Homework 1

Emre Çağan Kanlı - 2020402072

Introduction

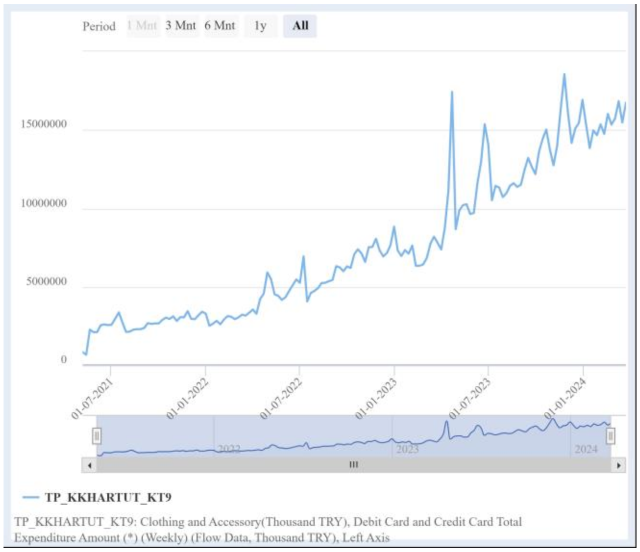
For this homework, 3 datasets from TCMB website /) and relevant datasets from Google Trends was used. The aim is to see if the datasets from TCMB are related with an independent dataset, possibly from Google Trends.

Model 1 - Clothing Expenditures and Holidays

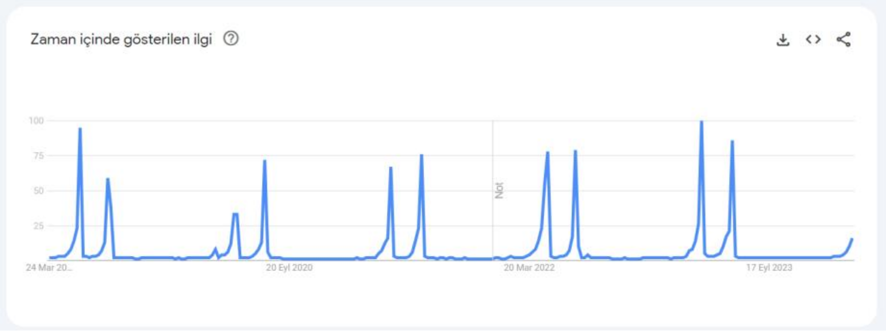
In Turkey, it is common for people to buy new clothes before the religious holidays, namely Ramadan and Eid. In this sense, the data between clothing expenditures and holidays might be investigated. To have data on how close the holiday dates are and how much people think about these holidays, we can assume that people search about them more when they are close.

Inspection of the Time Series

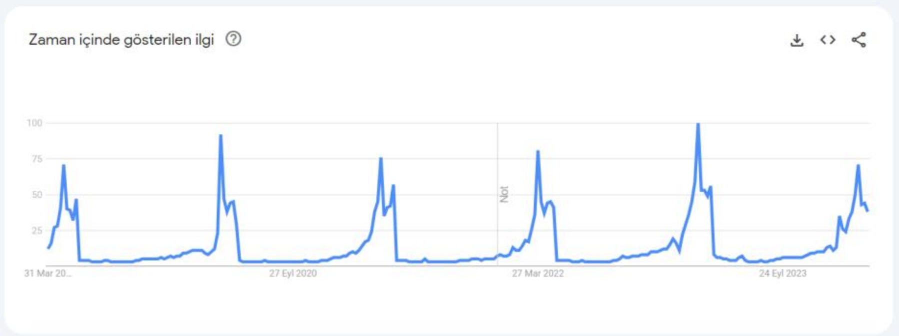
The clothing expenditure data can be pulled from TCMB Website and search data can be acquired from Google Trends. All datasets are weekly, ranging from 21/05/2021 to 22/03/2024 with 149 data points.



