



IE 360

Homework 1

3 Regression Models for Macroeconomic Variables

April 5th, 2024

Professor:

Mustafa Gökçe Baydoğan

Prepared by:

Saime Nur Benliler

2019402045

Table of Contents

Introduction

Research Questions

Correlation matrix between the target variables

First Regression Model – House Sales in Turkey

- Independent variables
- Data Manipulation
- Correlation between the independent variables
- Visualization of the variables– trends, seasonality etc
- Regression Model fitting
- Residual Analysis and Model Revision
- Final Model

Second Regression Model – Gold Production in Turkey

- Independent variables
- Data Manipulation
- Correlation between the independent variables
- Visualization of the variables – trends, seasonality etc
- Regression Model fitting
- Residual Analysis and Model Revision
- Final Model

Third Regression Model – Investment in Turkey from abroad

- Independent variables
- Data Manipulation
- Correlation between the independent variables
- Visualization of the variables – trends, seasonality etc
- Regression Model fitting
- Residual Analysis and Model Revision
- Final Model

Conclusion

Large Language Model Prompts

Introduction

This report aims to explore the relationship between various economic factors and trends in Turkey, using data from both the Central Bank of the Republic of Turkey (CBRT) and Google Trends. Three different regression models were developed to examine how the number of house sales in Turkey can be predicted by interest rates and people's online search behaviors related to housing and jobs; how gold production in Turkey is affected by international gold reserves, local events, and online searches for gold prices; and what factors influence foreign investment in Turkey, including online interest in investing, exchange rates, political climate, and inflation. The selection of these variables was based on the availability of data and their potential relevance to the economic phenomena being studied. By combining traditional economic indicators with modern data sources like Google Trends, this report seeks to provide insights into the factors driving the target variables by conducting time series regression analysis.

Research Questions

- First Regression - Housing Market Analysis:

How are the number of house sales in Turkey influenced by interest rates and online search interests related to housing, employment, and economic conditions?

- Second Regression - Gold Production Analysis

What factors contribute to gold production in Turkey, including international gold reserves, local mining incidents, and public interest in gold prices and economic policies?

- Third Regression - Foreign Investment Analysis:

To what extent do online search interest in investing, exchange rates, political climate, and inflation impact foreign investment in Turkey?

Each target variable and corresponding independent variable will be deep dived in its section.

Correlation Matrix Between Target Variables

Correlation between the target variables belonging to different categories from CBRT must have a correlation less than 0.5. As one can see from the correlation matrix in figure 1, the maximum absolute correlation coefficient is 0.24. Therefore, we can start with these target variables in our regression models.

	House Sales	Gold production	Investment in TR
House Sales	1.000000	-0.141218	0.129120
Gold production	-0.141218	1.000000	-0.241879
Investment in TR	0.129120	-0.241879	1.000000

Figure 1: Correlation matrix between target variables

First Regression - Housing Market Analysis:

Target Variable: Number of house sales in Turkey

Independent Variables:

Interest Rate: The cost of borrowing or the price paid for the rental of funds (often expressed as a percentage).

Sahibinden Search Interest: Online search volume for "Sahibinden" (a popular second-hand or first-hand car and house marketplace platform in Turkey), reflecting interest in housing market activities.

İş İlanı Search Interest: Online search volume for "İş İlanı" (job listings), indicating the level of employment-seeking activity which could relate to housing affordability and demand.

Fed Faiz Search Interest: Online search volume for "Fed Faiz" (Federal Reserve Interest Rate), suggesting public interest or concern about global economic policies that could affect local interest rates and housing market.

Konut Kredisi Search Interest: Online search volume for "Konut Kredisi" (mortgage loans), reflecting interest in obtaining financing for housing purchases.

Araba İlanı Search Interest: Online search volume for "Araba İlanı" (car listings), which may be included to gauge broader consumer interest or economic health indicators that could indirectly relate to housing market dynamics.

Data Manipulation

Two different functions are used to manipulate the Google Trends data and CBRT data. For the Google Trends data, the weekly csv file is converted into monthly data by taking the average of four weeks. All of the CBRT data used are monthly data, so the function for CBRT data loads the excel file.

Correlation Between Independent Variables

	House Sales	Interest Rate	Sahibinden Search Interest	İş İlanı Search Interest	Fed Faiz Search Interest	Konut Kredisi Search Interest	Araba İlanı Search Interest
House Sales	1.000000	-0.158355	0.187107	-0.411864	-0.026327	-0.089188	-0.156231
Interest Rate	-0.158355	1.000000	-0.788156	0.312840	-0.347840	-0.178256	0.464260
Sahibinden Search Interest	0.187107	-0.788156	1.000000	-0.069533	0.381406	0.493048	-0.421654
İş İlanı Search Interest	-0.411864	0.312840	-0.069533	1.000000	0.108942	0.047105	-0.084726
Fed Faiz Search Interest	-0.026327	-0.347840	0.381406	0.108942	1.000000	0.240640	-0.265887
Konut Kredisi Search Interest	-0.089188	-0.178256	0.493048	0.047105	0.240640	1.000000	-0.328842
Araba İlanı Search Interest	-0.156231	0.464260	-0.421654	-0.084726	-0.265887	-0.328842	1.000000

Figure 2: Correlation between independent variables in House Sales Regression

None of the independent variables have correlation larger than 0.5 with the house sales. The only thing is the interest rates for housing has relatively high correlation between "Sahibinden" word search in Google trends. However, it is not discarded because they individually have good correlation rate with house sales.

Visualization of the variables

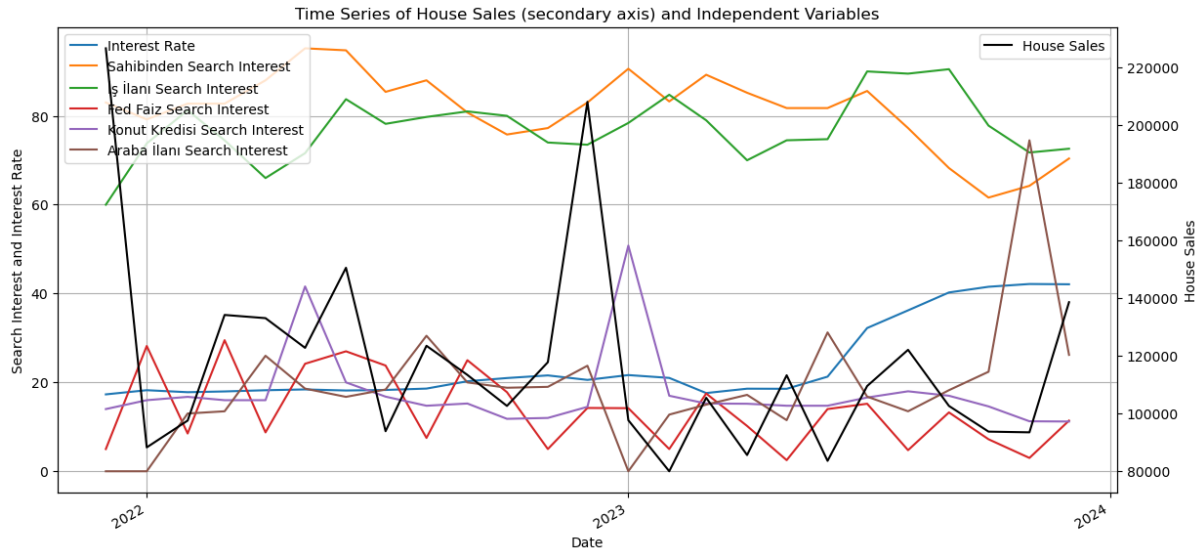


Figure 3: Visualisation of the variables in House Sales Regression

In figure 3, one can see the line graph belonging to the number of House sales over time in Turkey and the independent variables to be used in the regression model. For the house sales data, there is no trend observed for the data but it has a peak in the end of the year 2022. Also, a seasonality is observed in the “Fed Faiz” Google Trends data, which is expected because of the seasonal meetings held by FED to decide on the interest rate for the United States. Also, in the last half of the year 2023, an increasing trend in interest rate can be observed. Since the cost of borrowing money will also increase, we might expect a decrease in the number of houses sold if it has a significant effect on the target variable.

Regression Model fitting

By using the independent variables specified in the beginning of this section, an initial regression model is fitted in Python and the seasonality and trend variables are handled after the residual analysis. Below, you can find the some metrics belonging to initial regression model:

Mean Squared Error: 744500537.4062554

R-squared: 0.3751688255835991

Coefficients

Interest Rate: 2974.72601894

Sahibinden Search Interest: 3350.35148904

İş İlanı Search Interest: -2937.10134999

Fed Faiz Search Interest: 23.28055203

Konut Kredisi Search Interest: -1748.56408495

Araba İlanı Search Interest: -904.78589036

By looking at the signs of the coefficients, we can say that people have tendency to buy house with increasing interest rate. Also, when Fed news are released, the house market is affected positively in terms of the decrease in house prices, therefore, people tend to buy more houses. Interestingly, “iş ilanı” search interest and “konut kredisi” search interest affect the number of house sales with a negative relationship. It may be caused by the increasing interest rate also affects the mortgage rates and in the end of 2023, interest rate has increased.

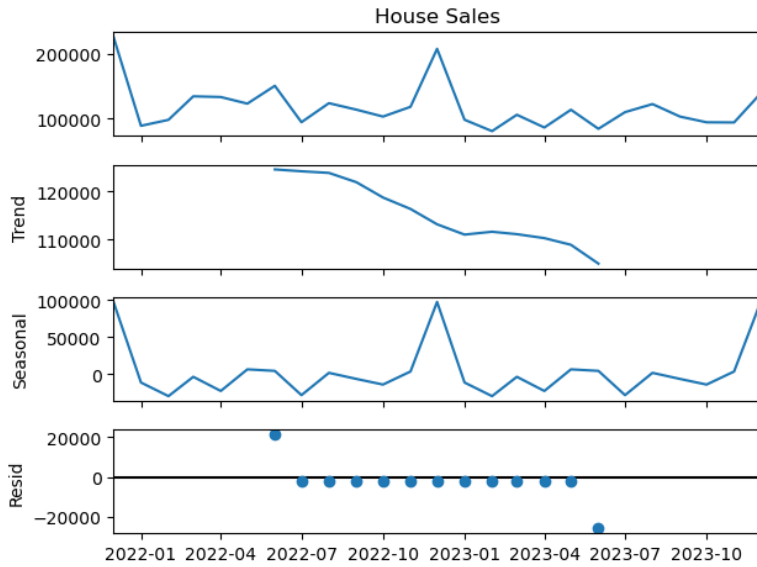


Figure 4 shows the trend and seasonality of the house sales data and gives the signs to make some adjustments on the regression model like using, autocorrelation function, differencing method, and taking logarithm of the house sales data.

