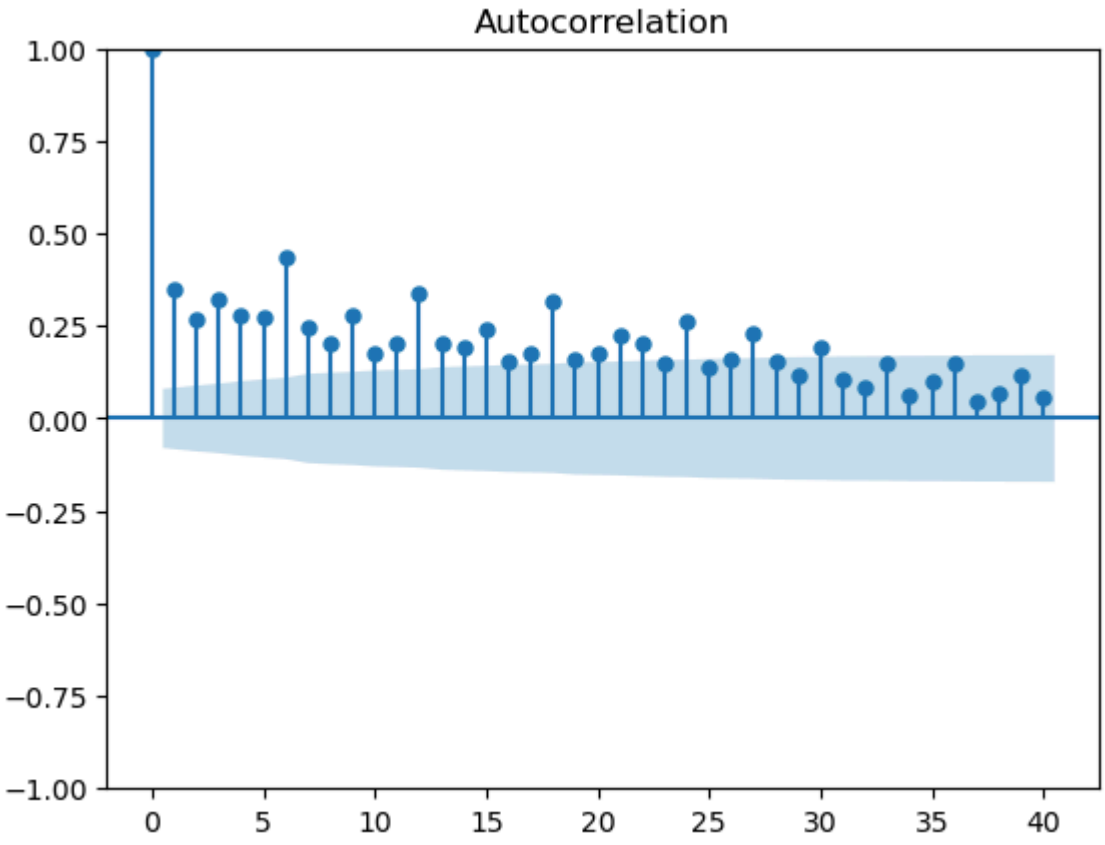
بررسی داده های روزانه

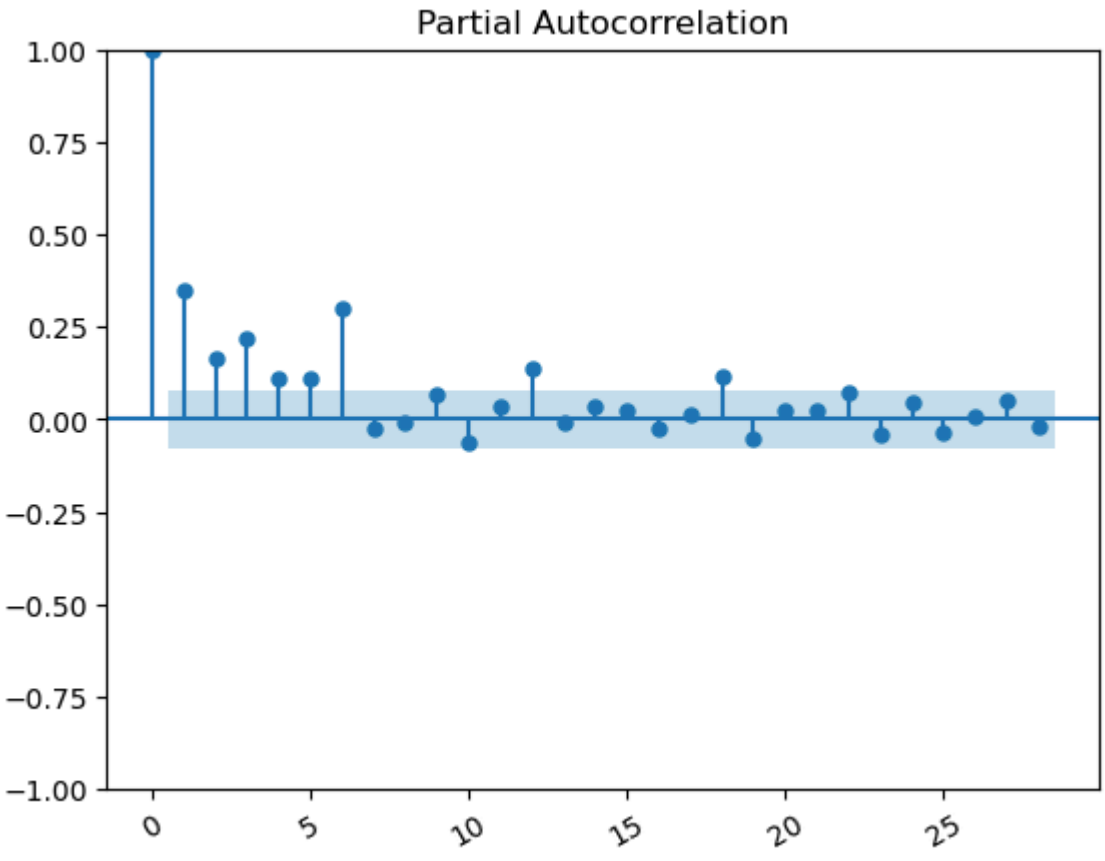

1e6
1.2
1.0
0.8
0.6
0.4
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec


1e6
3.0
2.5
2.0
1.5
1.0
q1
q2
q3
q4


daily_sales
200000
175000
150000
125000
100000
75000
50000
25000
0
2010-01
2010-04
2010-07
2010-10
2011-01
2011-04
2011-07
2011-10
2012-01


TotalPrice
Trend
Seasonal
Resid





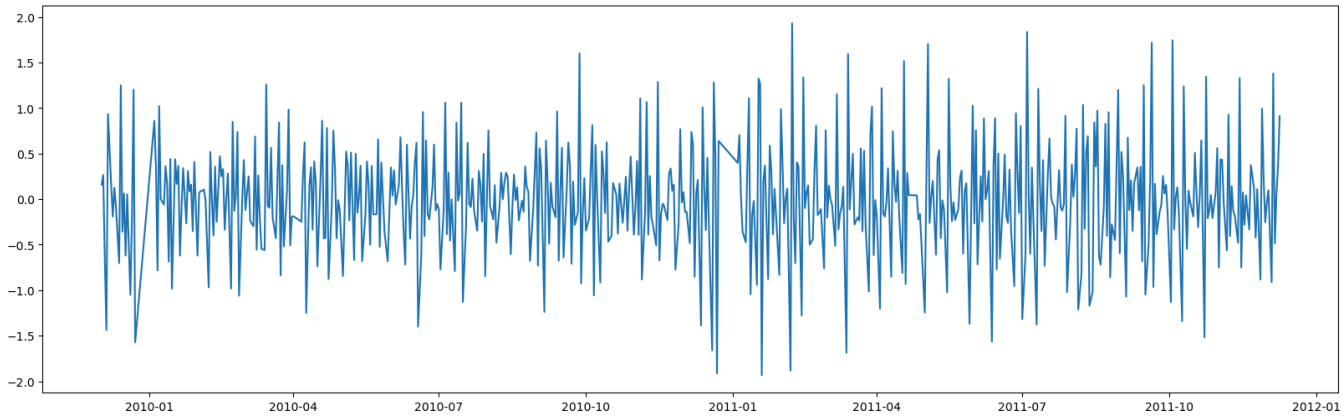
P_value : 0.8948919984918149

تست نشان میده که داده های مانا نیستند

InvoiceDate	TotalPrice
2009-12-01	10.772055
2009-12-02	10.930895
2009-12-03	11.195772
2009-12-04	10.544482
2009-12-05	9.109786
...	...
2011-12-05	11.354315
2011-12-06	10.868286
2011-12-07	10.919005
2011-12-08	11.277210
2011-12-09	12.186459