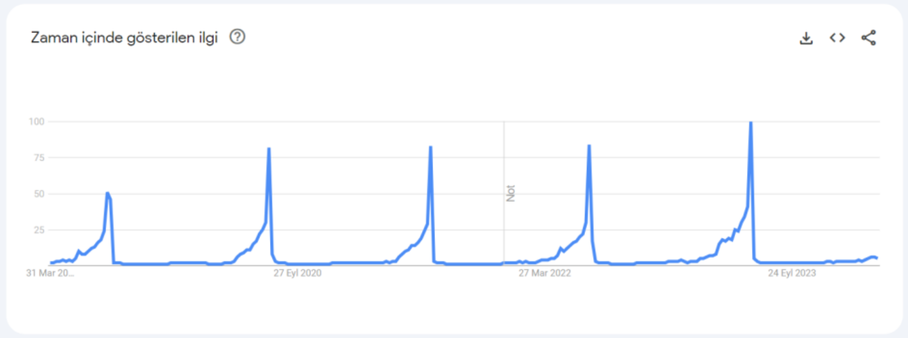
Clothing Data From TCMB Website



Bayram Search Google Trends



Ramazan Search Google Trends

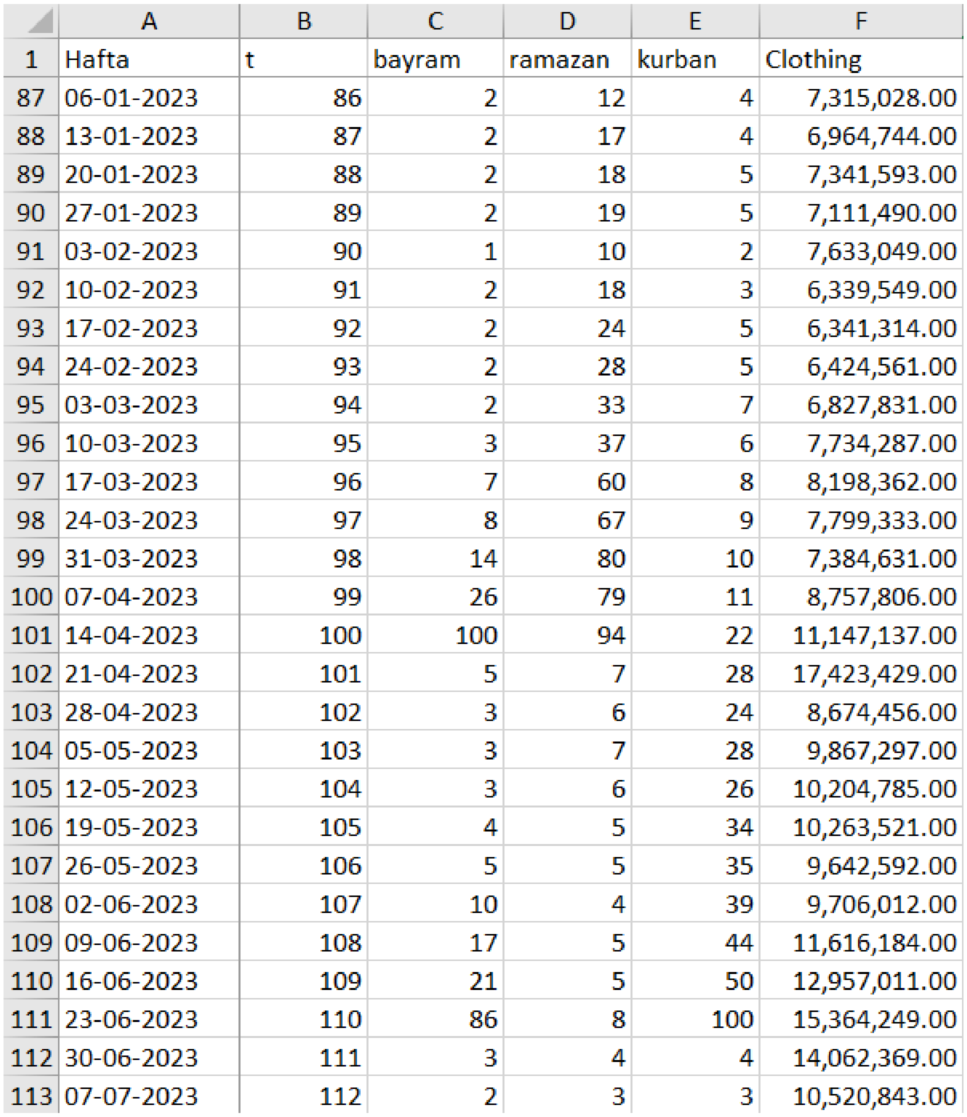


Kurban Search Google Trends

It can be argued that the peaks of Clothing Data matches with the Bayram dates and the combination of Ramazan and Kurban search trends.

Pulling and Organizing the Data

After downloading the time series from their respective websites, all datasets were combined into one excel sheet with an additional t value for time purposes. A constant feature for intercept was added afterwards in code.

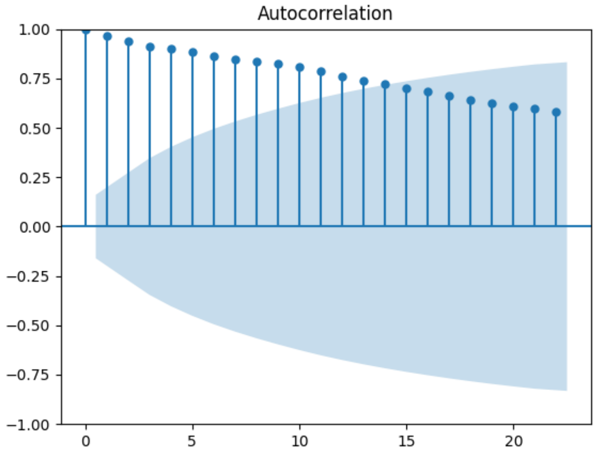


	A	B	C	D	E	F
1	Hafta	t	bayram	ramazan	kurban	Clothing
87	06-01-2023	86	2	12	4	7,315,028.00
88	13-01-2023	87	2	17	4	6,964,744.00
89	20-01-2023	88	2	18	5	7,341,593.00
90	27-01-2023	89	2	19	5	7,111,490.00
91	03-02-2023	90	1	10	2	7,633,049.00
92	10-02-2023	91	2	18	3	6,339,549.00
93	17-02-2023	92	2	24	5	6,341,314.00
94	24-02-2023	93	2	28	5	6,424,561.00
95	03-03-2023	94	2	33	7	6,827,831.00
96	10-03-2023	95	3	37	6	7,734,287.00
97	17-03-2023	96	7	60	8	8,198,362.00
98	24-03-2023	97	8	67	9	7,799,333.00
99	31-03-2023	98	14	80	10	7,384,631.00
100	07-04-2023	99	26	79	11	8,757,806.00
101	14-04-2023	100	100	94	22	11,147,137.00
102	21-04-2023	101	5	7	28	17,423,429.00
103	28-04-2023	102	3	6	24	8,674,456.00
104	05-05-2023	103	3	7	28	9,867,297.00
105	12-05-2023	104	3	6	26	10,204,785.00
106	19-05-2023	105	4	5	34	10,263,521.00
107	26-05-2023	106	5	5	35	9,642,592.00
108	02-06-2023	107	10	4	39	9,706,012.00
109	09-06-2023	108	17	5	44	11,616,184.00
110	16-06-2023	109	21	5	50	12,957,011.00
111	23-06-2023	110	86	8	100	15,364,249.00
112	30-06-2023	111	3	4	4	14,062,369.00
113	07-07-2023	112	2	3	3	10,520,843.00

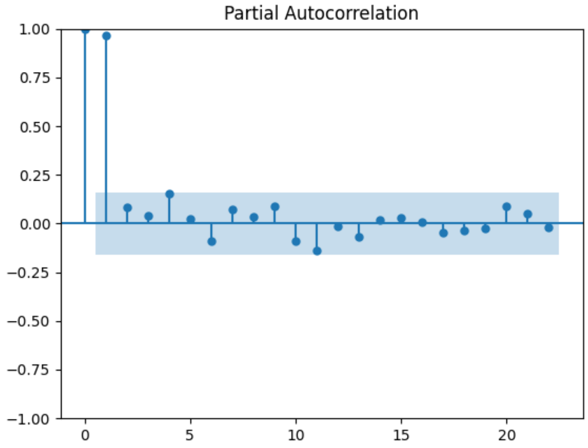
Some Observations From The Excel Containing All Data

Building the Model

Firstly, there appears to be a trend in the Clothing data. The Augmented Dickey–Fuller test gives a p-value of 99.3%. We can also see that the data is not stationary by checking the ACF and PACF plots, .

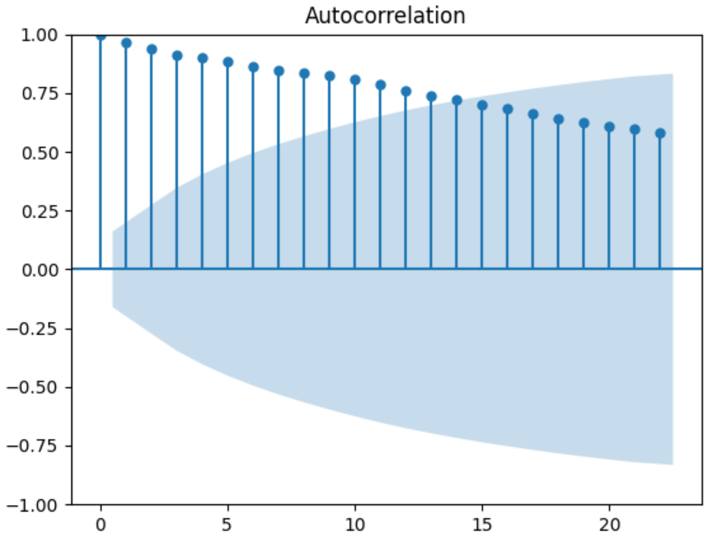


ACF Plot

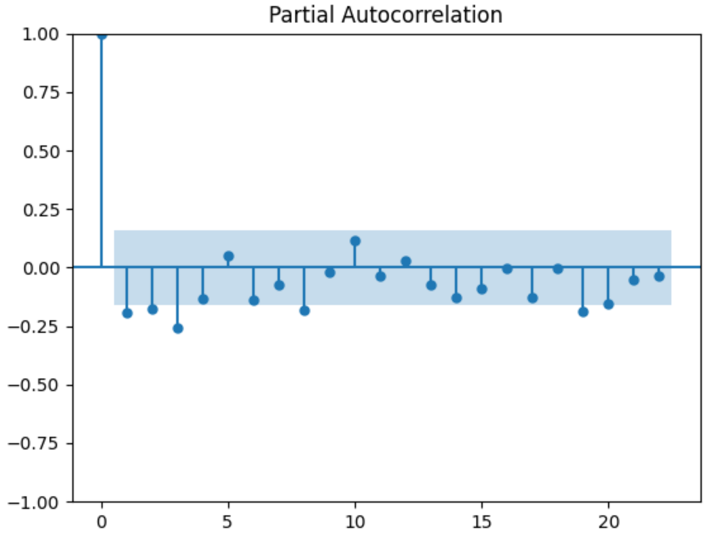


PACF Plot

Taking the first difference and plotting again, the ACF and PACF are closer to ideal. The Augmented Dickey–Fuller test gives 1.41e-07%, which indicates a stationary time series. The mean of the data is closer to 0.

