Data Analytics
Term Project Report

Name: Apurva Anand **A#:** A04285700

For my Term Project I chose a very interesting topic which would test the limits of my coding skill as well as what I've learnt throughout the semester. My topic was:

"Predicting conflicts in a country or overall world using socio-economic factors"

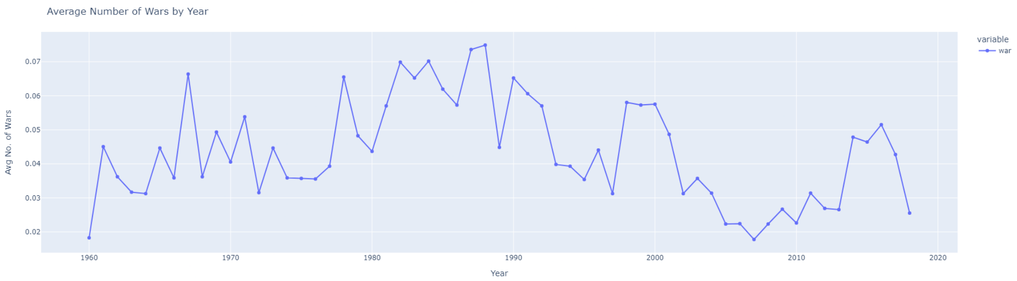
Some of the questions that I had in mind as I was starting this project:

- I wanted to understand how wars/conflicts were progressing?
- How did they start?
- Was there any indication that some sort of conflict was approaching?
- How do conflicts evolve over time? Are they periodical?
- Do economic factors have an impact on conflicts?
- Is a diverse state more prone to civil wars?
- Is there a link between relying too much on natural resources and instability?

The Dataset
I used the UCDP/PRIO Armed Conflict Dataset
Uppsala Conflict Data Program
Department of Peace and Conflict Research

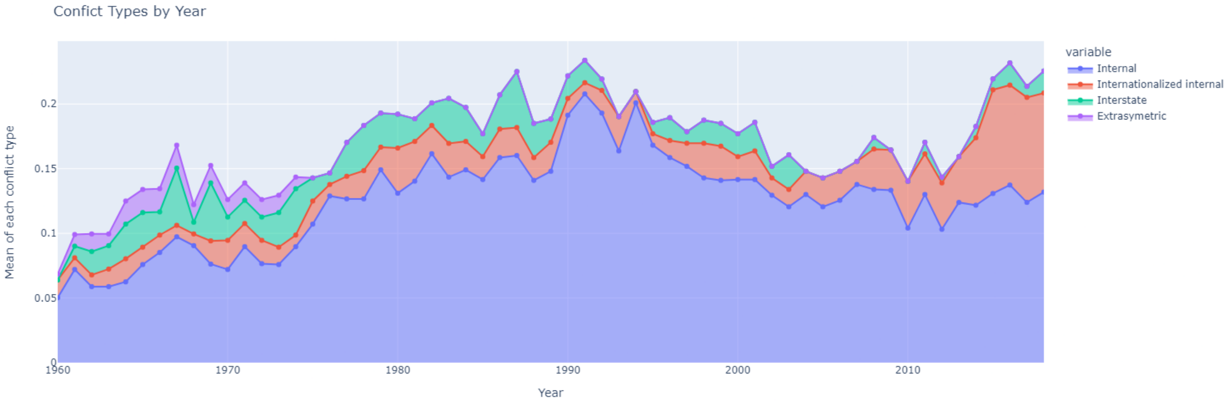
It is a highly detailed dataset which has data on every conflict.war a country has faced since 1960. I found another dataset which had used this and added features such as GDP, Unemployment, Debt, Account Balance, Corruption, Inflation, Religion, Ethnicity, Language, Infant mortality rate, Natural Resources etc.

Some of the questions I had gave answers to the kind of data that was needed. After collecting all my data I had to find the relevant ones. This included plotting some useful graphs, a lot of useless graphs and to understand the regional effect of conflict I had to plot a world choropleth map.

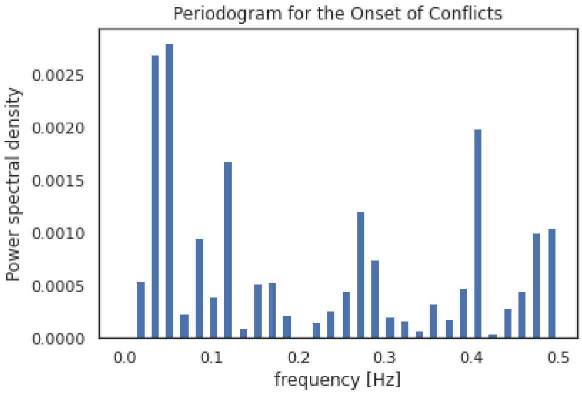


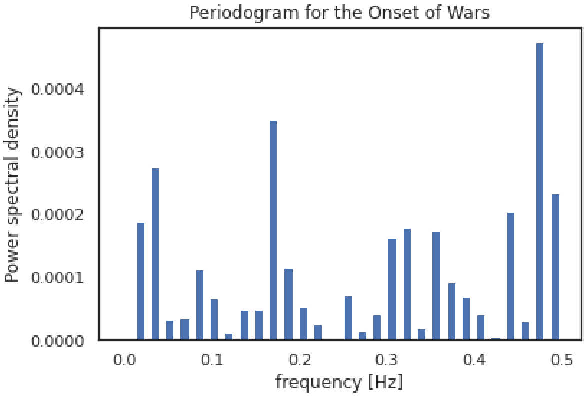
Conflict Types by Year.

