ITERATION 2

ISAS

Shuo Feng
UPI: sfen779
ID:836311644

Table of Contents

1. 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.

1.2. Assessment of the Situation

1.2.1. 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, except the SPSS Modeler Premium, it is provided by University of Auckland.

The data used will be accessed from the world bank.

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

1.2.2. 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 assume 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example, there might be missing values, or some records might not have been captured in their entirety, there might be outdated data, especially in dynamically changing environments.

1.2.3. 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.

1.2.4. 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.
2. **Talent Acquisition and Training:**
   Ongoing training and professional development to keep up with evolving technologies and methodologies.
3. **Data Acquisition and Integration:**
   Costs associated with acquiring relevant datasets, possibly from third-party vendors.
   Integrating disparate data sources, ensuring compatibility and consistency.
4. **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.
2. **Tailored Strategies:**
   Data analytics allows for the customization of strategies to specific regions, sectors, or demographics, ensuring more targeted and effective interventions.
3. **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 Reputation damage
Data understanding	10	IT system failure
Data preparation	15	IT system failure
Data transformation	5	IT system failure
Data-mining method selection	10	IT system failure Deadline exceeded Model building failure
Data-mining algorithm selection	15	IT system failure Deadline exceeded Model building failure
Data-mining	15	IT system failure Deadline exceeded
Interpretation	20	IT system failure Deadline exceeded

2. Data Understanding

2.1. Collecting Initial Data

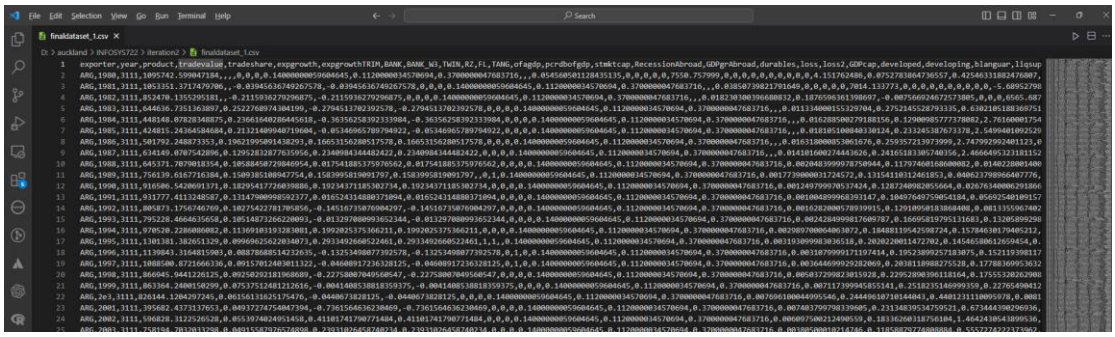
The data would be collected from:
https://datacatalog.worldbank.org/search/dataset/0041188

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

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.

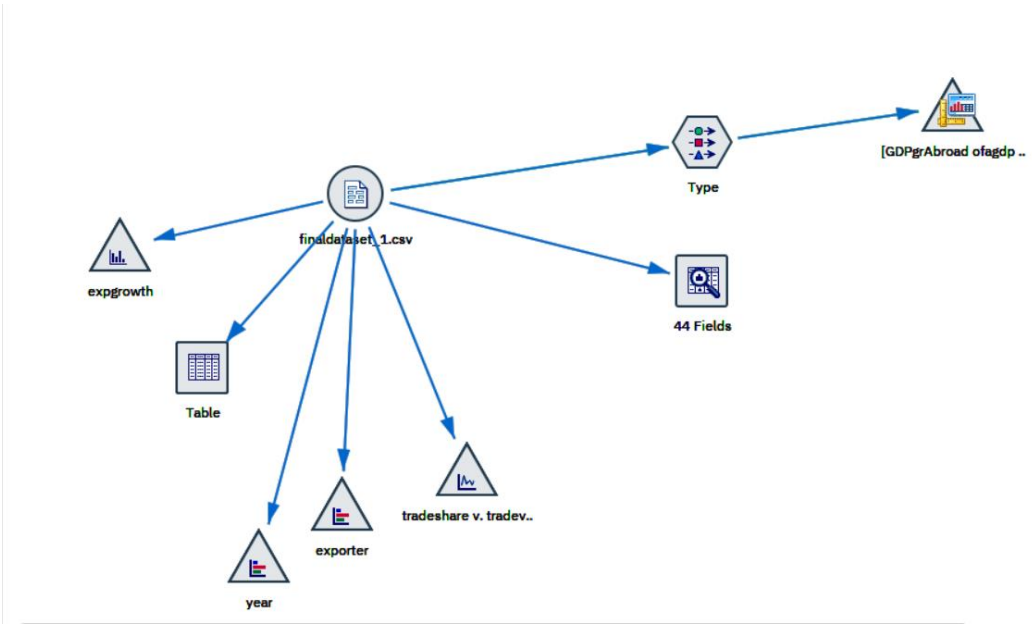
Open it in Visual Studio Code to see its original text looks:

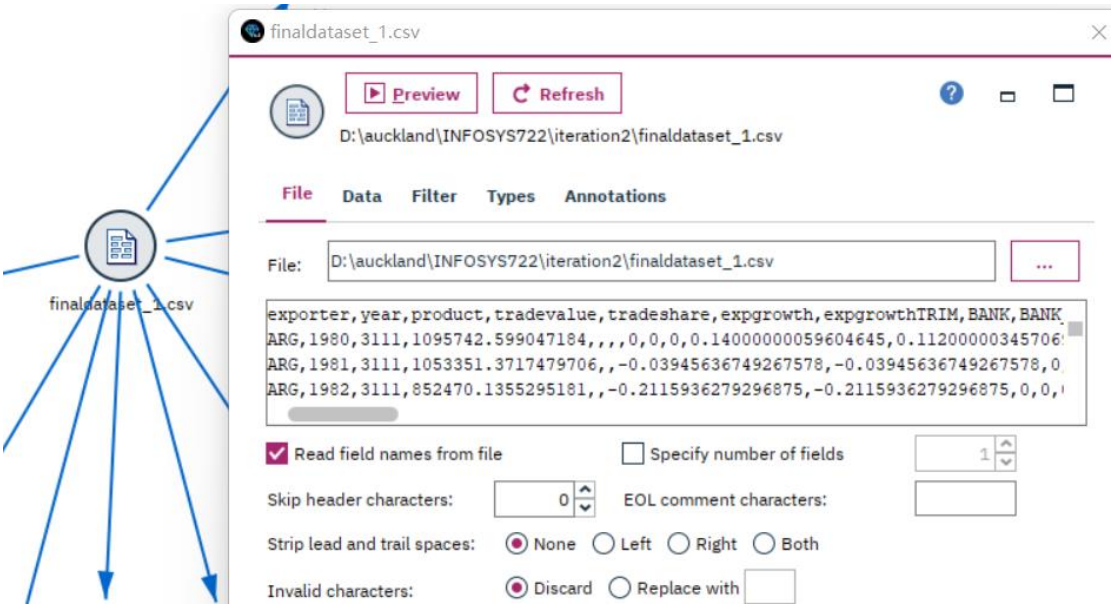


Most of the variables in this data have a numerical type, some flag type and some categorical type variables are also included.

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.

Import it into SPSS Modeler to get a better view:





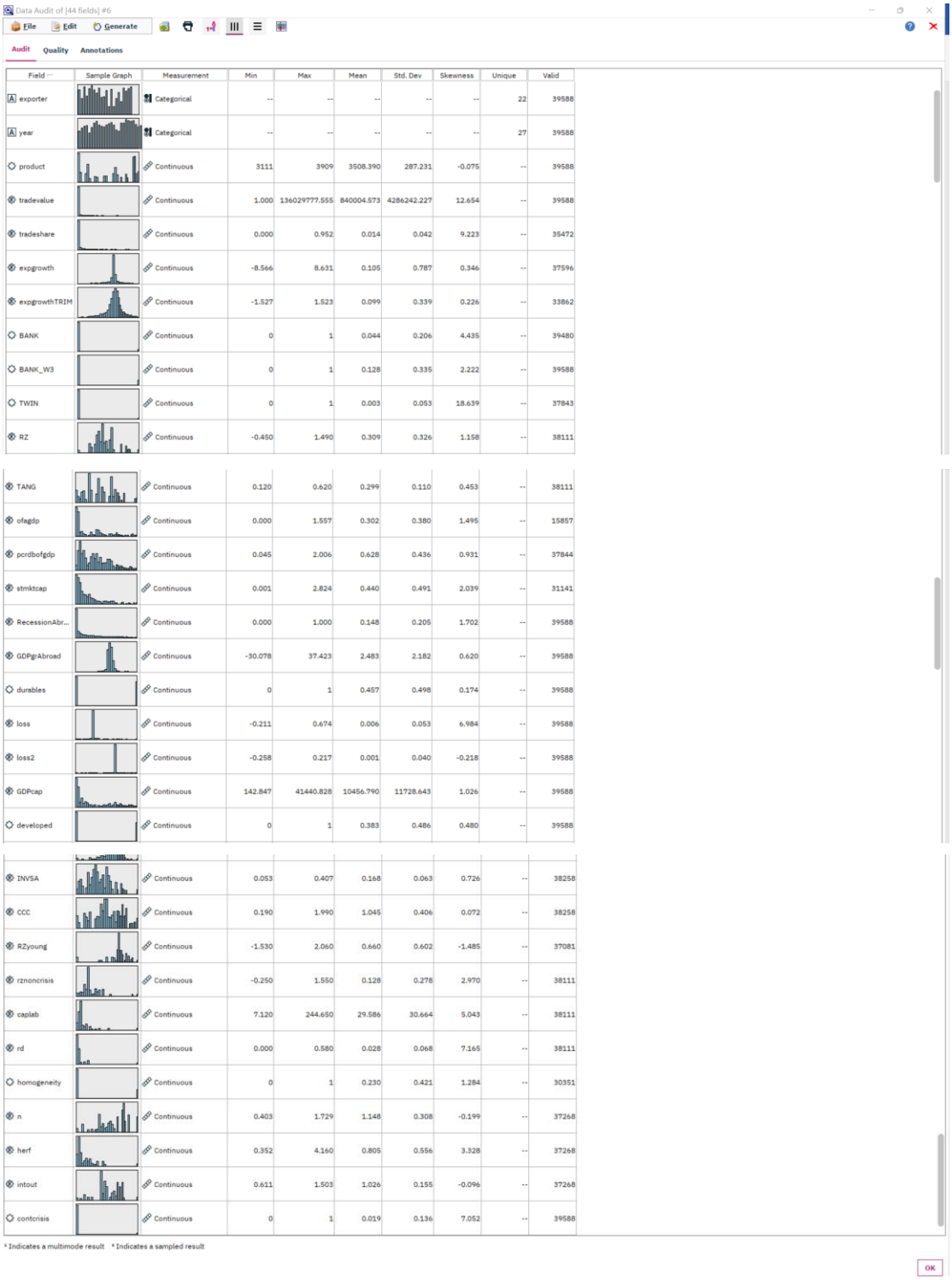
In this iteration, it is not necessary to fully understand the meaning of all headers. First, find the variables most related to “contcrisis” through analysis, and then understand the meaning of these related variables.

A data understanding stream is created in SPSS Modeler.

It could be seen that there are many missing values from the data audit in this stream.

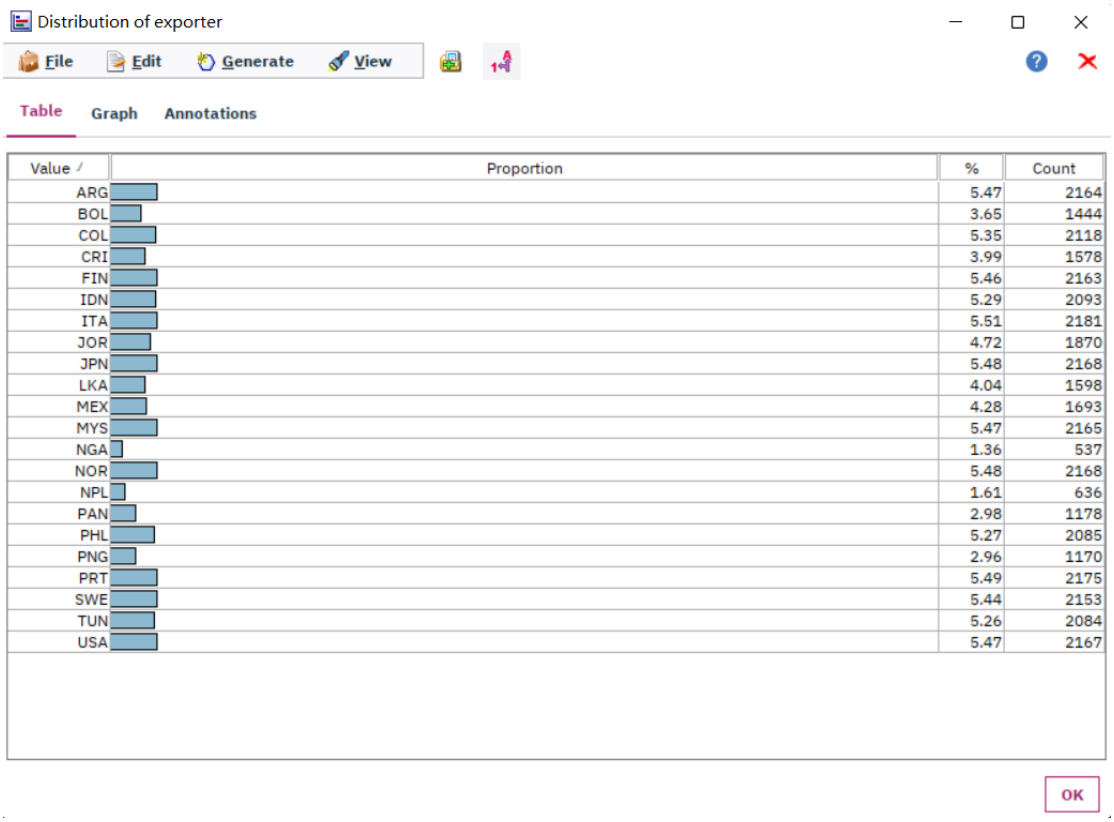
The maximum value, minimum value, mean value, standard deviation, skewness are also clearly listed in the data audit.

ITERATION 2, ISAS

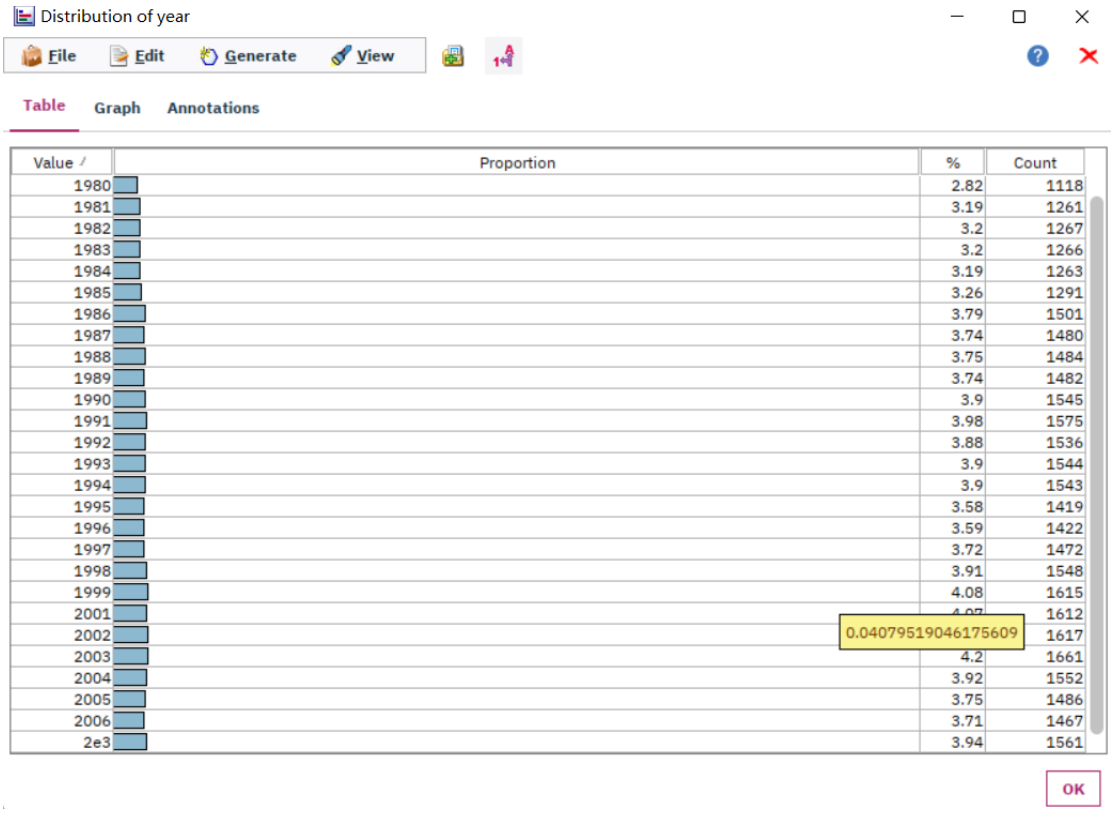


Here are some specific fields in this data:

Some Graphs is created to describe those fields.



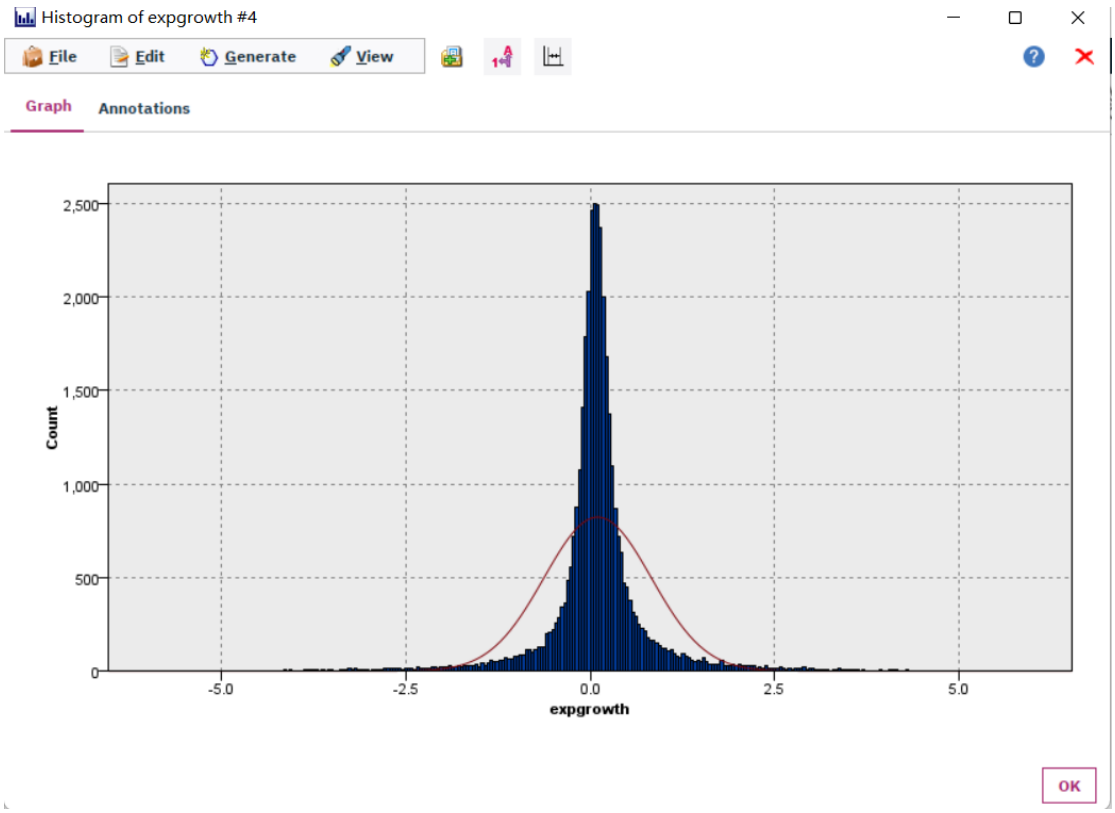
Value	Proportion	%	Count
ARG		5.47	2164
BOL		3.65	1444
COL		5.35	2118
CRI		3.99	1578
FIN		5.46	2163
IDN		5.29	2093
ITA		5.51	2181
JOR		4.72	1870
JPN		5.48	2168
LKA		4.04	1598
MEX		4.28	1693
MYS		5.47	2165
NGA		1.36	537
NOR		5.48	2168
NPL		1.61	636
PAN		2.98	1178
PHL		5.27	2085
PNG		2.96	1170
PRT		5.49	2175
SWE		5.44	2153
TUN		5.26	2084
USA		5.47	2167