Figure 4: Decomposition Results of House Sales Regression

Residual Analysis and Model Revision

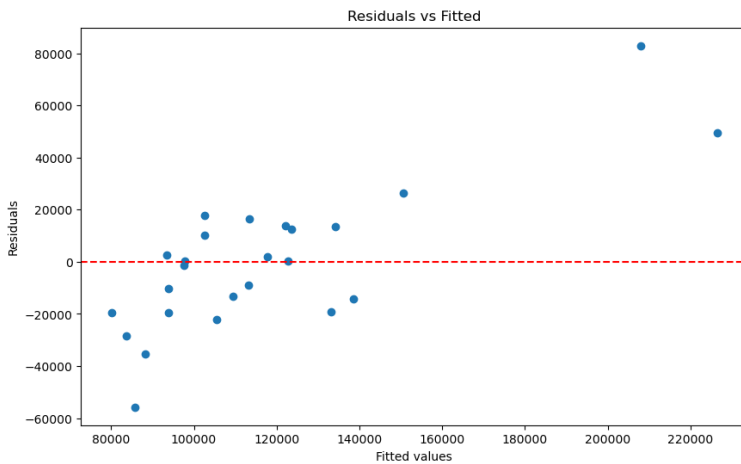


Figure 5: Residuals vs Fitted values in House Sales Regression Model

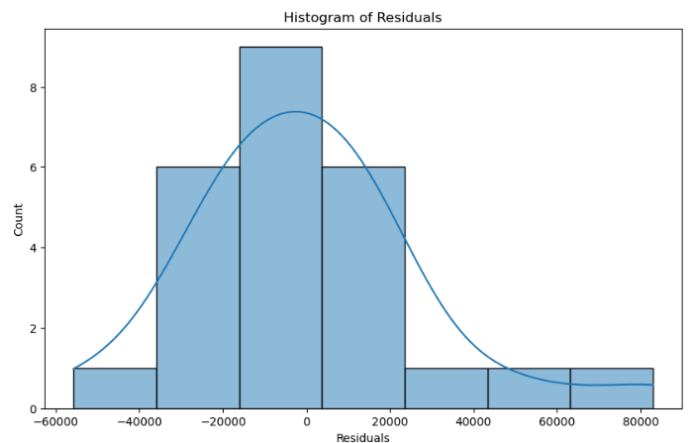


Figure 6: Histogram of the residuals

As the Figure 3 also suggests there is a trend in the residuals vs fitted scatter plot. We will apply differencing method and also check autocorrelation function.

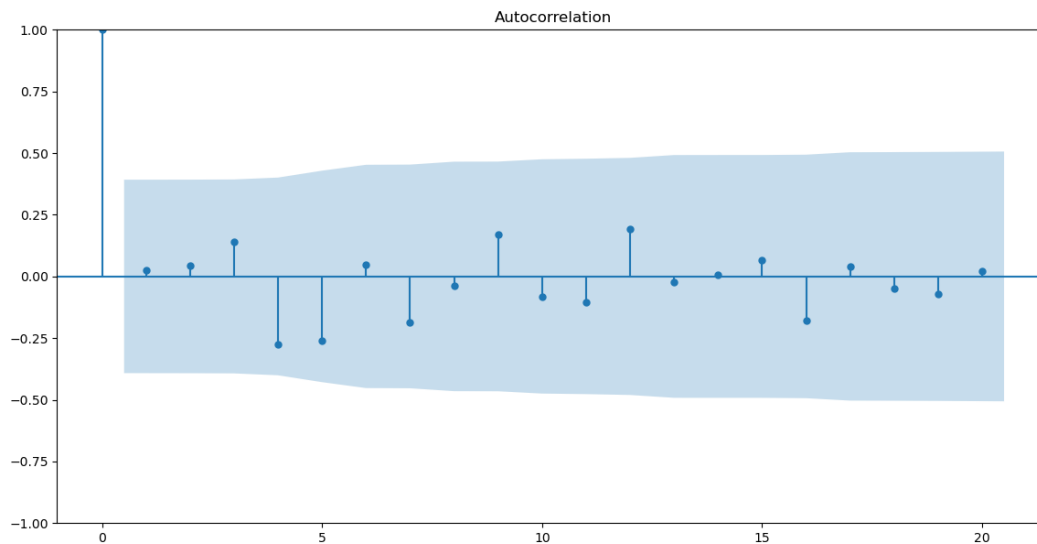


Figure 7: Autocorrelation Function in House Sales Regression

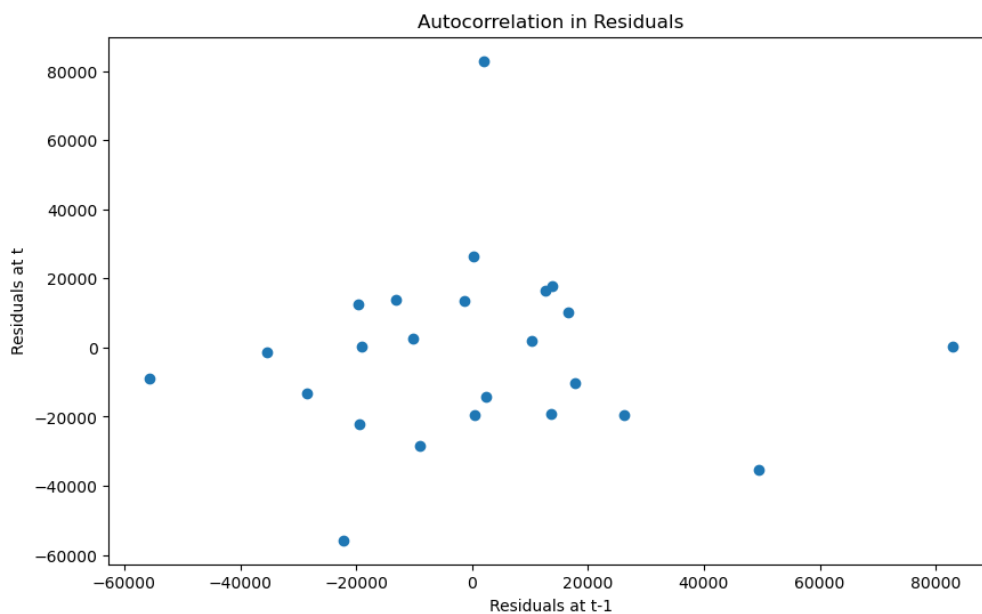


Figure 8: Residuals in t versus $t-1$

Autocorrelation function in figure 7 suggests that autocorrelation adjustment with lag 4 will give better results in terms of explaining the variability in the model.

To revise the model, we will first try the differencing method. By adding $X(\text{time})^2$ to the model, we achieved the 'Time' variable to act like a trend variable in the regression model. Afterwards, differencing method is used to get rid of the autocorrelation between t and $t-1$. In figure 9, one can see the residuals vs fitted values after using the differencing method. There is remaining increasing trend in the residuals. So, logarithm of house sales is regressed against the independent variables on top of the differencing method.

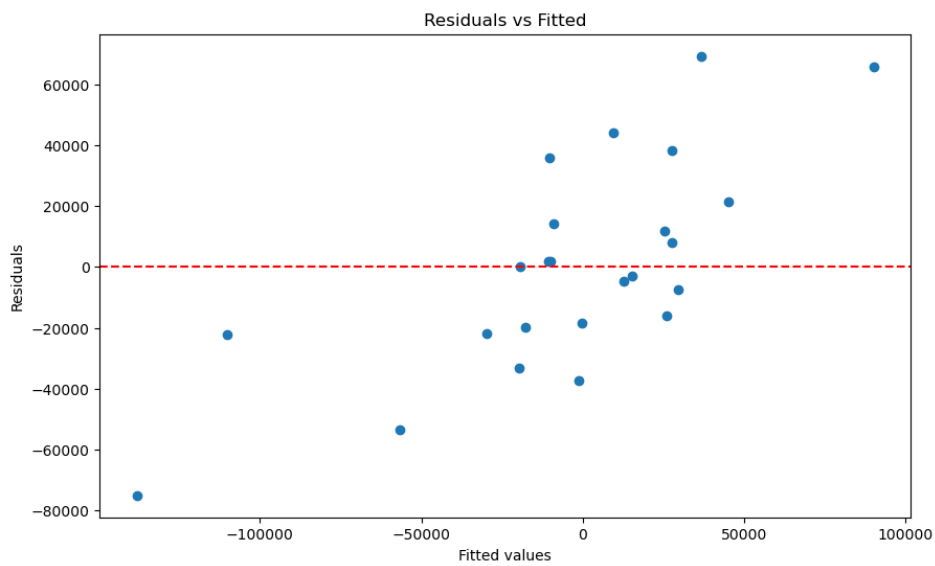


Figure 9: Residuals vs Fitted values after differencing method in House Sales Regression

When the logarithm approach is applied on top of the differencing method, the R^2 value improves drastically. Below you can see the important metrics of the regression after differencing and logarithm approaches are applied.

Mean Squared Error: 0.026874175256056954

R-squared: 0.7753439131692001

R^2 value has increased from 0.37 to 0.77. Also, the residuals started to act like randomly scattered and they resembled to the normal distribution as one can see from the figure 10 and figure 11.