rows × 1 columns 604

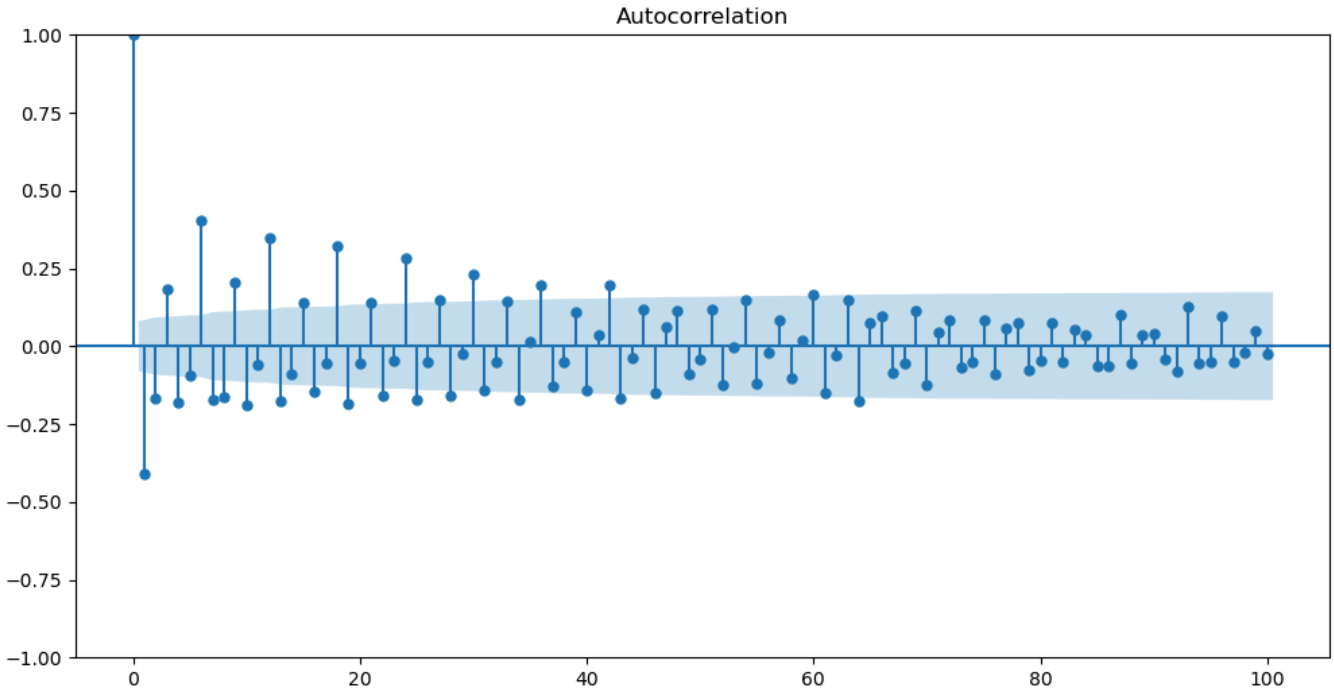
P_value : 0.4751320931105454
تست نشان میده که داده های مانا نیستند

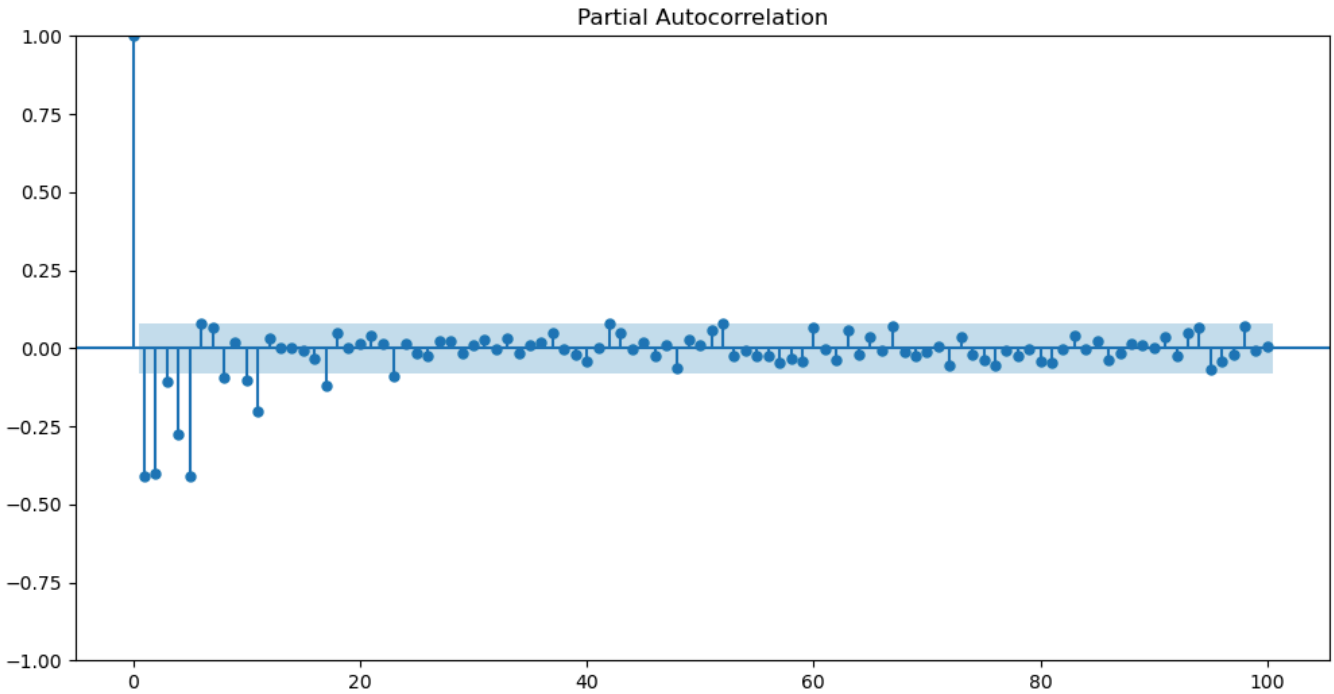


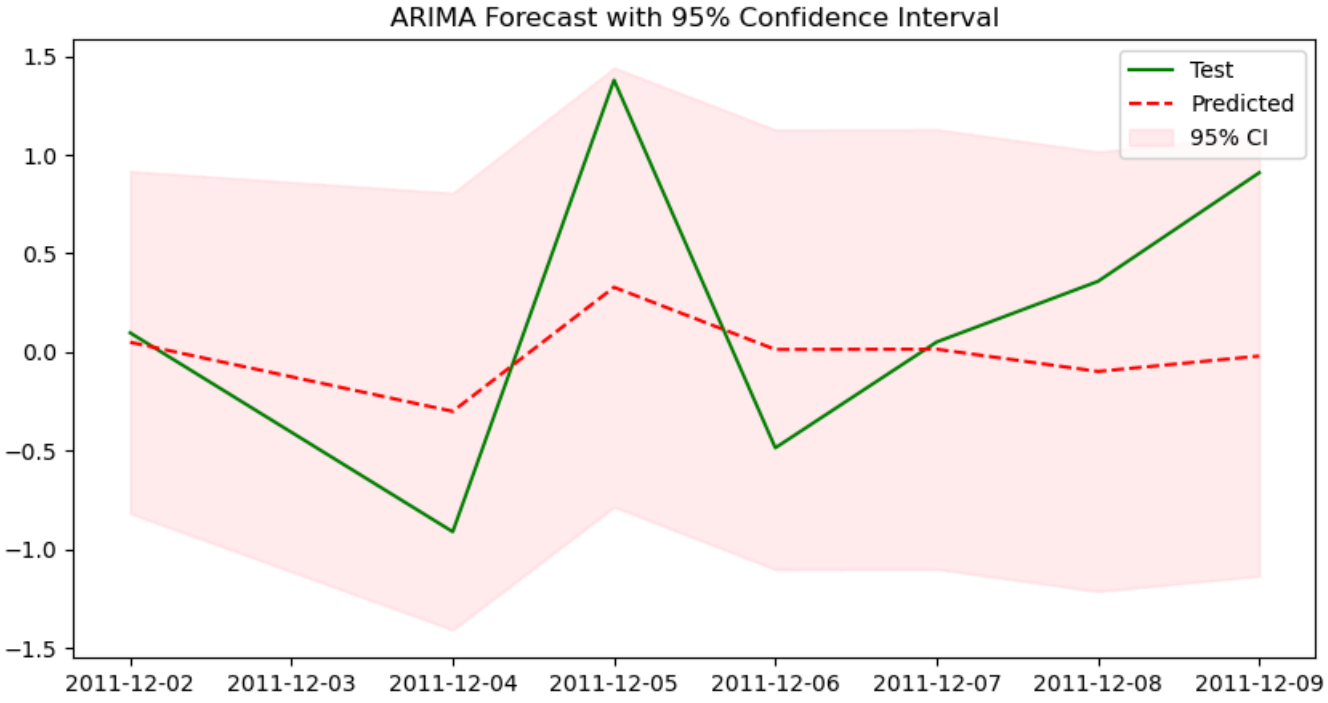
P_value : 4.577446411956347e-16
داده ها مانا هستند

