Visualizing with Guided Analytics: for End Users

Phil Winters
Maarit Widman
Paolo Tamagnini
Scott Fincher
“Any sufficiently advanced technology is indistinguishable from magic”

Arthur C. Clarke, 1962
Focused on AUTOMATING and GUIDING the machine learning process.
Guided Analytics

The benefits of Guided Analytics Machine Learning applications are clear.

But can Guided Analytics help us with day to day computing tasks?
Guided Visualization

“I need to make a graph for a presentation....”
### KNIME Guided Visualization: A Comparison

<table>
<thead>
<tr>
<th>The Good Old Days</th>
<th>Modern “Smart” tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Choose data to graph</td>
<td>• Interactive application</td>
</tr>
<tr>
<td>• Choose a graph style</td>
<td>• Choose data to graph</td>
</tr>
<tr>
<td>• Create the graph style</td>
<td>• Relevant graph styles presented</td>
</tr>
<tr>
<td>• Customize</td>
<td>• Graph created</td>
</tr>
</tbody>
</table>

#### Challenges

<table>
<thead>
<tr>
<th>Data must be clean</th>
<th>Black Box logic: not customizable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data TYPES must be correct</td>
<td>Data must be clean</td>
</tr>
<tr>
<td>Data/graph type must match</td>
<td>Data TYPES must be correct</td>
</tr>
<tr>
<td>Only one graph</td>
<td>Faster, but hard to tweak each graph</td>
</tr>
<tr>
<td>Takes time to get it right</td>
<td></td>
</tr>
</tbody>
</table>
Data Upload

Use the form below to upload the data you would like to visualize.

adult.csv
Guided Visualization

Data Upload

Use the form below to upload the data you would like to visualize.

Guide

Upload the dataset to visualize. The file must be in CSV format. The file will be uploaded to the server for further processing.
Guided Visualization

Data Upload
Use the form below to upload the data you would like to visualize.

flights.csv

Guide
Upload the dataset to visualize. The file must be in CSV format. The file will be uploaded to the server for further processing.
Select the Columns to Visualize

This application lets you visualize the data in a few clicks. You need to decide which columns of the uploaded data should be visualized. Follow the instructions in the panel on the right to select up to three columns.

**Column Selection Table**
Choose at most three columns

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>String</td>
</tr>
<tr>
<td>Month</td>
<td>String</td>
</tr>
<tr>
<td>DayOfWeek</td>
<td>String</td>
</tr>
<tr>
<td>Cancelled</td>
<td>String</td>
</tr>
<tr>
<td>Diverted</td>
<td>String</td>
</tr>
<tr>
<td>SecurityDelay</td>
<td>String</td>
</tr>
<tr>
<td>DelayTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSDepTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>ArrTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSArTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>FlightNum</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>ActualElapsedTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSElapsedTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>AirTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>AirDelay</td>
<td>Number (integer)</td>
</tr>
</tbody>
</table>

Showing 1 to 16 of 31 entries
Select the Columns to Visualize

This application lets you visualize the data in few clicks. You need to decide which columns of the uploaded data should be visualized. Follow the instructions in the panel on the right to select up to three columns.

Column Selection Table
Choose at least three columns.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>String</td>
</tr>
<tr>
<td>Month</td>
<td>String</td>
</tr>
<tr>
<td>DayOfWeek</td>
<td>String</td>
</tr>
<tr>
<td>Cancelled</td>
<td>String</td>
</tr>
<tr>
<td>Diverted</td>
<td>String</td>
</tr>
<tr>
<td>SecurityDelay</td>
<td>String</td>
</tr>
<tr>
<td>DayOfMonth</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>DepTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSDepTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>ArrTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSArrTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>FlightNum</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>ActualElapsedTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSElapsedTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>AirTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSArrTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>AirDelay</td>
<td>Number (integer)</td>
</tr>
</tbody>
</table>

Guide

Select between one and three column(s) for visualization. All available columns are listed in the Column Selection Table. To select a column, click the checkbox next to its name.

When you select a column, it is appended to the Title View next to the table. The Title view displays the domain (or variable type) of the selected columns, and a sample value. Visualizations are chosen based on domains specified for each column, so it is critical to provide correct domain information. To change the domain for a particular column, check the checkbox 'Convert Column Domain'. If the checkbox is selected, an additional input field page will be shown next that will allow you to change column domains as needed.

ActualElapsedTime
- Number (integer)
- Sample: 123
## Select the Columns to Visualize

This application lets you visualize the data in few clicks. You need to decide which columns of the uploaded data should be visualized. Follow the instructions in the panel on the right to select up to three columns.

### Column Selection Table

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>String</td>
</tr>
<tr>
<td>Month</td>
<td>String</td>
</tr>
<tr>
<td>DayOfWeek</td>
<td>String</td>
</tr>
<tr>
<td>Cancelled</td>
<td>String</td>
</tr>
<tr>
<td>Diverted</td>
<td>String</td>
</tr>
<tr>
<td>SecurityDelay</td>
<td>String</td>
</tr>
<tr>
<td>DayofMonth</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>DepTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSDepTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>ArrTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSArTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>FlightNum</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>ActualElapsedTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CRSElapsedTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>AirTime</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>AirDelay</td>
<td>Number (integer)</td>
</tr>
</tbody>
</table>

Selecting columns to visualize is based on the domain of the selected columns. For example, if you select "ActualElapsedTime," the visualization will show the number of elapsed time in seconds. Similarly, selecting "CRSElapsedTime" will display the number of elapsed time in seconds.
Select the Columns to Visualize

This application lets you visualize the data in a few clicks. You need to decide which columns of the uploaded data should be visualized. Follow the instructions in the panel on the right to select up to three columns.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>DepDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>Distance</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>TaxIn</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>TaxOut</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CarrierDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>WeatherDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>NASDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>LateAircraftDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>Origin</td>
<td>String</td>
</tr>
<tr>
<td>UniqueCenter</td>
<td>String</td>
</tr>
<tr>
<td>TailNum</td>
<td>String</td>
</tr>
<tr>
<td>Dest</td>
<td>String</td>
</tr>
<tr>
<td>CancelationCode</td>
<td>String</td>
</tr>
<tr>
<td>delay_class</td>
<td>String</td>
</tr>
<tr>
<td>date</td>
<td>String</td>
</tr>
</tbody>
</table>

Guide

Select between one and three column(s) for visualization. All available columns are listed in the Column Selection Table. To select a column, click the checkboxes next to its name.

When you select a column, it is appended to the Tile View next to the table. The Tile View displays the domain (or variable type) of the selected columns, and a sample value. Visualizations are chosen based on domains specified for each column, so it is critical to provide correct domain information. To change the domain for a particular column, check the checkbox 'Convert Column Domain'. If the checkbox is selected, an additional input/output page will be shown next that will allow you to change column domains as needed.
Select the Columns to Visualize

This application lets you visualize the data in few clicks. You need to decide which columns of the uploaded data should be visualized. Follow the instructions in the panel on the right to select up to three columns.

Column Selection Table

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>DepDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>Distance</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>TaxIn</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>TaxOut</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>CarrierDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>WeatherDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>NASDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>LateAircraftDelay</td>
<td>Number (integer)</td>
</tr>
<tr>
<td>Origin</td>
<td>String</td>
</tr>
<tr>
<td>UniqueCarrier</td>
<td>String</td>
</tr>
<tr>
<td>TailNum</td>
<td>String</td>
</tr>
<tr>
<td>Dest</td>
<td>String</td>
</tr>
<tr>
<td>CancellationCode</td>
<td>String</td>
</tr>
<tr>
<td>delay_class</td>
<td>String</td>
</tr>
<tr>
<td>date</td>
<td>String</td>
</tr>
</tbody>
</table>

Showing 17 to 31 of 31 entries

Guide

Select between one and three column(s) for visualization. All available columns are listed in the Column Selection Table. To select a column, click the checkboxes next to its name.

When you select a column, it is appended to the Tab View next to the table. The Tab View displays the domain (or variable type) of the selected columns, and a sample value. Visualizations are chosen based on domains specified for each column, so it is critical to provide correct domain information. To change the domain for a particular column, check the checkbox ‘Convert Column Domain’. If the checkbox is selected, an additional input portal page will be shown next that will allow you to change column domains as needed.
Convert Column Domains

Change one or more columns data types via the forms listed below.

