IBM Cognos Analytics
Version 12.0.x

Explorations User Guide

Contents

Chapter 1. Getting started with Explorations

Explorations

Explore is a flexible workspace where you can discover and analyze data.

You can also explore an existing visualization from a dashboard or story. Uncover hidden relationships and identify patterns that turn your data into insights. Correlated insights are represented by a green icon with a number on the x-axis, y-axis, or the title of a chart.

Uploading data

Upload a data asset to your **My content** folder to use in your exploration.

Procedure

1. Click the **Open menu** icon , and then click **Upload data**.
2. Browse to where you saved your data asset and select it.
   After the upload is complete, the data asset appears in the **My content** folder.

Starting an exploration from an existing dashboard or story

When you are working on a dashboard or story, you can create or edit an exploration directly from a visualization.

About this task

Complete these steps to open a visualization in a new exploration or to add to an existing exploration:

Procedure

1. Open an existing dashboard or story.
2. Select a visualization.
3. Click the **Explore** icon in the toolbar.
4. Select **New exploration** or **Add to existing**.

Starting a new exploration from the Open menu

From the **welcome** page, you can start a new exploration from the **New** menu.

Procedure

1. Click the **Open menu** icon , and then click **New**.
2. Click **Exploration**.
3. Select a data source and click **Add**.
   A starting points page is generated from the data source you selected.

Starting a new exploration from a data asset in the Content page

You can select the **Action menu** on a recently used data asset in any of the **Content** folders.

Procedure

1. In a folder in the **Content** page, such as **Team content** or **My content**, locate the data asset, such as a data module, that you want to use as a source for your exploration.
2. Select the data asset checkbox, and from its **Action menu** ⋮, select **Create exploration**.

Adding a data source

Add a data source to your exploration to explore its data.

Procedure

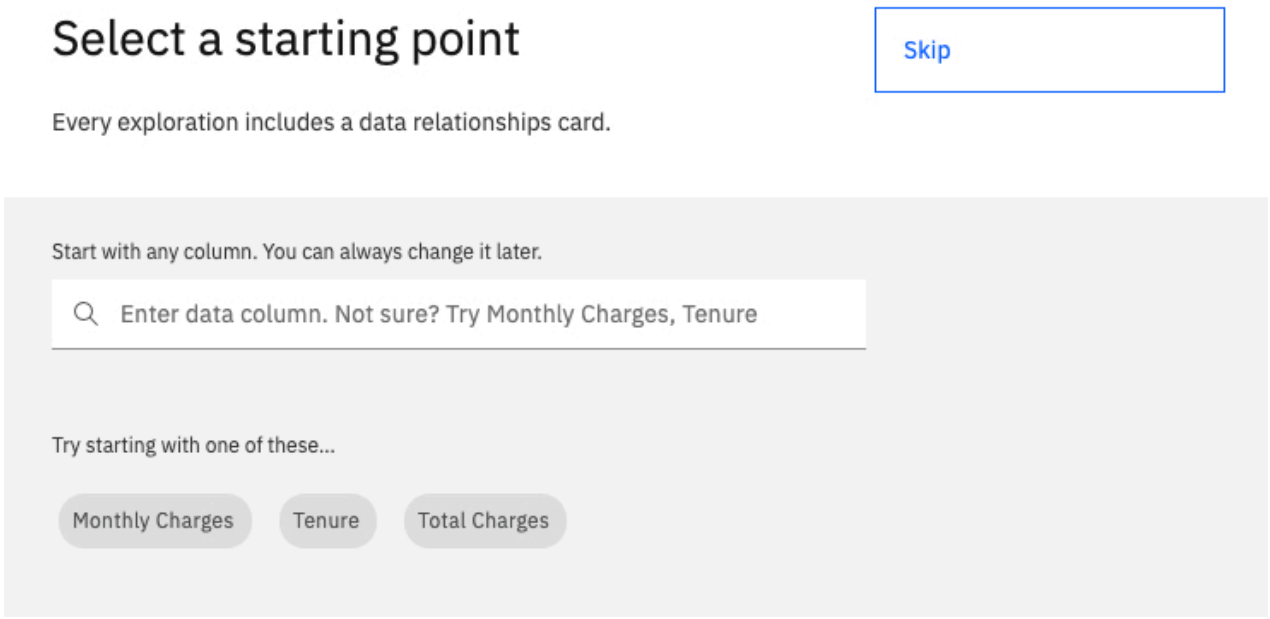
1. In the **Selected sources** pane, click the **Add a source** + icon.
2. Go to **My content** or the **Team content** folder, and select the data source that you want to add. Click **Add**.
3. Expand the data source in the **Selected sources** pane to see what's available.
4. Use the starting points page to generate a relationship diagram from your data.

Chapter 2. Exploring relationships

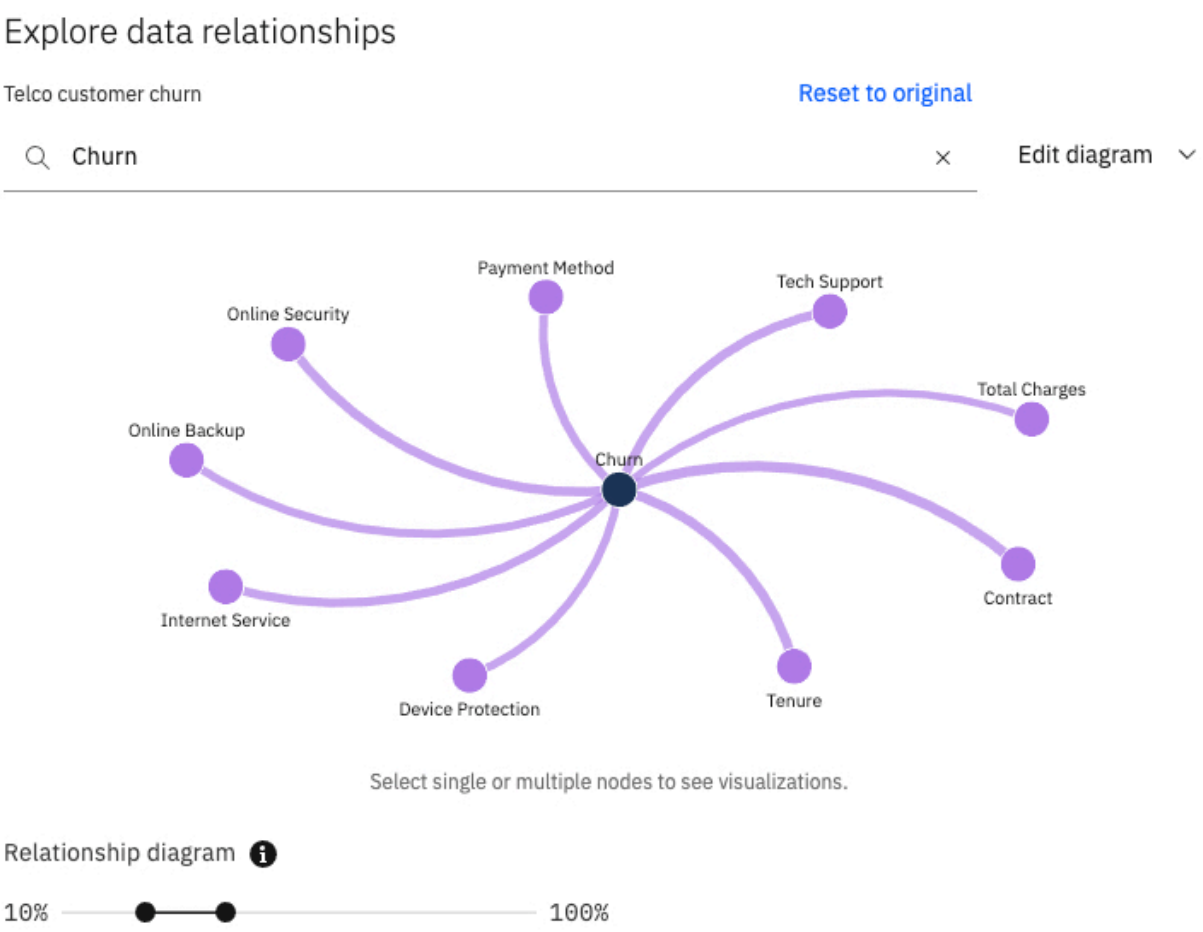
Explore relationships in your data

When you create an exploration, you can start from a data source. A starting points page is displayed with suggestions for how to get started.

You can type a column name that appears in your data source. Or, you can click one of the suggested columns that the system identifies as interesting. If you're not sure which column to start with, click **Skip** to see a relationship diagram with some suggested starting point visualizations.

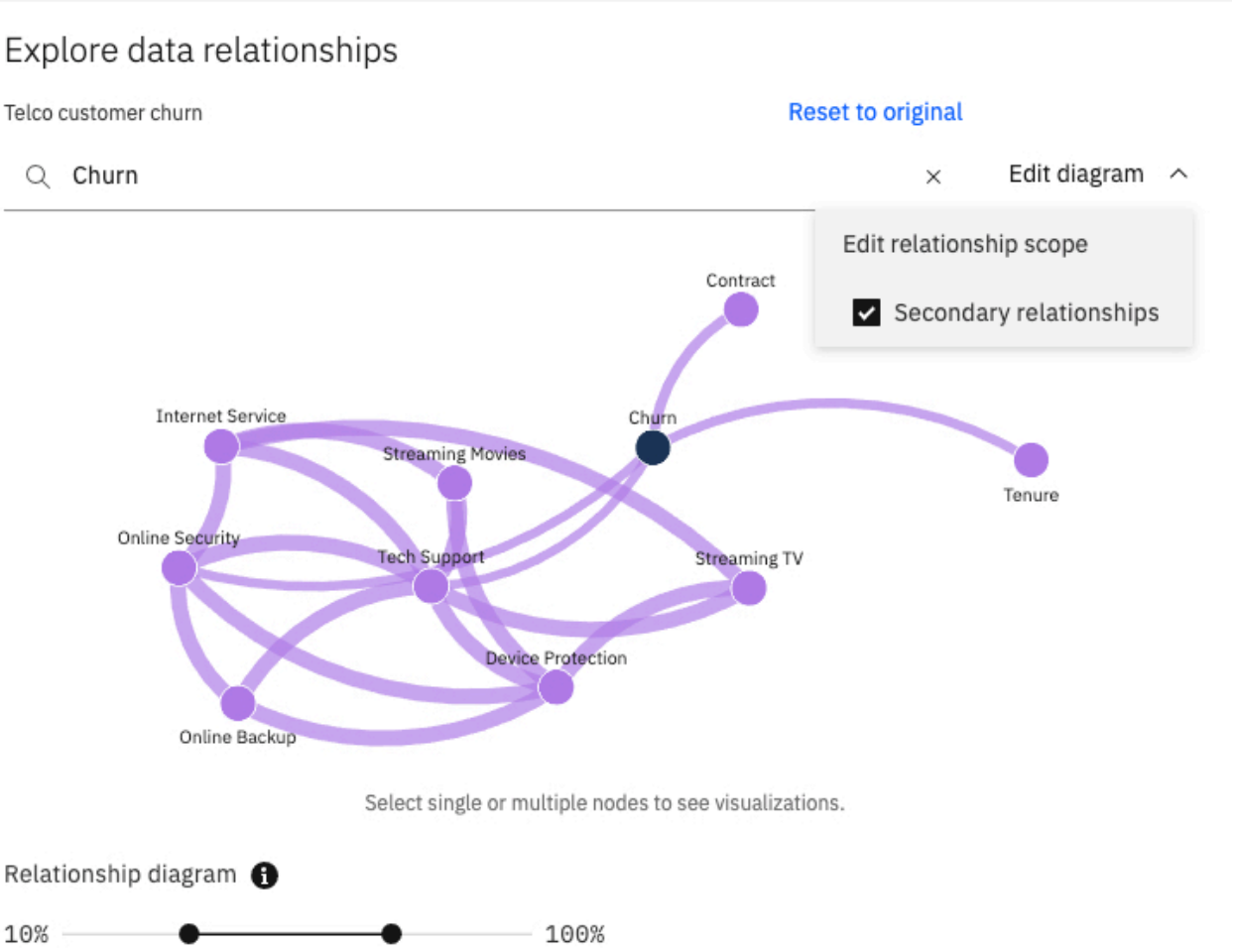


In the relationship diagram, the column that you start with is the prime focus and is represented by a dark blue node. Related fields are represented by purple nodes. Lines connect the nodes and represent relationships. The thickness of the line indicates the strength of the relationship.



The strongest primary relationships are displayed by default and are the direct relationships between the prime focus and the related fields. Secondary relationships are the relationships between other fields directly or indirectly related to the target.

