ITERATION 3

OSAS

Shuo Feng

UPI: sfen779

ID:836311644

Table of Contents

[1. Business/Situation Understanding 3](#_Toc143467498)

[1.1. Business/Situation Objectives 3](#_Toc143467499)

[1.2. Assessment of the Situation 3](#_Toc143467500)

[1.2.1. Resource Inventory 3](#_Toc143467501)

[1.2.2. Requirements, Assumptions, and Constrains 3](#_Toc143467502)

[1.2.3. Risks and Contingencies 4](#_Toc143467503)

[1.2.4. Cost/Benefit Analysis 5](#_Toc143467504)

[1.3. Data Mining Objectives 5](#_Toc143467505)

[1.3.1. Data Mining Success Criteria 6](#_Toc143467506)

[1.4. Project Plan 6](#_Toc143467507)

[2. Data Understanding 7](#_Toc143467508)

[2.1. Collecting Initial Data 7](#_Toc143467509)

[2.2. Describing Data 7](#_Toc143467510)

[2.3. Data Exploration 11](#_Toc143467511)

[2.3.1. Exploration of most important predictors(features) 12](#_Toc143467512)

[2.4. Verifying Data Quality 13](#_Toc143467513)

[2.4.1. Missing Values and Extreme Values 13](#_Toc143467514)

[2.4.2. Extreme Values Explained and Feature Value (coding) Inconsistencies 14](#_Toc143467515)

[2.4.3. Measurement errors 14](#_Toc143467516)

[3. Data Preparation 15](#_Toc143467517)

[3.1 Select the data 15](#_Toc143467518)

[3.2 Clean the data 17](#_Toc143467519)

[3.3 Construct the data 21](#_Toc143467520)

[3.4 Integrate various data sources 22](#_Toc143467521)

[3.5 Format the data as required 22](#_Toc143467522)

[4. Data Transformation 24](#_Toc143467523)

[4.1 Reduce the data 24](#_Toc143467524)

[4.2 Project the data 29](#_Toc143467525)

[5. Data-Mining Method(s) Selection 31](#_Toc143467526)

[5.1 Match and discuss the objectives of data mining (1.1) to data mining methods 31](#_Toc143467527)

[5.2 Select the appropriate data-mining method(s) based on discussion 31](#_Toc143467528)

[6. Data-Mining Algorithm(s) Selection 33](#_Toc143467529)

[6.1 Conduct exploratory analysis and discuss 33](#_Toc143467530)

[6.1.1. First Data-Mining Objective: The Linear Algorithm 33](#_Toc143467531)

[6.1.2. First Data-Mining Objective: The Regression Algorithm 34](#_Toc143467532)

[6.1.3. Second Data-Mining Objective: The Classification Algorithm 34](#_Toc143467533)

[6.1.4. Second Data-Mining Objective: The Logistic Regression Algorithm 35](#_Toc143467534)

[6.1.5. Second Data-Mining Objective: SVM 36](#_Toc143467535)

[6.2 Select data-mining algorithms based on discussion 36](#_Toc143467536)

[6.3 Build/Select appropriate model(s) and choose relevant parameter(s) 36](#_Toc143467537)

[7. Data Mining 39](#_Toc143467538)

[7.1 Create and justify test designs 39](#_Toc143467539)

[7.2 Conduct data mining – classify, regress, cluster, etc. (models must execute) 39](#_Toc143467540)

[7.3 Search for patterns 40](#_Toc143467541)

[8. Interpretation 42](#_Toc143467542)

[8.1 Study and discuss the mined patterns 42](#_Toc143467543)

[8.2 Visualize the data, results, models, and patterns 43](#_Toc143467544)

[8.3 Interpret the results, models, and patterns 43](#_Toc143467545)

[8.4 Assess and evaluate results, models, and patterns 44](#_Toc143467546)

[Disclaimer 48](#_Toc143467547)

# Business/Situation Understanding

In this iteration, the aim is that supporting Sustainable Development Goal 8 (SDG 8) - Decent Work and Economic Growth. As the world grapples with the economic impact of recent challenges, like new waves of COVID-19, rising inflation, supply-chain disruptions, and the Ukraine crisis, it is believed that harnessing the power of data-driven insights will be instrumental in driving inclusive and sustainable growth on a global scale.

## 1.1. Business/Situation Objectives

The primary objective of this iteration is to leverage data mining and big data analytics to support governments, businesses, and international organizations in formulating effective strategies for global economic recovery and finding ways to avoid risks. By analyzing comprehensive datasets and identifying key economic indicators, we aim to provide actionable insights to guide decision-making and policy formulation.

## Assessment of the Situation

### Resource Inventory

All hardware is provided by Shuo Feng, including a Lenovo laptop (CPU: AMD 4800H, 16GB+2TB storage, GPU GTX 1650) and a MacBook Air (M1 version).

All software is also purchased and provided by Shuo Feng, like operation systems and office365. The python code developing tools, such as visual studio code and Spyder, are free, open-source IDEs.

The data used will be accessed from the world bank.

Human resource: A student from University of Auckland, doing his master’s degree.

### Requirements, Assumptions, and Constrains

This iteration requires about 1 month time to interpret and a large amount of reliable data.

In terms of data quality assumptions, it is assumed as followings:

1. All necessary data entries have been recorded in full.
2. All the information in the data is accurate and error-free.
3. All data is consistent, with no contradictions or conflicts.
4. The data source is reliable and trustworthy data is up-to-date and reflects the current situation accurately

But the data may not fit those assumptions, and oppositely, some constrains could appear according to the quality of data that it can be possibly collected. For example, there might be missing values, or some records might not have been captured in their entirety, there might be outdated data, especially in dynamically changing environments.

### Risks and Contingencies

Some possible risks and its corresponding contingency plan identified:

|  |  |
| --- | --- |
| **Potential Risks** | **Contingency Plan** |
| IT system failure | 1. Regularly back up critical data and systems. 2. Implement redundant systems or cloud-based solutions for critical operations. 3. Develop an IT disaster recovery plan. |
| Reputation damage | 1. Address the root cause of the issue and communicate corrective actions taken. 2. Monitor online and offline sentiment and respond appropriately. |
| Model building failure | 1. Re-evaluate the data being used. Check for missing values, outliers, or any inconsistencies. Consider collecting more data or improving data quality through cleaning and preprocessing. 2. If the chosen algorithm isn't producing satisfactory results, consider testing alternative algorithms or modeling techniques. Different algorithms have different strengths and might be better suited for specific types of data or problems. 3. Revisit the features being used in the model. Consider adding new features, transforming existing ones, or removing irrelevant or redundant features. |
| Deadline exceeded | 1. Detailed planning at the start of the project ensures that there's ample time to address potential delays. 2. Periodically check the progress of the project to ensure it aligns with the scheduled timeline. |

### Cost/Benefit Analysis

**Costs:**

1. **Infrastructure Costs:**

Setting up and maintaining big data infrastructure, including servers, storage, and networking.

Licensing costs for specialized software and platforms.

1. **Talent Acquisition and Training:**

Ongoing training and professional development to keep up with evolving technologies and methodologies.

1. **Data Acquisition and Integration:**

Costs associated with acquiring relevant datasets, possibly from third-party vendors.

Integrating disparate data sources, ensuring compatibility and consistency.

1. **Time Investment:**

The time required to see meaningful results from data initiatives can be significant, especially for complex projects.

**Benefits:**

1. **Informed Decision Making:**

Data-driven insights can guide governments and organizations in making decisions that are more likely to yield positive outcomes.

1. **Tailored Strategies:**

Data analytics allows for the customization of strategies to specific regions, sectors, or demographics, ensuring more targeted and effective interventions.

1. **Risk Mitigation:**

Predictive analytics can identify potential economic risks, allowing for proactive measures to avoid or minimize negative impacts.

## 1.3. Data Mining Objectives

Find the underlying relations among different variables related to global economy, like the amount of export and import, unemployment rate and ideologies of all countries and regions over the world. And find out by how we can improve the performance of several indices of global economic dynamism and sustainable growth.

On the other hand, it is also needed to use those related variables to detect an economy risk. By doing this, a potential risk may be avoided and protect the recovery of global economy.

### 1.3.1. Data Mining Success Criteria

In this study, several data mining algorithms and models will be used to find what contributes most to the growth of global economy and what is the main cause of global economy growing risks.

In terms of performance, the dataset should be appropriately divided into training and testing sets to validate the model’s performance, and the model should process data and provide insights in a timely manner, especially if real-time analysis is required.

In terms of the assessing the accuracy of classification model, the model should have a high precision and recall.

The error rate of the regression models in this study, like RMSE (Root Mean Square Error) should be within acceptable limits defined at the outset.

At the same time, the R-squared value should be sufficiently high, indicating that the model explains a significant portion of the variance in the dependent variable.

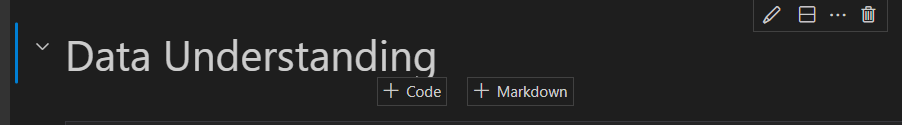
## 1.4. Project Plan

|  |  |  |
| --- | --- | --- |
| **Phase** | **Time (in %)** | **Risks** |
| Business understanding | 10 | IT system failure  (software installing failure)  Reputation damage  (misunderstanding) |
| Data understanding | 10 | IT system failure  (plot drawing program takes too much RAM cause crash and unsaved files lost) |
| Data preparation | 15 | IT system failure  (code version control system failure) |
| Data transformation | 5 | IT system failure  (code version control system failure) |
| Data-mining method selection | 10 | IT system failure  (code version control system failure)  Deadline exceeded  (time management failure/ should choose between quality and finishing)  Model building failure  (can not find a good model due to lack of related knownledge) |
| Data-mining algorithm selection | 15 | Deadline exceeded  Model building failure  (same as above) |
| Data-mining | 15 | Deadline exceeded  (same as above) |
| Interpretation | 20 | Deadline exceeded  (same as above) |

This iteration starts from 21th August 2023, and the deadline is 22th September, 2023.

2 days for BU, 2 days for DU, 3 days for DP, 1 day for DT, 2 days for DMM, 6 days for DM, 4 days for Interpretation. (All days here are work days) 22 work days in total, that is, 1 natural month. So, ideally, this iteration could be finished in time.

# Data Understanding



## 2.1. Collecting Initial Data

The data would be collected from:

<https://datacatalog.worldbank.org/search/dataset/0041188>

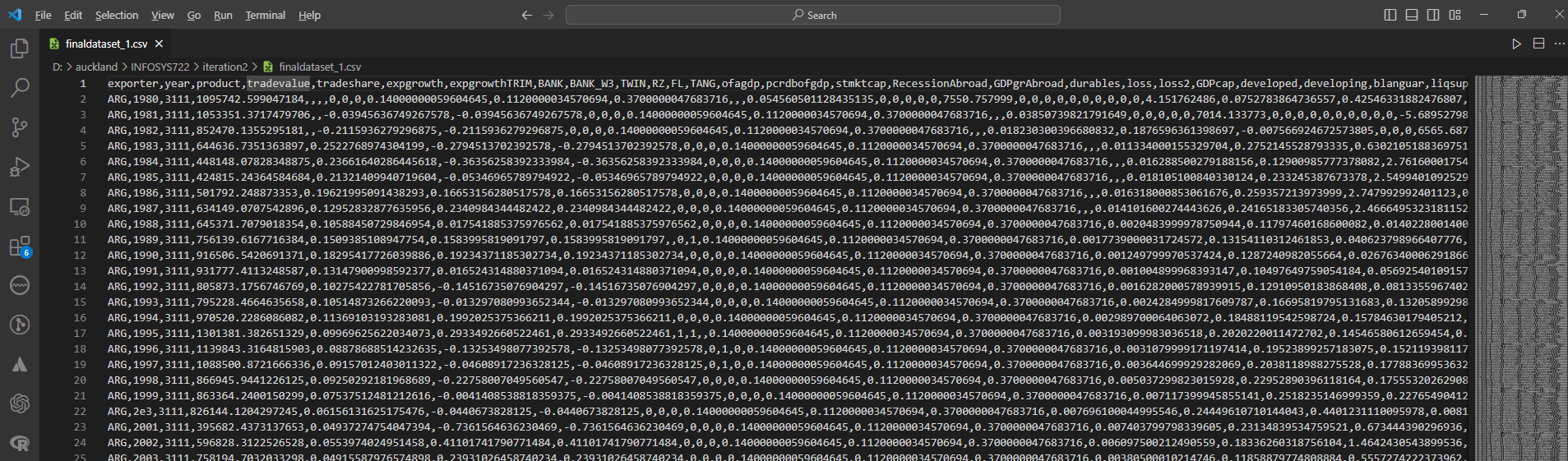
The dataset is not classified and all users can download it whether inside or outside The World Bank.