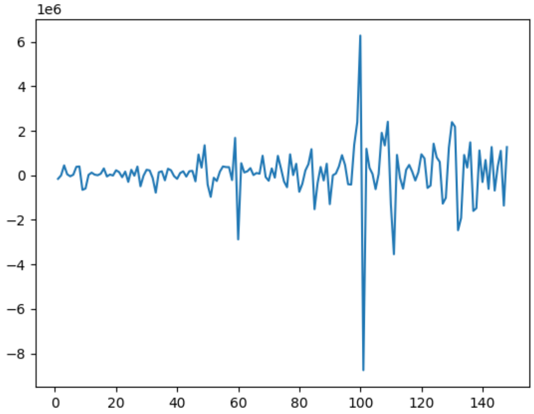


ACF Plot



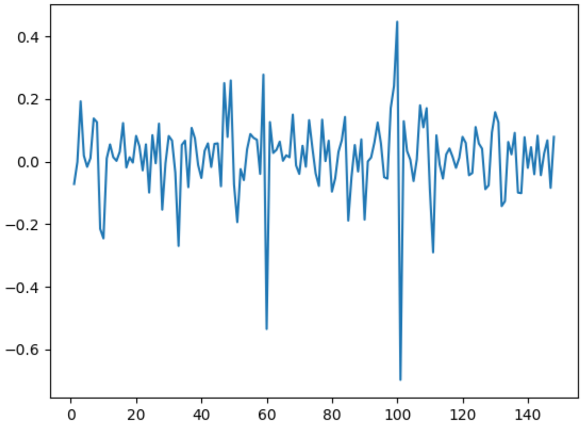
PACF Plot

The data can now be argued to be stationary, however the residual plot shows some increase in variance.



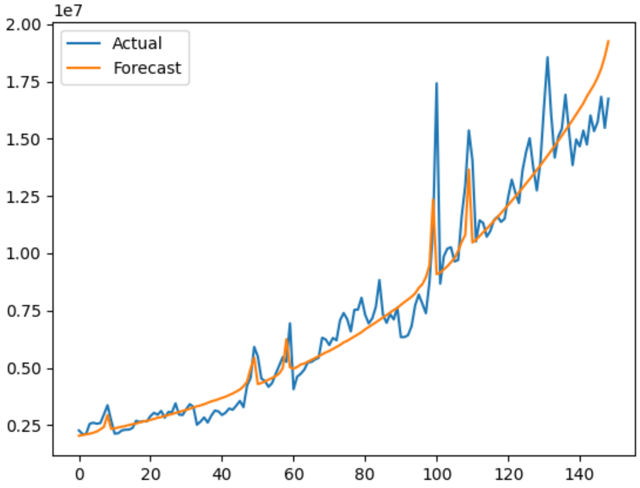
First Difference Plot

To help with this, a log Transformation can be used. Taking the first difference of the log transformed data, new Augmented Dickey–Fuller test gives 9.96e-13% and the plot looks more consistent.

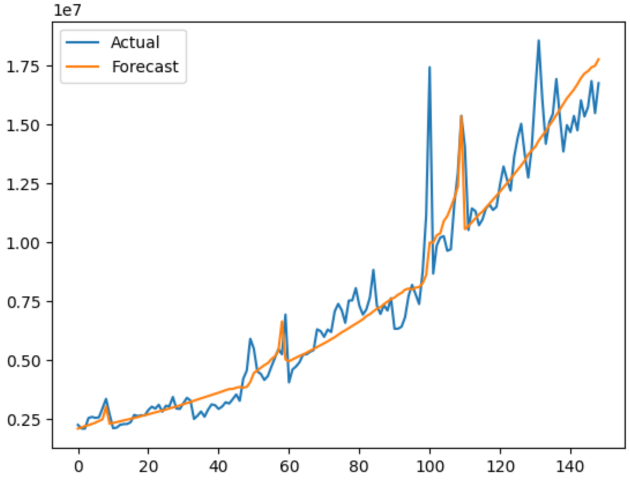


Log Transform First Difference Plot

To see which features to use we can check the correlations between the search data. We get .465 for Bayram and Ramazan, .675 for Bayram and Kurban and -.00145 for Ramazan and Kurban. This is not surprising as the Bayram data closesly resembles the sum of Ramazan and Kurban data. We can try two models with one having bayram and the other having ramazan and kurban.

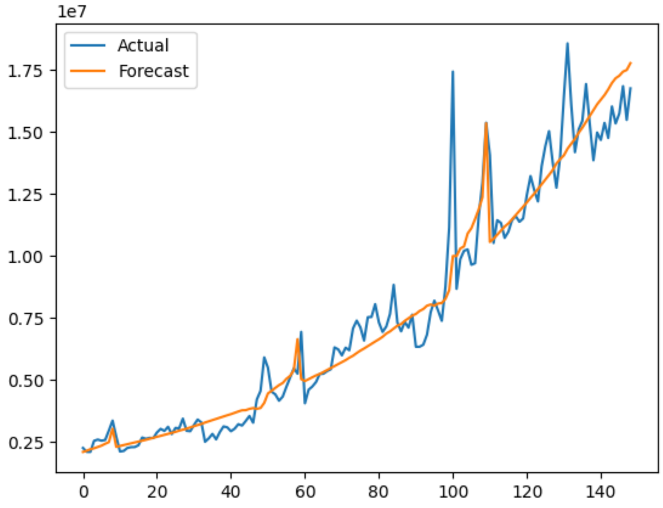


Model 1.1 with feature bayram



Model 1.2 with features ramazan and kurban

Both models perform well, but model 1.1 seems to capture the peaks better. Choosing it as the main model 1 and investigating more, we get the following summary.



Model 1 Summary

p-value of the F-statistic is very low, showing that the model is promising. The p-value of each feature is also low, indicating that they are significant. R^2 value is close to 1, showing a good fit. All in all, this looks like a good model.