**ActualElapsedTime**
- **Number (integer)**
- **Sample:** 123

**Converting to:**
- **Do not convert**

**CRSElapsedTime**
- **Number (integer)**
- **Sample:** 157

**Converting to:**
- **Do not convert**

**date**
- **String**
- **Sample:** 2007/03/05

**Converting to:**
- **Do not convert**

---

**Guide**

In this page you can see a form for each of the selected columns. It consists of the following:

- **Table View:** A table including:
  - The name of the column
  - The domain of the column
  - A sample value from a single row

- **Drop-Down Selection:** Select the correct domain for the column. If the domain is already correct, select "Do not convert".

- **Text Input:** To convert the column domain to a date you must define in the text box the pattern for the conversion. Note that:
  - **y:** year digit
  - **Y:** month digit
  - **d:** day digit
  - **H:** hour digit
  - **m:** minute digit
  - **s:** second digit
Convert Column Domains

Change one or more columns data types via the forms listed below.

### ActualElapsedTime
- **Convert to**
  - Do not convert
  - Date (and time) format

- **Sample**: 123

### CRSElapsedTime
- **Convert to**
  - Do not convert
  - Date (and time) format

- **Sample**: 117

### date
- **Convert to**
  - Do not convert
  - Do not convert Numeric
  - Date/Time

- **Sample**: 2007/03/05

---

**Guide**

In this page you can see a form for each of the selected columns. It consists of the following:

- **Text Box**: To convert the column domain to a date you must define in the text box the pattern for the conversion. Note that:
  - y: year digit
  - m: month digit
  - d: day digit
  - h: hour digit
  - m: minute digit
  - s: second digit

- **Drop-Down Selection**: Select the correct domain for the column. If the domain is already correct, select "Do not convert".
Convert Column Domains

Change one or more columns data types via the forms listed below.

**ActualElapsedTime**
- **Number (integer)**
- **Sample**: 123

**CRSElapsedTime**
- **Number (integer)**
- **Sample**: 197

**date**
- **String**
- **Sample**: 2007/03/05

---

**Guide**

In this page you can see a form for each of the selected columns. It consists of the following:

- **File View**: A file containing:
  - The name of the column
  - The domain of the column
  - A sample value from a single row

- **Drop-Down Selection**: Select the correct domain for the column. If the domain is already correct, select "Do not convert".

- **Custom Entry**: To convert the column domain to a date you must define in the last box the pattern for the conversion. Note that:
  - y: year digit
  - m: month digit
  - d: day digit
  - h: hour digit
  - m: minute digit
  - s: second digit
Convert Column Domains

Change one or more columns data types via the forms listed below.

**ActualElapsedTime**
- **Number (integer)**
- **Sample**: 123

Convert to:
- **Do not convert**
- **Date (and time) format**

**CRSElapsedTime**
- **Number (integer)**
- **Sample**: 177

Convert to:
- **Do not convert**
- **Date (and time) format**

**date**
- **String**
- **Sample**: 2007/03/06

Convert to:
- **Date/Time**
- **Date (and time) format**: yyyyMMdd
Customize the Visualizations

Visualizations displayed below are chosen based upon your previous selection of columns. The visualizations are highly interactive, and you can customize them to your needs as described in the sidebar. On the next page you will be able to export these charts to static images.

Title
Randomly sampled 2500 rows. Insert Title via button "Settings" in top right corner

Displayed Visualizations

- **Line Plot Moving Average**: Plots one or more numerical variables on the y-axis against a date on the x-axis. The line plot graphically represents the evolution of a time series. The moving average line shows the change of the past 21-day average of the values in the numerical variable(s). The 21-day average is calculated by averaging the value of the current record and the 20 records preceding it.
Customize the Visualizations

Visualizations displayed below are chosen based upon your previous selection of columns. The visualizations are highly interactive, and you can customize them to your needs as described in the sidebar. On the next page you will be able to export these charts to static images.

Title
Randomly sample 250 rows. Insert title via button "Settings" in top right corner.

Displayed Visualizations

- **Line Plot Moving Average**: Plots one or more numerical variables on the y-axis against a date on the x-axis. The line plot graphically represents the evolution of a time series. The moving average line shows the change of the prior 21-day average of the values at the numerical variables. The 21-day average is calculated by averaging the value of the current record and the 20 records preceding it.
Customize the Visualizations

Visualizations displayed below are chosen based upon your previous selection of columns. The visualizations are highly interactive, and you can customize them to your needs as described in the sidebar. On the next page you will be able to export these charts to static images.

Title
Randomly sampled 2500 rows. Insert title via button "Settings" in top right corner.
Customize the Visualizations

Visualizations displayed below are chosen based upon your previous selection of columns. The visualizations are highly interactive, and you can customize them to your needs as described in the sidebar. On the next page you will be able to export these charts to static images.

Comparing Time Series
Moving Average of Actual and CR5 Expired Times

Guide

On this page we display automatically the most appropriate charts for your selection of columns. Feel free to interact with the JavaScript visualization.

A configuration panel can be accessed in the top right corner of each chart. You can customize a number of things depending on the chart, such as the title, subtitle, legend, axis labels and other settings unique to the inspected visualization. Once you are done customizing the chart, your edits are applied to the visualizations by selecting Next.

Displayed Visualizations

- Line Plot Moving Average: Plots one or more numerical variables on the y-axis against a date on the x-axis. The line plot graphically represents the evolution of a time series. The moving average line shows the change of the prior 21-day average of the value in the numerical variable(s). The 21-day average is calculated by averaging the value of the current record and the 20 records preceding it.
Customize the Visualizations

Visualizations displayed below are chosen based upon your previous selection of columns. The visualizations are highly interactive, and you can customize them to your needs as described in the sidebar. On the next page you will be able to export these charts to static images.

Comparing Time Series

Moving Average of Actual and GIS Exposed Times

Guide

On this page we display automatically the most appropriate charts for your selection of columns. Feel free to interact with the JavaScript visualization.

A configuration panel can be accessed in the top right corner of each chart.

You can customize a number of things depending on the chart, such as the title, subtitle, legend, axis labels and other settings unique to the inspected visualization. Once you are done customizing the chart, your edits are applied to the visualizations by selecting Next.

Displayed Visualizations

- Line Plot: Moving Average: Plots one or more numerical variables on the y-axis against a date on the x-axis. The line plot graphically represents the evolution of a time series. The moving average line shows the change of the prior 21-day average of the values in the numerical variable(s). The 21-day average is calculated by averaging the value of the current record and the 20 records preceding it.
Download Visualizations

Below are the static versions of your customized visualizations. Use the button to the right of each image to download it in SVG format.

Comparing Time Series

Monthly Average of Actual and Off-Season Times

![Comparison Chart]

Download Line plot

Download
Download Visualizations

Below are the static versions of your customized visualizations. Use the button to the right of each image to download it in SVG format.

Comparing Time Series

Moving Average of Actual and CBO Deeped Times

Download Line plot

Guide

The visualizations are now displayed as static images showing all the customization you applied in the previous page. If you are happy with what you see, download the SVG file by clicking on the download button. If you need to make additional changes, go back to the previous page to apply more customizations.
Comparing Time Series

Expected Flight Duration vs Actual Flight Duration

Average Flight Duration

Date


MA(CRSElapsedTime) MA(ActualElapsedTime)
Guided Visualization

This workflow defines a fully automated web based application to select, customize and download a number of visualizations. The workflow was designed for business analysts to easily create charts by applying their domain knowledge. Each of the wrapped metanodes outputs a web page with which the business analyst can interact.

The Process Step by Step

1. Upload your data / Select one of the available datasets
2. Select the columns to visualize (maximum 3)
3. Convert the domain of the columns (OPTIONAL) 
4. Customize the visualizations interactively
5. Download the images of the customized charts
Guided Visualization

This workflow defines a fully automated web-based application to select, customize, and download a number of visualizations. The workflow was designed for business analysts to easily create charts by applying their domain knowledge. Each of the wrapped metanodes outputs a web page with which the business analyst can interact.