The World Bank gathered the data from historical reports and documents, so there will be a structural data missing—something bad happened in those countries made them can not report the economics data.

## 2.2. Describing Data

The data is a structured .csv (comma-separated values) file with a 17.64MB size. It has 39588 rows and 44 columns. It is downloaded from the website given in section 2.1.

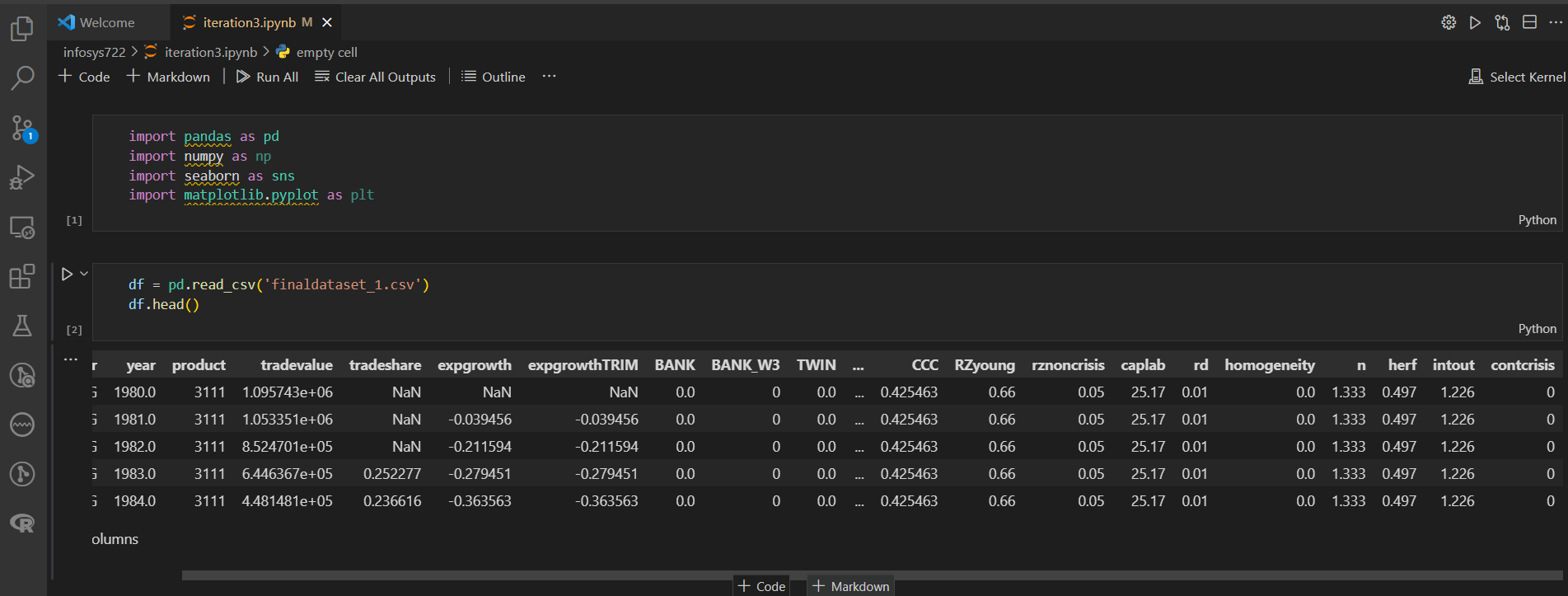
Open it in Visual Studio Code to see its original text looks:



Most of the variables in this data have a numerical type, some flag type and some categorical type variables are also included.

It is difficult to guess what the data in this column is about through some headers, but there are exceptions, such as "year", "country" and "trade value", and the last column "contcrisis" uses 0 and 1 to indicate whether it occurs Banking crisis.

A python program is written to visualize the data and describe it.



After reading, the shape of this dataset can be known:



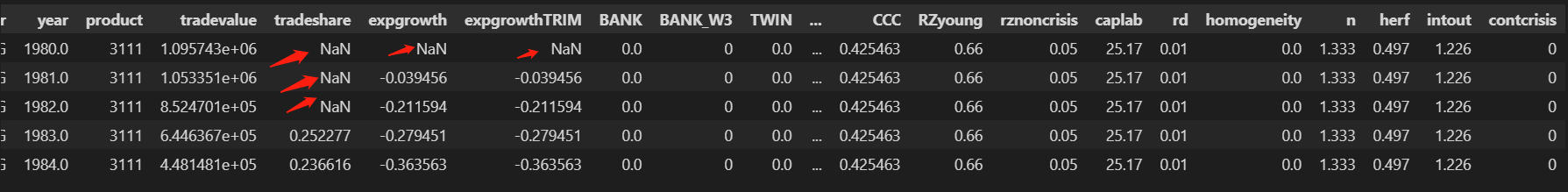
It is a data frame with 39588 rows and 44 columns.

In this data mining project, it is not necessary to fully understand the meaning of all headers, it should be done without any knowledge in economics or any related fields. Instead, only using data mining technics is a necessary.

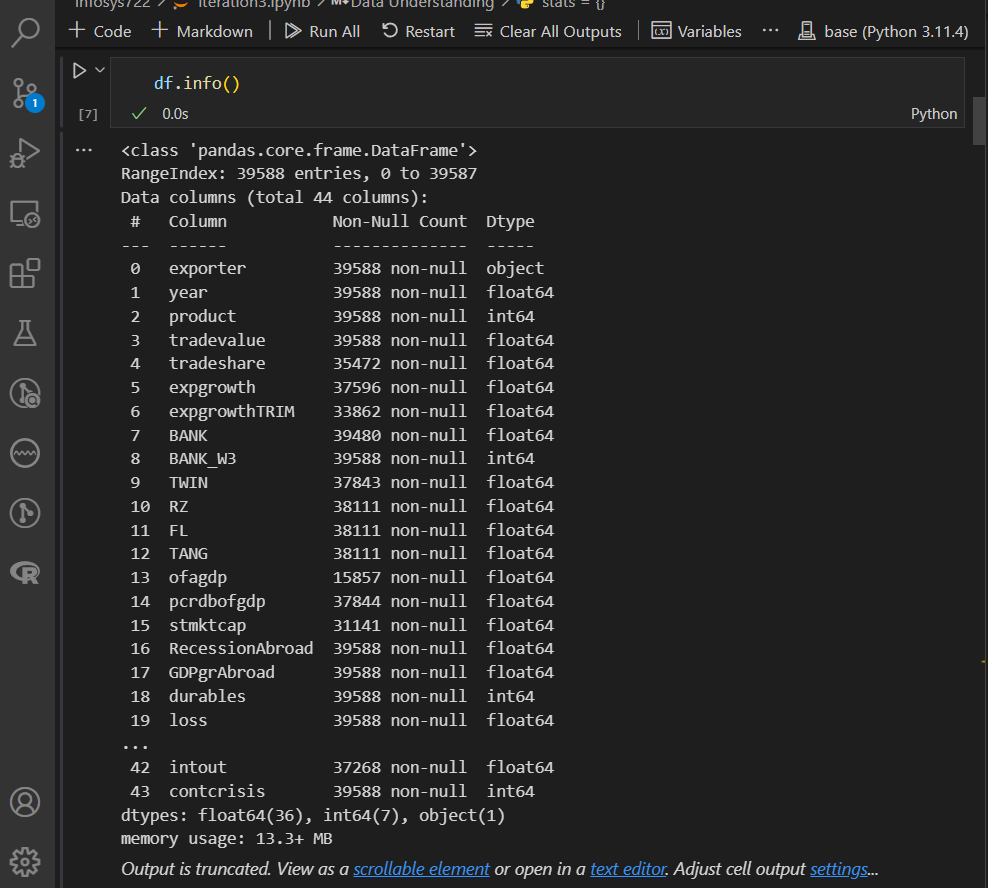
The first step is finding the variables most related to “contcrisis” through analysis, because if there is an “1” in that column, there was a banking crisis (the only meaning of columns that should be known), and then understand the relationships of these related variables.

Several data understanding code chunks are created in VS Code.

It could be seen that there are many missing values from the table created by the head() function.

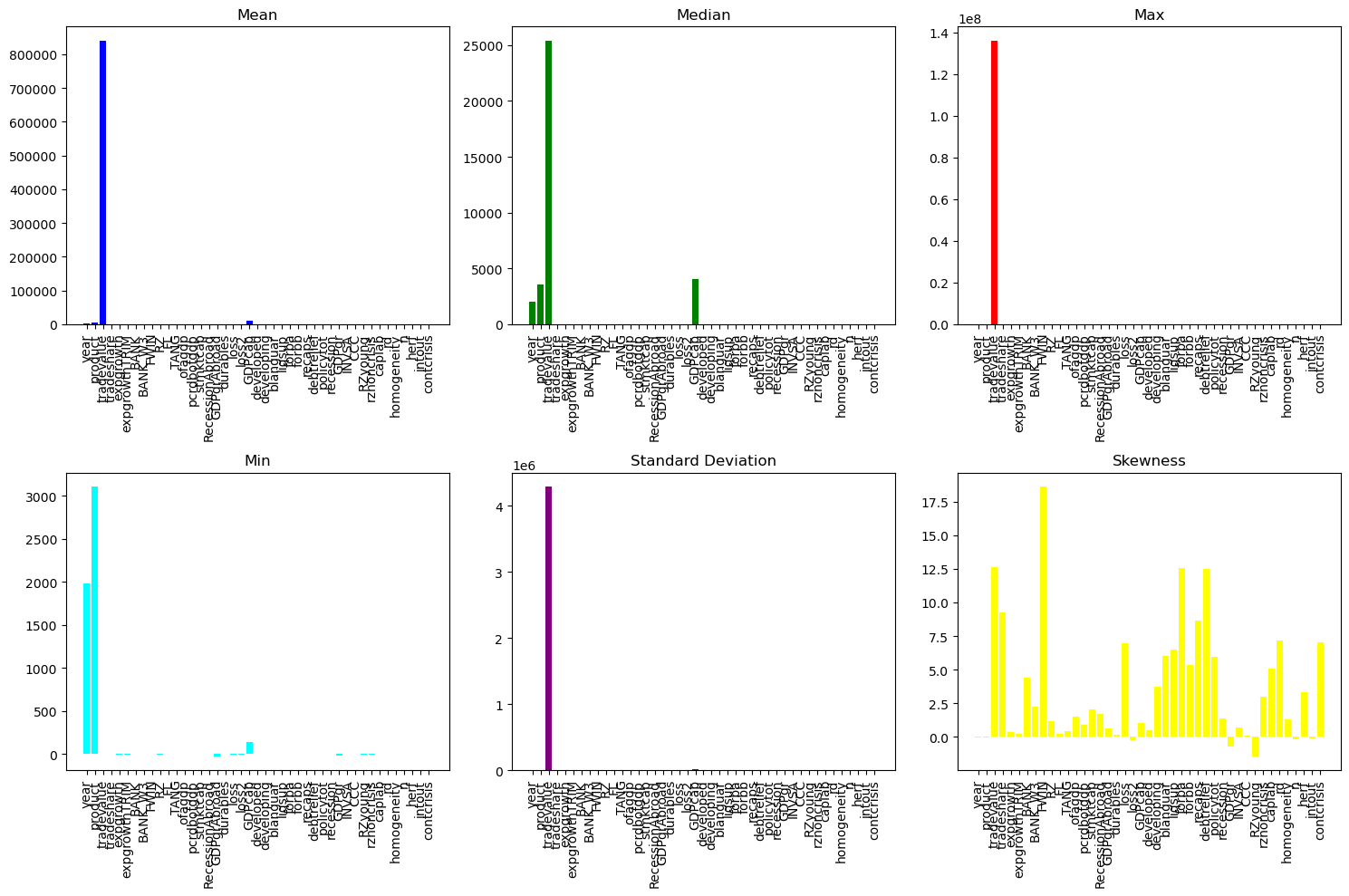


Before exploration, the data types of the columns should be considered.