Here we visualize what type of conflict is the most common.



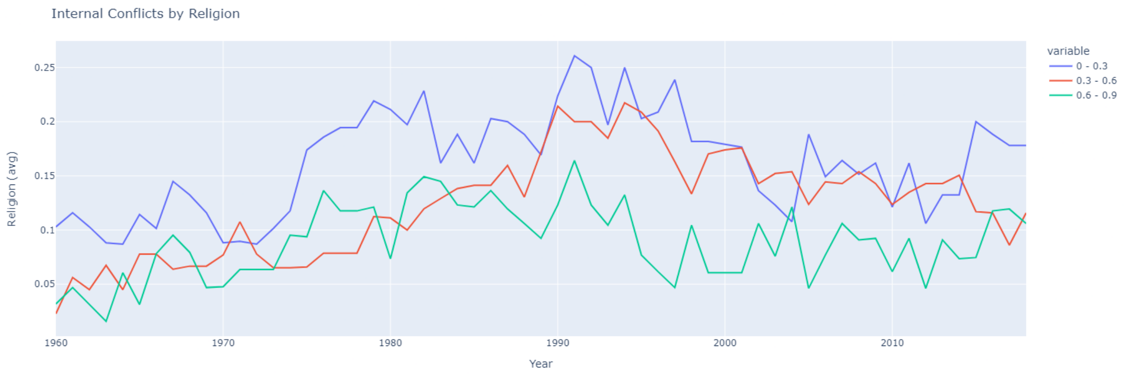
We observe that internal conflict is the most prevalent sort of conflict any country faces.



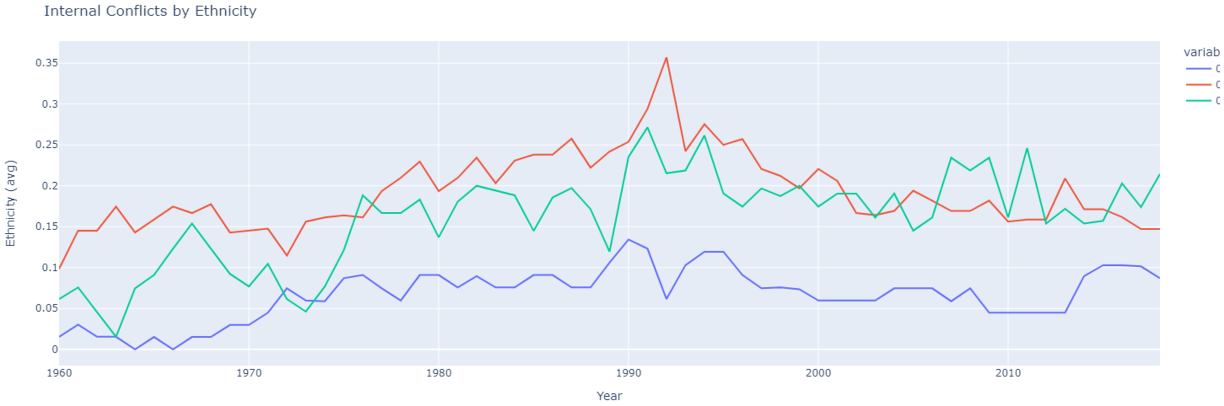


We plot a Periodogram to observe whether wars and conflicts have any periodicity, i.e do they repeat after a set period.

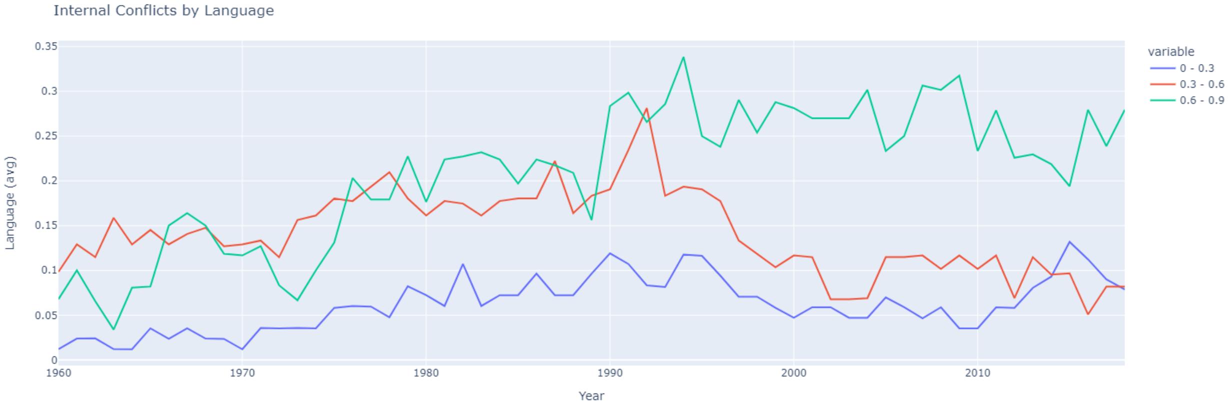
Are Civil Wars caused due to diversity of Language, Religion and Ethnicity?