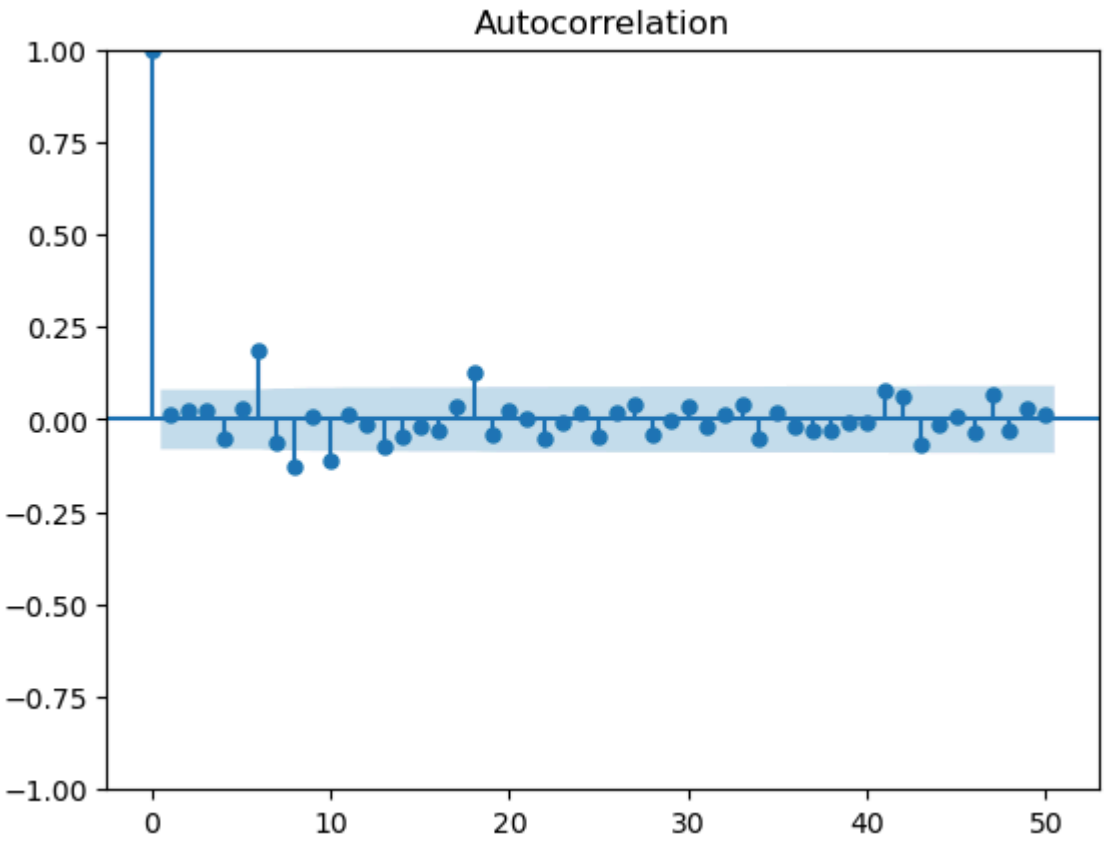
,0.24187989353131847)
,0.1
,52
({0.739 :'1%' ,0.574 :'2.5%' ,0.463 :'5%' ,0.347 :'10%'}