To view both primary relationships and secondary relationships, select the **Secondary relationships** check box under **Edit diagram**.



The relationship diagram plots these fields based on a statistical evaluation of related items. The relationship diagram is not a picture of the data model. However, the model might be an influencing factor in the analysis. To improve performance when there are many rows in the data source, the analysis is based on a representative sample of the entire data.

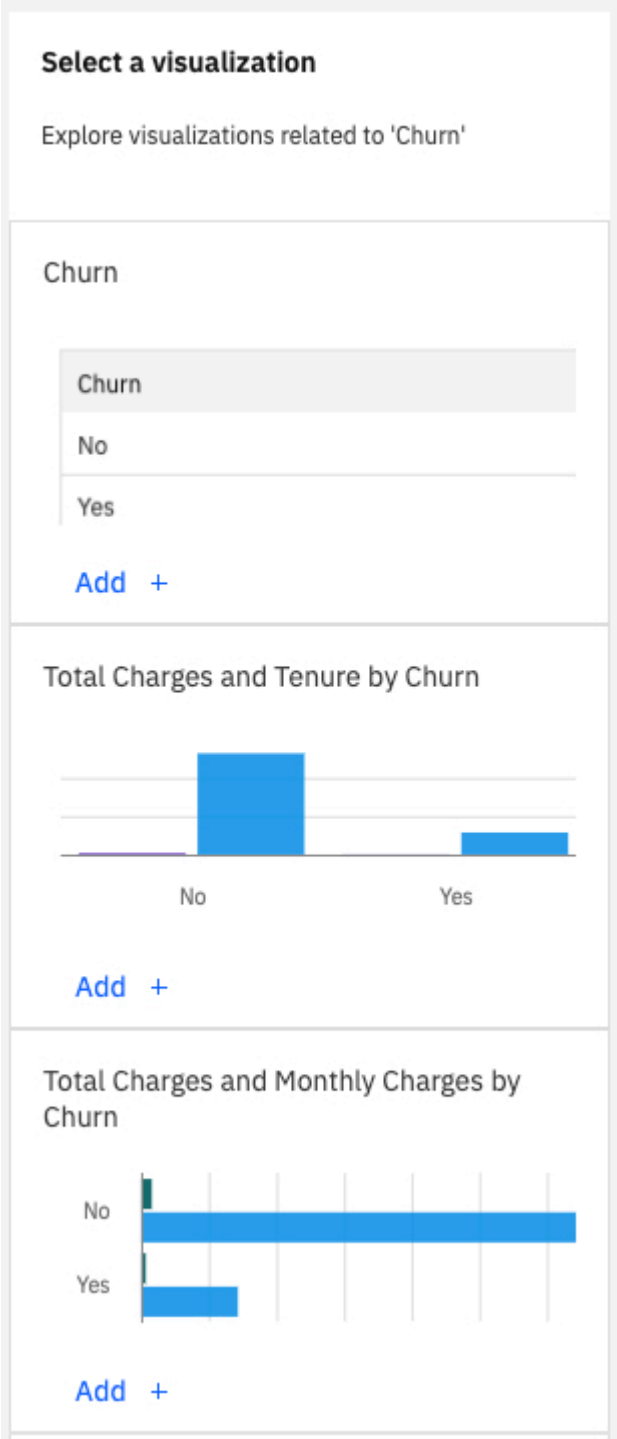
You can interact with the relationship diagram by selecting a node that you are interested in. As you do, the list of suggested starting point visualizations to the right of the diagram updates to include the nodes you selected. You can also use Ctrl+click to select multiple nodes.

Click **Reset to original** if you want to reset the scope and view of all the fields in the relationship diagram to the default setting.

Suggested starting points

Suggested starting point visualizations are displayed as thumbnails beside the relationship diagram. Select single nodes or multiple nodes in the relationship diagram to generate these visualizations.

Click a visualization if you want to add it to your exploration and view it at the same time. Click the plus icon + on the starting point visualization to add it to your list of cards and maintain the current view.



Opening the relationship diagram

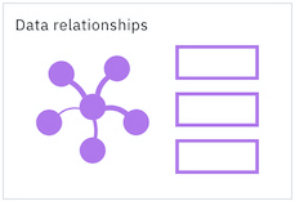
When you're viewing a visualization and you want to return to the relationship diagram, use the **Data relationships card** to return to the starting points view.

About this task

Complete the following steps to return to the starting points view to see a relationship diagram and the suggested starting points.

Procedure

1. Click **Explorations** icon in the side pane.

2. Click the **Data relationships card**.


Tip: The **Data relationships card** is also available from the **New card** menu on the toolbar.

Chapter 3. Visualizations

Visualizations

You can change the visualization type or change the columns that are used in the visualization.

Viewing cards in the navigation panel

View thumbnails of your visualizations, called cards, in the navigation panel to the left of the main view.

About this task

Cards are a collection of visualizations in your Exploration. Use the cards to open your visualizations to view details and modify them using the data slots.

Procedure

1. Click the **Explorations** icon in the side panel.
   
   Your cards are listed here.
2. Click the card that displays a visualization thumbnail that you want to open.
   The visualization opens in the main view.
3. View the generated text under the **Details** tab, or add more data items under the **Fields** tab.

Results

Viewing visualization details

When you open a visualization, it is displayed in the main exploration area. The exploration pane displays the **Details** tab, the **Fields** tab, and the **Properties** tab.

Visualization details

The **Details** tab displays text that is generated to describe aspects of the data represented in the visualizations. These details are not obvious from looking at the visualization. For example, the details might reveal an average of the values over time.

Fields

The **Fields** tab is where you can add columns to build and modify visualizations. Add a column to each mandatory field.

Properties

The **Properties** tab is where you can modify properties that apply to your visualizations.

Creating a single visualization

While you work with your exploration, you might decide that you need another visualization.

About this task

Complete the following steps to create a single visualization.

Procedure

1. On the toolbar, click **Create**.
2. Select **Single visualization**.
3. In the **Sources** window, expand the data asset that you want to use.

   If a different data asset is open, click the **Add a source** icon + next to the name of the data asset that is open.
4. To create a new visualization, complete one of the following actions:
   - Drag and drop data items onto the **Create a visualization** area.

     IBM® Cognos Analytics creates a visualization to match the data items. For example, when you add Year or Department, a table is created. Drag in a measure, such as Revenue, and a bar visualization is created.
   - Click **Choose a type** and select a visualization type. Then, add a data item to each field.

Creating a visualization using search in data fields

You can quickly build a visualization that uses search in data fields.

Procedure

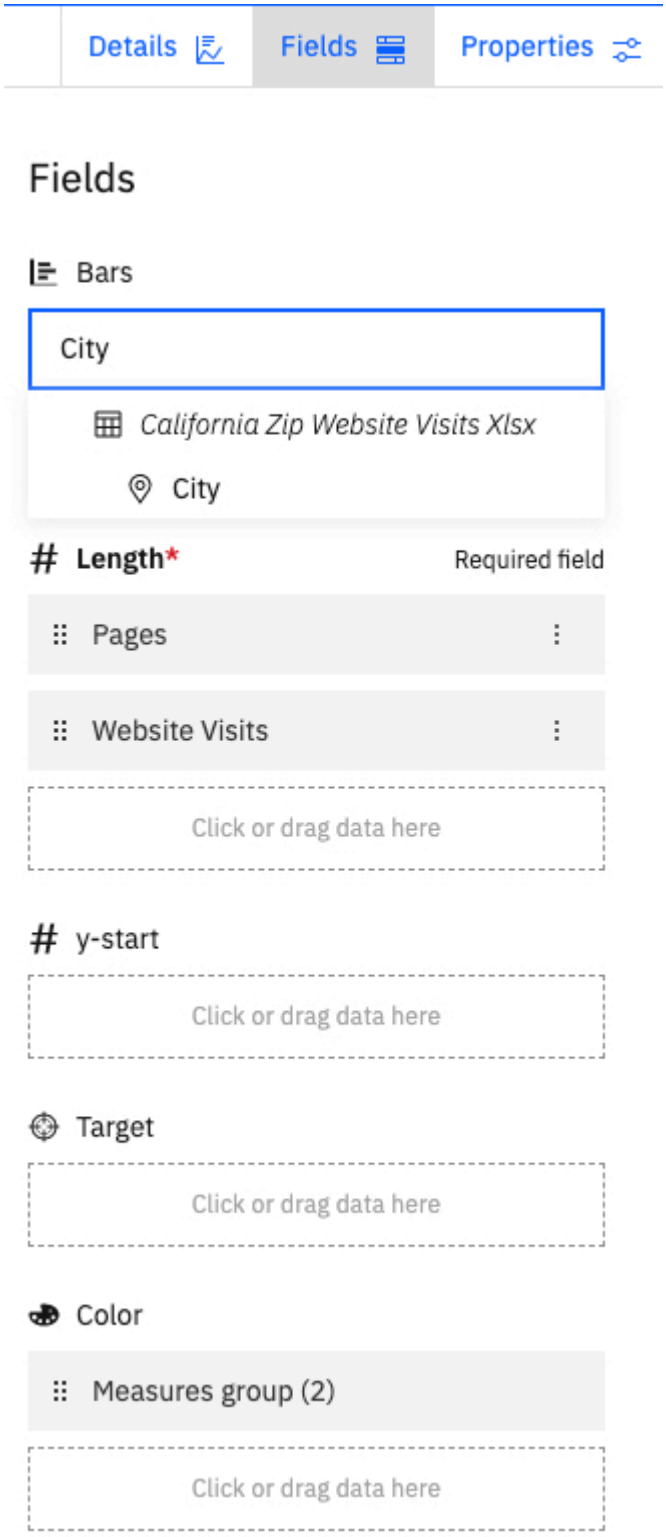
1. Click the **Explorations** icon in the side panel.

   Your cards are listed here.
2. Click the card that displays a visualization thumbnail that you want to open.

   The visualization opens in the main view.
3. Click the **Fields** tab.
4. Search for the data you are looking for. Click the data to populate the data field.

Results


Details
Fields
Properties
Fields
Bars
City
California Zip Website Visits Xlsx
City
Length*
Required field
Pages
Website Visits
Click or drag data here
y-start
Click or drag data here
Target
Click or drag data here
Color
Measures group (2)
Click or drag data here

Comparing two visualizations

You can create your own comparison to analyze the data between two visualizations. Or, you can start with a recommended comparison. In either case, a summary of key information and differences between the two visualizations is generated.

About this task

Complete the following steps to create a comparison between two visualizations.

Note: When you create a new visualization, you can select **Compare visualization** to get a blank comparison card with two slots for visualizations.

Procedure

1. Click the **Explorations** icon in the side pane.
   The **Cards** pane opens.
2. Select a card to create a comparison.
   A visualization is displayed.
3. On the toolbar, click **Compare**.
   The **How do you want to compare?** page is displayed with guidance on how to create your own comparison or start with a recommendation.
4. Click the plus icon + on a card thumbnail to add it to the list of cards in the navigation panel. Or, click the card thumbnail to add the new card and immediately view it.
5. Optionally, modify the data in one visualization to compare with the other visualization.
   a) Select one of the two visualizations.
   b) In the **Fields** tab, modify the visualization in some of the following ways, for example:
      - Remove filters.
      - Show top or bottom count.
      - Remove data items.
      - From the **Sources** pane, add new data items or filters. Or, use search in the data fields. For more information, see "Creating a visualization using search in data fields" on page 10.

Comparing two data points on a visualization

You can select two data points on an existing visualization and compare the data.

Procedure

1. Click the **Explorations** icon in the side pane.
   The **Cards** pane opens.
2. From the **Cards** pane, select the card that displays a visualization thumbnail that you want to open.
   The visualization opens in the main view.
3. Select two data points on the visualization by pressing **Ctrl** and clicking the two points.
4. Right-click one of selected points and then click **Compare by**.

   Note: To compare more than two points in the visualization, right-click one of the selected points and then click **Show by**.
5. Type a column to compare the two data points.
   A table displays information about how the two data points compare to each other.

Advanced data analytics

IBM Cognos Analytics is a business intelligence tool for managing and analyzing data. It includes various self-service features that make it possible for users to prepare, explore, and share data. As part of this offering Cognos Analytics includes a number of predictive, descriptive, and exploratory techniques, also known as numeric intelligence. Cognos Analytics uses many statistical tests to analyze your data. It is important to understand the definitions of these tests as they apply to Cognos Analytics.

For more information, see the *IBM Cognos Analytics Dashboard and Stories Guide*.

Choosing a different visualization type

Visualizations communicate comparisons, relationships, and trends. They emphasize and clarify numbers. To choose a visualization type, consider what you want the visualization to illustrate and what appeals to the audience for the visualization.

Before you begin

For more information on visualization types, see the visualization documentation in the *IBM Cognos Analytics Dashboards and Stories User Guide*.