We see that variation in Religion does not lead to more civil wars contrary to popular opinion.

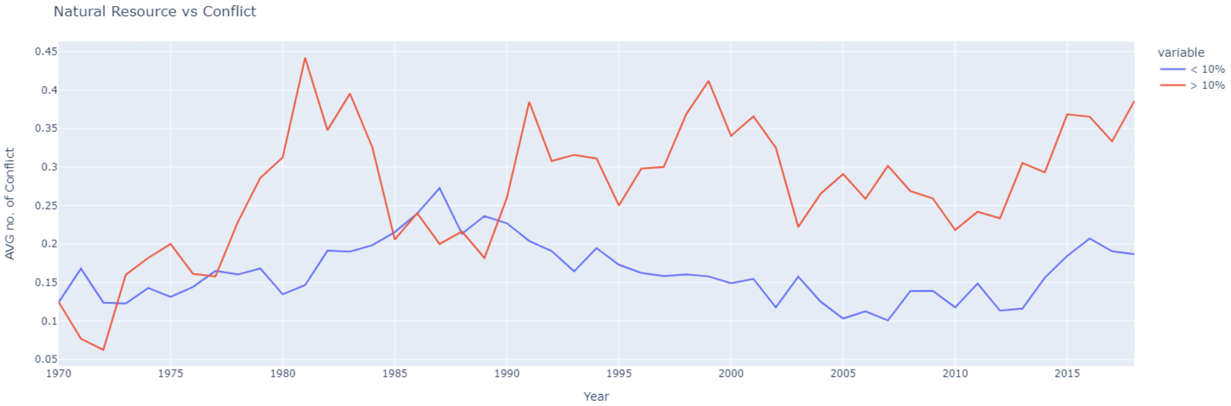


Same as we saw in the case of Religion, a country's variation in Ethnicity does not lead to more civil wars.

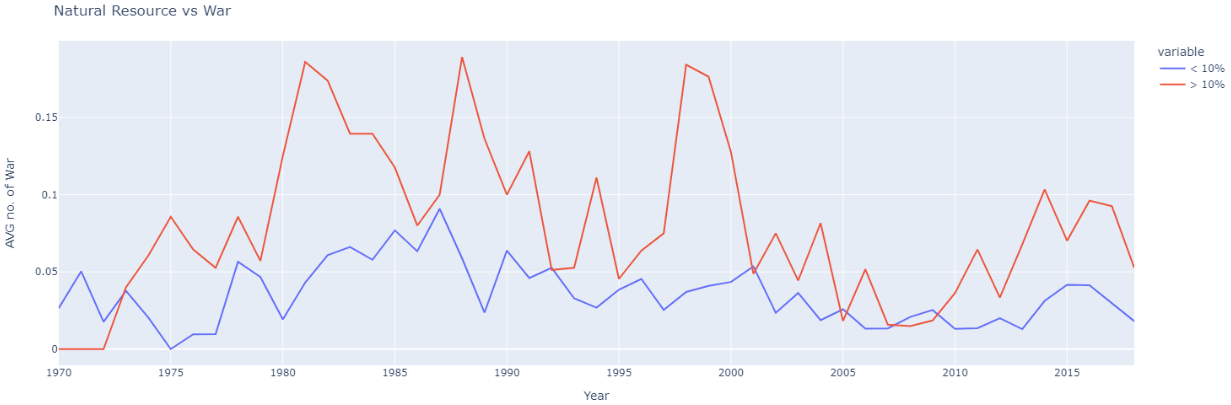


In the case of variation of Language in a country we see that there are more civil wars when there is more language variation.
We come to the conclusion that Language does play a role in internal conflict in a country.

Does a country's dependence on Natural Resources play a role in internal conflict



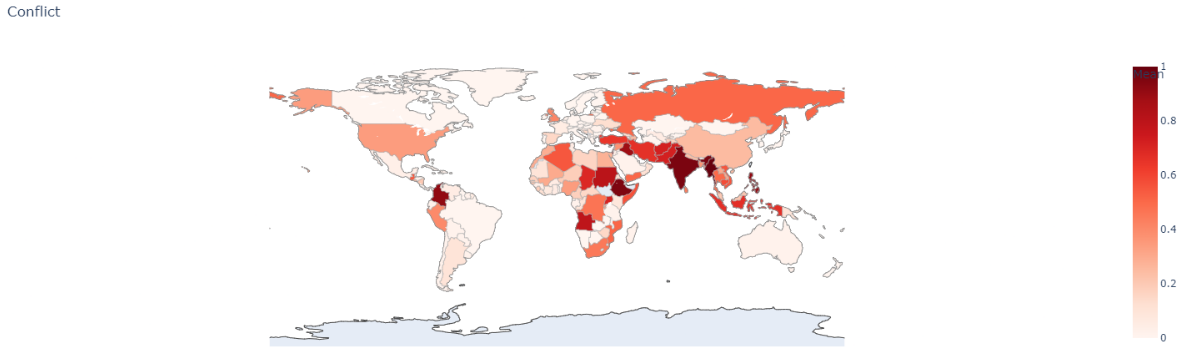
We see that countries with more than 10% dependence on it's natural resources experience more conflict.



