



Akai Kaeru

AK Analyst User Guide

Version 1.0.5 (March. 2, 2022)

Table of Contents

1. [Introduction](#)
2. [Launching the Workspace](#)
3. [Landing Page](#)
4. [AK Analyst Overview](#)
5. [AK Analyst Actions](#)
6. [Load Data Action](#)
7. [Load Data Lake Action](#)
8. [Data Transformer Action](#)
9. [Aggregate Action](#)
10. [Join Action](#)
11. [AK Pattern Mining Action](#)
12. [Rolling Regression Action](#) (Experimental)
13. [AK Pattern Browser Action](#)
14. [Visual Explorer Action](#)

Introduction

The AK Analyst is a No-code / Low-Code platform for data science geared toward analytics that involve complex high-dimensional data. The platform features an easy-to-use GUI that facilitates the creation of a wide variety of dedicated analytics pipelines. The pipelines are constructed by simply dragging, dropping, and connecting action widgets that map into a rich set of specially designed data analysis tools.

The platform allows you to load, clean, and transform data with visual feedback. It provides a wide set of innovative analysis components that can extract relations in high-dimensional data and perform predictive analytics with concise explanations of model behavior.

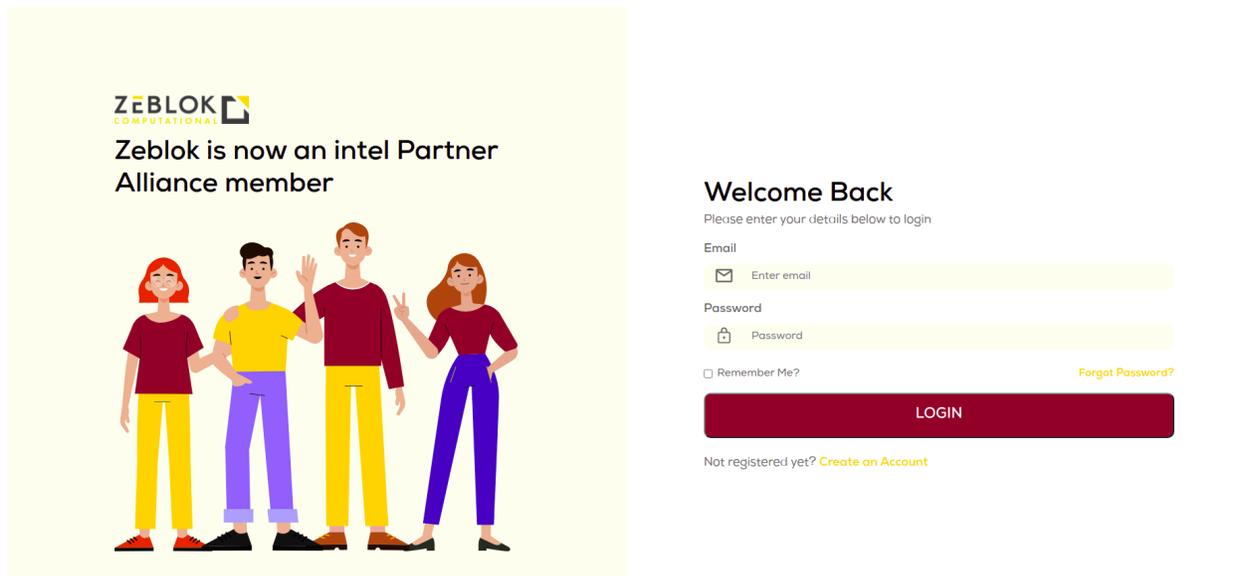
In this guide, we will introduce the AK Analyst platform, explain each action supported by the platform, and how these actions can be combined to create solution-focused data analysis pipelines.

For further information please contact us at info@akaikaeru.com

Launching the Workspace

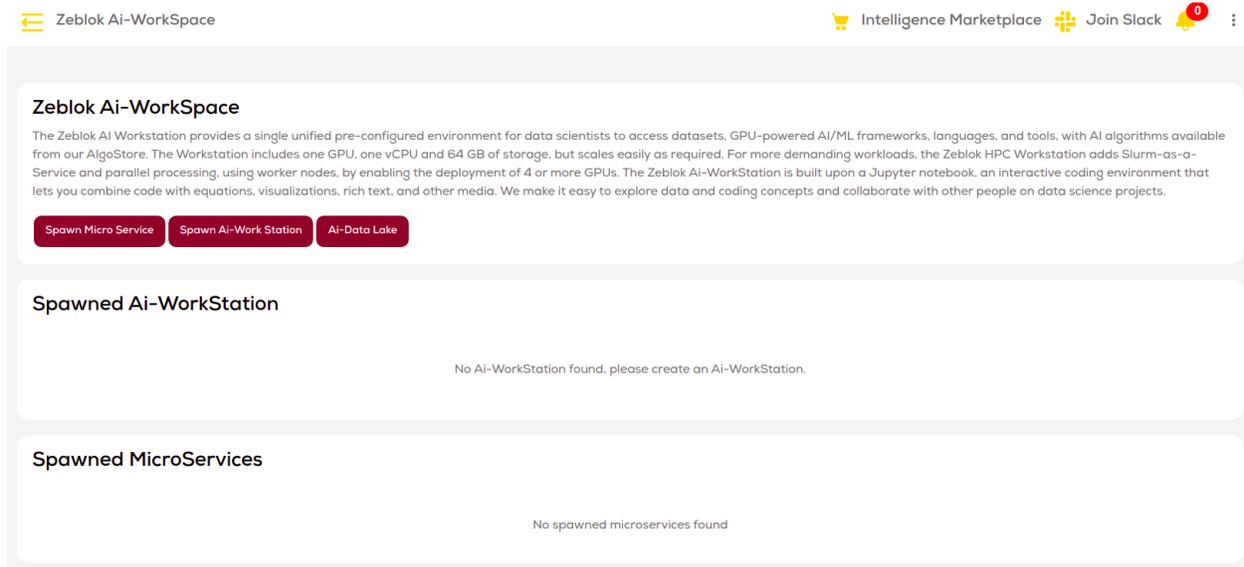
The following instructions are a step-by-step guide for launching the AI Rover workstation through the Zeblok computational AI platform.

Login to the Zeblok computational platform at <https://app.zbl-aws.zeblok.com/>



The image shows a composite of two parts of the Zeblok login interface. On the left, a yellow banner features the Zeblok logo and the text "Zeblok is now an intel Partner Alliance member" above an illustration of four diverse people. On the right, a white login form titled "Welcome Back" prompts the user to enter their email and password. It includes a "Remember Me?" checkbox, a "Forgot Password?" link, and a prominent red "LOGIN" button. Below the button, there is a link for "Not registered yet? Create an Account".

On the homepage click on the “Spawn Micro Service” button.



The image displays the Zeblok Ai-WorkSpace homepage. At the top, there is a navigation bar with a home icon, the text "Zeblok Ai-WorkSpace", and links for "Intelligence Marketplace", "Join Slack", and a notification bell. The main content area is divided into three sections: "Zeblok Ai-WorkSpace" with a descriptive paragraph and three buttons ("Spawn Micro Service", "Spawn Ai-Work Station", "AI-Data Lake"); "Spawned Ai-WorkStation" with a message "No Ai-WorkStation found, please create an Ai-WorkStation."; and "Spawned MicroServices" with a message "No spawned microservices found".