Procedure

1. From the **Cards** pane, select the card that represents the visualization you want to open.
2. Click the **Choose visualization type** icon  in the toolbar.
3. Click the visualization type that you want to use.

   Notice how each visualization type communicates data differently. For example, use a bar, column, or line visualization to compare a set of values. Use a line or area visualization to track relationships. Use a tree map or pie visualization to see the parts of a whole.

Area

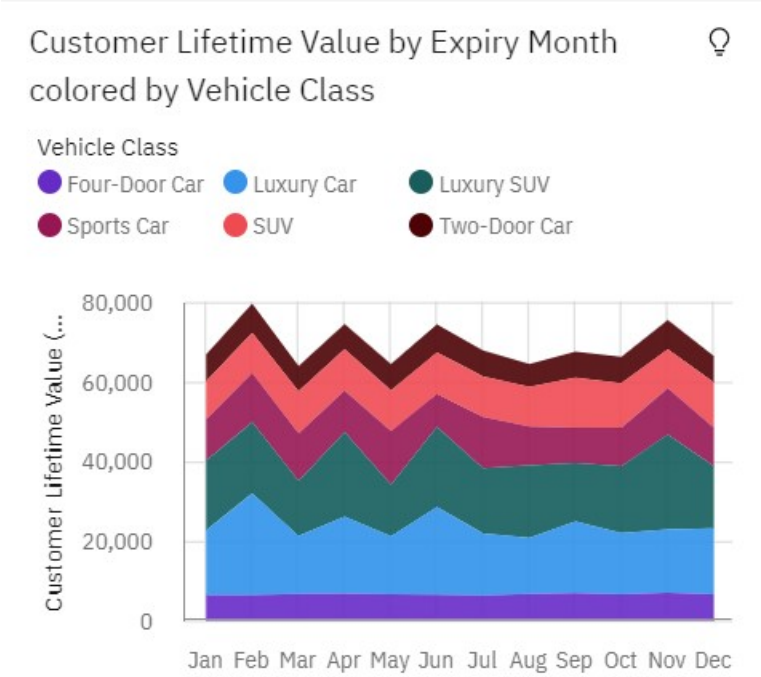
Use an area visualization to emphasize the magnitude of change over time.

Area charts are like line charts, but the areas below the lines are filled with colors or patterns. Stacked charts are useful for comparing proportional contributions in a category. They plot the relative value that each data series contributes to the total.

Because an area visualization stacks the results for each column or item, the total of all results is easily seen.

For example, an area visualization is excellent for looking at revenue over time across several products.

For example, this area visualization shows the customer lifetime value for each vehicle class per month. Because the area visualization stacks the results, you see the totals for each month.



The area visualization was created by dragging the following data items from the Sources panel:

- Drag **Expiry Month** type onto the **x-axis** field.
- Drag **Vehicle Class** onto the **Color** field.
- Drag **Customer Lifetime Value** onto the **y-axis** field

Samples

You can see an example of a word cloud visualization in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Customer lifetime value analysis**.

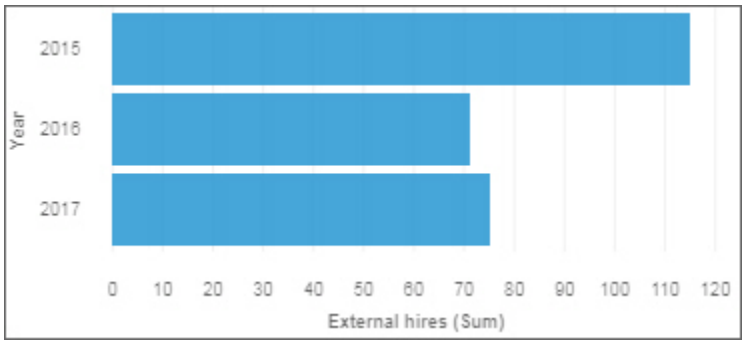
If any of the sample objects are missing, contact your administrator.

Bar

Use a bar visualization to compare values by one or more columns, such as sales for products or sales for products each month.

Bar visualizations use horizontal data markers that are arranged in groups to compare individual values. You can use bar visualizations to compare discrete data or to show trends over time.

A bar visualization can show change over a specific time period or can compare and contrast two or more columns in a time period or over time. If there are so many bars that the labels are impossible to read, filter the data to focus on a subset of the data or use a tree map.



Use the **Target** field to show measures that need to be compared against a target value.

Use the **y-start** field to define where the measure must start.

Box plot

You can use box plots for identifying outliers and for comparing distributions.

You can create a box plot to show the median, quartiles, and outlier and extreme values for a variable. The inter-quartile range is the difference between the 75th and 25th percentiles and corresponds to the length of the box. The middle line is the 50th percentile.

Above and under each box whiskers give additional information about the spread of the data.

Far out values are represent by adding "o" signs beyond the whiskers.

The mean score in a box plot is presented by a "+" sign.

Use the key field in a box plot visualization to determine for which items you want to identify outliers and compare distributions. In our example, we use **Province or State** from the GOSales data source.

The next box plot shows the gross profit statistics for various markets.

Note: Drill-through is not available for a box plot visualization.

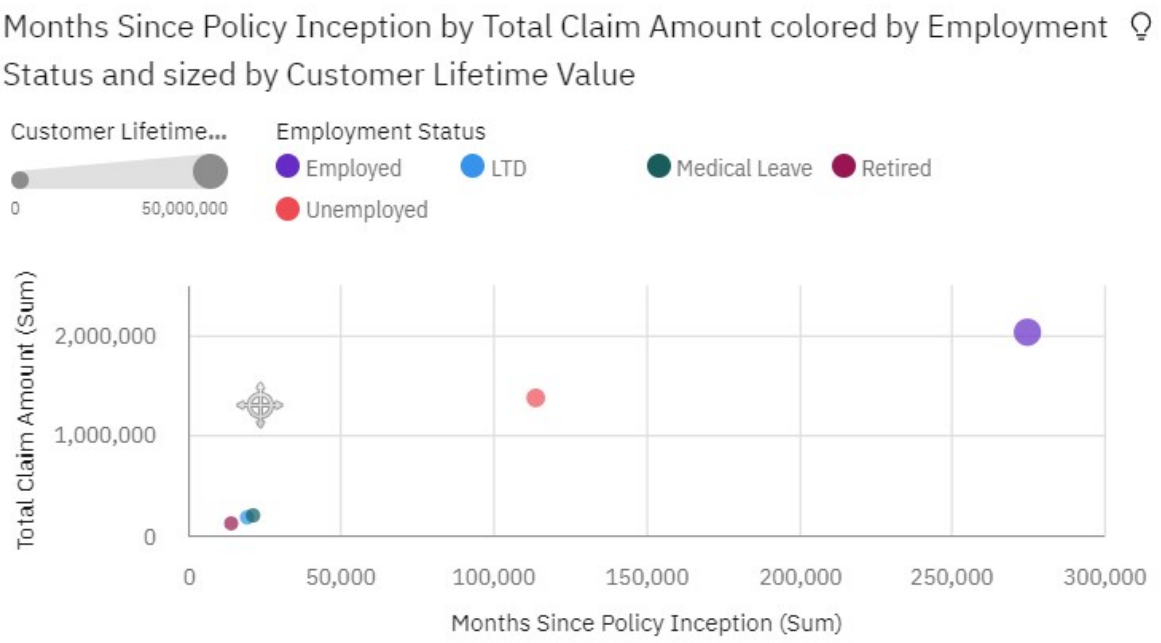
Bubble

Use a bubble visualization to show relationships among columns that contain numeric values, such as revenue and profit.

A bubble visualization uses data points and bubbles to plot measures anywhere along a scale. One measure is plotted along each axis. The size of the bubble represents a third measure. Use bubble visualizations to represent financial data or any data where measure values are related.

The bubbles are in different sizes and colors. The x-axis represents one measure. The y-axis represents another measure, and the size of the bubbles represents the third measure. In the example shown below, color is represented by an identifier.

The example that is shown represents the months since the policy inception.



Create the Bubble visualization by dragging the following data items from **Customer Analysis** in the **Sources** pane :

- Drag **Months Since Policy Inception** onto the **x-axis** field.
- Drag **Total Claim Amount** onto the **y-axis** field.
- Drag **Customer Lifetime Value** onto the **Size** field.
- Drag **Employment Status** onto **Color**

You can customize the bubble chart. For example, to make the x-axis of the bubble chart appear as it does in the sample, do the following steps:

1. Click the visualization, then in the **Data** pane, click the **<Total Claim Amount>** data item.
2. Click
3. Next to **Data format**, click and set the following options:
   - **Format type**: Currency
   - **Currency symbol**: K
   - **Currency symbol position**: End
   - **Number of decimal places**: 0
   - **Scale**: -3 (this presents values in thousands).
4. Click **OK**.

To change the size of the visualization, click the visualization, then set the following option in the properties pane:

- **Size - Width**: 700 px, **Height**: 300 px

Click to close the **Properties** pane.

Samples

You can see examples of visualizations in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Standard reports** > **Customer lifetime value analysis**.

If any of the sample objects are missing, contact your administrator.

Bullet

Use bullet charts to show measures that need to be compared against a target value.

In a call center, a bullet chart can be used to measure metrics like call volume, call answer speed, and percentage of abandoned calls.

In manufacturing, a bullet chart can be used to track metrics like number of defects and orders that are shipped.

In a fitness context, a bullet chart can be used to measure metrics like steps that are taken and calories that are burnt.

Bullet visualizations compare an actual measure (the bullet) to targeted measure (the target). Bullet visualizations also relate the compared measures against colored regions in the background that provide more qualitative measurements, such as good, satisfactory, and poor. Bullet visualizations can be shown at small sizes while still effectively conveying information.

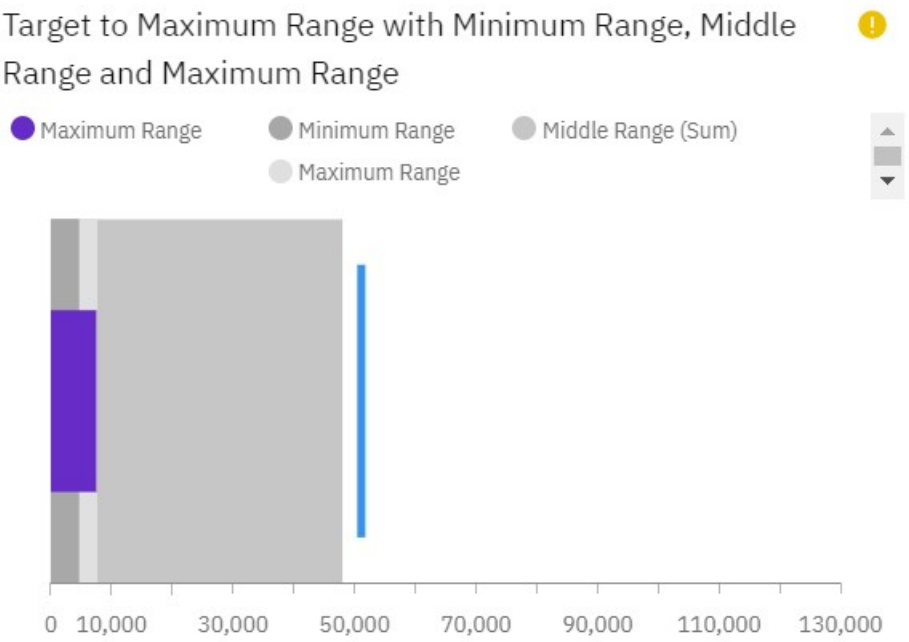
A bullet visualization features a single, primary measure. For example, current year-to-date revenue. And compares that measure to one or more other measures to enrich its meaning. For example, compared to a target. The primary measure is displayed in the context of a qualitative range of performance, such as poor, satisfactory, and good.

If you select a bullet visualization, then specify the following fields:

- The **Actual bar** field specifies the actual measure.
- The **Target** field specifies the target measure.
- The **Minimum range** field specifies the minimum qualitative range.
- The **Middle range** field specifies the middle qualitative range.
- The **Maximum range** field specifies the higher qualitative range.

Note: Drill-through is not available for a bullet visualization.

Make sure the minimum, middle, and maximum ranges relate to the actual and target measure.



The bullet visualization was created by dragging the following data items from the Sources panel:

- Drag **Minimum Range** onto the **Minimum range** field.
- Drag **Middle Range** onto the **Middle range** field.
- Drag **Maximum Range** onto the **Maximum range** field
- Drag **Maximum Range** onto the **Actual bar** field
- Drag **Target** onto the **Target** field

Samples

You can see an example of a bullet visualization in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Customer lifetime value analysis**.

If any of the sample objects are missing, contact your administrator.