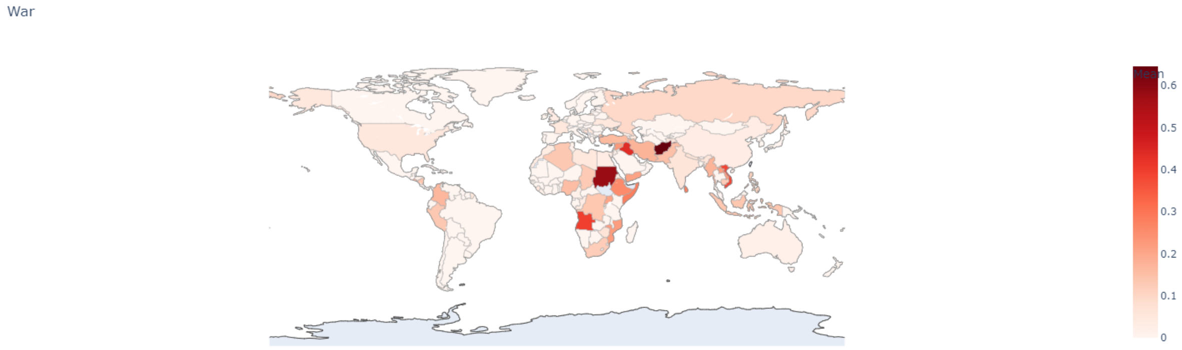
We see that countries with more than 10% dependence on its natural resources experience more wars.

We come to the conclusion that a country's dependence on its natural resources does lead to more wars and conflicts.

Plotting the conflict and wars faced by all the countries since 1960



We see that the Middle East , Africa and Latin America faced a lot of conflicts.



We see that Africa and the Middle East have faced more wars since 1960.

Machine Learning Model

1. Vector Autoregression (VAR)

We will use the VAR model on only data from the USA.

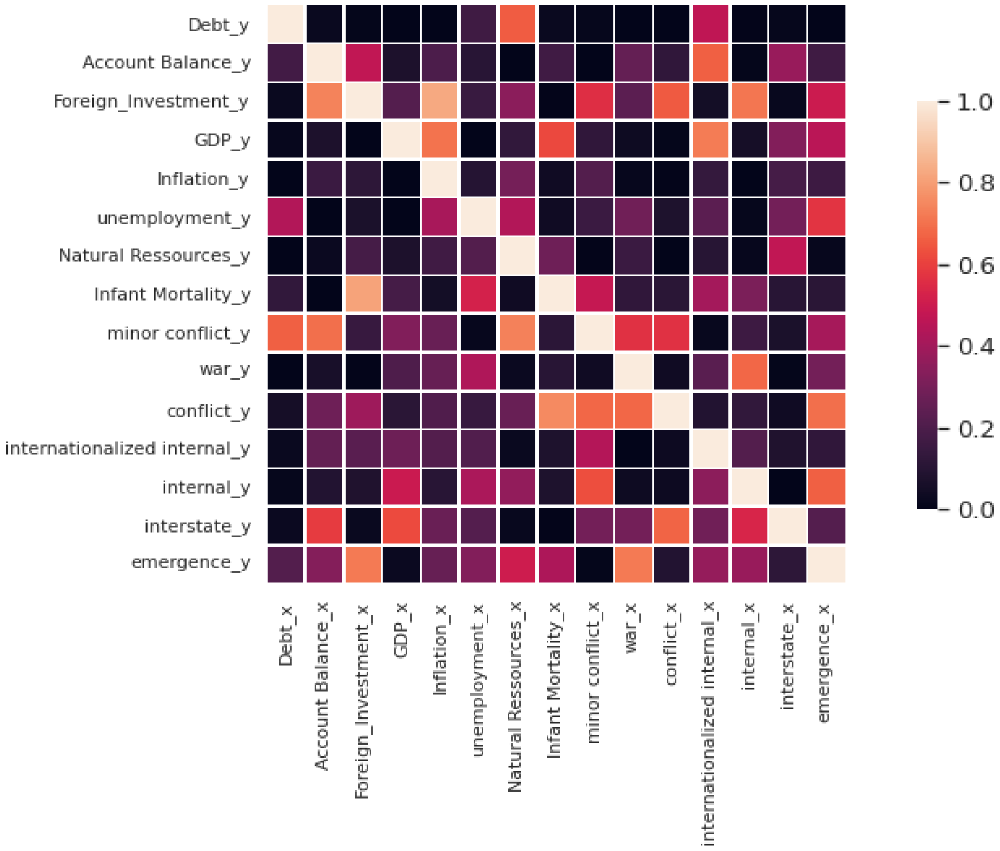
Granger Causality Test

The basis behind Vector AutoRegression is that each of the time series in the system influences each other. That is, you can predict the series with past values of itself along with other series in the system.