Select Akai Kaeru Explainable AI MicroService

Select Your MicroService... Select Your DataCenter... Select Your Plan... Select Your Namespace... Configure Your MicroService

Search...

 mongoDB testing_1: 1649446540 Select MicroService	 Akai Kaeru: AK Analyst Select MicroService	 OVMS Select MicroService	 jupyter HPC Select MicroService
 intel Open MPI, Horovod*, and Jupyter* Notebook Select MicroService	 intel TensorFlow* & oneDNN with Jupyter* Notebook Select MicroService	 OpenVINO DEEP LEARNING WORKBENCH Import a model Perform baseline inference Intel DL workbench Select MicroService	 mongoDB MongoDB Select MicroService

Select Zeblok AWS Ohio DataCenter

Select Your MicroService... Akai Kaeru Explainable AI... Select Your DataCenter... Select Your Plan... Select Your Namespace... Configure Your MicroService

 amazon web services Zeblok AWS Ohio Ohio, US Category: Enterprise Select DataCenter	 Zeblok Azure Redmond, US Category: Enterprise Select DataCenter	 Advantech Oregon, US Category: Edge Select DataCenter
---	--	---

Select Plan

Select Your MicroService.....
Akai Kearu Explainable AI

Select Your DataCenter.....
Zeblok AWS Ohio

Select Your Plan.....

Select Your Namespace.....

Configure Your MicroSer.....



C1-0-1-1-1-MS 

0 GPU

1 vCPU

1 GB RAM

1 GB Storage

Select Plan

Select Namespace as akaikaerupublicaws

Select Your MicroService.....
Akai Kearu Explainable AI

Select Your DataCenter.....
Zeblok AWS Ohio

Select Your Plan.....
C1-0-1-1-1-MS

Select Your Namespace.....
akaikaerupublicaws

Configure Your MicroSer.....

Select Namespace

akaikaerupublicaws

Click Next

Select Your MicroService.....
Akai Kearu Explainable AI

Select Your DataCenter.....
Zeblok AWS Ohio

Select Your Plan.....
C1-0-1-1-1-MS

Select Your Namespace.....
akaikaerupublicaws

Configure Your MicroSer.....

Ports

5000

Environment Variables

Example: FILE_UPLOAD_PATH=/public/uploads:AWS_KEY=xxctedgaj

Arguments

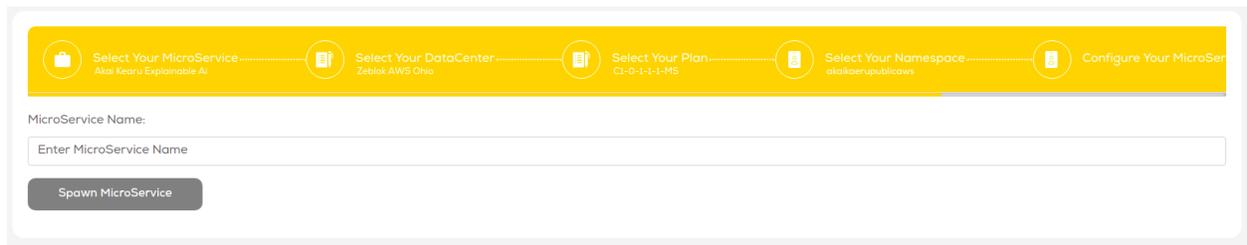
Example: --config_path=/models/name;--port=9000

Command

If your run command looks like this: ["bash", "-c", "token='password' "] enter bash. -c. token = "password"

NEXT

Enter a descriptive name and click “Spawn MicroService”



Select Your MicroService
Akai Kearsu Explainable AI

Select Your DataCenter
Zeblok AWS Ohio

Select Your Plan
CI-0-1-1-1-MS

Select Your Namespace
akaikearsupublicaws

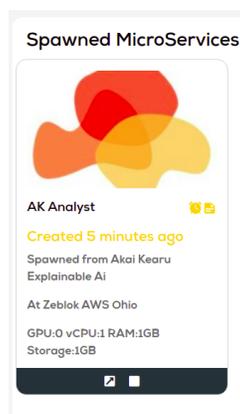
Configure Your MicroSer

MicroService Name:

Enter MicroService Name

Spawn MicroService

Back on the homepage you will see the spawned MicroService. Click [🔗](#) to open.



Spawned MicroServices

AK Analyst 

Created 5 minutes ago

Spawned from Akai Kearsu Explainable AI

At Zeblok AWS Ohio

GPU:0 vCPU:1 RAM:1GB Storage:1GB

Landing Page

Upon launching the AK analyst microservice the initial page that comes up will be the loading page. From this page you will be able to (1) start a new project, (2) load an existing project, (3) upload a project, or (4) launch a pre-built sample pipeline. Clicking on any of these icons will bring up the AK Analyst.



The AK Analyst loading page

AK Analyst Overview

The AK Analyst platform opens with our easy-to-use drag-and-drop pipeline designer by which you will construct your analytics pipelines. An example of this interface with a complete pipeline already set up is shown below. It consists of four main panels and components:

Actions: Actions are components that allow you to perform tasks on your data. They must be dragged into the pipeline canvas and connected to each other to form an analytics pipeline. Connections can be made by holding the mouse down over an action's output port (i.e. right ►) and releasing it after it snaps to another action's input port (i.e. left ►).

Pipeline Canvas: The pipeline canvas is where pipelines reside and can be edited.

Action Configuration: This panel is used to configure each action in the pipeline. Clicking on an action will load its configuration in the panel where it can be edited according to your needs.

Action Output: This panel displays the output for each panel. If the output is a data frame you can download that data frame from this panel.

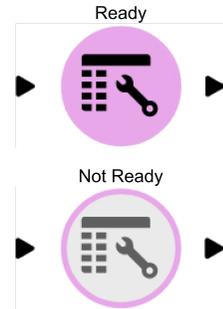
The screenshot displays the AK Analyst interface with four main panels highlighted by green boxes and labels:

- Action Configuration:** Shows settings for data selection (File: sp500_stocks.csv), encoding (utf-8), column delimiter, and line delimiter. It also includes a 'Raw Data Preview' section with a list of stock symbols and their corresponding data points.
- Actions:** A vertical sidebar containing icons for various actions, categorized into 'Analysis & Prediction' and 'Visual Derivation'.
- Pipeline Canvas:** A central workspace where a pipeline is visualized. It starts with a data source icon (sp500_stocks.csv) connected to a central action icon, which then branches into three parallel paths, each ending with a visualization icon.
- Action Output:** A panel at the bottom right showing a 'Data Preview' table with columns for various financial metrics and a 'Download' button.

asset_turnover	dividend	eps	liquidity	on_bal_vol	operating_leverage	picf	p/e	payout_ratio	peg	price	price/book	price/cashflow	price/earnings	price/earnings/growth	profit_m
0.42	1.82	1.93	1.03	-7254700.00	2.09	73.47	113.18	0.85	-450.82	89.95	1.51	23.93	46.61	-1.55	7.01
0.91	1.23	2.70	1.33	-75216000.00	0.73	81.32	-37.14	0.19	-720.14	42.19	1.46	10.90	15.63	-2.79	7.60
0.93	0.00	19.47	0.63	7286700.00	-44.15	-72.61	-80.92	0.00	-99.66	56.99	0.75	3.58	2.93	0.01	19.31

AK Analyst Actions

Actions in the AK Analyst are the core components that allow you to perform tasks on your data. Actions are represented by colored circles with an icon at the center. They also have input and/or output ports to connect them to other actions. The color of an action indicates its category. An action's color changes to gray when it is not ready which indicates that it may not have an input or is not configured correctly.



There are four categories of actions:

- **Input/Output** – Actions for moving data in and out of the platform
- **Data Manipulation** – Actions for performing data manipulations such as column transformations, row aggregations, various cleaning operations, and dataset merging.
- **Analysis and Prediction** – Actions for performing an analysis on the data and for building predictive models. This includes our proprietary AK Pattern Mining algorithm.
- **Visual exploration** – Actions that allow you to visually explore the data at every stage of your pipeline.