The Process Step by Step

1. Upload your data / Select one of the available datasets
2. Select the columns to visualize (maximum 3)
3. Convert the domain of the columns (OPTIONAL)
4. Customize the visualizations interactively
5. Download the images of the customized charts
Convert selected columns to the selected types
If conversion fails, return the original data type
Guided Visualization

This workflow defines a fully automated web based application to select, customize and download a number of visualizations. The workflow was designed for business analysts to easily create charts by applying their domain knowledge. Each of the wrapped metanodes outputs a web page with which the business analyst can interact.

The Process Step by Step

1. Upload your data / Select one of the available datasets
2. Select the columns to visualize (maximum 3)
3. Convert the domain of the columns (OPTIONAL)
4. Customize the visualizations interactively
5. Download the images of the customized charts

 Ports

Input Ports
0 Table storing the selected columns to be visualized.

Output Ports
0 Table storing the customized charts in SVG format.
Guided Visualization for the Casual User

Standalone.....

Or as the part of a bigger Guided Analytics Application
KNIME Guided Visualization: A Comparison

The Good Old Days
• Choose data to graph
• Choose a graph style
• Create the graph style
• Customize

Challenges
• Data must be clean
• Data TYPES must be correct
• Data/graph type must match
• Only one graph
• Takes time to get it right

Modern “Smart” tools
• Interactive application
• Choose data to graph
• Relevant graph styles presented
• Graph created

Challenges
• Black Box logic: not customizable
• Data must be clean
• Data TYPES must be correct
• Faster, but hard to tweak each graph

KNIME Guided Visualization
• Interactive application
• Choose data to graph
• Data Inconsistencies caught / corrected
• Relevant graph styles presented
• Graph created
• Customizable process
• Tailorable Graphs
Visualizing with Guided Analytics:
for End Users  AND  Experts.

Phil Winters
Maarit Widman
Paolo Tamagnini
Scott Fincher
Guided Exploration

• “I have no idea what this massive quantity of new data is.....”
**Guided Exploration**

This workflow defines a fully automated web based application to show relevant visualization iteration after iteration. The workflow was designed for data scientist to easily create a dashboard and find relationships between columns.

**Guided Visualization**

This workflow defines a fully automated web based application to select, customize and download a number of visualizations. The workflow was designed for business analysts to easily create charts by applying their domain knowledge. Each of the wrapped metanodes outputs a web page with which the business analyst can interact.

**The Process Step by Step**

1. Upload your data / Select one of the available datasets
2. Inspect the visualizations in "Guided Exploration" Wrapped Metanode
3. Select at the bottom of the View the columns you would like to remove
4. Upload Pre-Processing Guided Exploration Reference Column Filter
   - Select Columns If Switch
   - Convert Domains
   - Customize Download

**Ports**

**Input Ports**

0. A table listing the columns to visualize.
1. A table with the data. This table should have all the columns mentioned in the other inputs of the wrapped metanode.
2. A table listing all columns with different scores related to missing, constant and unique values.

**Output Ports**

0. The same data that was used as input to visualize the charts.
1. Those are the columns the user flagged. You can do different things with this output depending on the purpose of the web application.
Guided Exploration

The visualizations below were automatically generated to provide insight into your dataset. They were selected by detecting properties and relationships within and between data columns.

Univariate Column Detection

Each visualization takes into account a single column detected by a different statistical test.

- **Most Uniform Histogram**
  - Possible Row ID
  ![Histogram](image)
  - DayOfWeek
  - DayOfMonth
  - DayOfYear

- **Most Skewed Histogram**
  - CarrierDelay
  - TotalDelay
  - TotalArrDelay

- **Most Spiked Histogram**
  - Some bins are way bigger than others

- **Most Leptokurtic Box Plot**
  - Possible profusion of outliers

Guide

Generated Visualizations

In this page the workflow automatically displays some visualizations to explore the data.

Univariate Column Detection

- **Most Uniform Histogram**
  - Shows a histogram of the column that is most uniformly distributed. Such a distribution will appear mostly "flat" to the viewer, and often occurs when every value is (mostly) unique - for example, in a raw identifier column.

- **Most Skewed Histogram**
  - Shows the histogram of the column with the largest skewness. Skewed distributions are...
Guided Exploration

The visualizations below were automatically generated to provide insight into your dataset. They were selected by detecting properties and relationships within and between data columns.

Univariate Column Detection

Each visualization takes into account a single column detected by a different statistical test.

- **Most Uniform Histogram**: This shows a histogram of the column that is most uniformly distributed. Such a distribution will appear mostly "flat" to the viewer, and often occurs when every value is (mostly) unique - for example, a row identifier column.

- **Most Skewed Histogram**: This shows the histogram of the column with the largest skewness. Skewed distributions are asymmetric relative to the mean of the distribution, and the skewness measures the degree of that asymmetry.

- **Most Spiked Histogram**: This shows the histogram of the column with the largest difference between a particular binned value and the average binned value. If the difference is sufficiently large, it may indicate an anomalous feature in your dataset which should be investigated further.

- **Most Leptokurtic Box Plot**: This shows a box plot of the column with the largest kurtosis, which is a measure of how heavy-tailed a particular distribution is. The column shown here will usually have a large number of outliers relative to other features in the dataset.

Multivariate Column Detection

- **Most Correlated Scatter Plot**: This shows the two numeric columns most positively correlated with one another, where the first column increases, the second column does also. Highly positively correlated columns may indicate that one of the features is a candidate for removal.

- **Most Inverse Correlated Scatter Plot**: This shows the two numeric columns most negatively correlated with one another, where the first column increases, the second column decreases. Highly negatively correlated columns may indicate that one of the features is a candidate for removal.
Multivariate Column Detection

Each visualization takes into account a relationship between two or more columns detected by a different statistical test.

Correlated Numerical Columns

- **Most Inverse Correlated Scatter Plot**: Greater the x lower the y and vice versa.
- **Most Correlated Scatter Plot**: Greater the x greater the y and vice versa.

Correlated Categorical Columns

- **Most Correlated Sunburst Chart**: This shows the most correlated categorical columns. The inner and outer rings should appear connected in some fashion, as the correlation should bind together the different possible string value pairs. The vertical slider enables you to set the number of observations to display in the outer ring, for better readability.

Related Columns with Different Domains

- **Most Diversified Stacked Histogram**: This chart displays a selected categorical and numerical column, based on maximization of the computed correlation between categorical columns and binned numerical columns. The visualization shows two subsets of rows in two stacked histograms of the numerical feature, which should appear quite different. The more different the two stacked histograms, the stronger the relationship between the two columns.
- **Most Diversified Conditional Bar Plot**: This chart displays a selected categorical and numerical column based on performing several ANOVA tests. The visualization shows two subsets of rows in two different box plots of the numerical feature, which should appear quite different. The more different the two box plots, the stronger the relationship between the two columns.
- **Most Correlated Parallel Coordinates Plot**: This plot shows the relationship between the two most correlated numerical and categorical columns, where the categorical information is encoded by replacing string values with associated frequencies.
- **Most Correlated Missing Values Parallel Coordinates Plot**: This plot is similar to its predecessor, and again measures correlation between numerical and categorical columns.
Multivariate Column Detection

Each visualization takes into account a relationship between **two or more** columns detected by a different statistical test.

**Correlated Numerical Columns**

- **Most Inverse Correlated Scatter Plot**: Greater the x lower the y and vice versa

- **Most Correlated Scatter Plot**: Greater the y greater the x and vice versa

**Correlated Categorical Columns**

- **Most Correlated Sunburst Chart**: Inner Circle: Origin - Outer Circle: Unique Carrier
Related Columns with Different Domains

Most Diversified Stacked Histogram
Distributions of Delay partitioned by delay, class

Most Diversified Conditional Box Plot

Related Columns with Different Domains
Related Columns with Different Domains

Most Diversified Stacked Histogram
Distributions of Delay by delay_class

Most Diversified Conditional Box Plot

Most Correlated Parallel Coordinates Plot

Most Correlated Missing Values Parallel Coordinates Plot
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

Ctrl + Left Click Bins of Columns to be Flagged

Bed Quality Rank

No matching records found

Removing Columns from the Dashboard

Use the Horizontal Bar Chart on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality. For example, several missing values, many constant numerical values, or highly varying categorical values.

Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary.
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

Ctrl + Left Click Bins of Columns to be Flagged

Data Explorer
<table>
<thead>
<tr>
<th>Column</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Overall Sum</th>
<th>No. zeros</th>
<th>No. missings</th>
</tr>
</thead>
<tbody>
<tr>
<td>DayOfMonth</td>
<td>1</td>
<td>31</td>
<td>15.596</td>
<td>8.895</td>
<td>79.120</td>
<td>0.058</td>
<td>-1.204</td>
<td>80094</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DepTime</td>
<td>2</td>
<td>2400</td>
<td>1398.974</td>
<td>483.948</td>
<td>234205.243</td>
<td>-0.004</td>
<td>-0.905</td>
<td>7234095</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CRSDepTime</td>
<td>20</td>
<td>2359</td>
<td>1389.823</td>
<td>466.647</td>
<td>217759.596</td>
<td>0.069</td>
<td>-1.036</td>
<td>7168776</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ArTime</td>
<td>2</td>
<td>2400</td>
<td>1520.233</td>
<td>526.811</td>
<td>279746.796</td>
<td>-0.534</td>
<td>0.010</td>
<td>7889308</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>CRSArTime</td>
<td>1</td>
<td>2359</td>
<td>1534.829</td>
<td>489.347</td>
<td>239460.297</td>
<td>-0.327</td>
<td>-0.365</td>
<td>8040020</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FlightNum</td>
<td>1</td>
<td>7820</td>
<td>2083.988</td>
<td>1931.693</td>
<td>3731436.061</td>
<td>1.035</td>
<td>-0.044</td>
<td>10776301</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ActualElapsedTime</td>
<td>26</td>
<td>558</td>
<td>135.565</td>
<td>71.765</td>
<td>5150.233</td>
<td>1.336</td>
<td>2.126</td>
<td>699076</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>CRSElapsedTime</td>
<td>35</td>
<td>578</td>
<td>135.410</td>
<td>71.217</td>
<td>5071.925</td>
<td>1.394</td>
<td>2.338</td>
<td>705424</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AirTime</td>
<td>14</td>
<td>539</td>
<td>109.749</td>
<td>69.226</td>
<td>4792.215</td>
<td>1.413</td>
<td>2.423</td>
<td>566193</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>ArDelay</td>
<td>-54</td>
<td>923</td>
<td>12.556</td>
<td>40.063</td>
<td>1605.025</td>
<td>5.799</td>
<td>81.662</td>
<td>64774</td>
<td>149</td>
<td>12</td>
</tr>
<tr>
<td>DepDelay</td>
<td>-28</td>
<td>339</td>
<td>13.436</td>
<td>36.788</td>
<td>1335.345</td>
<td>7.190</td>
<td>118.783</td>
<td>69479</td>
<td>465</td>
<td>0</td>
</tr>
<tr>
<td>Distance</td>
<td>36</td>
<td>4502</td>
<td>795.950</td>
<td>583.182</td>
<td>340101.720</td>
<td>1.311</td>
<td>1.515</td>
<td>4115342</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TaxiIn</td>
<td>0</td>
<td>74</td>
<td>6.322</td>
<td>4.373</td>
<td>19.121</td>
<td>4.380</td>
<td>37.931</td>
<td>32589</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>TaxiOut</td>
<td>3</td>
<td>233</td>
<td>19.427</td>
<td>12.596</td>
<td>157.140</td>
<td>4.488</td>
<td>38.841</td>
<td>100455</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CarrierDelay</td>
<td>0</td>
<td>844</td>
<td>4.943</td>
<td>23.691</td>
<td>561.253</td>
<td>17.307</td>
<td>505.711</td>
<td>25862</td>
<td>4423</td>
<td>0</td>
</tr>
<tr>
<td>WeatherDelay</td>
<td>0</td>
<td>202</td>
<td>0.897</td>
<td>8.066</td>
<td>65.572</td>
<td>13.492</td>
<td>223.658</td>
<td>4640</td>
<td>5040</td>
<td>0</td>
</tr>
<tr>
<td>NASDelay</td>
<td>0</td>
<td>257</td>
<td>0.025</td>
<td>14.890</td>
<td>221.870</td>
<td>7.080</td>
<td>79.360</td>
<td>20821</td>
<td>4236</td>
<td>0</td>
</tr>
<tr>
<td>LateAircraftDelay</td>
<td>0</td>
<td>312</td>
<td>6.885</td>
<td>21.196</td>
<td>449.281</td>
<td>6.017</td>
<td>49.312</td>
<td>295956</td>
<td>4484</td>
<td>0</td>
</tr>
</tbody>
</table>

Showing 1 to 18 of 18 entries
### Data Explorer

<table>
<thead>
<tr>
<th>Column</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Overall Sum</th>
<th>No. zeros</th>
<th>No. missings</th>
</tr>
</thead>
<tbody>
<tr>
<td>DayOfWeek</td>
<td>1</td>
<td>31</td>
<td>15.596</td>
<td>8.895</td>
<td>79.120</td>
<td>0.056</td>
<td>-1.204</td>
<td>80954</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DepTime</td>
<td>2</td>
<td>2400</td>
<td>1398.974</td>
<td>483.948</td>
<td>234205.243</td>
<td>-0.004</td>
<td>-0.905</td>
<td>7234095</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CRSDepTime</td>
<td>20</td>
<td>2359</td>
<td>1369.823</td>
<td>465.647</td>
<td>217759.596</td>
<td>0.069</td>
<td>-1.036</td>
<td>7166775</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ArrTime</td>
<td>2</td>
<td>2400</td>
<td>1550.232</td>
<td>528.911</td>
<td>275746.766</td>
<td>-0.534</td>
<td>0.010</td>
<td>7669308</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>