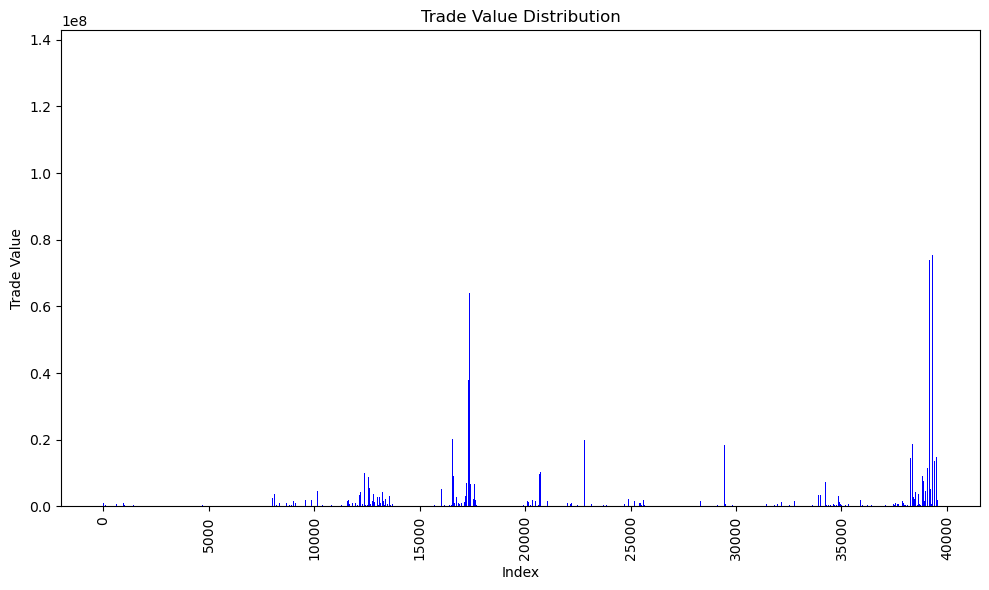


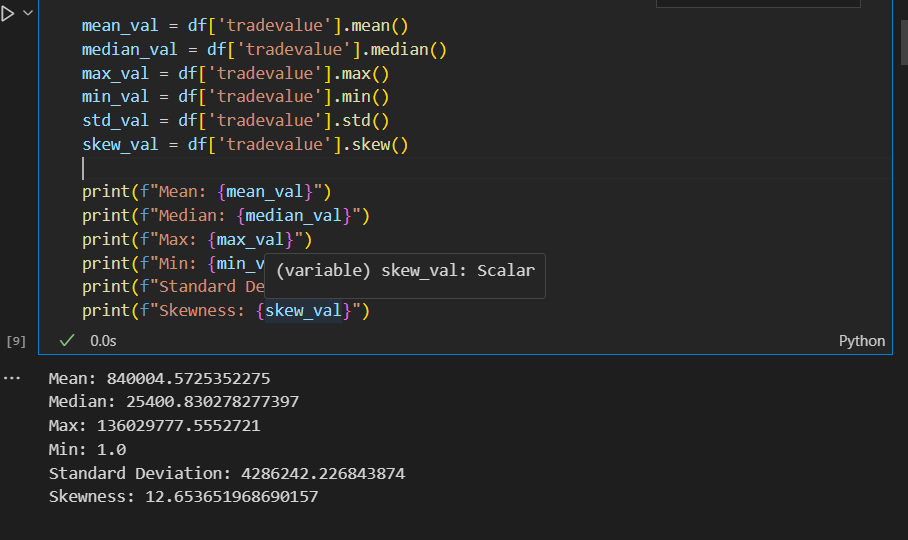
For the columns do not have Dtype in “int64” or “float64”, a calculation is not needed.

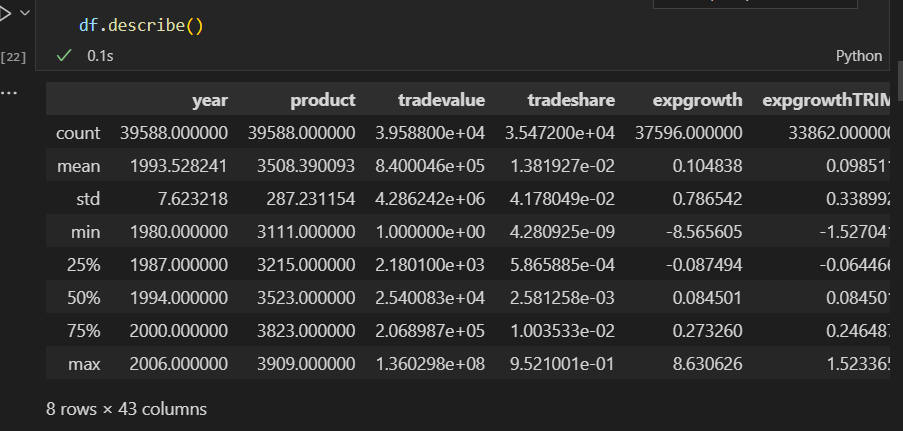
After some calculation, the maximum value, minimum value, mean value, standard deviation, skewness is clearly shown.



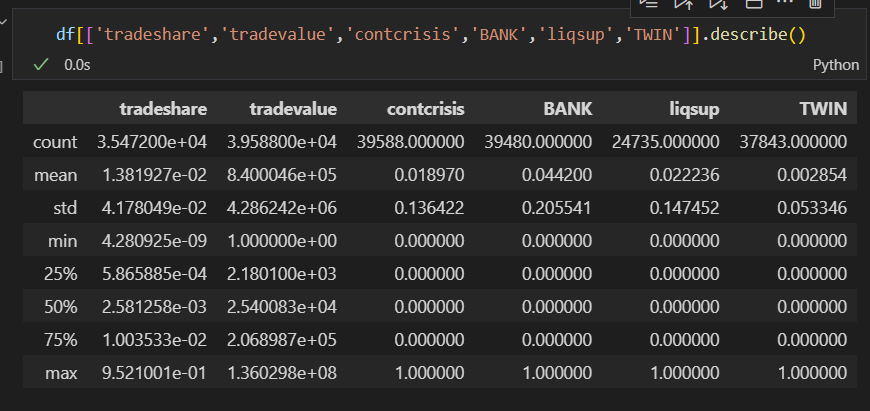
From this, it is obvious “tradevalue” is obviously far larger than the others, so a better way to show all those statistical values is to show them by column, not by rows. At the meantime, this implies “tradevalue” may play a very important role in detecting the banking crisis.

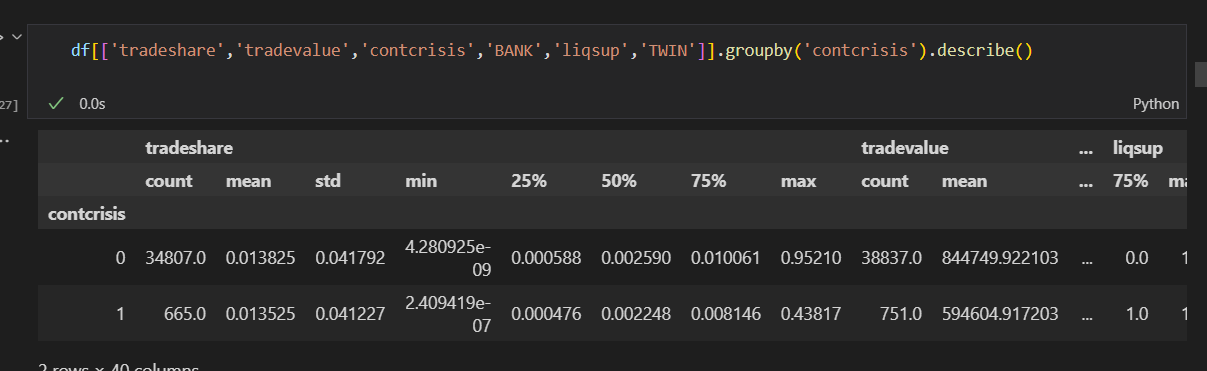




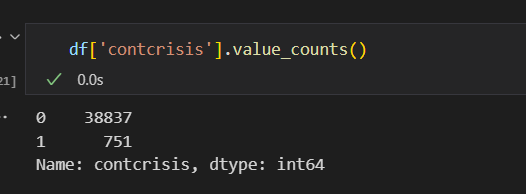


Here are some specific fields in this data (all above could also be done with “describe” function):



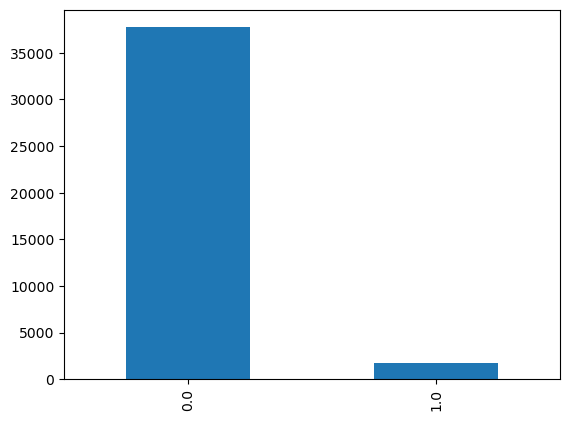


The good news from value counting: the banking crisis do not happen a lot.

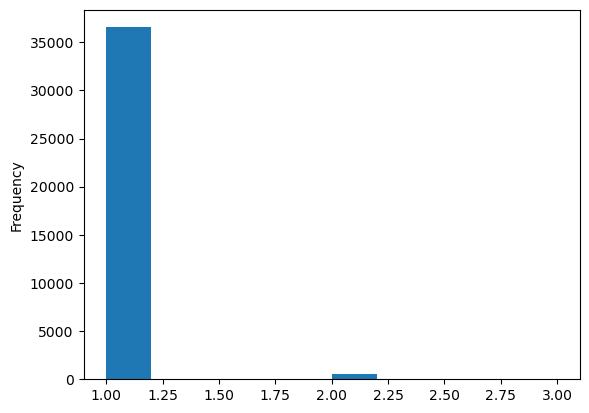


Some Graphs is created to describe those fields.

This graph below shows that most of the records -- over 35000 has the “BANK” variable with the value 0.

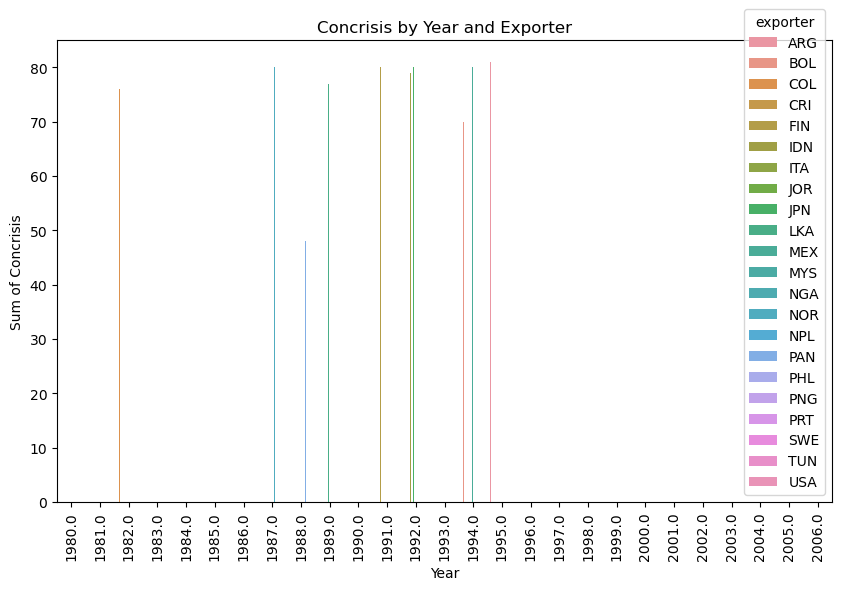


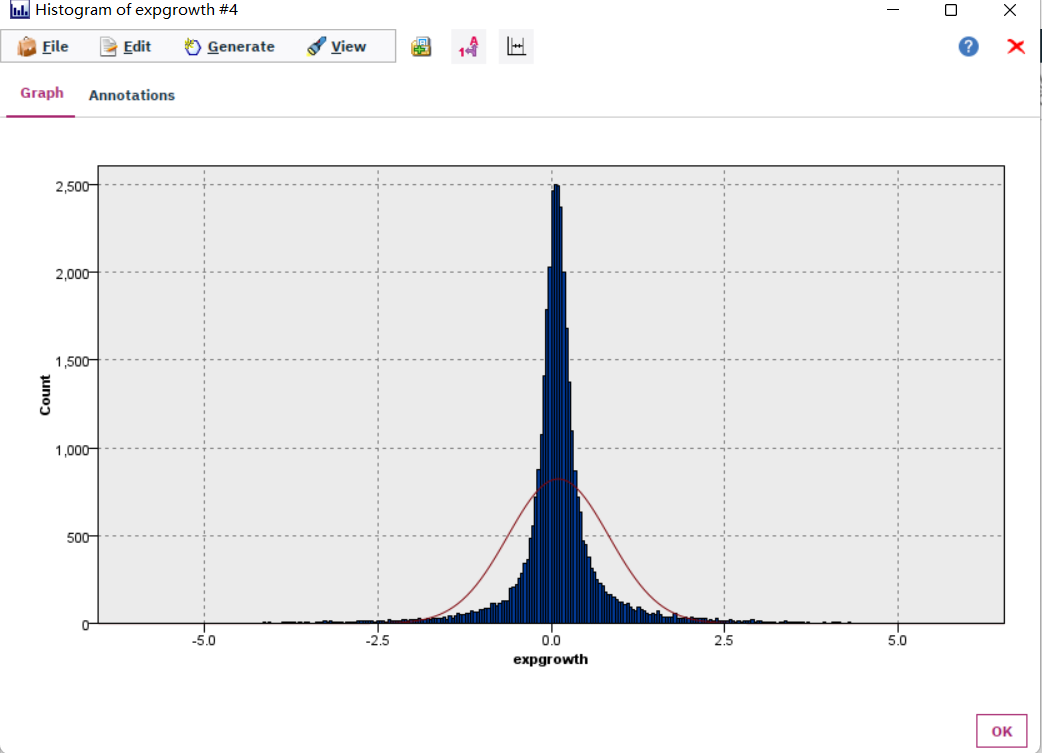
It could be seen that there are some outliers in export growth rate. Around 2, but most of the record have a export growth rate less than 1.25.