Load Data



Load Data Lake



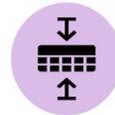
Transform Data



Merge Tables



Aggregate Rows



AK Pattern Miner



Rolling Regression



AK Pattern Browser



Visual Data Explorer



Load Data Action



The load data action is one of the first actions that you will use. It allows you to import data into the AK Analyst. It currently supports the loading of structured data files such as delimited text files (e.g. CSV, TSV).

The configuration panel (shown on the right) for this action allows you to specify various file loading parameters:

Encoding type: Encoding to use when reading/writing (e.g. utf-8).

Column delimiter: Delimiter to use to indicate the end of a column.

Line delimiter: Delimiter to use to indicate the end of a row.

Header row: The row at which the header is located.

Skip rows after header: The number of rows to skip after the header.

Escape character: One-character string used to escape other characters.

Comment character: One-character string used to indicate the remainder of the line should not be parsed.

Thousands separator: Character to recognize as the thousands separator.

Decimal character: Character to recognize as decimal point.

NA Values: Strings in the file that should be treated as NaN/NA

Skip Empty Rows: A flag that if set tells the loader if an empty row should be skipped when reading in the file. If it is not set, empty rows are read in with NaN/NA values.

ACTION CONFIGURATION

Select Data

Select File

Options

Encoding

Column Delimiter

Line Delimiter

Advanced Options

Header Row

Skip Rows after Header

Escape Character

Comment Character

Thousands Separator

Decimal Character

NA Values

Skip Empty Rows

Load Data Lake Action



The load data lake action enables importing data from a data lake hosted in the cloud. The configuration panel is very similar to the load data action – with the only difference being the addition of “Datalake Credentials” which is needed to access data stored in the cloud.

The “Datalake Credentials” include:

IP Address: The IP address of the data lake to access.

Username: The data lake username.

Secret Key: The secret key associated with the data lake.

Bucket: The bucket to access.

File Path: File path within the data lake bucket to import into the pipeline.

The remaining “Options” and “Advanced Options” are identical to the load data action.

ACTION CONFIGURATION

Datalake Credentials

IP Address	127.0.0.1:9000
Username	johndoe
Secret Key	*****
Bucket	datalake
File Path	johndoe/sp500_data.csv

Options

Encoding	utf_8
Column Delimiter	,
Line Delimiter	

Advanced Options

Data Transformer Action



The data transformer action provides a visual interface for transforming and cleaning the data. Upon clicking on the data transformer icon, the Action Configuration panel will update to display a button called “Launch Data Transformer” (see left). There is also an option to use sampling which, when enabled, will use a random sample of the data within the data transformer.

Upon clicking the “Launch Data Transformer” button, the main display will navigate to the data transformer visual interface. This interface initially contains two panels – the “Attributes” panel (left) and the “Data Preview” panel (right).

ACTION CONFIGURATION

Sample Options

Use Sample # Samples 1000

Launch Data Transformer

Transformation Summary

No transforms applied.

ATTRIBUTES

- symbol ▲ Nominal
- asset_turnover Numeri...
- liquidity ▲ Numeri...
- dividend Numeri...
- eps Numeri...
- return Numeri...
- on_bal_vol Numeri...
- operating_leverage ▲ Numeri...
- payout_ratio Numeri...
- price Numeri...
- price/book ▲ Numeri...
- price/cashflow ▲ Numeri...
- price/earnings Numeri...
- price/earnings/gr... Numeri...

Add Derived Attribute Transform Multiple

DATA PREVIEW

Select an attribute from the left to view details.

symbol	asset_turnover	liquidity	dividend	eps	return	on_bal_vol	operating_leverage	payout_ratio	price	price/book	price/cashflow	price/earnings	price/earnings/growth
JBHT	1.70	1.61	1.08	3.66	-0.08	-4200100.00	37.00	0.23	77.61	2.89	10.29	21.21	1.34
CPB	1.00	0.75	2.48	2.21	8.86	4629500.00	3.62	0.56	50.30	2.98	13.31	22.76	-1.55
TAP	0.42	1.03	1.82	1.93	-2.14	-7254700.00	2.09	0.85	89.95	1.51	23.93	46.61	-1.55
VRTX	0.41	2.78	0.00	-2.31	-26.58	2449000.00	-0.42	0.00	130.36	15.76	-85.91	-56.43	2.13
BWA	0.91	1.33	1.23	2.70	-31.14	-7521600.00	0.73	0.19	42.19	1.46	10.90	15.63	-2.79
UAL	0.93	0.63	0.00	19.47	-14.73	7286700.00	-44.15	0.00	56.99	0.75	3.58	2.93	0.01
WM	0.63	0.93	3.02	1.65	-0.21	1559200.00	1.49	0.93	51.00	1.29	9.26	30.91	-0.76
FIS	0.25	1.49	1.66	2.19	0.12	4639900.00	-4.77	0.47	62.46	0.74	15.63	28.52	-4.19
GT	1.00	1.24	0.73	1.12	-11.14	-19155200.00	-1.94	0.22	34.18	0.80	5.46	30.52	-0.35
GS	0.05	0.00	1.36	12.14	-11.42	-18675800.00	0.00	0.21	187.75	0.00	12.16	15.47	-0.54
NWL	0.81	1.25	1.73	1.29	-11.50	-18509800.00	-0.17	0.59	43.85	2.23	20.87	33.99	-7.65
GE	0.24	0.00	3.25	-0.61	-8.06	171119500.00	0.00	-1.51	28.28	0.00	14.09	-46.37	0.33
GD	0.98	1.17	0.00	9.08	-2.21	-13387400.00	3.72	0.00	142.00	2.33	18.25	15.64	0.70
VAR	0.86	1.83	0.00	4.09	-4.46	1463300.00	-2.40	0.00	71.57	3.22	15.19	17.50	2.58
GM	0.75	1.09	4.20	5.91	-11.47	-65664000.00	-64.31	0.23	32.85	0.42	4.35	5.56	0.02
ALK	0.86	0.92	1.02	6.56	-12.69	-3734400.00	8.15	0.12	78.70	2.14	6.37	12.00	0.25
MAS	1.26	1.33	1.25	1.02	-6.85	-16986500.00	-0.99	0.36	29.70	3.18	14.44	28.12	-0.51
MAR	2.38	0.43	0.00	3.15	-8.38	-4635400.00	3.29	0.00	70.20	6.57	13.10	22.28	0.93
MAT	0.87	1.94	6.69	1.08	-1.51	37988500.00	3.24	1.41	22.71	1.58	10.58	21.03	-0.82
SNI	0.45	1.77	1.67	4.66	12.48	-2722300.00	0.85	0.20	55.14	1.25	8.78	11.83	0.55
XRAY	0.61	2.51	0.47	1.76	-1.35	946100.00	1.86	0.16	61.32	2.19	17.30	34.84	-1.63
SIG	0.91	3.29	0.57	4.75	-2.12	-1010000.00	0.03	0.15	127.20	2.04	35.93	26.78	6.43
XYL	0.78	2.44	0.02	1.87	-2.41	-1271700.00	0.45	0.00	36.64	1.73	14.28	19.59	8.96
TSN	1.80	1.52	0.86	2.95	1.84	23947200.00	5.12	0.14	49.39	0.53	4.01	16.74	0.68
AFL	0.18	0.00	2.51	5.85	-5.66	-13305200.00	0.52	0.27	62.93	0.00	4.00	10.76	-1.08

Rows per page: 25 ▾ 1-25 of 100 < >

100.00% of the data remains after filtering.
Attribute Count: 27 (27) | Data Item Count: 374 (374)

Initial view of the data transformer visual interface

Clicking on an attribute will update the interface to show more information about the clicked attribute. Specifically, the interface will replace the “Data Preview” panel with 3 more panels – “Attribute Details”, “Notifications”, and “Transforms”.