**Histogram**

- **CRSArrTime**: 1 2359 1554.929 489.347 239469.297 -0.327 -0.365 86040020 0 0
- **FlightNum**: 1 7820 2003.988 1931.693 3731436.061 1.035 -0.044 10779301 0 0
- **ActualElapsedTime**: 28 558 135.506 71.765 5150.233 1.336 2.128 6596776 0 12
<table>
<thead>
<tr>
<th>Column</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Overall Sum</th>
<th>No. zeros</th>
<th>No. missings</th>
</tr>
</thead>
<tbody>
<tr>
<td>DayOfMonth</td>
<td>1</td>
<td>31</td>
<td>15.596</td>
<td>8.895</td>
<td>79.120</td>
<td>-1.204</td>
<td>-0.204</td>
<td>80094</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DepTime</td>
<td>2</td>
<td>2400</td>
<td>1399.974</td>
<td>483.948</td>
<td>234205.243</td>
<td>-0.004</td>
<td>-0.905</td>
<td>7234095</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CRSDepTime</td>
<td>20</td>
<td>2359</td>
<td>1389.823</td>
<td>456.647</td>
<td>217759.596</td>
<td>0.069</td>
<td>-1.036</td>
<td>7168775</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ArrTime</td>
<td>2</td>
<td>2400</td>
<td>1520.233</td>
<td>526.511</td>
<td>279746.796</td>
<td>-0.534</td>
<td>0.019</td>
<td>7600308</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>CRSArrTime</td>
<td>1</td>
<td>2359</td>
<td>1554.829</td>
<td>489.347</td>
<td>239460.297</td>
<td>-0.327</td>
<td>-0.365</td>
<td>8040020</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FlightNum</td>
<td>1</td>
<td>7820</td>
<td>2083.988</td>
<td>1931.693</td>
<td>3731436.061</td>
<td>1.035</td>
<td>-0.044</td>
<td>10776301</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ActualElapsedTime</td>
<td>26</td>
<td>555</td>
<td>135.506</td>
<td>71.765</td>
<td>5150.233</td>
<td>1.336</td>
<td>2.128</td>
<td>6990756</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>CRSElapsedTime</td>
<td>35</td>
<td>578</td>
<td>136.419</td>
<td>71.217</td>
<td>5071.925</td>
<td>1.394</td>
<td>2.338</td>
<td>705424</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AirTime</td>
<td>14</td>
<td>539</td>
<td>109.749</td>
<td>69.226</td>
<td>4792.215</td>
<td>1.413</td>
<td>2.423</td>
<td>566193</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>ArrDelay</td>
<td>-54</td>
<td>923</td>
<td>12.556</td>
<td>4.063</td>
<td>1605.025</td>
<td>5.799</td>
<td>81.662</td>
<td>64774</td>
<td>149</td>
<td>12</td>
</tr>
<tr>
<td>DepDelay</td>
<td>-28</td>
<td>939</td>
<td>13.436</td>
<td>36.788</td>
<td>1335.345</td>
<td>7.190</td>
<td>118.783</td>
<td>69679</td>
<td>465</td>
<td>0</td>
</tr>
<tr>
<td>Distance</td>
<td>36</td>
<td>4502</td>
<td>795.950</td>
<td>583.182</td>
<td>340101.720</td>
<td>1.311</td>
<td>1.513</td>
<td>415134</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TaxiIn</td>
<td>0</td>
<td>74</td>
<td>6.322</td>
<td>4.373</td>
<td>19.121</td>
<td>4.380</td>
<td>37.931</td>
<td>32569</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>TaxiOut</td>
<td>3</td>
<td>233</td>
<td>19.427</td>
<td>12.536</td>
<td>157.140</td>
<td>4.488</td>
<td>38.841</td>
<td>100455</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CarrierDelay</td>
<td>0</td>
<td>844</td>
<td>4.943</td>
<td>23.691</td>
<td>556.253</td>
<td>17.357</td>
<td>595.711</td>
<td>25562</td>
<td>4423</td>
<td>0</td>
</tr>
<tr>
<td>WeatherDelay</td>
<td>0</td>
<td>202</td>
<td>0.897</td>
<td>0.898</td>
<td>0.757</td>
<td>13.492</td>
<td>236.158</td>
<td>464</td>
<td>5840</td>
<td>0</td>
</tr>
<tr>
<td>NASDelay</td>
<td>0</td>
<td>257</td>
<td>4.025</td>
<td>14.890</td>
<td>221.870</td>
<td>7.080</td>
<td>70.360</td>
<td>20821</td>
<td>4296</td>
<td>0</td>
</tr>
<tr>
<td>LateAircraftDelay</td>
<td>0</td>
<td>312</td>
<td>6.085</td>
<td>21.196</td>
<td>449.281</td>
<td>6.017</td>
<td>49.312</td>
<td>29996</td>
<td>4484</td>
<td>0</td>
</tr>
</tbody>
</table>

Showing 1 to 18 of 18 entries
Removing Columns from the Dashboard

**Horizontal Bar Chart to Remove Columns**

Ctrl + Left Click Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Directed
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
- delay_class
- TailNum
- FlightNum
- AirTime
- Month
- DayOfWeek
- CRSOrDepTime
- Time
- Distance
- DayOfWeek
- CRSDepTime
- UniqueCarrier
- Eta
- DepDelay
- TaxOut
- ActualDepTime
- CRSActualTime
- AirTime
- Dist
- AirDelay
- DayOfMonth
- data

Use the Horizontal Bar Chart on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality, for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

**Horizontal Bar Chart to Remove Columns**

- **Ctrl + Left Click Bins of Columns to be Flagged**

  - Year
  - Cancelled
  - Cancellation Code
  - Security Delay
  - Directed
  - Weather Delay
  - Late Aircraft Delay
  - Carrier Delay
  - NAS Delay
  - delay_class
  - Tailnum
  - FlightNum
  - AirTime
  - Month
  - DayOfWeek
  - CRS Airline Time
  - Taxis
  - Distance
  - Day of Week
  - CRS Departure Time
  - Unique Carrier
  - Origin
  - Dest
  - Ticket
  - Actual Departure Time
  - CRSElapsed Time
  - AirTime
  - Dist
  - Air Delay
  - Days of Month
  - data

**Bad Quality Rank**

Use the **Horizontal Bar Chart** on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality; for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the **Data Explorer** to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

**Horizontal Bar Chart to Remove Columns**

- **Ctrl + Left Click Bins of Columns to be Flagged**
  - Year
  - CancellationCode
  - SecurityDelay
  - Directed
  - WeatherDelay
  - LateAircraftDelay
  - CarrierDelay
  - NASDelay
  - delay_class
  - TailNum
  - FlightNum
  - AirTime
  - Month
  - DepTime
  - CRSArrTime
  - TqfFlt
  - Distance
  - DayOfWeek
  - CRSDepTime
  - UniqueCarrier
  - Origin
  - Dest
  - TaxOn
  - AirlineCode
  - CRSElapsedTime
  - AirTime
  - Dist
  - AirDelay
  - DayOfMonth
  - Total

Use the **Horizontal Bar Chart** on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with a relatively poor quality, for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the **Data Explorer** to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

**Horizontal Bar Chart to Remove Columns**

Ctrl + Left Click Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Directed
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
- delay_class
- TailNum
- FlightNum
- AirTime
- Month
- DayTime
- CRSFlightTime
- Tolk
- Distance
- DayOfWeek
- CRSDepTime
- UniqueCarrier
- Origin
- Dest
- TaxiIn
- ActualDepartureTime
- CRSElapsedTime
- AirTime
- Dist
- AirDelay
- DaysOfMonth

Use the **Horizontal Bar Chart** on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality; for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the **Data Explorer** to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.

---

© 2019 KNIME AG. All Rights Reserved.
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

Use the Horizontal Bar Chart on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality, for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

**Horizontal Bar Chart to Remove Columns**

Press `Ctrl + Left Click` Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Diverted
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
- delay_class
- TailNum
- FlightNum
- ArrTime
- Month
- DayTime
- CRSArrTime
- Tarin
- Distance
- DayOfWeek
- CRSDepTime
- UniqueCarrier
- Origin
- DepDelay
- TaxiOut
- ActualElapsedTime
- CRSElapsedTime
- ArrTime
- Dest
- AirDelay
- DayofMonth
- data

Use the **Horizontal Bar Chart** on the side to flag columns you would like to remove in the next iteration by pressing `Ctrl + click`. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with a relatively poor quality; for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the **Data Explorer** to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

Ctrl + Left Click Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Diverted
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
- delay_class
- Tailnum
- Flightnum
- AirTime
- Month
- DayTime
- CRSArrTime
- Tostart
- Distance
- DayOfWeek
- CRSDepTime
- UniqueCarrier
- Origin
- Dest
- TaxOut
- ActualDepTime
- CRSElapsedTime
- AirTime
- Dist
- AirDelay
- DayOfMonth
- data

Use the **Horizontal Bar Chart** on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality, for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the **Data Explorer** to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

Ctrl + Left Click Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Diverted
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
- delay_class
- TailNum
- FlightNum
- AirTime
- Month
- DayOfWeek
- CRSFlightTime
- T obt
- Distance
- DayOfWeek
- CRSDepTime
- UniqueCarrier
- Origin
- Dest
- TaxiOut
- ActualElapsedTime
- CRSElapsedTime
- AirTime
- Dist
- AirDelay
- DayofMonth

Use the **Horizontal Bar Chart** on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality, for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the **Data Explorer** to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

Ctrl + Left Click Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Diverted
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
- delay_class
- TailNum
- FlightNum
- AirTime
- Month
- DayOfWeek
- CRSDepTime
- CRSElapsedTime
- UniqueCarrier
- Origin
- Dest
- TaxiOut
- ActualArrTime
- ActUALDrpoffDistance

Use the Horizontal Bar Chart on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality; for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

- Ctrl + Left Click Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Diverted
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
- delay_class
- TailNum
- FlightNum
- AirTime
- Month
- DayTime
- CRSArrTime
- Taxi
- Distance
- DayOfWeek
- CRSDLalt
- UniqueCarrier
- Origin
- Dest
- TaxiOut
- AirTime
- Dist
- AirDelay
- DayofMonth
- data

---

Removing Columns from the Dashboard

Use the Horizontal Bar Chart on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality; for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

