ITERATION 4

BDAS

Shuo Feng

UPI: sfen779

ID:836311644

Table of Contents

[1. Business/Situation Understanding 4](#_Toc143467498)

[1.1. Business/Situation Objectives 4](#_Toc143467499)

[1.2. Assessment of the Situation 4](#_Toc143467500)

[1.2.1. Resource Inventory 4](#_Toc143467501)

[1.2.2. Requirements, Assumptions, and Constrains 5](#_Toc143467502)

[1.2.3. Risks and Contingencies 5](#_Toc143467503)

[1.2.4. Cost/Benefit Analysis 6](#_Toc143467504)

[1.3. Data Mining Objectives 7](#_Toc143467505)

[1.3.1. Data Mining Success Criteria 7](#_Toc143467506)

[1.4. Project Plan 9](#_Toc143467507)

[2. Data Understanding 10](#_Toc143467508)

[2.1. Collecting Initial Data 10](#_Toc143467509)

[2.2. Describing Data 10](#_Toc143467510)

[2.3. Data Exploration 18](#_Toc143467511)

[2.3.1. Exploration of most important predictors(features) 21](#_Toc143467512)

[2.4. Verifying Data Quality 23](#_Toc143467513)

[2.4.1. Missing Values and Extreme Values 23](#_Toc143467514)

[2.4.2. Extreme Values Explained and Feature Value (coding) Inconsistencies 24](#_Toc143467515)

[2.4.3. Measurement errors 24](#_Toc143467516)

[3. Data Preparation 25](#_Toc143467517)

[3.1 Select the data 25](#_Toc143467518)

[3.2 Clean the data 27](#_Toc143467519)

[3.3 Construct the data 30](#_Toc143467520)

[3.4 Integrate various data sources 30](#_Toc143467521)

[3.5 Format the data as required 31](#_Toc143467522)

[4. Data Transformation 32](#_Toc143467523)

[4.1 Reduce the data 32](#_Toc143467524)

[4.2 Project the data 34](#_Toc143467525)

[5. Data-Mining Method(s) Selection 36](#_Toc143467526)

[5.1 Match and discuss the objectives of data mining (1.1) to data mining methods 36](#_Toc143467527)

[5.2 Select the appropriate data-mining method(s) based on discussion 37](#_Toc143467528)

[6. Data-Mining Algorithm(s) Selection 39](#_Toc143467529)

[6.1 Conduct exploratory analysis and discuss **错误!未定义书签。**](#_Toc143467530)

[6.1.1. First Data-Mining Objective: The Linear Algorithm **错误!未定义书签。**](#_Toc143467531)

[6.1.2. First Data-Mining Objective: The Regression Algorithm **错误!未定义书签。**](#_Toc143467532)

[6.1.3. Second Data-Mining Objective: The Classification Algorithm **错误!未定义书签。**](#_Toc143467533)

[6.1.4. Second Data-Mining Objective: The Logistic Regression Algorithm **错误!未定义书签。**](#_Toc143467534)

[6.1.5. Second Data-Mining Objective: SVM **错误!未定义书签。**](#_Toc143467535)

[6.2 Select data-mining algorithms based on discussion **错误!未定义书签。**](#_Toc143467536)

[6.3 Build/Select appropriate model(s) and choose relevant parameter(s) **错误!未定义书签。**](#_Toc143467537)

[7. Data Mining 44](#_Toc143467538)

[7.1 Create and justify test designs 44](#_Toc143467539)

[7.2 Conduct data mining – classify, regress, cluster, etc. (models must execute) 45](#_Toc143467540)

[7.3 Search for patterns 48](#_Toc143467541)

[8. Interpretation 51](#_Toc143467542)

[8.1 Study and discuss the mined patterns 51](#_Toc143467543)

[8.2 Visualize the data, results, models, and patterns 52](#_Toc143467544)

[8.3 Interpret the results, models, and patterns 54](#_Toc143467545)

[8.4 Assess and evaluate results, models, and patterns 55](#_Toc143467546)

[Disclaimer 59](#_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 project is to support the United Nations' Sustainable Development Goal 8 (SDG 8) - Decent Work and Economic Growth. We aim to understand the relationship between banking crises and exports and determine how banking crises can impact economic growth and employment opportunities. By leveraging data mining techniques, we hope to provide actionable insights that can guide policymakers in strengthening economic resilience and promoting sustainable growth.

## 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.

We have access to a dataset from Kaggle that provides information on banking crises and exports. Additionally, we will utilize data processing tools like PySpark and visualization tools such as Matplotlib and Seaborn.

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

To achieve our objective, we need to:

1. Clean and preprocess the dataset.
2. Conduct exploratory data analysis.
3. Build predictive models.
4. Extract actionable insights.

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
5. The dataset is representative of the global situation.
6. Banking crises have a direct or indirect impact on exports and, consequently, on economic growth.

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. Furthermore, the dataset might not cover all countries or all relevant years. And external factors not present in the dataset can also influence exports and economic growth.

### 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. |
| Biases in the data | In case of data insufficiency, we might need to source additional data. |

### 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

Our primary data mining goal is to determine the relationship between banking crises and exports. We aim to:

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.

Predict the likelihood of a country experiencing a banking crisis based on economic indicators.

Understand the impact of banking crises on a country's exports and overall economic 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.

These goals align with our business objective of supporting SDG 8 by providing insights into factors that can hinder economic growth.

### 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.

|  |  |  |
| --- | --- | --- |
| **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)  Biases in the data |
| Data preparation | 15 | IT system failure  (code version control system failure)  Biases in the data |
| Data transformation | 5 | IT system failure  (code version control system failure)  Biases in the data |
| 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 knowledge) |
| 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) |

## 1.4. Project Plan

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:

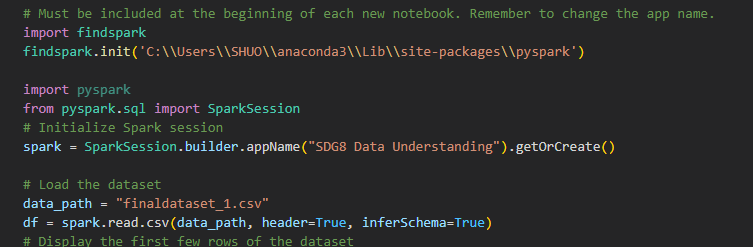
<https://datacatalog.worldbank.org/search/dataset/0041188>

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

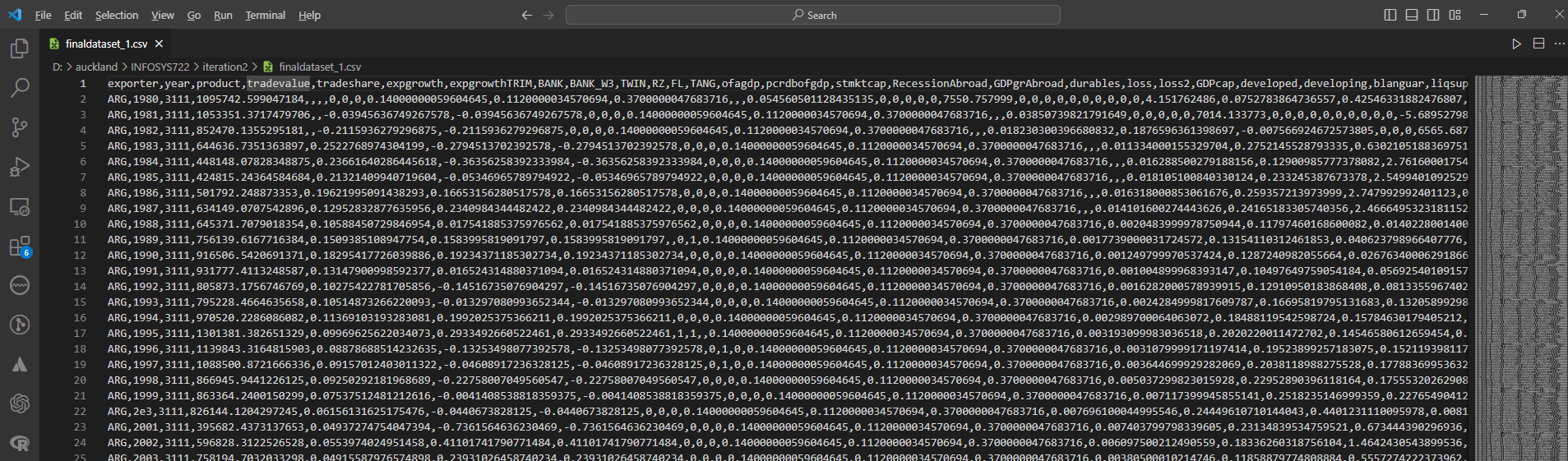
The World Bank gathered the data from historical reports and documents, so there will be a structural data missing—something bad happened in those countries made them can not report the economic data.

## 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. It is a large dataset, so PySpark is used in this iteration.



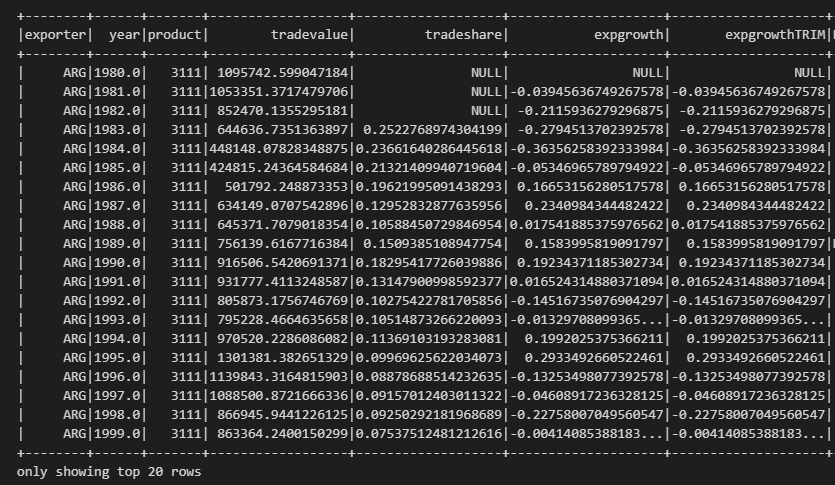
Open it in Visual Studio Code to see its original text looks:

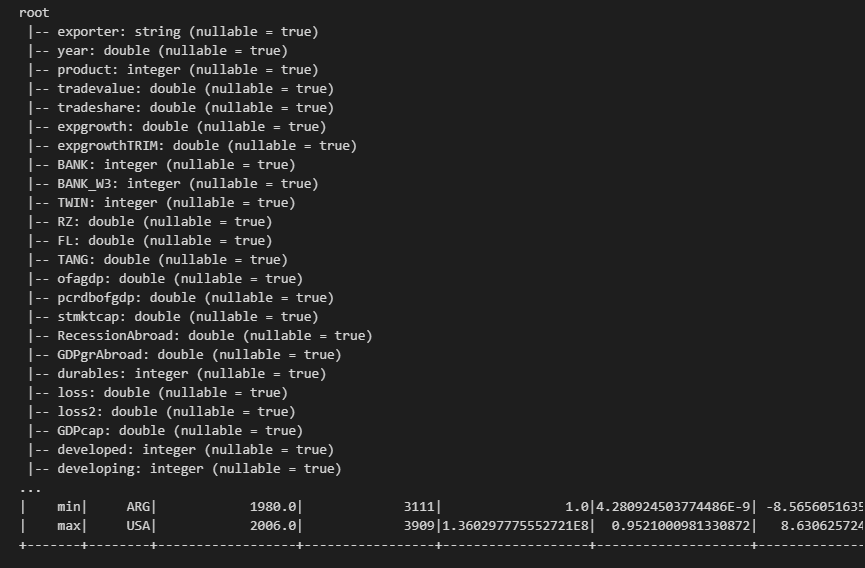


Most of the variables in this data have a numerical type, some flag type and some categorical type variables are also included.

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.



To understand the dataset's structure and content, we'll explore its columns, data types, and basic statistics.

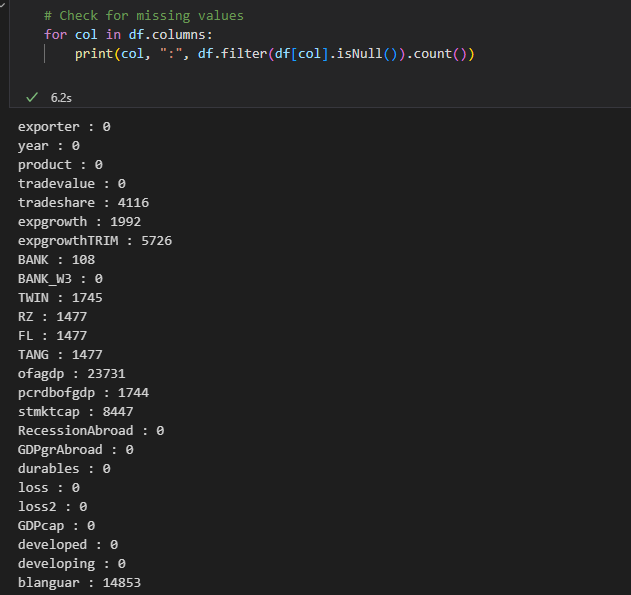
It is a data frame with 39588 rows and 44 columns.

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.