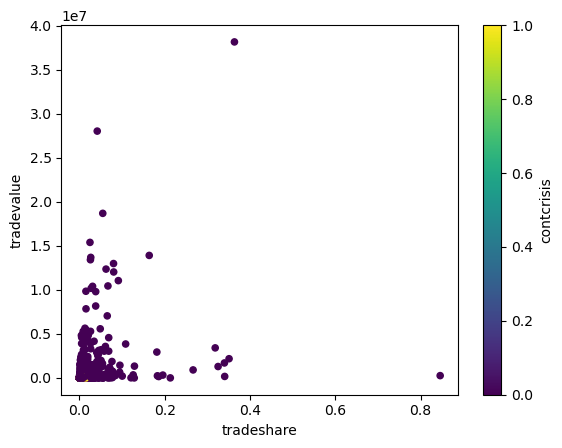
## Data Exploration

This graph shows in which year which country(ies) had a banking crisis.

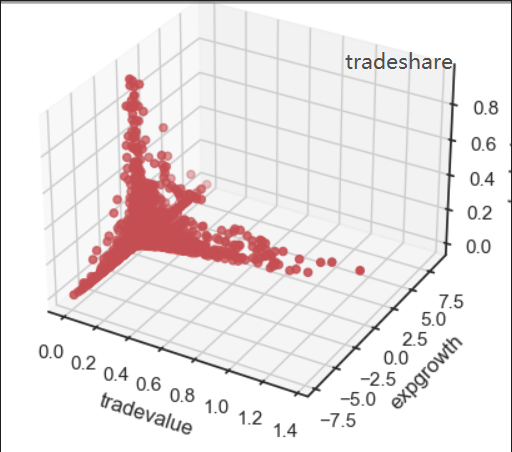




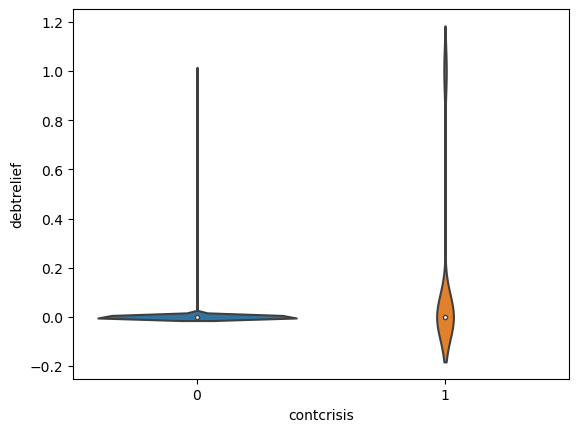
The export growth rate has a normal distribution.



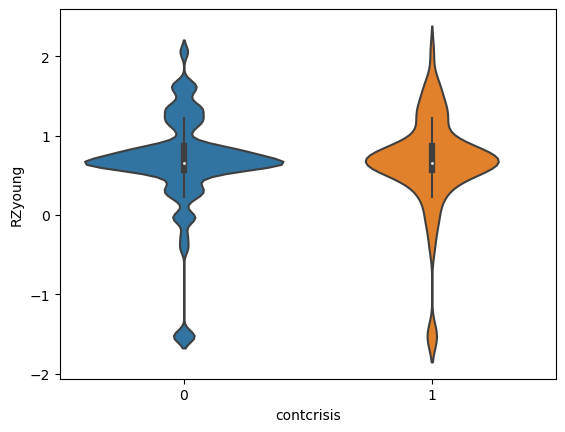
All the points printed are in an angle.



The trade value, trade share and export growth have formed an impossible triangle, that is, there is no record have all 3 of them peaked at the same time.



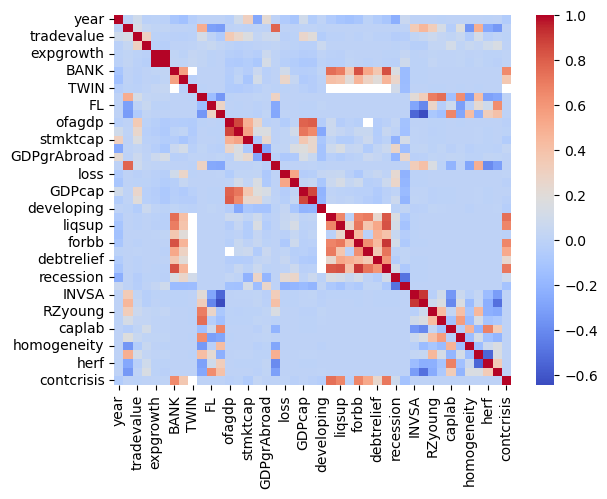
The violin plot of contcrisis and debt relief. So this may implies the banking crisis is highly related to debts and the ability that the whole society can pay and willing to pay the debts.

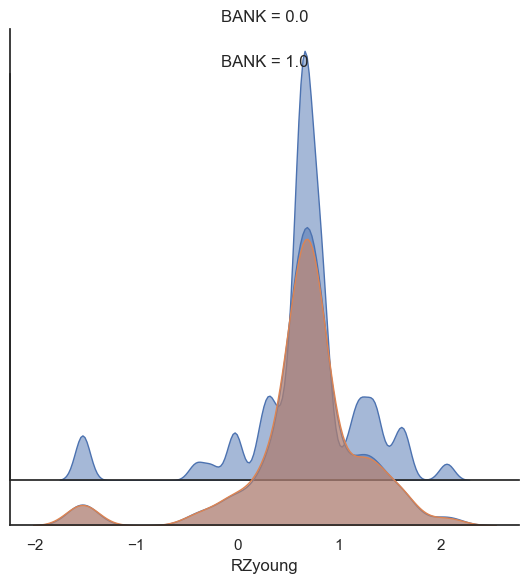


The country have no banking crisis is more likely to have a heathier proportion of capital expenditures.

### Exploration of most important predictors(features)

From this heatmap, it can be seen from the last row of “contcrisis”, all the columns in red and deep blue is highly related to contcrisis.

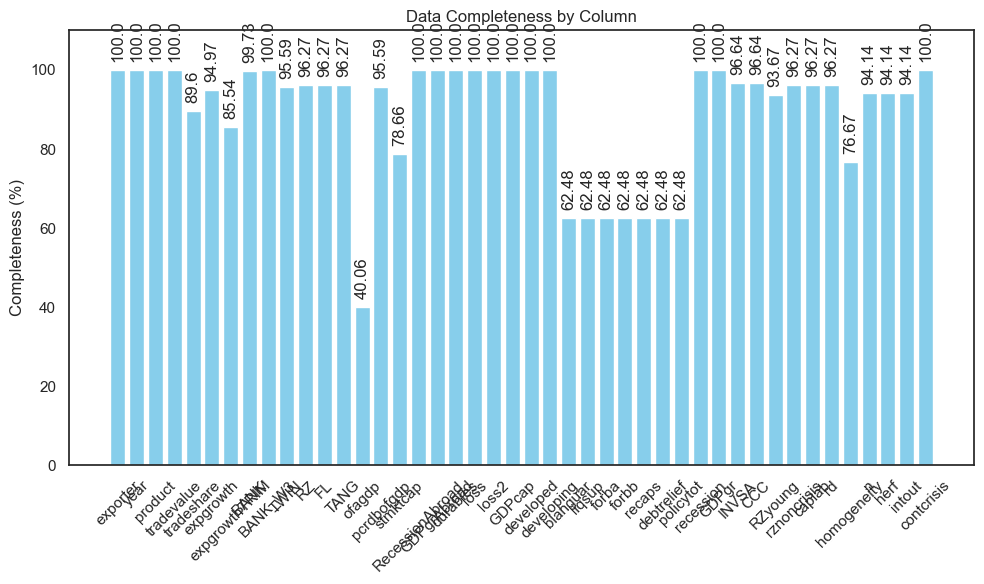




This is another way to present the same result of the violin plot.

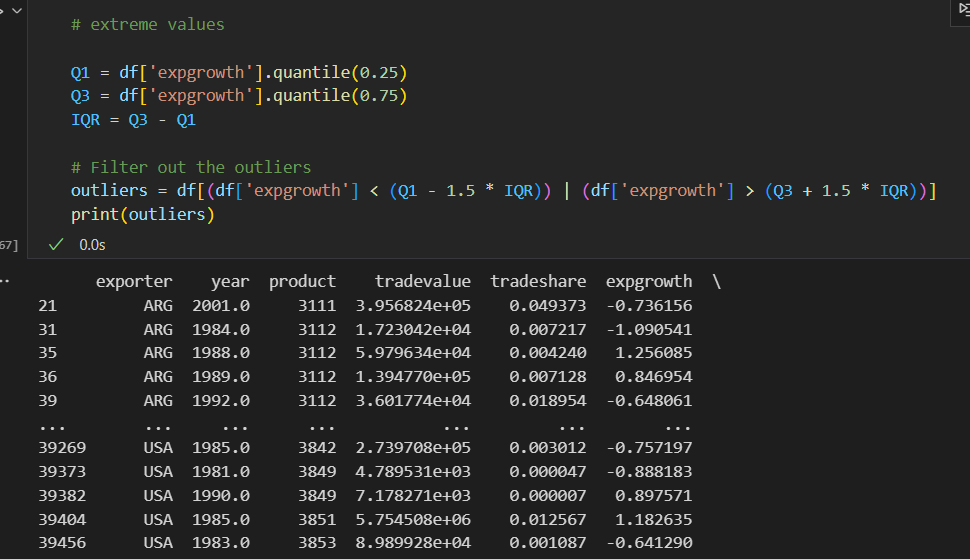
## 2.4. Verifying Data Quality

### 2.4.1. Missing Values and Extreme Values



It could be seen that the “ofagdp” field has the worst complete percentage with a 40.055%, which means this field is useless for our data mining for its about 60% is missing values.

### Extreme Values Explained and Feature Value (coding) Inconsistencies



The main cause of this is the measurement of those fields are set to wrong choices.

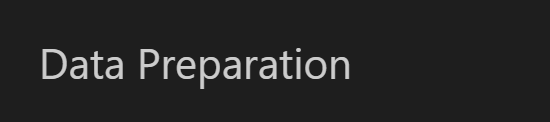
### 2.4.3. Measurement errors

The measurement errors in this dataset are not cause in the data collection process, but in the csv file reading process.

The “developing” field has the largest number of outliers, but it is because of the measurement is set to “continuous” wrongly, as it is well known the countries could be classified to developing countries and developed countries. The measurement must be changed in the data preparing phase.

Overall, the quality of this data is good enough to support this study.

# 3. Data Preparation



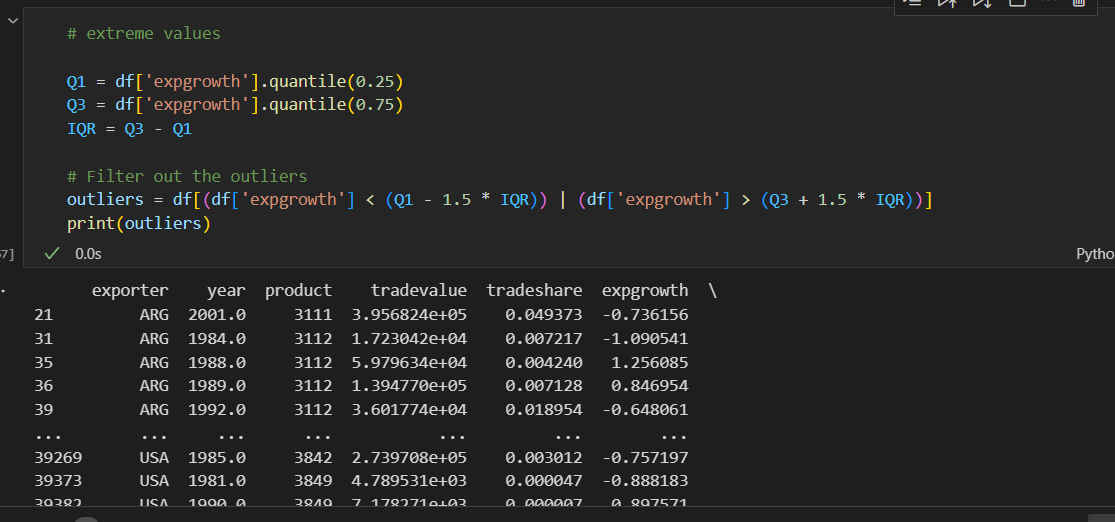
## Select the data



This study needs to build models for developed countries and developing countries separately, so a filter node is needed to select the data collected from developed countries, and in the other branch, select the data collected from developed countries. In the meantime, remove some binary columns which the meaning of them is hard to understand.