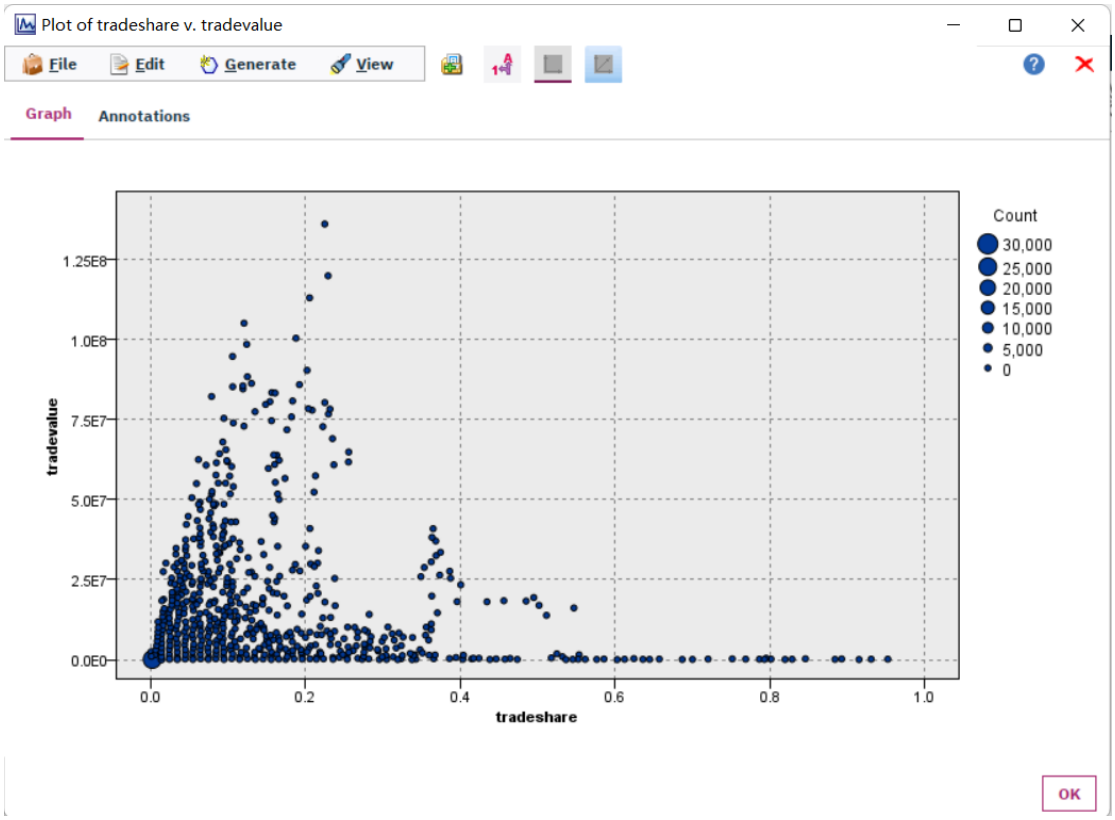
Value	Proportion	%	Count
1980		2.82	1118
1981		3.19	1261
1982		3.2	1267
1983		3.2	1266
1984		3.19	1263
1985		3.26	1291
1986		3.79	1501
1987		3.74	1480
1988		3.75	1484
1989		3.74	1482
1990		3.9	1545
1991		3.98	1575
1992		3.88	1536
1993		3.9	1544
1994		3.9	1543
1995		3.58	1419
1996		3.59	1422
1997		3.72	1472
1998		3.91	1548
1999		4.08	1615
2001		[illegible]	1612
2002		0.04079519046175609	1617
2003		4.2	1661
2004		3.92	1552
2005		3.75	1486
2006		3.71	1467
2e3		3.94	1561

Here, 2e3 should represent the year 2000.

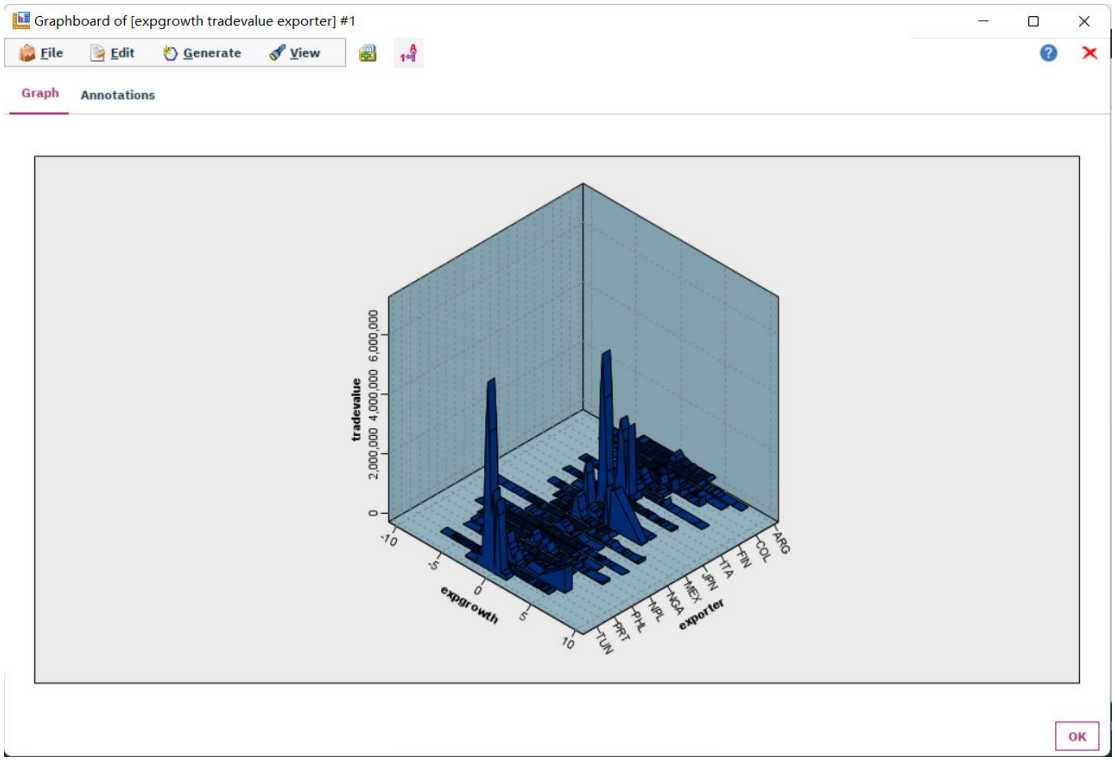
2.3. Data Exploration



The export growth rate has a normal distribution.

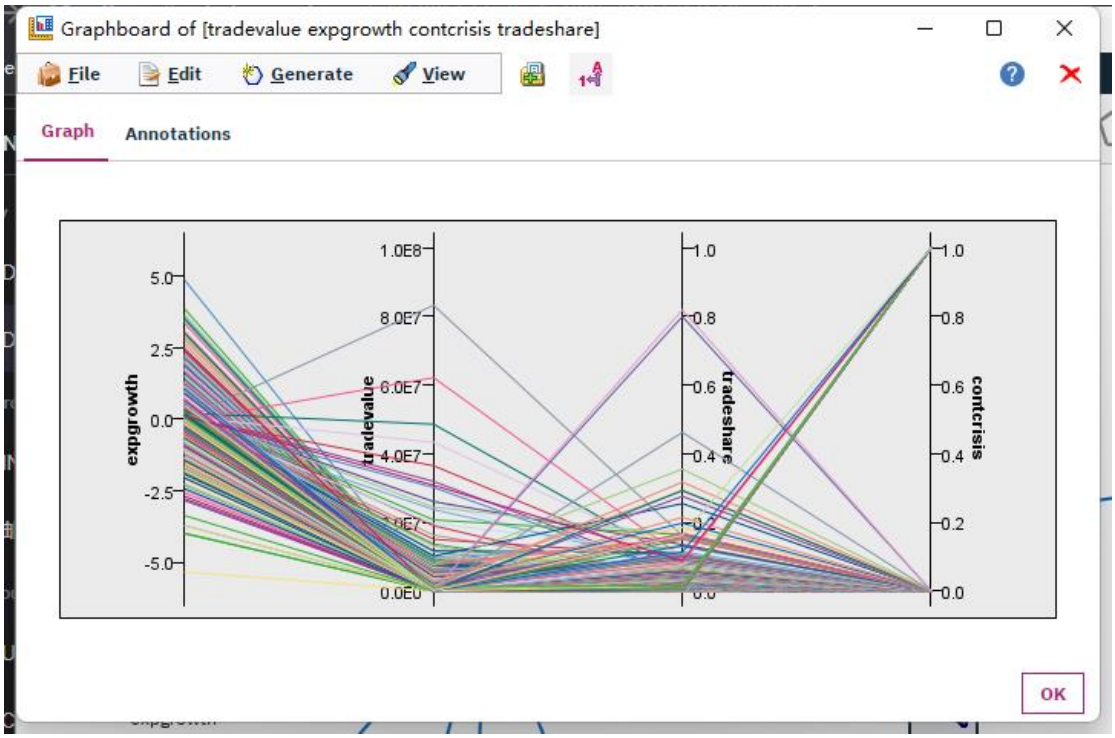


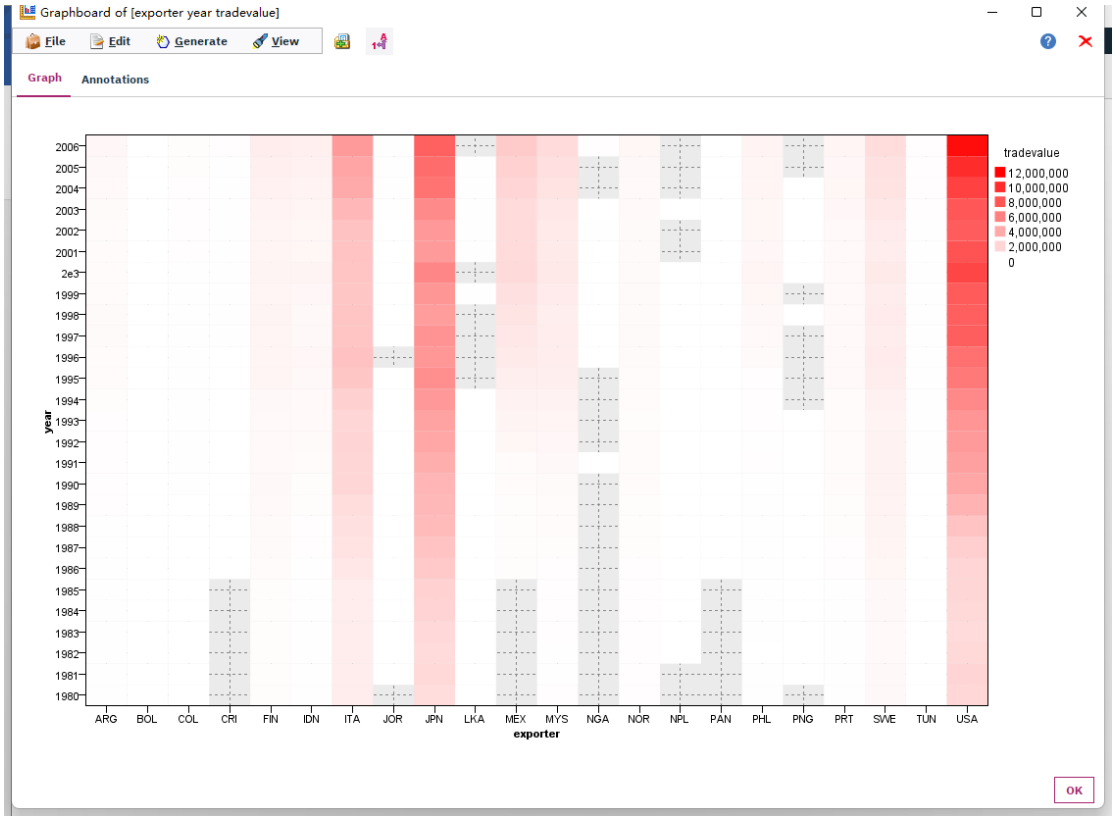
All the points printed are in an angle.



There are 2 obvious peaks in the data of JPN and TUN.

2.3.1. Exploration of most important predictors(features)

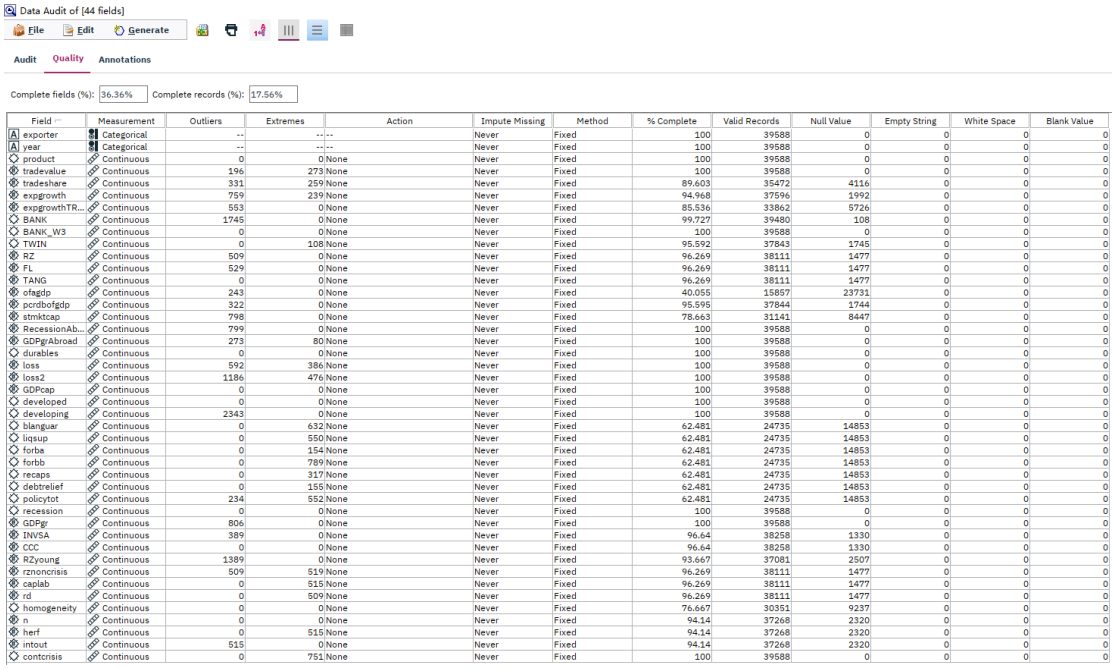




2.4. Verifying Data Quality

2.4.1. Missing Values and Extreme Values

The

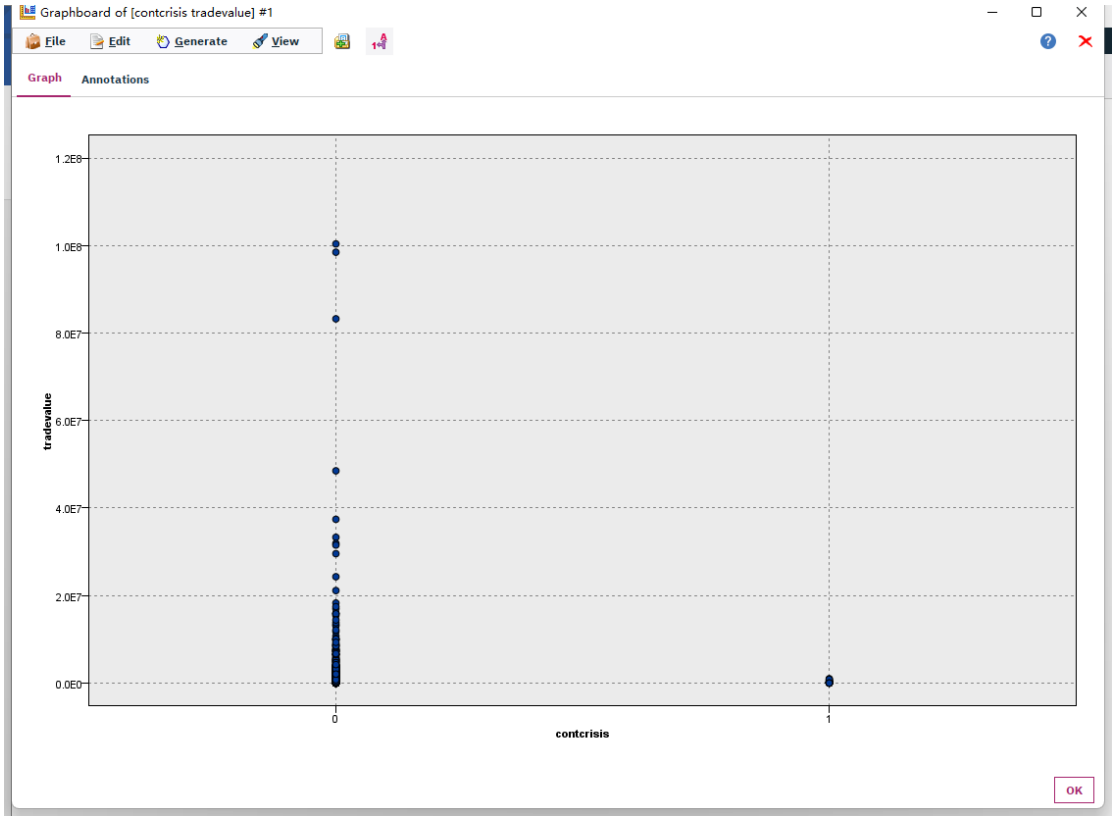


Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
exporter	Categorical	--	--	--	Never	Fixed	100	39588	0	0	0	0
year	Categorical	--	--	--	Never	Fixed	100	39588	0	0	0	0
product	Continuous	0	0	None	Never	Fixed	100	39588	0	0	0	0
tradevalue	Continuous	196	273	None	Never	Fixed	100	39588	0	0	0	0
tradeshare	Continuous	331	259	None	Never	Fixed	89.603	35472	4116	0	0	0
expgrowth	Continuous	759	239	None	Never	Fixed	94.968	37596	1992	0	0	0
expgrowthTR...	Continuous	553	0	None	Never	Fixed	85.536	33862	5726	0	0	0
BANK	Continuous	1745	0	None	Never	Fixed	99.727	39480	108	0	0	0
BANK_W3	Continuous	0	0	None	Never	Fixed	100	39588	0	0	0	0
TWIN	Continuous	0	108	None	Never	Fixed	95.592	37843	1745	0	0	0
RZ	Continuous	509	0	None	Never	Fixed	96.269	38111	1477	0	0	0
FL	Continuous	529	0	None	Never	Fixed	96.269	38111	1477	0	0	0
TANG	Continuous	0	0	None	Never	Fixed	96.269	38111	1477	0	0	0
ofagdp	Continuous	243	0	None	Never	Fixed	40.055	15857	23731	0	0	0
pcrdbofgdp	Continuous	322	0	None	Never	Fixed	95.595	37844	1744	0	0	0
stmktcap	Continuous	798	0	None	Never	Fixed	78.663	31141	8447	0	0	0
RecessionAb...	Continuous	799	0	None	Never	Fixed	100	39588	0	0	0	0
GDPgrAbroad	Continuous	273	80	None	Never	Fixed	100	39588	0	0	0	0
durables	Continuous	0	0	None	Never	Fixed	100	39588	0	0	0	0
loss	Continuous	592	386	None	Never	Fixed	100	39588	0	0	0	0
loss2	Continuous	1186	476	None	Never	Fixed	100	39588	0	0	0	0
GDPcap	Continuous	0	0	None	Never	Fixed	100	39588	0	0	0	0
developed	Continuous	0	0	None	Never	Fixed	100	39588	0	0	0	0
developing	Continuous	2343	0	None	Never	Fixed	100	39588	0	0	0	0
blanguar	Continuous	0	632	None	Never	Fixed	62.481	24735	14853	0	0	0
liqsup	Continuous	0	550	None	Never	Fixed	62.481	24735	14853	0	0	0
forba	Continuous	0	154	None	Never	Fixed	62.481	24735	14853	0	0	0
forbb	Continuous	0	789	None	Never	Fixed	62.481	24735	14853	0	0	0
recaps	Continuous	0	317	None	Never	Fixed	62.481	24735	14853	0	0	0
debtrelief	Continuous	0	155	None	Never	Fixed	62.481	24735	14853	0	0	0
policytot	Continuous	234	552	None	Never	Fixed	62.481	24735	14853	0	0	0
recession	Continuous	0	0	None	Never	Fixed	100	39588	0	0	0	0
GDPgr	Continuous	806	0	None	Never	Fixed	100	39588	0	0	0	0
INVSA	Continuous	389	0	None	Never	Fixed	96.64	38258	1330	0	0	0
CCC	Continuous	0	0	None	Never	Fixed	96.64	38258	1330	0	0	0
RZyoung	Continuous	1389	0	None	Never	Fixed	93.667	37081	2507	0	0	0
rznoncrisis	Continuous	509	519	None	Never	Fixed	96.269	38111	1477	0	0	0
caplab	Continuous	0	515	None	Never	Fixed	96.269	38111	1477	0	0	0
rd	Continuous	0	509	None	Never	Fixed	96.269	38111	1477	0	0	0
homogeneity	Continuous	0	0	None	Never	Fixed	76.667	30351	9237	0	0	0
n	Continuous	0	0	None	Never	Fixed	94.14	37268	2320	0	0	0
herf	Continuous	0	515	None	Never	Fixed	94.14	37268	2320	0	0	0
intout	Continuous	515	0	None	Never	Fixed	94.14	37268	2320	0	0	0
contcrisis	Continuous	0	751	None	Never	Fixed	100	39588	0	0	0	0

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.