با توجه به دو آزمون گذشته سری زمانی ما، مانا هستند


Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
20
40
60
80
100


Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
20
40
60
80
100

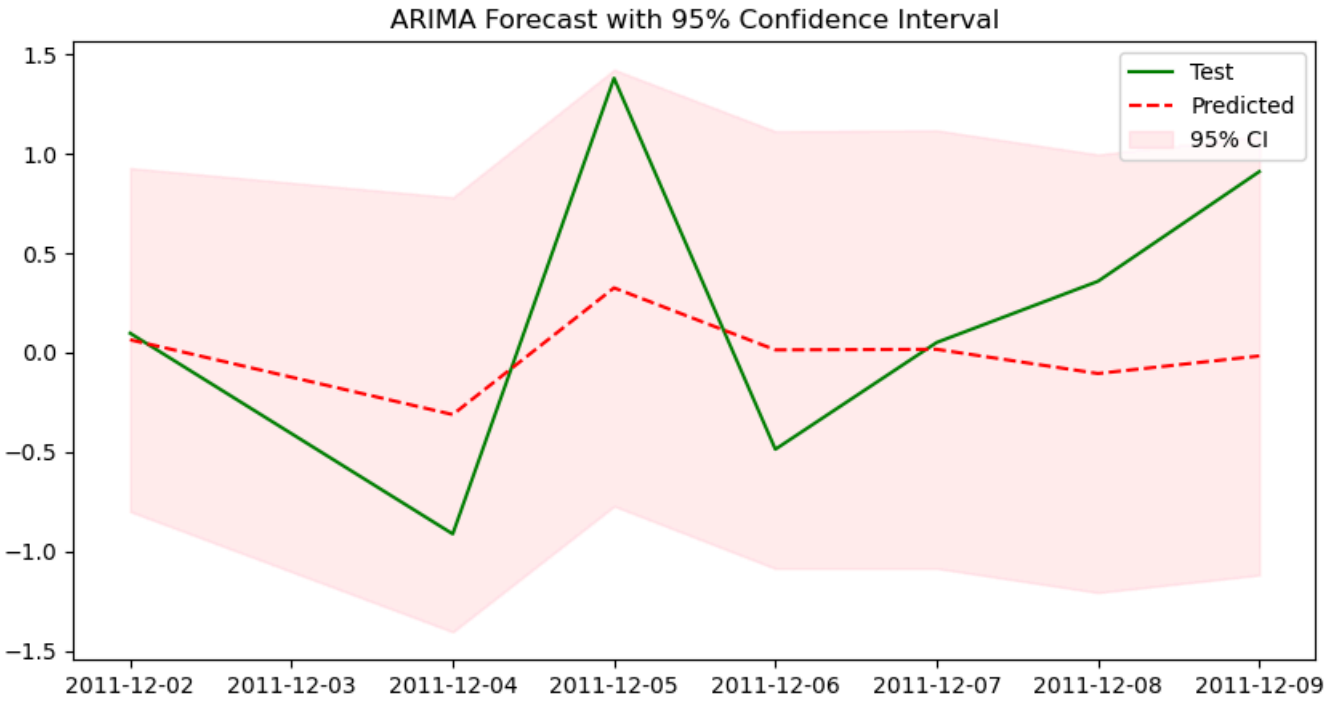


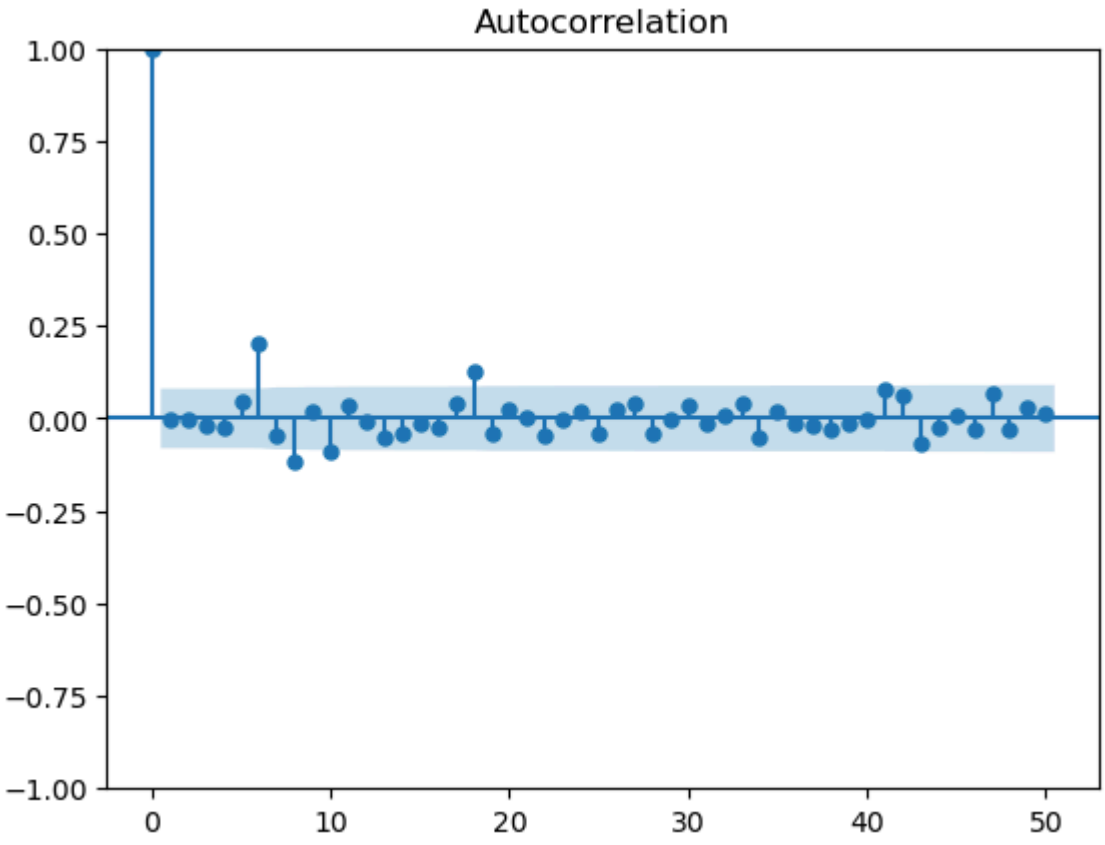


lb_pvalue	lb_stat	
0.715137	0.133200	**1**
0.808803	0.424399	**2**
0.854973	0.776945	**3**
0.667793	2.371447	**4**
0.722360	2.854798	**5**
0.000439	24.410788	**6**

	lb_stat	lb_pvalue
7	26.906732	0.000346
8	36.523471	0.000014
9	36.544620	0.000032
10	44.397507	0.000003

```
(0.8384036320090842 ,0.20393584678184795)
```

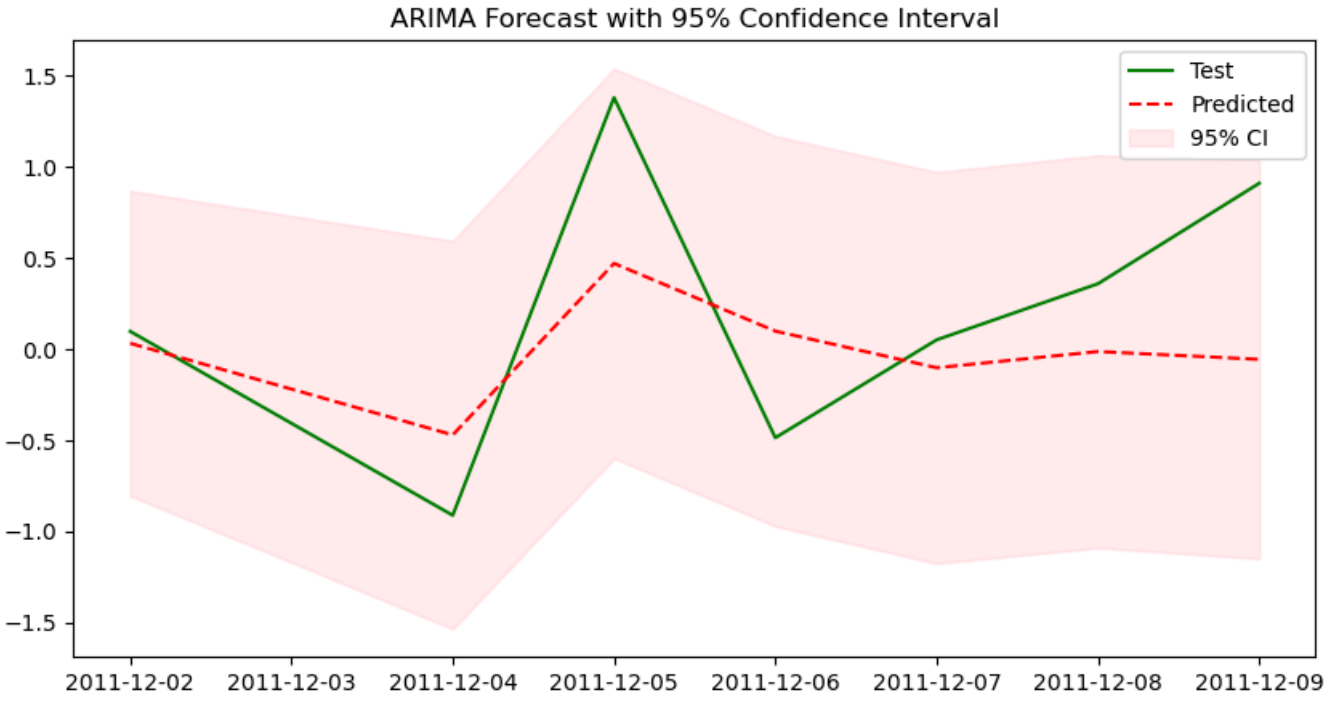




	lb_stat	lb_pvalue
1	0.005023	0.943498
2	0.008889	0.995566
3	0.289999	0.961898
4	0.720733	0.948748
5	1.953146	0.855589
6	27.462873	0.000119
7	28.646676	0.000168
8	37.134517	0.000011
9	37.334678	0.000023
10	42.063374	0.000007

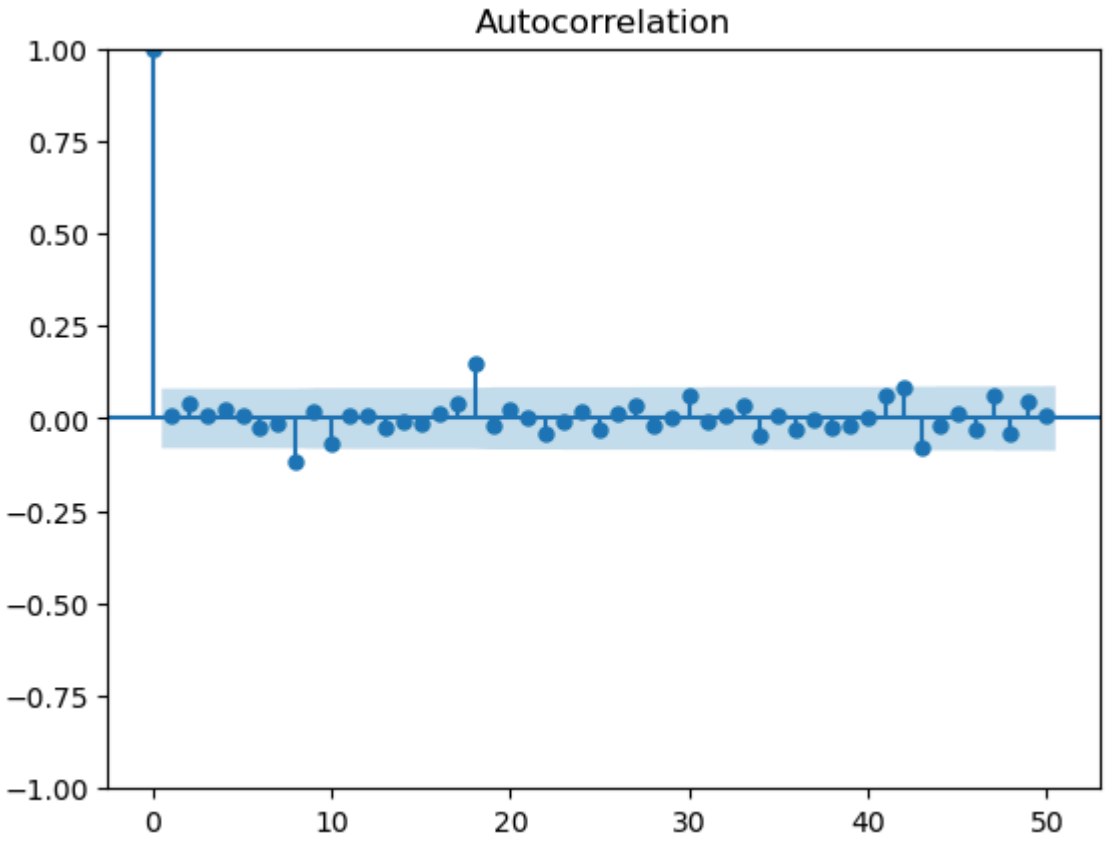
```
(0.7748311230208311 ,0.28606136345886335)
```