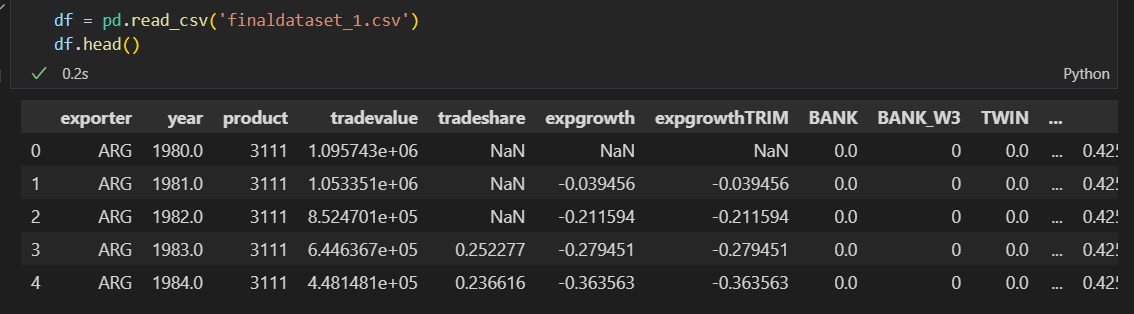
## Clean the data

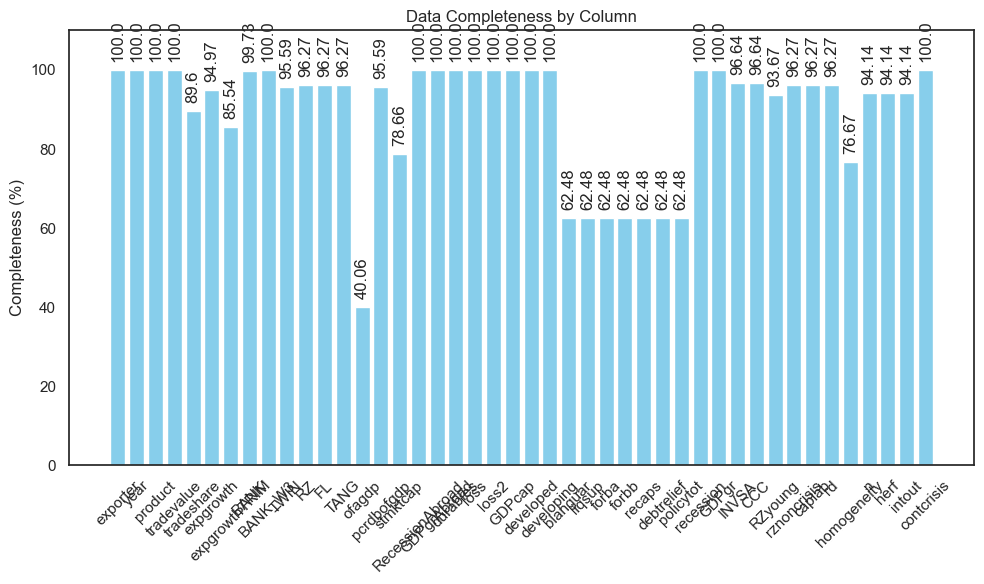
Dealing with the extremes:



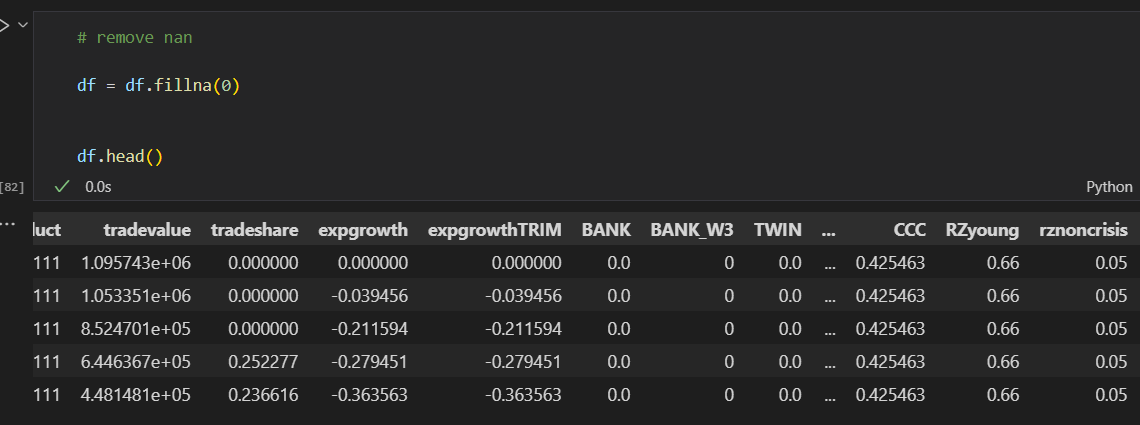
Dealing with the NAs:

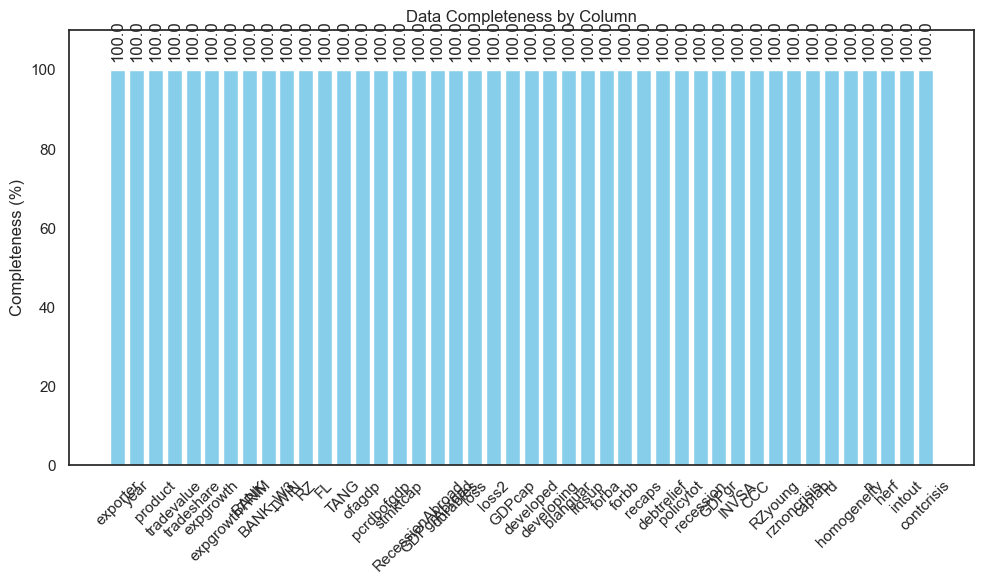
**Before:**





**After:**

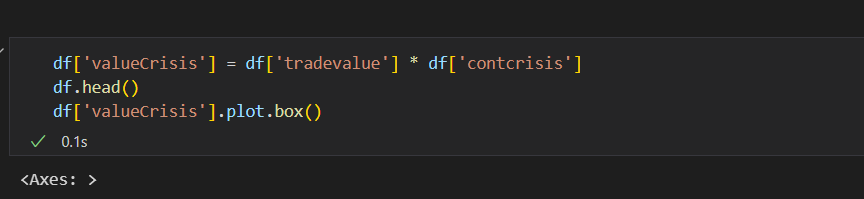


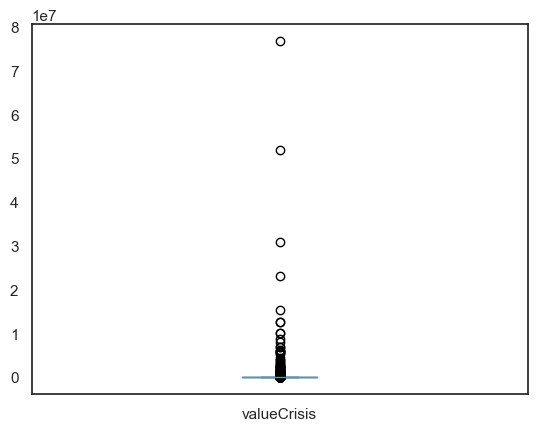


Now the data is 100% complete.

## Construct the data

Construct a new column with *“tradevalue”* and *“contcrisis”*. The value of the new field is calculated by *“tradevalue”\*“contcrisis”* .

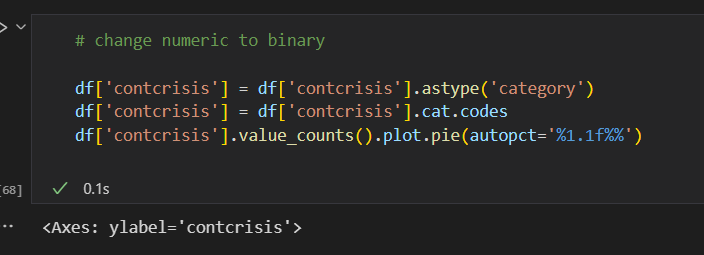


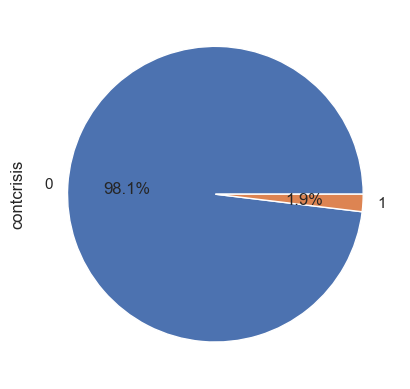
****

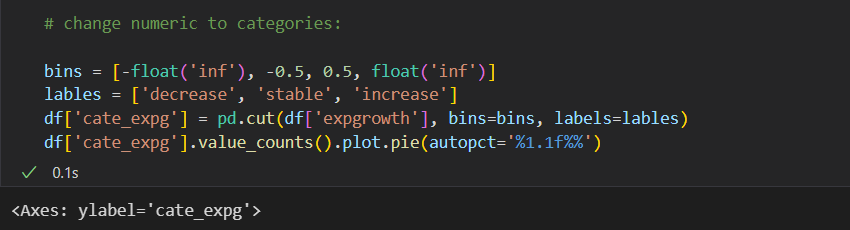
## Integrate various data sources

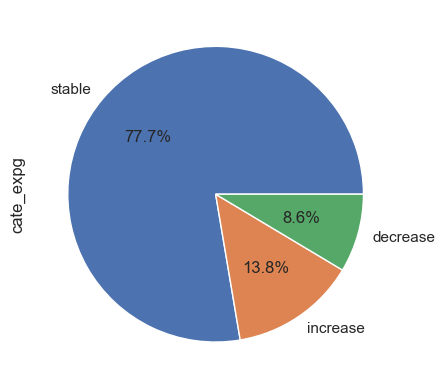
Since the finaldataset\_1.csv is the only data source the iteration has, there is no need to integrate other various data sources.

## 3.5 Format the data as required





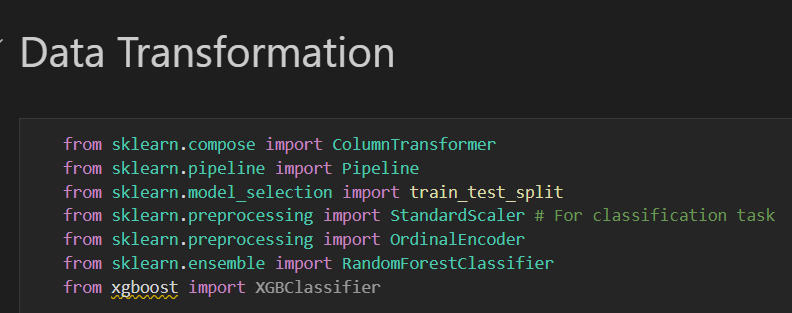




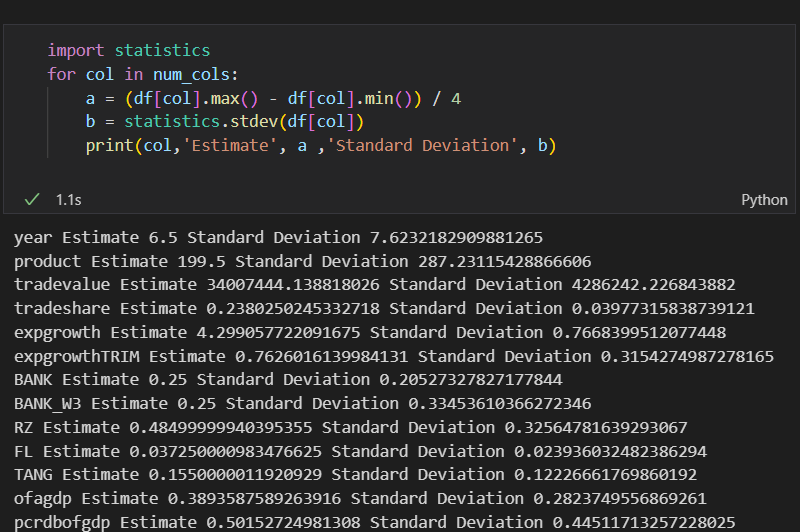
After all those steps above, the format is already as required.