2.4.2. Extreme Values Explained and Feature Value (coding) Inconsistencies



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

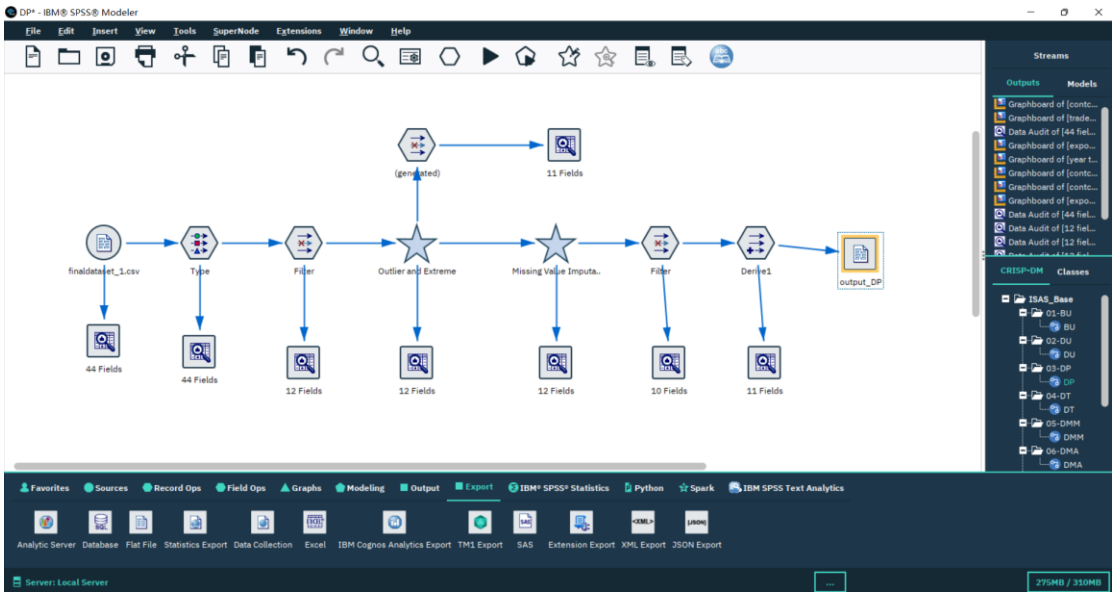
2.4.3. Measurement errors

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

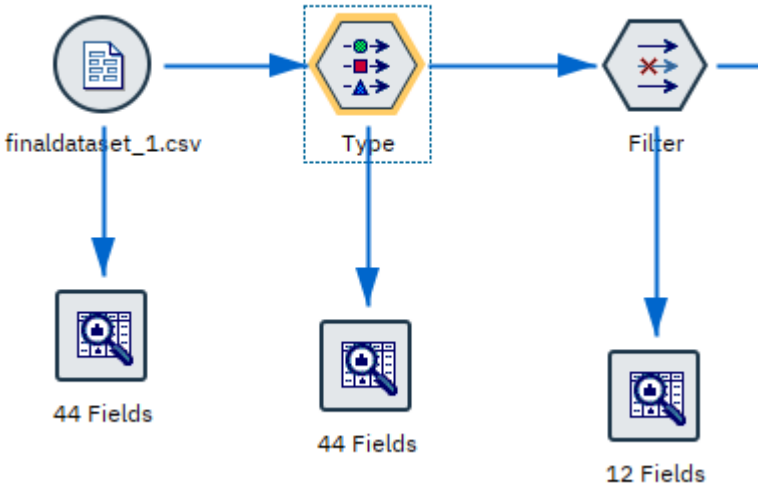
The “developing” field has the largest number of outliers, but it is because of the measurement is set to “continuous” wrongly, as it is well known the countries could be classified to developing countries and developed countries. The measurement must be changed in the data preparing phase.

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

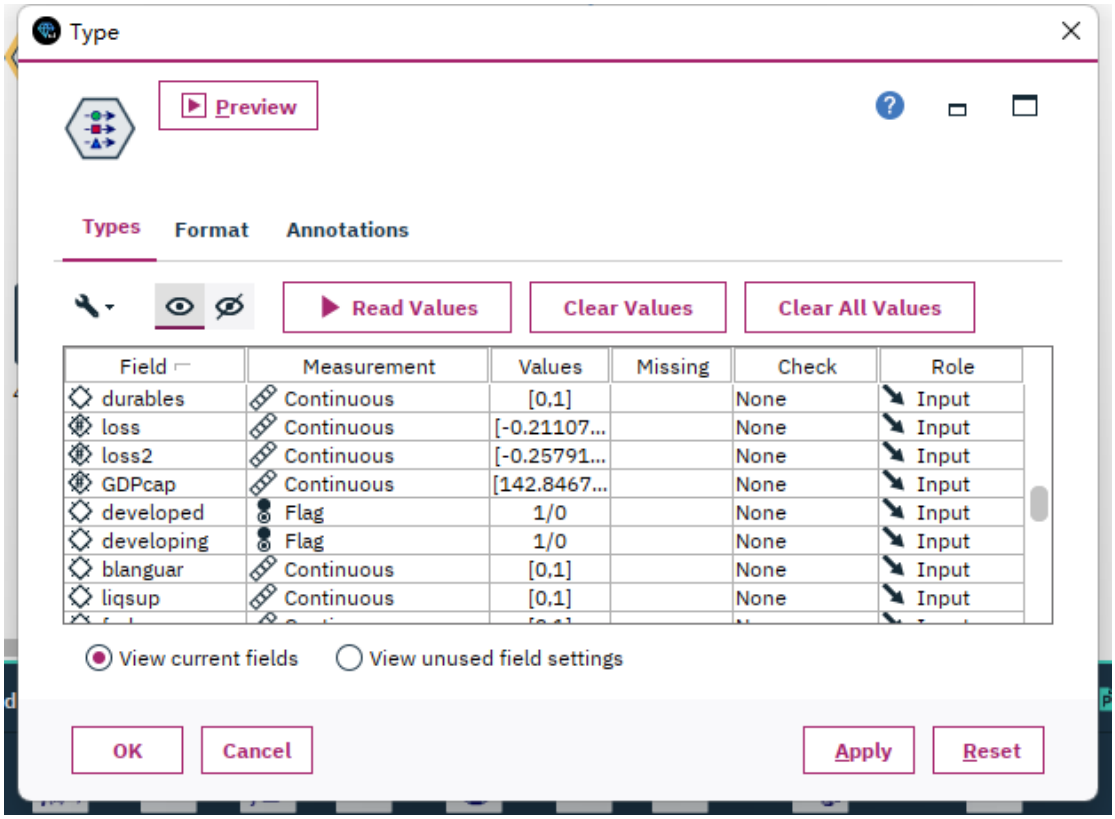
3. Data Preparation



3.1 Select the data

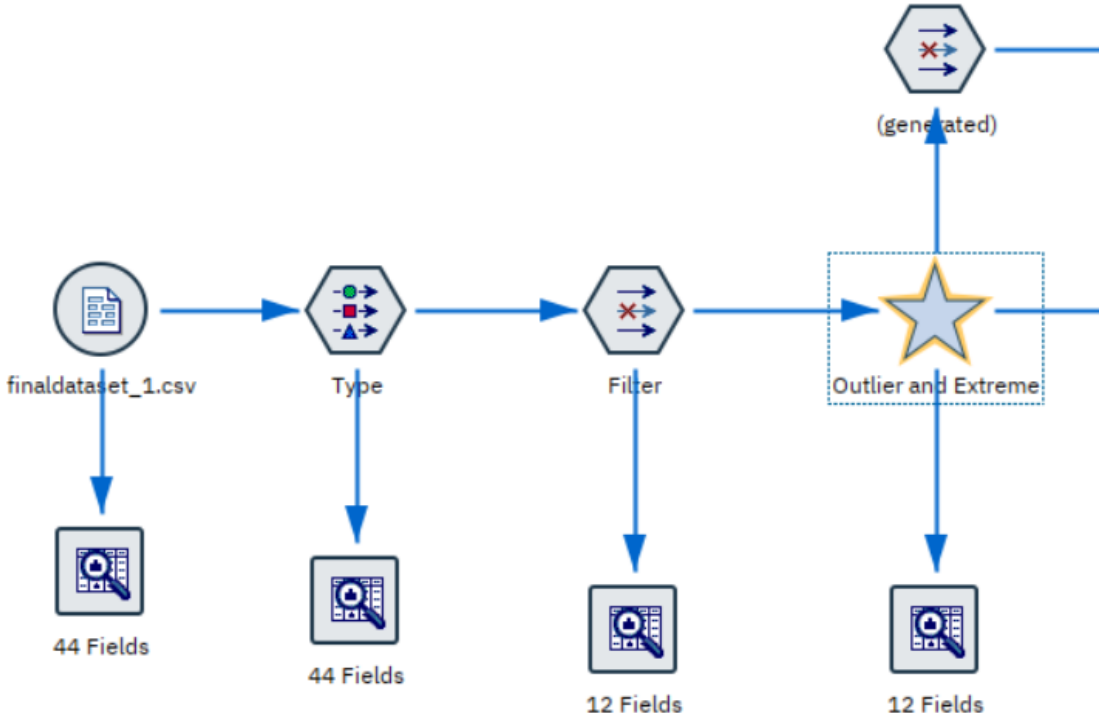


First make some change to the “measurement” field in the type node.

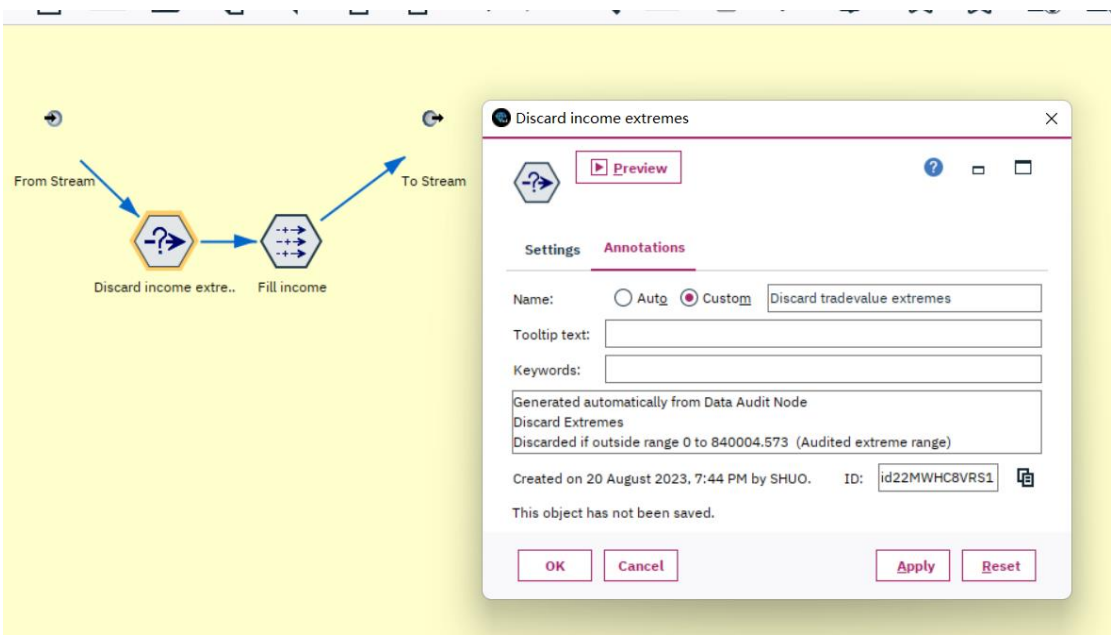


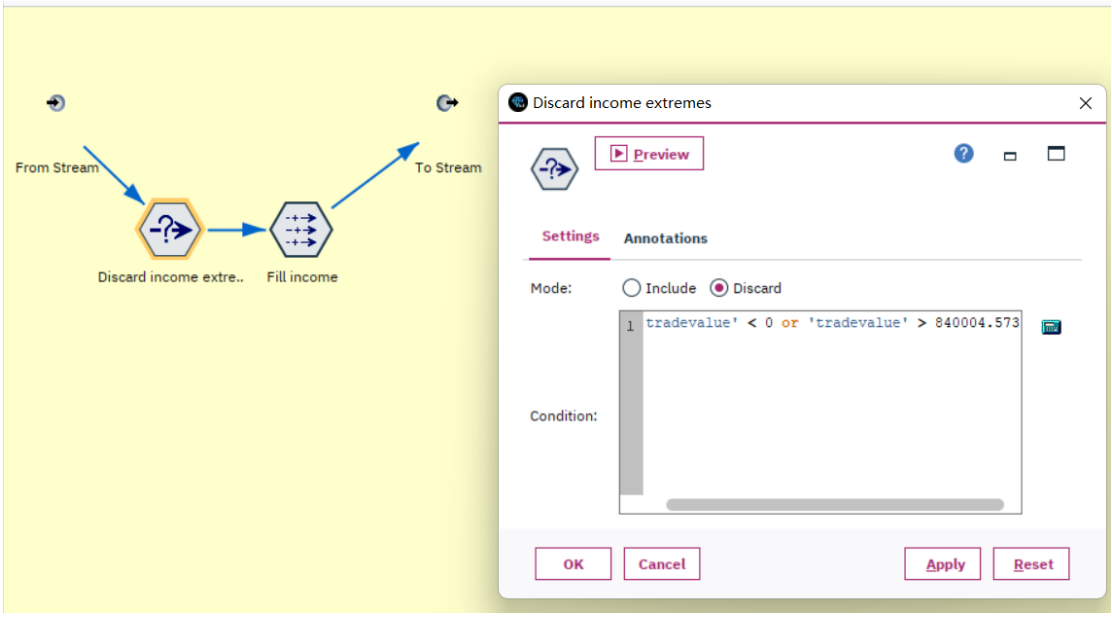
This study needs to build models for developed countries and developing countries separately, so a filter node is needed to select the data collected from developed countries, and in the other branch, select the data collected from developed countries.

3.2 Clean the data

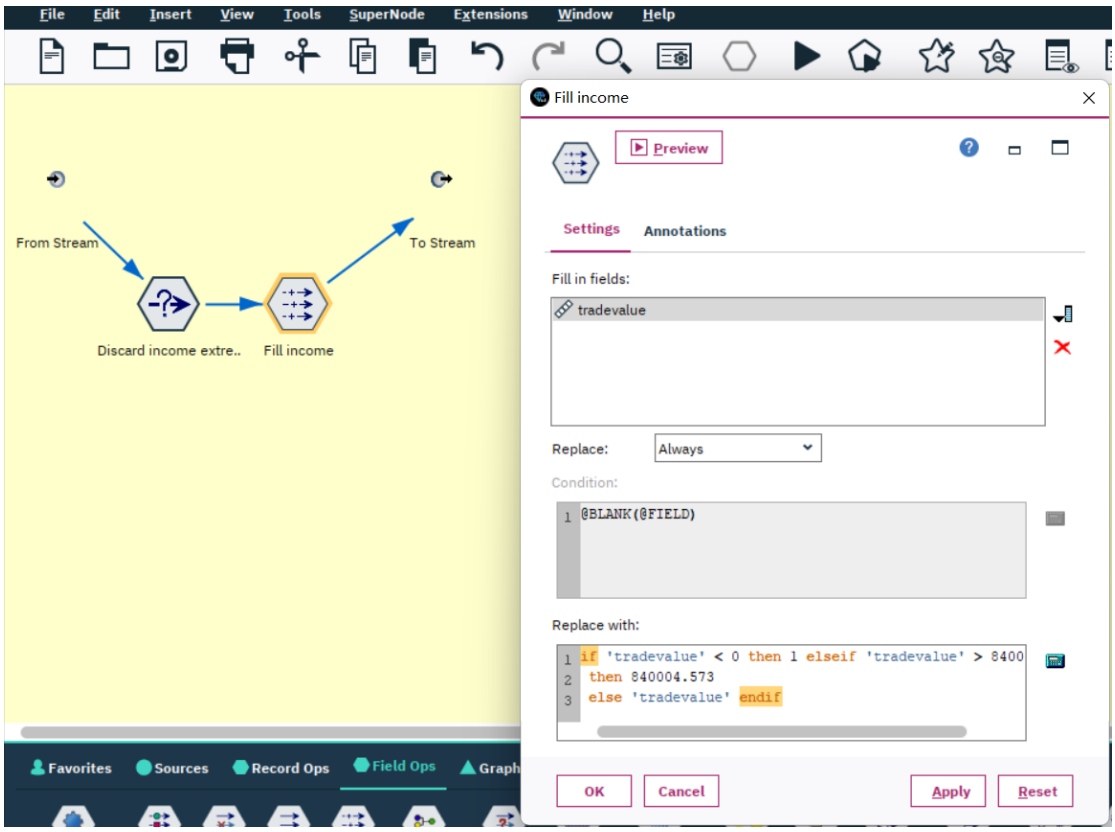


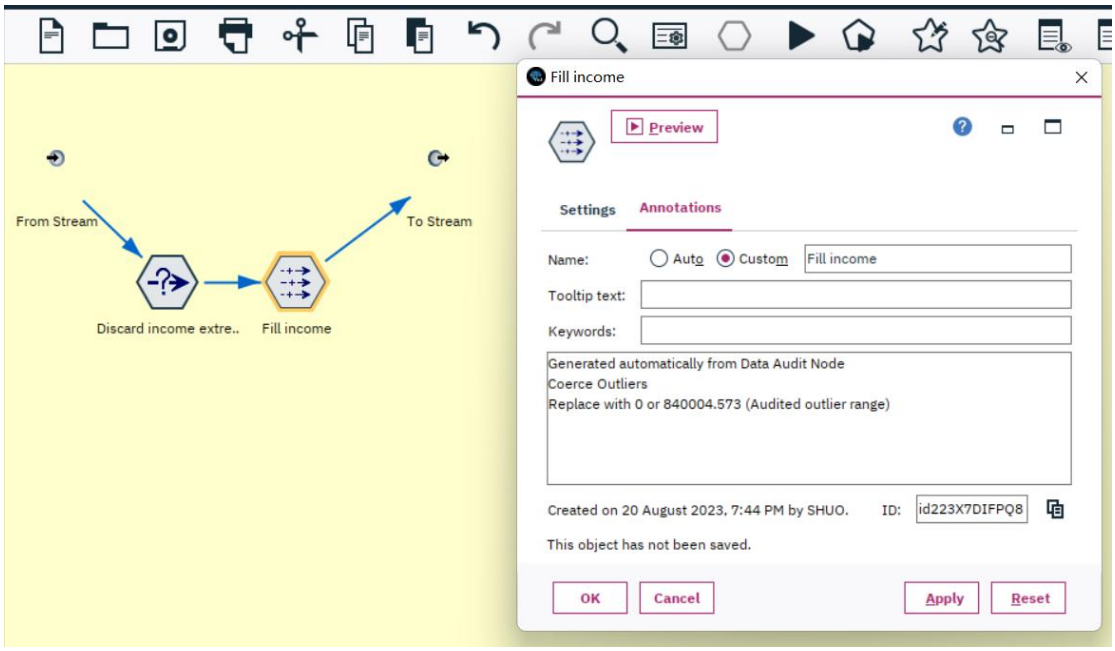
Dealing with the extremes:





Dealing with the NAs:





Data Audit of [12 fields] #1

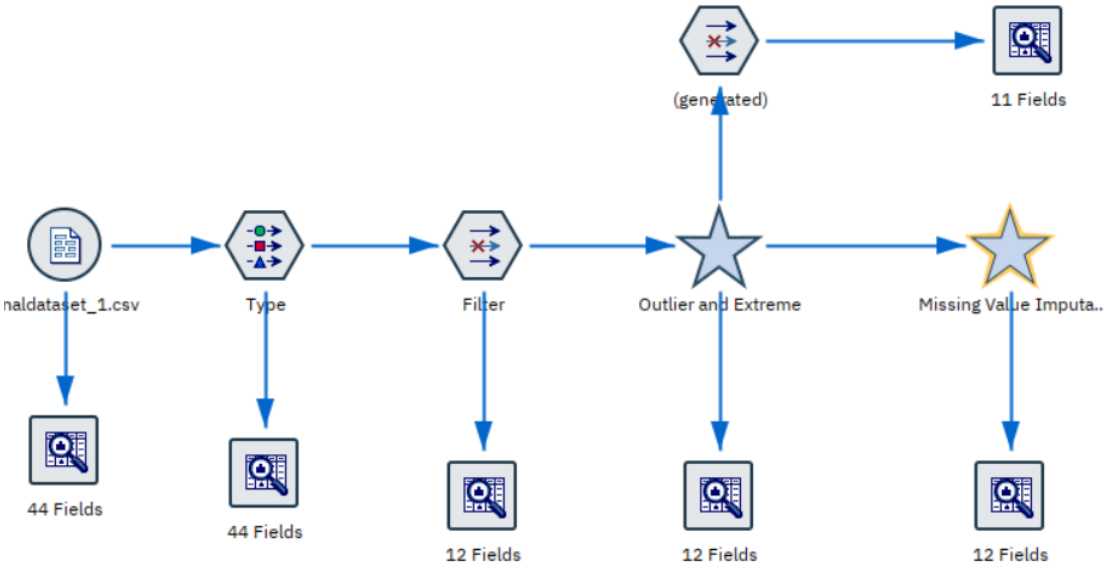
File Edit Generate

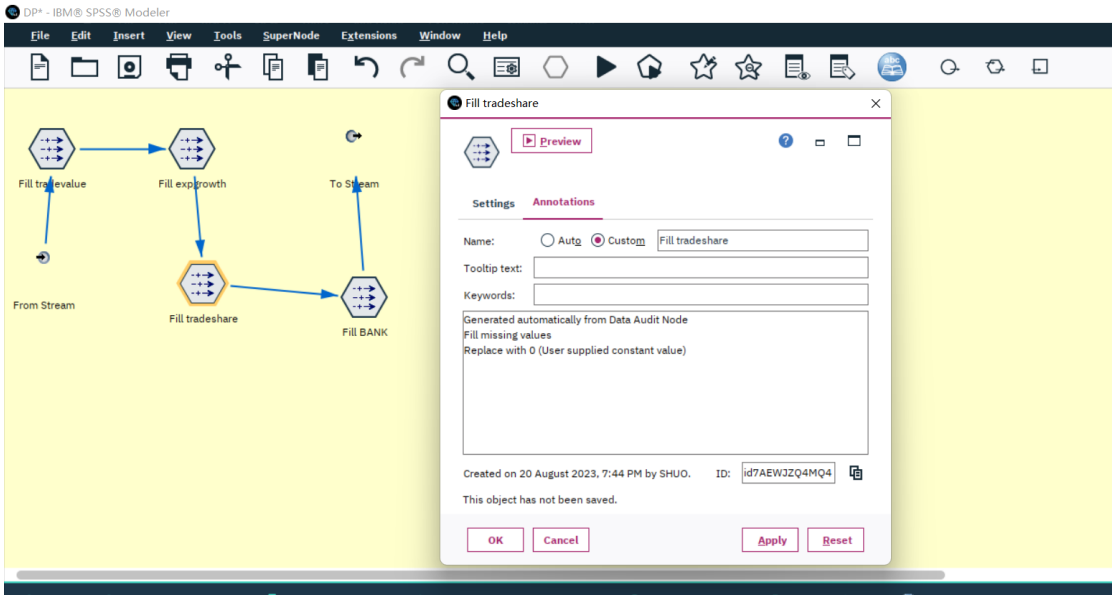
Audit Quality Annotations

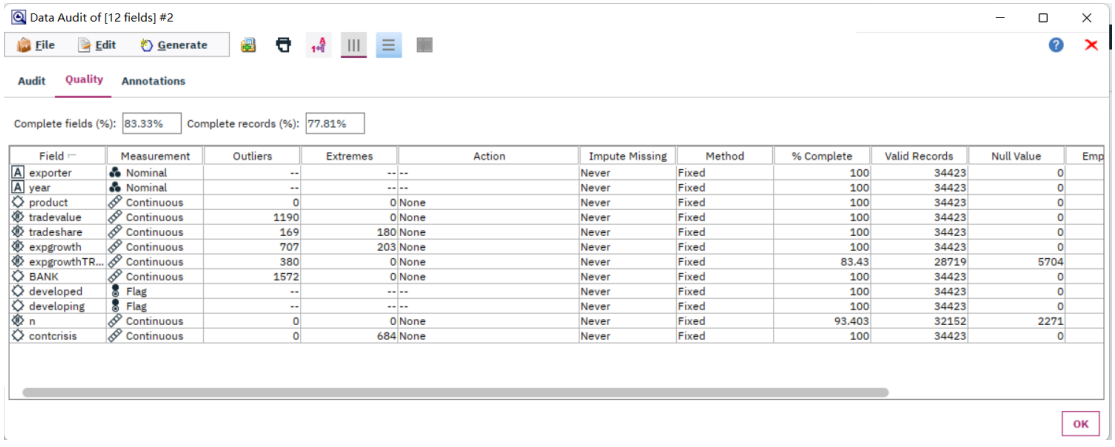
Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
exporter		Nominal	--	--	--	--	--	22	34423
year		Nominal	--	--	--	--	--	27	34423
product		Continuous	3111	3909	3492.327	288.527	0.017	--	34423
tradevalue		Continuous	1.000	839663.207	87134.190	160181.321	2.563	--	34423
tradeshare		Continuous	0.000	0.952	0.009	0.036	13.678	--	30361
expgrowth		Continuous	-8.566	8.631	0.107	0.843	0.297	--	32437
expgrowthT...		Continuous	-1.527	1.523	0.101	0.362	0.193	--	28719
BANK		Continuous	0	1	0.046	0.209	4.345	--	34317
developed		Flag	0	1	--	--	--	2	34423
developing		Flag	0	1	--	--	--	2	34423
n		Continuous	0.403	1.729	1.139	0.309	-0.192	--	32152

[1] Indicates a multimode result [2] Indicates a sampled result

OK



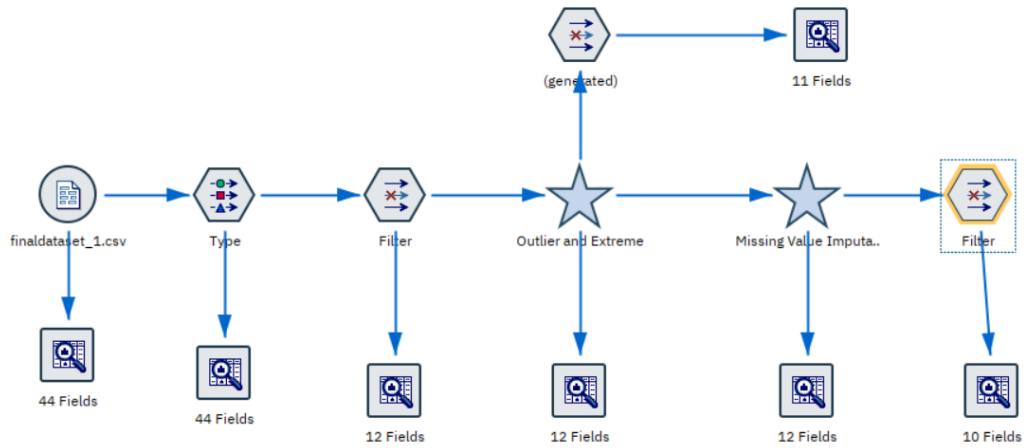




Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Emp
exporter	Nominal	--	--	--	Never	Fixed	100	34423	0	
year	Nominal	--	--	--	Never	Fixed	100	34423	0	
product	Continuous	0	0	None	Never	Fixed	100	34423	0	
tradevalue	Continuous	1190	0	None	Never	Fixed	100	34423	0	
tradeshare	Continuous	169	180	None	Never	Fixed	100	34423	0	
expgrowth	Continuous	707	203	None	Never	Fixed	100	34423	0	
expgrowthTR...	Continuous	380	0	None	Never	Fixed	83.43	28719	5704	
BANK	Continuous	1572	0	None	Never	Fixed	100	34423	0	
developed	Flag	--	--	--	Never	Fixed	100	34423	0	
developing	Flag	--	--	--	Never	Fixed	100	34423	0	
n	Continuous	0	0	None	Never	Fixed	93.403	32152	2271	
contcrisis	Continuous	0	684	None	Never	Fixed	100	34423	0	

The data quality increased a lot, the complete fields increases from 36.36% to 83.33% and the complete records increased from 17.56% to 77.81%.

There are still 2 fields with many null values, filter them out.



Data Audit of [10 fields]

File Edit Generate

Audit Quality Annotations

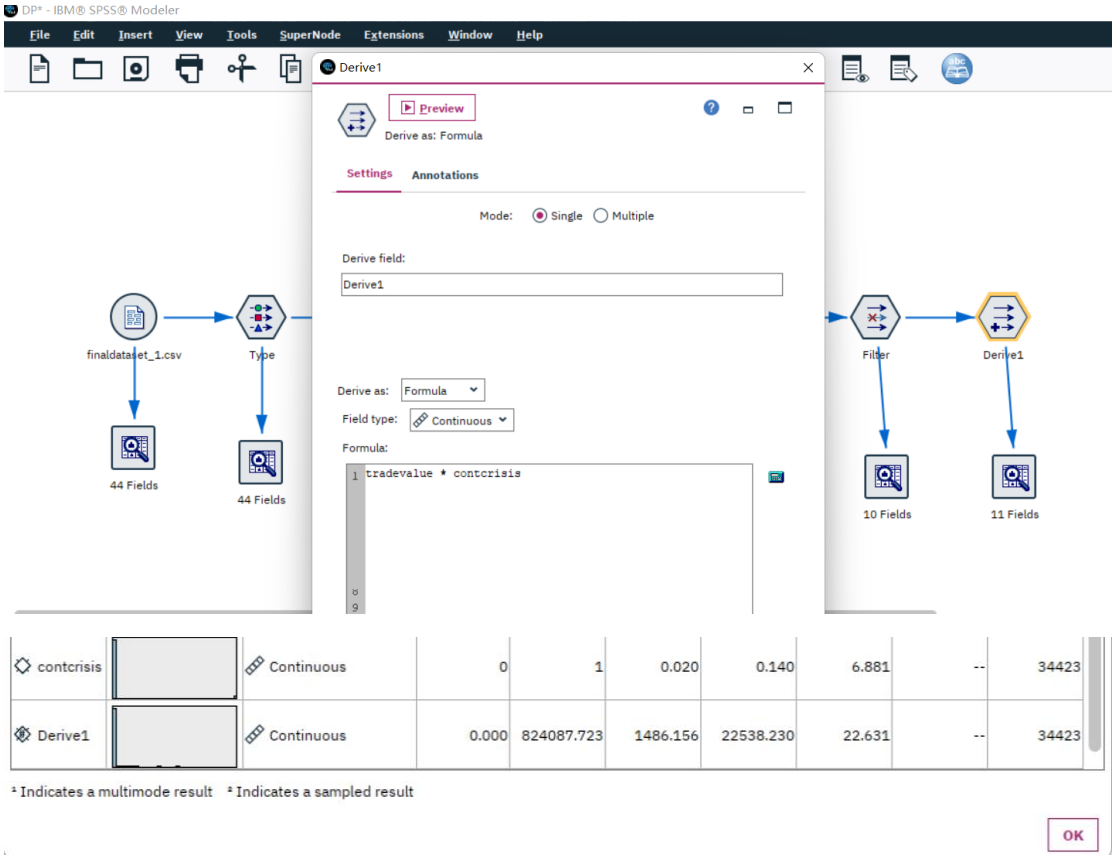
Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute M
exporter	Nominal	--	--	--	Never
year	Nominal	--	--	--	Never
product	Continuous	0	0	None	Never
tradevalue	Continuous	1190	0	None	Never
tradeshare	Continuous	169	180	None	Never
expgrowth	Continuous	707	203	None	Never
BANK	Continuous	1572	0	None	Never
developed	Flag	--	--	--	Never
developing	Flag	--	--	--	Never
contcrisis	Continuous	0	684	None	Never

Now the data is 100% complete.

3.3 Construct the data

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

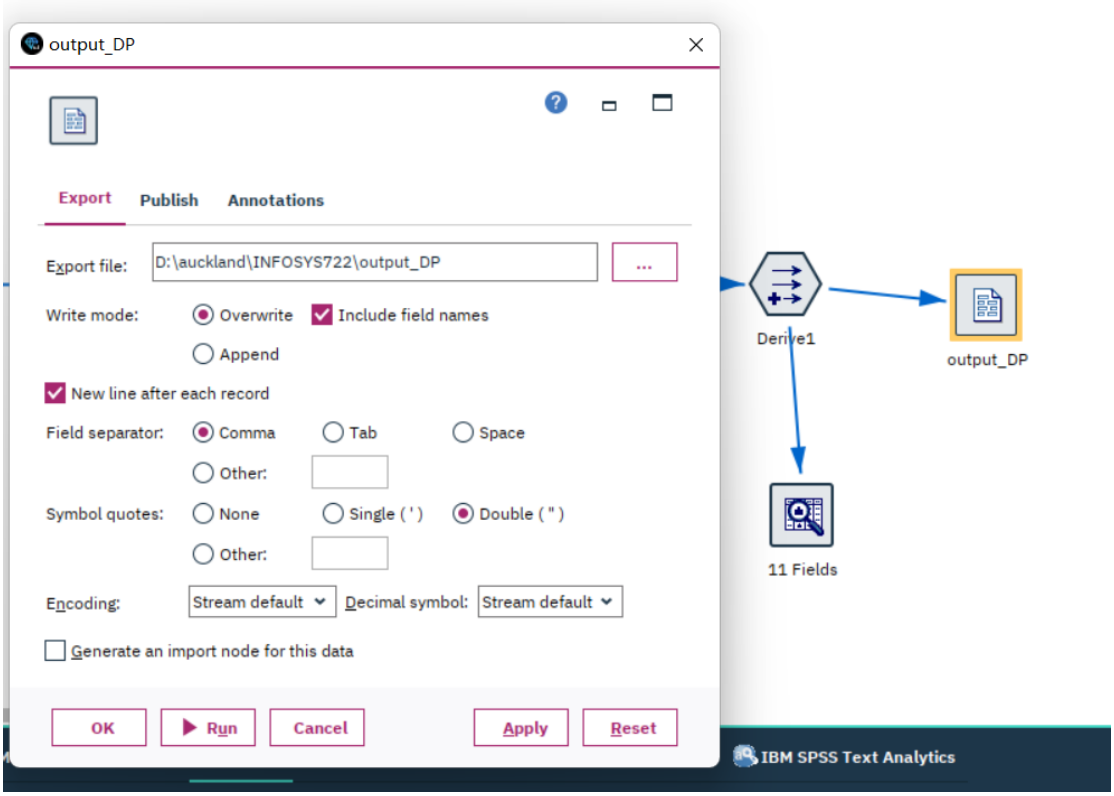


3.4 Integrate various data sources

Since the finaldataset_1.csv is the only data source the iteration has, there is no need to integrate other various data sources.

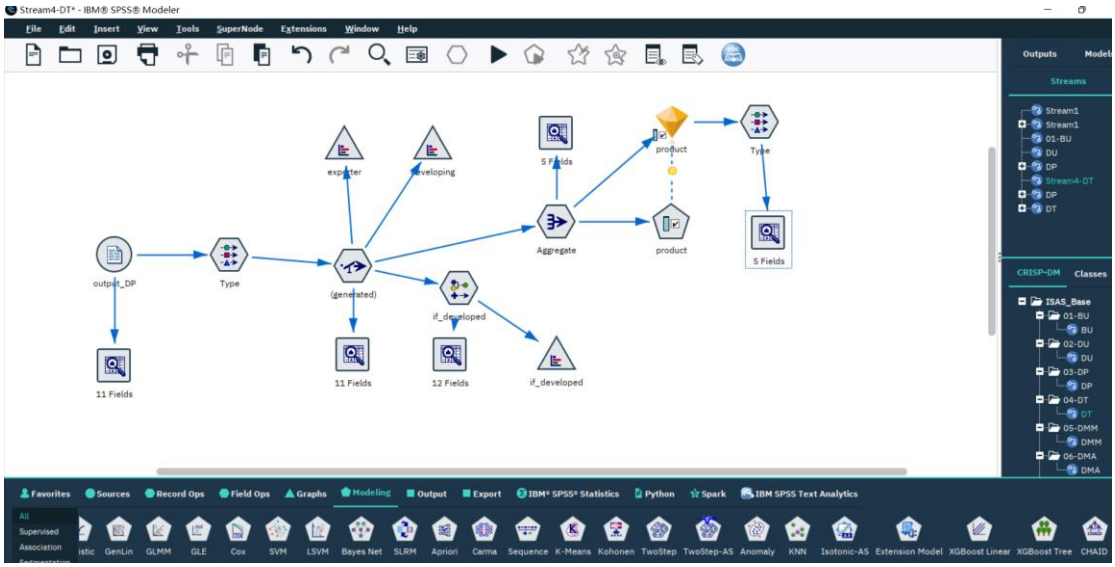
3.5 Format the data as required

After data cleaning and data constructing , the format is already as required.

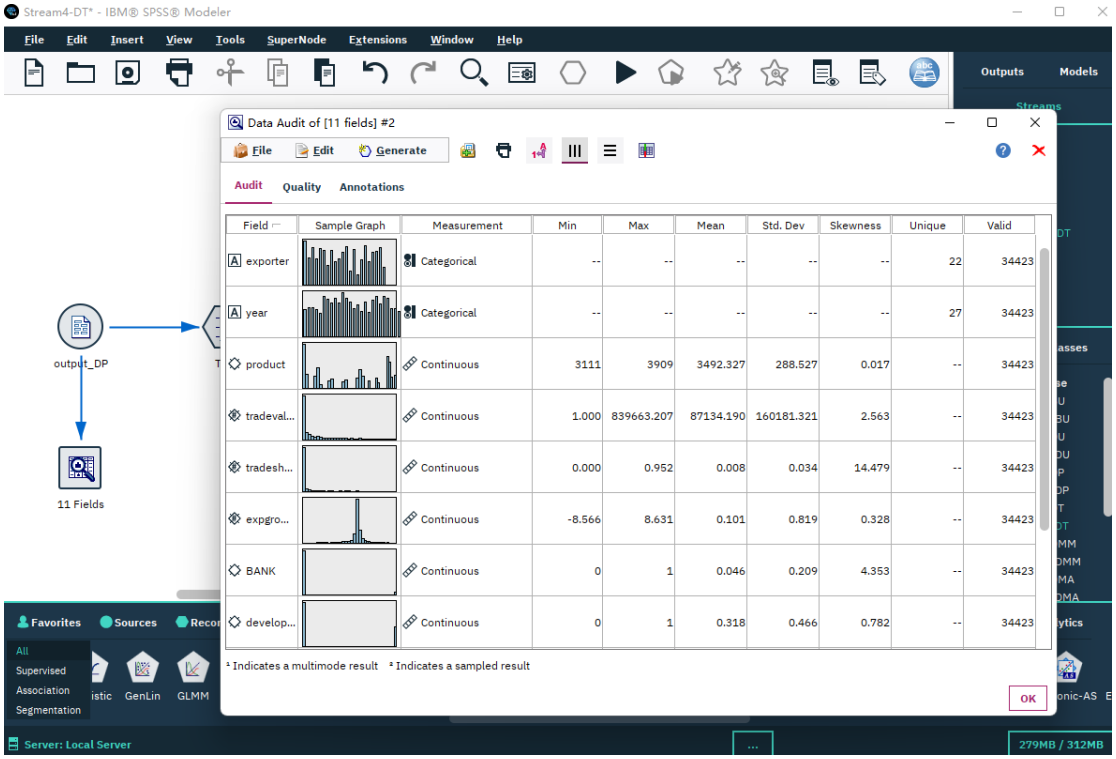


Save to disk after all the data preparation is done.