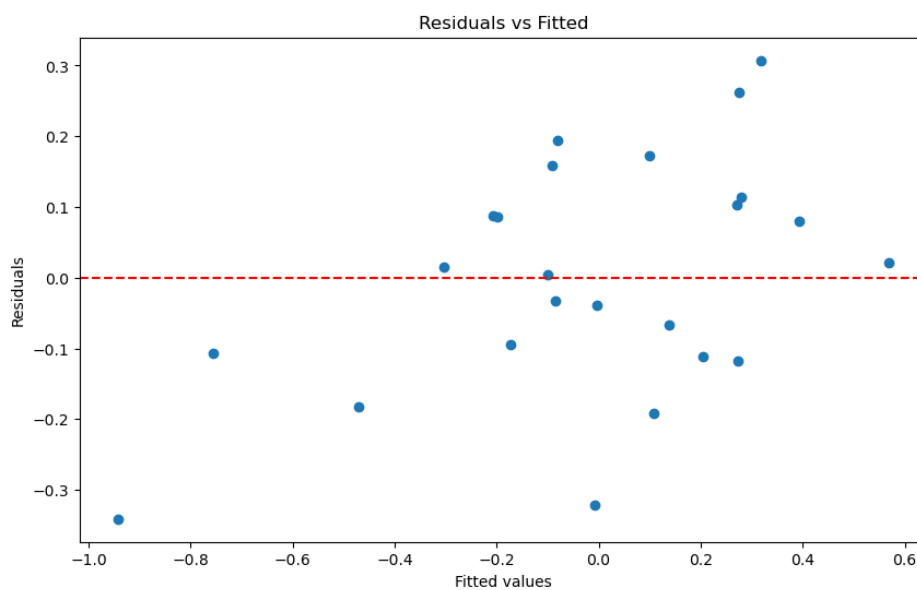


Figure 10: Residuals vs Fitted values after logarithm method in House Sales Regression

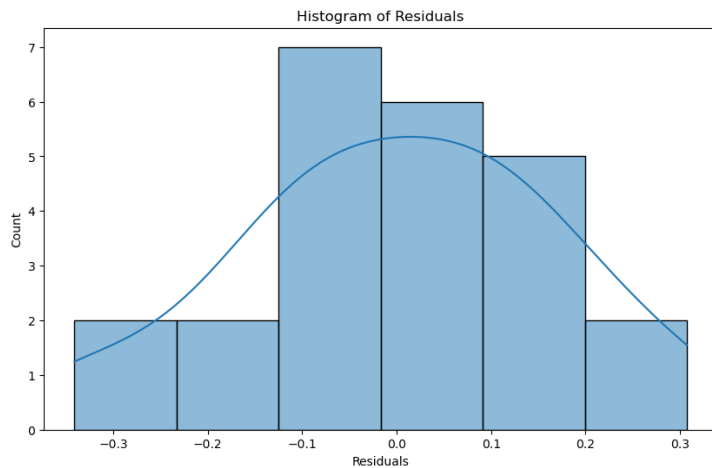


Figure 11: Histogram of residuals after logarithm method in House Sales Regression

However, there is still room for improvement. As we have decided when analyzing figure 7, autocorrelation effect can be handled via adjustment with lag 4. House sales variables are shifted by 4 periods. Then we have the best R^2 value as you can see from the below metrics:

Mean Squared Error: 0.00910721556713814

R-squared: 0.9032318853448605

Intercept: -15.919840850249242

Coefficients:

Interest Rate: -3.15515144e-06

Sahibinden Search Interest: 1.76043612e-02

İş İlanı Search Interest: 2.49639430e-02

Fed Faiz Search Interest: -4.68969155e-03

Konut Kredisi Search Interest: -8.75186902e-03

Araba İlanı Search Interest: -2.45021480e-02

Final Model

When we look at the coefficients in the regression model, the signs of the coefficient make more sense compared to the initial coefficients. For example, when the interest rate increases, the number of houses sold will decrease because the cost of borrowing money will increase. Also, when people search for Sahibinden platform more, the number of houses sold will increase. The number of “İş ilanı” search also increases the number of houses sold because it is a representation of the wealth. Moreover, when FED increases the interest rate, the TL will be devalued and there will be a decrease in the house sales. “Araba ilanı” search interest is also logical because people prefer either car or house in a 2-year period and the cars protected their value better in the past two years. The only unexpected result belongs to the “Konut kredisi” search interest because I expected it to have a positive relation between house sales. But, it can be reasonable if we think of the increasing interest rate because it affects the mortgages.

In figure 12, one can see the final residual performance in the House Sales regression model. It is close to the White noise series and Figure 13 is closer to the Normal distribution. Also, the final R^2 value performed well (0.90).

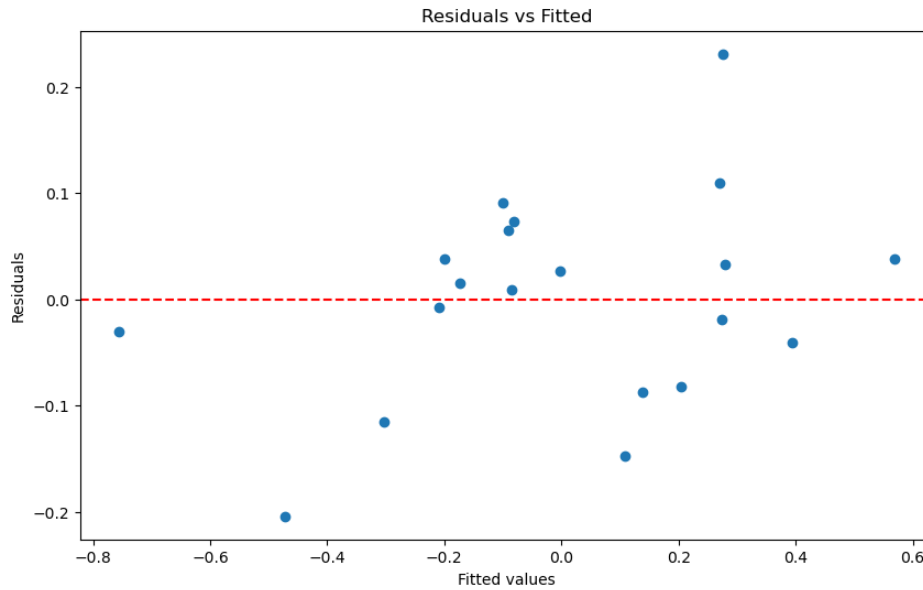


Figure 12: Ultimate Residuals vs Fitted Graph for the House Sales Model

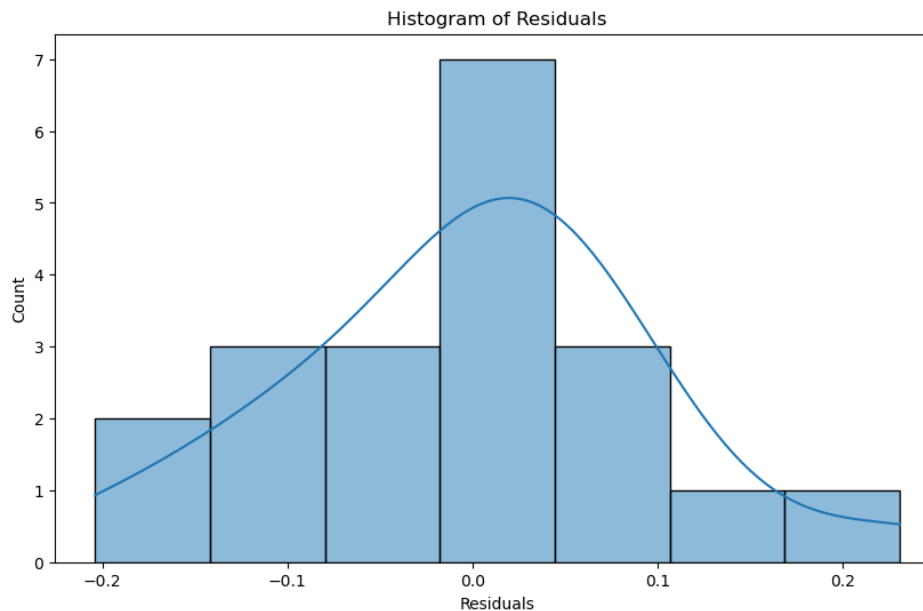


Figure 13: Ultimate Residuals vs Fitted Graph for the House Sales Model

Second Regression - Gold Production Analysis:

Target Variable: Gold Production in Turkey

Independent Variables:

International Gold Reserve: The amount of gold held by international central banks and major financial organizations, influencing global gold prices and potentially national production decisions.

Altın kaç TL Search Interest: Online search volume for "Altın kaç TL" (How much is gold in Turkish Lira?), indicating public interest in gold prices, possibly reflecting or affecting local gold market dynamics.

Elazığ Maden Olayı Search Interest: Online search volume related to a mining incident in Elazığ, which could impact perceptions of mining safety, regulations, or gold production capacities.

Fed Faiz Search Interest: As above, indicating the global economic climate's influence on gold as an investment commodity.

Exchange Rate: The value of Turkey's currency in relation to others, affecting gold production costs and profitability.

Data Manipulation

As in the previous regression model, the same 2 function for Google Trends data and CBRT data are used to organize data.

Correlation Between Independent Variables

	Gold Production	International Gold Reserve	Altın kaç TL Search Interest	Elazığ Maden Olayı Search Interest	Fed Faiz Search Interest	Exchange Rate
Gold Production	1.000000	-0.021554	0.295240	-0.147589	-0.125673	-0.304648
International Gold Reserve	-0.021554	1.000000	-0.289371	0.131253	-0.239947	0.638112
Altın kaç TL Search Interest	0.295240	-0.289371	1.000000	-0.484116	-0.139616	-0.701420
Elazığ Maden Olayı Search Interest	-0.147589	0.131253	-0.484116	1.000000	0.204457	0.551499
Fed Faiz Search Interest	-0.125673	-0.239947	-0.139616	0.204457	1.000000	-0.155597
Exchange Rate	-0.304648	0.638112	-0.701420	0.551499	-0.155597	1.000000

Figure 14: Correlation matrix for the independent variables and target variable for the Gold Production Regression Model

As one can see from the figure 13, all independent variables have small correlation coefficients with the target variable Gold Production in Turkey. Exchange rate has more than 0.5 correlation coefficient with some of the independent variables such as international gold reserve, “Altın kaç TL” search interest and “Elazığ maden olayı” search interest. However, it is not discarded from the analysis because exchange rate will have a valuable insight for the Gold production due foreign currency based raw material procurement and Gold export from Turkey. Other than the correlation coefficient of exchange rate between these three variables, all of the independent variables have less than 0.5 correlation coefficient among each other.