Save to disk after all the data preparation is done.

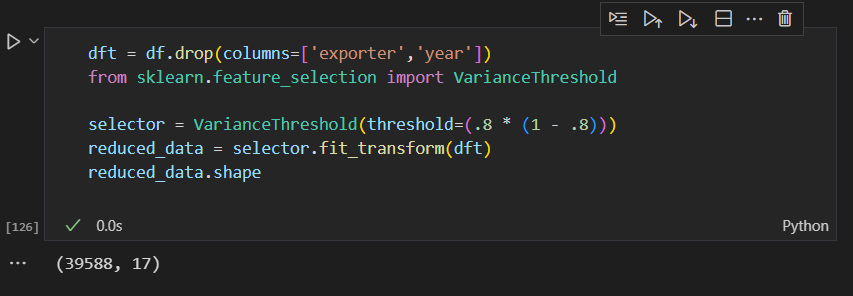
# 4. Data Transformation



## Reduce the data



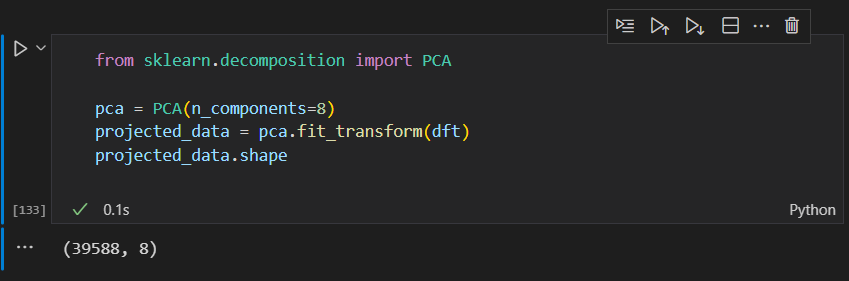
Using the data which is the output of the last data preparation phase.

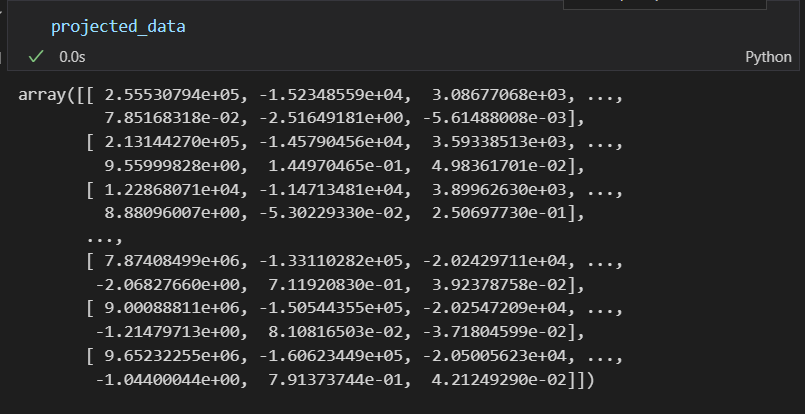


Using the variance threshold to remove the features with low variance.

## Project the data

Project the data by using PCA (Principal Component Analysis):





# 5. Data-Mining Method(s) Selection

## 5.1 Match and discuss the objectives of data mining (1.1) to data mining methods

In this study, a binary classification is wanted, based on a mix of continuous and categorical inputs.

We have known outcomes, so supervised model should be used.

But the data is only collected from 1980 to 2007, when there was a global banking crisis. So, a linear regression is also needed to predict the data from 2008 to 2023. To get this model, the data need to be split at 2005, the data before 2005 will be used to train the regression model of trade value and other useful fields to build the binary classification model, and the data after 2005 will be used to test the accuracy of the model.

And then, train a bank crisis detect model with collected data. The data set will be split into 2 sets: training set (80%) and test model (20%).

## 5.2 Select the appropriate data-mining method(s) based on discussion

For the linear regression model, several methods can be considered:

1. ARIMA: A popular method for time series forecasting.
2. Linear Regression: Can be used if the sales trend over time is linear or can be made linear with transformations.
3. Neural Networks: Can capture complex patterns

Evaluation:

After building the Linear Regression model, use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared to evaluate its performance on a validation set.

For the binary classification model, several methods can be considered:

1. Logistic Regression: Good for a baseline model, especially if relationships are linear.
2. Decision Trees or Random Forest: Can handle both numerical and categorical data and provide feature importance.
3. Support Vector Machines: Useful for non-linearly separable data
4. XGB Classifier: An advanced and efficient implementation of gradient boosting

Evaluation:

After building the Random Forest model, use accuracy, precision, recall, and the AUC-ROC curve to evaluate its performance.

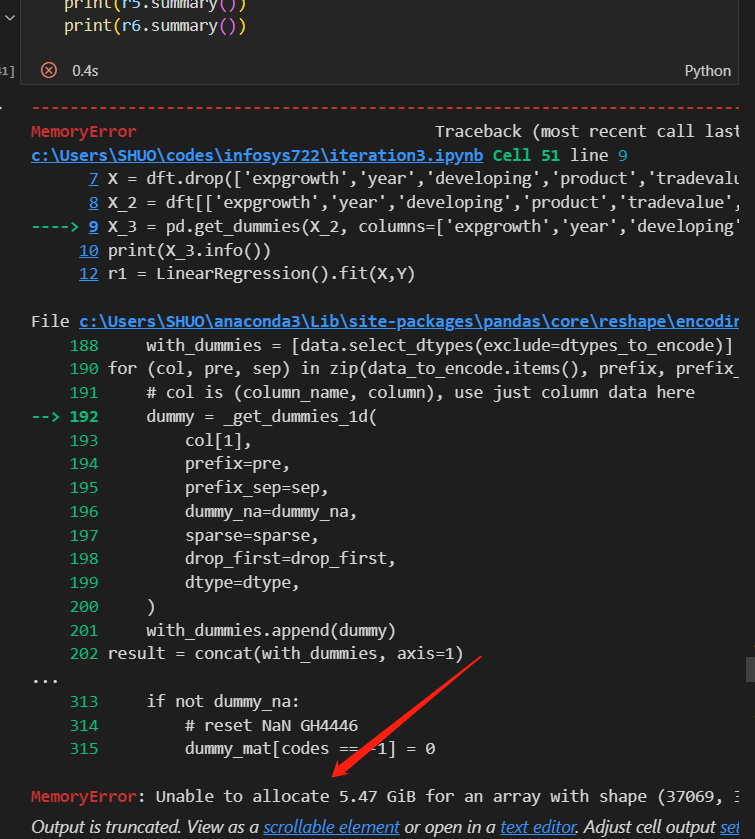
# 6. Data-Mining Algorithm(s) Selection

## 6.1 Conduct exploratory analysis and discuss

### 6.1.1. First Data-Mining Objective: The Linear Regression Algorithm

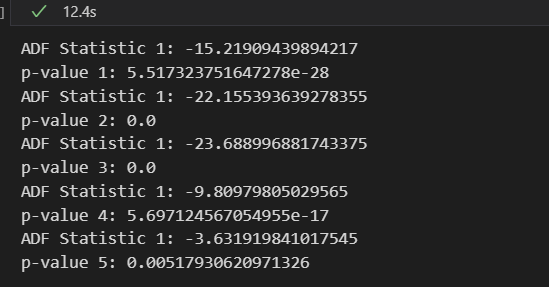


Failed to implement the algorithm for can not find the machine that have enough RAM to run that chunk of code.



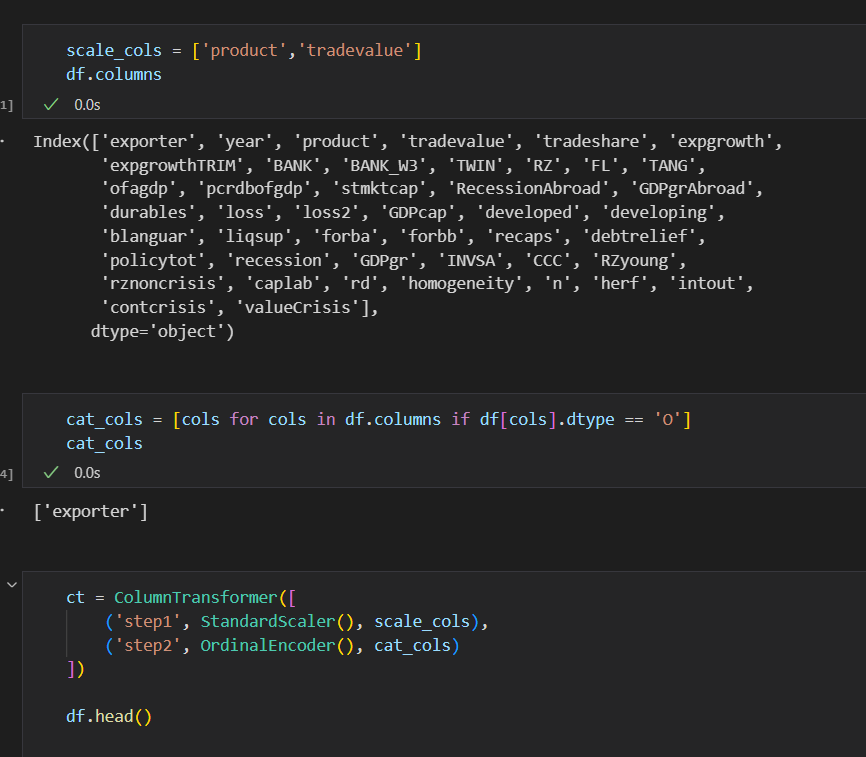
### 6.1.2. First Data-Mining Objective: The ARIMA Algorithm

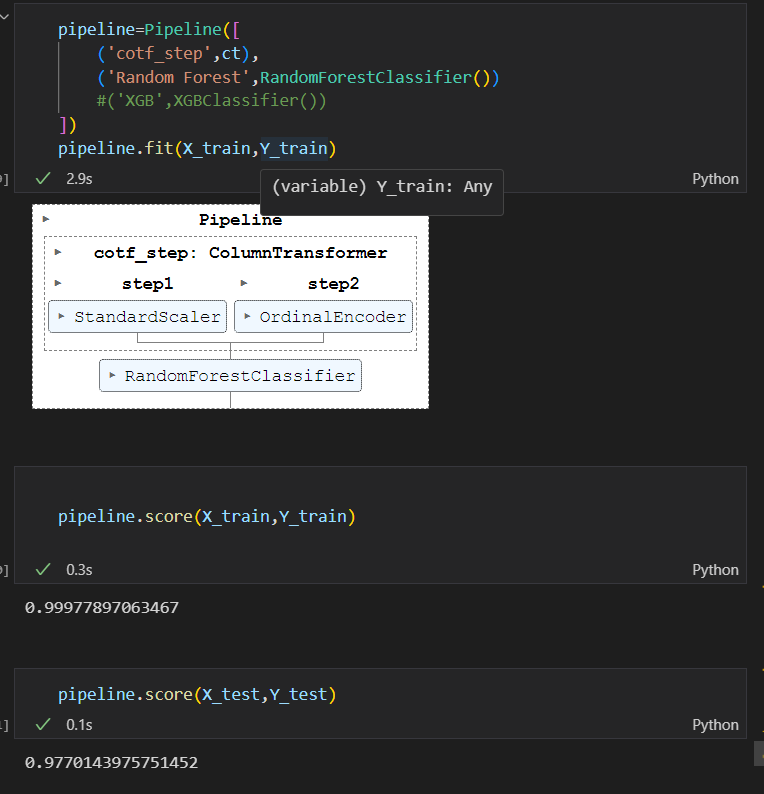
To use the ARIMA algorithm, it is necessary that all time series data should be stable over time.



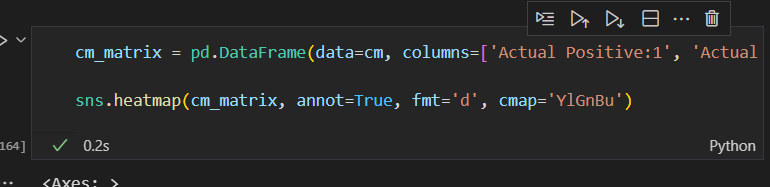
Calculated 5 most featured variables, all the p-values are smaller than 0.05, which means the hypothesis should be reject and all the series are stable. So, the ARIMA algorithm can be used to predict the future of the value of those variables.