Conclusion

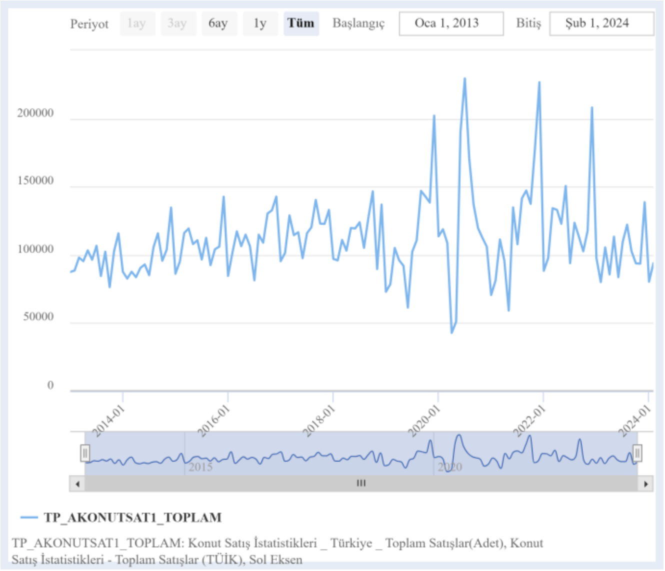
The significance of coefficients, p-value of the F-statistic, the visualization and everything points to the bayram search data being influential on predicting the Clothes expenditure. The reason why bayram data captured the peaks better than ramazan and kurban might be the fluctuations towards the end. Since the bayram search data has no values there, it is no influenced by those fluctuations. However, it might be the case that the ramazan and kurban data fit too much to those data at the endthat they lose the ability to predict peaks.

Model 2 - House Sales, Seasonality and House for Sale Search

People buy houses all the time, but the amount might be investigated. A good candidate of a feature predicting how much people think about buying houses might be when people search House for Sales.

Inspection of the Time Series

TCMB data from House Sales is a monthly time series, from 2013-01 to 2024-02 with 134 datapoints. Plotting the House Sales Time Series, we can try to make some suggestions.



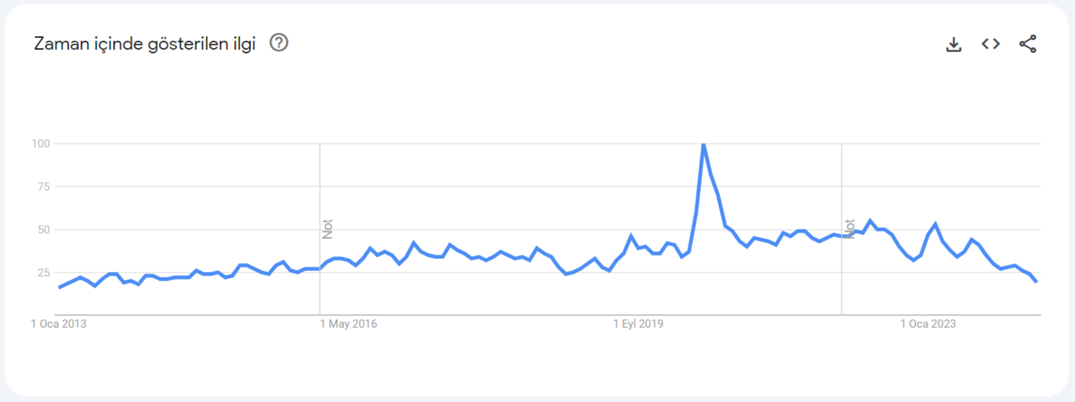
Başlangıç Oca 1, 2013

Bitiş Şub 1, 2024

TP_AKONUTSAT1_TOPLAM: Konut Satış İstatistikleri _ Türkiye _ Toplam Satışlar(Adet), Konut Satış İstatistikleri - Toplam Satışlar (TÜİK), Sol Eksen

House Sales Time Series Plot

There appears to be some seasonality, which we can see by noticing a peak every december, for example. The question we will try to answer is whether seasonilty really has an effect her and can house fo sale search data cover the remaining variance.

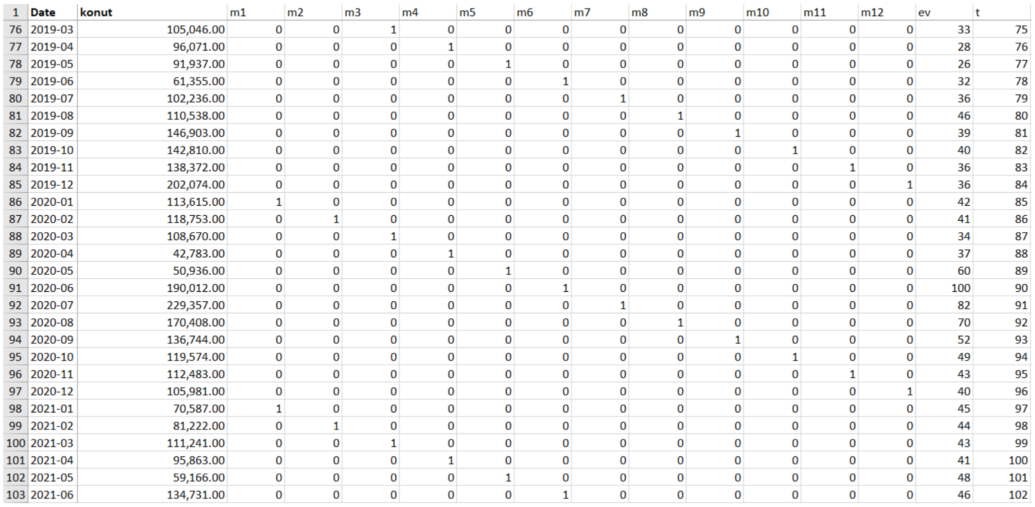


House for Sale Search Plot

The peak in the summer 2020 on both plots suggest that there might be a correlation between the dependent and independant datt.

Pulling and Organizing the Data

After downloading the time series from their respective websites, all datasets were combined into one excel sheet with an additional t value for time purposes and dummy variable for each month. A constant feature for intercept was added afterwards in code for one of the models.