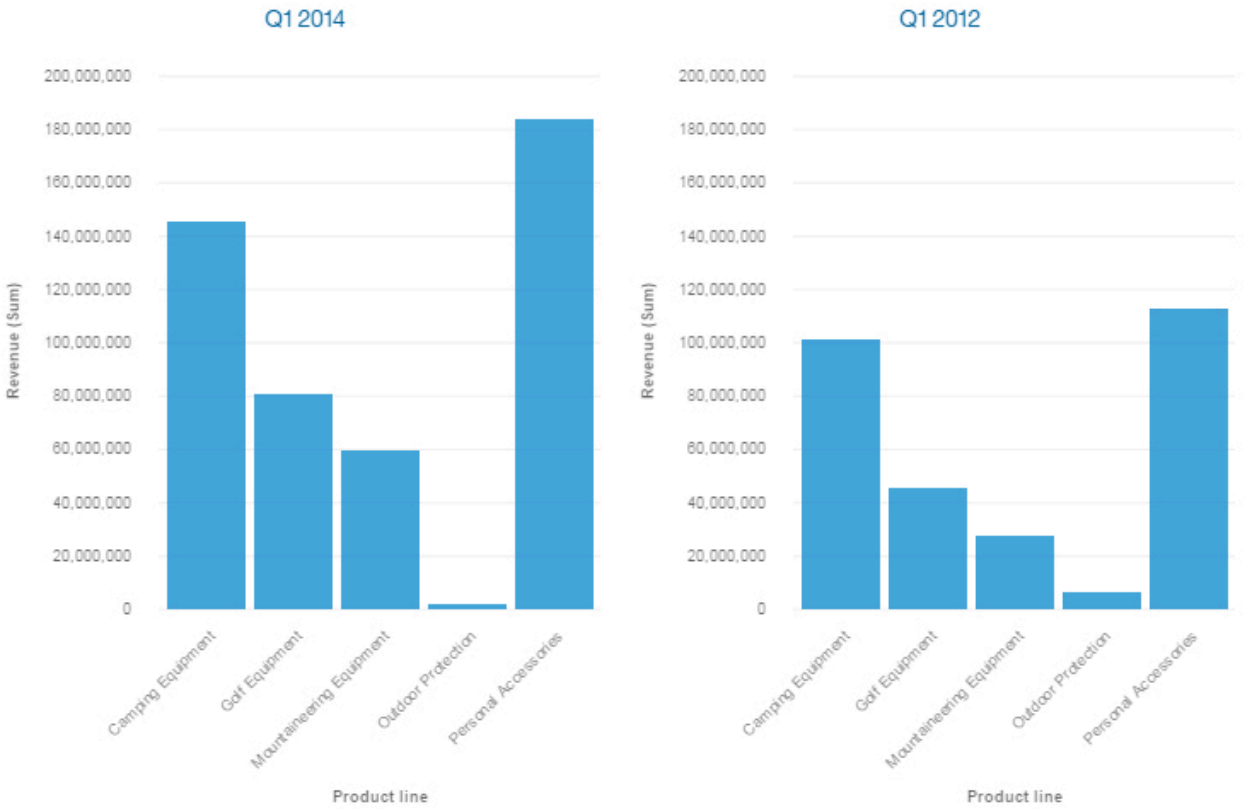
Column

Use a column visualization to compare values by one or more columns, such as sales for products or sales for products each month.

Column visualizations use vertical data markers that are arranged in groups to compare individual values. Use column visualizations to compare discrete data or show trends over time.

A column visualization shows change over a specific time period or can compare and contrast two or more columns in a time period or over time. If there are so many bars that the labels are impossible to read, filter the data to focus on a subset of the data or use a tree map.

For example, revenue for each product line is grouped by quarter, which emphasizes performance in each quarter.



Use the **Target** field to show measures that need to be compared against a target value.

Use the **y-start** field to define where the measure must start.

Crosstab

Use a crosstab when you want to show the relationships between three or more columns. Crosstabs show data in rows and columns with information summarized at the intersection points.

For example, this crosstab shows the income for states by gender.

Income for State and Gender

		Income
Arizona	F	33,258,643
	M	30,442,757
	Summary	**63,701,400**
California	F	59,787,475
	M	58,523,207
	Summary	**118,310,682**
Nevada	F	18,390,070
	M	15,451,922
	Summary	**33,841,992**
Oregon	F	50,521,333
	M	47,165,161

Starting from Cognos® Analytics version 11.1.4, you can drag data from the **Selected sources** pane and insert data in a column/row or drop the data on top of existing data to replace it.

Data player

Use a data player to see an animation of the impact of a column on the other visualizations.

Decision tree

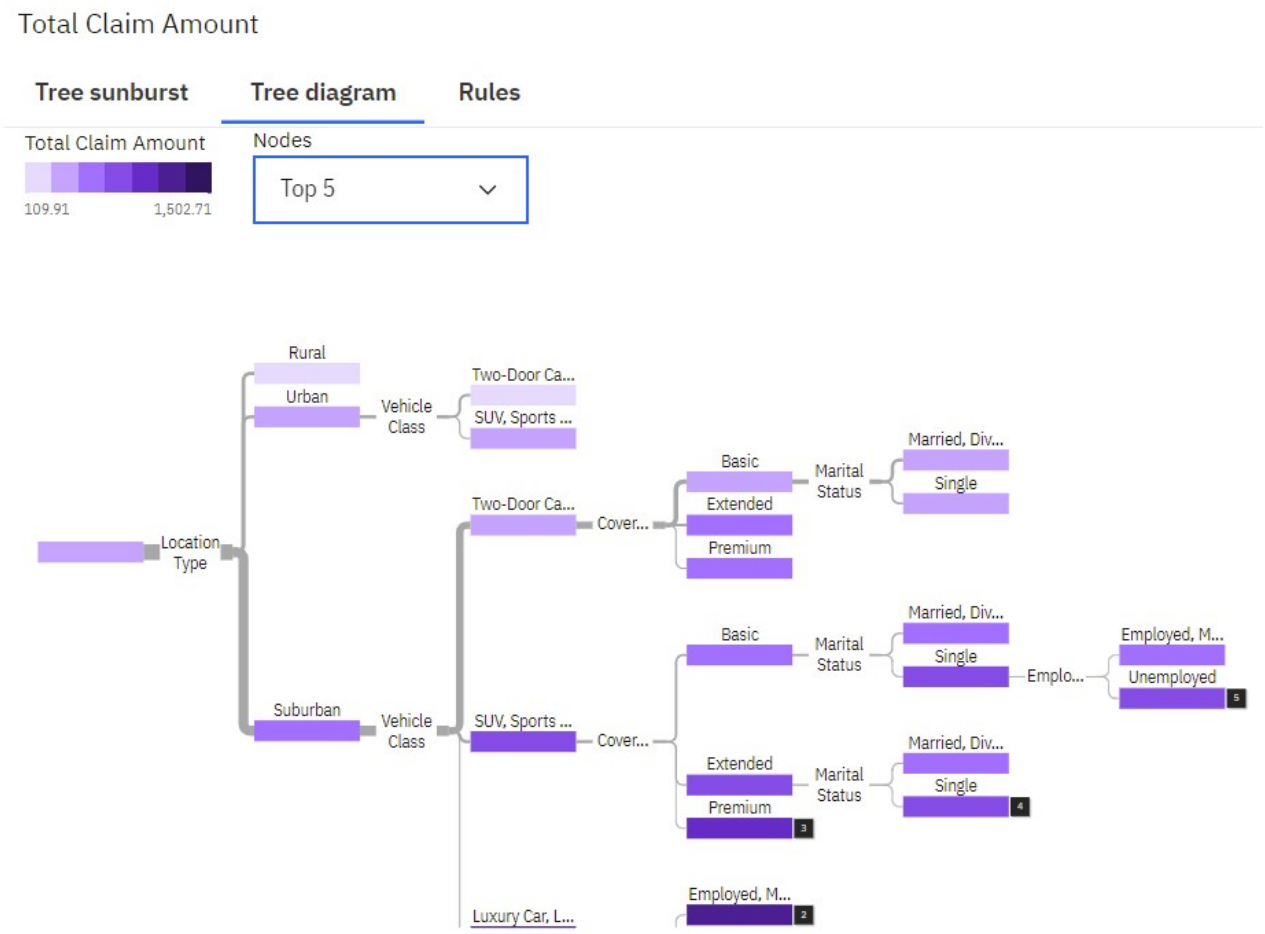
A decision tree shows a connected hierarchy of boxes to represent the values of records.

Records are segmented into groups, which are called nodes. Each node contains records that are statistically similar to each other with respect to the target field. For example, a node might contain the records for males who have more than 18 years of education. Nodes can then be used to predict a target's field value. For example, the node about males and education might be used to predict salary.

Each branch in a decision tree corresponds to a decision rule. For more info about decision rules, see "Viewing decision rules" on page 24

To improve performance, due to number of rows in the data source, the analysis is based on a representative sample of the entire data.

For example, a decision tree visualization can look like this:



Note: Filters are not supported for decision tree visualizations.

For more information, see "Exploring a decision tree visualization" on page 20.

Exploring a decision tree visualization

A decision tree visualization is used to illustrate how underlying data predicts a chosen target and highlights key insights about the decision tree.

About this task

The predictive strength of a decision tree determines the degree to which the decisions represented by each branch that is shown in the tree, predicts the value of the target.

Decision trees have a single target. If the target field of the decision tree is continuous, then the key insight indicators highlight unusually high or low groups. If the target field of the decision tree is categorical, then the key insight is the mode of the node. The mode of the node is the most frequently occurring category or categories of the target field within the group.

To improve performance, due to number of rows in the data source, the analysis is based on a representative sample of the entire data.

When you review a decision tree:

- If you want to see all the drivers, use either the **Tree diagram** tab or the **Rules** tab.
- If you want to focus on key drivers, use the **Tree sunburst** tab.

To edit or add key drivers, click the  on the target field.

Insights are different depending on the type of your target. If you are predicting a continuous measure, for example income, age, or profit, then the decision tree shows within the node the average value of the target given the conditions so far within the group that is represented by the node. For example, if you have a tree that is predicting income and you have a branch that has gender and then city. If you follow the path from male to Chicago, then the value that is in the Chicago node, is the average income of males in Chicago.

Procedure

1. If you have a continuous measure, the following example illustrates the decision tree.

   The color shows whether the value of the node is associated with high, medium, or low values of the target. The color of the node is based on the average of the target for the measure. The higher the average value of the target for a node, the darker the color.

   For example, shown next is the detailed visualization for Total Claim Amount on automobile insurance policies. A strong predictor for a high claim amount is claims that originate from policy holders who live in a suburban location, drive a luxury car, and are employed. A predictor of low claim amount is claims that originate from policy holders who live in a rural location.

   The minimap helps you move around the areas of the tree. The minimap is especially useful if there are many nodes.

   In this example, the top five highest target values are indicated with a number. You can choose between the following options:

   - Full tree. No highest, or lowest values are indicated specifically.

   - Top five highest target values. The top five highest target values are shown.