Using Granger's Causality Test, it's possible to test this relationship before even building the model.

Granger’s causality tests the null hypothesis that the coefficients of past values in the regression equation is zero.

In simpler terms, the past values of time series (X) do not cause the other series (Y). So, if the p-value obtained from the test is lesser than the significance level of 0.05, then, you can safely reject the null hypothesis.



Some Observations:

- Minor conflicts are strongly affected by the debt, the account balance and the Natural resources.
- The conflicts in general are strongly influenced by the infant mortality rates
- Interstate conflicts are triggered by the GDP and the Account balance.

Vector Autoregression

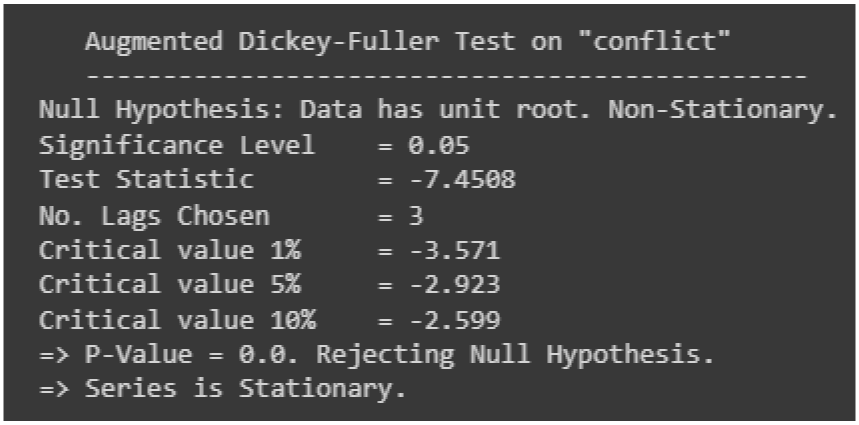
Vector Autoregression (VAR) is a multivariate forecasting algorithm that is used when two or more time series influence each other.
It is considered as an Autoregressive model because each variable (Time Series) is modeled as a function of the past values, that is the predictors are nothing but the lags (time delayed value) of the series.
All variables in a VAR enter the model in the same way:
Each variable has an equation explaining its evolution based on its own lagged values, the lagged values of the other model variables, and an error term. VAR modeling does not require as much knowledge about the forces influencing a variable as do structural models with simultaneous equations: The only prior knowledge required is a list of variables which can be hypothesized to affect each other intertemporally.

Split data into TRAIN and TEST

The VAR model will be fitted on df_train_usa and then used to forecast the next 4 observations. These forecasts will be compared against the actuals present in test data.

To do the comparisons, we will use multiple forecast accuracy metrics, as seen later in this report.
Check for stationarity and make the time series stationary

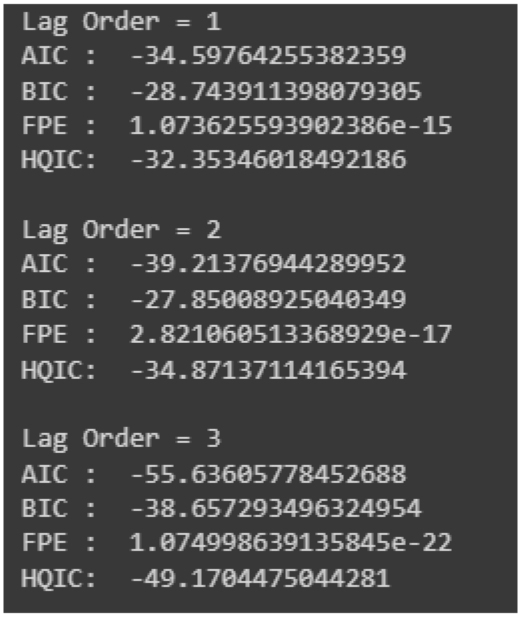
Since the VAR model requires the time series you want to forecast to be stationary, it is customary to check all the time series in the system for stationarity. A stationary time series is one whose characteristics like mean and variance does not change over time.
We use Augmented Dickey-Fuller Test.
After performing this test we differentiate the time series twice.
We perform the Augmented Dickey-Fuller Test again and get satisfactory results.



```
   Augmented Dickey-Fuller Test on "conflict"
   -----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -7.4508
No. Lags Chosen       = 3
Critical value 1%     = -3.571
Critical value 5%     = -2.923
Critical value 10%    = -2.599
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.
```

Selecting Order (P) of model

To select the right order of the VAR model, we iteratively fit increasing orders of the VAR model and pick the order that gives a model with least AIC.