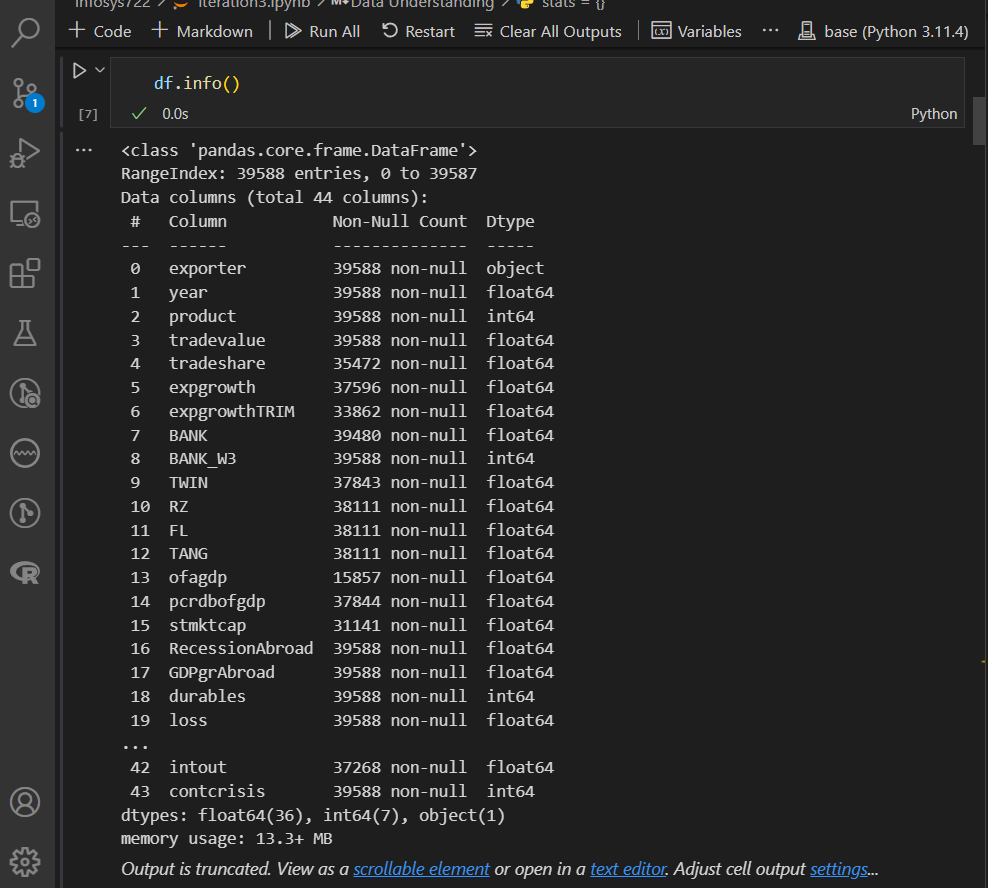
The first step is finding the variables most related to “contcrisis” through analysis, because if there is an “1” in that column, there was a banking crisis (the only meaning of columns that should be known), and then understand the relationships of these related variables.

Several data understanding code chunks are created in VS Code.

It could be seen that there are many missing values. This may be caused by the countries did not have a chance to do the census.

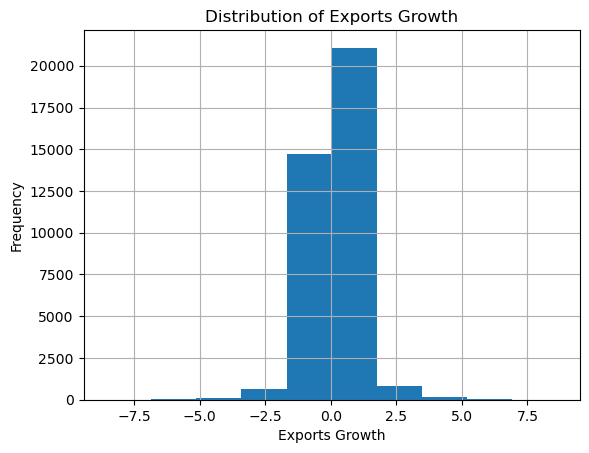


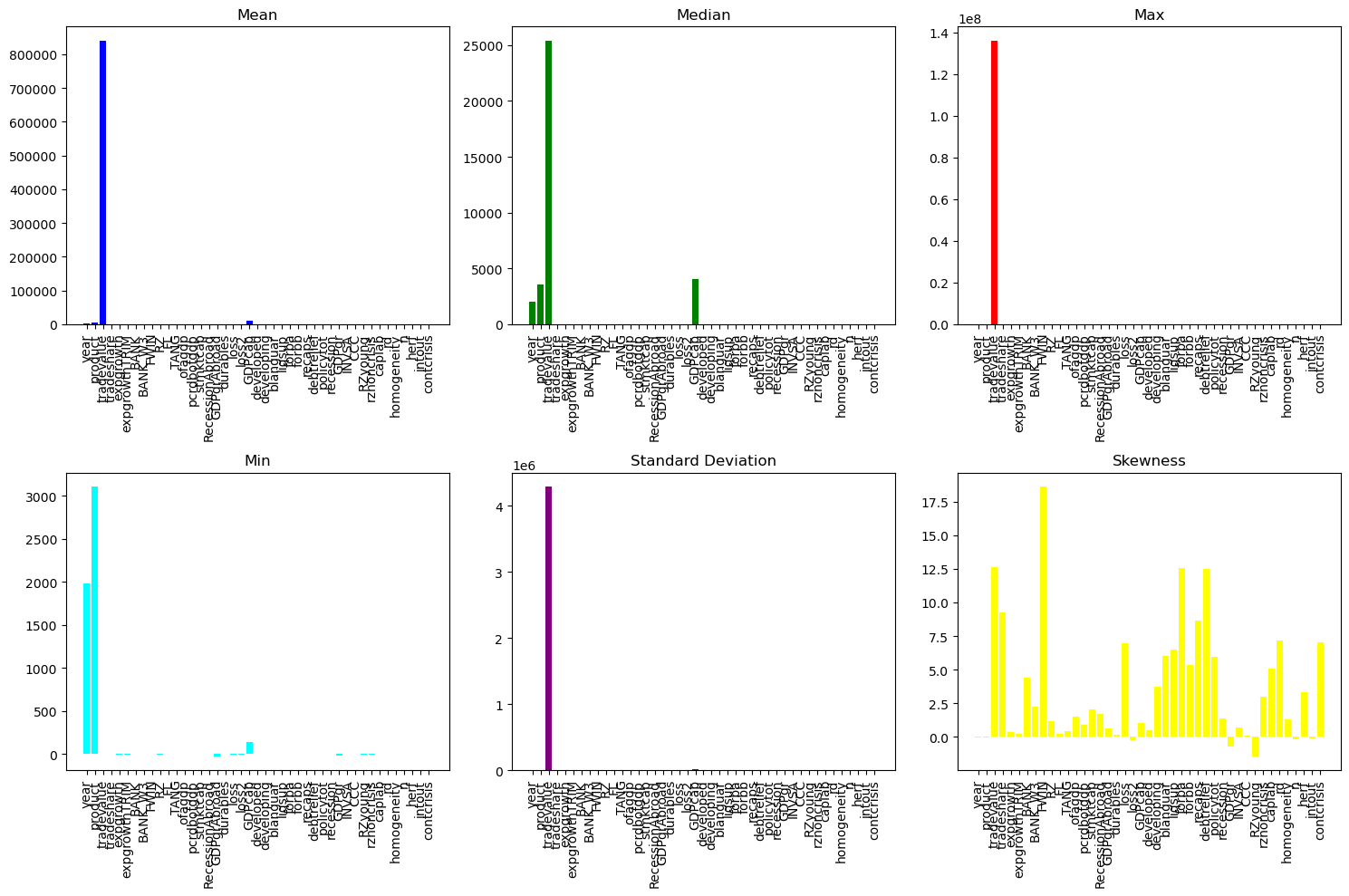
Before exploration, the data types of the columns should be considered.



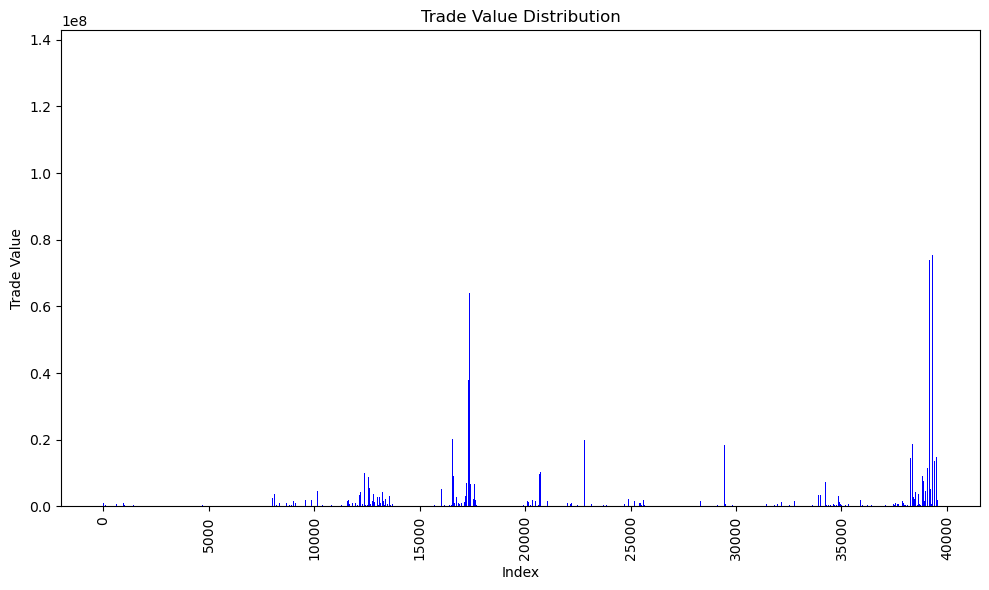
For the columns do not have Dtype in “int64” or “float64”, a calculation is not needed.

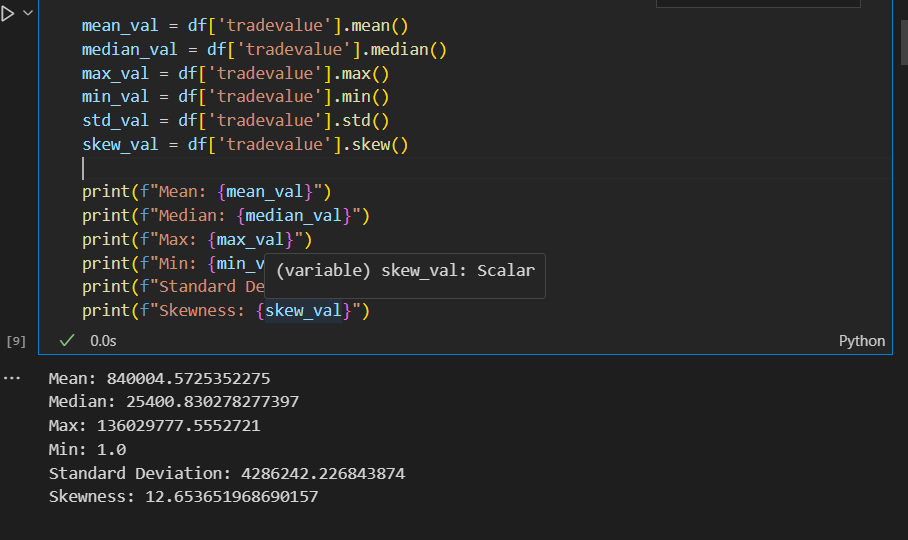
After some calculation, the maximum value, minimum value, mean value, standard deviation, skewness is clearly shown.

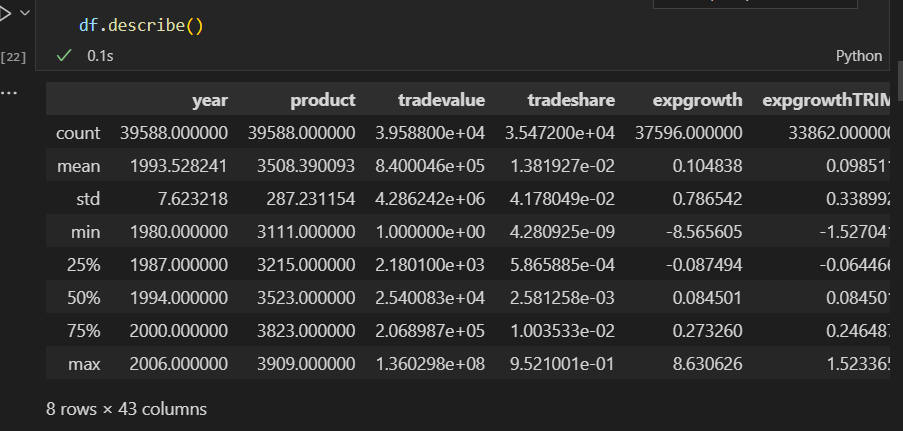




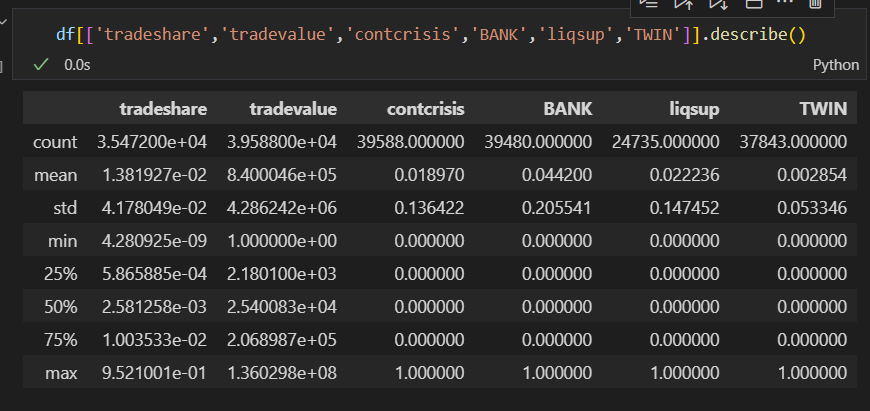
From this, it is obvious “tradevalue” is obviously far larger than the others, so a better way to show all those statistical values is to show them by column, not by rows. At the meantime, this implies “tradevalue” may play a very important role in detecting the banking crisis.