- Ctrl + Left Click Bins of Columns to be Flagged

<table>
<thead>
<tr>
<th>Column</th>
<th>Bad Quality Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td></td>
</tr>
<tr>
<td>Cancellation</td>
<td></td>
</tr>
<tr>
<td>CancellationCode</td>
<td></td>
</tr>
<tr>
<td>SecurityDelay</td>
<td></td>
</tr>
<tr>
<td>Diverted</td>
<td></td>
</tr>
<tr>
<td>WeatherDelay</td>
<td></td>
</tr>
<tr>
<td>LateAircraftDelay</td>
<td></td>
</tr>
<tr>
<td>CarrierDelay</td>
<td></td>
</tr>
<tr>
<td>NASDelay</td>
<td></td>
</tr>
<tr>
<td>delay_class</td>
<td></td>
</tr>
<tr>
<td>TailNum</td>
<td></td>
</tr>
<tr>
<td>FlightNum</td>
<td></td>
</tr>
<tr>
<td>AirTime</td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td></td>
</tr>
<tr>
<td>DepTime</td>
<td></td>
</tr>
<tr>
<td>CRSArrTime</td>
<td></td>
</tr>
<tr>
<td>Tcode</td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td></td>
</tr>
<tr>
<td>DayOfWeek</td>
<td></td>
</tr>
<tr>
<td>CRSDepTime</td>
<td></td>
</tr>
<tr>
<td>UniqueCarrier</td>
<td></td>
</tr>
<tr>
<td>Origin</td>
<td></td>
</tr>
<tr>
<td>DepDelay</td>
<td></td>
</tr>
<tr>
<td>TaxiOut</td>
<td></td>
</tr>
<tr>
<td>Actual_gates_and_time</td>
<td></td>
</tr>
<tr>
<td>CRS_elapsed_time</td>
<td></td>
</tr>
<tr>
<td>AirTime</td>
<td></td>
</tr>
<tr>
<td>Dist</td>
<td></td>
</tr>
<tr>
<td>AirDelay</td>
<td></td>
</tr>
<tr>
<td>DayOfWeek</td>
<td></td>
</tr>
<tr>
<td>data</td>
<td></td>
</tr>
</tbody>
</table>

Use the **Horizontal Bar Chart** on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality, for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the **Data Explorer** to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Removing Columns from the Dashboard

Horizontal Bar Chart to Remove Columns

Ctrl + Left Click Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Diverted
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
- delay_class
- TailNum
- FlightNum
- AirTime
- Month
- DayTime
- CRSLatTime
- Tstp
- Distance
- DayOfWeek
- CRSDepTime
- UniqueCarrier
- Origin
- DepDelay
- TaxCD
- ActualDepTime
- CRSElapsedTime
- AirTime
- Dist
- ArrDelay
- DayOfWeek
Removing Columns from the Dashboard

**Horizontal Bar Chart to Remove Columns**

Ctrl + Left Click Bins of Columns to be Flagged

Year
Cancelled
CancellationCode
SecurityDelay
Cancelled
Diverted
WeatherDelay
LateAircraftDelay
CarrierDelay
NASDelay
delay_class
Tailnum
FlightNum
AirTime
Month
DepTime
CRSDepTime
Tout
Distance
DayOfWeek
CRSArrTime
UniqueCarrier
Origin
DepDelay
TaxiOut
ActualDepDelay
CRSElapsedTime
AirTime
Dist
ArrDelay
DayOfWeek
data

**Bad Quality Rank**

Use the **Horizontal Bar Chart** on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality, for example, several missing values, many constant numerical values, or highly varying categorical values.

Use the **Data Explorer** to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Guided Exploration

This workflow defines a fully automated web-based application to show relevant visualization iteration after iteration. The workflow was designed for data scientists to easily create a dashboard and find relationships between columns.

The Process Step by Step
1. Upload your data / Select one of the available datasets
2. Inspect the visualizations in "Guided Exploration" Wrapped Metanode
3. Select at the bottom of the View the columns you would like to remove

Reference Column Filter

This node allows columns to be filtered from the first table using the second table as reference table. Expanding on the dialog settings, either the columns from the reference table are included or excluded in the output table.

Dialog Options

Ex/Include columns from reference table
Includes or excludes all columns from the reference table into the resulting table from the first input.

Ensure compatibility of column types
Ensures that the matching columns don’t only have the same name but also the same type. Columns are only included or excluded if the column type of the first table is a super-type of the column type from the second table. If not selected, only the column names need to match.

Guided Visualization

This workflow defines a fully automated web-based application to select, customize, and download a number of visualizations. The workflow was designed for business analysts to easily create charts by applying their domain knowledge. Each of the wrapped metanodes outputs a web page with which the business analyst can interact.

The Process Step by Step
1. Select the columns to visualize (maximum 3)
2. Convert the domain of the columns (OPTIONAL)
3. Customize the visualizations interactively
4. Download the images of the customized charts
<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Index</th>
<th>Color Handler</th>
<th>Size Handler</th>
<th>Filter Handler</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
<th>Value 7</th>
<th>Value 8</th>
<th>Value 9</th>
<th>Value 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month</td>
<td>String</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfWeek</td>
<td>String</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Hour</td>
<td>Number</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Minute</td>
<td>Number</td>
<td>3</td>
<td>8</td>
<td>6</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfMonth</td>
<td>Number</td>
<td>4</td>
<td>9</td>
<td>7</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>5</td>
<td>10</td>
<td>8</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>WeekDay</td>
<td>Number</td>
<td>6</td>
<td>11</td>
<td>9</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>MonthDay</td>
<td>Number</td>
<td>7</td>
<td>12</td>
<td>10</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>8</td>
<td>13</td>
<td>11</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>9</td>
<td>14</td>
<td>12</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>10</td>
<td>15</td>
<td>13</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>11</td>
<td>16</td>
<td>14</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>12</td>
<td>17</td>
<td>15</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>13</td>
<td>18</td>
<td>16</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>14</td>
<td>19</td>
<td>17</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>DayOfYear</td>
<td>Number</td>
<td>15</td>
<td>20</td>
<td>18</td>
<td>false</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>
Guided Exploration

This workflow defines a fully automated web based application to show relevant visualization iteration after iteration. The workflow was designed for data scientists to easily create a dashboard and find relationships between columns.

Guided Visualization

This workflow defines a fully automated web based application to select, customize and download a number of visualizations. The workflow was designed for business analysts to easily create charts by applying their domain knowledge. Each of the wrapped metanodes outputs a web page with which the business analyst can interact.

The Process Step by Step

1. Upload your data / Select one of the available datasets
2. Inspect the visualizations in "Guided Exploration" Wrapped Metanode
3. Select at the bottom of the View the columns you would like to remove
4. Pre-Processing Guided Exploration
5. Reference Column Filter

The Process Step by Step

1. Select the columns to visualize (maximum 3)
2. Convert the domain of the columns (OPTIONAL)
3. Customize the visualizations interactively
4. Download the images of the customized charts

Reference Column Filter

This node allows columns to be filtered from the first table using the second table as reference table. Expanding on the dialog settings, either the columns from the reference table are included or excluded in the output table.

Dialog Options

In/Exclude columns from reference table includes or excludes all columns from the reference table into the resulting table from the first input.

Ensure compatibility of column types Ensures that the matching column don't take the same name but also the same type. Columns are only included or excluded if the column type of the first table is a super-type of the column type from the second table. If not selected, only the column names need to match.

Ports

Input Ports
0 Table from which columns are to be included or excluded
1 Table with the columns used as reference

Output Ports
0 Table with filtered columns
Guided Exploration for the Data Scientist

**Guided Exploration**
This workflow defines a fully automated web-based application to show relevant visualization iteration after iteration. The workflow was designed for data scientists to easily create a dashboard and find relationships between columns.

**Guided Visualization**
This workflow defines a fully automated web-based application to select, customize and download a number of visualizations. The workflow was designed for business analysts to easily create charts by applying their domain knowledge. Each of the wrapped metanodes outputs a web page with which the business analyst can interact.