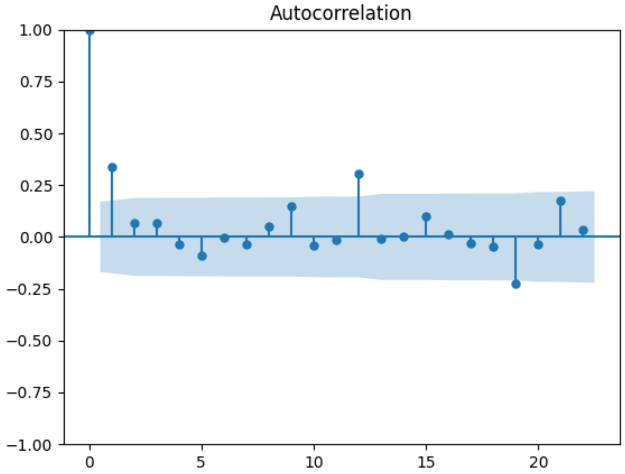


1	Date	konut	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	ev	t
76	2019-03	105,046.00	0	0	1	0	0	0	0	0	0	0	0	0	33	75
77	2019-04	96,071.00	0	0	0	1	0	0	0	0	0	0	0	0	28	76
78	2019-05	91,937.00	0	0	0	0	1	0	0	0	0	0	0	0	26	77
79	2019-06	61,355.00	0	0	0	0	0	1	0	0	0	0	0	0	32	78
80	2019-07	102,236.00	0	0	0	0	0	0	1	0	0	0	0	0	36	79
81	2019-08	110,538.00	0	0	0	0	0	0	0	1	0	0	0	0	46	80
82	2019-09	146,903.00	0	0	0	0	0	0	0	0	1	0	0	0	39	81
83	2019-10	142,810.00	0	0	0	0	0	0	0	0	0	1	0	0	40	82
84	2019-11	138,372.00	0	0	0	0	0	0	0	0	0	0	1	0	36	83
85	2019-12	202,074.00	0	0	0	0	0	0	0	0	0	0	0	1	36	84
86	2020-01	113,615.00	1	0	0	0	0	0	0	0	0	0	0	0	42	85
87	2020-02	118,753.00	0	1	0	0	0	0	0	0	0	0	0	0	41	86
88	2020-03	108,670.00	0	0	1	0	0	0	0	0	0	0	0	0	34	87
89	2020-04	42,783.00	0	0	0	1	0	0	0	0	0	0	0	0	37	88
90	2020-05	50,936.00	0	0	0	0	1	0	0	0	0	0	0	0	60	89
91	2020-06	190,012.00	0	0	0	0	0	1	0	0	0	0	0	0	100	90
92	2020-07	229,357.00	0	0	0	0	0	0	1	0	0	0	0	0	82	91
93	2020-08	170,408.00	0	0	0	0	0	0	0	1	0	0	0	0	70	92
94	2020-09	136,744.00	0	0	0	0	0	0	0	0	1	0	0	0	52	93
95	2020-10	119,574.00	0	0	0	0	0	0	0	0	0	1	0	0	49	94
96	2020-11	112,483.00	0	0	0	0	0	0	0	0	0	0	1	0	43	95
97	2020-12	105,981.00	0	0	0	0	0	0	0	0	0	0	0	1	40	96
98	2021-01	70,587.00	1	0	0	0	0	0	0	0	0	0	0	0	45	97
99	2021-02	81,222.00	0	1	0	0	0	0	0	0	0	0	0	0	44	98
100	2021-03	111,241.00	0	0	1	0	0	0	0	0	0	0	0	0	43	99
101	2021-04	95,863.00	0	0	0	1	0	0	0	0	0	0	0	0	41	100
102	2021-05	59,166.00	0	0	0	0	1	0	0	0	0	0	0	0	48	101
103	2021-06	134,731.00	0	0	0	0	0	1	0	0	0	0	0	0	46	102

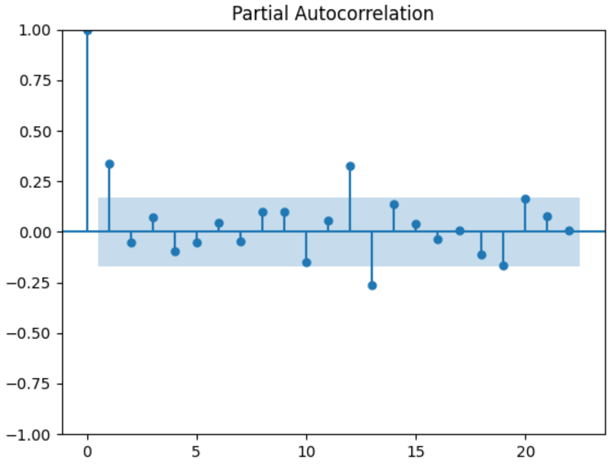
Some Observations From The Excel Containing All Data

Building the Model

The p-value from the Augmented Dickey–Fuller test at first comes out to be 13.4%. This is not low enough, but seasonality plays a big role here. Checking the ACF and PACF plots also show that the non-stationary situation is not that significant.

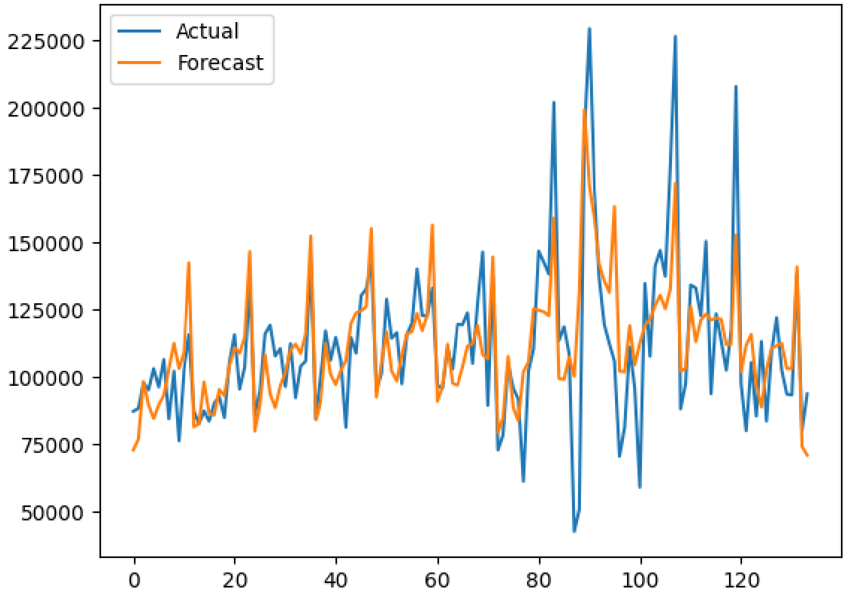


ACF Plot



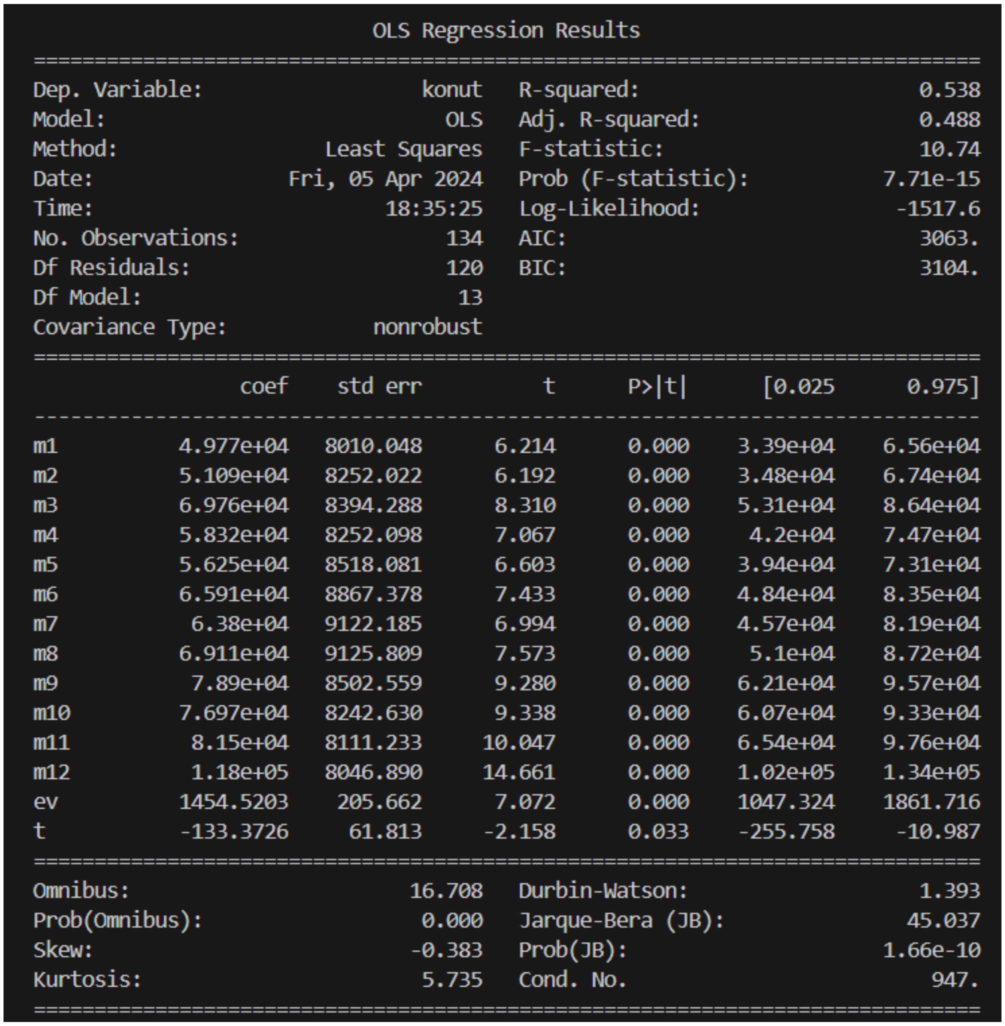
PACF Plot

There are two ways of implementing seasonality: Using all dummy variables and no intercept or using one less dummy variable and using and intercept. The plots of these two will be the same, with the same MAPE of 14.3%, but their interpretations and statistics may vary.