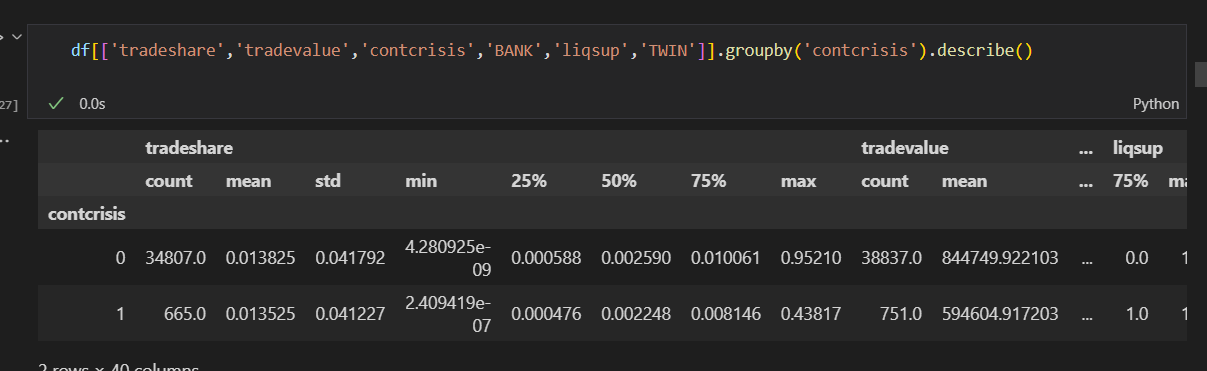




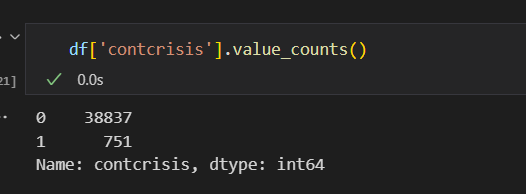


Here are some specific fields in this data (all above could also be done with “describe” function):



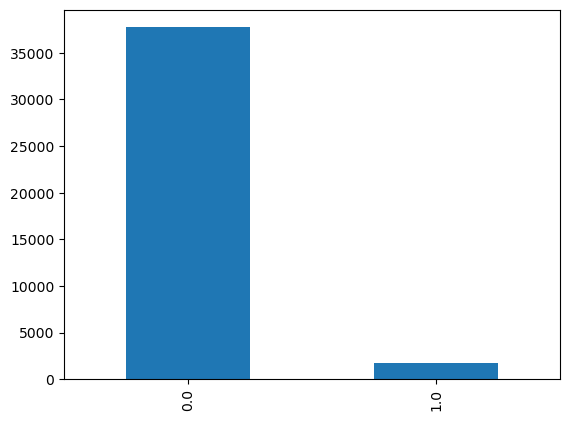


The good news from value counting: the banking crisis do not happen a lot.

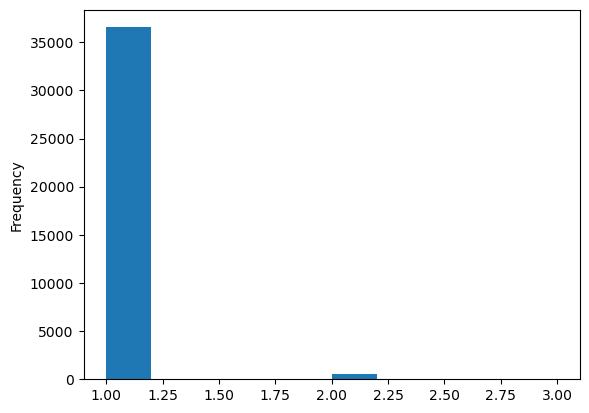


Some Graphs is created to describe those fields.

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

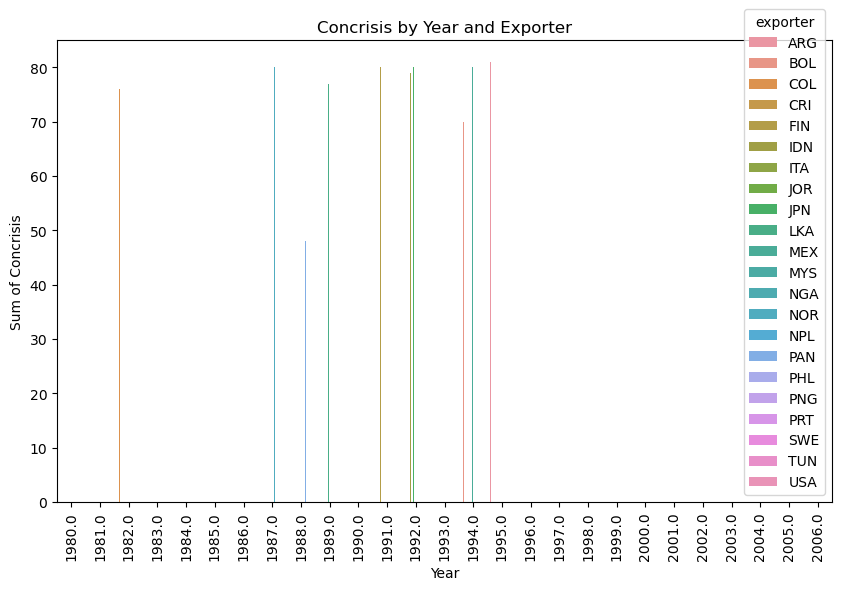


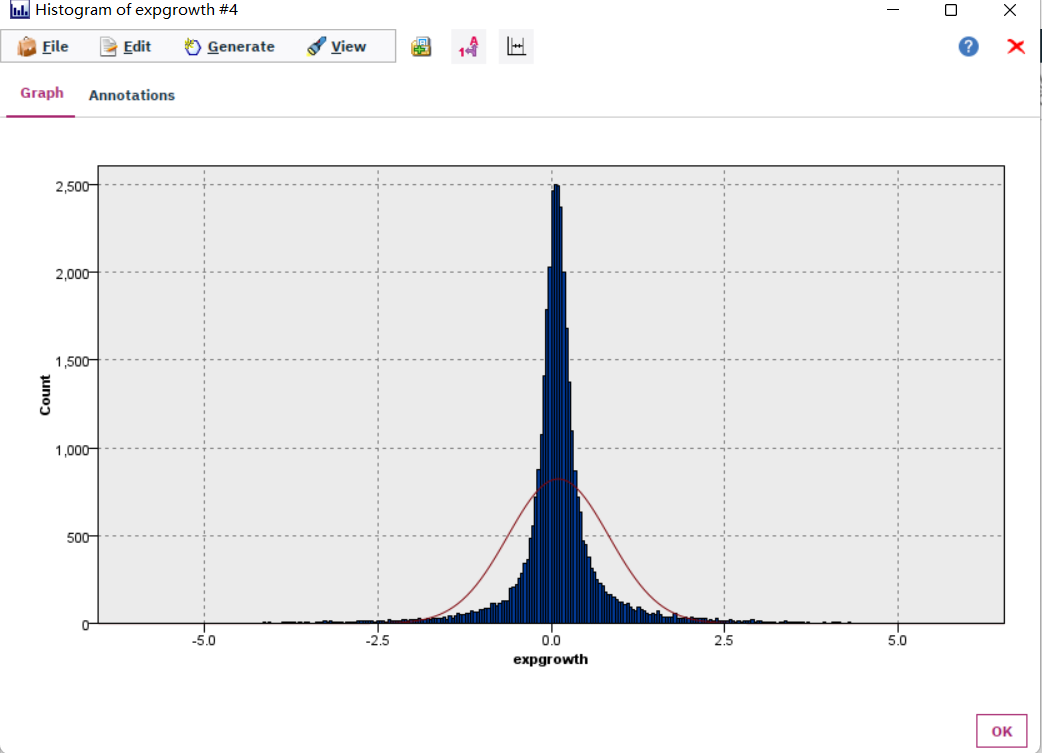
It could be seen that there are some outliers in export growth rate. Around 2, but most of the record have export growth rate less than 1.25.



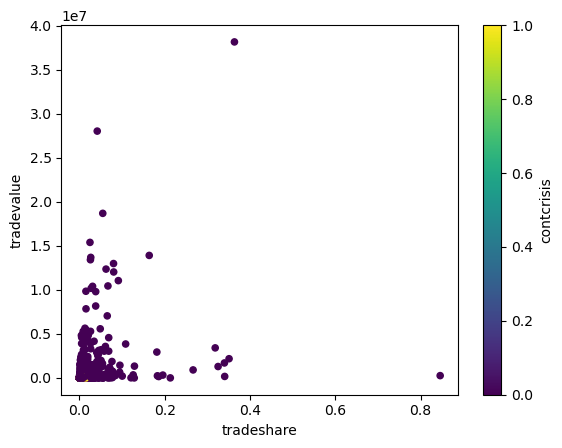
## Data Exploration

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

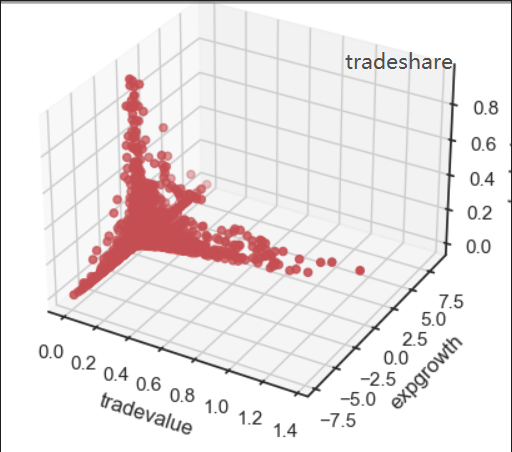




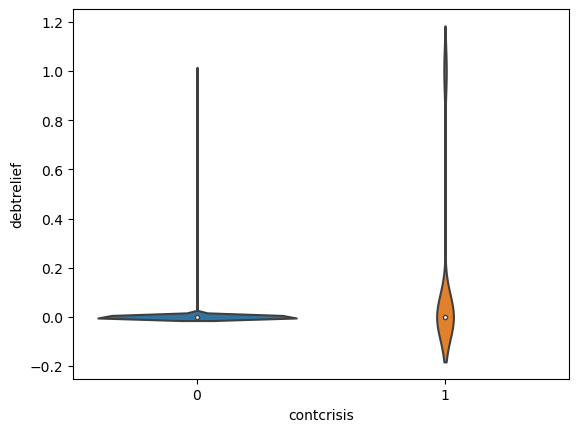
The export growth rate has a normal distribution.



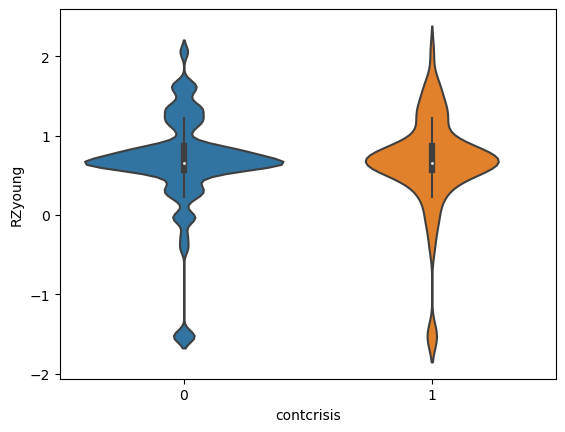
All the points printed are in an angle.



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.



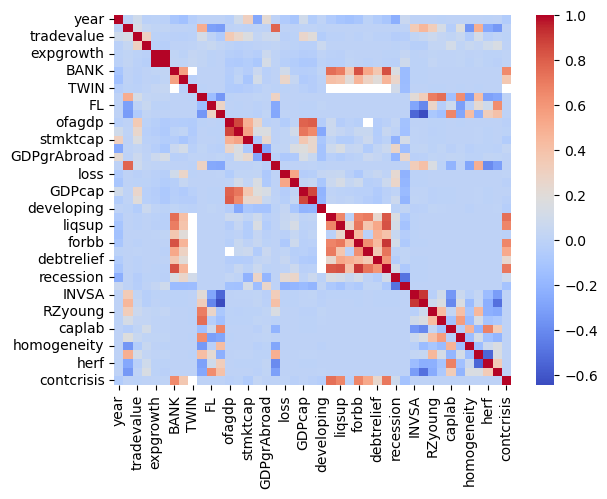
The violin plot of contcrisis and debt relief. So this may implies the banking crisis is highly related to debts and the ability that the whole society can pay and willing to pay the debts.

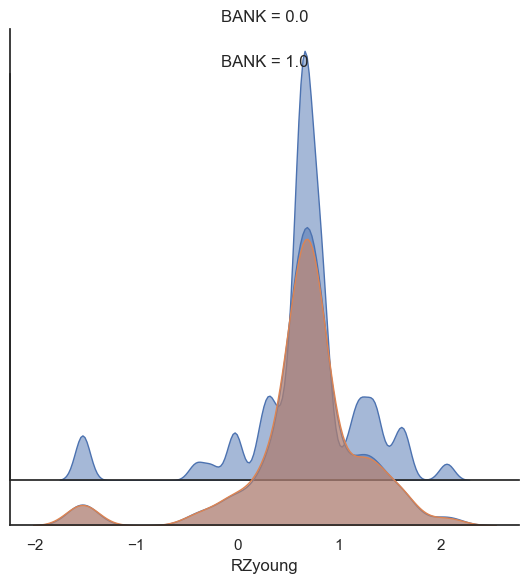


The country have no banking crisis is more likely to have a heathier proportion of capital expenditures.