Visualization of the variables

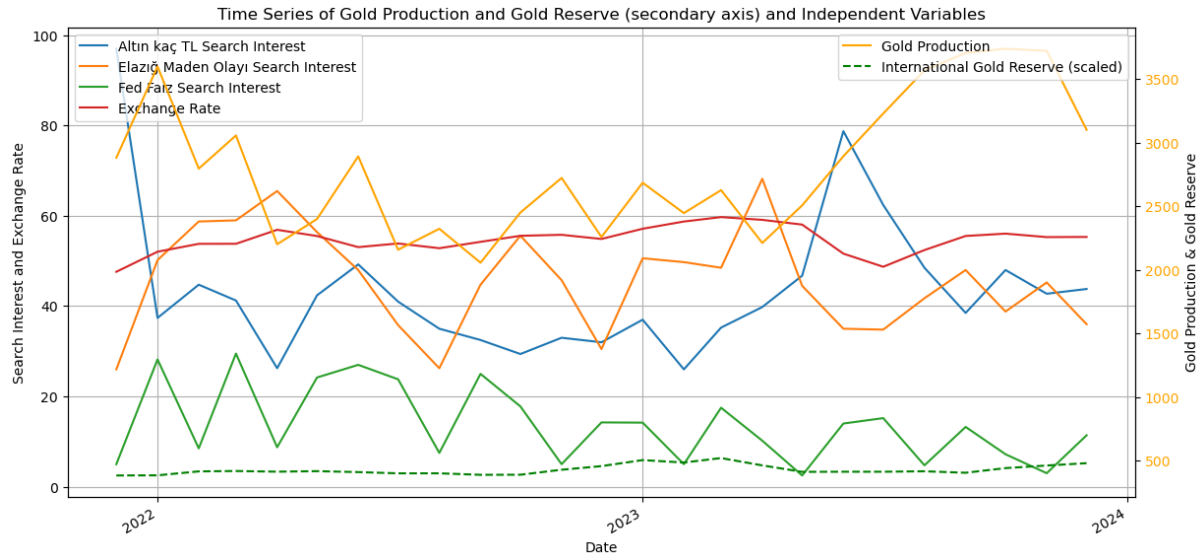


Figure 15: Visualisation of the variables in the Gold Production Regression Model

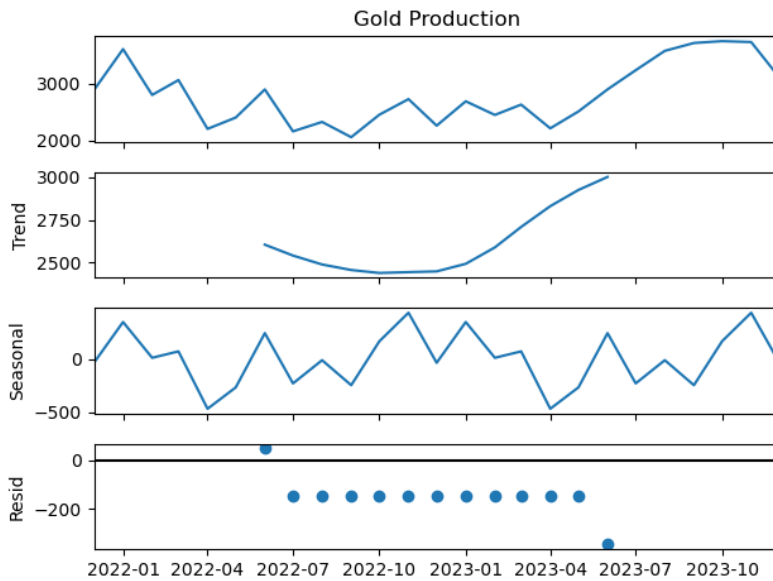
As one can see from the figure 15, gold production in Turkey has an increasing trend especially in the end of 2023. “Altın kaç TL” Search Interest has also increasing trend in the second half of 2023. Also, a seasonality is observed in the “Fed Faiz” Google Trends data, which is expected because of the seasonal meetings held by FED to decide on the interest rate for the United States. Interestingly, “Elazığ Maden Olayı” Search interest has a seasonality which I was expecting to see a peak point in the time it happened.

Regression Model fitting

At first, initial regression model is fitted and residuals are observed. Afterwards, the methods to get rid of effects of trend and seasonality are chosen. Below you can see the important metrics for the first fitted model:

Mean Squared Error: 223602.9023629319
R-squared: 0.18692258686716068
International Gold Reserve: 4.57090703e-0
Altın kaç TL Search Interest: 2.14020337e-01
Elazığ Maden Olayı Search Interest: 1.01604307e+01
Fed Faiz Search Interest: -1.22149977e+01
Exchange Rate: -1.19680162e+02

At first, R-squared value is so small, which indicates that we need to handle autocorrelation and trends in the line chart we have analyzed. Secondly, when the signs of the coefficients are analyzed, we see negative sign in front of Fed Faiz search interest, which is natural because when Fed increases the interest rate, it is more appealing to put money in the dollar instead of Gold. So, it might decrease the gold production. Similar idea applies for also the exchange rate. When exchange rate increases, it is more logical to put money on the foreign currencies.



As seen from figure 16, There is an increasing trend in the gold production residuals data, so, we can get rid of this trend by using time differencing. Also, as we did before, we can take the logarithm of the gold production data. Autocorrelation function can be utilized for the seasonality.

Figure 16: Decomposition Results of Gold Production Regression

Residual Analysis and Model Revision

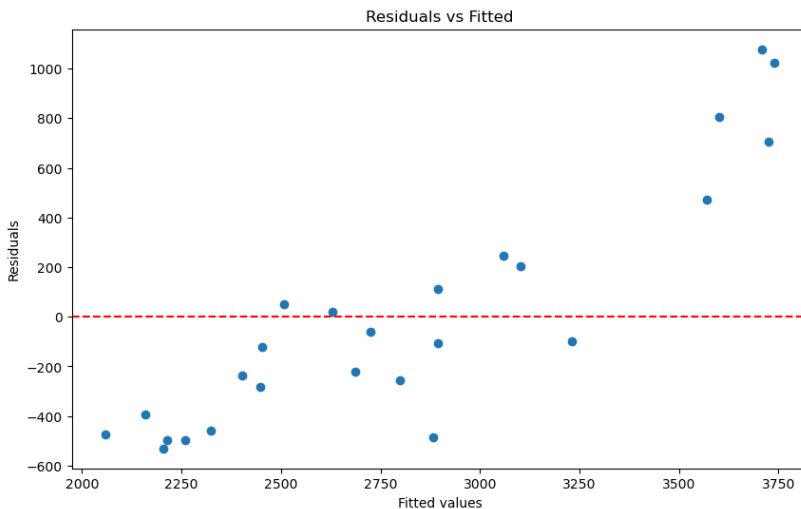


Figure 17: Residuals vs Fitted graph for Gold Production Regression

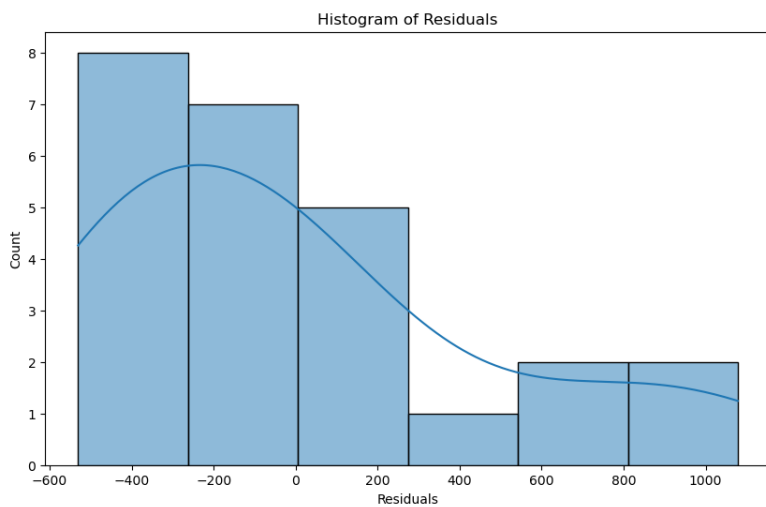


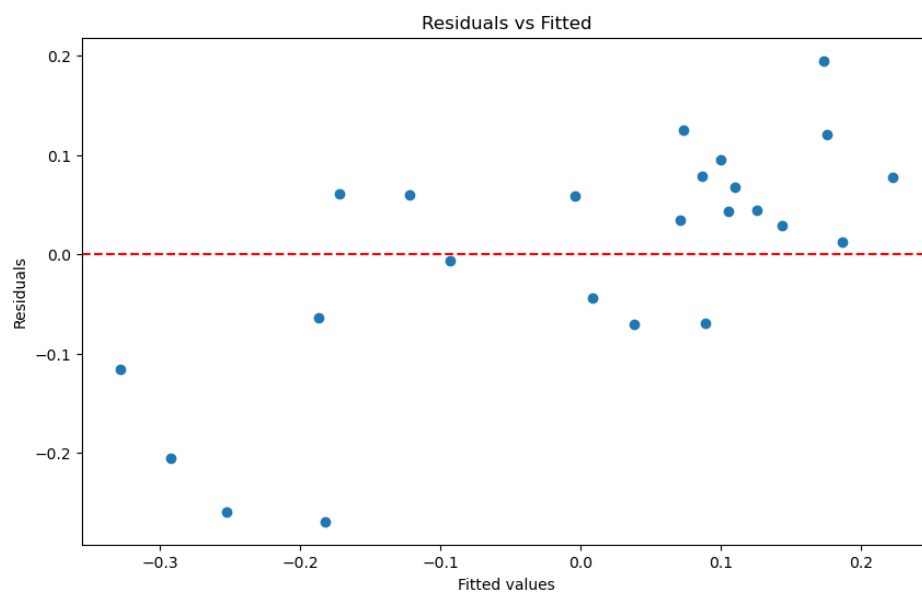
Figure 18: Histogram of Residuals for the Gold Production Regression

As one can see from the initial analysis of the regression, the residuals are not obeying either constant variance assumption or normal distribution assumption. First, differencing method for the trend in the data is used and logarithm of Gold production is taken. For the differencing method, By adding $X(\text{time})^2$ to the model, we achieved the 'Time' variable to act like a trend variable in the regression model. After applying these changes, the important metrics for our regression will be as follows:

Mean Squared Error: 0.013422475152818572

R-squared: 0.4798271852495699

The R-squared value improved well but there is still an unexplained variability in the model with our regression. As one can see from the figure 19, residuals are still concentrated on the top right corner. Therefore, autocorrelation function will be analysed.