**The Process Step by Step**
1. Upload your data / Select one of the available datasets
2. Inspect the visualizations in "Guided Exploration" Wrapped Metanode
3. Select, at the bottom of the View, the columns you would like to remove

**The Process Step by Step**
1. Select the columns to visualize (maximum 3)
2. Convert the domain of the columns (OPTIONAL)
3. Customize the visualizations interactively
4. Download the images of the customized charts

**Reference Column Filter**
This node allows columns to be filtered from the first table using the second as reference table. Depending on the dialog settings, either the columns from the reference table included or excluded in the output table.

**Dialog Options**
- In/Exclude columns from reference table
- Includes or excludes all columns from the reference table into the resulting table from the first input.

**Ports**
- **Input Ports**
  0 Table from which columns are to be included or excluded
  1 Table with the columns used as references
- **Output Ports**
  0 Table with filtered columns
Guided Exploration

This workflow defines a fully automated web based application to show relevant visualization iteration after iteration. The workflow was designed for data scientists to easily create a dashboard and find relationships between columns.

The Process Step by Step
1. Upload your data / Select one of the available datasets
2. Inspect the visualizations in "Guided Exploration" Wrapped Metanode
3. Select at the bottom of the View the columns you would like to remove

Guided Visualization

This workflow defines a fully automated web based application to select, customize and download a number of visualizations. The workflow was designed for business analysts to easily create charts by applying their domain knowledge. Each of the wrapped metanodes outputs a web page with which the business analyst can interact.

The Process Step by Step
1. Select the columns to visualize (maximum 3)
2. Convert the domain of the columns (OPTIONAL)
3. Customize the visualizations interactively
4. Download the images of the customized charts
Guided Exploration

This workflow delivers a fully automated web-based application for the KNIME WebPortal to show relevant visualizations based on data. The workflow was designed to analyze data and create a dashboard and filter relationships between columns. The workflow named "Guided Exploration" is stored in source code and outputs a web page at each iteration. If the data content exceeds columns, the dashboard is updated in the next iteration. Only the first five iterations can be executed from KNIME Analytics Platform using the command: "Do one loop step (Ctrl + A + F5)."

The Process Step by Step:
1. Upload your data. Select one of the available datasets.
2. Inspect the visualizations in "Guided Exploration". Use the "Wrap" button. If you would like to improve the visualizations, simply click on the "Wrap" button and click on the "View" button.
3. To do one step (Ctrl + A + F5) on the Recursive Loop End.
4. Access the "Wrapped" Metanode's view.

Ports
- **Input Ports**
  - A table listing the columns to visualize.
  - A table with the data. This table should have all the columns mentioned in the other inputs of the Wrapped Metanode.
  - A table including all columns with different scoring related to the purpose of the web application; these columns can be used as input to visualize the charts.
- **Output Ports**
  - The same data that was used as input to visualize the charts.
  - Those are the columns the user selected. You can do different things with this output depending on the purpose of the web application.
Guided Visualization & Exploration

This is a temporary copy of "02_Guided_Exploration". Use "Save As..." to store it to your local workspace or if you are currently logged in.

Guided Exploration

The workflow defines a fully automated web-based application for the KNIME WebPortal to show relevant visualizations for the page after execution. The workflow was designed for data scientists to easily create dashboards and add relationships between columns. The wrapped node "Guided Exploration" is placed in an execute loop and outputs a web page at each iteration. If the data subset of the page is updated, the dashboard is updated in the next iteration. Only the first five iterations can be executed using KNIME Analytics Platform.

The Process: Step by Step

1. Upload your data / Select one of the available datasets
2. Inspect the visualizations in "Guided Exploration" in the workflow engine
3. Select at the bottom of the View the column you would like to remove
4. Right-click the node and click "View"
5. To use one step ("View + Add") on the Recursive Loop End
6. Reopen the wrapped node view

Guided Exploration Dashboard

Pre-Processing

Boundary Exploration

Reference Column Filter

CASE Switch Data Split

Recursive Loop Start

sandy check

end loop

Recursive Loop End

Variable to Table Row
Data Upload

Use the form below to upload the data you would like to visualize.

Guide

Upload the dataset to visualize. The file must be in CSV format. The file will be uploaded to the server for further processing.
Guided Exploration

The visualizations below were automatically generated to provide insight into your dataset. They were selected by detecting properties and relationships within and between data columns.

Univariate Column Detection

Each visualization takes into account a single column detected by a different statistical test.

- **Most Uniform Histogram**: This shows a histogram of the column that is most uniformly distributed. Such a distribution will appear mostly "flat" to the viewer, and often occurs when every value is (mostly) unique - for example, in a row identifier column.

- **Most Skewed Histogram**: This shows the histogram of the column with the largest skewness. Skewed distributions are asymmetric relative to the mean of the distribution, and the skewness measures the degree of that asymmetry.

- **Most Spiked Histogram**: This shows the histogram of the column with the largest difference between a particular binned value and the average binned value. If the difference is sufficiently large, it may indicate an anomalous feature in your data, which should be investigated further.

- **Most Leptokurtic Box Plot**: This shows a box plot of the column with the largest kurtosis, which is (in some sense) a measure of how heavy-tailed a particular distribution is. The column shown here will have usually have a large number of outlying values relative to other features in the dataset.

Multivariate Column Detection

- **Most Correlated Scatter Plot**: This shows the two numeric columns most positively correlated with one another, where the first column increases, the second column does also. Highly positively correlated columns may indicate that one of the features is a candidate for removal.
Removing Columns from the Dashboard

Use the Horizontal Bar Chart on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality. For example, several missing values, many constant numerical values, or highly varying categorical values.

Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Horizontal Bar Chart to Remove Columns

Ctrl + Left Click Bins of Columns to be Flagged

- Year
- Cancelled
- CancellationCode
- SecurityDelay
- Diverted
- WeatherDelay
- LateAircraftDelay
- CarrierDelay
- NASDelay
delay_class
- TailNum
- FlightNum
- AirTime
- Month
- DayTime
- CRSDate
- TEMP
- Distance
- DayOfWeek
- DepDelay
- ActualDepTime
- DepTime
- UniqueCarrier
- Origin
- AddDelay
- TaxIn
- Distance
- DistanceActual
- CRSActualTime
- ArrTime
- Dest
- ArrDelay
- DayOfMonth
date

Removing Columns from the Dashboard

Use the Horizontal Bar Chart on the side to flag columns you would like to remove in the next iteration by using Ctrl + click. The columns in the bar chart are sorted by a ranking. A higher rank indicates a column with relatively poor quality. For example, several missing values, many constant numerical values or highly varying categorical values.

Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Guided Exploration

The visualizations below were automatically generated to provide insight into your dataset. They were selected by detecting properties and relationships within and between data columns.

Univariate Column Detection

Each visualization takes into account a single column detected by a different statistical test.

- Most Uniform Histogram: This shows a histogram of the column that is most uniformly distributed. Such a distribution will appear mostly "flat" to the viewer, and often occurs when every value is (mostly) unique - for example, in a row identifier column.
- Most Skewed Histogram: This shows the histogram of the column with the largest skewness. Skewed distributions are asymmetric relative to the mean of the distribution, and the skewness measures the degree of that asymmetry.
- Most Spiked Histogram: This shows the histogram of the column with the largest difference between a particular binned value and the average binned value. If the difference is sufficiently large, it may indicate an anomalous feature in your data, which should be investigated further.
- Most Leptokurtic Box Plot: This shows a box plot of the column with the largest kurtosis, which is (in some sense) a measure of how heavy-tailed a particular distribution is. The column shown here will usually have a large number of outlying values relative to other features in the dataset.