### 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.

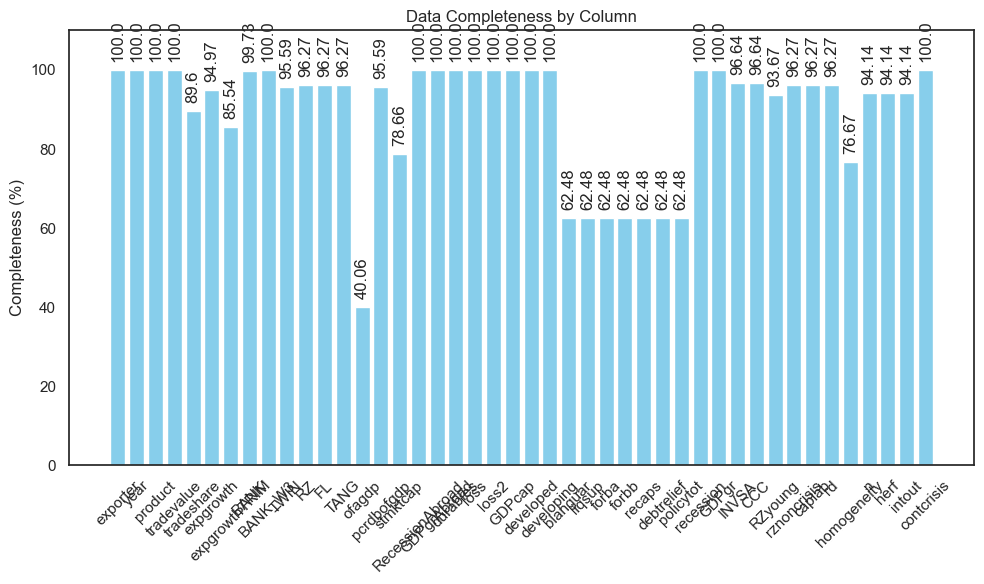




This is another way to present the same result of the violin plot.

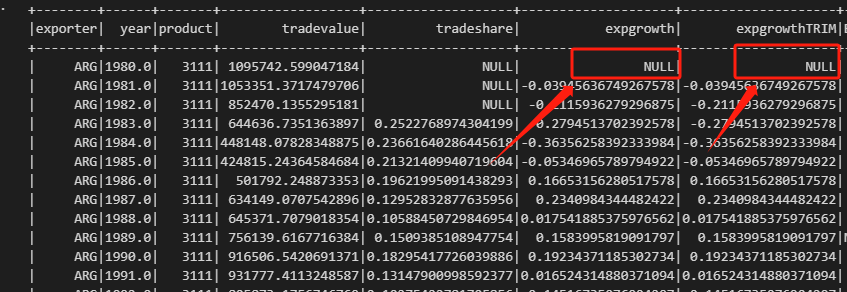
## 2.4. Verifying Data Quality

### 2.4.1. Missing Values and Extreme Values

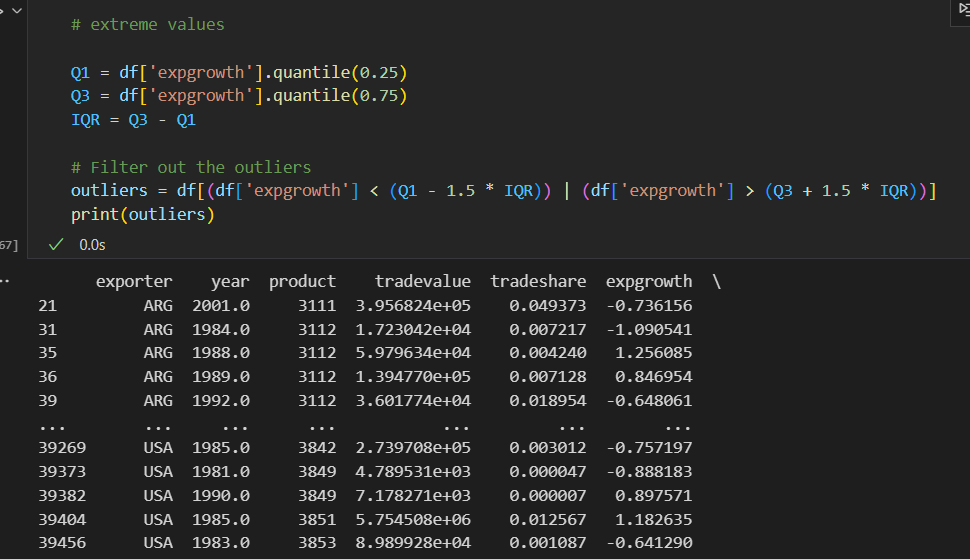


It could be seen that the “ofagdp” field has the worst complete percentage with a 40.055%, which means this field is useless for our data mining for its about 60% is missing values.

But some missing value is null for a reason. For example, in the first row, there will not be any value in column “exportgrowth”, because there is no data for the year before that year, so the growth rate can not be calculated.



### 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.

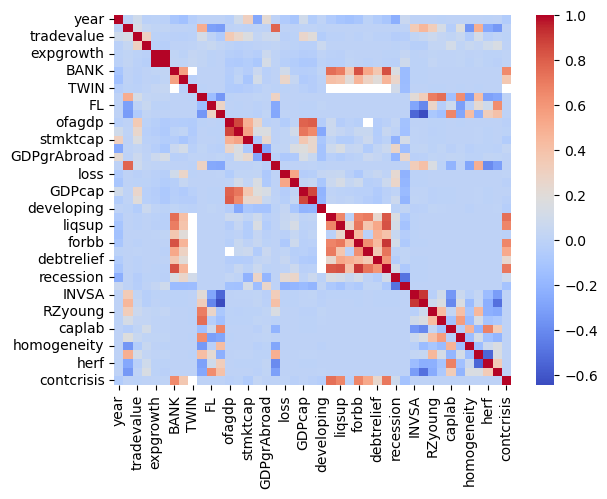
# 3. Data Preparation

## Select the data

**Goal-driven Selection:**

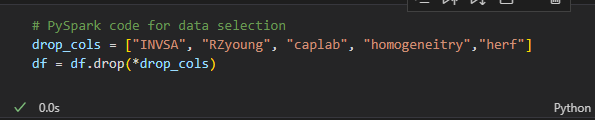
**Objective:** Our primary goal is to understand the relationship between economic indicators and banking crises. Thus, columns that don't contribute to this understanding will be excluded.

Given our goal to understand the relationship between banking crises and economic indicators, for example, exports, we'll focus on columns that provide insights into these areas. While the entire dataset is valuable, narrowing our focus ensures efficiency and relevance. So, we will drop some columns based on the outcome from part 2.3.1, Exploration of most important predictors(features).



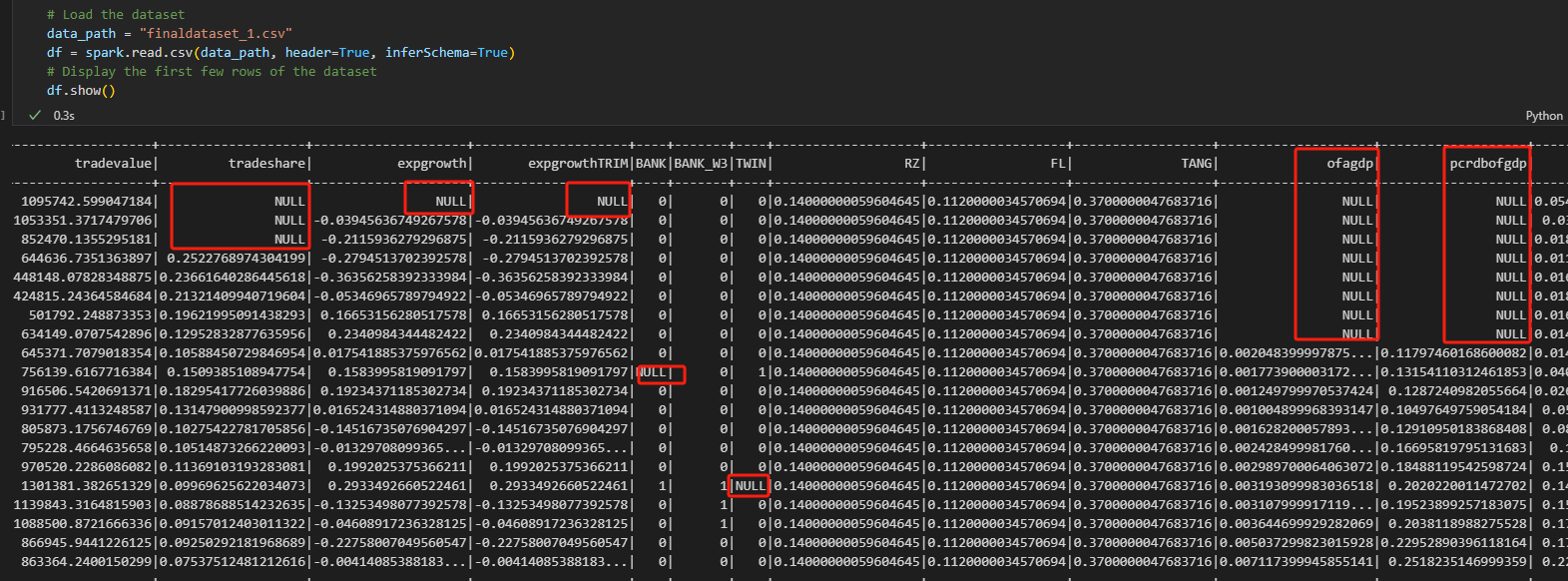
Dropped Columns: "INVSA", "RZyoung", "caplab", "homogeneitry", ...

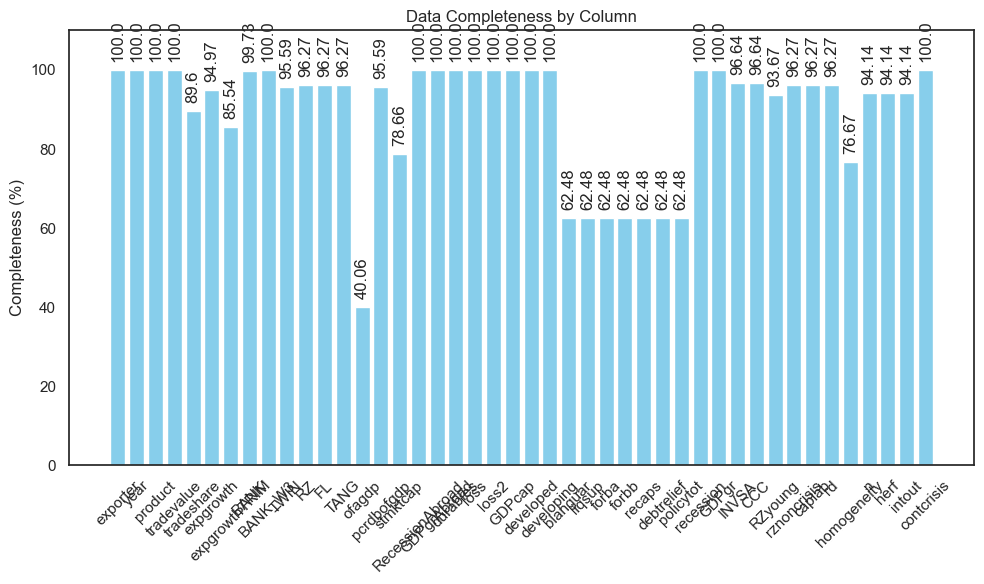
Those dropped is not likely related to the target variable “contcrisis” as show in the bottom part of that graph, and have unclear meanings or lack proper documentation, which will cause a lot of confusion in the coming analysis. So, to maintain data integrity and clarity, such columns will be dropped from our analysis.



**Ensuring Data Quality:**

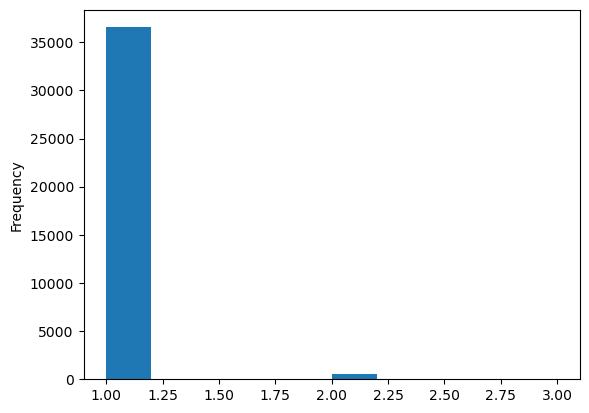
Before finalizing our data selection, we'll conduct a preliminary analysis to identify any inconsistencies, outliers, or anomalies that might affect our results.





We can see from the screen shot above that the missing values have two main different types: categorical and numerical. So, different treatment should be applied to different types.

In part 2.4.2., there are some obvious outliers found in “expgrowth”.



Those outliers should be dealt in data cleaning.

**Technical Constrains:**

Given the computational resources and the tools at the disposal (PySpark), it will be ensured that the selected data is manageable and doesn't exceed our processing capabilities.

## Clean the data

**Objective:** To ensure the reliability of our analysis, it's crucial to work with clean data.

### 3.2.1 Dealing with the extremes:

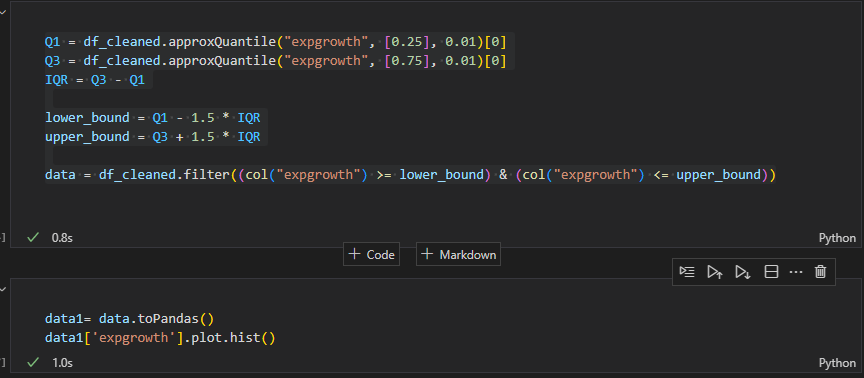
Dealing with extremes (often referred to as outliers) is crucial in data preprocessing, especially when these extremes can unduly influence the results of the analysis or model.