*Figure 19:
Residuals vs
Fitted graph after
using logarithm
approach and
differencing
method*

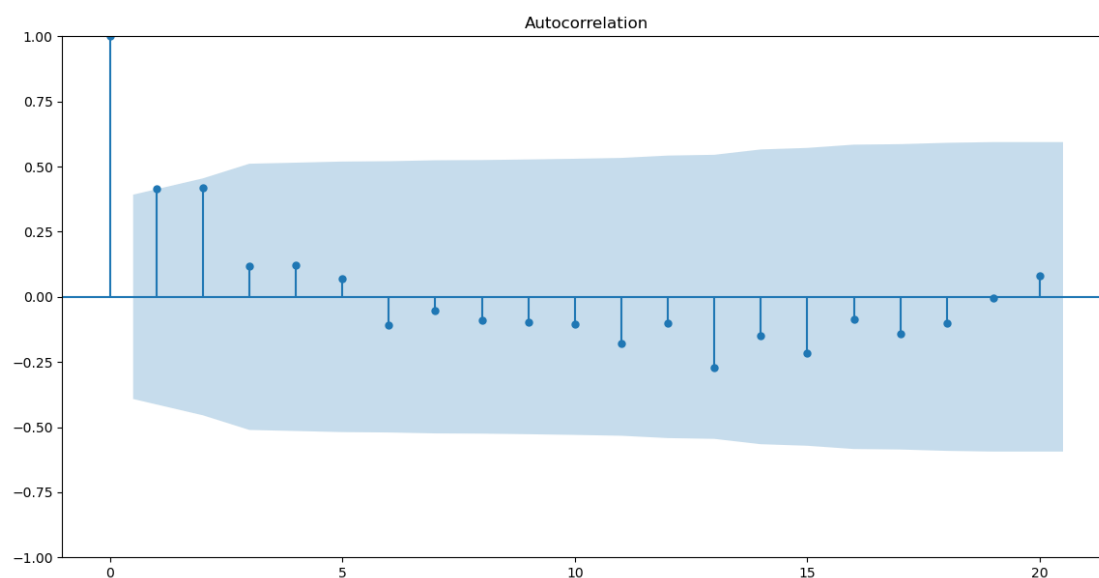
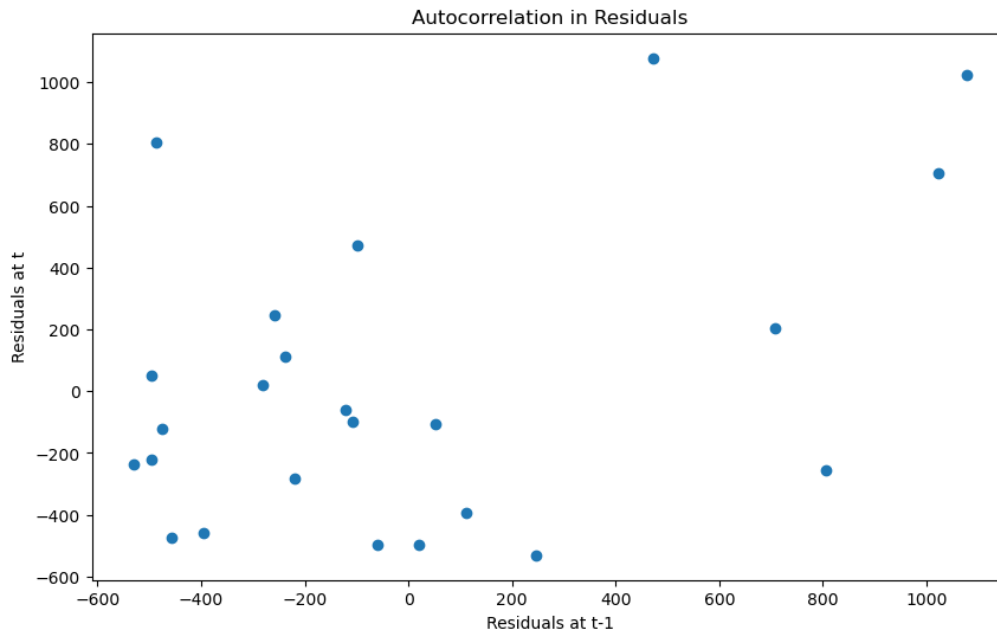


Figure 20: Autocorrelation function for the Gold Production Regression



*Figure 21:
Autocorrelation in
residuals in t
versus $t-1$*

Autocorrelation function in figure 20 and 21 suggests that autocorrelation adjustment with lag 6 will give better results in terms of explaining the variability in the model. Below you can see the performance metrics after adding the autocorrelation variable with lag 6.

Mean Squared Error: 0.003817486773003428

R-squared: 0.8028771434298683

Coefficients:

International Gold Reserve: -1.46987860e-03

Altın kaç TL Search Interest: -1.18101174e-05

Elazığ Maden Olayı Search Interest: -3.03751208e-04

Fed Faiz Search Interest: 3.58249861e-03

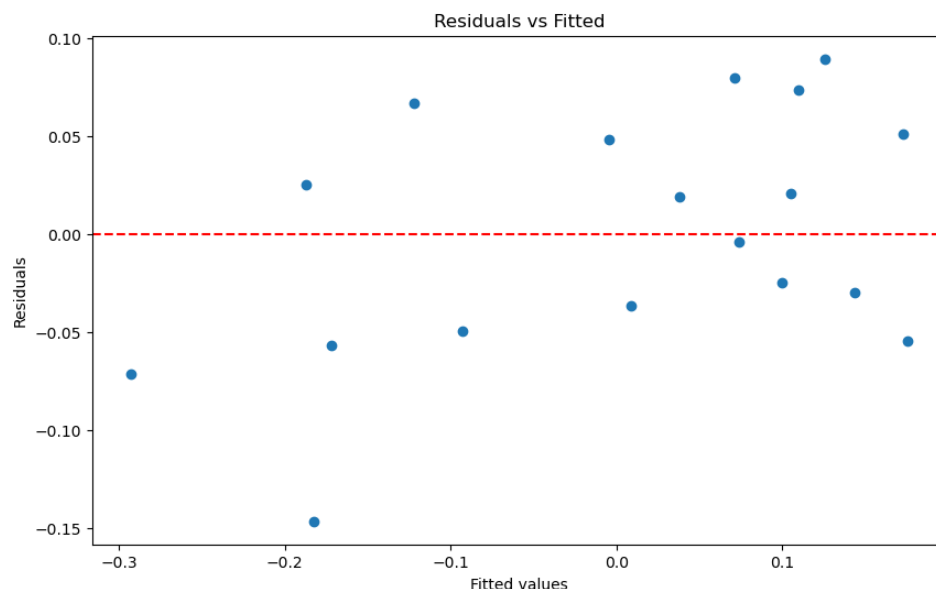
Exchange Rate: -3.14436964e-03

Final Model

In the regression model predicting gold production in Turkey, various factors were shown to influence production levels, as indicated by the sign of their coefficients. An increase in international gold reserves was associated with a decrease in gold production, suggested by a negative coefficient, possibly reflecting reduced domestic production due to higher global supplies or lower prices. Similarly, higher online search interest in the price of gold in Turkish Lira and increased public concern regarding mining incidents in Elazığ were both found to be negatively associated with gold production, implying that public interest in gold prices and safety concerns might not lead to increased production, perhaps due to market or operational constraints. Conversely, a positive relationship was observed between search interest in Federal Reserve interest rates and gold production, indicating that anticipations or reactions to global economic policies could encourage higher production, potentially as a response to economic uncertainties or to take advantage of favorable gold prices. Lastly, the negative coefficient for the exchange rate suggested that an appreciation of Turkey's currency is linked to a decrease in gold production, potentially due to the reduced competitiveness of

gold exports or lower local currency returns from international sales, thereby possibly disincentivizing production.

The R-squared value improved very well, which means we can explain the variability by our model.



*Figure 22:
Ultimate
Residuals vs
Fitted graph for
Gold Production
Regression*

Finally, as seen from figure 22, residuals are relatively randomly scattered around zero and we got rid of the trend

Third Regression - Foreign Investment Analysis:

Target Variable: Investment in Turkey from abroad

Independent Variables:

Invest in TR Search Interest: Online search volume for investing in Turkey, indicating international interest in Turkish investment opportunities.

Exchange Rate: Influences the attractiveness of investment in Turkey by affecting the value of potential returns.

Politics TR Search Interest: Online search volume related to Turkish politics, which could indicate concerns or interest in the political stability and policies affecting investment climate.

Fed Faiz Search Interest: Reflects concern or interest in global economic policy's impact on investment climates, including Turkey.

Inflation TR Search Interest: Online search volume for inflation in Turkey, indicating concerns about the economic stability and the value of investments.

Data Manipulation

As in the previous regression model, the same 2 function for Google Trends data and CBRT data are used to organize data.