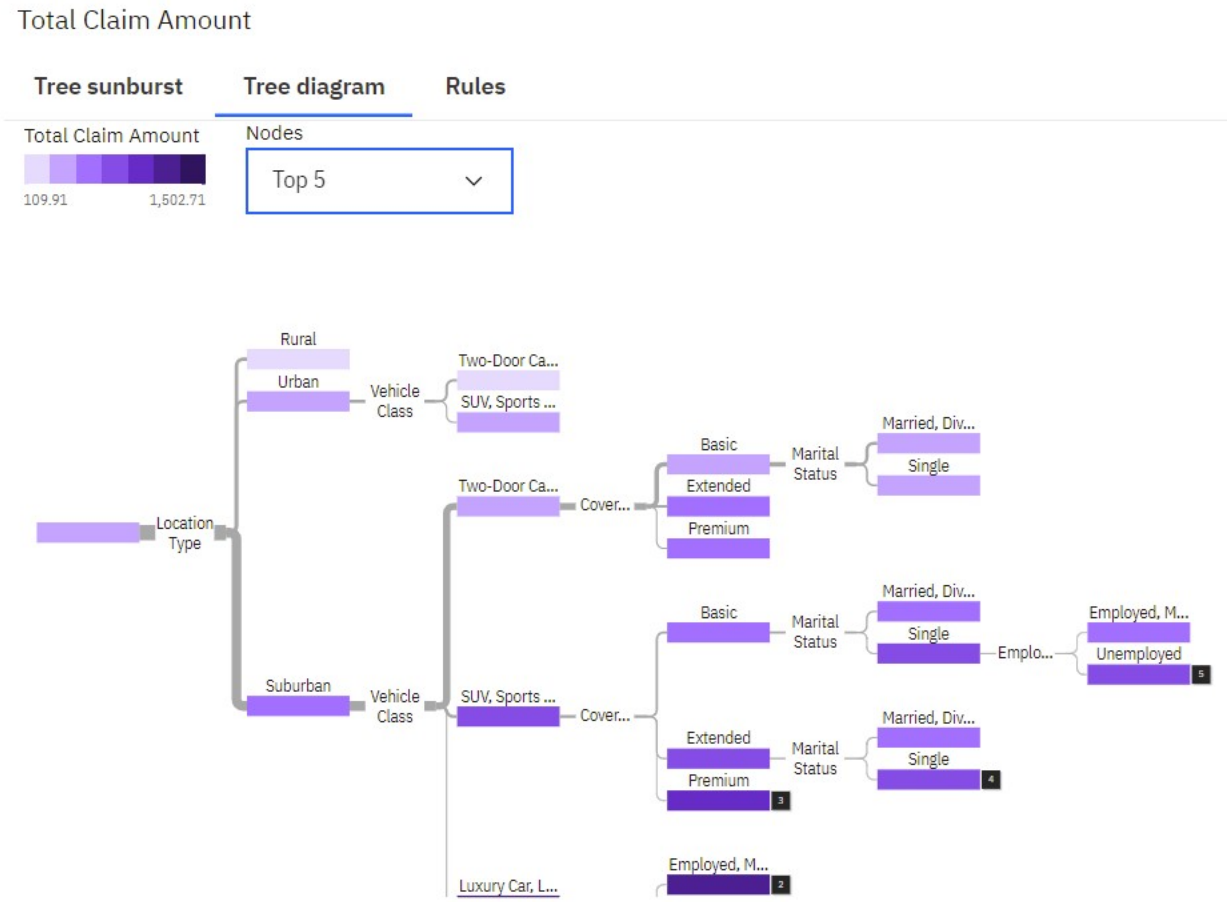
     

   - Top five lowest target values. The five lowest target values are shown.

   If you have a categorical measure, select the category for which you want to see the top five or lowest five targets from the **Top 5 nodes for:** menu or from the **Bottom 5 nodes for:** menu.

   In case you zoomed in too far, the top five or bottom five nodes are not visible.

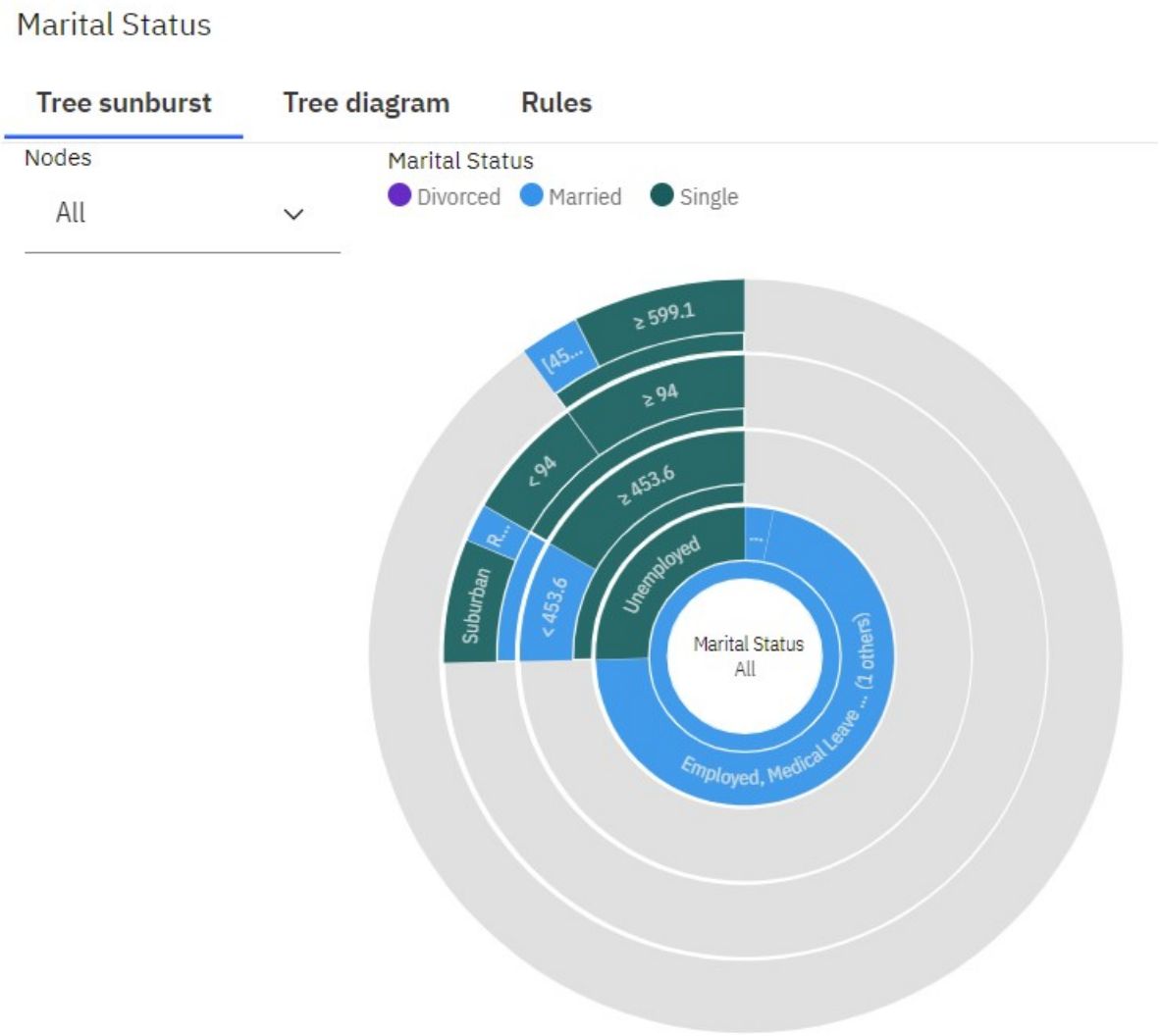


2. If you have a categorical measure, the following example illustrates the decision tree.

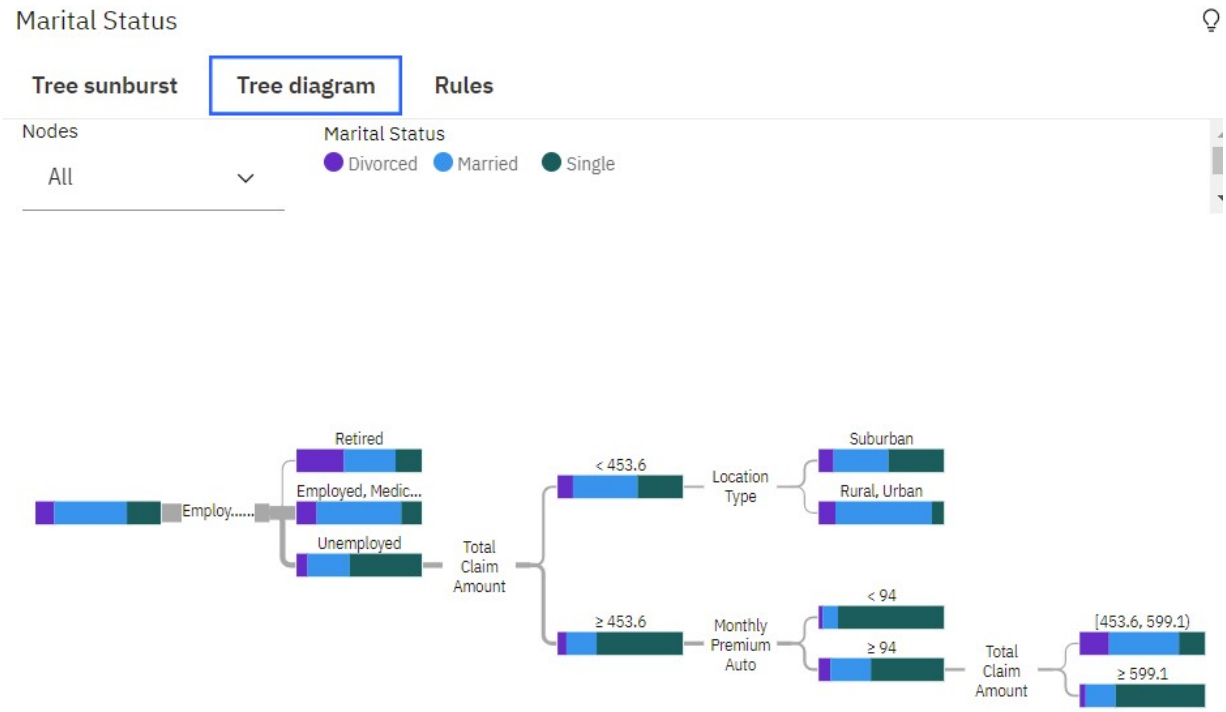
   The color shows which field value or values are represented the most.

   In the **Tree sunburst** tab, you can see that if the measures within the decision tree are strong predictors for a target value or target values, then the colors prevail in that node. The non-significant values are left out.

   For example, shown next is the detailed visualization of the marital status in the **Tree sunburst** tab. It shows that being employed is a strong predictor for being married.



In the **Tree diagram** tab, the nodes visually show the distribution of the people by marital status.



Viewing decision rules

A decision rule predicts an outcome in the target field. Viewing the decision rules helps you determine which conditions are likely to result in a specific outcome.

For example, consider some hypothetical decision rules that might predict churn. These rules might identify classifications based on the ranges for customer age and number of previous claims. From these rules, you might observe that customers who have no or 1 claim and are older than 50 are more likely to churn.

Each branch in a decision tree corresponds to a decision rule.

Procedure

1. In a decision tree, tap **Rules**.
2. Review the decision rules.
3. To return to the visualization, tap **Tree diagram**.

Driver analysis

A driver analysis visualization shows you the key drivers, or predictors, for a target. The closer the driver is to the right, the stronger that driver is.

IBM Cognos Analytics uses sophisticated algorithms to deliver highly interpretable insights that are based on complex modeling. You don't have to know which statistical tests to run on your data. Cognos Analytics picks the right tests for the data.

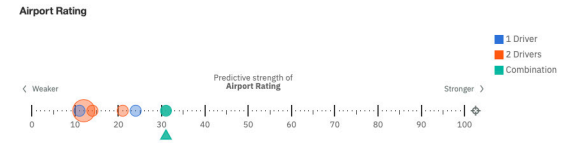
Key drivers for both continuous and categorical targets are available in the driver analysis visualization in dashboards and explorations.

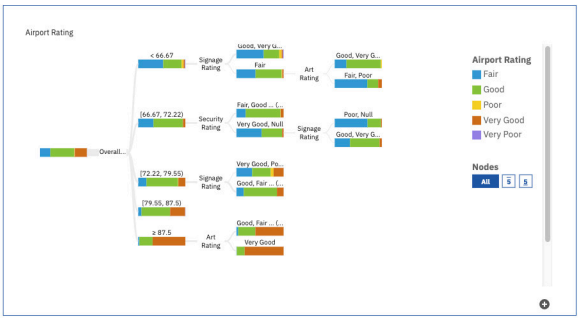
For more info, see *Statistical tests* documentation in the *IBM Cognos Analytics Dashboards and Stories User Guide*.

For example, this driver analysis visualization shows that the combination of overall satisfaction, signage rating, security rating, and art rating are the strongest drivers of the target airport rating.

To edit or add key drivers, click the  on the target data slot.

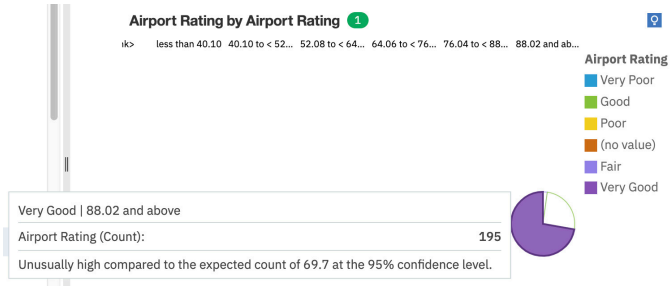
To improve performance, due to number of rows in the data source, the analysis is based on a representative sample of the entire data.





If you hover over a data point, then the driver analysis visualization shows what drives the overall airport rating.

If you click a data point in the tree, other recommended visualizations are shown.