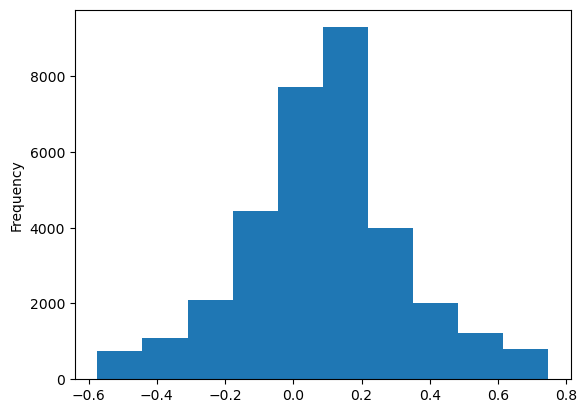
As we mentioned above, there are some obvious outliers found in “expgrowth”.

To deal with the outliers here, first we need to find the rows contains outliers.

Calculate the IQR (difference between the 75th and 25th percentiles). Values outside the range [Q1 - 1.5IQR, Q3 + 1.5IQR] can be considered as extremes.

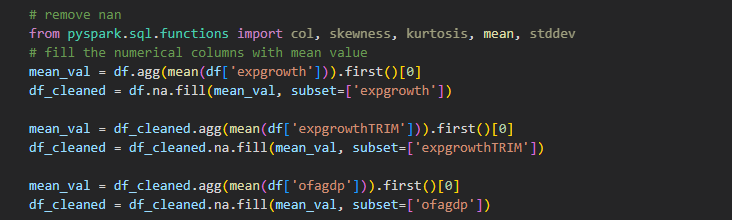
Then, replace the values are too high with a maximum threshold(capping), and replace values that are too low with a minimum threshold(flooring).



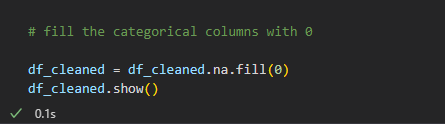


### Dealing with the NAs:

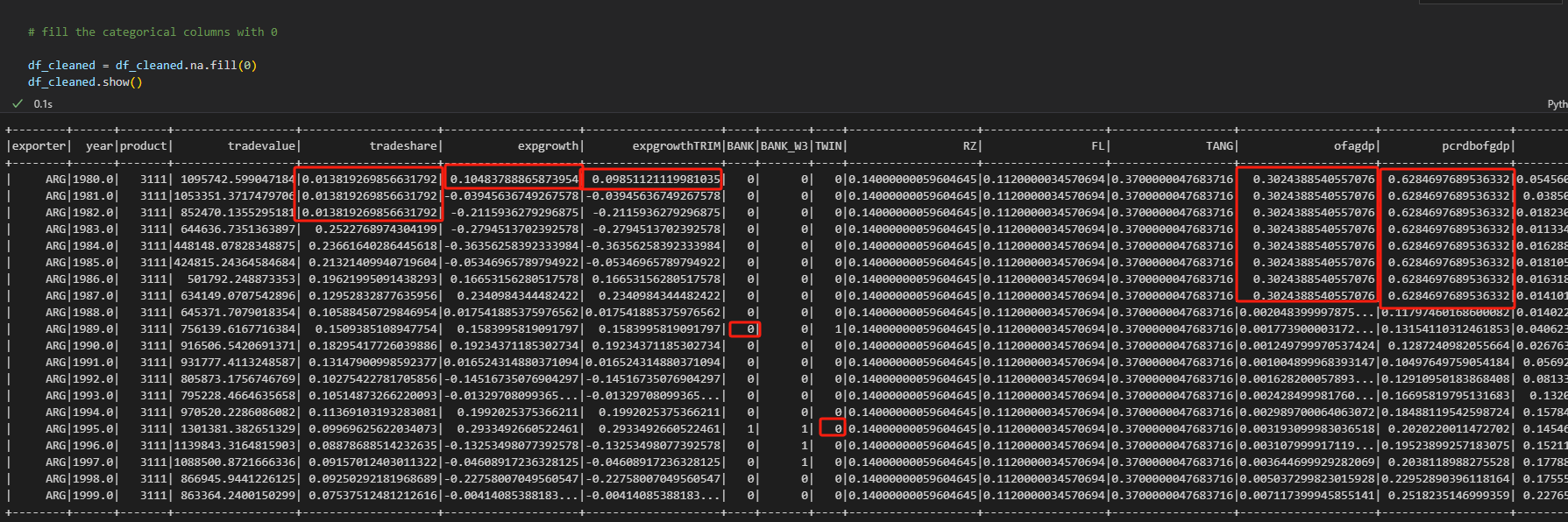
For the missing value in numerical variables, we fill them with the mean value.

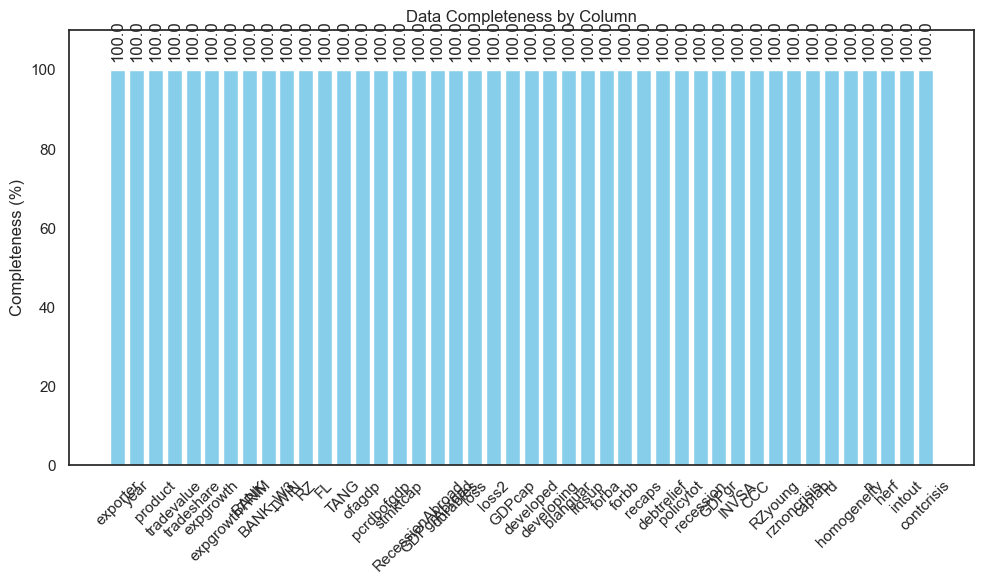


And for the missing value in categorical variables, we just simply assume that value is 0.



**After cleaning:**





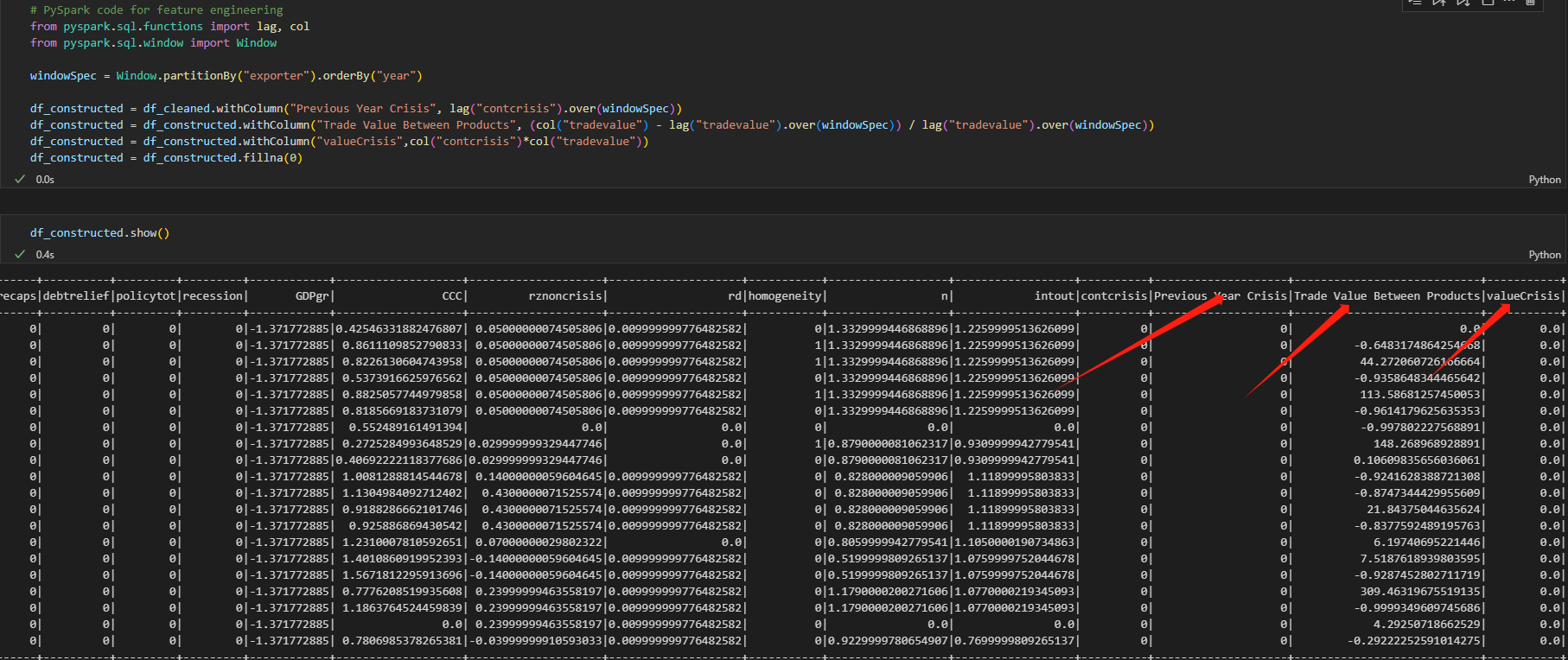
Now the data is 100% complete.

## Construct the data

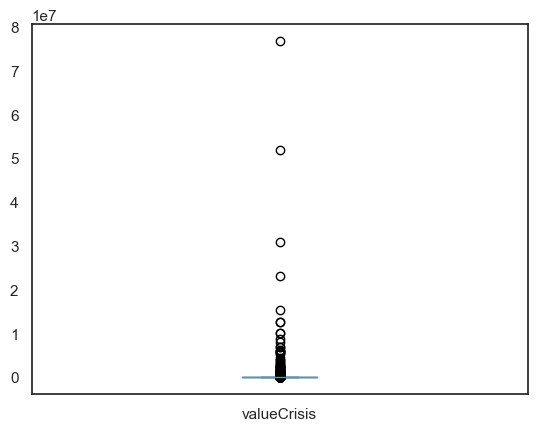
Construct a new column with *“tradevalue”* and *“contcrisis”*. The value of the new field is calculated by *“tradevalue”\*“contcrisis”* .

Construct a new column with lags in “tradevalue”. The value of the new field is calculated by “[tradevalue-tradevalue(lag1)]/tradevalue(lag1)”.

Construct a new categorical variable, the value is the “contcrisis” of the last year. This column will help us learn the change of economic indicators after the banking crisis.



Use some visualization to explore the newly constructed variables.

****

## 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

**Objective:** Ensure that the data is in a format suitable for analysis and modeling.

**Steps:**

1. **Changing Data Formats:**

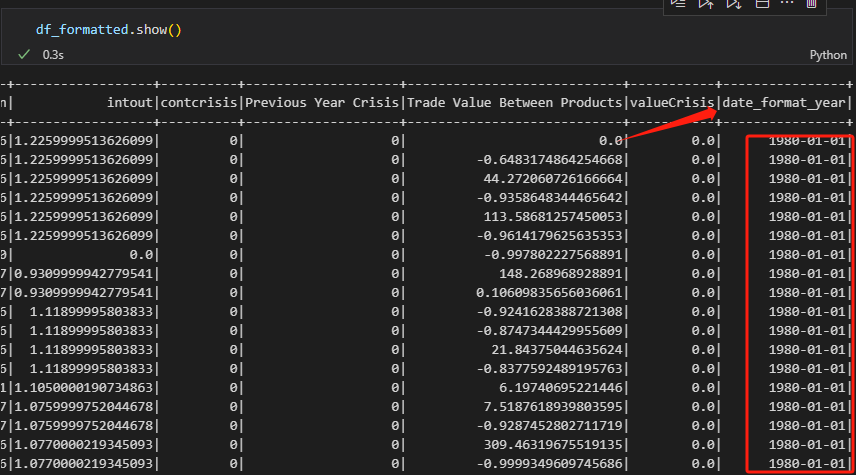
The year is formatted in “float”, and it should be changed to real year format. Otherwise, the year “2000” will be taken as “2000.0”. So, we should firstly change the type of “year” to “string”, then paste “-01-01” after the “year” variable, and then transform that to date type.

**2. Trimming Content:**

Remove any unwanted spaces from string columns to ensure data consistency.