Multivariate Column Detection

- Most Correlated Scatter Plot: This shows the two numeric columns most positively correlated with one another, where the first column increases, the second column does also. Highly positively correlated columns may indicate that one of the features is a candidate for removal.
Cut + Left Click Bins of Columns to be Flagged

- OnAir_Cas8
- TailNum
- FlightNum
- AirTime
- Meta
- DepTime
- CRSArrTime
- TaxiIn
- Distance
- DayOfWeek
- CRSDepTime
- UniqueCarrier
- Origin
- TaxiOut
- ActualElapsedTime
- CRSElapsedTime
- AirTime
- Dest
- date

Data Explorer

Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
**Data Explorer**

- **Crtl + Left Click Bins of Columns to be Flagged**
  - Columns are sorted by a ranking. A higher rank indicates a column with relatively poor quality. For example, several missing values, many constant numerical values, or highly varying categorical values.
  - Use the Data Explorer to explore both numerical and nominal columns. Clicking on a column name will provide additional information about the data in that column, including summary statistics and a histogram.
Guided Exploration

This workflow defines a fully automated web-based application for the KNIME WebPortal to show relevant visualization features after iteration. The workflow was designed for data science to easily create dashboards and find relationships between columns. The wrapped node is "Guided Exploration" and it loads an automatic loop to open a web page at each iteration. If the data scientist discovers columns on the dashboard is updated in the next iteration. Only the first two iterations can be executed from KNIME Analytics Platform using the command "Do one step" (Ctrl+Alt+F9).

The Process Step by Step:
1. Upload your data / Select one of the available datasets
2. Import the visualizations in "Guided Exploration" node
3. Select all the features in the view of the column you would like to remove
4. Select the columns and click on "Remove"
5. "Do one step loop" (Ctrl+Alt+F9) on the Recursive Loop End
6. Repeat the "Ipt" step until you reach the desired number of iterations

Guided Exploration Dashboard

Pre-Processing

Column Selection

Filtering

Variable Transformation

Variable Table

Reference Column Filter

Output Ports

0 (same data that was used as input to visualize the charts)
1 (these are the columns the user selected. You can do different things with these output depending on the purpose of the web application).
This is a job running on KNIME Server (knime-server).

1. Job started by WebPortal. CLI operations are not allowed. Nodes following the currently active wrapped metanode (WebPortal page) are not executed.

Guided Exploration

The workflow defines a fully automated Web-based application via the KNIME WebPortal to show relevant visualizations and actions on demand. The workflow was designed for data scientists to easily create dashboards and Visualizations between columns. The wrapped metanode “Guided Exploration” initiates an executable loop and outputs a web page at each iteration. If the data scientist discards columns, the dashboard is updated in the next iteration. Only the first five columns can be executed from KNIME Analytics Platform using the command “Do one loop step” (D1 + All + F5).

The Process Step by Step

1. Upload your data / Select one of the available datasets
2. Inspect the visualizations in “Guided Exploration” Wrapped Metanode
3. Select the visualizations you want to include in the report
4. Apply the settings and Close the View
5. To see one loop step (D1 + All + F5) on the Recursive Loop End
6. Reopen the Wrapped Metanode view
This is a job running on KNIME Server (knime-server).

Job started by 'WebPortal'. Edit operations are not allowed. Nodes following the currently active wrapped metanode (WebPortal page) are not executed.

Guided Exploration

The workflow defines a fully automated web-based application for the KNIME WebPortal to show relevant visualizations for the workflows. The workflow is designed for data transformaton tasks. It creates a dashboard and a list of nodes between columns. The wrapped metanode (Guided Exploration) runs in a loop and outputs a web page at each iteration. If the data contains an index column, the dashboard is updated to the next iteration. Only the last iteration can be executed from KNIME Analytics Platform using the command "Do one loop step" (Ctrl + Alt + F6).

The Process Step by Step

1. Upload your data / Select one of the available datasets
2. Inspect the visualizations in "Guided Exploration" wrapped metanode
3. Select at the bottom of the workflow the column you would like to remove
4. Apply the settings and close the view
5. "Do one loop step" (Ctrl + Alt + F6) on the Recursive Loop End
6. Open the Wrapped Metanode view

Guided Exploration Dashboard

Pre-Processing

- Import Data
- Filter Data

Data Exploration

- Summary Statistics
- Correlation Matrix

Reference Column Filter

- Save Reference Columns
- Remove Reference Columns

Configurations

- Execute
- Cancel
- Reset
- Edit Node Descriptions...
This is a metanode of a job running on KNIME Server.

Job started by 'WebPortal'. Edit operations are not allowed. Nodes following the currently active wrapped metanode (WebPortal page) are not executed.
Guided Exploration Highlights

• No coding was required to create this application/dashboard
• Quite useful to use Nested Wrapped Metanodes
• In the first iterations: obvious anomalies and relationships
• After some iterations: strong and unexpected patterns
• You can customize the dashboard to your needs
• You can customize the recursive loop as well
  – Adding additional Composite Views
  – Adding more functions
Visualizing with Guided Analytics:
for Muggles AND Wizards.

Phil Winters
Maarit Widman
Paolo Tamagnini
Scott Fincher
Guided Analytics Examples available on the KNIME Workflow Hub

• Guided Analytics Use Case Examples
  – Inventory analysis
  – Recommendation Engine for Retail
  – Price Benchmarking
  – Customer Experience and Sentiment Analysis
  – Image Recognition for Retail
  – Visualizing GIT Statistics
  – Teacher Bots
Guided Analytics Blueprints on the KNIME Workflow Hub

• Complete extendable applications

• Machine Learning Automation
• Exploration
• Visualization

What’s in common among these blueprints?
Guided Analytics Blueprints: Features in Common

• Consistent Workflows
  – Layout
  – Documentation
  – Style
  – Way of Working
  – Examples come with data

• Works on:
  – KNIME Analytics Platform
  – KNIME WebPortal via KNIME Server
Guided Analytics Blueprints: Features in Common

• Consistent look and feel
  – New layout panel for arranging elements
  – CSS classes specified in initial wrapped metanode, and passed downstream to subsequent nodes
  – HTML passed via flow variables (with CSS) such as header, as well as dynamically populating sidebar guide
Highlight: Dynamically populating sidebar

Guide

On this page we display automatically the most appropriate charts for your selection of columns. Feel free to interact with the JavaScript visualizations!

A configuration panel can be accessed in the top right corner of each chart. You can customize a number of things depending on the chart, such as the title, subtitle, legend, axis labels and other settings unique to the inspected visualization. Once you are done customizing the chart, your edits are applied to the visualizations by selecting Next.

Displayed Visualizations

- **Line Plot Moving Average**: Plots one or more numerical variables on the y-axis against a date on the x-axis. The line plot graphically represents the evolution of a time series. The moving average line shows the change of the prior 21-day average of the values in the numerical variable(s). The 21-day average is calculated by averaging the value of the current record and the 20 records preceding it.

- **Line Plot Cumulative Sum**: Plots one or more numerical variables on the y-axis against a date variable on the x-axis. The line plot graphically represents the evolution of a time series. The cumulative sum line shows how the total sum of the values in the numerical column accumulates over time.
Guided Analytics Blueprints: Features in Common

• Modular and Reusable “Applications” built with components to be used, extended, enhanced!

• Wrapped Metanodes: **Interaction**
  – *Upload*
  – *Select Columns*
  – *Select Domains*
  – *Download*
  – *Etc.*
Guided Analytics Blueprints: Features in Common

- Regular Metanodes: Logic/automation/organization functionality.
  - Example: Pre-Processing
    - Detect Visualizations
      - Activate Numerical Analysis
      - Activate Categorical Analysis
    - Feature Quality Calculations
      - Map Columns to Task
      - Calculate Feature Quality Metrics
What did we learn?

• Modularize with metanodes from the start
  – Massively speeds new application development through reuse

• Use Metanode template repository via KNIME Server!
  – Document metanodes
  – Versioning, instantiation, reuse.
  – (we didn’t only because of stand-alone public examples)
What did we learn?

• Think before you build!
  – People do the most amazing (and unexpected!) things...
  – More user interaction
    • more user choice
    • more potential for unknown behavior
    • more workflow interaction control logic may be required
  – Little user interaction
    • less user choice
    • less unknown behavior to manage
    • less workflow interaction controlling logic
  – Focus on only what is *absolutely necessary*

• Even with minimal interaction, you can still build very good looking applications!
What did we learn?

• First things first: mockup your workflow to represent automation and interaction in the application!
Workflows Available on Workflow Hub
Guided Analytics for....

• Other common (or uncommon) applications?
• Data Blending?

• Talk to us!

But most importantly: Try it yourself!
Visualizing with Guided Analytics: for End Users AND Experts.

Phil Winters
Maarit Widman
Paolo Tamagnini
Scott Fincher
The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.