The screenshot displays the following components:

- ATTRIBUTES:** A list of attributes including 'symbol', 'asset_turnover', 'liquidity', 'dividend', 'eps', 'return', 'on_bal_vol', 'operating_leverage', 'payout_ratio', 'price', 'price/book', 'price/cashflow', 'price/earnings', and 'price/earnings/gr...'. The 'liquidity' attribute is selected and highlighted.
- ATTRIBUTE DETAILS:**
 - Name: liquidity
 - Type: Index, Numerical, Nominal, Ordinal (with 'Numerical' selected)
 - Summary Statistics:

Min	25%	50%	75%	Max	Mean	Std.
0.2	1	1.5	2.2	11.2	1.8	1.3
 - Histogram showing the distribution of 'liquidity' values.
 - Buttons: Remove outliers, Remove Filter, Apply Filter, Filter, Clamp, Normalize, Log, Custom, Make Duplicate, Delete.
- NOTIFICATIONS:**
 - Warning: Missing values present (15.24%).
 - Imputation options: Drop, Zero, Mean, Interpolate, Pad.
 - Buttons: Resolve, Apply.
- TRANSFORMS:**
 - price: Log (Base: 1...)
 - eps: Clamp (LB: 0.00, UB: 49.45)
 - rsi: Normalize (LB: 0.00, UB: 1.00)
 - Buttons: Show All, Hide, Disable / Remove Transform (X icon).

Visual interface after clicking on the “liquidity” attribute

The “Attribute Details” panel contains relevant information about the selected attribute including summary statistics and a histogram showing the distribution of the selected attribute. From this panel various transformations can be applied via the control panel at the bottom. The data type of the selected attribute can also be changed.

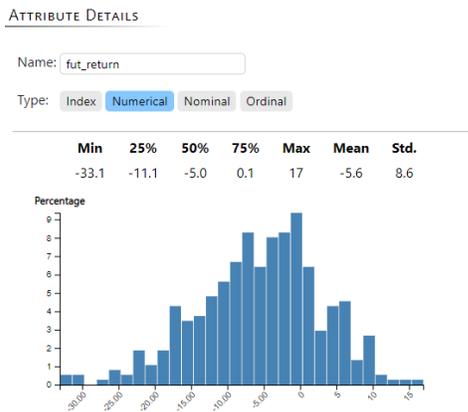
The “Notifications” panel contains information and warnings about the selected attribute. Information is marked with **i** and indicates potentially useful information (e.g. correlated attributes). The **⚠** icon indicates a potentially serious issue with the selected attribute. This can include missing values (e.g. NaNs), collinear attributes, or high cardinality for nominal attributes. The “Resolve” button to the right of the warning provides a hint on how to deal with these issues. Note that these are merely suggestions and it is not required that all warnings be addressed.

Attribute Details

The AK Analyst supports 4 attribute types - Numerical, Nominal, Ordinal, and Index. Based on the attribute type the “Attribute Details” panel appearance changes.

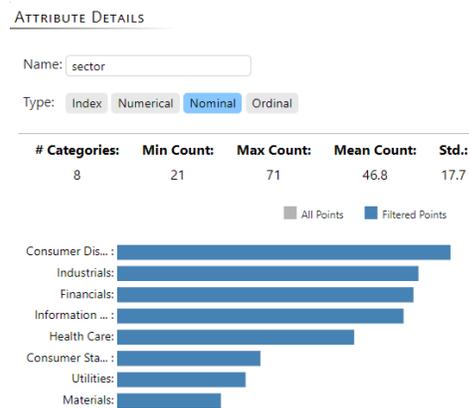
Numerical Attributes

For numerical or quantitative attributes, the AK analyst reports summary statistics values and a histogram showing the distribution of the data.



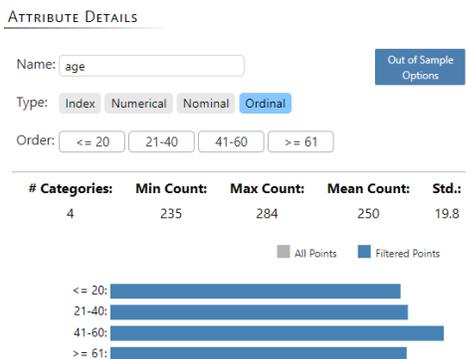
Nominal Attributes

For nominal or categorical attributes, the AK analyst reports summary statistics values and a bar chart showing the count of each category. The bars are ordered in descending order of the count.



Ordinal Attributes

Ordinal attributes, are identical to the nominal attributes but here the user has a control to specify the order of the attributes. The bars are ordered based on the user-specified ordering.



Index Attributes

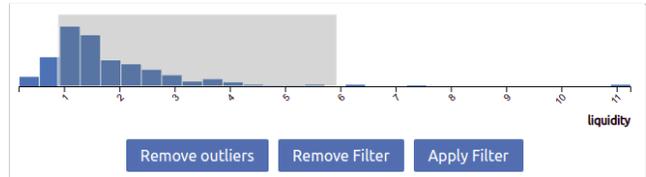
Index attributes are attributes that are used to index the rows of the data set and should have unique values for each row. Thus, we do not show any summary for these attributes.



Numerical Attribute Transforms

The following transformations are available for “Numerical” attributes.

Filter: A filter transformation can be applied by clicking and dragging the edges of the gray box. This will update the histogram to show only those values within the gray box. Clicking on “Remove outliers” will update the gray box to automatically exclude outliers. In both cases Clicking “Apply Filter” applies the filter transform and updates the data.



Clamp: The clamp transform replaces any value below/above “Clamp Min”/“Clamp Max” with the value at “Clamp Min”/“Clamp Max”.

Clamp Min: Clamp Max:

Normalize: The normalize transform normalizes the data to between “Lower Bound” and “Upper Bound.”

Lower Bound: Upper Bound:

Log: The log transform applies a log transformation with the specified base.

Base:

Custom: The custom transform provides a way to enter python code to perform custom transformations not provided elsewhere by the interface.

Enter custom python operation:

liquidity =
x[x<10] = x[x<10] + y[x<10]
Here x and y should be attribute names

Nominal Attribute Transforms

The following transformations are available for “Nominal” attributes:

Filter: The filter transform allows for inclusive / exclusive filtering by category name. In the figure on the right, for example, clicking on the “Apply” button filters the dataset to only include those data items whose *sector* values are “Utilities” or “Materials.”



Utilities: [Orange bar]

Materials: [Orange bar]

Consumer Dis...: [Blue bar]

Industrials: [Blue bar]

Financials: [Blue bar]

Information ...: [Blue bar]

Health Care: [Blue bar]

Consumer Sta...: [Blue bar]

Filter Type: Include | v

Categories to filter by: Utilities x Materials x | v

Apply

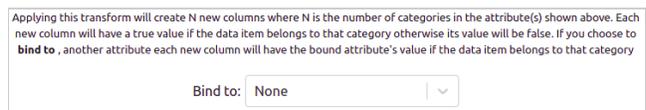
Replace: The replace transform provides a way of merging categories into another larger class of categories (e.g. replacing a set of infrequent categories with an “Other” category).



New Category Name: Other

Categories to be merged: Select... | v

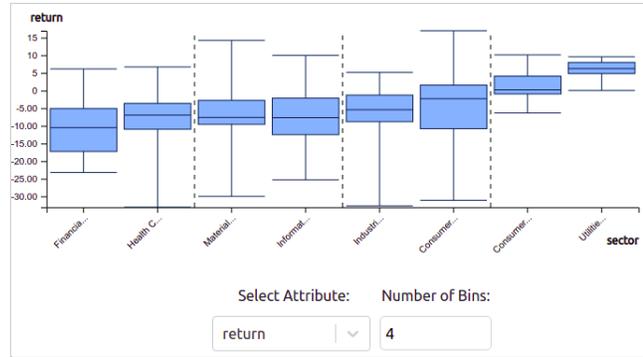
One-Hot Encode: One-hot encoding creates a new set of attributes (i.e. 1 for each category level). If “Bind to” is set to “None”, then each newly created attribute will be binary (i.e. 1 if the data item belongs to that category and 0 otherwise). If “Bind to” is set to another attribute, then the newly created attribute will take on the bound attribute’s value whenever the data item belongs to that category and NaN otherwise.