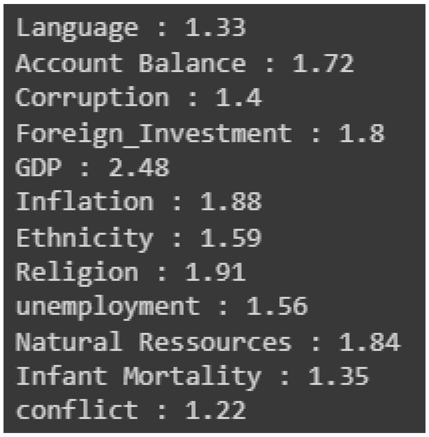
```
Lag Order = 1
AIC :  -34.59764255382359
BIC :  -28.743911398079305
FPE :  1.073625593902386e-15
HQIC:  -32.35346018492186

Lag Order = 2
AIC :  -39.21376944289952
BIC :  -27.85008925040349
FPE :  2.821060513368929e-17
HQIC:  -34.87137114165394

Lag Order = 3
AIC :  -55.63605778452688
BIC :  -38.657293496324954
FPE :  1.074998639135845e-22
HQIC:  -49.1704475044281
```

We continue with Lag Order 3 and train the model.

Check for Serial Correlation of Residuals (Errors) using Durbin Watson Statistic

Serial correlation of residuals is used to check if there is any leftover pattern in the residuals (errors). If there is any correlation left in the residuals, then, there is some pattern in the time series that is still left to be explained by the model. In that case, the typical course of action is to either increase the order of the model or induce more predictors into the system or look for a different algorithm to model the time series.

The value of this statistic can vary between 0 and 4. The closer it is to the value 2, then there is no significant serial correlation. The closer to 0, there is a positive serial correlation, and the closer it is to 4 implies negative serial correlation.

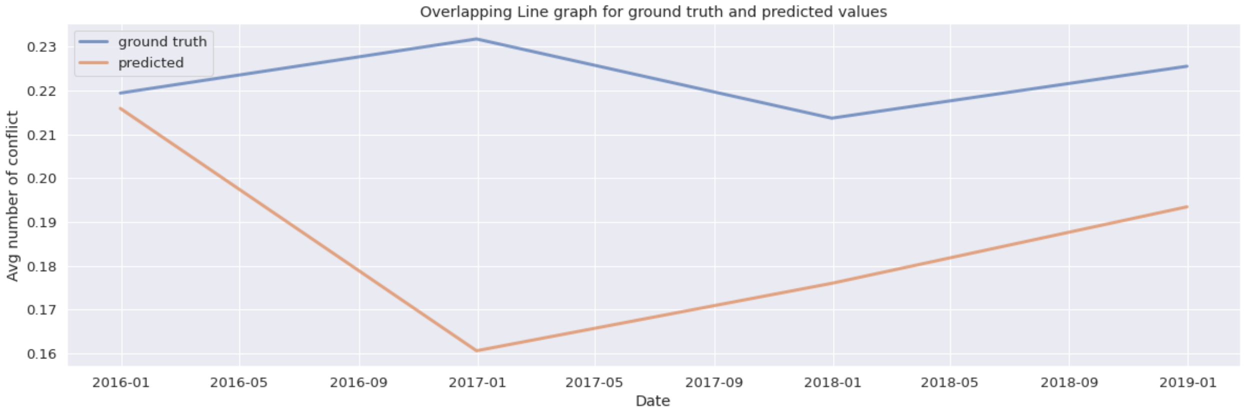


```
Language : 1.33
Account Balance : 1.72
Corruption : 1.4
Foreign_Investment : 1.8
GDP : 2.48
Inflation : 1.88
Ethnicity : 1.59
Religion : 1.91
unemployment : 1.56
Natural Ressources : 1.84
Infant Mortality : 1.35
conflict : 1.22
```

Forecasting

In order to forecast, the VAR model expects up to the lag order number of observations from the past data.

This is because the terms in the VAR model are essentially the lags of the various time series in the dataset, so you need to provide it with as many of the previous values as indicated by the lag order used by the model.
We reverse the differencing that we did earlier to get real values and plot it.



Metrics

Metric	Value
MSE	0.0018
MAE	0.0361
RMSE	0.0434
Med AE	0.0349

ARIMA

An ARIMA model is a class of statistical models for analyzing and forecasting time series data.

It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration.

This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

- AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.
- I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
- MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The parameters of the ARIMA model are defined as follows:

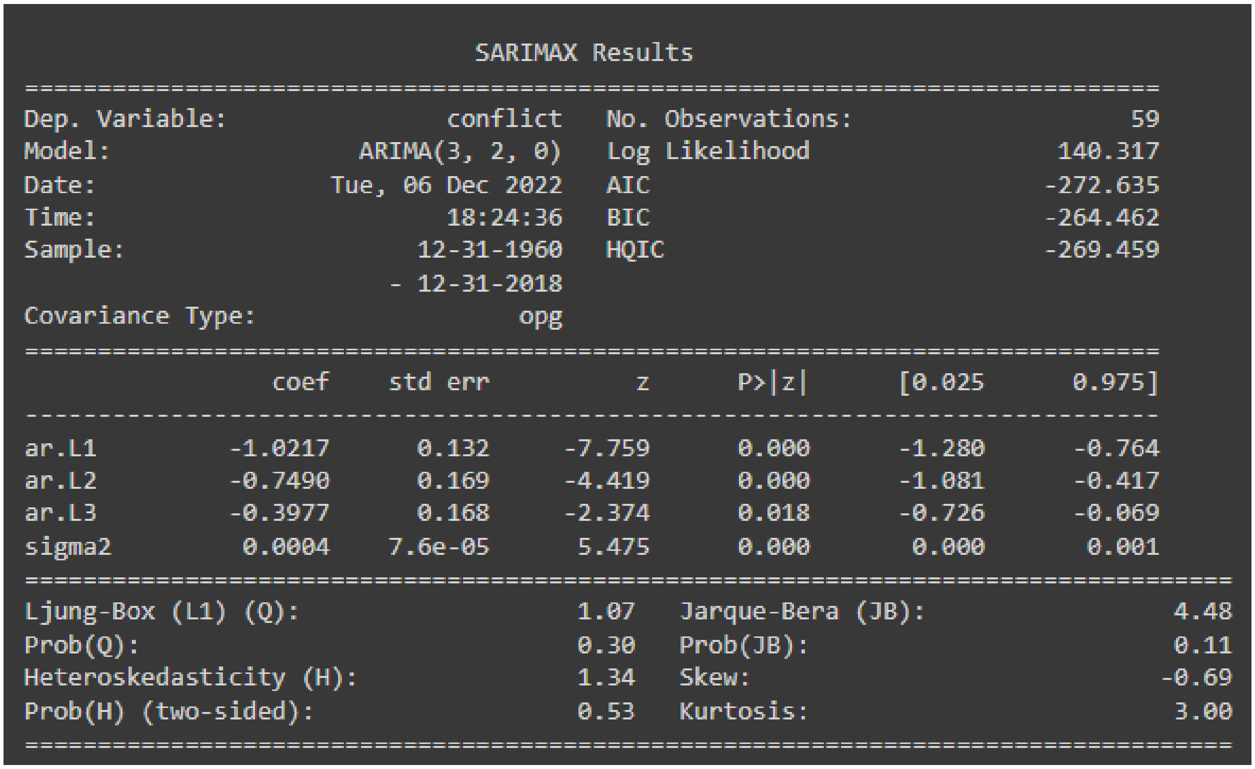
- p: The number of lag observations included in the model, also called the lag order.
- d: The number of times that the raw observations are differenced, also called the degree of differencing.
- q: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model.

Experimentation

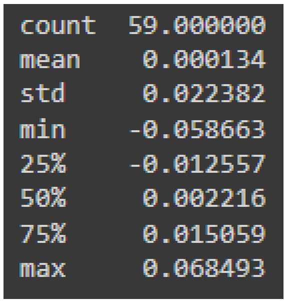
We have some clue of the stationarity of the time series from the previous tests we ran, so I do some experimentation while fitting the the model. I find the best parameters to be
P = 3
D = 2
Q = 0

Summary of fitting the models gives us:



```
                               SARIMAX Results
==============================================================================
Dep. Variable:               conflict   No. Observations:                   59
Model:                 ARIMA(3, 2, 0)   Log Likelihood                 140.317
Date:                Tue, 06 Dec 2022   AIC                           -272.635
Time:                        18:24:36   BIC                           -264.462
Sample:                    12-31-1960   HQIC                          -269.459
                         - 12-31-2018
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.0217      0.132     -7.759      0.000      -1.280      -0.764
ar.L2         -0.7490      0.169     -4.419      0.000      -1.081      -0.417
ar.L3         -0.3977      0.168     -2.374      0.018      -0.726      -0.069
sigma2         0.0004    7.6e-05      5.475      0.000       0.000       0.001
===================================================================================
Ljung-Box (L1) (Q):                   1.07   Jarque-Bera (JB):                 4.48
Prob(Q):                              0.30   Prob(JB):                         0.11
Heteroskedasticity (H):               1.34   Skew:                            -0.69
Prob(H) (two-sided):                  0.53   Kurtosis:                         3.00
===================================================================================
```

The distribution of the residual errors is displayed. The results show that indeed there is a very small bias in the prediction (a non-zero mean in the residuals).



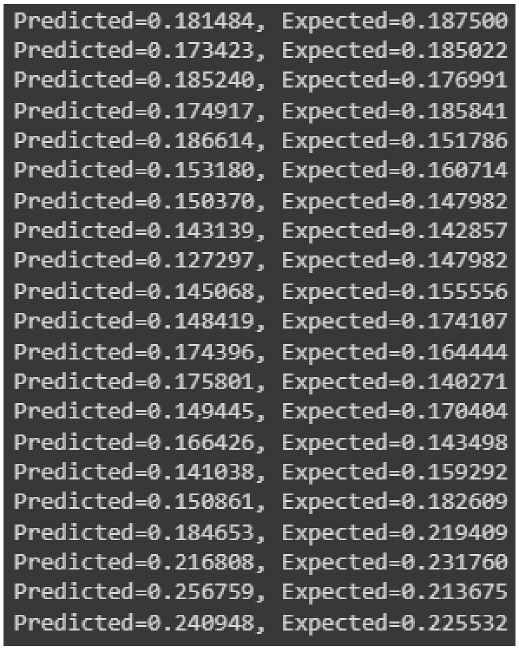
```
count   59.000000
mean     0.000134
std      0.022382
min     -0.058663
25%     -0.012557
50%      0.002216
75%      0.015059
max      0.068493
```

Forecasting

We split the training dataset into train and test sets, use the train set to fit the model, and generate a prediction for each element on the test set.