Model 2 Plot

Since the p-values of the coefficients when there is no constant is better, we can take that as the base model and comment on it.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  konut   R-squared:                       0.538
Model:                            OLS   Adj. R-squared:                  0.488
Method:                 Least Squares   F-statistic:                     10.74
Date:                Fri, 05 Apr 2024   Prob (F-statistic):           7.71e-15
Time:                        18:35:25   Log-Likelihood:                -1517.6
No. Observations:                 134   AIC:                             3063.
Df Residuals:                     120   BIC:                             3104.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
m1          4.977e+04   8010.048      6.214      0.000    3.39e+04    6.56e+04
m2          5.109e+04   8252.022      6.192      0.000    3.48e+04    6.74e+04
m3          6.976e+04   8394.288      8.310      0.000    5.31e+04    8.64e+04
m4          5.832e+04   8252.098      7.067      0.000     4.2e+04    7.47e+04
m5          5.625e+04   8518.081      6.603      0.000    3.94e+04    7.31e+04
m6          6.591e+04   8867.378      7.433      0.000    4.84e+04    8.35e+04
m7           6.38e+04   9122.185      6.994      0.000    4.57e+04    8.19e+04
m8          6.911e+04   9125.809      7.573      0.000     5.1e+04    8.72e+04
m9           7.89e+04   8502.559      9.280      0.000    6.21e+04    9.57e+04
m10         7.697e+04   8242.630      9.338      0.000    6.07e+04    9.33e+04
m11          8.15e+04   8111.233     10.047      0.000    6.54e+04    9.76e+04
m12          1.18e+05   8046.890     14.661      0.000    1.02e+05    1.34e+05
ev          1454.5203    205.662      7.072      0.000    1047.324    1861.716
t           -133.3726     61.813     -2.158      0.033    -255.758     -10.987
==============================================================================
Omnibus:                       16.708   Durbin-Watson:                   1.393
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.037
Skew:                          -0.383   Prob(JB):                     1.66e-10
Kurtosis:                       5.735   Cond. No.                         947.
==============================================================================
```

Model 2 Summary

Conclusion

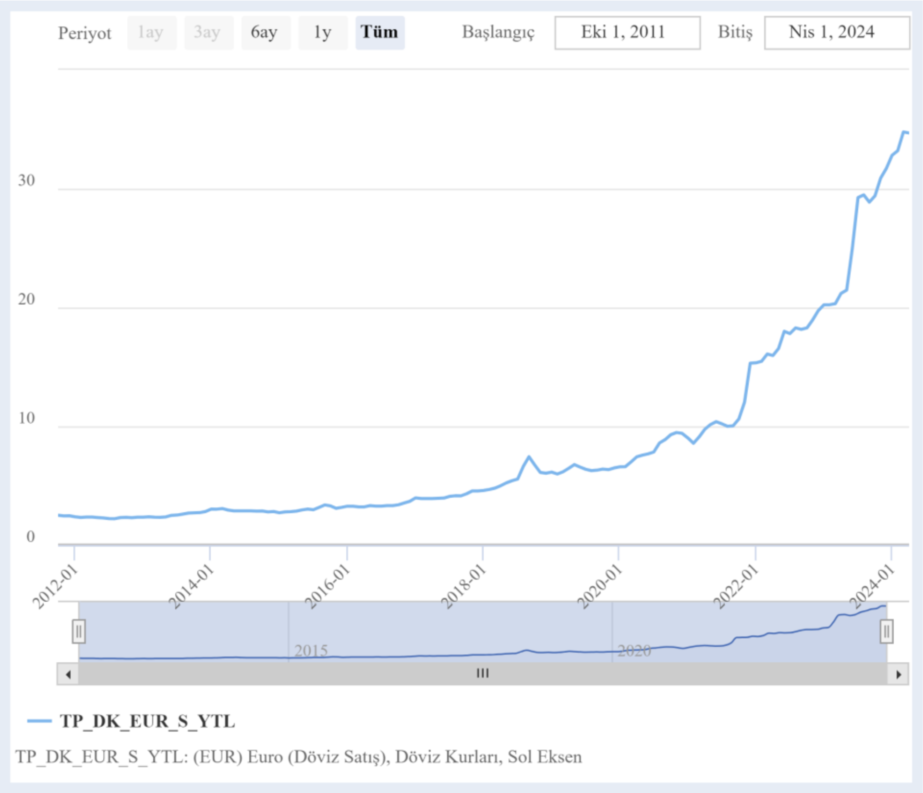