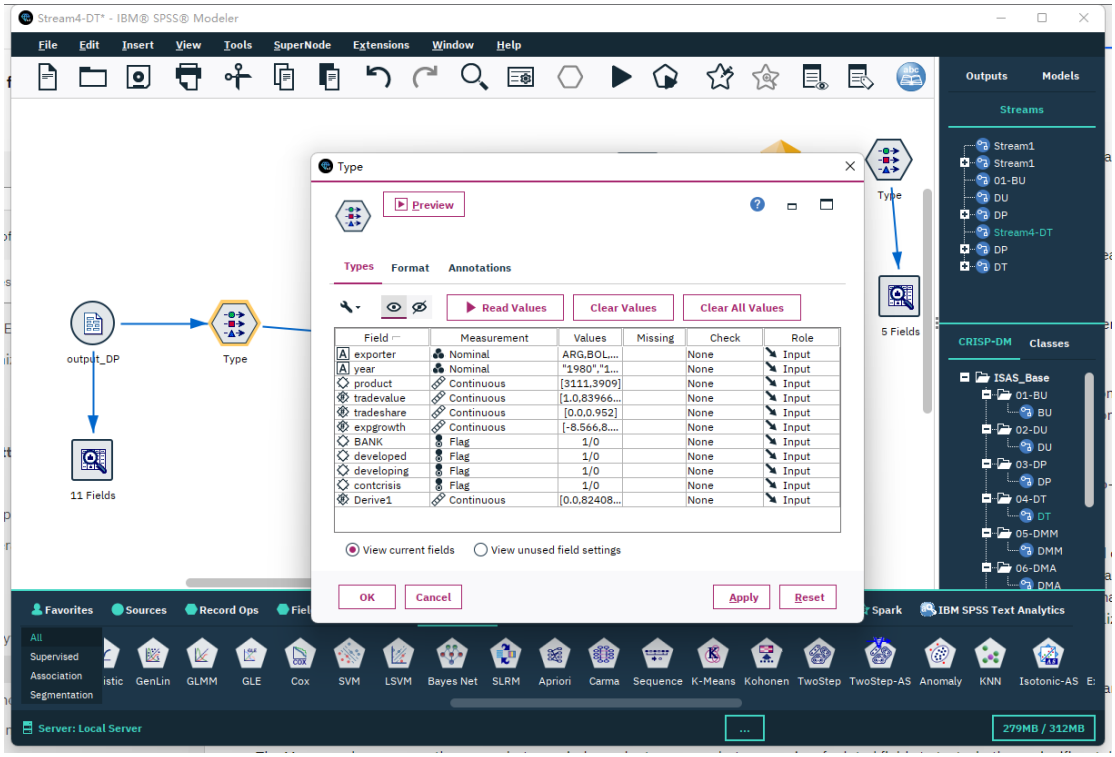
4. Data Transformation



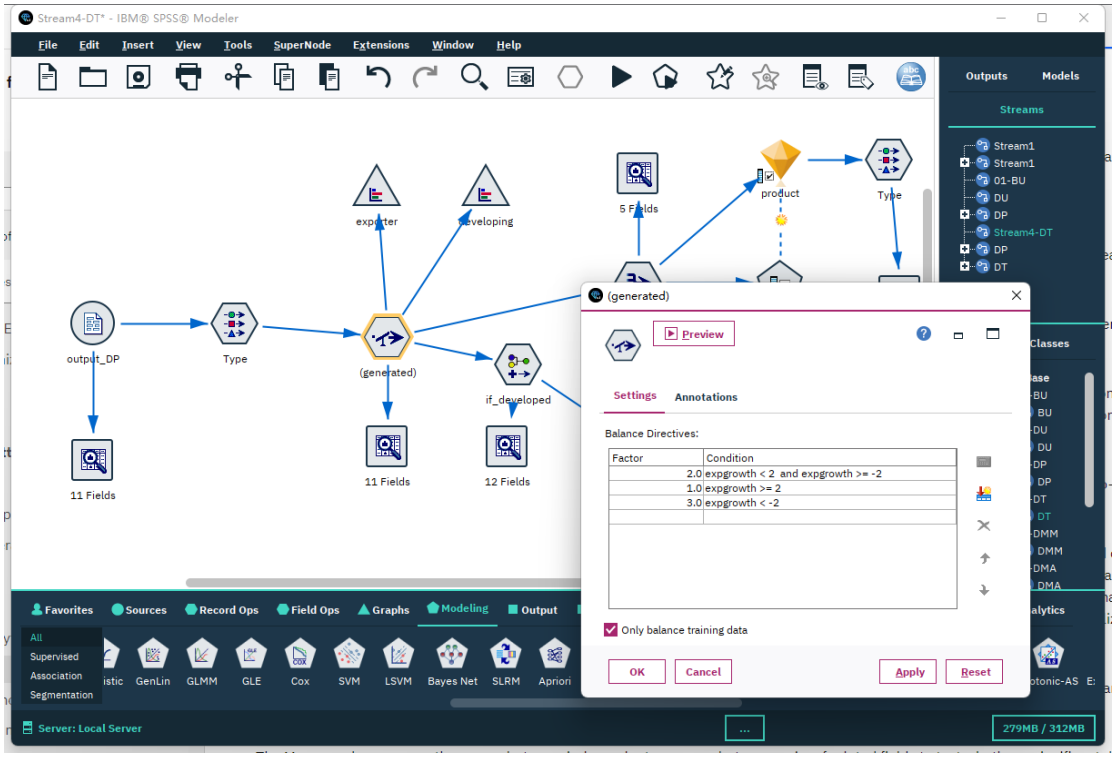
4.1 Reduce the data



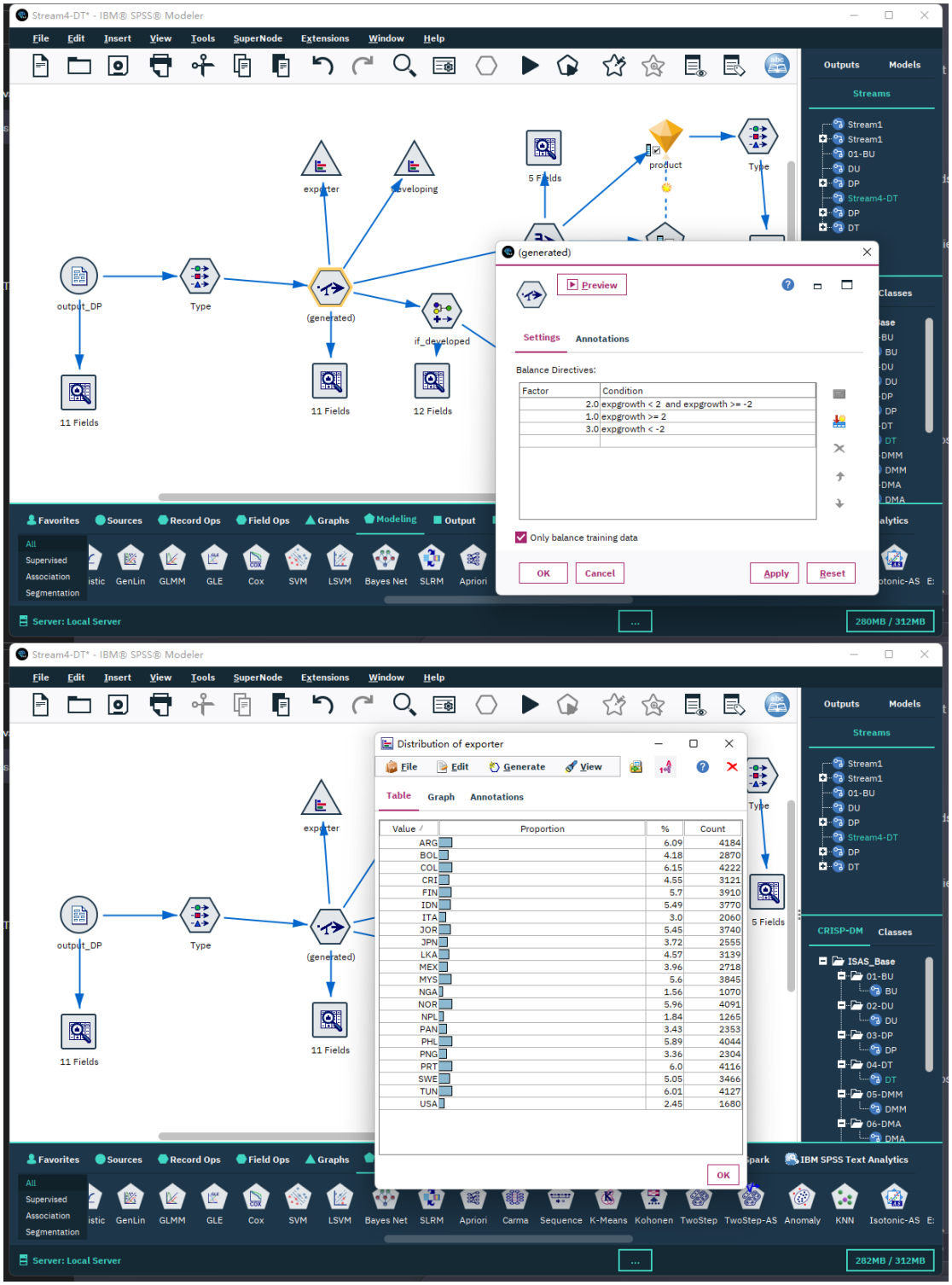
Using data audit node to visualize the input of this phase, which is also the output of the last data preparation phase.



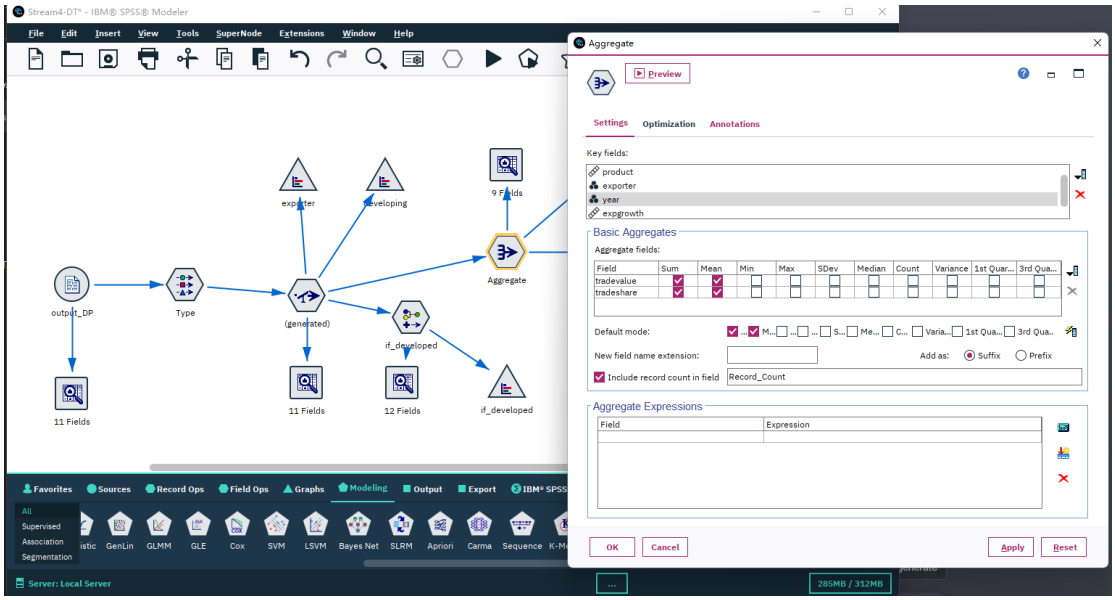
Organize the measurements of the fields in Type node.



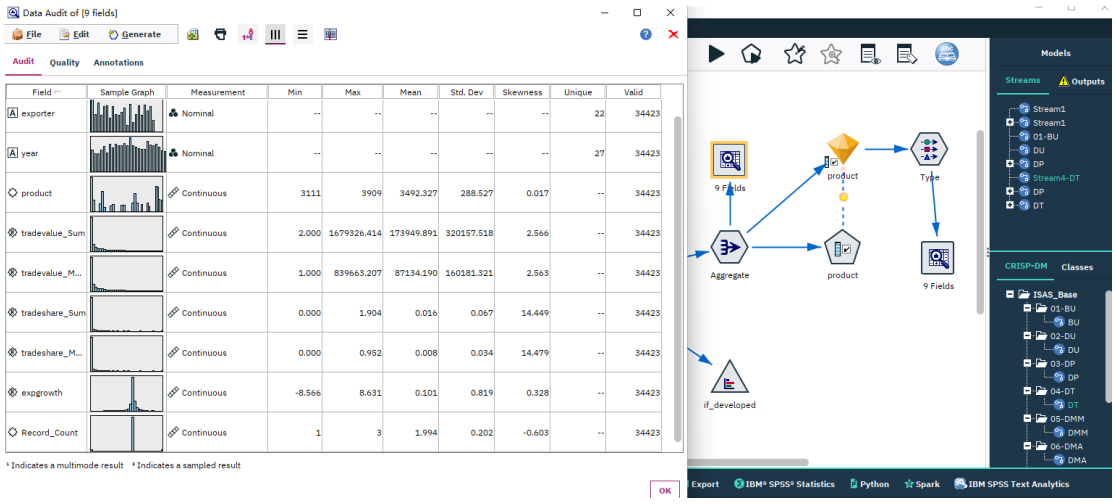
Set the different export growth speed.

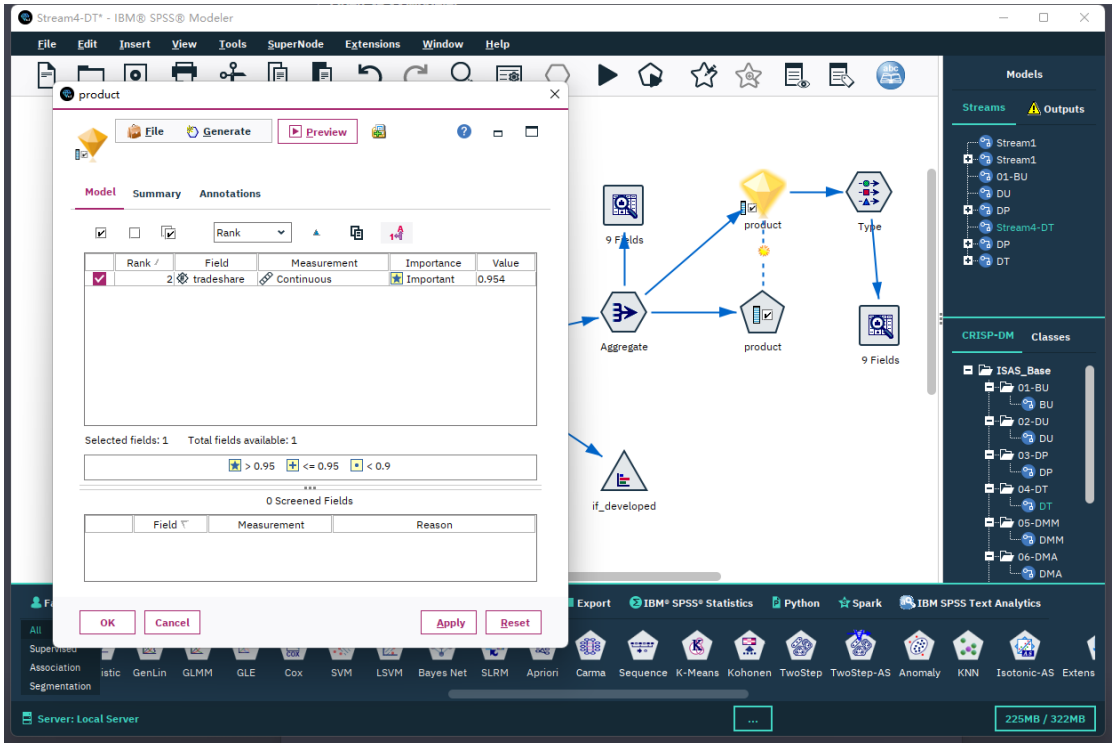


Using aggregate node to summarize data, calculating the mean and sum.

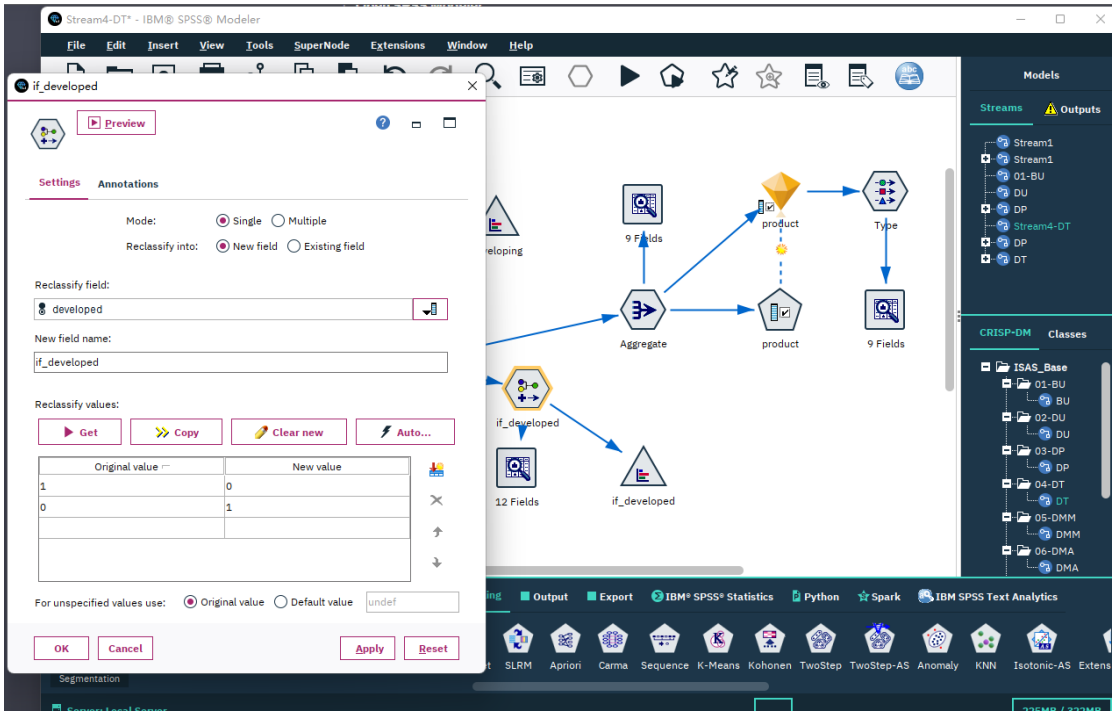


The result of aggregate:

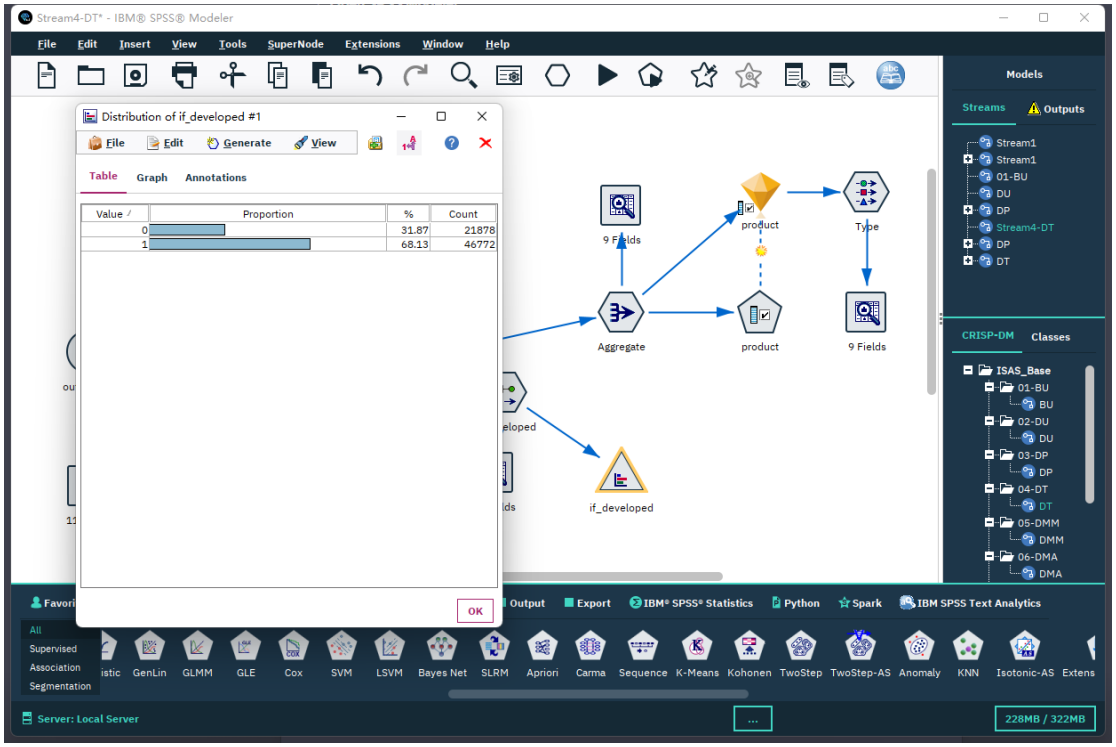




Combine the developing field and develop field:

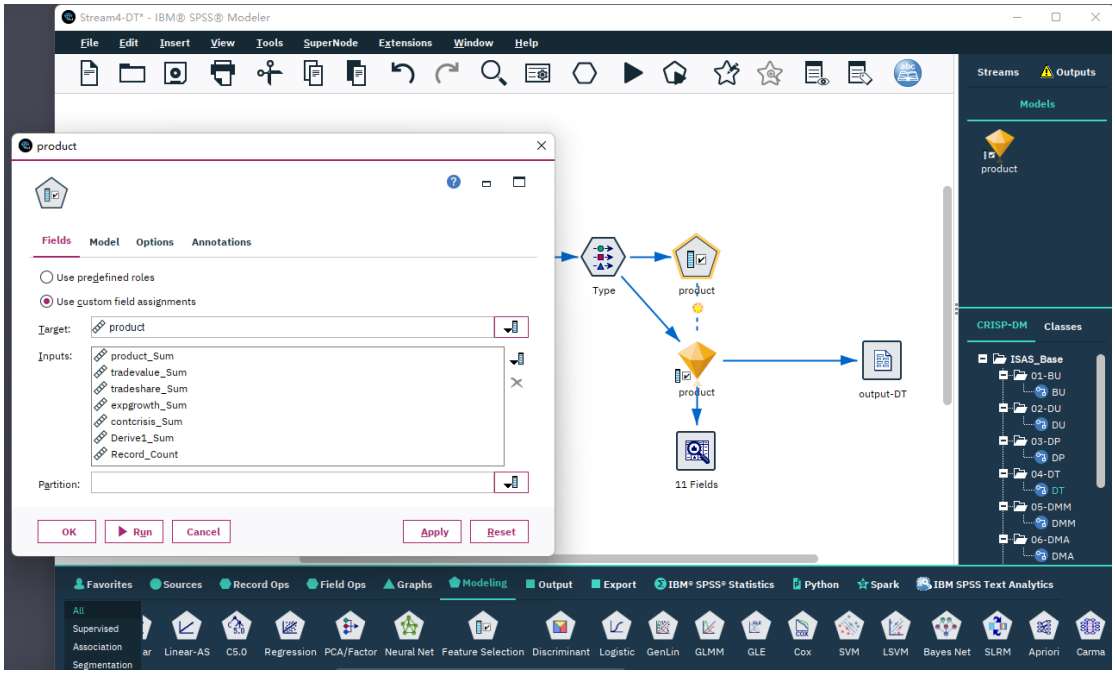


The result:



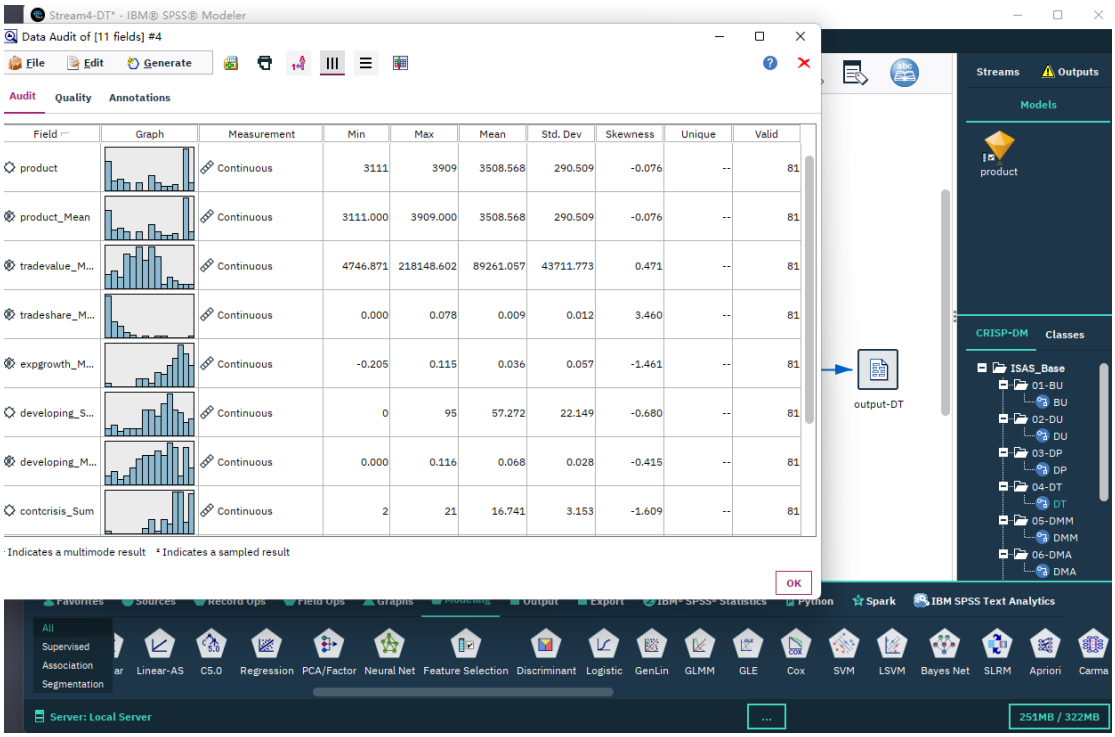
4.2 Project the data

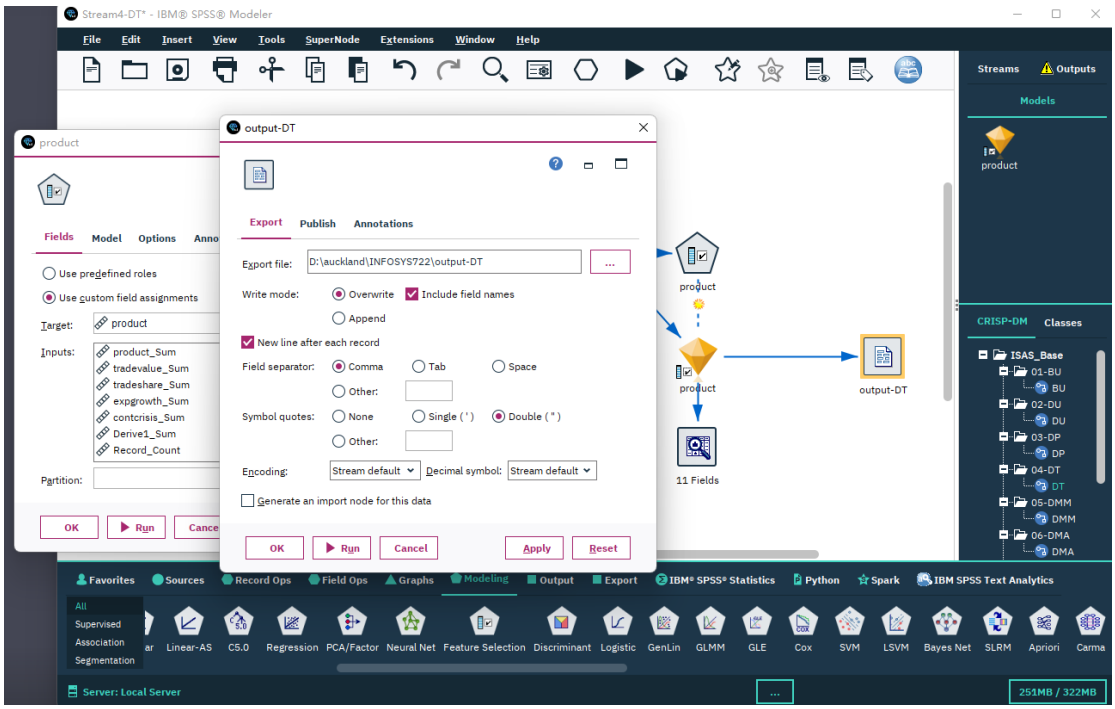
Project the data by using Feature Selection node:



Export all the data in this phase to a csv file named “output-DT.csv”

ITERATION 2, ISAS





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

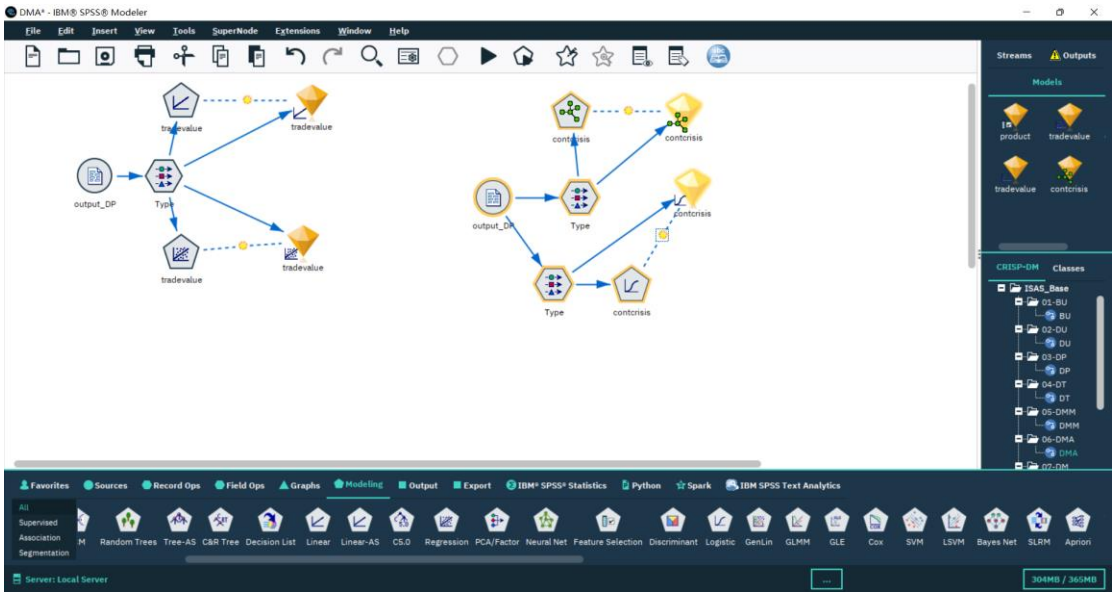
Evaluation:
After building the Linear Regression model, use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared to evaluate its performance on a validation set.

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

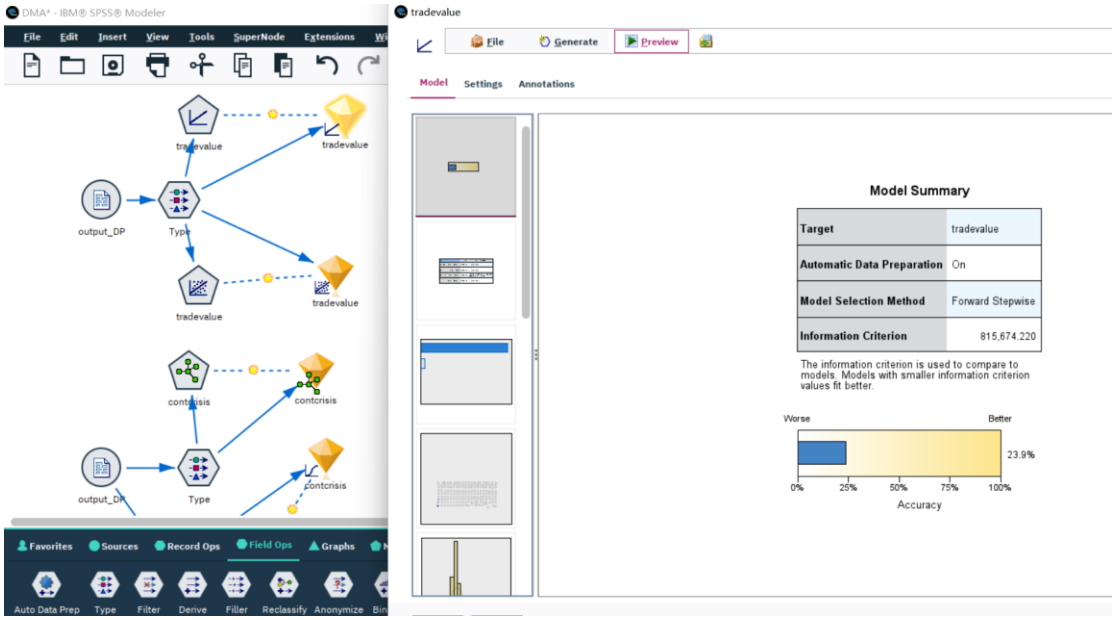
Evaluation:
After building the Random Forest model, use accuracy, precision, recall, and the AUC-ROC curve to evaluate its performance.

6. Data-Mining Algorithm(s) Selection

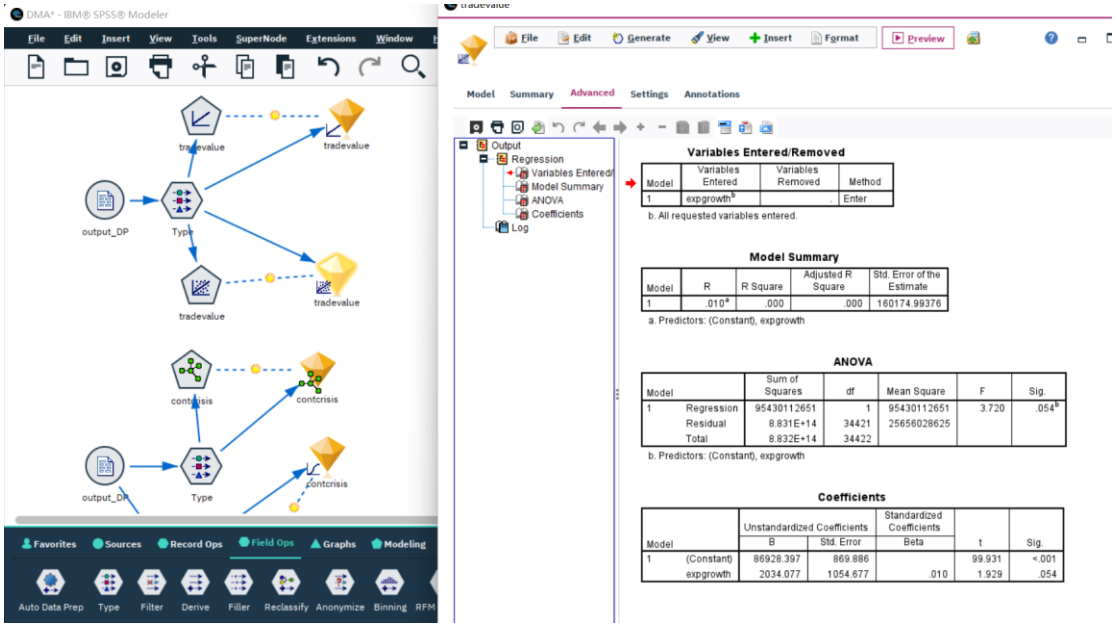


6.1 Conduct exploratory analysis and discuss

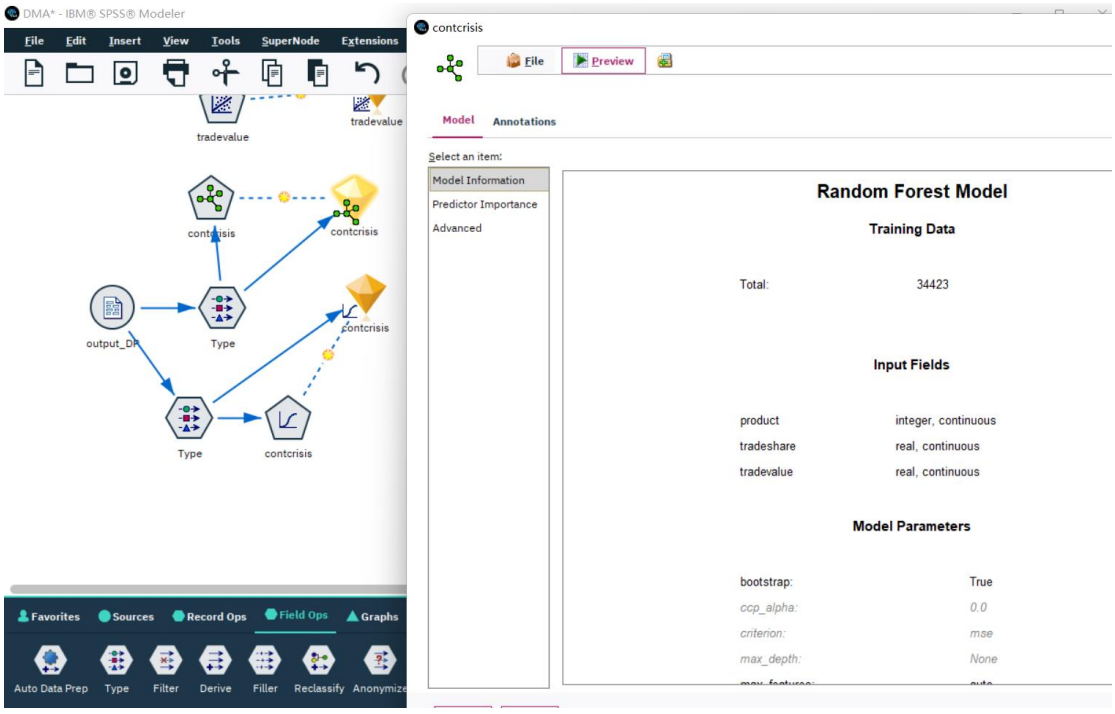
6.1.1. First Data-Mining Objective: The Linear Algorithm



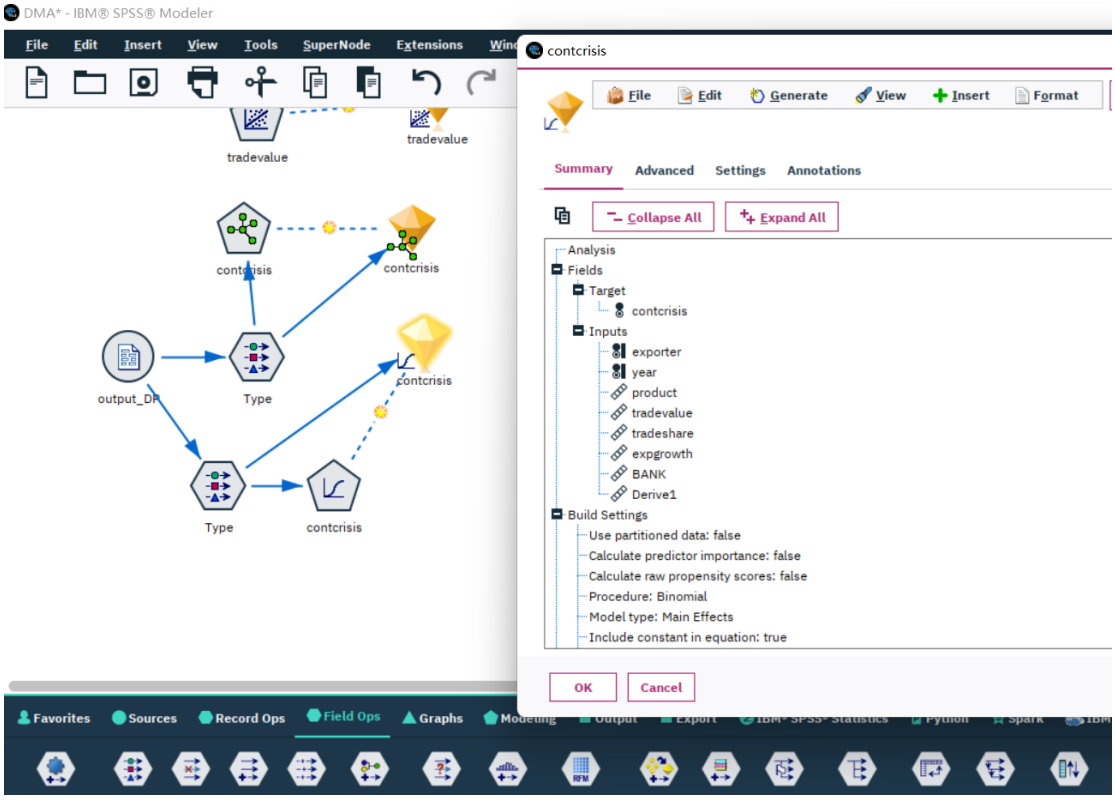
6.1.2. First Data-Mining Objective: The Regression Algorithm