Applying this transform will create N new columns where N is the number of categories in the attribute(s) shown above. Each new column will have a true value if the data item belongs to that category otherwise its value will be false. If you choose to **bind to**, another attribute each new column will have the bound attribute's value if the data item belongs to that category

Bind to: None | v

Rank: The rank transform provides a way of ranking the categories based on another numerical attribute and bin the categories based on this ranking. In the figure on the right, for example, the *sector* categories are ranked / ordered based on *return*. Consecutive categories are grouped (indicated by dashed lines) into 4 bins (e.g. “Financials” and “Health Care” are grouped in the same bin). This transformation creates a new attribute called `sector_rank4_return` with categories “rank_0”, “rank_1”, etc. based on this binning.



Custom: The custom transform provides a way to enter python code to perform custom transformations not provided elsewhere by the interface.

```
sector = x**2
x[x<10] = x[x<10] + y[x<10]
Here x and y should be attribute names
```

Ordinal Attribute Transforms

Attributes that have “Ordinal” data types contain the same set of transformations as “Nominal” attributes. The main difference is the addition of a drag-and-drop mechanism for setting the order of the categories. The order is specified from left to right (e.g. in the left image “Female” is rank 1 and “Male” is rank 2). For subsequent actions in the pipeline, ordinal categories will be replaced with their integer rank internally (e.g. in the pattern mining action, instead of Female / Male there will be 1 and 2).

ATTRIBUTE DETAILS

Name:

Type: Index Numerical Nominal Ordinal

Order:

[Out of Sample Options](#)

If the data transformer was launched with sampling turned on, then a button will appear on the top right to handle categories that are out of sample. Clicking on this button will trigger a pop-up that allows you to enter additional categories not included in the current sample (see below). In subsequent actions, any out-of-sample categories not specified will be appended to the current ranking in alphabetical order.

OUT OF SAMPLE OPTIONS

You are working with a sample / subset of your dataset. Sampling improves run-time performance. You may see differences in the summary statistics for the full dataset. Some categories in nominal / ordinal columns may not have been sampled. To handle issues related sampling please use the options below:

Sampled Categories:

×

[Add Unsampld Category](#)

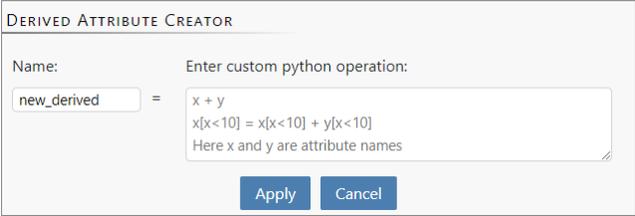
Any additional categories not included in this sample will be appended to the list above in alphabetical order.

[Set Ordering](#) [Close](#)

Out of sample options pop-up

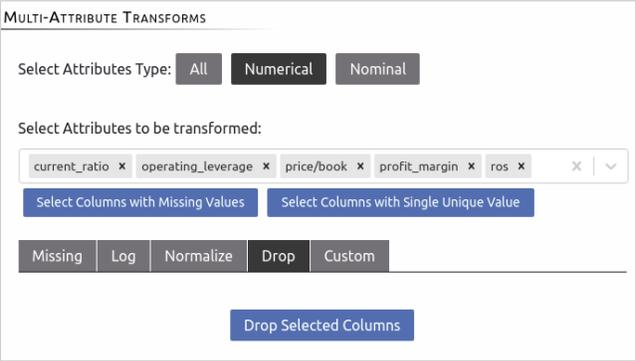
Derived Attribute Transform

Derived attributes can also be created by clicking on the “Add Derived Attribute” button. This brings up the “Derived Attribute Creator” pop-up. Similar to the “Custom” transform, python code can be entered to create a new attribute based on some combination of existing attributes.

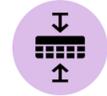


Multi-Attribute Transform

Clicking on the “Transform Multiple” button brings up the “Multi-Attribute Transform” pop-up. This is a convenience feature which allows for applying a single transformation to multiple attributes. It also includes buttons for selecting all columns with missing values or single unique values – allowing the user to handle them all simultaneously.



Aggregate Action



The aggregate action allows you to combine multiple rows in a table into a single row. To use this action drag it into the canvas and connect it to any data source.

To be able to aggregate rows you must select a key column that will contain repeating keys that indicate which rows should be combined. You also must select aggregation functions for the remainder of the columns that tell the software how to combine the rows. An example is shown below.

The screenshot shows the AGGREGATOR interface. The 'Select Column to Aggregate Over' dropdown is set to 'state'. The 'Select Aggregation Functions for Columns' section shows settings for 'fips' (Drop), 'county' (One-Hot Encoding), 'num_driving_de...' (Mean), 'num_deaths' (Mean), 'years_of_po...' (Mean), 'quartile' (Mean), 'ypil_rate_a...' (Mean), and another 'ypil_rate_a...' (Mean). The 'Data Preview' table shows columns for '20th_percentile_income', '80th_percentile_income', and '95percent_ci_high_10' through '95percent_ci_high_15'. The table contains 20 rows of data. At the bottom, there are 'Apply' and 'Done' buttons.

20th_percentile_income	80th_percentile_income	95percent_ci_high_10	95percent_ci_high_11	95percent_ci_high_12	95percent_ci_high_13	95percent_ci_high_14	95percent_ci_high_15
32967.00	145096.00	33.37	22.40	5.11	69.62	13.11	30.22
15758.00	79880.00	28.25	72.00	19.43	36.58	47.33	51.86
24106.00	110322.00	31.37	16.50	3.58	73.54	18.79	35.70
13274.00	73248.00	41.32	49.60	16.33	42.80	41.50	53.46
18925.00	91734.00	35.03	47.90	16.23	44.99	32.38	44.15
23317.00	96679.00	67.57	25.40	9.85	76.73	18.55	31.35
13452.00	94403.00	48.27	3.60	7.56	90.29	15.71	26.81
13116.00	55403.00	53.33	30.20	22.76	58.96	51.43	81.32
17319.00	81147.00	28.31	40.90	8.98	63.96	30.20	42.88
30015.00	120732.00	48.27	25.90	15.55	77.93	12.42	26.53
27517.00	117877.00	42.22	20.40	8.37	63.62	15.57	36.93
32158.00	121695.00	23.85	18.00	10.52	70.98	10.53	18.69
25313.00	80417.00	68.30	NaN	16.49	100.00	30.10	23.41
17843.00	80429.00	52.49	59.20	8.45	53.15	44.06	44.33
25378.00	114959.00	34.52	25.70	5.46	66.23	20.20	35.03
48666.00	214239.00	18.82	3.90	6.52	83.77	6.31	16.42
16734.00	93305.00	13.55	77.70	23.55	42.99	39.31	45.83
26892.00	122800.00	35.29	19.90	7.59	66.48	18.14	33.34
22750.00	95040.00	37.85	44.60	12.57	69.64	25.08	50.59
15288.00	95733.00	59.58	44.90	11.30	48.72	36.64	62.09

For numerical variables, you can choose between computing the mean, max, min, standard deviation, variance, and the sum total of all the rows. You can also choose to just select the first or last value or use the count (i.e. number of rows) belonging to a key with or without NaN values in the numeric column.