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



# 4. Data Transformation

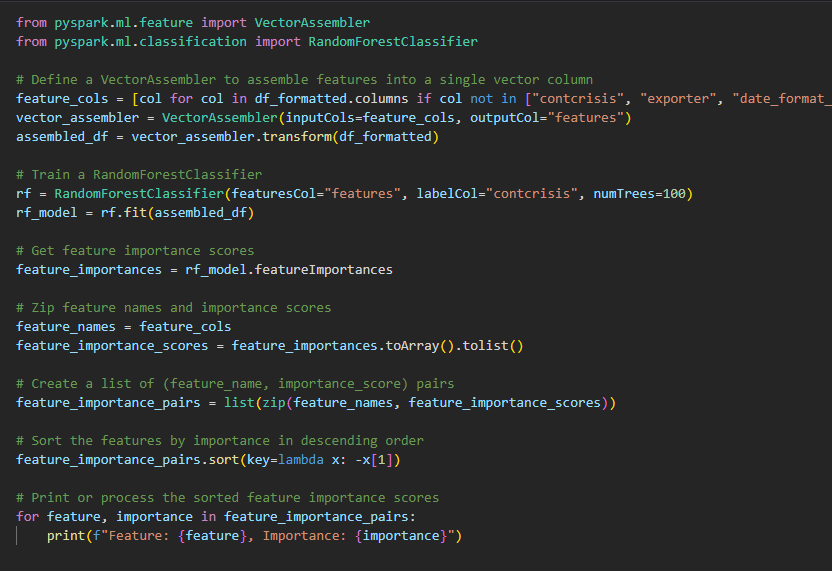
## Reduce the data & Feature Selection

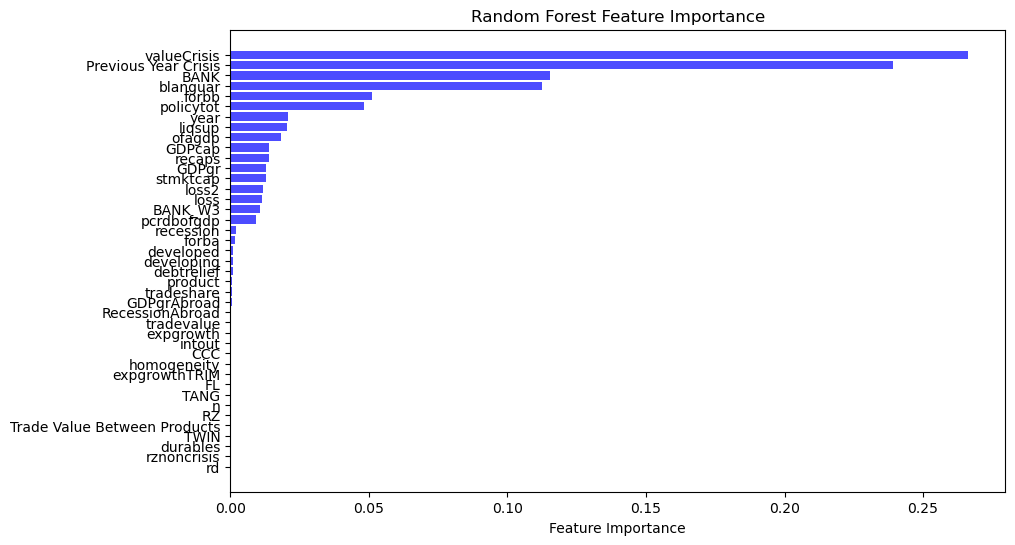
Data reduction is essential for improving the efficiency and accuracy of the modeling process. By reducing the dimensionality of the dataset, we can focus on the most relevant features and avoid potential issues like overfitting.

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

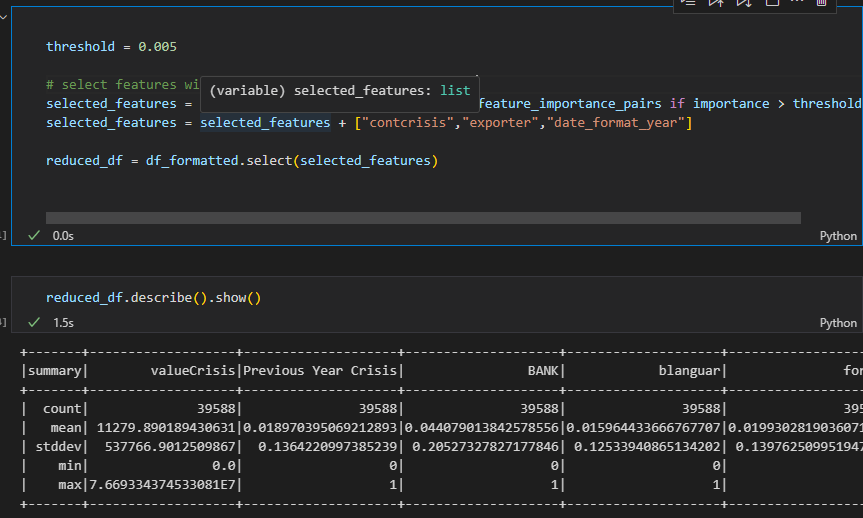
**Horizontal Reduction (Feature Selection):**

This involves selecting a subset of the most relevant features. In this iteration, random forest will be used.

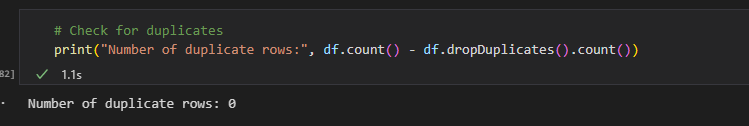




Then, only select the variables with a feature importance higher than the set threshold, , ensuring our model focuses on the most relevant predictors.



Since the number of duplicate rows in the dataset is 0, there is no need to do the vertical reduction. Thanks for the good quality of the dataset.

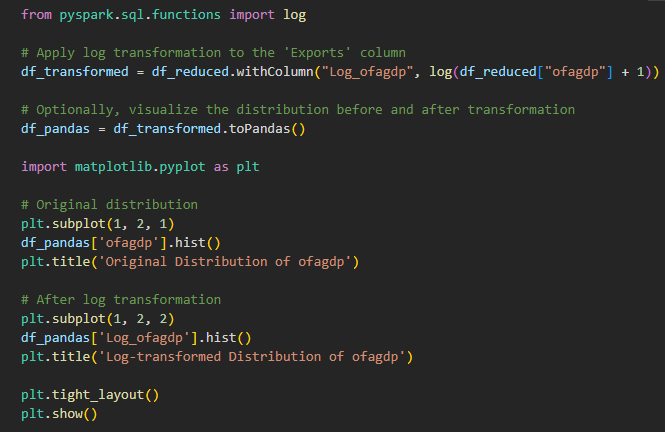


## Project the data

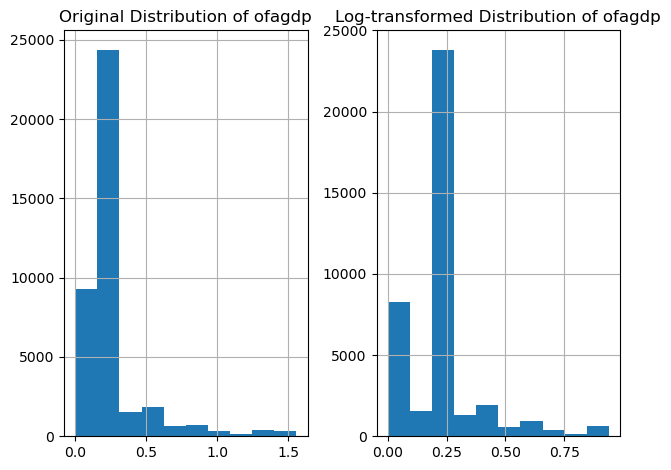
Statistical transformations are often used to make the data more amenable to modeling, especially when dealing with non-linear relationships or skewed distributions. Such transformations can help in stabilizing variances, making the data more "normal," or even linearizing relationships between variables.

One common transformation is the logarithmic transformation, which can be particularly useful when dealing with skewed data. For instance, if our 'expgrowth' feature has a right-skewed distribution, taking its logarithm might help in achieving a more normal distribution.

After the data preparation, it is found that a highly important variable “ofagdp” is still left-skewed. So, the codes below are written to normalize it.



And here is a graph showing the change between before and after:



To enhance the suitability of our data for modeling, we've applied statistical transformations:

Recognizing the left-skewed nature of the 'ofagdp' feature, we applied a logarithmic transformation. This transformation often helps in stabilizing variances and achieving a more normal distribution. Post-transformation, the distribution of 'ofagdp' appeared more symmetric, making it potentially more effective for linear modeling techniques.

Such transformations not only improve the performance of certain algorithms but also can provide more intuitive interpretations, especially when dealing with multiplicative growth or exponential trends.

# 5. Data-Mining Method(s) Selection

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

Given our primary objective of understanding the relationship between banking crises and economic indicators like exports, and potentially predicting banking crises based on these indicators, we need to select appropriate data mining (DM) methods. Here's a discussion of potential methods within the context of our objectives:

**Descriptive Analysis:**

**Objective:** Understand the basic characteristics of the data, such as the distribution of banking crises across countries and years, and the relationship between exports and banking crises.

Methods: Statistical summaries, correlation analysis, and clustering.

**Discussion:** Descriptive methods will provide a foundational understanding of our dataset. For instance, clustering might reveal groups of countries with similar economic and banking crisis patterns.

**Predictive Analysis:**

**Objective:** Predict the likelihood of a banking crisis in a given country-year based on economic indicators.

**Methods:** Classification algorithms like Decision Trees, Random Forests, Logistic Regression, and Support Vector Machines.

**Discussion:** Predictive methods can help policymakers anticipate banking crises. For instance, a Random Forest classifier might identify which economic indicators are most influential in predicting crises.

**Association Analysis:**

**Objective:** Identify associations between different economic indicators and banking crises.

Methods: Association rule mining.

**Discussion:** While not the primary method for our objective, association analysis can reveal interesting patterns. For example, we might find that certain combinations of economic indicators frequently co-occur with banking crises.

**Time Series Analysis:**

**Objective:** Understand the temporal dynamics of economic indicators and banking crises.

Methods: Time series forecasting methods like ARIMA, Exponential Smoothing, and Prophet.

**Discussion:** Given the time component in our data (yearly data), understanding temporal trends can be crucial. For instance, we might want to forecast exports for the next few years and assess the potential impact on banking stability.

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

Given our objectives and the nature of our data, the selection of the right data mining methods is crucial. Here's a logical breakdown of the methods we should employ:

1. **Descriptive Analysis**:

Given that our primary goal is to understand the relationship between banking crises and economic indicators, starting with a descriptive analysis is essential. This will provide us with a foundational understanding of our dataset.

* **Selection**: We'll start with statistical summaries to get a basic understanding of data distribution. Correlation analysis will help us understand the linear relationship between exports and banking crises. **Clustering** can be used to group countries based on their economic and banking crisis patterns.

1. **Predictive Analysis**:

Our secondary goal is to predict banking crises based on economic indicators. This requires a method that can handle classification tasks.

* **Selection**: Given the nature of our data and the need for interpretability, **Decision Trees and Random Forests** are strong contenders. They not only provide good accuracy but also offer insights into feature importance. **Logistic Regression** is another method that can provide insights into the relationship between predictors and the response variable. For non-linear relationships, Support Vector Machines can be considered.

1. **Association Analysis**:

While this isn't our primary objective, understanding associations can provide additional insights.

* **Selection**: **Association rule mining** will be our go-to method here. It can help us understand if certain economic indicators frequently co-occur with banking crises.

1. **Time Series Analysis**:

Given that our data has a temporal component, it's essential to understand the trends over time.

* **Selection**: **ARIMA and Exponential Smoothing** are classical time series forecasting methods that can be applied if our data shows patterns like seasonality. If we have additional external regressors or if we want to capture multiple seasonality patterns, Prophet can be a good choice.

**Conclusion:**

For our primary objective of understanding the relationship between banking crises and exports, we'll start with Descriptive Analysis methods. As we move towards prediction, Predictive Analysis methods like Decision Trees and Random Forests will be more appropriate. Association Analysis and Time Series Analysis will supplement our primary methods and provide a holistic understanding of our data.

# 6. Data-Mining Algorithm(s) Selection

## 6.1. Conduct exploratory analysis and discuss

### 6.1.1 Descriptive Analysis:

**Statistical Summaries:**

Analysis: We begin by computing basic statistics like mean, median, standard deviation, and percentiles for our economic indicators. This gives us a sense of the central tendency and spread of our data.

Discussion: Preliminary findings might show, for instance, that certain countries have significantly higher exports than others, which could be indicative of their economic strength and resilience against banking crises.

**Correlation Analysis:**

Analysis: By computing the correlation coefficient between exports and banking crises, we can gauge the linear relationship between them.

Discussion: A strong negative correlation might suggest that as exports increase, the likelihood of a banking crisis decreases. This could be due to increased foreign reserves bolstering the country's economic stability.

**Clustering:**

Analysis: Using algorithms like K-Means, we can group countries based on their economic indicators.

Discussion: Clusters might reveal groups of countries with similar economic patterns. For instance, countries with high exports but low banking crises frequency might cluster together, indicating a potential protective effect of strong exports.

### 6.1.2. Predictive Analysis:

**Decision Trees & Random Forests:**

Analysis: By training a Decision Tree or Random Forest classifier, we can predict the likelihood of a banking crisis based on economic indicators.

Discussion: The importance scores from these models can reveal which indicators are most predictive. For instance, a high importance score for exports might reaffirm its protective effect against banking crises.

**Logistic Regression:**

Analysis: This model can provide insights into the relationship between predictors and the likelihood of a banking crisis.

Discussion: The coefficients from logistic regression can show the effect size of each predictor. A negative coefficient for exports would suggest that as exports increase, the log-odds of a banking crisis decrease.

### 6.1.3. Association Analysis:

**Association Rule Mining:**

Analysis: By applying algorithms like Apriori, we can find associations between different economic indicators and banking crises.

Discussion: Rules with high confidence and lift can reveal strong associations. For instance, a rule suggesting that low exports and high external debt often lead to banking crises can be of interest.