به دلیل اینکه در لگ ششم یک خودهمبستگی معنادار داریم از p=6 استفاده می کنیم

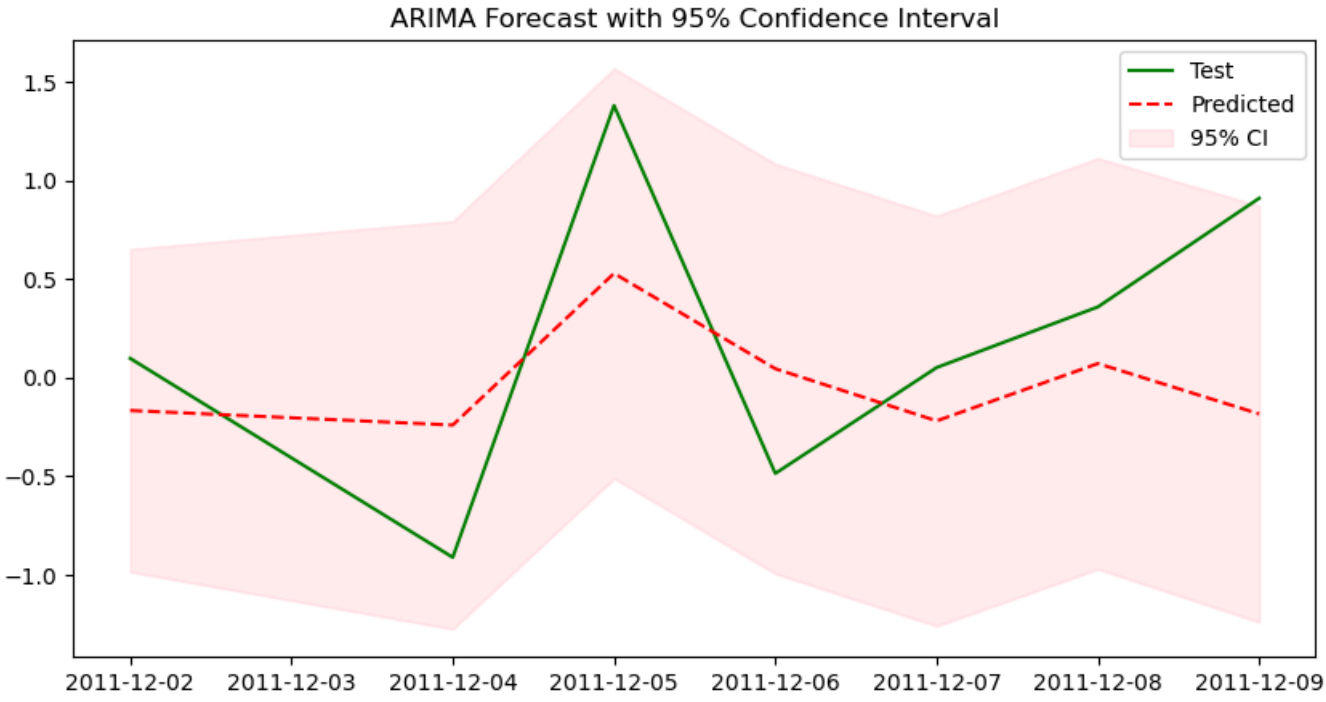



```
lb_stat  lb_pvalue
0.864403   0.029163    1
0.635926   0.905345    2
0.819544   0.924365    3
0.876270   1.211084    4
0.940610   1.243768    5
0.955102   1.564414    6
0.973752   1.718893    7
0.295876   9.578335    8
0.364381   9.830502    9
0.235203   12.797714   10
0.302727   12.857713   11
0.375724   12.908529   12
```

0.427698	13.263136	13
0.503921	13.288788	14
0.573466	13.373594	15
0.636423	13.493061	16
0.623725	14.607560	17
0.063305	27.918582	18
0.081529	28.099690	19
0.099839	28.419309	20
0.128567	28.422645	21
0.132448	29.450297	22
0.164229	29.502284	23
0.194049	29.724565	24
0.214523	30.264915	25
0.252502	30.371630	26
0.262528	31.211892	27
0.299937	31.392347	28
0.346835	31.399575	29
0.291459	33.737240	30
0.334373	33.785516	31
0.379670	33.819622	32
0.389729	34.636300	33
0.368935	36.138891	34
0.413039	36.185849	35
0.432790	36.775502	36
0.479434	36.776264	37
0.504952	37.228452	38
0.542814	37.402928	39
0.587126	37.417518	40
0.525380	39.766987	41
0.383003	44.096800	42
0.277143	48.004735	43
0.303197	48.303956	44
0.337281	48.402349	45
0.350903	49.074703	46
0.292038	51.800104	47
0.289743	52.926723	48
0.277470	54.367278	49
0.310532	54.403901	50

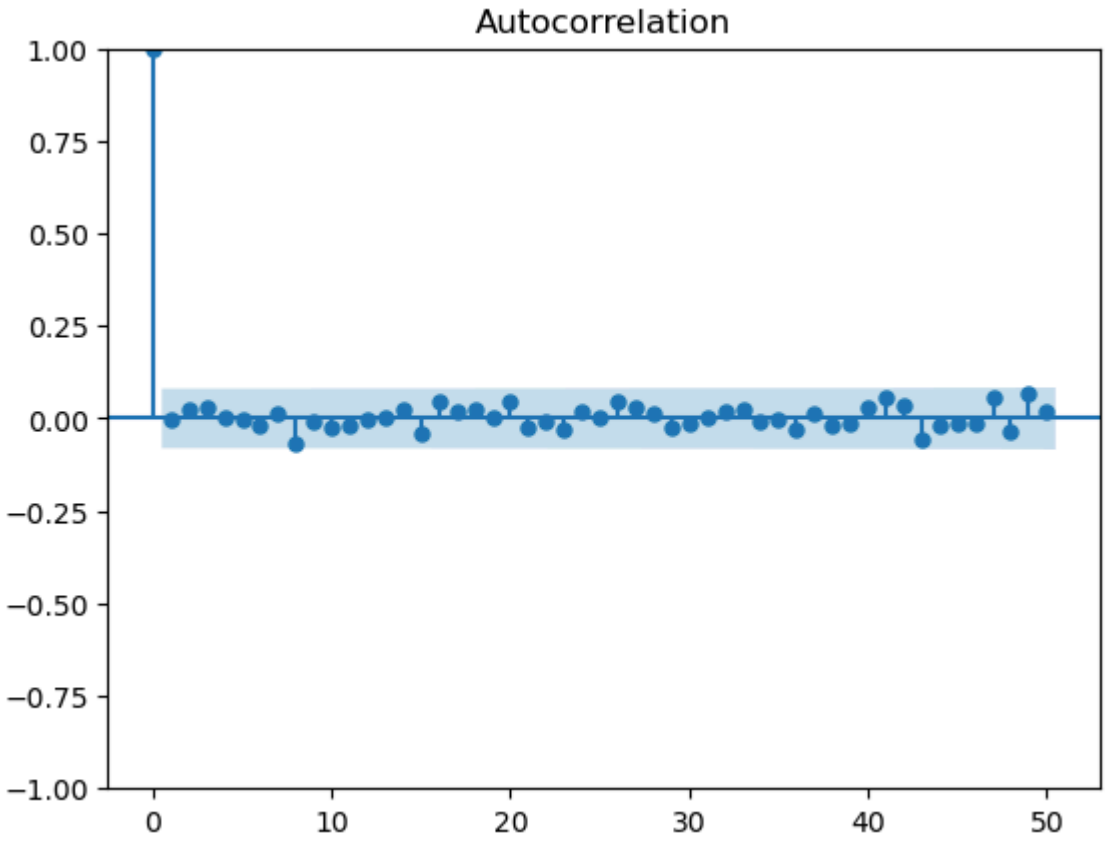


که همچنان در بعضی لگ ها هنوز معناداری داریم و از مدل قوی تری استفاده می کنیم



lb_stat	lb_pvalue	
0.899809	0.015851	1
0.819366	0.398449	2
0.819635	0.923988	3
0.920446	0.928436	4
0.966887	0.944886	5
0.979951	1.135496	6
0.990587	1.214580	7
0.869256	3.863130	8