For nominal variables, you can choose between selecting the most frequent value or performing a one-hot encoding and aggregating the one-hot encoded columns. With one-hot encoding, you can choose to apply numerical aggregations (e.g. mean, max, etc.) to them as well as bind other attributes to them and perform aggregations (see data transformer section for more information on binding). An example is shown on the right

where the *county* attribute is one-hot encoded and two aggregations are performed. The first aggregation does not bind to any other attribute (i.e. it is a standard one-hot encoding) and aggregates by computing the mean for each encoded column. The second aggregation binds to the *num_driving_deaths* attribute (i.e. each encoded column takes on the value of *num_driving_deaths* where it is 1 and NaN where it is 0) and aggregates by taking the mean (excluding NaNs).

The image shows a configuration interface for data aggregation. It features two rows of settings for the 'county' attribute. The first row has 'county' selected, 'One-Hot Encoding' as the aggregation type, 'None' as the bind target, and 'Mean' as the aggregation function. The second row has 'county' selected, 'num_driving_dea...' as the bind target, and 'Mean' as the aggregation function. A blue '+' button is at the bottom, and an 'X' button is on the right side of the second row.

Attribute	Aggregation Type	Bind	Aggregation Function
county	One-Hot Encoding	None	Mean
county	num_driving_dea...	num_driving_dea...	Mean

Join Action



The join action allows you to join two data tables along a column. To join the two tables, they should share one or more key columns. The key columns indicate which rows match in both tables.

To use this action, drag it into the canvas and connect two data sources to it. This is illustrated in the right panel of the figure below. The left panel shows how to set up the configuration for the join action. At least one pair of keys is required. In the example, *id* in both sources serves as the key. More key columns can be specified, if necessary, by clicking the button with the '+' sign. Finally, a join type must be selected to complete the configuration. Four join types are supported – left, right, inner, and outer.

The screenshot displays a software interface with a dark blue header containing 'File' and 'Help' menus. The main area is divided into three panels:

- ACTION CONFIGURATION:** This panel is used to set up the join action. It features a 'Select Join Column' section with two input fields: 'patient_health.csv' and 'patient_demograp hics.csv'. Both fields have 'id' selected in a dropdown menu, and a '+' button is located to the right of the second field. Below this is the 'Join Type' section, which includes four radio buttons: 'Left' (selected), 'Right', 'Inner', and 'Outer'. At the bottom of this panel are 'Run' and 'Reset' buttons.
- ACTIONS:** This panel shows a list of available actions. Under the 'File I/O' category, there are two icons: a green one with an upward arrow and a purple one with a downward arrow. Under the 'Data Manipulation' category, there are four icons: a purple one with a key, a purple one with a downward arrow, a purple one with an upward arrow, and a purple one with a join symbol.
- Canvas:** The rightmost panel is a grid-based canvas. It shows two green data source icons labeled 'patient...' connected by arrows to a central purple join action icon. The join action icon is highlighted with a dashed border, indicating it is the active element.

AK Pattern Miner



The AK Pattern Mining action uses our proprietary automated, AI-driven process which extracts statistically well-defined groups of data items. A group is defined by a set of data items that are considered ‘interesting’, i.e., perform unusually high or unusually low in terms of a user-specified target variable, and at the same time are defined by a set of common feature value ranges.

Input Data Properties

Item Count	374
Feature Count	25

Mined Patterns - Details

Pattern Count	36
Largest Pattern	284
Smallest Pattern	27
Maximum Feature Count	3
Minimum Feature Count	1

To use this action, drag it into the canvas and connect a data source to it. This action can connect to any action that outputs a data table. An example is shown in the figure below. Next, configure the action as shown in the panel on the left below. The configuration parameters are:

Sample Option: You can opt to mine on a random sample of your data or the entire data by configuring the sample option. Sampling is necessary when the data set is too large (i.e. too large to fit in RAM at any one time).

Target: This is the attribute whose performance you want to study based on the ranges of other attributes in the dataset.

Mine Type: *Numeric* or *Binary*. Select *Numeric* if your target is a continuous type numerical variable. Select *Binary* if your target is a two-class variable (e.g. Yes and No, Present and Absent). We do not support multi-class variables (e.g. Red, Blue, Green), however, these can be converted to two-class variables and analyzed by converting all values except one to a different value (e.g. Red and Other).

Max Pattern: The approximate maximum number of patterns to mine for.

Threshold: The minimum effect size for a pattern to be considered 'interesting'. The effect size is the common language effect size for the *Numeric* mine type (default is 0.6) and the odds ratio for the *Binary* mine type (default is 2).

Holdout: Number of holdout sets to test patterns against. Increasing this value reduces false positives but can increase false negatives.

Min Pattern Size: The minimum number of data points a pattern must contain. This is specified as a percentage of the total number of data points.

Rolling Regression (Experimental)



The rolling regression action provides a means for modeling time varying relationships. It is marked experimental, as it is not yet fully supported. We still include it, however, as it can be useful for smaller sized datasets and will work as expected for most cases.

The configuration panel (see right) for this action include:

Sample Options: For performing rolling regression on a sample of the dataset.

Target: The target variable to model.

Predictors: The list of covariate attributes to model the target variable.

Window: Rolling window size to use to model the target variable.

Confidence Interval: The confidence interval (in %) to include in the output.

Feature Selection: When toggled, a backwards elimination method is used to include the significant covariates for each window.

ACTION CONFIGURATION

Sample Options

Use Sample # Samples 1000

Regression Parameters

Target qyld_returns ▼

Predictors return_dji x return_gspc x return_ixic x return_nya x return_rut x return_stoxx x return_vix x x ▼

Window 60

Confidence Interval 95

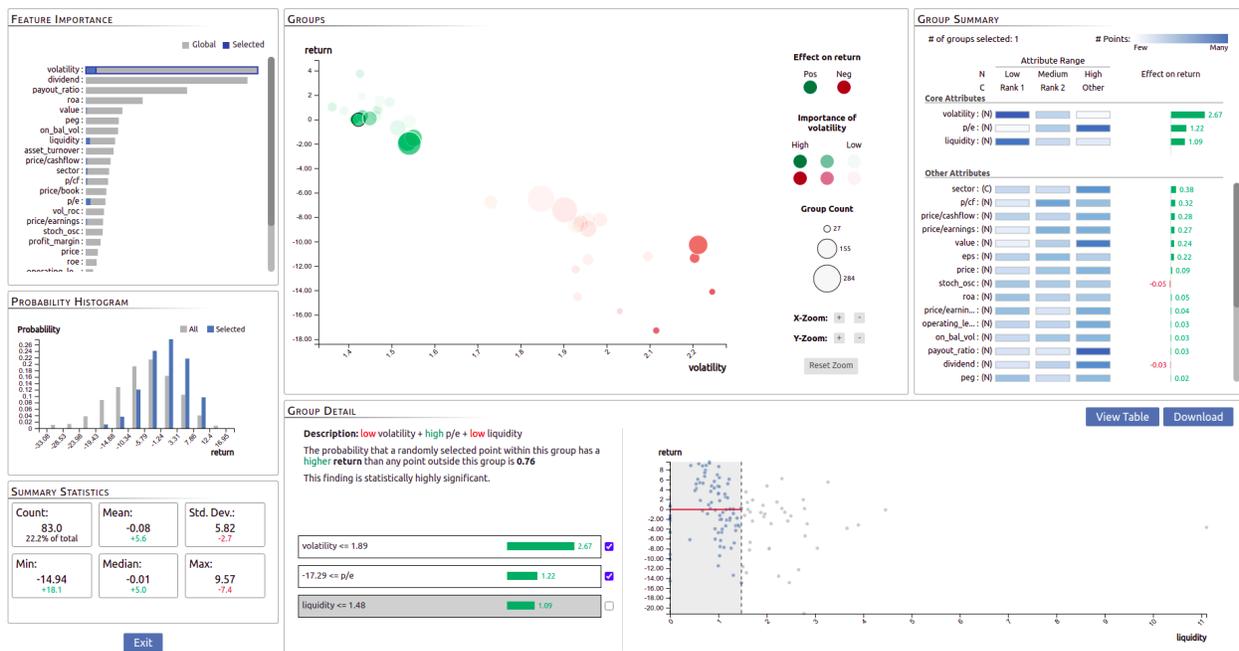
Feature Selection

The rolling regression action takes as input a dataframe and outputs a dataframe which appends the predictors' beta coefficients and lower / upper confidence intervals to the input dataframe.