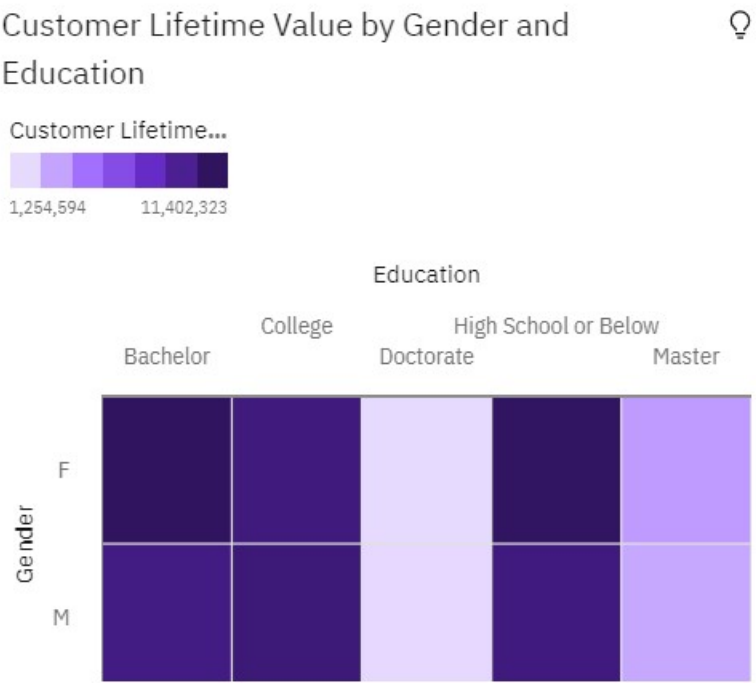
Note: Filters are not supported for driver analysis visualizations.

Heat map

Use a heat map visualization to visualize the relationship between columns, represented in a matrix type view.

A heat map visualization uses color and intensity of the color to show the relationship between two columns.

For example, this heat map visualization shows the average customer lifetime value by gender and education.



Create the heatmap visualization by dragging the following data items from the **Sources** panel :

- Drag **Gender** onto the **Rows** field.
- Drag **Education** onto the **Columns** field.
- Drag **Customer Lifetime Value** onto the **Heat** field.

Samples

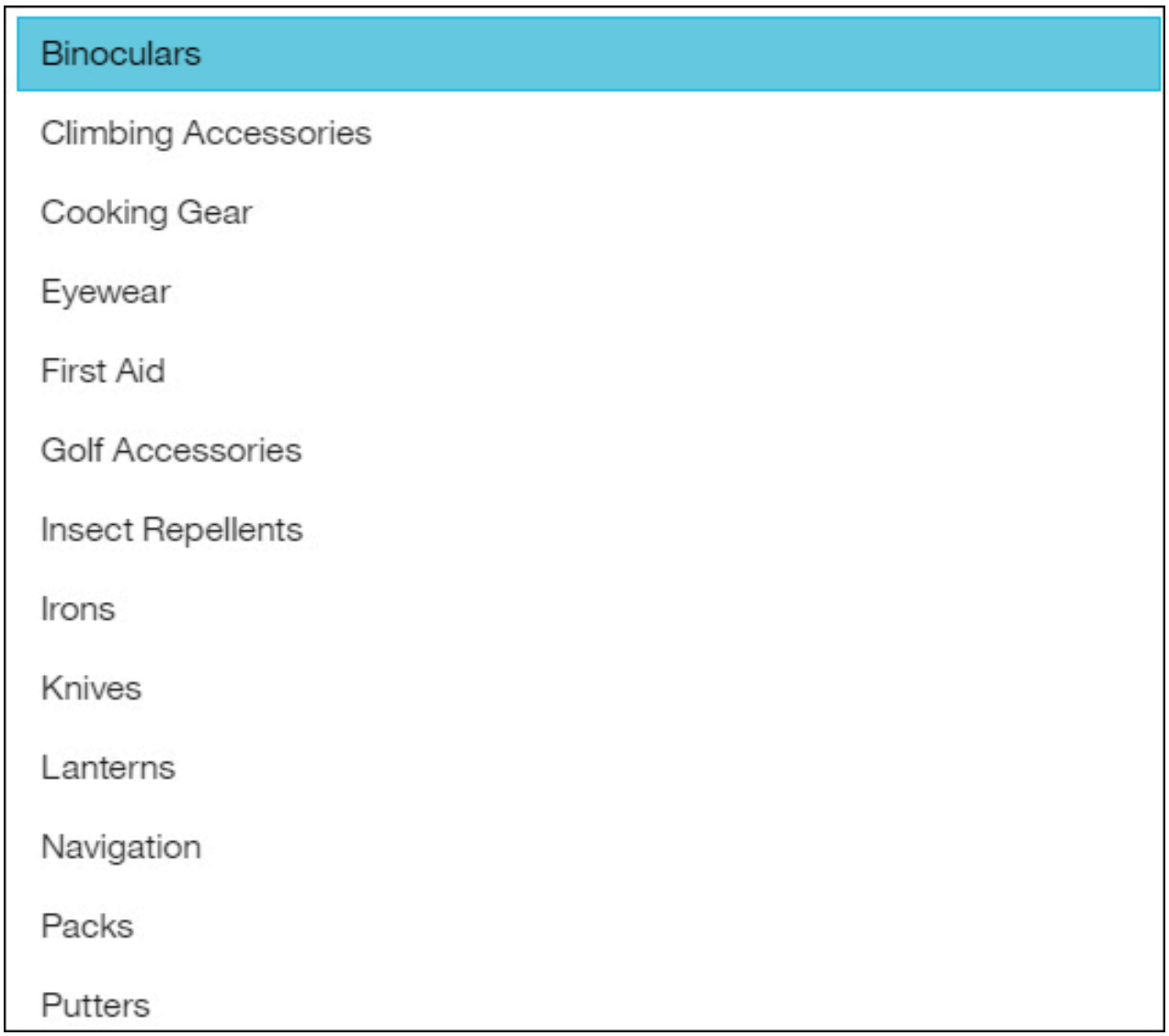
You can see examples of visualizations in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Standard reports** > **Customer lifetime value analysis**.

If any of the sample objects are missing, contact your administrator.

Hierarchy

Use a hierarchy when you want to see the data in rows and columns.

For example, this hierarchy shows product types.



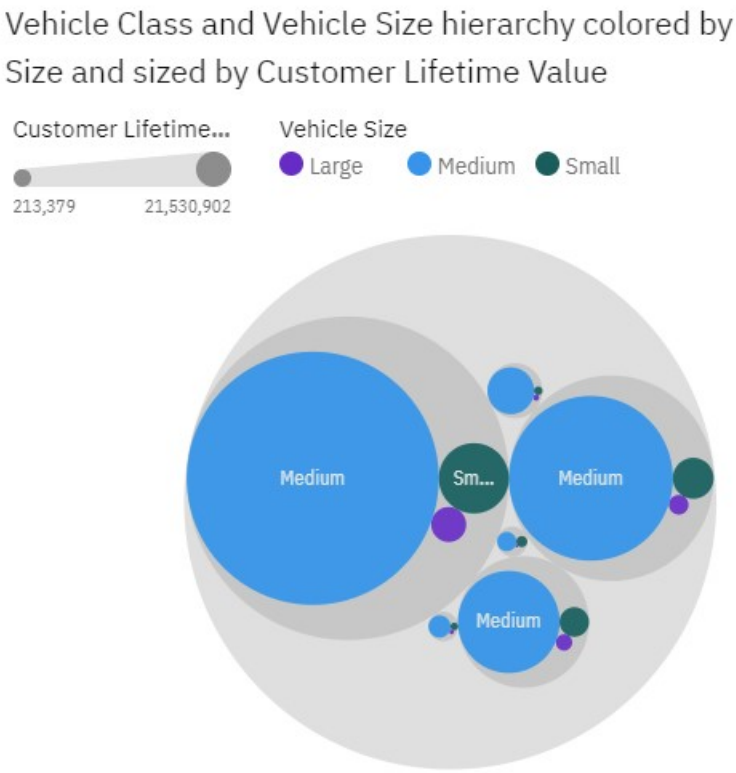
Hierarchy bubble

Use a hierarchy bubble visualization when you want to show relationships among columns that contain values, such as net loss. It is similar to the bubble visualization but the bubbles are tightly packed instead of spread over a grid. The bubbles use nesting to represent the hierarchy. A hierarchy bubble visualization shows a large amount of data in a small space.

The size of each bubble shows a quantitative dimension of each data point. It shows many levels within a hierarchy and relationships between groups based on assigned attributes. It uses bubble size and color to convey comparative information about categories.

The bubbles are in different sizes and colors.

For example, this hierarchy bubble visualization shows customer lifetime value by vehicle class per vehicle size. Each bubble is a different vehicle class in one of the three vehicle size. The size of each bubble is determined by the customer lifetime value of that vehicle class. The colors of the bubbles are determined by the vehicle size.



The hierarchical packed bubble visualization was created by dragging the following data items from the Sources panel:

- Drag **Vehicle Class** and **Vehicle Size** onto the **Bubbles** field.
- Drag **Customer Lifetime Value** onto the **Size** field.
- Drag **Vehicle Size** onto the **Color** field

Samples

You can see an example of a word cloud visualization in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Customer lifetime value analysis**.

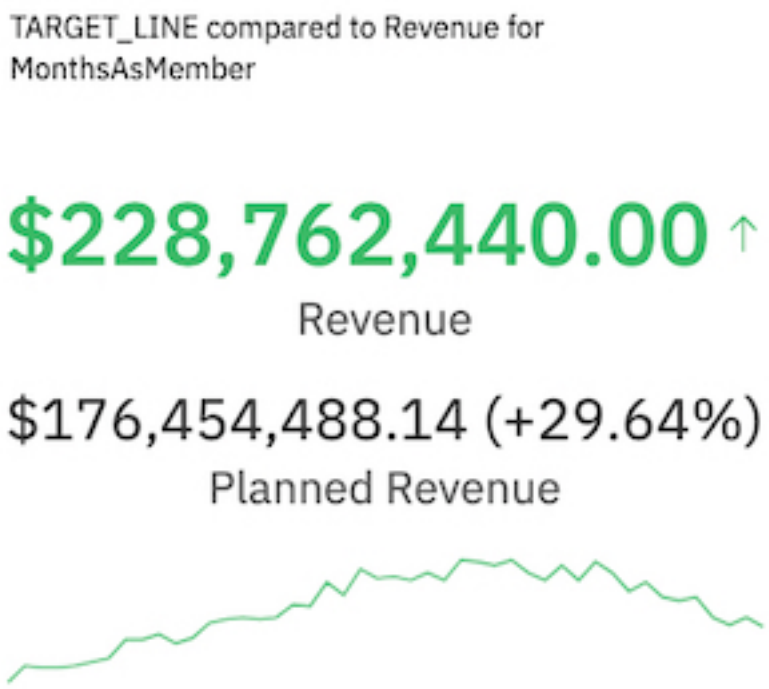
If any of the sample objects are missing, contact your administrator.

KPI

Use a KPI visualization to display a key performance indicator (KPI) that contains two related measures, such as revenue and planned revenue. Optionally, you can display a sparkline and a meaningful shape in your KPI visualizations.

A KPI visualization compares a base value to a target value and shows the variance between the two measures.

For example, this KPI visualization shows the actual revenue in green with an up arrow to indicate that revenue is up compared to the target. In this case, the target value is planned revenue. A sparkline displays the shape of the variation over time and is the same color as the base value.



Create a similar KPI visualization by dragging measures from your own data source to the fields in an empty KPI visualization:

1. On the toolbar, select **New card**.
2. Select the blank **Single** card.
3. Click **Choose a type** and select the **KPI** visualization type. Then, add a data item to each field.
4. Drag a measure onto the **Base value** field. This value is the actual target.
5. Drag a measure onto the **Target value** field.
6. Drag another measure onto the **Time** field. This value creates a sparkline for your KPI visualization. You can add multiple measures, for example Years and Months, to the **Time** field.

Use the properties to customize a KPI visualization. For example, the properties are set by default to display a green conditional color when the target is met and a red conditional color if the target is not met. To display the actual target in another color, under properties, expand the rule and then select a different **Text color**.

Complete the following steps to edit a conditional color rule and select a custom colors:

1. Select the KPI visualization on your exploration.
2. Click the **Properties** tab.
3. Under **Rules**, expand the rule that you want to edit.
4. From **Text color**, select a color.

The following information describes the KPI properties under **Rule style**:

- **Text color**

  Set the color for the value, sparkline, and indicator shape.