Each p-value of the coefficients are very small, so all of them are significant. The low p-value of ev feature shows that House for Sales search was significant in predicting the House Sales. The p-value of the F-statistic is also low, which means that the model is significant. The coefficient of trend is negative, but not very large compared to other numbers, so it can be argued that there is a small trend down. The R^2 value is low, but this can be attributed to the low trend level.

Model 3 - Euro Exchange Rate and President

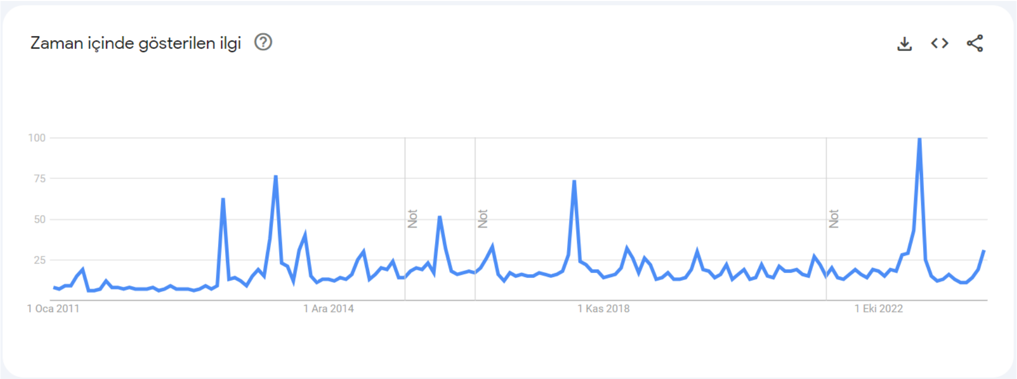
President Recep Tayyip Erdoğan is arguably the most influencial person in Turkey in many aspects. As he is the one having the last say in many decisions, including economy, what he is doing and saying might influance the exchange rates.

Inspection of the Time Series

The exchange rate of Euro can be used as a standart foreign currency. The data from TCMB goes from 28-05-2021 to 29-03-2024, with 149 datapoints. How much people think that the president is influancing the economy might be related with how much they search his name on Google.



Euro Time Series

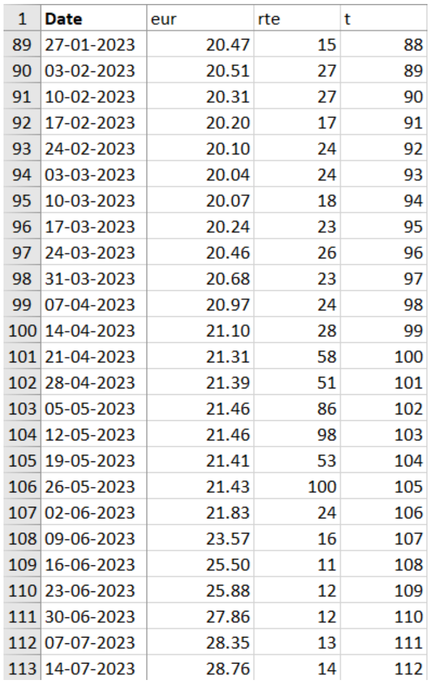


Recep Tayyip Erdoğan Search Time Series

The jumps of Euro plot show a relation with the peaks of search data, can be investigated.

Pulling and Organizing the Data

After downloading the time series from their respective websites, all datasets were combined into one excel sheet with an additional t value for time purposes.

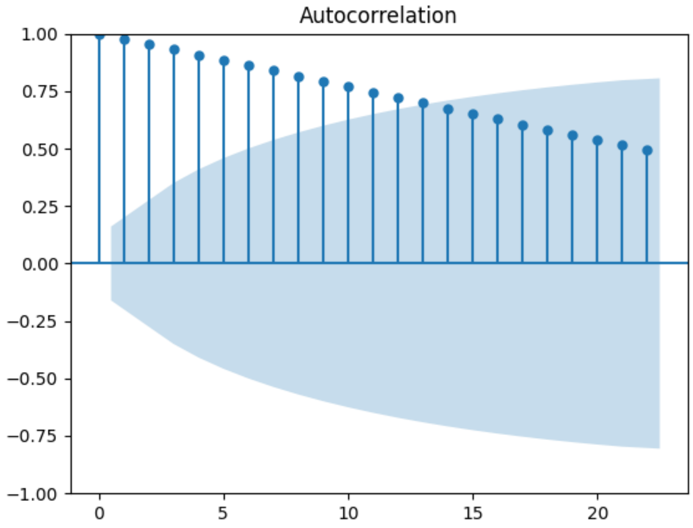


1	Date	eur	rte	t
89	27-01-2023	20.47	15	88
90	03-02-2023	20.51	27	89
91	10-02-2023	20.31	27	90
92	17-02-2023	20.20	17	91
93	24-02-2023	20.10	24	92
94	03-03-2023	20.04	24	93
95	10-03-2023	20.07	18	94
96	17-03-2023	20.24	23	95
97	24-03-2023	20.46	26	96
98	31-03-2023	20.68	23	97
99	07-04-2023	20.97	24	98
100	14-04-2023	21.10	28	99
101	21-04-2023	21.31	58	100
102	28-04-2023	21.39	51	101
103	05-05-2023	21.46	86	102
104	12-05-2023	21.46	98	103
105	19-05-2023	21.41	53	104
106	26-05-2023	21.43	100	105
107	02-06-2023	21.83	24	106
108	09-06-2023	23.57	16	107
109	16-06-2023	25.50	11	108
110	23-06-2023	25.88	12	109
111	30-06-2023	27.86	12	110
112	07-07-2023	28.35	13	111
113	14-07-2023	28.76	14	112

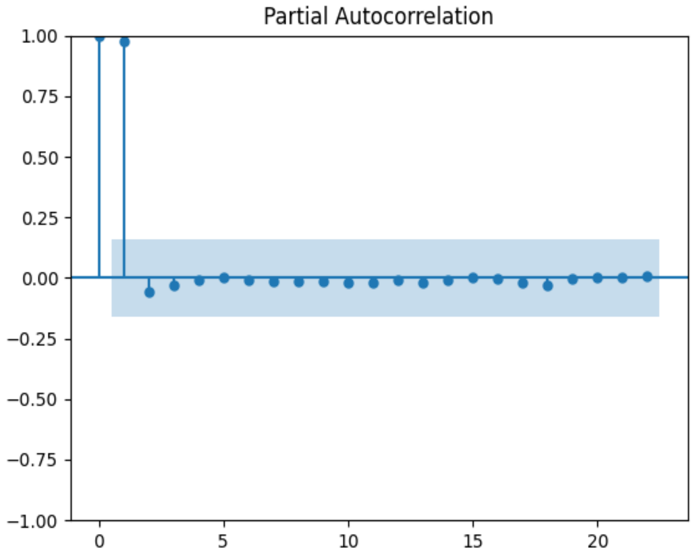
Some Observations From The Excel Containing All Data

Building the Model

The p-value from the Augmented Dickey–Fuller test at first comes out to be 98.2% because there is a clear trend.