AK Pattern Browser



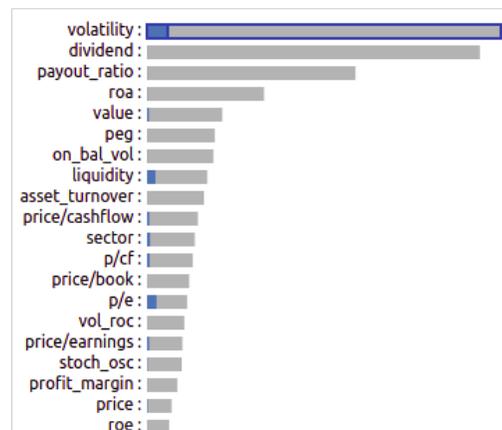
The AK Pattern Browser action brings up a visual interface for exploring a set of patterns mined via the AK Pattern Miner. It includes a high-level overview of the drivers of the specified target attribute as well as detailed information about specific patterns / groups. It also provides a way of exploring potential confounders (e.g. is a pattern based on *level of education* confounded by *age* since children have less education than adults?).



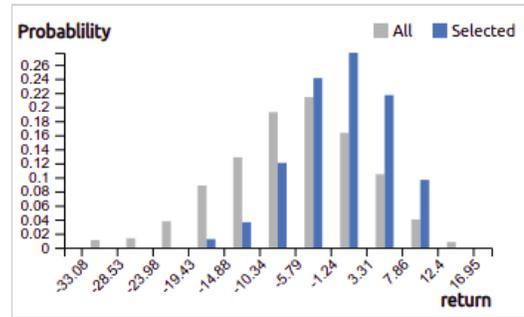
The AK Pattern Browser visual interface

The AK Pattern Browser interface includes the following panels:

Feature Importance: The feature importance panel shows a bar for each attribute indicating the relative predictive power of each attribute. The gray bars show the global importance while the blue bars indicate the local feature importance (i.e. the features important to a specific pattern / group). In the figure on the right, for example, while *dividend* is the second most important feature globally, it is not important for the selected group.



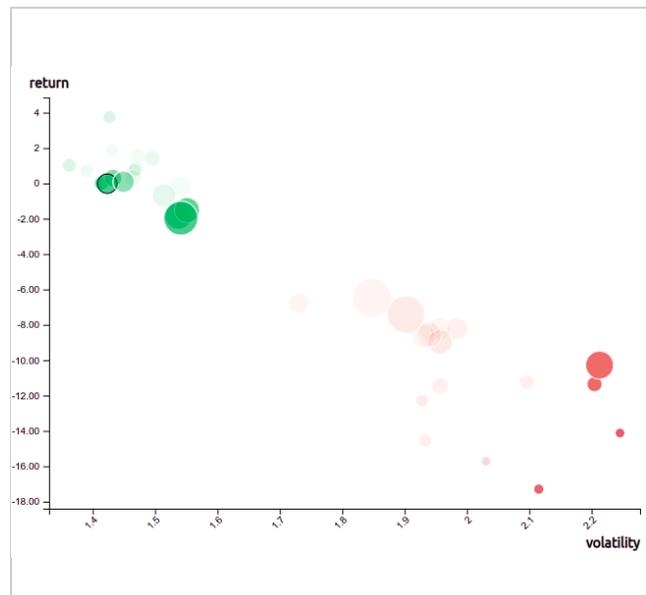
Probability Histogram: The probability histogram shows the distribution of the target variable for the full dataset (gray) and the data that falls within the selected pattern (blue).

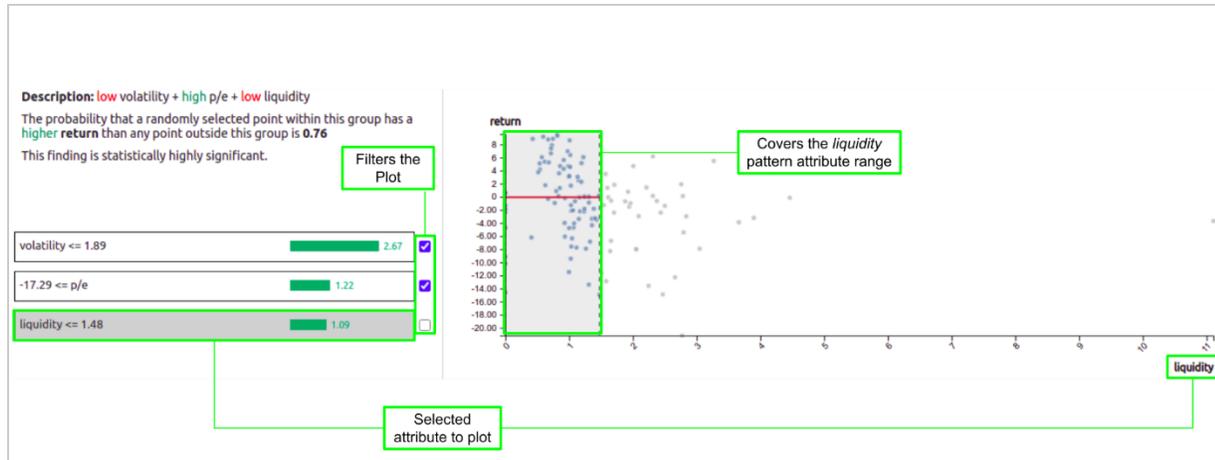


Summary Statistics: The summary statistics panel shows the statistics for the selected pattern (e.g. the count / size of the pattern, mean and standard deviation of the target variable within the pattern, etc) along with how it compares to the summary statistics for the full dataset (i.e. whether it is higher / lower than the full dataset)

Count: 83.0 22.2% of total	Mean: -0.08 +5.6	Std. Dev.: 5.82 -2.7
Min: -14.94 +18.1	Median: -0.01 +5.0	Max: 9.57 -7.4

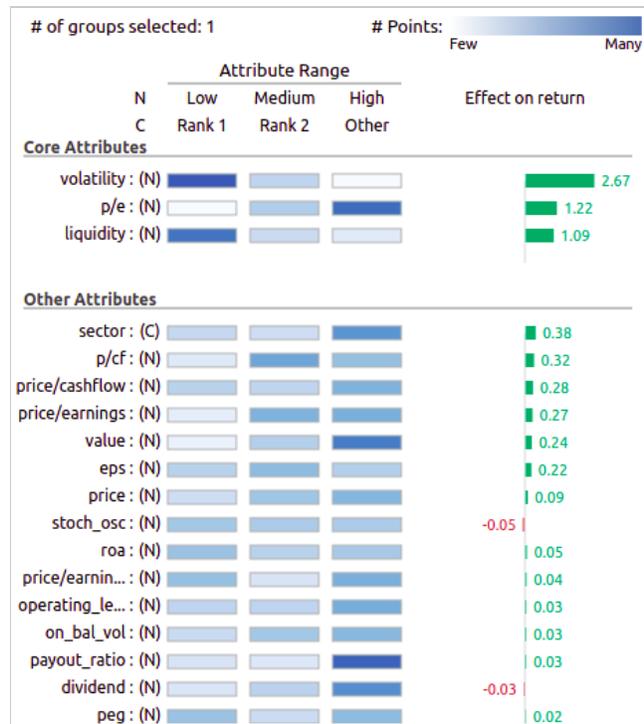
Group Bubble Chart: Each pattern in this chart is represented by a colored circle. The color is based on the target variable (i.e. whether the target variable is unusually high / low). The size of the circle is based on the number of points that fall within the pattern. The position is based on the median values for the x and y axes. The x-axis can be set by clicking on the corresponding attribute in the feature importance panel. Finally, the opacity indicates how important the x-axis attribute is as a defining characteristic (e.g. *volatility* is not important for the patterns where *volatility* is between 1.7 and 2).