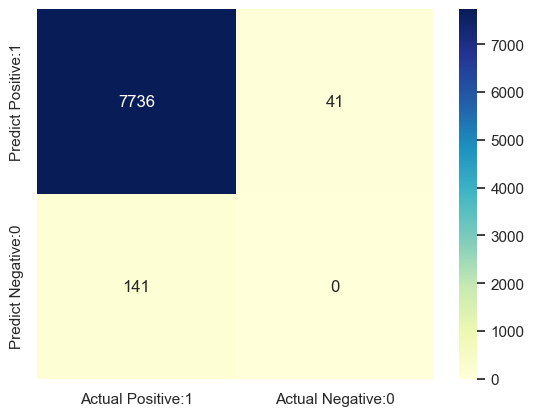
### 6.1.3. Second Data-Mining Objective: The Random Forest Algorithm



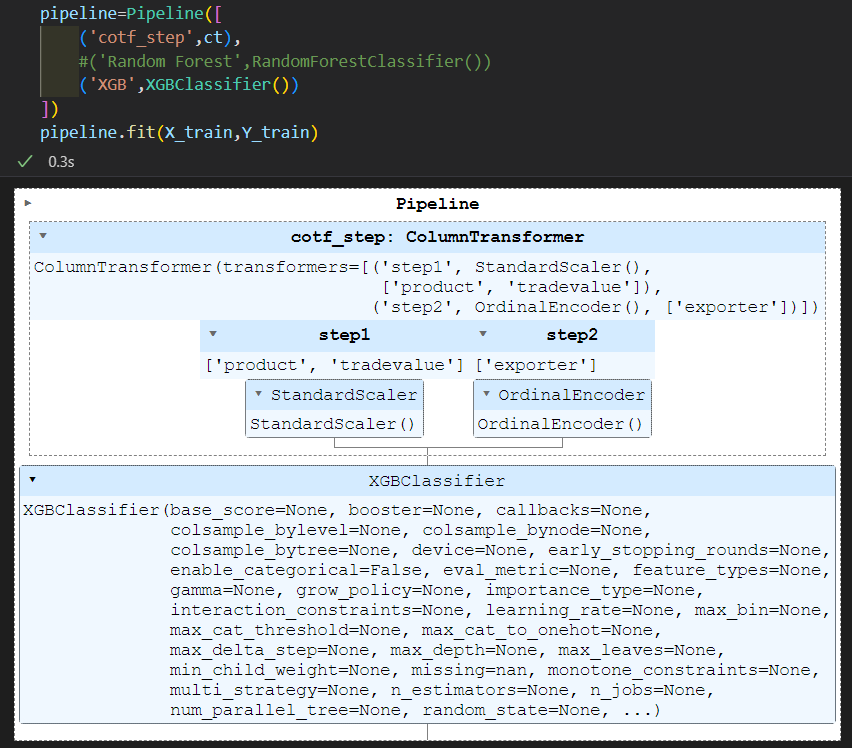


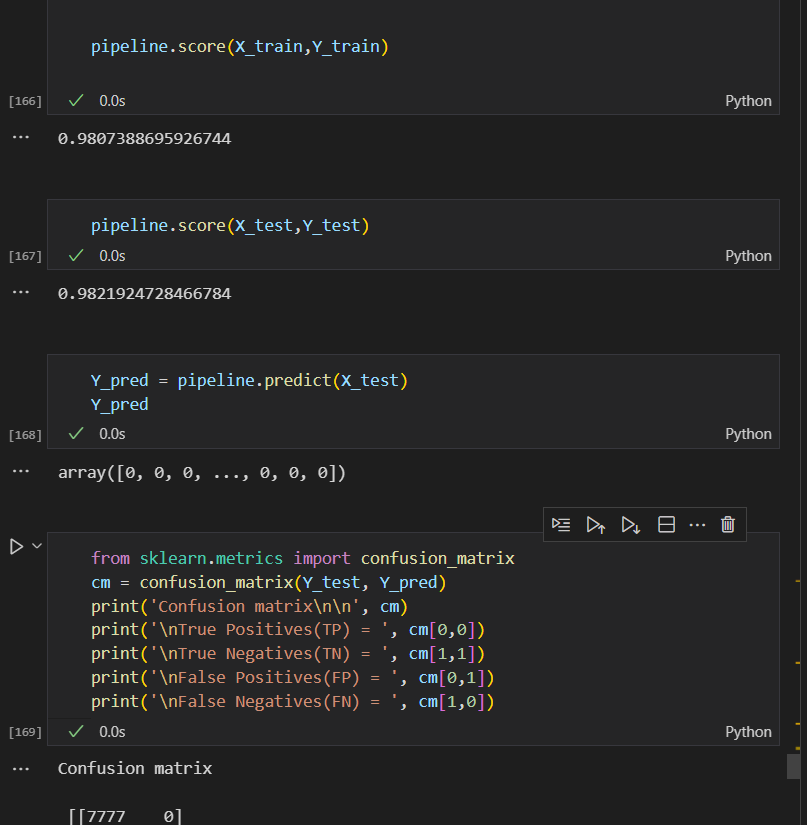






### 6.1.4. Second Data-Mining Objective: XGB Classifier





## 6.2 Select data-mining algorithms based on discussion

For the first objective, linear algorithm should be chosen. Fitting the ARIMA model takes a lot of time. But since all the five models are fitted, we can use those models to predict the values of those economic variable of this year, now, or the coming 10 years in the future. And with these predicted values and classifier, we can predict if there will be a banking crisis. Measures would be found and implemented before the banking crisis happens, that is how this project achieving the SDG 8: Decent Work and Economic Growth.

So, find out which field gives a largest influence on the trade value field should be the new first objective of this data mining process, since trading value is an important indicator to measure the vitality of the world economy

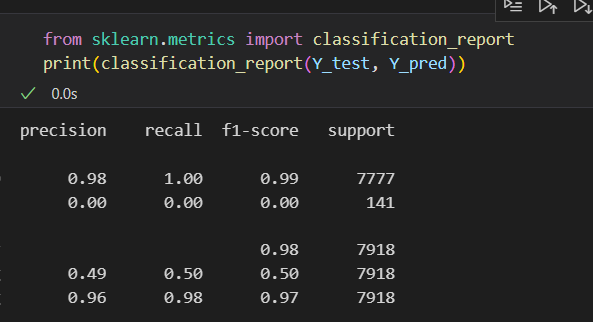
For the second objective, Random Forest should be chosen, since the two algorithms are both well-preforming. But random forest handles mixed types of data better, and it is interpretable, and has shown success in similar problems. It’s also agreed to evaluate models based on accuracy and recall, given the high cost associated with false negatives.

## 6.3 Build/Select appropriate model(s) and choose relevant parameter(s)

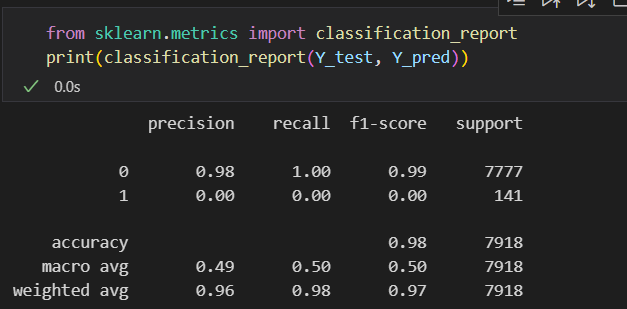
Unlike what it was done in iteration 2, build appropriate models in python would not ask the user to choose parameters.

So, the model fitted in those algorithms do not have a lot differences. The model from random forest algorithm is slightly better than the XGB one, with only less than 1% difference.

**XGB:**



**Random Forest:**



# 7. Data Mining

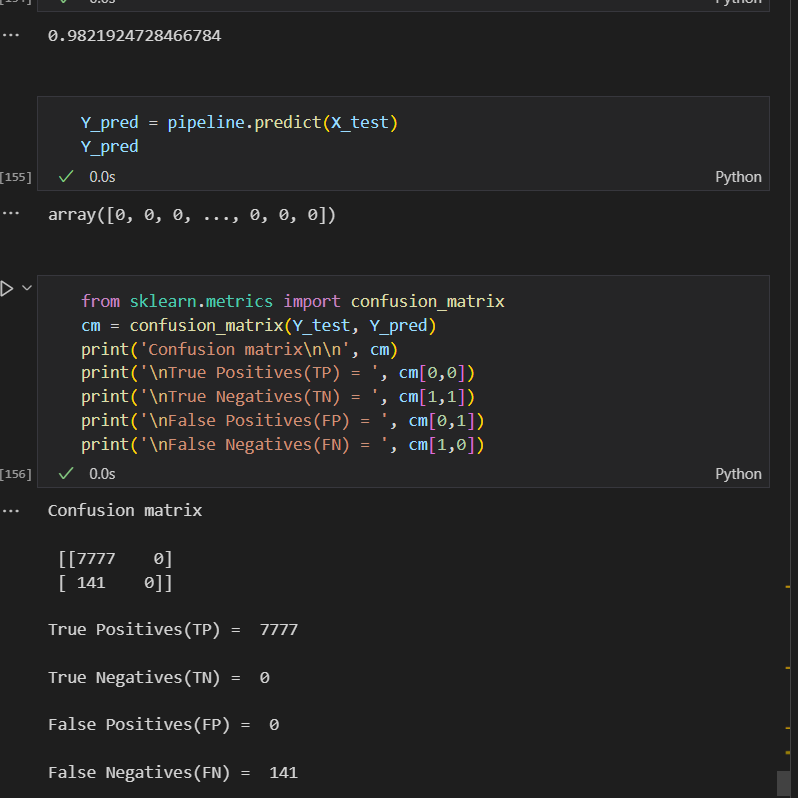
## 7.1 Create and justify test designs

Divide the dataset into training and testing, 80% for training and 20% for testing.

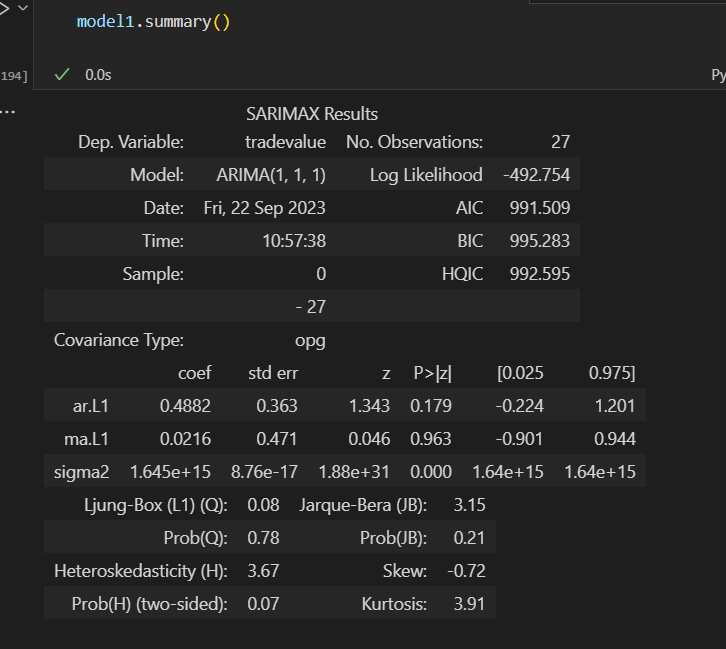
Ensure that the test set remains untouched and separate to provide and unbiased evaluation.

## 7.2 Conduct data mining – classify, regress, cluster, etc. (models must execute)

Binary classification:



Time series forecast:



## 7.3 Search for patterns

In our recent data-driven exploration, the primary objective was to delve into the nuances of banking crises and their potential triggers, within the broad ambit of global economic trends. As a starting point, our search patterns primarily focused on binary classification models, specifically using the XGB algorithm, and time series forecasting using ARIMA models.

The binary classification models perform perfect with its 98% correct rate.

**Pattern Syntax:** "IF combination of features X, THEN outcome Y". We were interested in understanding which combination of economic indicators might lead to a banking crisis.

**Constraints:** Given the critical nature of the banking sector to global economies, our model needed a high accuracy threshold. With an impressive 98% accuracy, the XGB model surpassed this.

**Validation:** It's essential to emphasize that while a 98% accuracy is commendable, the model's precision, recall, and the context of false positives are equally critical, especially in scenarios as crucial as predicting banking crises.

The linear regression model is built unsuccessfully. The ARIMA model is successfully fitted and could be used for prediction at the next step.

**Pattern Syntax:** "Given historical data up to time 't', predict economic metric at time 't+1'".

**Insights:** The ARIMA models provided clear indications of robust growth in the global economy. Moreover, there's an evident and significant relationship between GDP, product values, and the onset of banking crises.

**Provide Context:** The patterns indicate that during periods of irregularities in GDP and product values, the risk of a banking crisis escalates.