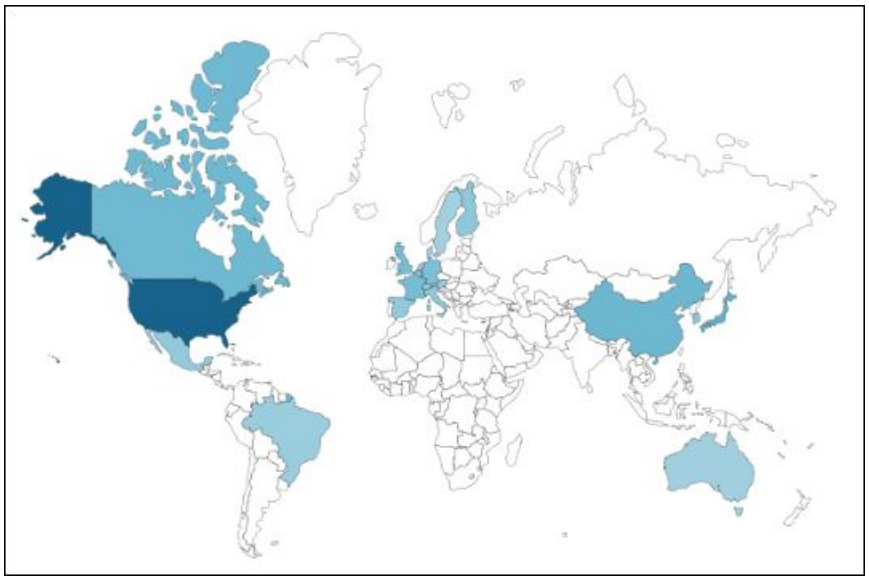
- **Indicator**

  Select a shape to display on the KPI visualization when the when the rule is met. For example, you might want to display a down arrow when your base value falls below a certain threshold compared to the target value.

Legacy map

Use a legacy map when you want to see patterns in your data by geography. You can use a legacy map when you are not connected to the internet.

For example, this legacy map visualization shows revenue by retailer country with the darker color indicating higher revenue.



For more information, see https://www.ibm.com/support/knowledgecenter/SSEP7J_11.1.0/com.ibm.swg.ba.cognos.ug_ca_legacymaps.doc/ug_ca_legacymaps.pdf.

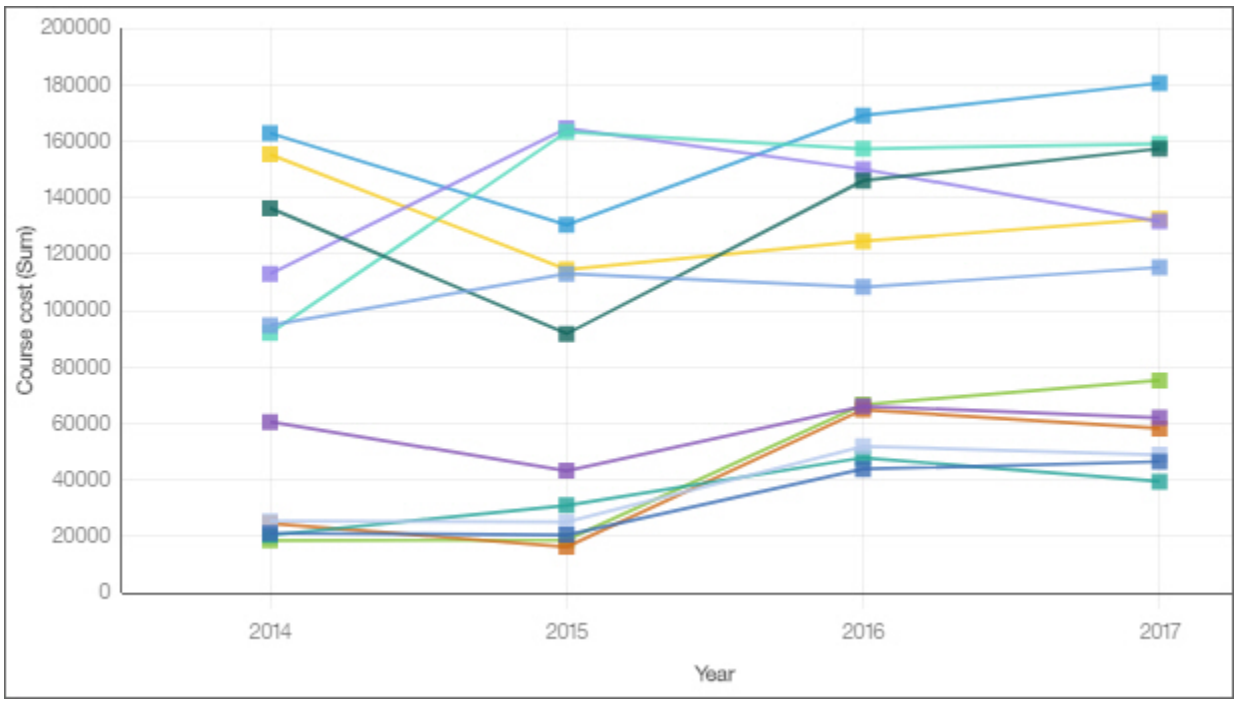
Line

Use a line visualization to show trends over time.

A line visualization can compare trends and cycles, infer relationships between variables, or show how a single variable is performing over time.

For an effective line visualization, use a time column in the x-axis, such as years, quarters, months, or days. If the x-axis shows something else, such as Canada, Netherlands, UK, and US, use a bar or column visualization.

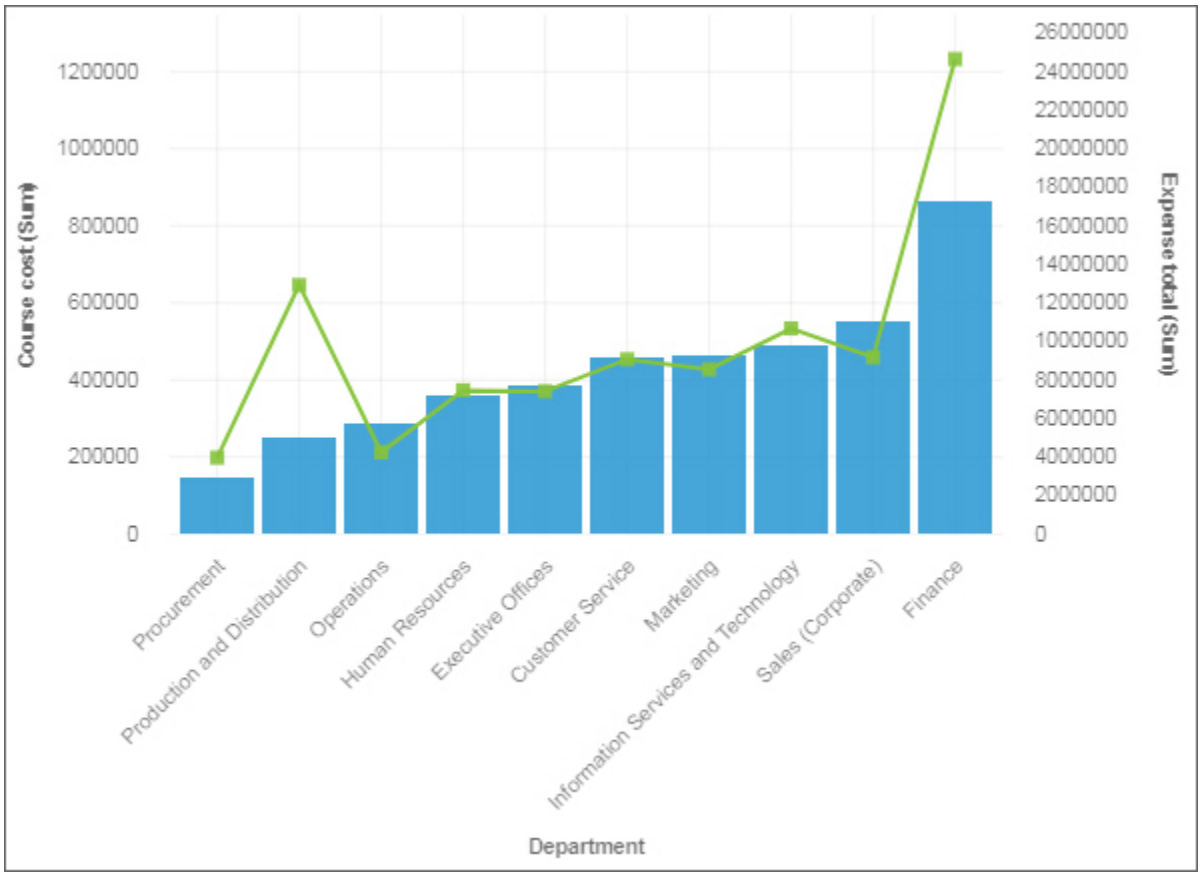
For example, this line visualization shows the trend in course costs by department over year.



Line and column

Use a line and column visualization to highlight relationships between multiple data series by combining bars and lines with one visualization.

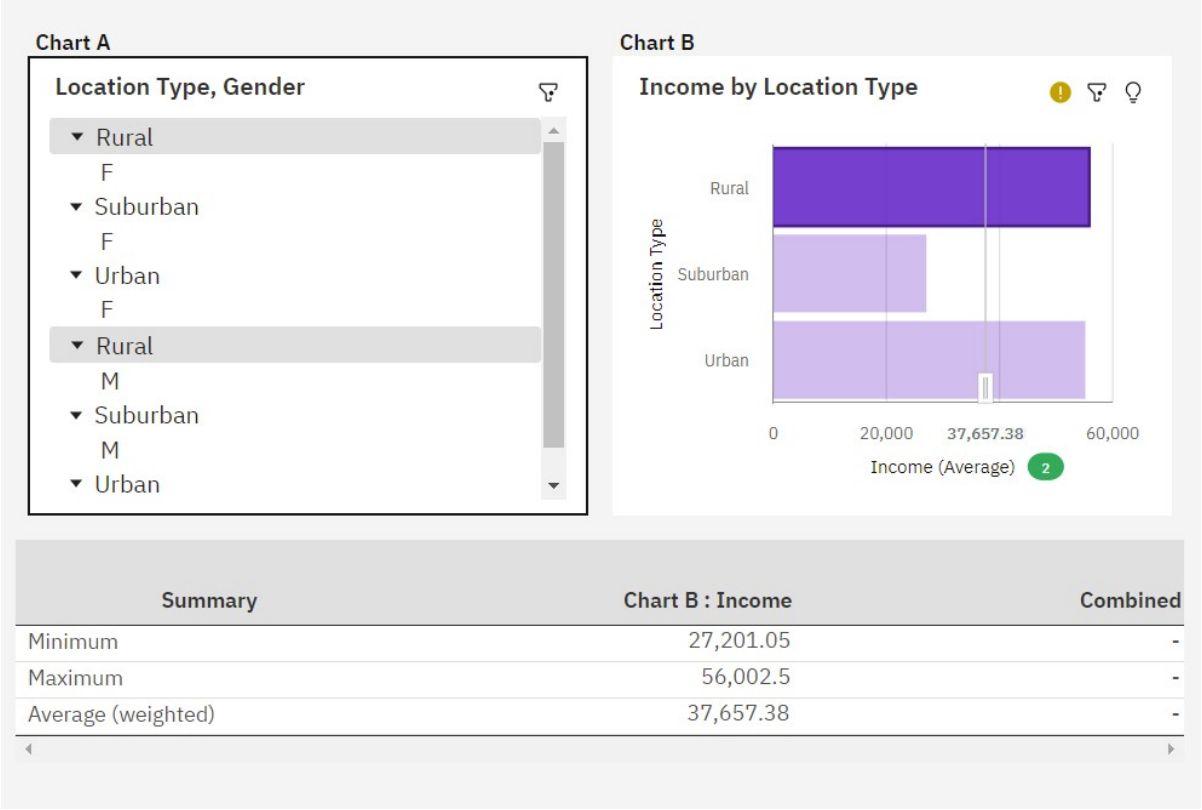
For example, this line and column visualization shows the relationship between course cost and expense totals by department.



List

Use a list visualization to create an overview the data in a hierarchical way.

You can use a list visualization as a filter widget. The next example show how you can use the list visualization as a filter widget.



Map

Use a map when you want to see patterns in your data by geography.

Your data asset must contain geographical data, such as countries, states, provinces, or continents.

Note: Maps do not show animations if you set your ease of access system settings to not display animations.