Correlation Between Independent Variables

	Investment in TR	Invest in TR Search Interest	Exchange Rate	Politics TR Search Interest	Fed Faiz Search Interest	Inflation TR Search Interest
Investment in TR	1.000000	0.155492	-0.561380	-0.539553	0.475487	0.535247
Invest in TR Search Interest	0.155492	1.000000	-0.337088	-0.277179	-0.048183	0.210638
Exchange Rate	-0.561380	-0.337088	1.000000	0.801836	-0.155597	0.012490
Politics TR Search Interest	-0.539553	-0.277179	0.801836	1.000000	-0.359635	-0.054644
Fed Faiz Search Interest	0.475487	-0.048183	-0.155597	-0.359635	1.000000	0.355051
Inflation TR Search Interest	0.535247	0.210638	0.012490	-0.054644	0.355051	1.000000

Figure 23: Correlation Matrix for the independent variables and the target variable in the Investment in Turkey Regression Model

As seen from figure 23, investment in Turkey only have a correlation higher than 0.5 with inflation in Turkey. However, it is not discarded from the analysis because the effect of inflation has a complex relationship with the investors abroad and should be examined via regression model. For example, if an investor made an investment in a company working with a large inventory, they need to adjust their investment with the adjusted stock value of the company. However, if the investors abroad made an investment in properties like land, building, inflation may affect them positively. Other than inflation, our target variable doesn't have a large correlation with the other independent variables.

Visualization of the variables

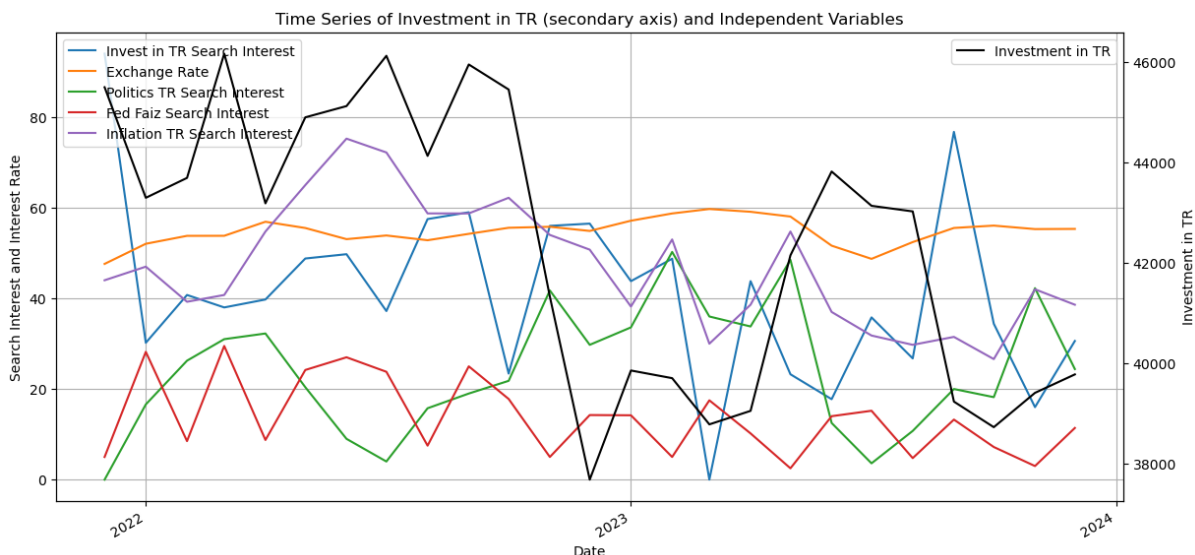


Figure 24: Visualisations of the variables in Investment in Turkey Regression

Seasonal fluctuations are apparent in the search interest rates, especially noticeable in the "Invest in TR Search Interest" and "Fed Faiz Search Interest," which exhibit peaks and troughs at regular intervals, possibly indicating cyclical patterns in public interest or economic events. The "Investment in TR," shown on the secondary axis, has a decreasing pattern in the end of the year 2022, and an increasing pattern in the mid of 2023.

Regression Model fitting

At first, initial regression model is fitted, and residuals are observed. Afterwards, the methods to get rid of effects of trend and seasonality are chosen. Below you can see the important metrics for the first fitted model:

Mean Squared Error: 2430782.4503818364

R-squared: 0.6691962466433052

Invest in TR Search Interest: -19.35874465

Exchange Rate: -525.05069235

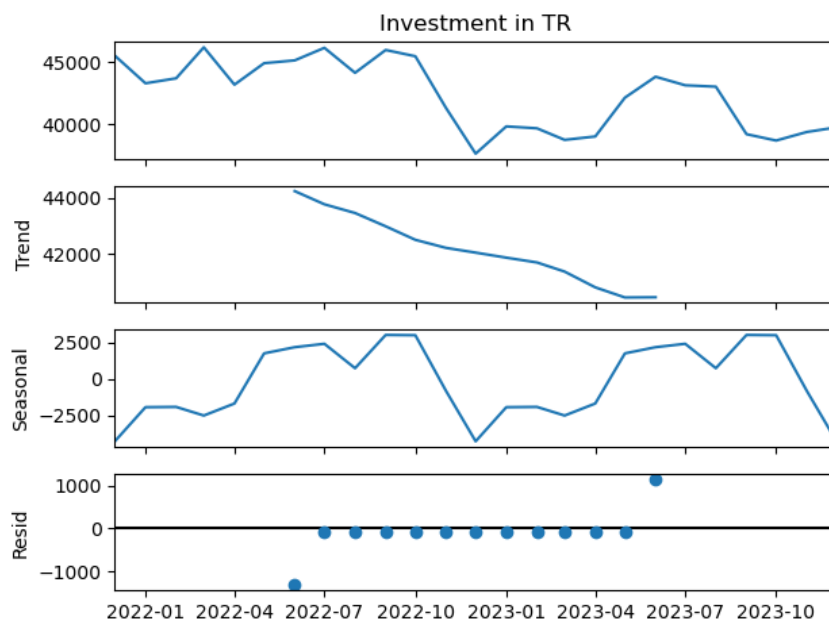
Politics TR Search Interest: -7.31289445

Fed Faiz Search Interest: 62.74555671

Inflation TR Search Interest: 102.33693547

From the signs of the coefficients, it is observed that when people looked up more information about investing in Turkey, when the exchange rate went up, or when there was more interest in Turkish politics, the amount of investment from abroad went down. On the other hand, when there was more interest in the U.S. Federal Reserve's interest rates or in inflation in Turkey, investments from other countries went up. These results suggest that people's interest in economic and political issues, as well as changes in the value of money, can have different effects on how much money comes into Turkey from outside.

For an initial model with no adjustment on the seasonality and trend in the data, 0.67 is a good start for the R-squared value.



As seen from figure 25, There is a decreasing trend in the investment in Turkey residuals data, so, we can get rid of this trend by using time differencing. Also, as we did before, we can take the logarithm of the gold production data. Autocorrelation function can be utilized for the seasonality.

Figure 25: Decomposition Results of Investment in Turkey

Residual Analysis and Model Revision

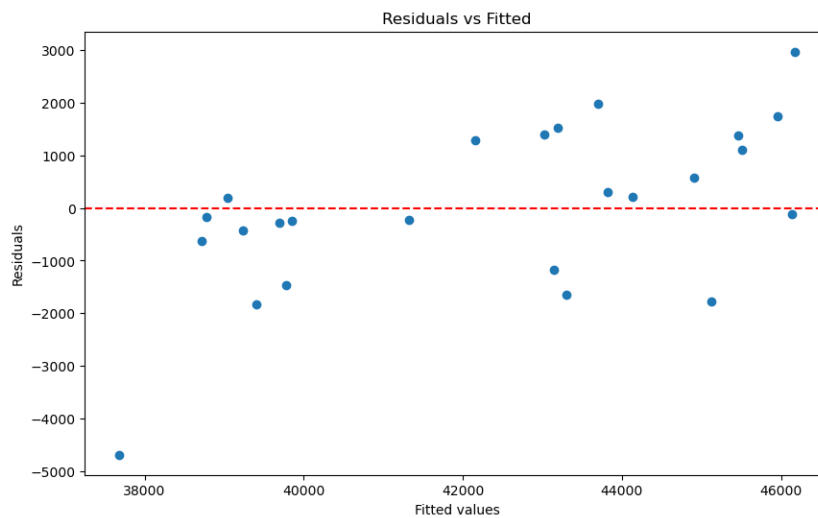


Figure 26: Residuals vs Fitted values scatter plot for Investment in Turkey Regression Model

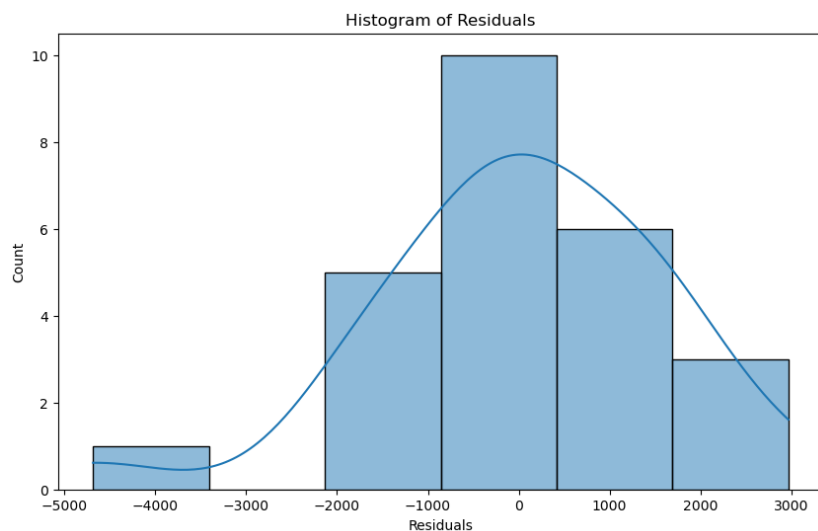


Figure 27: Histogram of Residuals for Investment in Turkey Regression Model

As one can see from the initial analysis of the regression, the residuals are not obeying either constant variance assumption or normal distribution assumption. In the residual vs fitted values plot, the trend component is not as significant as the previous two models, therefore, first the autocorrelation function will be examined, then, the adjustments will be made.