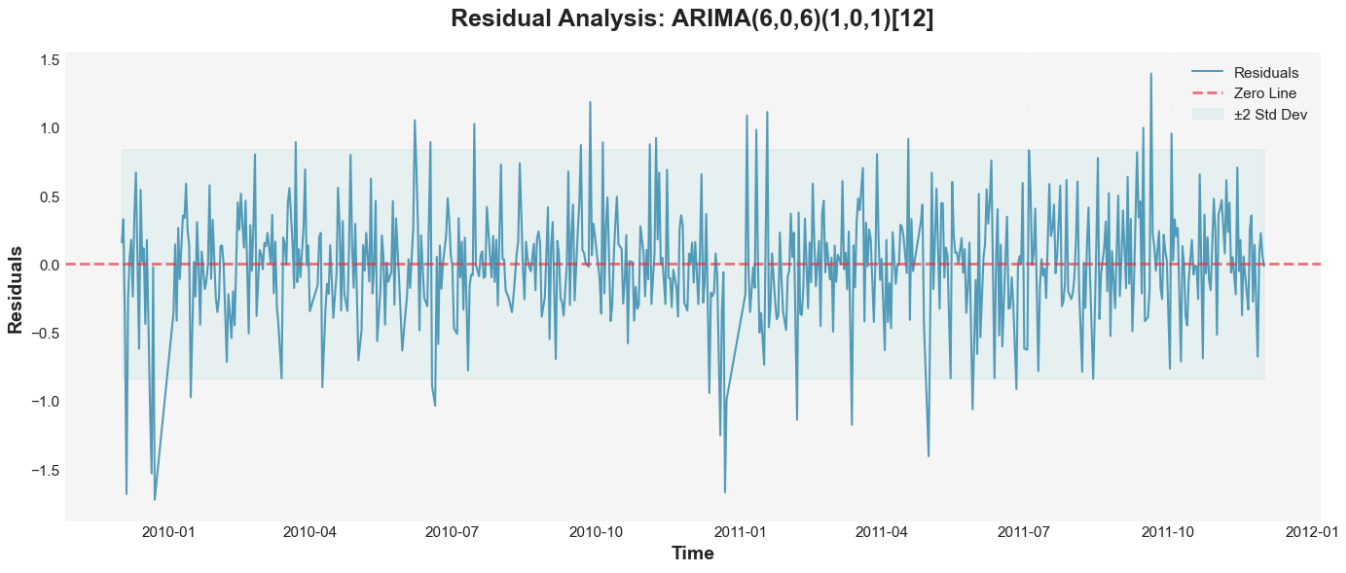
0.918320	3.892891	9
0.937300	4.211534	10
0.957500	4.378777	11
0.975542	4.380879	12
0.986372	4.384472	13
0.988584	4.785117	14
0.982648	5.817813	15
0.973624	6.981091	16
0.980606	7.213936	17
0.983661	7.629190	18
0.989970	7.636425	19
0.983367	8.959918	20
0.986893	9.273553	21
0.991401	9.334605	22
0.992283	9.829333	23
0.994543	10.000890	24
0.996645	10.002946	25
0.994942	11.176412	26
0.994996	11.808710	27
0.996661	11.884750	28
0.997382	12.190054	29
0.998258	12.279627	30
0.998924	12.286393	31
0.999240	12.471177	32
0.999404	12.784395	33
0.999620	12.853329	34
0.999774	12.854691	35
0.999779	13.421926	36
0.999855	13.539379	37
0.999899	13.725097	38
0.999931	13.884626	39
0.999932	14.452639	40
0.999734	16.683232	41
0.999715	17.399032	42
0.999187	19.573998	43
0.999365	19.847779	44
0.999551	19.973858	45
0.999683	20.100752	46
0.999184	22.260189	47
0.999162	22.984765	48
0.997368	25.850699	49
0.997980	26.025364	50

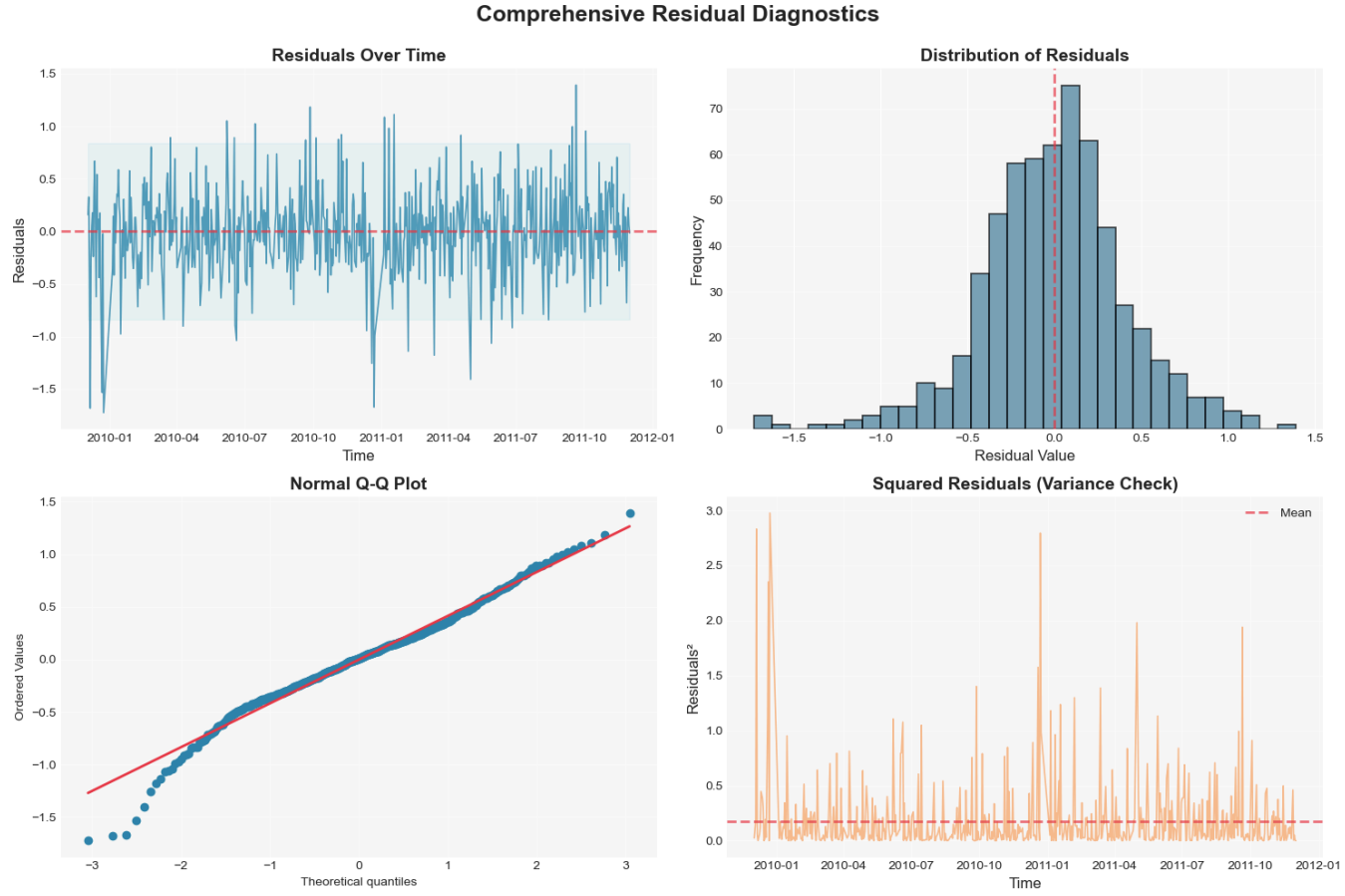


همانطور که در نمودار خودهمبستگی قابل مشاهده هست نمودار خودهمبستگی که داریم در فضای معنادار قرار داره و مدل ما کارش رو انجام دارد و ضرایب باکس-پیرز ما هم بالاتر از ۰.۱۵ هستند که لگ ها ما در موقعیتی خوبی قرار دارند