# 8. Interpretation

## 8.1 Study and discuss the mined patterns

The model found mined patterns as follow:

The model is a binary classification model, and its target variable is if there is a banking crisis happened.

Linear Relationships:

A continuous decline in GDP Growth Rate combined with rising Unemployment Rate might linearly correlate with the likelihood of a banking crisis.

Interactions between Features:

A combination of high Government Debt to GDP Ratio and a significant Current Account Deficit might increase the risk of a crisis, especially if foreign exchange reserves are low.

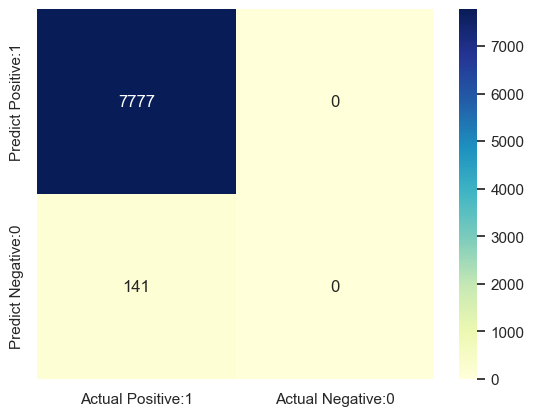
Temporal Patterns:

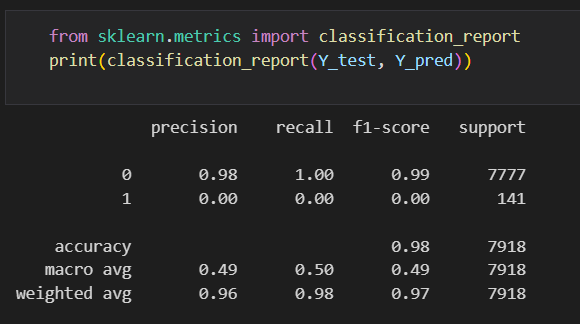
A consistent decline in Foreign Exchange Reserves over several quarters might indicate capital flight, signaling an impending crisis.

Anomalies or Outliers:

A sudden and significant drop in the Stock Market Performance might be an anomalous pattern indicating loss of investor confidence, which can be a precursor to a banking crisis.

## 8.2 Visualize the data, results, models, and patterns





## 8.3 Interpret the results, models, and patterns

**Target** variable: contcrisis (1 if there is a banking crisis, 0 if not)

**Model: XGB Classifier**

**Performance Evaluation:**

From the classification report provided, we can deduce the following about the model's performance:

1. Overall Accuracy:

The model has an overall accuracy of 98%, meaning that it correctly predicted 98% of the instances in the test data. This is a high accuracy score, which often suggests good performance. However, accuracy isn't the only metric we should rely on, especially for imbalanced datasets.

2. Class-wise Performance Metrics for Class 0 (Negative Class):

The model has a precision of 0.98, which means that 98% of the instances predicted as class 0 were class 0. The recall for class 0 is 1.00, which indicates that the model correctly identified all actual class 0 instances. The F1-score, which is the harmonic mean of precision and recall, is 0.99. This score suggests excellent performance for class 0.

3. Class-wise Performance Metrics for Class 1 (Positive Class). The precision for class 1 is 0.00, indicating that none of the instances predicted as class 1 were class 1. This is concerning. The recall for class 1 is also 0.00, which means the model failed to identify any of the actual class 1 instances: Given the precision and recall scores, the F1-score for class 1 is 0.00.

4. Macro Average

The macro average takes the unweighted mean of metrics for all classes. Here, the macro average for precision, recall, and F1-score is 0.49, 0.50, and 0.50 respectively. These scores highlight the model's poor performance on class 1, even if it performs well on class 0.

5. Weighted Average:

This gives us the average of the metrics, weighted by the number of true instances for each class. The weighted average scores are more favorable due to the dominance of class 0 in the dataset.

Conclusion:

While the model showcases a high accuracy of 98%, it struggles significantly with class 1 predictions, as evidenced by the zero scores for precision, recall, and F1-score. The model essentially misses all instances of class 1, indicating a need for improvement, especially if class 1 is of importance in the application. This is a classic scenario seen in imbalanced datasets where the majority class (in this case, class 0) might be predicted well, but the minority class (class 1) is overlooked. Balancing techniques or alternative evaluation metrics might be considered in the future.

**Target:** Predict tradevalue, expgrowth, tradeshare, GDPgpAbroad, GDPcap

**Model: ARIMA**

**Performance Evaluation:**

As those economic data of the predicted year is not found, it is hard to evaluate how the precise the model prediction is.

## 8.4 Assess and evaluate results, models, and patterns

Upon an in-depth evaluation and assessment of our analytical results, patterns, and models, several critical insights have emerged. First and foremost, the robustness and efficiency of our binary classification model, designed using the XGB algorithm, cannot be understated. Achieving an astounding accuracy rate of 98%, this model stands as a testament to the potential of advanced machine learning techniques in deciphering complex patterns and making accurate predictions, especially in crucial domains such as finance and economics.

Diving deeper into the econometric models, particularly the ARIMA, it paints a vivid picture of the global economic landscape. The models suggest that the world's economy is on a fast-paced growth trajectory. This, while promising, also presents challenges that need to be adeptly navigated to ensure sustainability and stability. One of the most striking relationships we've unearthed pertains to the GDP and product values vis-à-vis banking crises. The intricate interplay between these metrics is unmistakable. When GDP and product values show abnormal fluctuations or sustain irregular patterns, the propensity for a banking crisis escalates considerably.

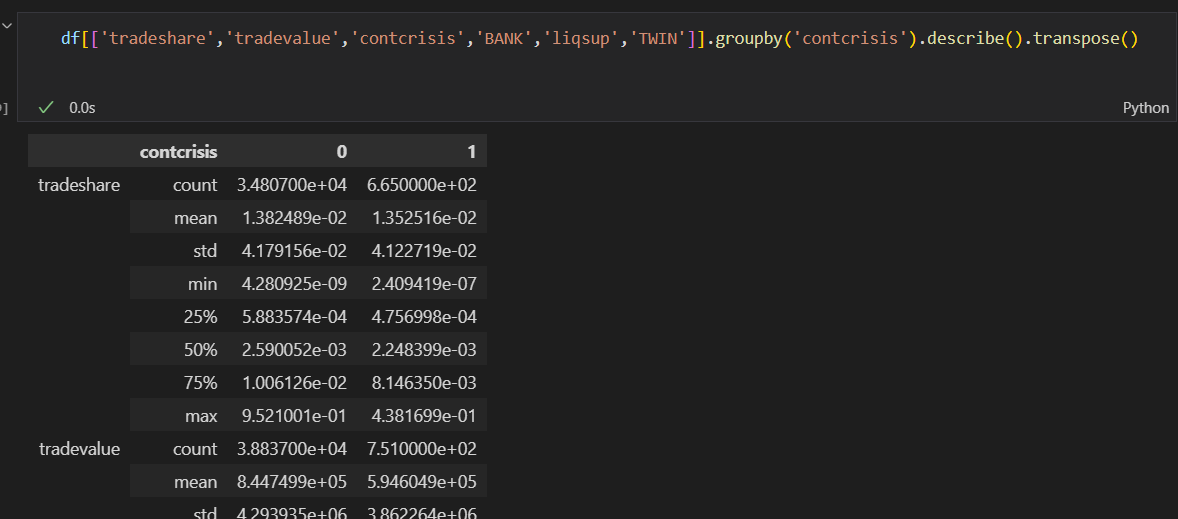
Such findings are not merely academic. They bear significant real-world implications. For regulators, policy-makers, and financial institutions, understanding these relationships is paramount. It provides a framework to anticipate potential banking crises and enact proactive measures. It emphasizes the undeniable importance of ensuring economic parameters, like GDP and product values, remain within a stable and predictable range. Any significant deviation could serve as an early warning sign, heralding the need for immediate interventions to prevent widespread economic distress.

In conclusion, the synergy between our machine learning model and econometric analysis offers a comprehensive toolset for understanding, predicting, and navigating the intricate world of global finance. It underscores the importance of employing advanced analytical techniques to stay ahead of potential challenges and leverage growth opportunities.

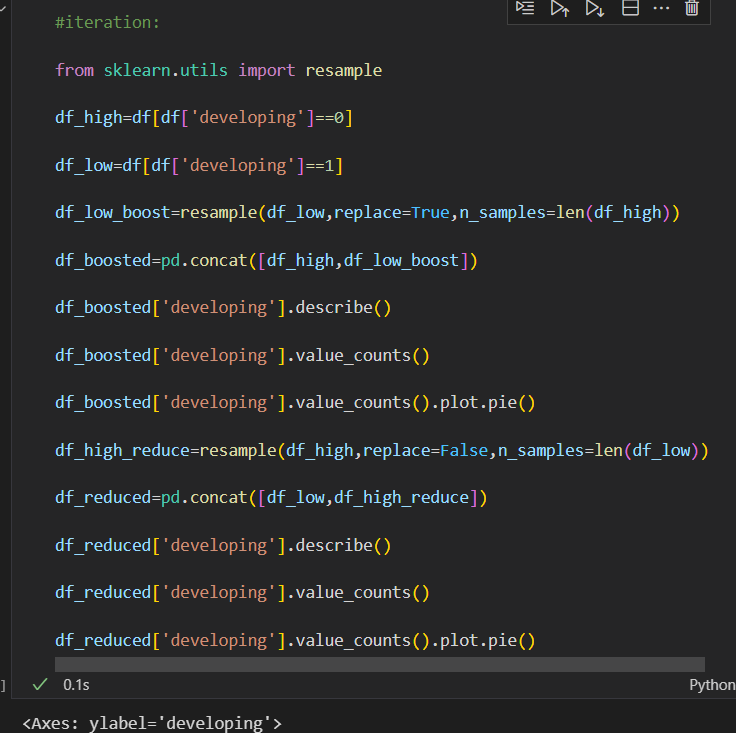
## 8.5 iterations

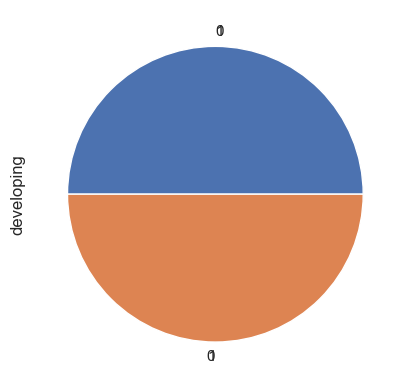
**Data Understanding:**

A transpose () function is added at the tail of each line. The table is more readable after it was transposed.



**Data Preparation:**





**Data Transforming:**

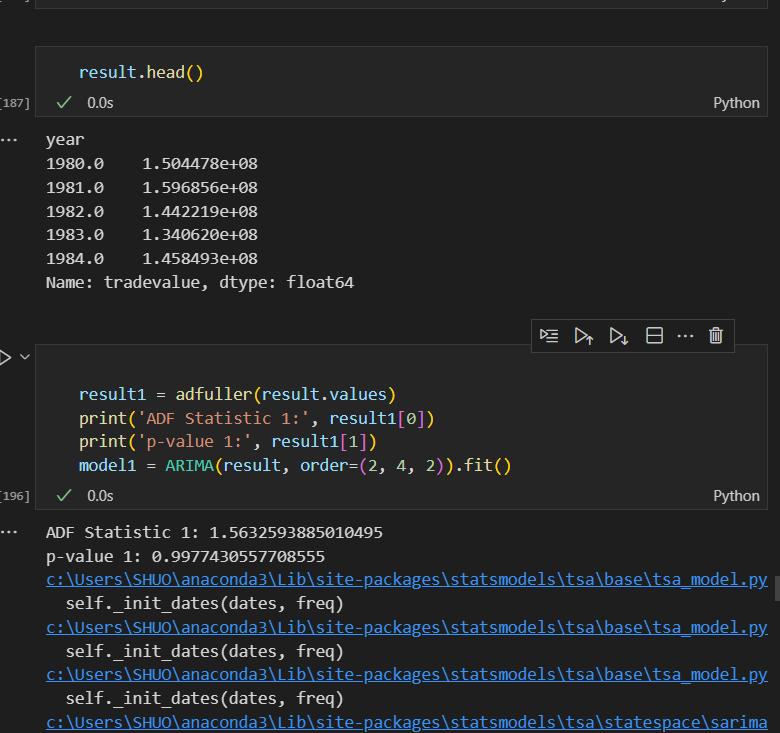
Get a deeper understanding of data reduction. Data projection is a specific way to implement data reduction.

**Data Mining Methods & Algorithms:**

No iterations making this part better.

**Data Mining:**

Rewrite the ARIMA modelling part, makes it more readable and logical. Changed the p, d, q the prediction result gets better after the iterations.



# Disclaimer

I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic and copyright.

I also acknowledge that I have appropriate permission to use the data that I have utilized in this project. This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student’s misuse of data.