As seen from the figures 28 and 29, the data is autocorrelated because it is concentrated on the plot showing the residuals at t and $t-1$. Also, for the first 2 component of autocorrelation function, there is a decreasing pattern in the positive side, so autocorrelation adjustment with lag 2 can be used to get rid of the effect of dependency on the past data.

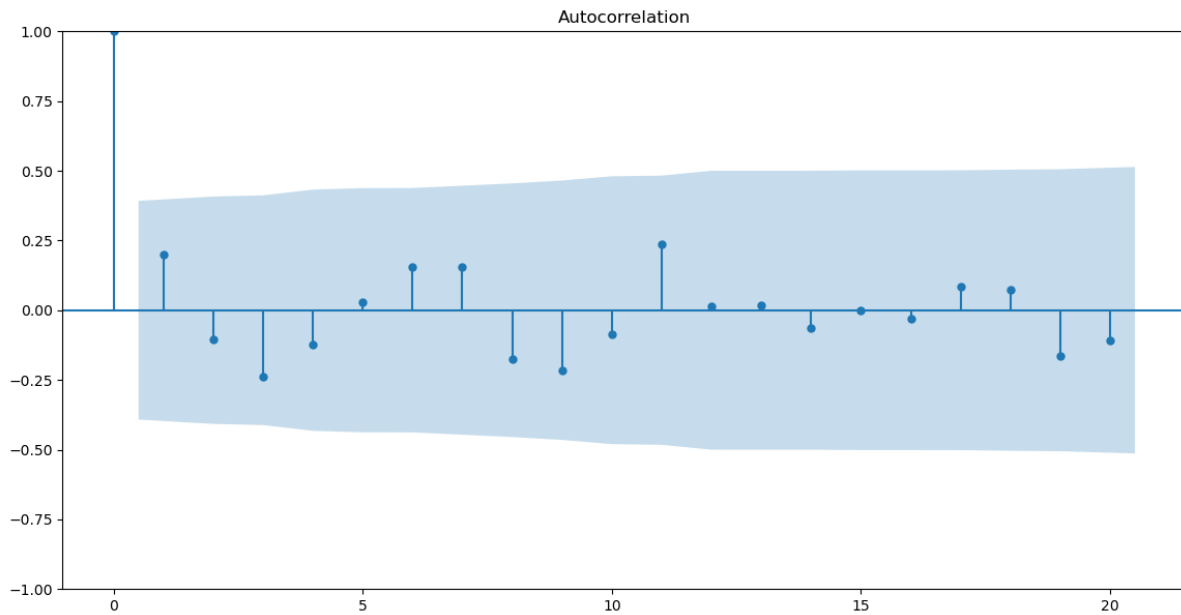


Figure 28: Autocorrelation function for Investment in Turkey Regression

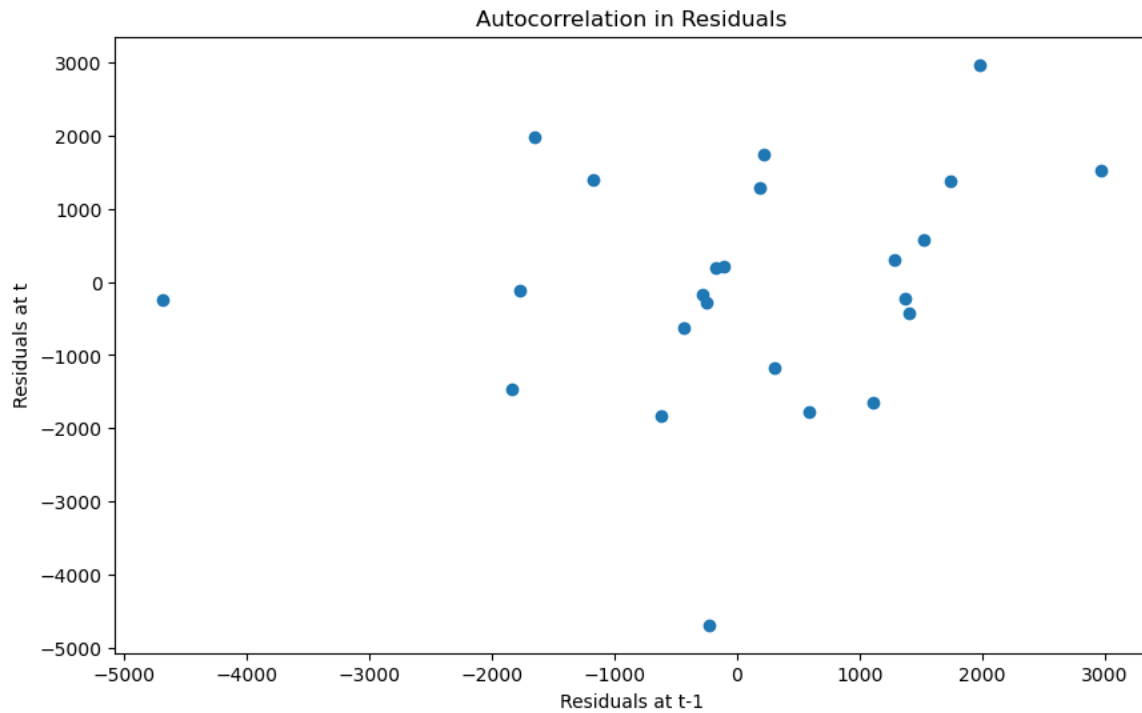


Figure 29: Residuals at t versus $t-1$ for Investment in Turkey Regression

After shifting the data with lag 2, as discussed above, an improved R-squared value is obtained. Below, you can see the ultimate model metrics after applying autocorrelation lagging methodology.

Mean Squared Error: 2150190.5133476737

R-squared: 0.7216221664097295

Coefficients:

Invest in TR Search Interest: -4.00720584e+01

Exchange Rate: -4.93073319e+02

Politics TR Search Interest: 1.54454693e+01

Fed Faiz Search Interest: 1.05979988e+02

Inflation TR Search Interest: 9.88968679e+01

Final Model

In the regression analysis focused on investment in Turkey, the decision to employ only the autocorrelation method, without resorting to differencing or logarithmic transformation, was guided by the nature of the data and the specific characteristics of the investment time series. The investment data exhibited stability and lacked the type of trends or volatility that would necessitate stabilizing transformations. By applying the autocorrelation method, any underlying patterns in the timing of investments were effectively captured, allowing for an analysis that preserved the original scale and interpretability of the investment data. This approach ensured a straightforward understanding of the relationship between investment levels and the independent variables. From the figure 30, we can conclude that the residuals act like white noise series.

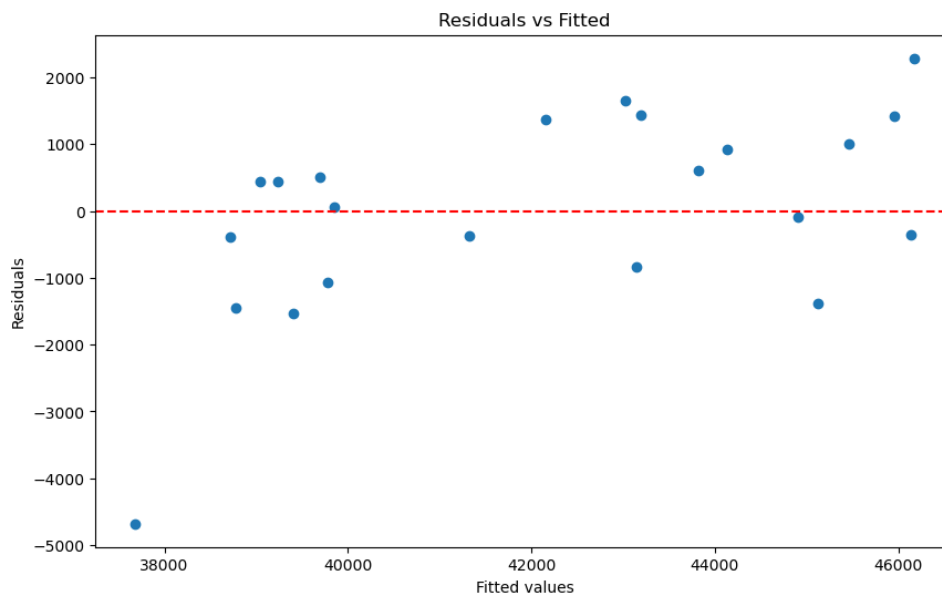


Figure 30: Ultimate residuals vs fitted values scatter plot for Investment in Turkey

Conclusion

In conclusion, this homework looked at three different areas of Turkey's economy: house sales, gold mining, and investments from other countries. We used techniques like taking logarithms, making differences in the data, and looking at how the data points are related to each other over time to make sure our models worked well with the trends and seasonal patterns we saw. These methods helped us create better models to understand how different factors affect these parts of the economy. By using both usual economic data and new types of data from Google Trends, we learned that our search interest in Google search engine can also reflect on the economic variables. Moreover, the most important remark for this

homework is the fact that working with time series data always need adjustment on the dependencies on the past data such as autocorrelation handling because once the autocorrelation lag is added to the model, the R-squared values have improved drastically.

Large Language Model Prompts

Prompt: how to delete day component in date in dataframes

ChatGPT Response:

To remove the day component from a datetime object in a DataFrame and keep only the year and month components, you can use the `.dt.to_period('M')` method on your datetime series. This converts the datetime objects to Period objects, effectively removing the day component and leaving you with year-month granularity. If you later need to convert these Period objects back to datetime, you can use `.dt.to_timestamp()` which will set the day to the first of the month by default.

Here's how you can apply this to a DataFrame:

```
import pandas as pd
```

```
# Assuming df is your DataFrame and it has a 'Date' column in datetime format
```

```
df['Date'] = pd.to_datetime(df['Date']) # Ensure 'Date' is in datetime format
```

```
df['YearMonth'] = df['Date'].dt.to_period('M')
```

```
# If you need to perform operations that require datetime objects (like setting an index),
```

```
# you can convert the 'YearMonth' back to datetime, which will default the day to the first of each month
```

```
df['YearMonth'] = df['YearMonth'].dt.to_timestamp()
```

Example:

Let's assume you have a DataFrame with a Date column and you want to remove the day component, leaving only the year and month.