بررسی نهایی مدل

بررسی باقی مانده های مدل

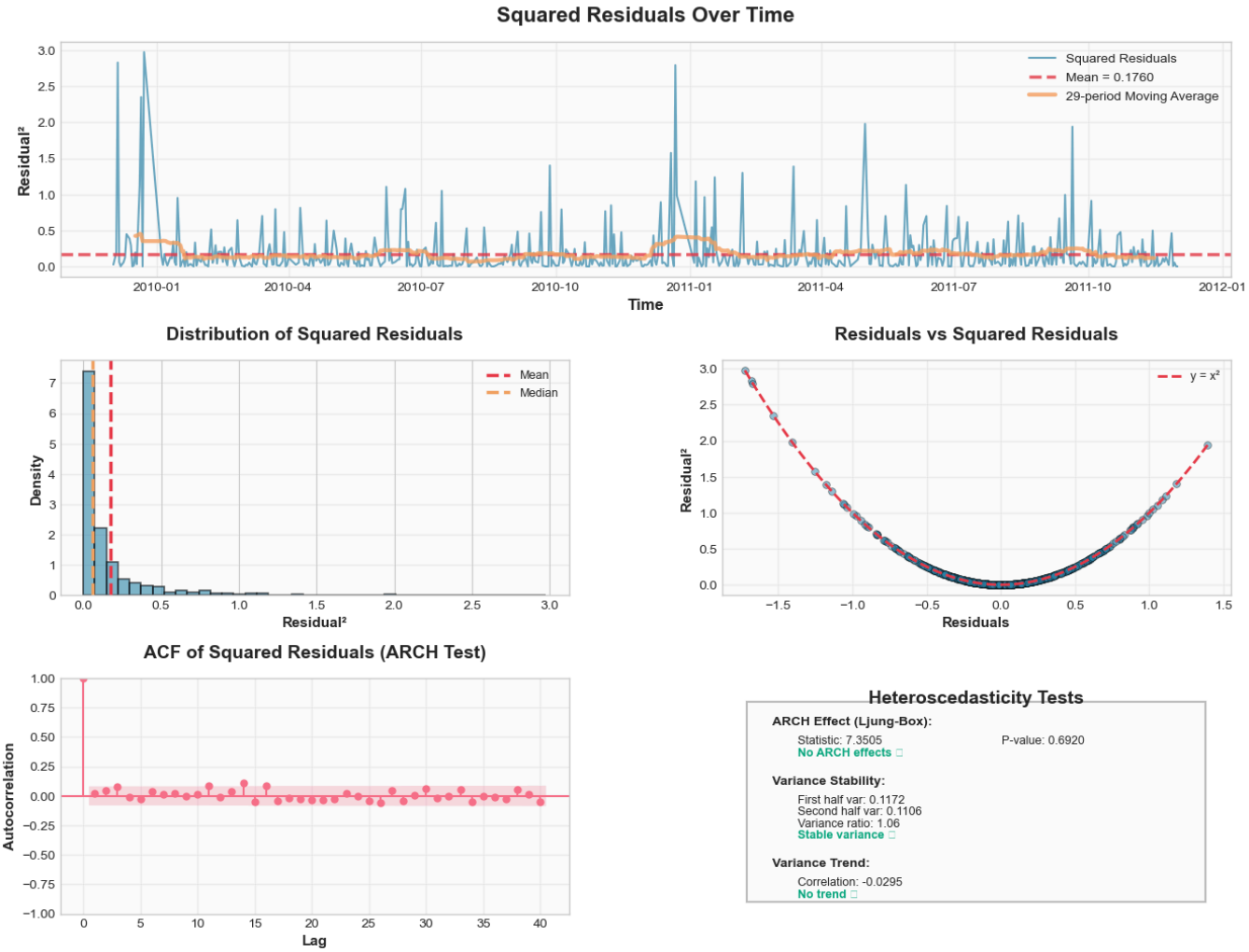




توان دوم باقی مانده ها

- در بررسی توان دوم باقی مانده ها مشاهده شد که واریانس ما ثابت هست که از طریق رسم نمودار و بررسی خودهمبستگی آن اطمینان حاصل شد

Squared Residuals Analysis (Heteroscedasticity Check)

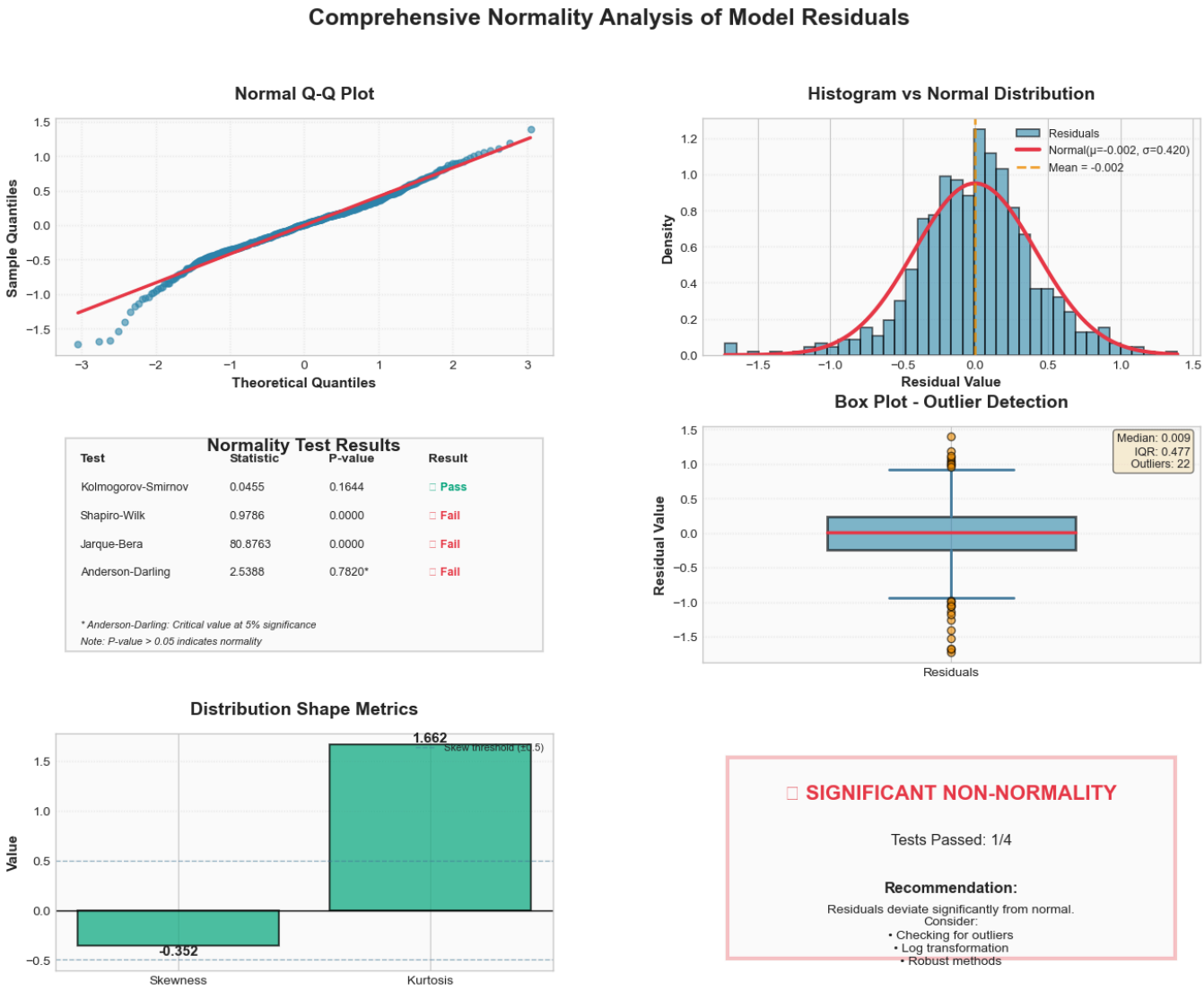



```
======================================================================
SQUARED RESIDUALS INTERPRETATION
======================================================================

Mean of squared residuals: 0.175996
Std of squared residuals:  0.337097
Variance ratio (1st/2nd half): 1.06
Time correlation: -0.0295

----------------------------------------------------------------------
:WHAT TO LOOK FOR
----------------------------------------------------------------------
GOOD: Random scatter around mean, no patterns ✓
GOOD: Constant variance over time (ratio < 2) ✓
GOOD: No autocorrelation in squared residuals (ARCH test p > 0.05) ✓
GOOD: Moving average is roughly flat ✓

BAD: Increasing/decreasing trend over time ✗
BAD: Periods of high variance followed by low variance ✗
BAD: Autocorrelation in squared residuals (volatility clustering) ✗
BAD: Fan-shaped pattern (variance increases with level) ✗
```

```
======================================================================
OVERALL: Residuals show HOMOSCEDASTICITY (constant variance) ✓
Your model's variance assumptions are satisfied!
======================================================================
```