A rolling forecast is required given the dependence on observations in prior time steps for differencing and the AR model. A crude way to perform this rolling forecast is to re-create the ARIMA model after each new observation is received.

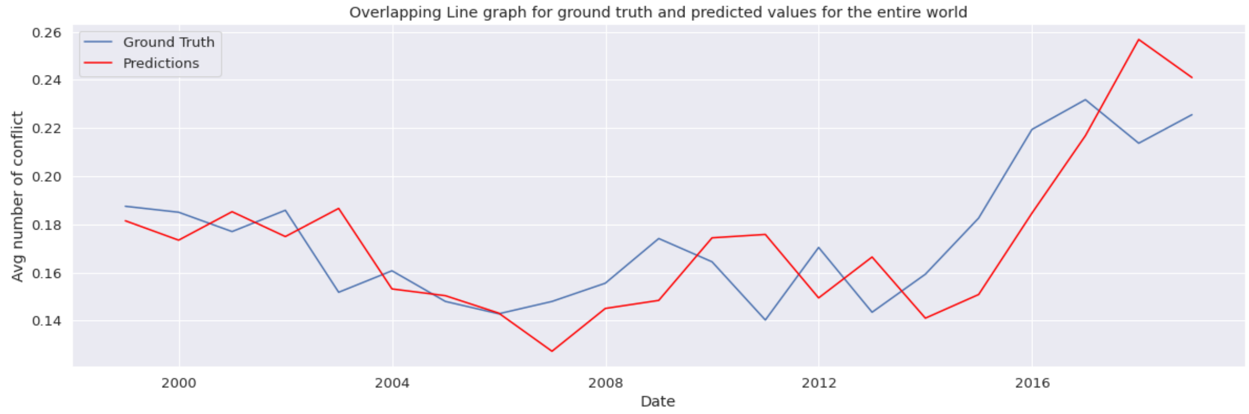
We manually keep track of all observations in a list called history that is seeded with the training data and to which new observations are appended each iteration.



```
Predicted=0.181484, Expected=0.187500
Predicted=0.173423, Expected=0.185022
Predicted=0.185240, Expected=0.176991
Predicted=0.174917, Expected=0.185841
Predicted=0.186614, Expected=0.151786
Predicted=0.153180, Expected=0.160714
Predicted=0.150370, Expected=0.147982
Predicted=0.143139, Expected=0.142857
Predicted=0.127297, Expected=0.147982
Predicted=0.145068, Expected=0.155556
Predicted=0.148419, Expected=0.174107
Predicted=0.174396, Expected=0.164444
Predicted=0.175801, Expected=0.140271
Predicted=0.149445, Expected=0.170404
Predicted=0.166426, Expected=0.143498
Predicted=0.141038, Expected=0.159292
Predicted=0.150861, Expected=0.182609
Predicted=0.184653, Expected=0.219409
Predicted=0.216808, Expected=0.231760
Predicted=0.256759, Expected=0.213675
Predicted=0.240948, Expected=0.225532
```

Results

A line plot is created showing the expected values (blue) compared to the rolling forecast predictions (red). We can see the values show some trend and are in the correct scale.



Metrics	Value
MSE	0.0004
MAE	0.0183
RMSE	0.0218
Med AE	0.0154

Conclusions

I found this project to be very helpful in truly understanding various aspects of data analytics and I had a lot of fun doing this. There are a few extra things I wanted to try out but I was limited by time and my own skills. I believe that I have created decent models which give very less error and I can only hope that you find this to be a decent project too.

References

1. Davies, S., Pettersson, T., & Öberg, M. (2022). Organized violence 1989–2021 and drone warfare. Journal of Peace Research, 59(4), 593-610.
2. Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M., & Strand, H. (2002). Armed conflict 1946-2001: A new dataset. Journal of peace research, 39(5), 615-637.
3. Chadefaux, T. (2017). Conflict forecasting and its limits. Data Science, 1(1-2), 7-17.
4. Hegre, H., Karlsen, J., Nygård, H. M., Strand, H., & Urdal, H. (2013). Predicting armed conflict, 2010–2050. International Studies Quarterly, 57(2), 250-270.
5. https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/
6. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
7. https://plotly.com/python/choropleth-maps/
8. https://colab.research.google.com/drive/1vn7ky0nsfNSs8rJbvWkVob0tT3xIhM4l?usp=sharing#scrollTo=1TFyKQe3ZCt8
9. https://stackoverflow.com/