As you can see, the YearMonth column now represents the first day of each month, effectively removing the specific day component from the original Date column. This can be particularly useful for time series analyses where the specific day is not relevant, and you're more interested in monthly trends or patterns.

Prompt: before fitting the regression equation, I want to visualize the independent variables and dependent variable in a time series chart. Do you have suggestion?

ChatGPT Response:

Visualizing time series data can provide valuable insights into trends, seasonality, and relationships between variables. In Python, you can use libraries such as matplotlib, seaborn,

or plotly for visualization. Matplotlib is widely used and quite versatile, so I'll show you how to create a time series chart with it.

Here's a general way to plot your time series data using matplotlib:

```
import matplotlib.pyplot as plt
import matplotlib.dates as mdates

# Let's assume 'merged_df' is your DataFrame and it's indexed by date
# 'House Sales' is your dependent variable and the rest are independent variables

# First, we plot the dependent variable 'House Sales'
plt.figure(figsize=(14, 7))
plt.plot(merged_df.index, merged_df['House Sales'], label='House Sales', color='blue')

# Next, plot each of the independent variables on the same chart
for column in merged_df.columns.drop('House Sales'):
    plt.plot(merged_df.index, merged_df[column], label=column)

# Formatting the plot
plt.title('Time Series of House Sales and Independent Variables')
plt.xlabel('Date')
plt.ylabel('Value')
plt.legend()
plt.grid(True)

# Improve formatting of date axis
plt.gca().xaxis.set_major_locator(mdates.YearLocator())
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%Y'))
plt.gcf().autofmt_xdate() # Rotate date labels to prevent overlap
```



```
plt.show()
```

This code creates a time series chart where the dependent variable 'House Sales' is plotted alongside each independent variable. Each series is labeled for clarity, and we use a grid to make it easier to follow the values across the chart. The date axis is also formatted to show only the year, preventing overlap of date labels.

Before running this code, ensure that 'merged_df' has a `DateTimeIndex`, and that all columns are properly named and contain numeric data suitable for plotting. If your `DataFrame` contains a lot of independent variables, the chart may become cluttered, so you may need to plot only a subset of variables or create separate charts for different sets of variables.

Please adjust the variable names and `DataFrame` name to fit your data.

Prompt: can you give me a library to do regression

ChatGPT Response:

Certainly, if you'd like to try an alternative to `statsmodels` for regression analysis, you can use `scikit-learn`, a powerful machine learning library in Python that includes support for linear regression among many other algorithms.

Here's how you can use `scikit-learn` to perform linear regression:

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Prepare the data
# Assuming 'Date' is not in the DataFrame or has been converted to a numeric feature
X = merged_df_house_sales.drop(columns=['House Sales'])
y = merged_df_house_sales['House Sales']

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the Linear Regression model
```

```

regressor = LinearRegression()

# Fit the model to the training data
regressor.fit(X_train, y_train)

# Predict on the test set
y_pred = regressor.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')

# To get the regression coefficients
print('Intercept:', regressor.intercept_)
print('Coefficients:', regressor.coef_)

```

Note that scikit-learn doesn't provide detailed statistics out-of-the-box like statsmodels does. For example, you won't get p-values for your coefficients directly. If you need such detailed statistics, you may have to calculate them manually or stick with statsmodels.

However, scikit-learn is great for making predictions and can be used with many different types of regression models, not just linear regression. It's also typically used in a machine learning context where you may split your data into training and test sets, train the model on the training set, and then evaluate its performance on the test set.

Please ensure you have scikit-learn installed in your Python environment (pip install scikit-learn), and replace merged_df_house_sales with the actual name of your DataFrame when you run the code.

Prompt: check the trend and seasonality, maybe use autocorrelation function

ChatGPT Response:

To analyze trend and seasonality in your time series data, you can use the Autocorrelation Function (ACF). The ACF helps identify the presence of seasonality by showing the correlation of a time series with its own lagged values.

In Python, you can use the statsmodels library to plot the ACF. Here's how to do it:

```
from statsmodels.graphics.tsaplots import plot_acf

# Assuming 'House Sales' is the series you want to analyze
house_sales_series = merged_df_house_sales['House Sales']

# Plot the Autocorrelation Function (ACF)
fig, ax = plt.subplots(figsize=(14, 7))
plot_acf(house_sales_series, ax=ax, lags=40)
plt.show()
```

This will plot the autocorrelation of the 'House Sales' series with lags up to 40. Peaks in the ACF plot at specific lags indicate seasonality. The lag at which the ACF first becomes insignificant is an indication of the season length.

You might also want to decompose the series to explicitly visualize the trend and seasonal components:

```
from statsmodels.tsa.seasonal import seasonal_decompose

# Decompose the time series
decomposition_results = seasonal_decompose(house_sales_series, model='additive',
period=12)

# Plot the decomposed components
decomposition_results.plot()
plt.show()
```

The seasonal_decompose function breaks down the time series into trend, seasonal, and residual components. The period parameter specifies the length of the seasonality cycle (e.g.,

12 for monthly data with annual seasonality). The residual component shows the noise in the data after the trend and seasonal components have been removed.

These tools will provide a good starting point for understanding the underlying patterns in your time series data and guide your choices for building a time series regression model, such as whether to include lagged terms to account for autocorrelation or seasonal dummies to model seasonality.

Prompt: please conduct a residual analysis

ChatGPT Response:

After fitting a regression model, it's important to perform a residual analysis to check that the residuals (the differences between the observed values and the values predicted by the model) behave in a way that's consistent with the assumptions of the model. Here's a step-by-step approach for conducting residual analysis in Python using statsmodels:

Plot the residuals: This can help check for any patterns that might suggest a problem with the model.

Check for normality: If the residuals are normally distributed, it suggests that the model's assumptions are likely being met.

Check for autocorrelation: Autocorrelation in the residuals can suggest that the model is missing a key component, like a trend or seasonal effect.

Assuming you've already fitted an OLS model using statsmodels and stored it in a variable called `model`, here's how you can perform the residual analysis:

```
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns

# Get the residuals from the model
residuals = model.resid

# Plot the residuals to check for patterns
plt.figure(figsize=(10, 5))
plt.plot(residuals, marker='o', linestyle='')
plt.title('Residuals from OLS model')
```

```
plt.xlabel('Index')
plt.ylabel('Residuals')
plt.axhline(y=0, color='r', linestyle='--')
plt.show()
```

Plotting a histogram and a Q-Q plot to check for normality of residuals

```
plt.figure(figsize=(10, 5))
sns.histplot(residuals, kde=True)
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()
```

```
sm.qqplot(residuals, line='s')
plt.title('Q-Q Plot of Residuals')
plt.show()
```

Using Durbin-Watson test to check for autocorrelation in residuals

```
from statsmodels.stats.stattools import durbin_watson
dw_test_stat = durbin_watson(residuals)
print(f'Durbin-Watson test statistic: {dw_test_stat}')
```

Values close to 2 suggest no autocorrelation, values < 2 indicate positive autocorrelation, and values > 2 indicate negative autocorrelation.

The histogram and Q-Q plot should give you a visual sense of whether the residuals are normally distributed. The Durbin-Watson statistic tests for the presence of autocorrelation in the residuals; a value close to 2.0 indicates there is no autocorrelation.

Note: Before running this code, make sure you've already created your OLS model with `model = sm.OLS(...).fit()` and that model contains the results of your fitted model.

Please run this analysis in your local Python environment and interpret the results to understand the behavior of your model's residuals. If you find patterns or non-normality in

the residuals, or if the Durbin-Watson test suggests autocorrelation, you may need to refine your model, possibly by including additional predictors, incorporating lagged variables, or transforming your variables.

Prompt: I want to get rid of the effect of increasing trend in residulas. How can we do that?

ChatGPT Response:

To mitigate the effect of an increasing trend in the residuals of a regression model, you can try several approaches. The goal is to either incorporate the trend directly into the model or transform the data to stabilize the trend. Here are a few strategies:

1. Include a Time/Trend Variable

If the trend is linear, you can add a time variable as a predictor in your model. This time variable can simply be a sequence of integers representing the time steps in your data (e.g., 1, 2, 3, ...).

```
# Assuming 'X' is your predictors DataFrame
```

```
X['Time'] = np.arange(len(X))
```

```
# Now, 'Time' acts as a trend variable in your regression model
```

2. Differencing

Differencing the target variable can help remove trends and seasonality, making the series more stationary. You can difference the series by subtracting the previous observation from the current observation.

```
y_diff = y.diff().dropna() # First difference of the target variable
```

```
# If using 'X' with additional predictors, ensure to align 'X' with 'y_diff'
```

```
X_aligned = X.iloc[1:, :] # Drop the first row to align with 'y_diff'
```

3. Log Transformation

If the trend is exponential, applying a log transformation to the target variable can help linearize the growth, making it easier for the linear model to fit the data.

```
y_log = np.log(y)
```

4. Polynomial Trend

If the trend is not linear but still systematic, you can model it with a polynomial term. This involves creating new predictors that are powers of the time variable.

```
X['Time'] = np.arange(len(X))  
X['Time^2'] = X['Time']**2 # Add a quadratic term  
# You can add higher-degree terms as needed
```

5. Detrending by Model

You can fit a model (like a simple linear regression) to capture the trend and then work with the residuals of this model as your detrended series.

```
from sklearn.linear_model import LinearRegression  
  
# Fit a model to capture the trend  
trend_model = LinearRegression().fit(X[['Time']], y)  
y_trend = trend_model.predict(X[['Time']])
```

```
# Calculate detrended series  
y_detrended = y - y_trend
```

After applying any of these approaches, you would use the transformed y (e.g., y_{diff} , y_{log} , or $y_{\text{detrended}}$) as your dependent variable in the regression analysis. Remember, each approach has its own assumptions and implications, so you should choose one that makes the most sense for your data and the nature of the trend you're observing.

These methods can help make your residuals more homoscedastic (having constant variance) and uncorrelated with time, which are key assumptions for linear regression models.