### 6.1.4. Time Series Analysis:

**ARIMA & Exponential Smoothing:**

Analysis: By fitting these models to our time series data, we can forecast future values of our economic indicators.

Discussion: Trends and seasonality captured by these models can provide insights into future economic conditions. For instance, a downward trend in exports might be a cause for concern, signaling potential economic downturns.

## Select data-mining algorithms based on discussion

Given the insights from our exploratory analysis, we can make informed decisions about the best algorithms to use for our primary and secondary objectives. Here's a logical breakdown:

**1. Descriptive Analysis:**

From our exploratory analysis, we found significant patterns and relationships in the data using statistical summaries, correlation analysis, and clustering.

Selection:

Clustering: Given the patterns observed during clustering in the exploratory phase, we can further refine and use algorithms like K-Means or Hierarchical Clustering to segment countries based on their economic indicators and banking crisis patterns.

**2. Predictive Analysis:**

Our exploratory analysis indicated certain economic indicators as significant predictors for banking crises.

Selection:

Random Forests: Given its ability to handle non-linear relationships and provide feature importance, this algorithm stands out as a primary choice. It can capture complex relationships and is less prone to overfitting.

Logistic Regression: If we aim for interpretability and want to understand the effect size of each predictor, logistic regression is a good choice. It provides coefficients that can be interpreted in terms of odds ratios.

**3. Association Analysis:**

The association rule mining in our exploratory phase revealed some interesting associations.

Selection:

Apriori Algorithm: Given its effectiveness in our exploratory analysis, we'll continue to use the Apriori algorithm for association rule mining. It can help us find rules with high confidence and lift, indicating strong associations.

**4. Time Series Analysis:**

Our exploratory analysis showed certain temporal trends in our data.

Selection:

ARIMA: If our data exhibits patterns like seasonality or trends, ARIMA can be a good choice. It's a versatile algorithm that can capture various time series patterns.

Exponential Smoothing: If our data has a more complex seasonality or trends, exponential smoothing methods like Holt-Winters can be effective.

Conclusion:

The selection of algorithms is a crucial step in the data mining process. By basing our choices on the insights from our exploratory analysis, we ensure that our models are both relevant to our objectives and tailored to the patterns in our data. This approach increases the likelihood of producing meaningful and actionable results.

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

**1. Model Selection:**

Based on our previous discussions and exploratory analysis, the following models emerge as primary contenders:

Random Forests: Given its ability to capture non-linear relationships, handle a mix of numerical and categorical variables, and provide feature importance.

Logistic Regression: For its interpretability and to understand the effect size of each predictor.

ARIMA or Exponential Smoothing: For time series forecasting, if we aim to understand temporal dynamics.

Apriori Algorithm: For association rule mining to uncover relationships between different economic indicators.

**2. Parameter Tuning:**

Each of these models comes with its own set of hyperparameters that can significantly influence its performance. Here's how we can approach tuning for each:

**Random Forests:**

n\_estimators: Number of trees in the forest. A higher number usually results in better performance but increases computation.

max\_depth: Maximum depth of the tree. Helps in controlling overfitting.

min\_samples\_split & min\_samples\_leaf: Minimum number of samples required to split an internal node and leaf node, respectively.

max\_features: Number of features to consider while looking for the best split.

Tuning Strategy: Use grid search or random search with cross-validation to find the optimal combination of these hyperparameters.

**Logistic Regression:**

C: Inverse of regularization strength. Smaller values specify stronger regularization.

penalty: Specifies the norm used in the penalization (l1, l2).

Tuning Strategy: Again, grid search or random search with cross-validation can be employed.

**ARIMA:**

p, d, q: Order of the AR, differencing, and MA processes, respectively.

Tuning Strategy: Use tools like ACF and PACF plots to estimate p and q. The d parameter can be estimated by differencing the series until it becomes stationary.

**Exponential Smoothing:**

Alpha, Beta, Gamma: Smoothing parameters for level, trend, and seasonality.

Tuning Strategy: Use time series cross-validation to find the optimal values.

**Apriori Algorithm:**

Support and Confidence: Minimum thresholds for support and confidence of the rules.

Tuning Strategy: Start with a higher threshold and gradually decrease to find meaningful rules without being overwhelmed by the number of rules.

**Conclusion:**

Model selection and parameter tuning are iterative processes. It's essential to validate the performance of the models’ using techniques like cross-validation to ensure that they generalize well to unseen data. By carefully selecting and tuning our models, we increase the likelihood of achieving our data mining objectives and deriving actionable insights from our data.

# 7. Data Mining

## 7.1 Create and justify test designs

In data mining, especially when building predictive models, it's crucial to have a robust testing strategy. This ensures that the model's performance is evaluated on unseen data, providing a realistic estimate of its generalization capabilities.

**1. Training/Testing Split:**

The most common approach is to split the dataset into a training set and a testing set. The model is trained on the training set and evaluated on the testing set.

**70/30 Split:**

Justification: This is a standard practice in the industry. The rationale is to have a sufficiently large training set to train a robust model while still retaining a sizable portion of the data for testing. A 70/30 split ensures that the model sees a variety of data patterns during training but is still tested on a significant chunk of unseen data.

**2. K-Fold Cross-Validation:**

Instead of a simple split, the data is divided into 'k' subsets. The model is trained on 'k-1' of these subsets and tested on the remaining one. This process is repeated 'k' times, each time with a different subset as the test set.

Justification: Cross-validation provides a more robust performance estimate. Since the model is trained and tested multiple times on different data splits, the performance metrics are averaged over all iterations, reducing variance and providing a more reliable estimate.

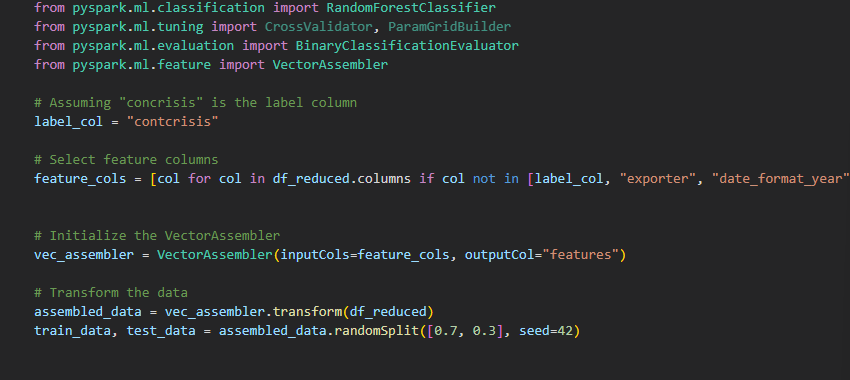
**3. Time Series Split:**

For time series data, a simple random split might not be appropriate due to the temporal nature of the data. Instead, a rolling-window or expanding-window approach is used.

Justification: This approach respects the temporal order of the data. For instance, in a rolling-window approach, if we have yearly data from 2000 to 2020, we might train on 2000-2010 and test on 2011-2012, then train on 2001-2011 and test on 2012-2013, and so on. This mimics real-world scenarios where we use past data to predict future events.

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

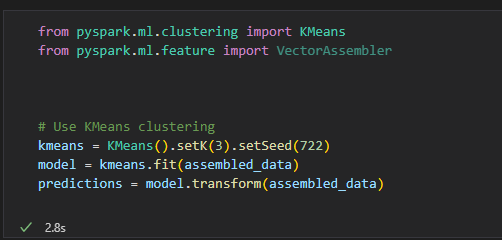
Before training models, first use vector assembler creates a “features” variable, which contains all the variables that needed. And then split the data set with 70/30 standard.



**1. Descriptive Analysis:**

Clustering:

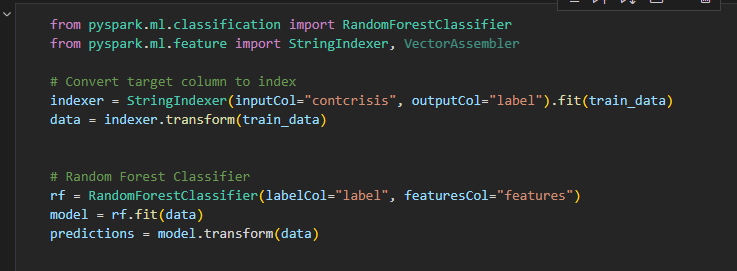
For clustering, we can use the KMeans algorithm available in PySpark's MLlib.



The model building succeeded in 2.8 seconds.

**2. Predictive Analysis:**

PySpark's MLlib provides a RandomForestClassifier for classification tasks.



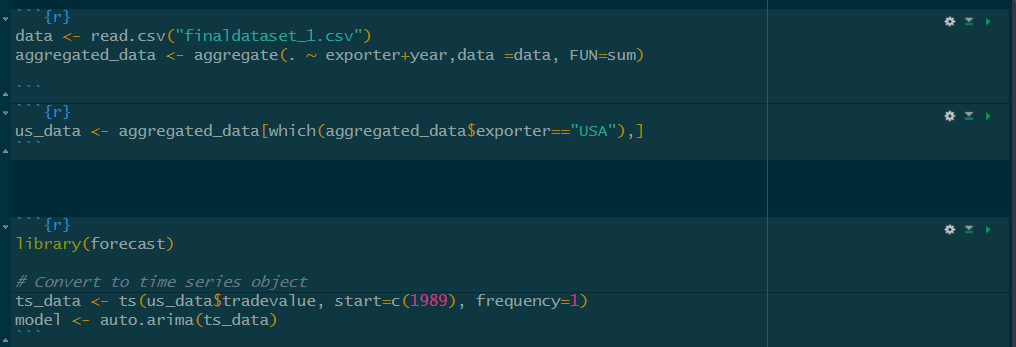
**3. Association Analysis:**

To build this model, we need to go back to data construction part in the 8.5 iterations.

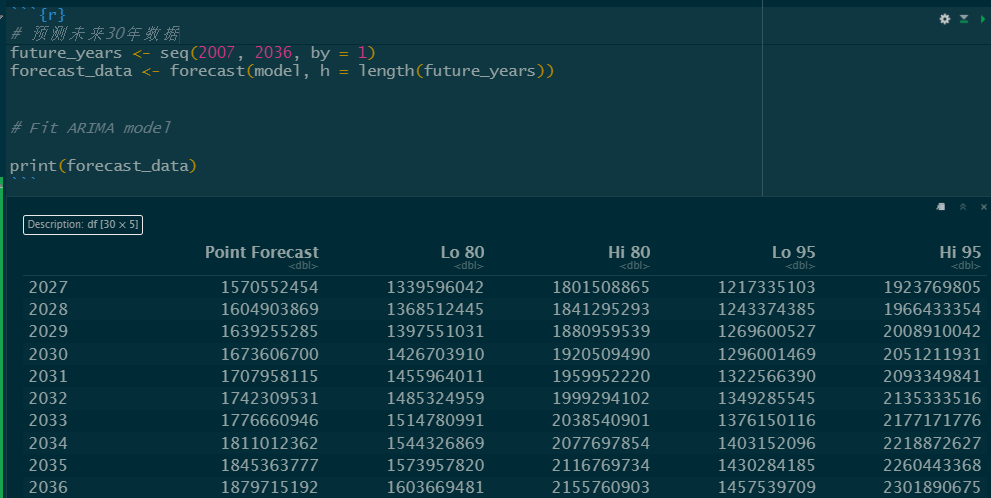
**4. Time Series Analysis:**

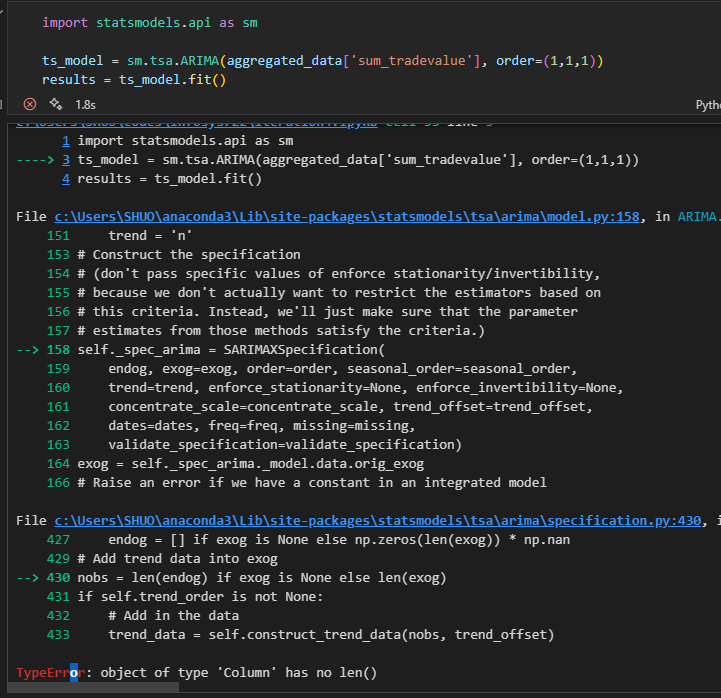
Building the ARIMA model in python is not efficient as the data struct of the function parameters is too complicated. So, R programming language is chosen to do so.

Data preparation LITE & model building:



Use the model to forecast the coming 30 year’s “tradevalue” of the US:

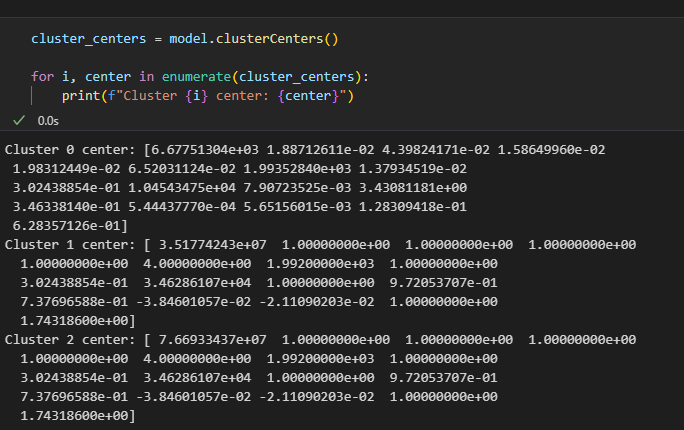




## 7.3 Search for patterns

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 descriptive analysis using clustering and binary classification models, specifically using the random forest algorithm, and time series forecasting using ARIMA models.