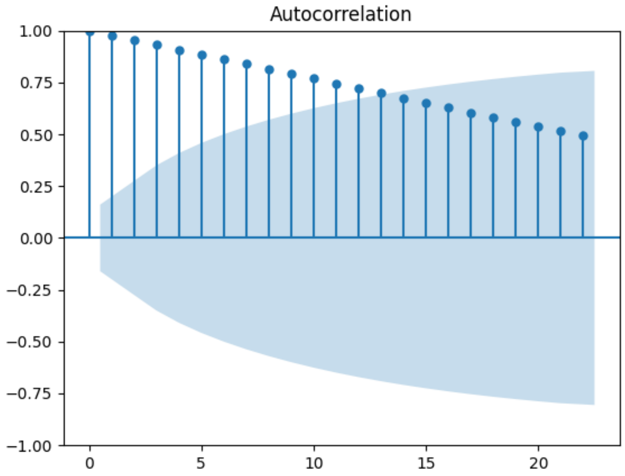


ACF Plot

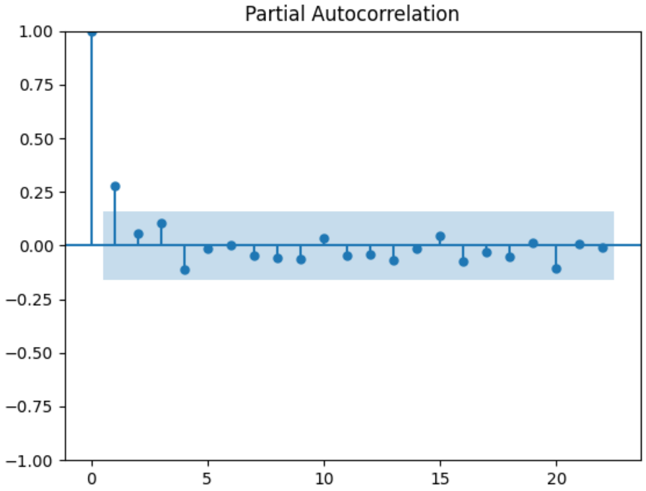


PACF Plot

Taking the first difference and trying again, we get a p-value of 8.00e-15.

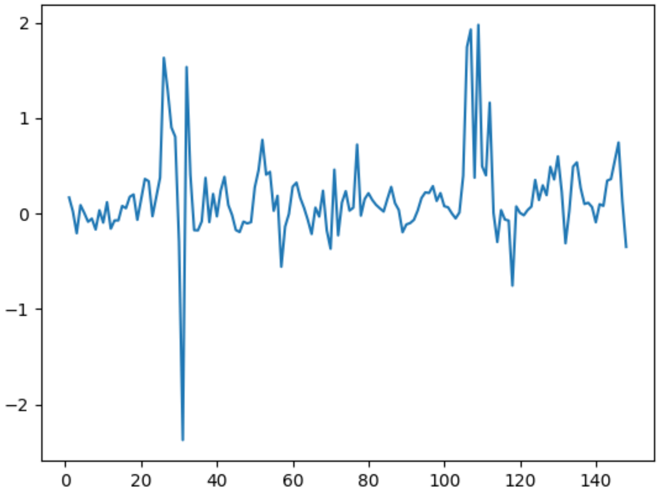


ACF Plot



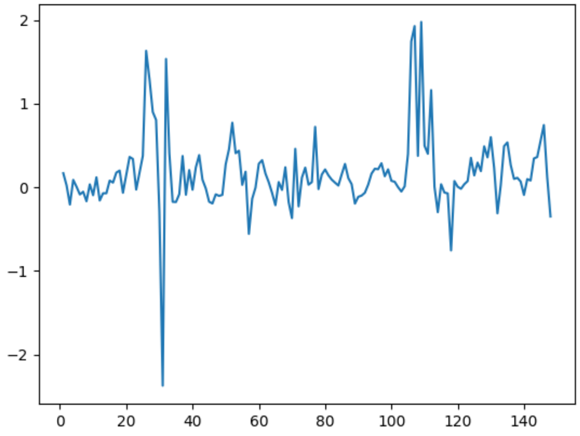
PACF Plot

Residuals also look stationary.



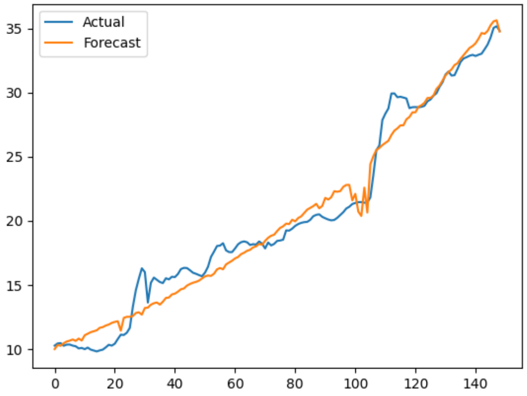
Residuals

Building a linear model with all the features, we get the following plot.



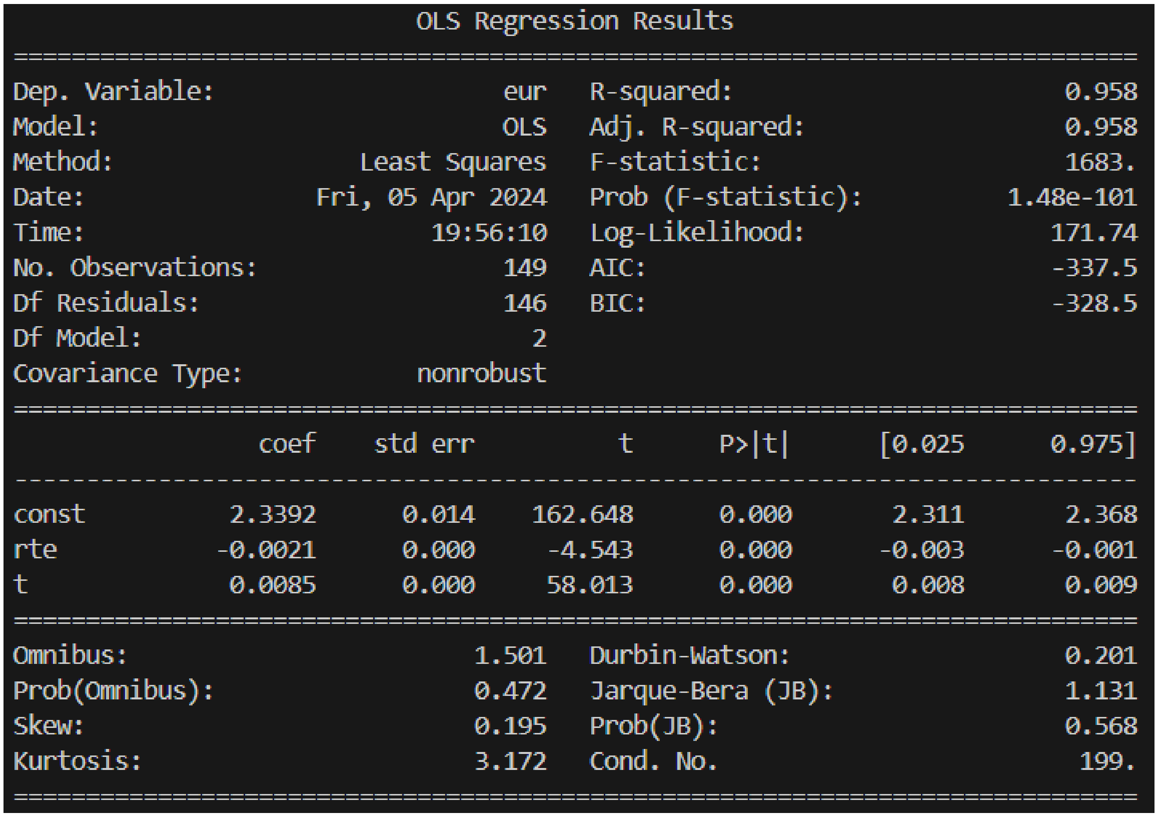
Model 3.1 Plot

However, it is easy to see that this is not a good fit. We can try a log transform to capture an exponential growth.



Model 3.2 Plot

The summary of this second plot is as follows.



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    eur   R-squared:                       0.958
Model:                            OLS   Adj. R-squared:                  0.958
Method:                 Least Squares   F-statistic:                     1683.
Date:                Fri, 05 Apr 2024   Prob (F-statistic):          1.48e-101
Time:                        19:56:10   Log-Likelihood:                 171.74
No. Observations:                 149   AIC:                            -337.5
Df Residuals:                     146   BIC:                            -328.5
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.3392      0.014    162.648      0.000       2.311       2.368
rte           -0.0021      0.000     -4.543      0.000      -0.003      -0.001
t              0.0085      0.000     58.013      0.000       0.008       0.009
==============================================================================
Omnibus:                        1.501   Durbin-Watson:                   0.201
Prob(Omnibus):                  0.472   Jarque-Bera (JB):                1.131
Skew:                           0.195   Prob(JB):                        0.568
Kurtosis:                       3.172   Cond. No.                         199.
==============================================================================
```

Model 3 Summary

Low p-values for coefficient and F-statistic and high R^2 value might indicate a good fit. However, the ceofficient of the search data is negative, which is contradictery to the initial assumption.

Conclusion

Although the statistics are good and the plot is not far off, the sign of the coefficients tell a different story. We can argue that the correlation between the independant and dependant data is not very strong.