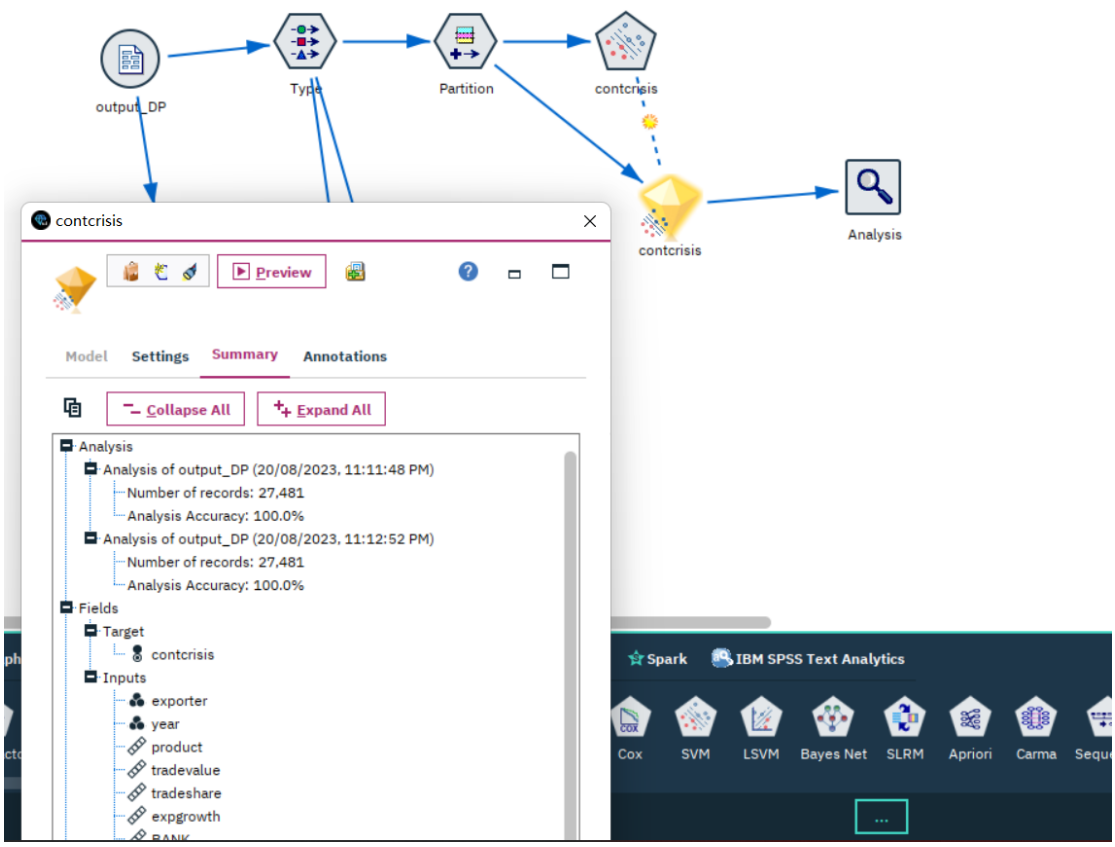
6.1.3. Second Data-Mining Objective: The Classification Algorithm



6.1.4. Second Data-Mining Objective: The Logistic Regression Algorithm



6.1.5. Second Data-Mining Objective: SVM



6.2 Select data-mining algorithms based on discussion

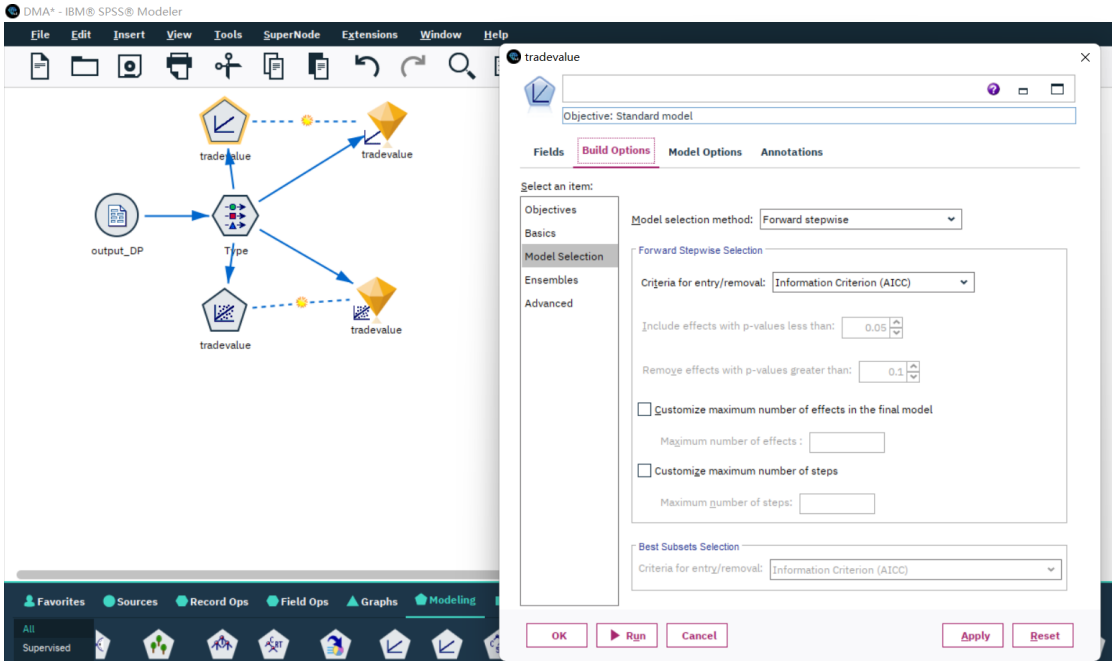
For the first objective, linear algorithm should be chosen. For the time series analysis failed because the data format of YEAR field is still string. This must be solved in the next iteration.

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, SVM should be chosen. This algorithm handles mixed types of data well, are 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.

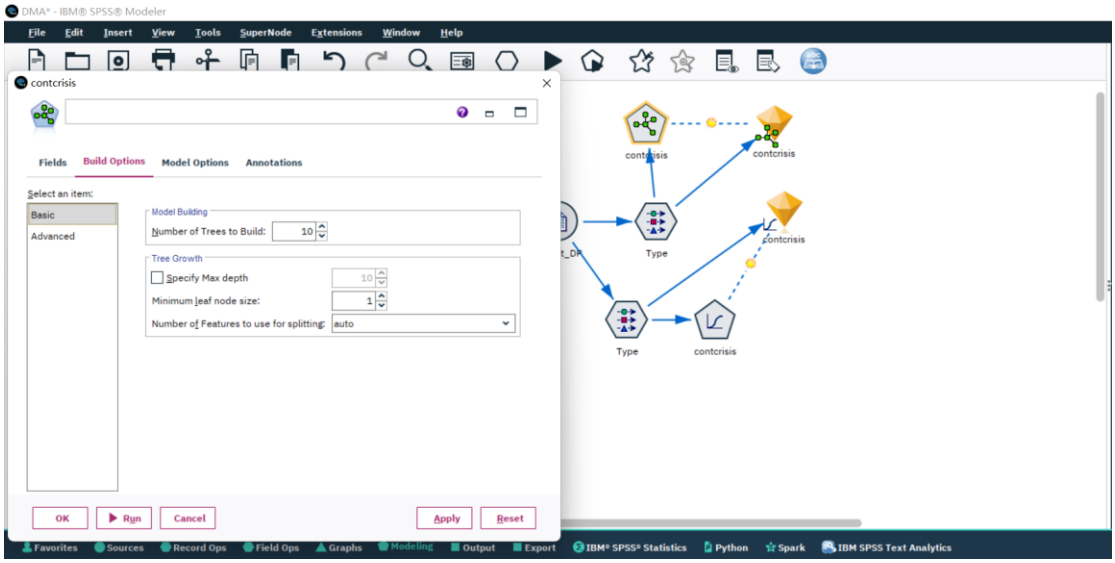
6.3 Build/Select appropriate model(s) and choose relevant parameter(s)

For the linear model:

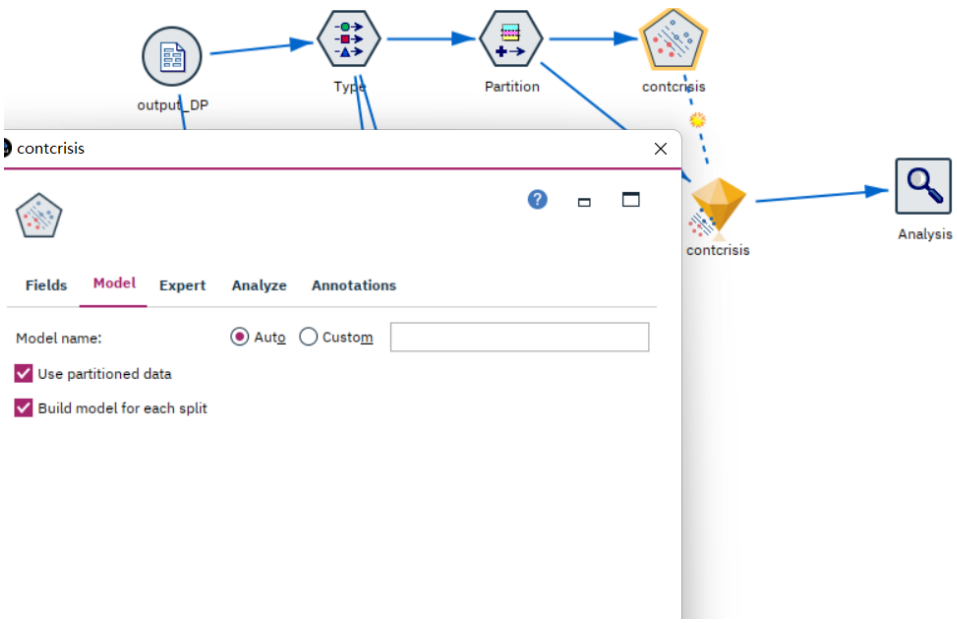


The model selection method is set to “forward stepwise”, criteria for entry/removal is set to”AICC" the others are set to default.

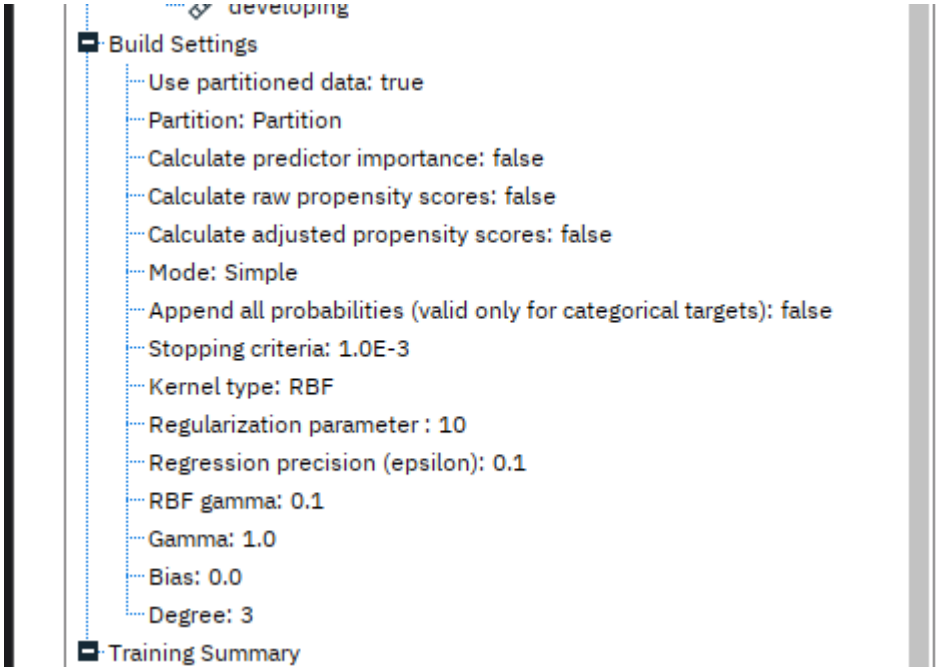
For the random forest, the number of trees to build is set to 10, minimum leaf node size is set to 1, number of feature to use for splitting is set to “auto”



For the SVM, only need to set to “auto”.



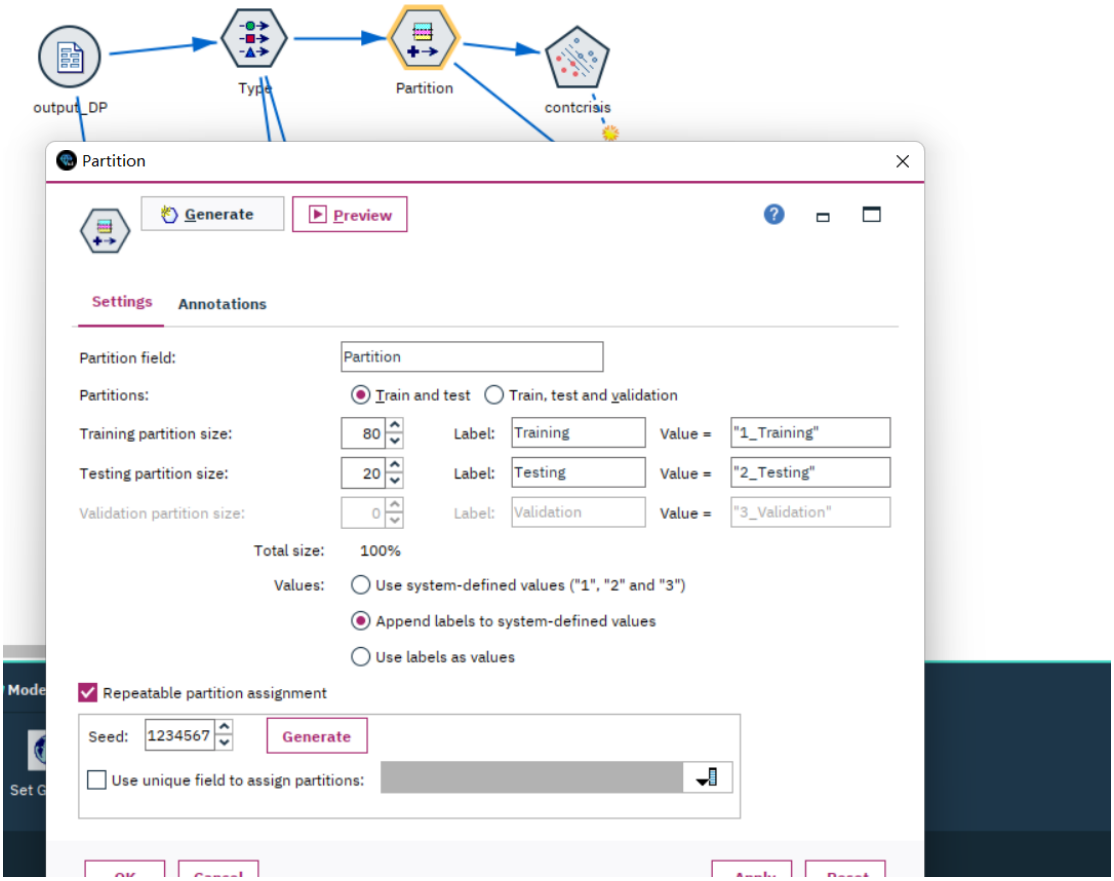
The default build settings are as follow:



7. Data Mining

7.1 Create and justify test designs

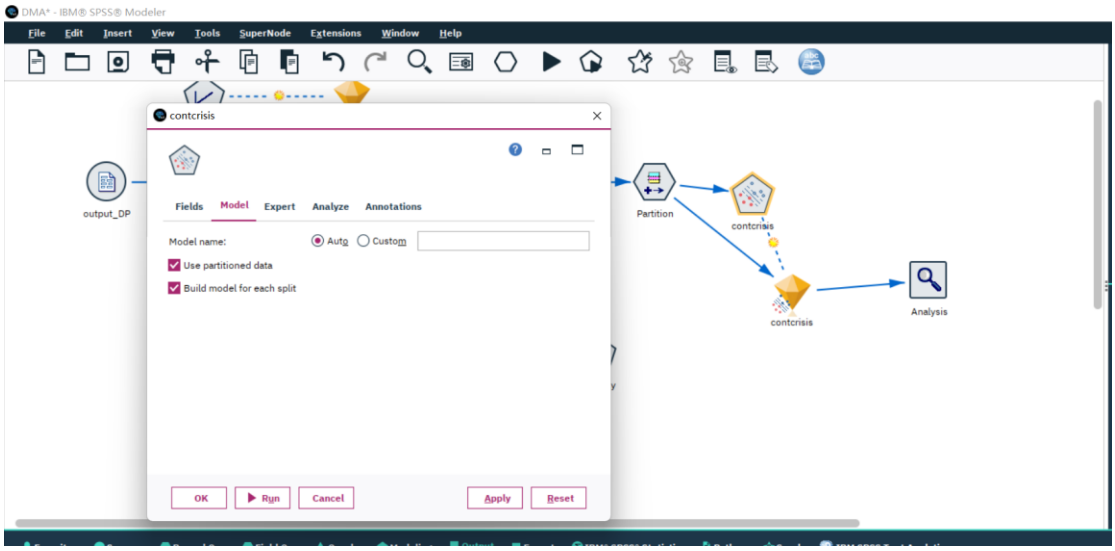
Divide the dataset into training and testing, 80% for training and 20% for testing.



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

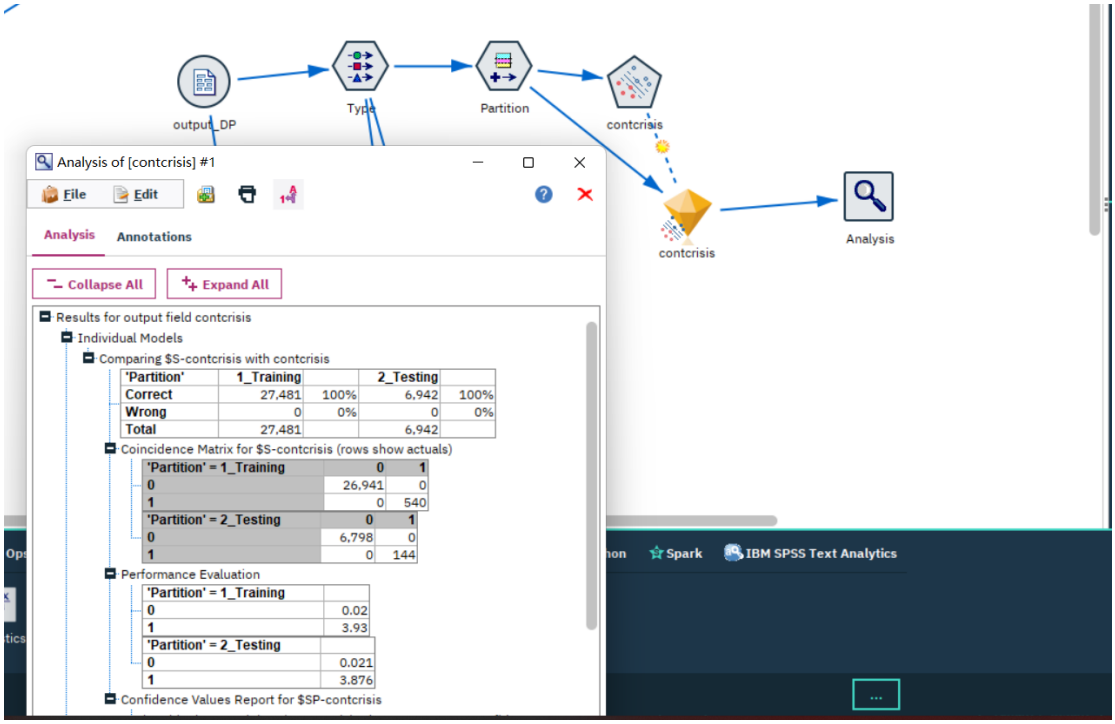
7.2 Conduct data mining – classify, regress, cluster, etc. (models must execute)

Binary classification:

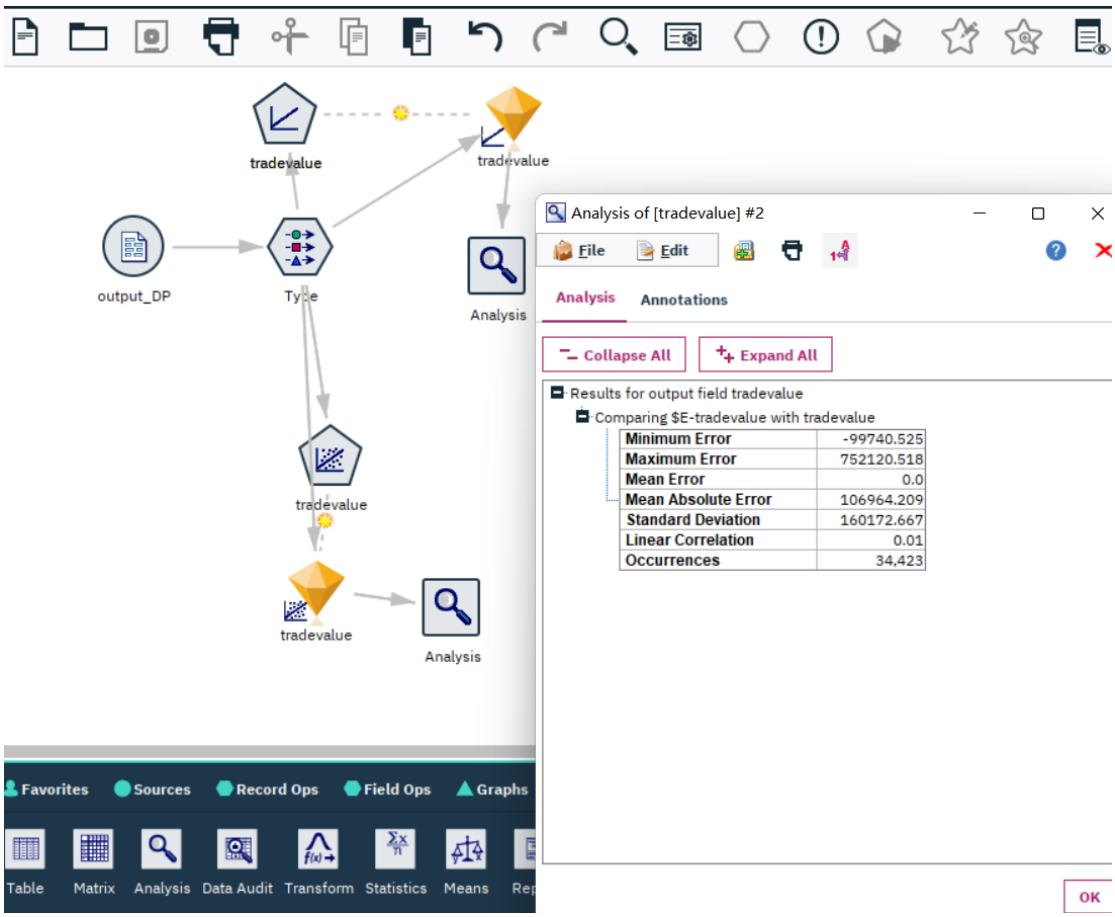


7.3 Search for patterns

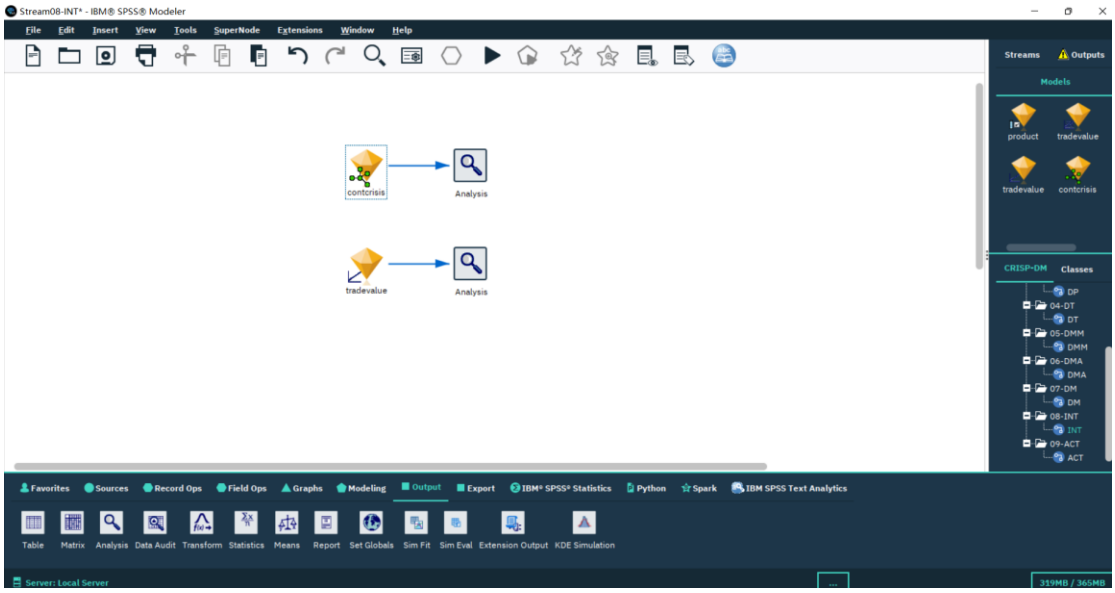
The binary classification model performs perfect with its 100% correct rate.



The linear model and regression model are both built unsuccessfully. Another iteration is needed.


tradevalue
tradevalue
output_DP
Type
Analysis
tradevalue
tradevalue
Analysis
Analysis of [tradevalue] #2
File
Edit
Analysis
Annotations
Collapse All
Expand All
Results for output field tradevalue
Comparing $E-tradevalue with tradevalue
Minimum Error -99740.525
Maximum Error 752120.518
Mean Error 0.0
Mean Absolute Error 106964.209
Standard Deviation 160172.667
Linear Correlation 0.01
Occurrences 34,423
OK
Favorites
Sources
Record Ops
Field Ops
Graphs
Table
Matrix
Analysis
Data Audit
Transform
Statistics
Means

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.

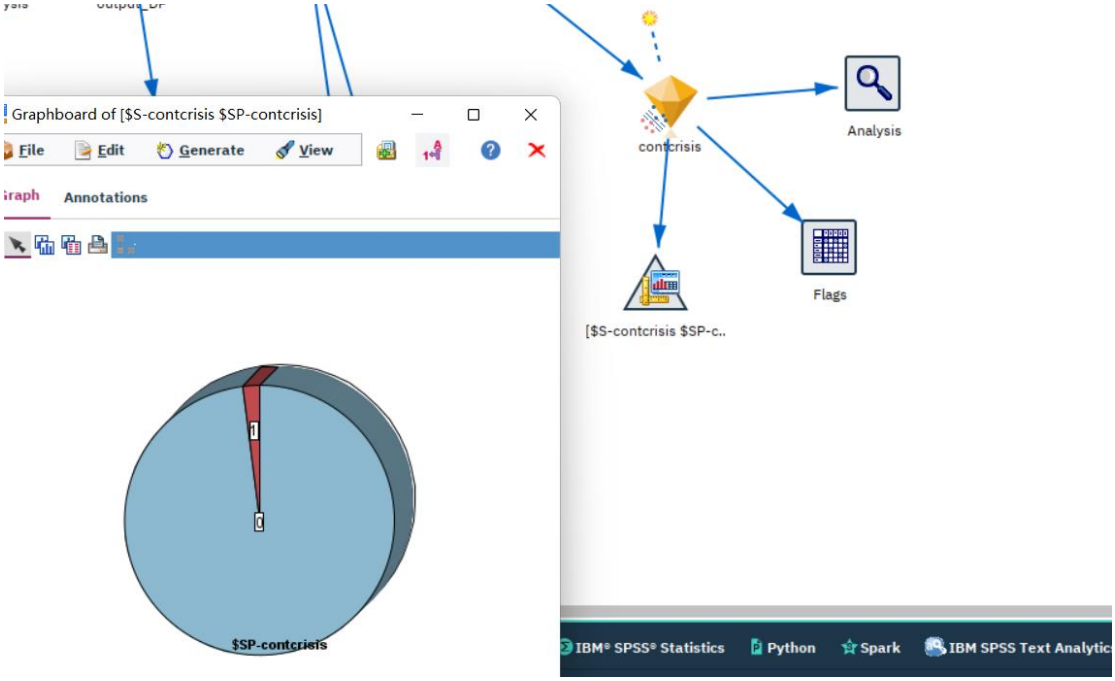
Linear Relationships:
A continuous decline in GDP Growth Rate combined with rising Unemployment Rate might linearly correlate with the likelihood of a banking crisis.

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: Churn (1 if there is a banking crisis, 0 if not)

Model: SVM

Model results:

Individual Models

Comparing $S-contcrisis with contcrisis

'Partition'	1_Training		2_Testing	
Correct	27,481	100%	6,942	100%
Wrong	0	0%	0	0%
Total	27,481		6,942	

Coincidence Matrix for $S-contcrisis (rows show actuals)

'Partition' = 1_Training	0	1
0	26,941	0
1	0	540

'Partition' = 2_Testing	0	1
0	6,798	0
1	0	144

Performance Evaluation

Accuracy: 100%

AUC: 1.0
Gini: 1.0
Performance Evaluation:

8.4 Assess and evaluate results, models, and patterns

Performance Evaluation

'Partition' = 1_Training	
0	0.02
1	3.93
'Partition' = 2_Testing	
0	0.021
1	3.876

Confidence Values Report for $SP-contcrisis

'Partition' = 1_Training: $S-contcrisis always correct. No confidence report

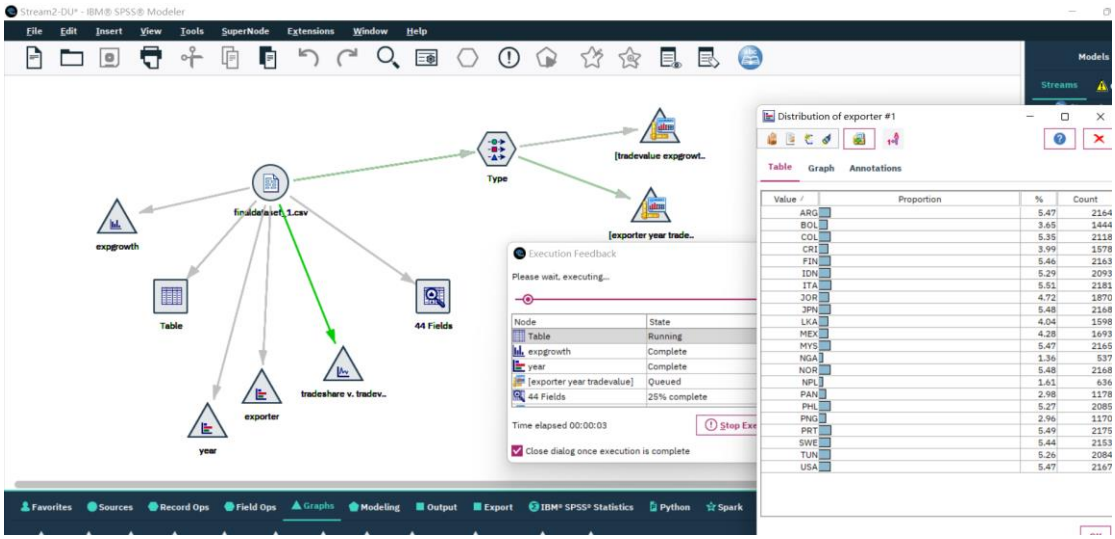
'Partition' = 2_Testing: $S-contcrisis always correct. No confidence report

Evaluation Metrics

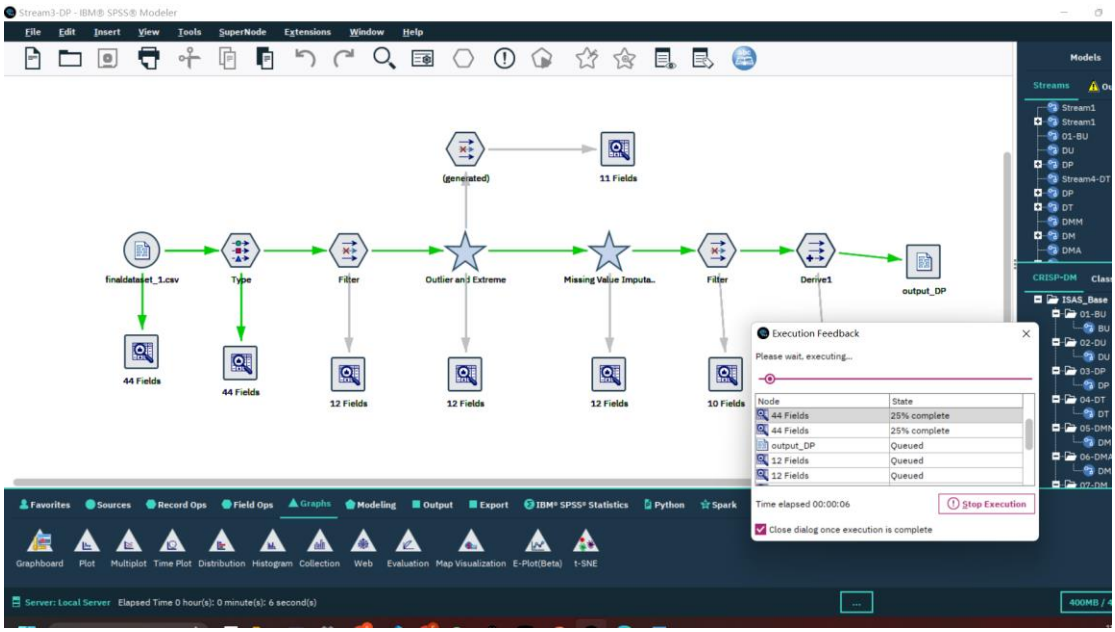
'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$S-contcrisis	1.0	1.0	1.0	1.0

8.5 iterations

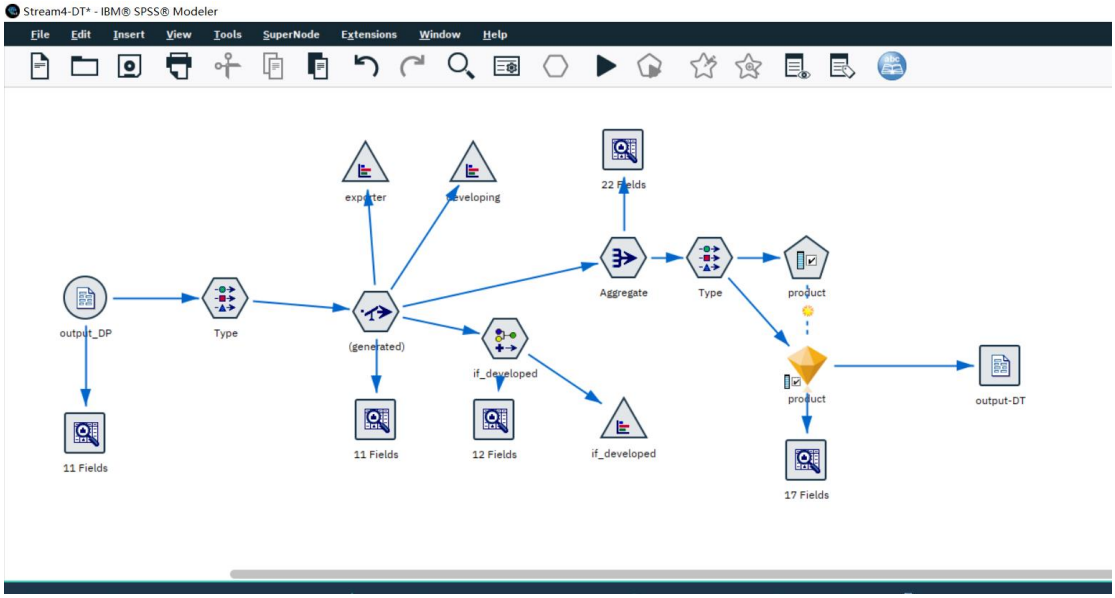
Data Understanding:



Data Preparation:

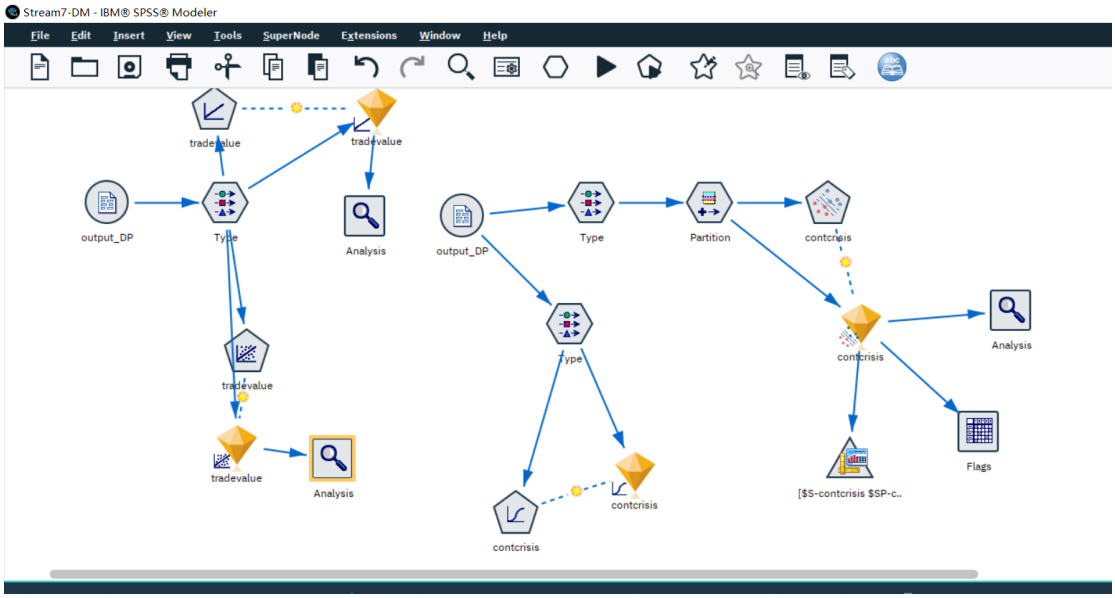


Data Transforming:

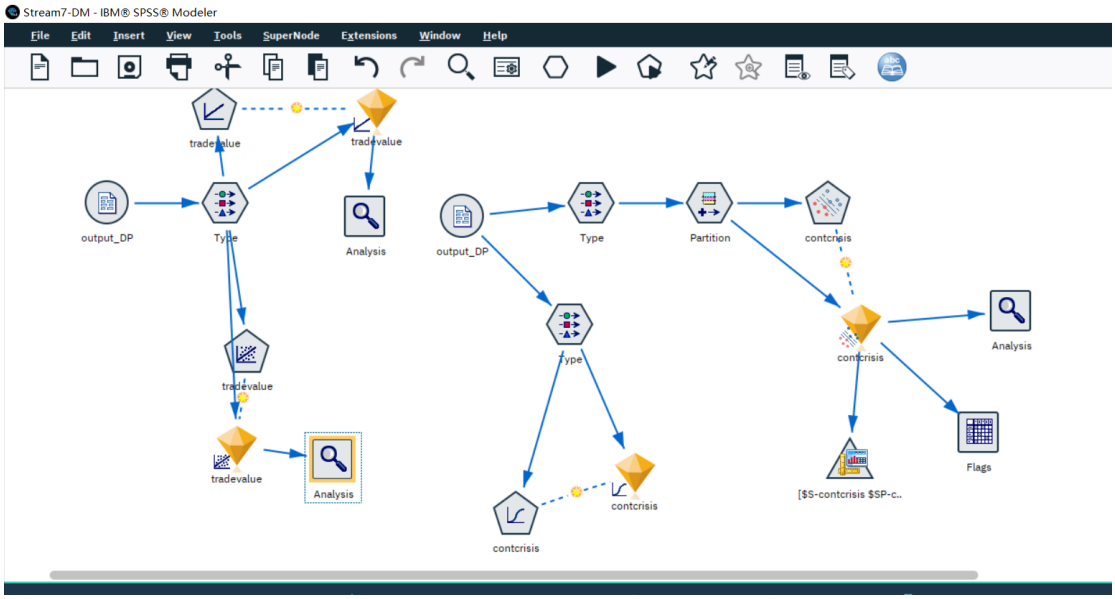


Data Mining Methods & Algorithms:

ITERATION 2, ISAS

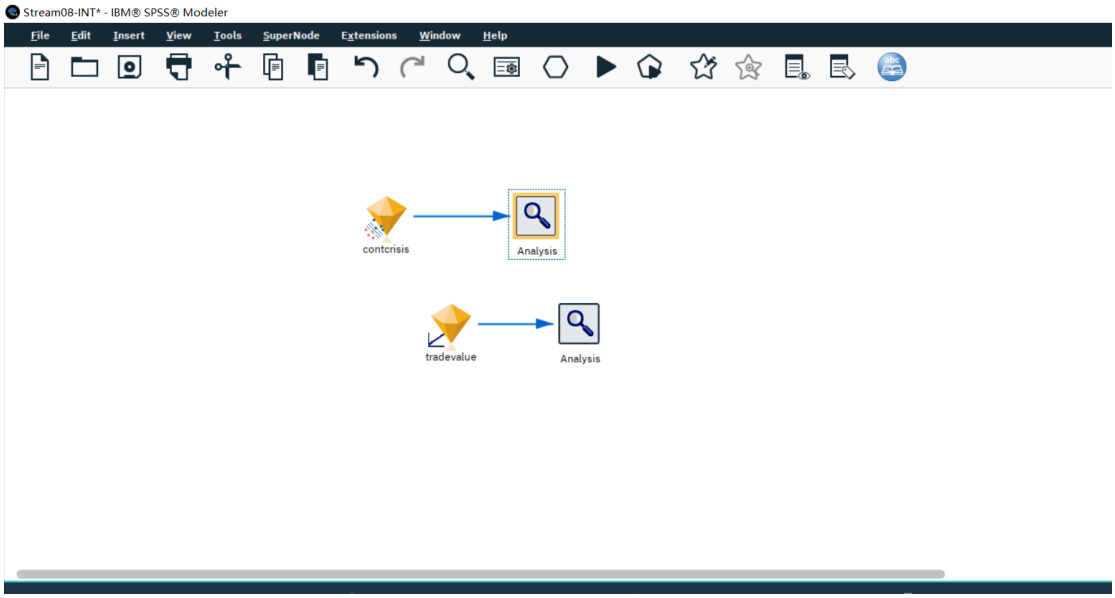


Data Mining:



Interpret:

ITERATION 2, ISAS



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.