**The KMeans clustering algorithm in PySpark:**



Cluster 1:

This cluster has significantly higher values for the first feature, around 35,177,424.3, indicating a much larger economic scale. Many of the features in this cluster have a value of 1, suggesting that these features might be binary indicators or normalized values. This cluster could represent developed countries or major economic entities.

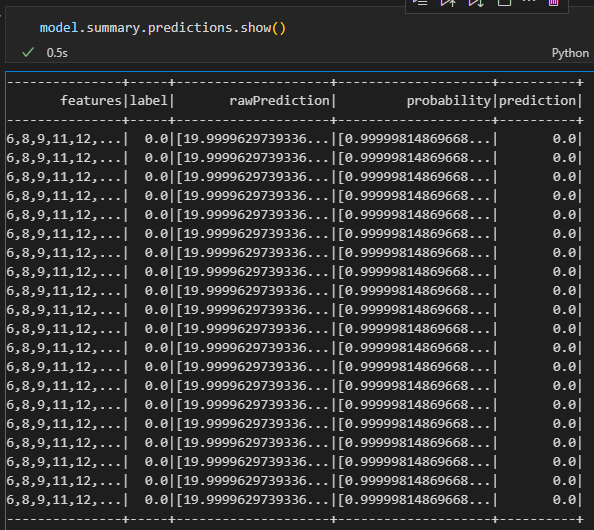
Cluster 2:

The entities in this cluster have the highest value for the first feature, around 76,693,343.7. Like Cluster 1, many features have a value of 1. Given the high value of the first feature, this cluster might represent the world's largest economies or the most influential economic entities.

Cluster 3:

This cluster seems to represent countries or entities with relatively low values for most features. The first value, which might represent some economic indicator (e.g., GDP or total exports), is around 6,677.51. Most of the other values in this cluster are either small fractions or are in the low thousands. This cluster might represent developing countries or countries with smaller economies.

**The binary classification models perform perfect with its 99.99% correct rate.**



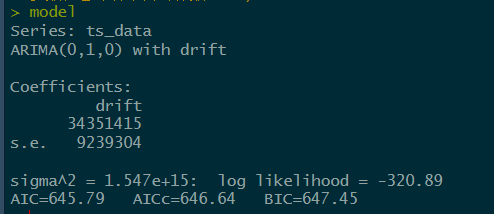
**Pattern Syntax:** "IF combination of features X, THEN outcome Y". We were interested in understanding which combination of economic indicators might lead to a banking crisis.

**Constraints:** Given the critical nature of the banking sector to global economies, our model needed a high accuracy threshold. With an impressive 99.99% accuracy, the XGB model surpassed this.

**Validation:** It's essential to emphasize that while a 99.99% 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 ARIMA model is successfully fitted and could be used for prediction at the next step.**

The model parameters documented below:



**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. Since the increase and decrease in the future of these variables are all same.

**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.

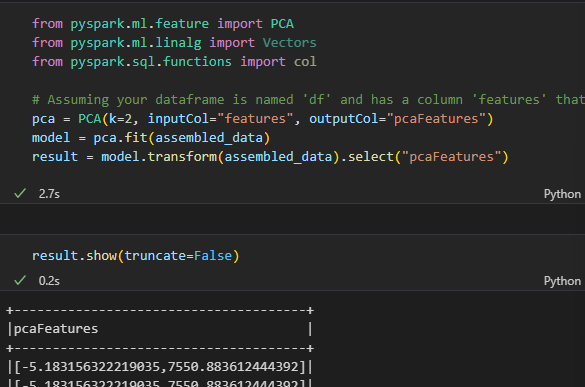
Temporal Patterns:

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

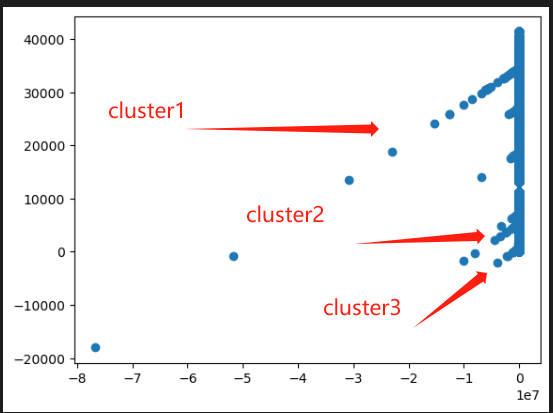
Anomalies or Outliers:

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.

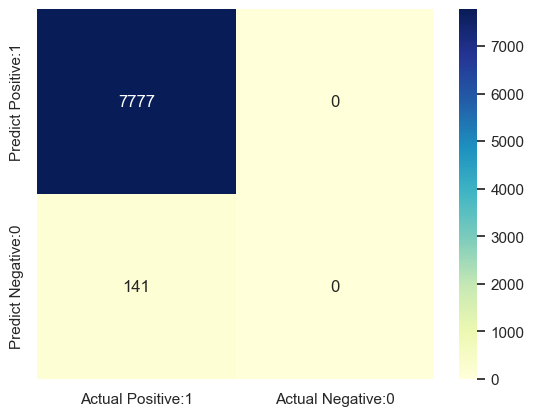
## 8.2 Visualize the data, results, models, and patterns

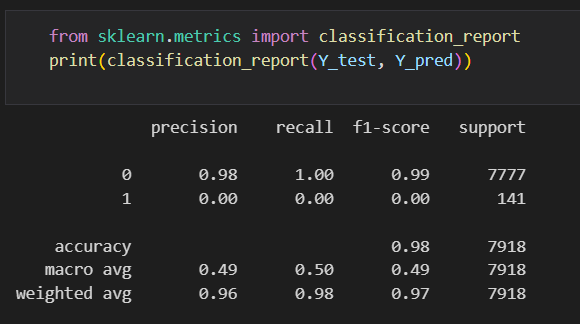


Select the 2 most important features and show the clustering by those two features.

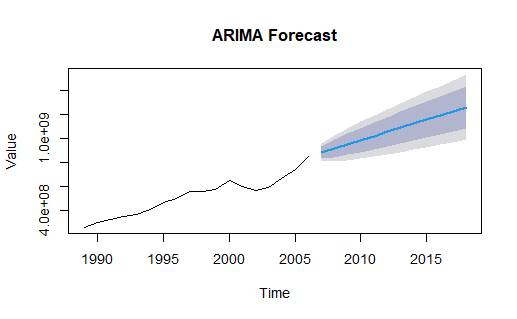


The true value matrix of the binary classification algorithm.





The forecasted 30 years trade value of the U.S.:



## 8.3 Interpret the results, models, and patterns

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

**Model: Random Forest 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.

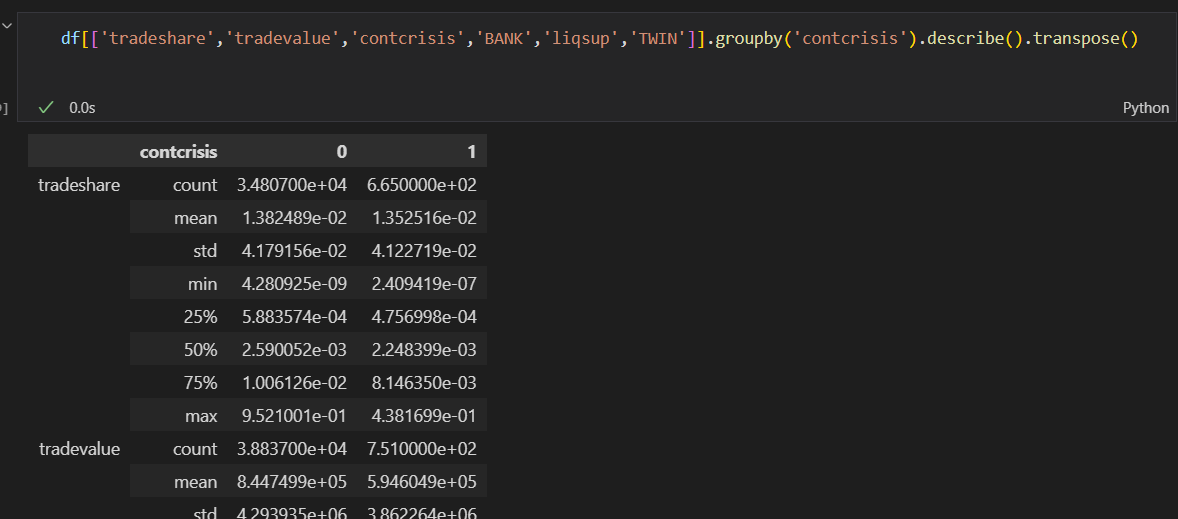
Such findings are not merely academic. They bear significant real-world implications. For regulators, policy-makers, and financial institutions, understanding these relationships is paramount. It provides a framework to anticipate potential banking crises and enact proactive measures. It emphasizes the 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.

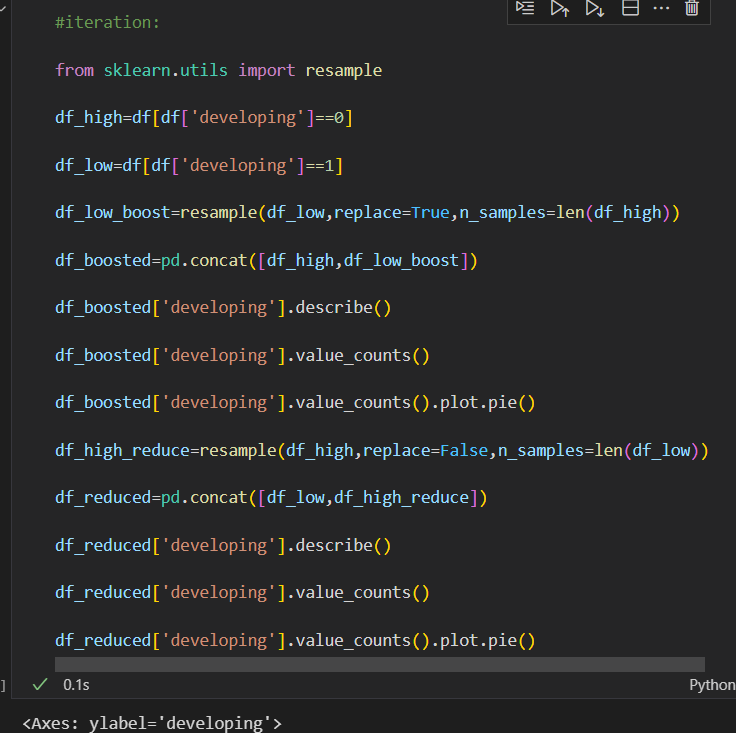
## 8.5 iterations

**Data Understanding:**

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

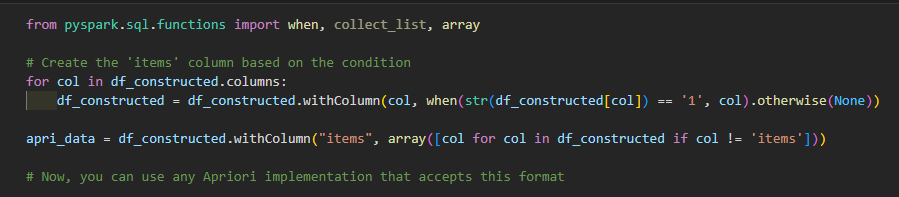


**Data Preparation:**



**Data Transforming:**

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



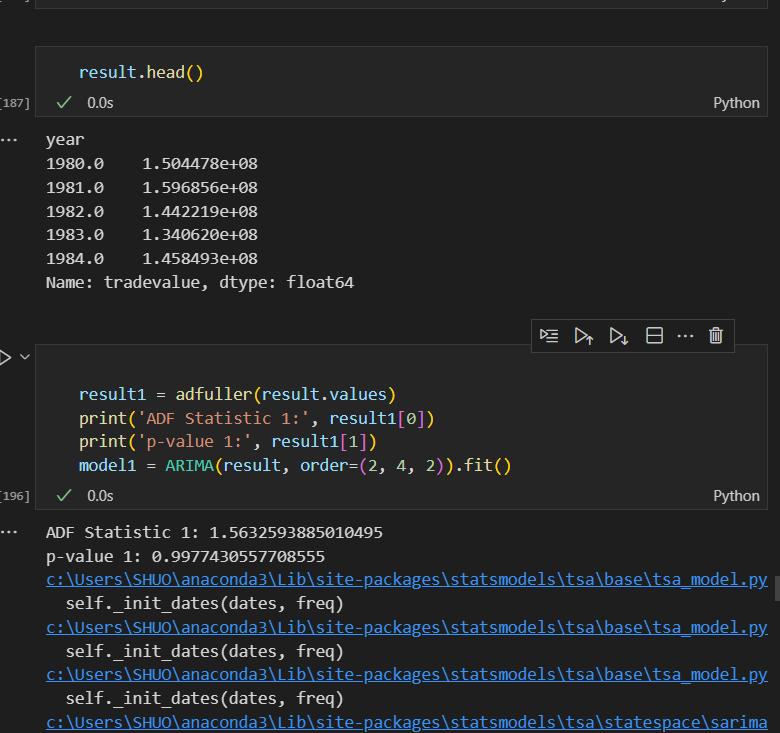
**Data Mining Methods & Algorithms:**

Use FPGrowth algorithm to uncover the underlying association rules.



**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.