نرمالیتی باقی مانده ها

- در این تست ها فقط در کلموگروف-اسمیرنوف مورد قبول شد و در تست ها دیگر مورد قبول در نظر گرفته نشد

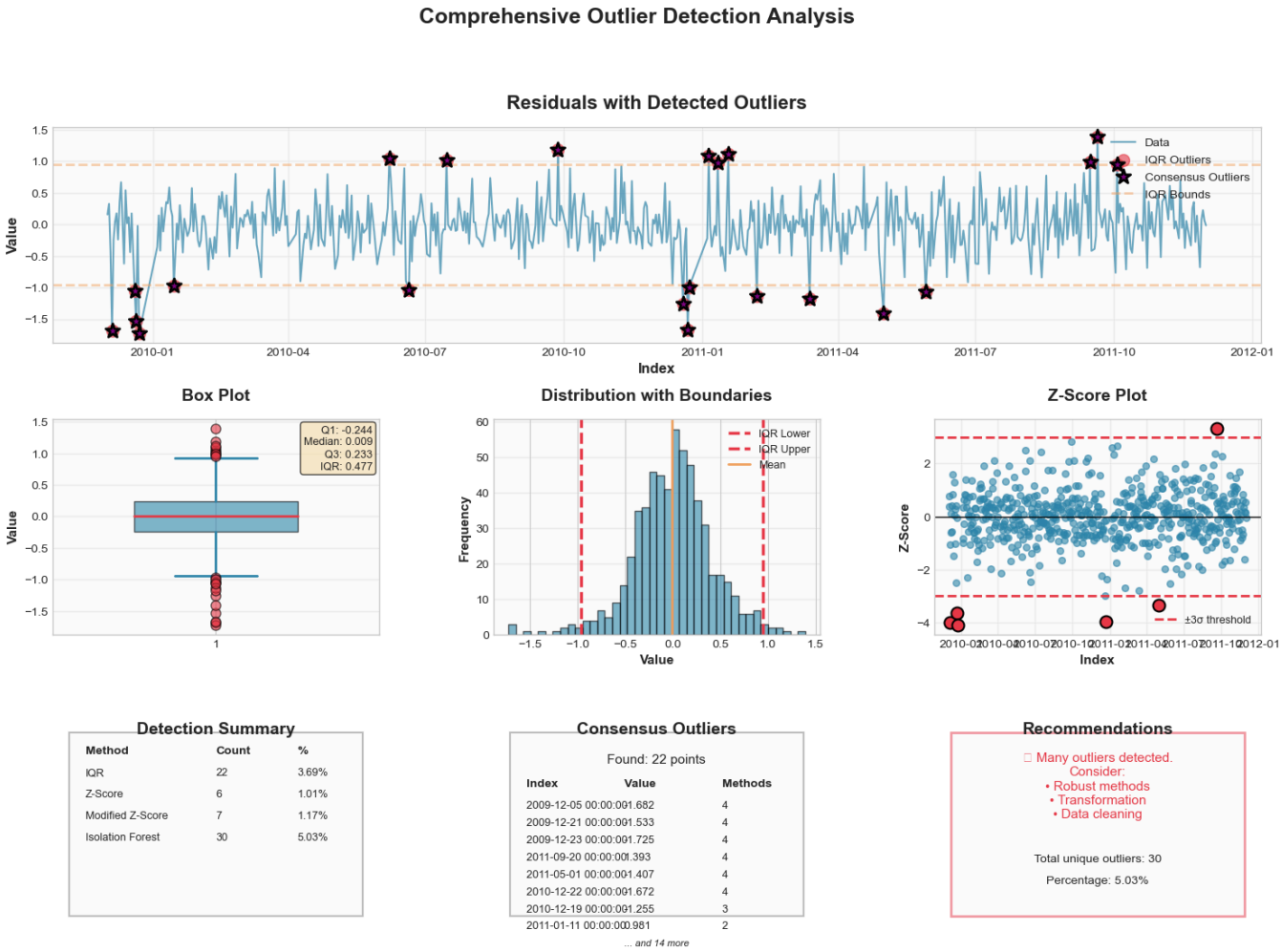



```
============================================================
NORMALITY TEST SUMMARY
============================================================
Kolmogorov-Smirnov: stat=0.0455, p=0.1644
Shapiro-Wilk:       stat=0.9786, p=0.0000
Jarque-Bera:        stat=80.8763, p=0.0000
Anderson-Darling:   stat=2.5388

Skewness: -0.3521
Kurtosis: 1.6616
```

```
Tests Passed: 1/4
============================================================
```

داده های پرت در باقیمانده ها

- به نظر من وجود داده های پرت در باقی مانده به خاطر پیچیدگی داده ای که داشتم قابل حدس بود و رفتاری قابل پیش بینی




```
============================================================
                OUTLIER DETECTION SUMMARY
============================================================

Total observations: 596

            Method  Outliers_Detected Percentage
               IQR                 22      3.69%
           Z-Score                  6      1.01%
  Modified Z-Score                  7      1.17%
  Isolation Forest                 30      5.03%

============================================================
CONSENSUS OUTLIERS (detected by 2+ methods): 22
============================================================
  Index 2009-12-05 00:00:00: Value = -1.6824, Detected by 4 methods
```

```
Index 2009-12-21 00:00:00: Value = -1.5332, Detected by 4 methods
Index 2009-12-23 00:00:00: Value = -1.7251, Detected by 4 methods
 Index 2011-09-20 00:00:00: Value = 1.3927, Detected by 4 methods
Index 2011-05-01 00:00:00: Value = -1.4071, Detected by 4 methods
Index 2010-12-22 00:00:00: Value = -1.6716, Detected by 4 methods
Index 2010-12-19 00:00:00: Value = -1.2552, Detected by 3 methods
 Index 2011-01-11 00:00:00: Value = 0.9815, Detected by 2 methods
 Index 2011-09-15 00:00:00: Value = 0.9966, Detected by 2 methods
Index 2011-05-29 00:00:00: Value = -1.0642, Detected by 2 methods
Index 2011-03-13 00:00:00: Value = -1.1778, Detected by 2 methods
Index 2011-02-06 00:00:00: Value = -1.1398, Detected by 2 methods
 Index 2011-01-18 00:00:00: Value = 1.1122, Detected by 2 methods
Index 2010-12-23 00:00:00: Value = -0.9927, Detected by 2 methods
 Index 2011-01-05 00:00:00: Value = 1.0864, Detected by 2 methods
Index 2009-12-20 00:00:00: Value = -1.0579, Detected by 2 methods
 Index 2010-09-27 00:00:00: Value = 1.1836, Detected by 2 methods
 Index 2010-07-15 00:00:00: Value = 1.0241, Detected by 2 methods
Index 2010-06-20 00:00:00: Value = -1.0380, Detected by 2 methods
 Index 2010-06-07 00:00:00: Value = 1.0514, Detected by 2 methods
Index 2010-01-15 00:00:00: Value = -0.9758, Detected by 2 methods
 Index 2011-10-03 00:00:00: Value = 0.9545, Detected by 2 methods

==========================================================================
                                                          :IQR BOUNDS
                                                  Lower: -0.9600
                                                   Upper: 0.9482
==========================================================================
```