Group Detail: Clicking on a circle in the group bubble chart updates the group detail panel to show more detailed information about the selected pattern. This includes each of the pattern constraints (e.g. *volatility* less than 1.89, *p/e* greater than -17.29, and *liquidity* < 1.48). The green bars are based on Shapley values and indicate the relative contribution of each of the constraints to the change in the target variable (e.g. *return*).

Group Summary: The group summary panel shows all feature's distributions for the selected group. It distinguishes between the "Core Attributes" (i.e. attributes that define the pattern) and the "Other Attributes." Each feature is divided into 3 bins – low, medium and high. Each bin is then colored from white to blue based on the number of points within the bin. In the figure on the right, for example, the selected pattern has low *volatility*, high *p/e* and low *liquidity*. Similar to the group detail panel, the green bars are based on Shapley values and show the relative contribution of the attribute on the change in the target variable.



Visual Explorer Action



The visual explorer action allows you to visually explore data in the AK analyst. This action can connect to any action that outputs a data table. The action currently enables two visualizations - scatterplot and line chart.

Scatterplot

The scatterplot can visualize up to four variables simultaneously. To use the scatterplot, select it as the base chart and then configure the parameters as shown in the figure below.

The parameters that can be selected are:

X Attribute - The attribute to which the x-axis maps.

Y Attribute - The attribute to which the y-axis maps.

Color - The color of the scatterplot points. Note this will be overridden if a color attribute is selected.

Radius Attribute - The attribute to which the radius size maps. Must be numerical.

Radius Size - The radius of the scatter plot points.

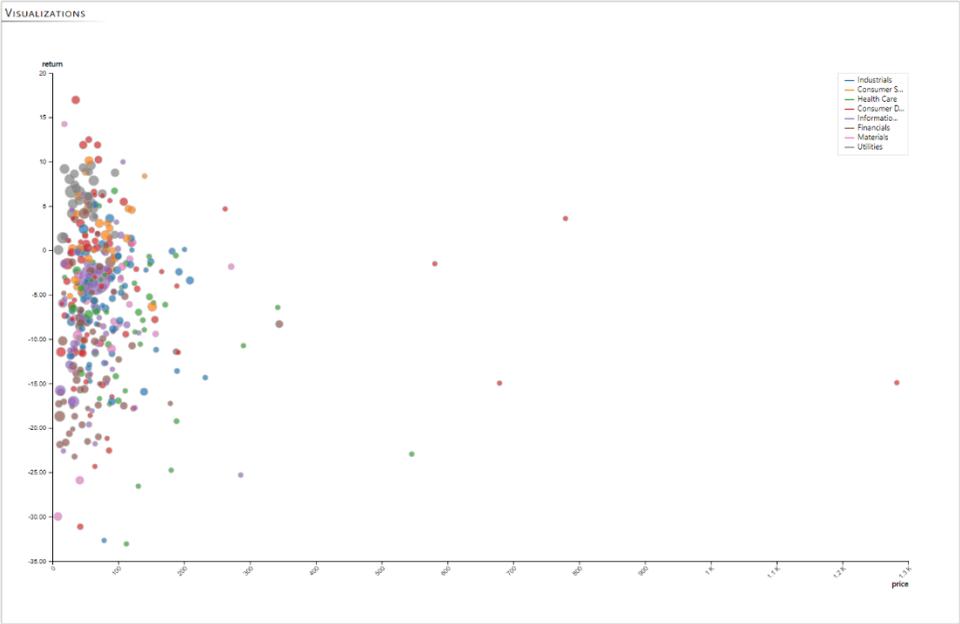
Radius Size (Max) - The maximum radius of the scatter plot points. This is only used if the radius attribute is selected and it must be larger than radius size.

Color Attribute - The attribute to which the color of points maps. If a numerical value is selected a single color scale is used ranging from white to that color. If a nominal variable is selected then a multi-color categorical color scale is selected.

The image shows a configuration window titled "CONFIGURATOR" for a scatterplot. It contains several settings:

- Base Chart:** A dropdown menu set to "Scatterplot".
- Scatterplot Options:** A section header.
- X Attribute:** A dropdown menu set to "price".
- Y Attribute:** A dropdown menu set to "return".
- Color:** A color selection bar currently showing a dark blue color.
- Radius Attribute:** A dropdown menu set to "dividend".
- Radius Size:** A text input field containing the value "4".
- Radius Size (Max):** A text input field containing the value "25".
- Color Attribute:** A dropdown menu set to "sector".

Example: An example of a scatterplot created with this action is shown below. Numerical attributes “price” and “return” have been mapped to the x-axis and y-axis respectively. Most of the items are concentrated to the lower end of the “price” range while the “return” is varying more. The numerical attribute “dividend” has been mapped to the radius of each point/circle – larger circles indicate bigger dividends. And color has been mapped to the nominal attribute “sector” which indicates which industry sector each data item (stock in this case) belongs to.



Line Chart

The line chart visualizes multiple attributes that change with a time attribute or something similar with lines. To use the line chart, select it as the base chart and then configure the parameters as shown in the figure below. The parameters that can be selected are:

X Attribute - The attribute to which the x-axis maps. For a line chart, this is most often a variable representing time or similar.

Line - You can add multiple lines with the following parameters for each line.

Y Attribute - The attribute to which the y-axis maps.

Color - The color of the line.

Lower Bound Attribute - The attribute which serves as a lower bound for the y-attribute.

Upper Bound Attribute - The attribute which serves as an upper bound for the y-attribute. When both the upper and lower are selected an interval will be highlighted around the line based on their values.

Marker Type - The shape assigned to a marker along the line.

Condition Join - The logical operator used to join multiple conditions that determine if a marker is plotted.

Condition - The condition used to determine whether a marker is plotted. The condition consists of two attribute names and a logical operator. For example, $x > y$ where x and y are attributes in the data.

CONFIGURATOR

Base Chart: Line Chart

Line Chart Options

X Attribute: date

Line 1

Y Attribute: Close_Price_qyld

Lower Bound Attribute: None

Upper Bound Attribute: None

Marker Type: Square

Condition Join: OR

Condition: < Close_P

Condition: > Close_P

Add Condition

Line 2

Y Attribute: Close_Price_qyld

Lower Bound Attribute: Close_Price_qyld_ci_lb

Upper Bound Attribute: Close_Price_qyld_ci_ut

Marker Type: None

Add Y Attribute

Example: An example of a line chart created with this action is shown below. Here dates have been mapped to the x-axis and a close price of a fund qyld (Close_price_qyld) has been mapped to the y-axis. Two lines have been added to this plot the actual close price (blue line) and a predicted close price (orange). Additionally, a confidence interval for the predicted close price has been added using the lower and upper bound attribute fields and is displayed as an orange halo. Markers have also been added using the marker and condition configuration parameters. These markers are blue squares indicating when the actual price broke the confidence interval i.e. the price was lower or higher than the lower bound or upper bound respectively.