Maps in Cognos Analytics support the following continents:

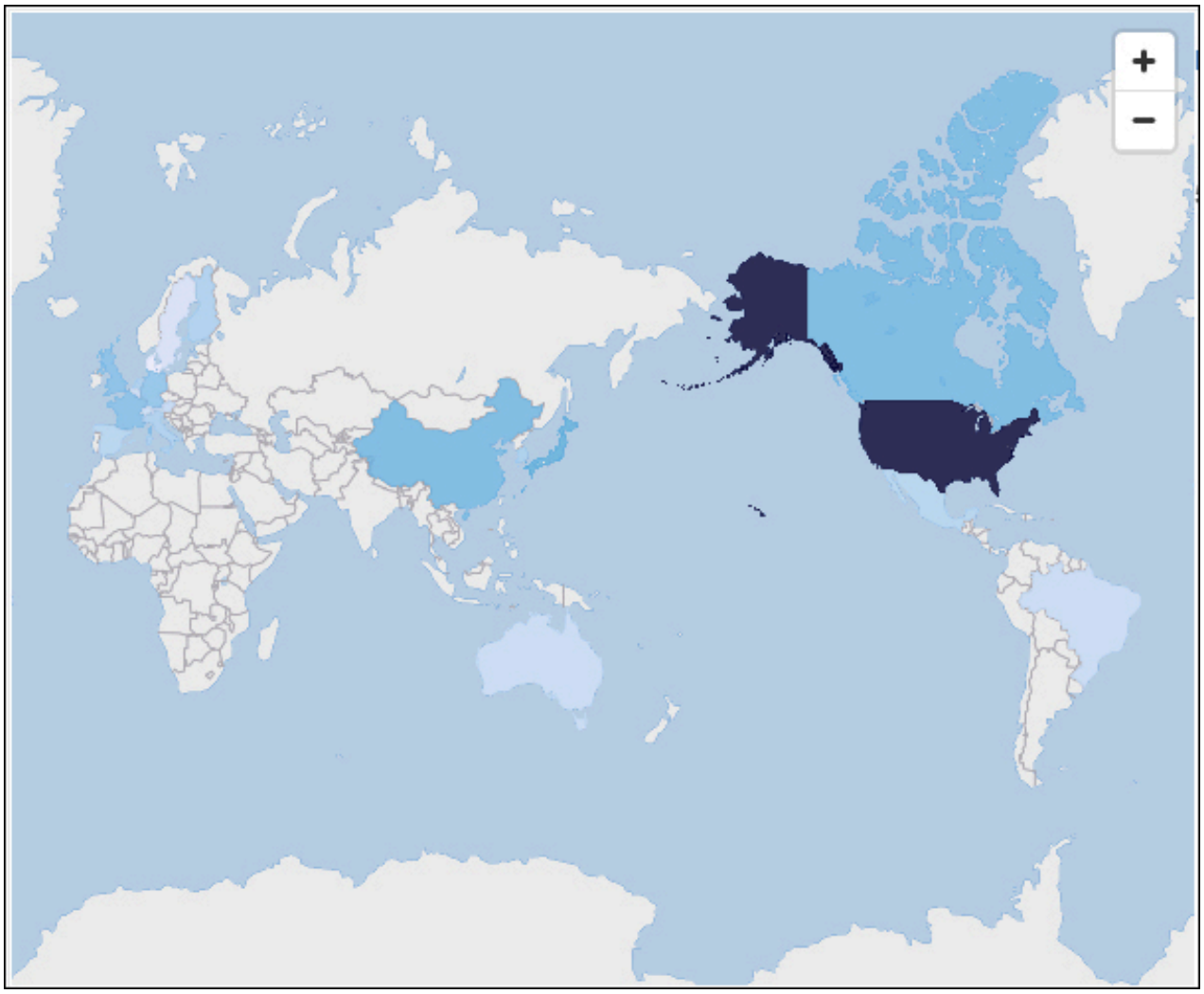
- North America
- South America
- Africa
- Asia
- Europe
- Antarctica
- Oceania

To determine whether a column is mappable, Cognos Analytics analyzes a sample of 2000 values in the location column, looking for recognizable place names. If 80% or more are recognized as map values, Cognos Analytics produces a map.

For example, you have four countries in your location column: Brazil, China, India, and Russia. The typographical error for India means that only 75% of the values are recognizable place names and you

will not see a map as a starting point. But if you have five countries and one has a typographical error in it, you see a map.

For example, this map visualization shows revenue by retailer country with the darker color indicating higher revenue.



Marimekko

A marimekko visualization is similar to a stacked column visualization. It shows data through varying heights and includes an added dimension of data through varying column widths. The width of the columns is based on the value that is assigned to the width field. Individual segment height is a percentage of the respective column total value.

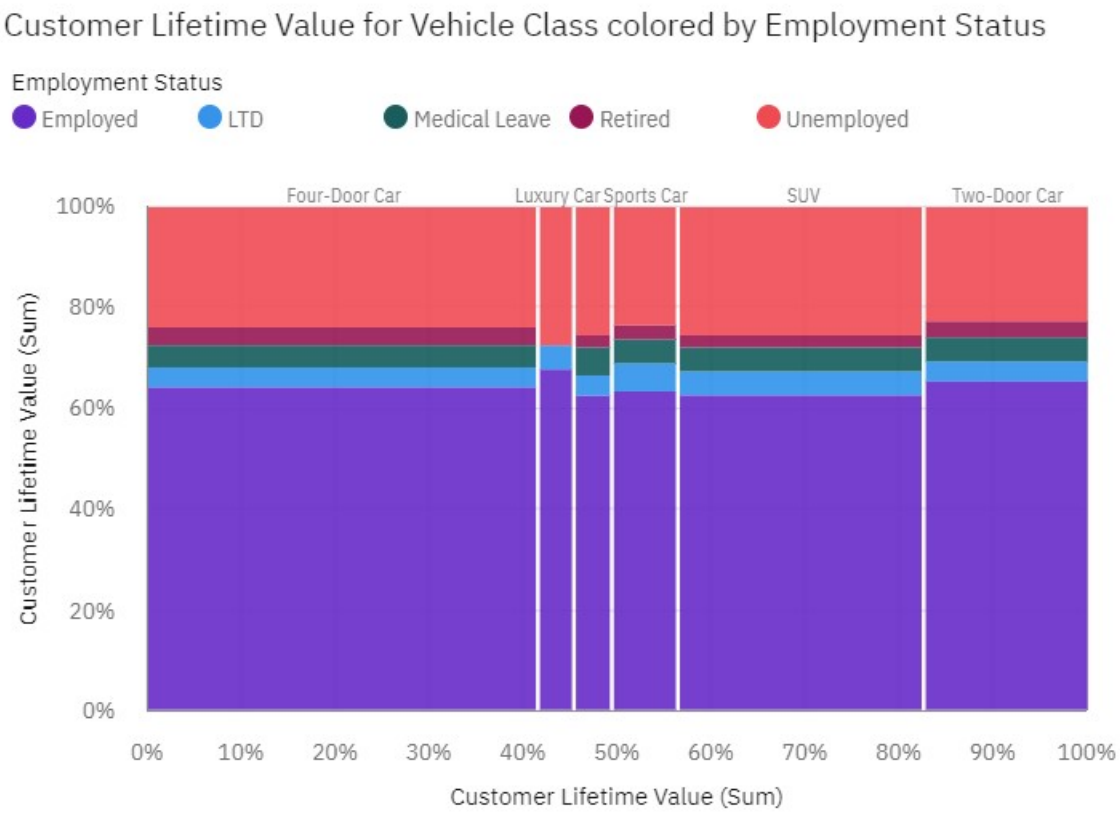
You can quickly spot large segments, such as a specific vertical that has a large share of a region. You can also identify white space such as an under-represented vertical in a specific region.

The marimekko visualization is useful for part-to-whole comparisons, where you need to show an extra measure/variable.

The marimekko visualization allows data to be depicted along two dimensions simultaneously. For example, market segments are often arrayed along the x-axis, with the width of each column corresponding to the financial value of a segment. You use marimekko visualizations in cases, for example, where you want to show the revenue contribution per product line. Or the gross domestic product per country.

The marimekko visualization can display total or partial number. If you want to use stacked percentages instead of number, then use the **Display as stacked percentage chart** option.

The following example shows the contribution of customer lifetime value and employment status in different vehicle classes with the option **Display as stacked percentage chart** enabled.



The marimekko visualization was created by dragging the following data items from the Sources panel:

- Drag **Vehicle Class** type onto the **Bars** field.
- Drag **Customer Lifetime Value** onto the **Length** field.
- Drag **Employment Status** onto the **Color** field

Samples

You can see an example of a word cloud visualization in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Customer lifetime value analysis**.

If any of the sample objects are missing, contact your administrator.

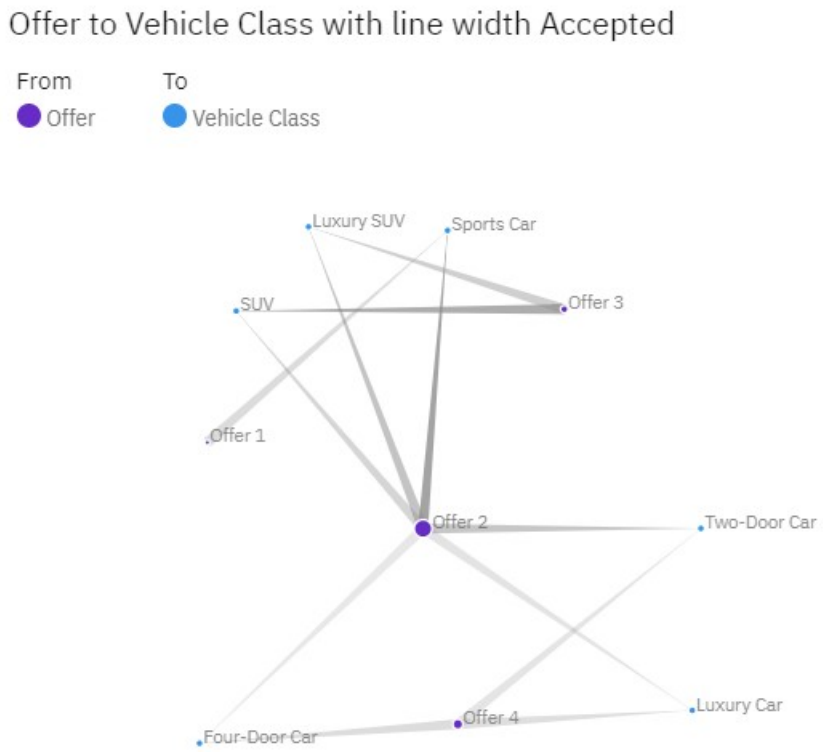
Network

Use a network visualization when you want to see the connections among columns in your data asset. A network visualization is a good choice to show connections, networks, and points of intersection.

Network visualizations display a set of nodes, represented by symbols, and links, represented by paths, to show the relationship between entities or items.

Use the **From** and **To** fields to define the relationship that you want to investigate.

For example, a network visualization can show offer acceptance by Vehicle Class.



Create the Network visualization by dragging the following data items from the **Offers** section in the **Sources** pane  :

- Drag **Offer** onto the **From** field.
- Drag **Vehicle Class** onto the **To** field.
- Drag **Accepted** onto the **Line width** field.

Next, set the size and node properties.

1. Click the visualization, then click . Set set the following options in the **Properties** pane:
   - **Size - Width**: 500 px, **Height**: 300 px
   - **Nodes minimum size**: 20
   - **Nodes maximum size**: 100
2. Click to close the **Properties** pane.

Samples

You can see examples of visualizations in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Standard reports** > **Customer lifetime value analysis**.

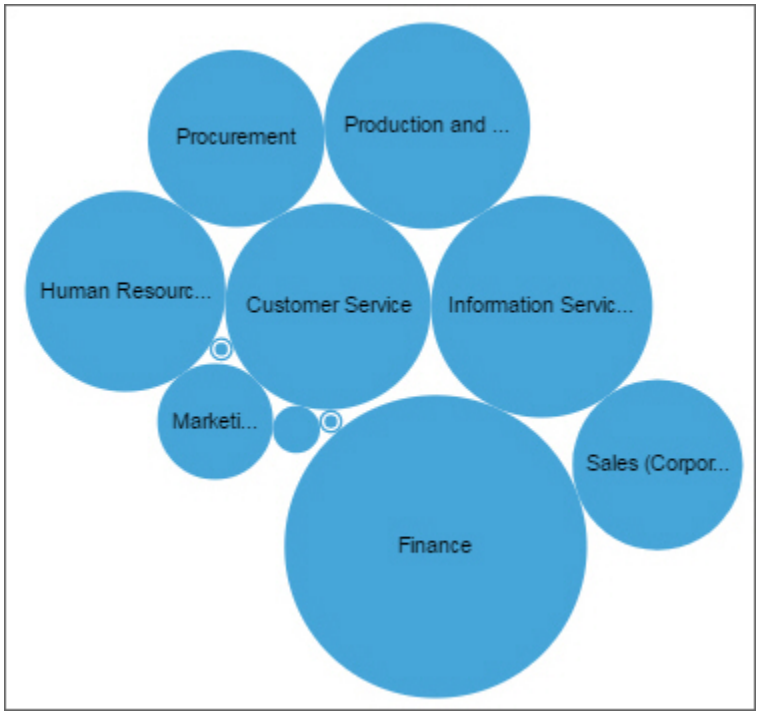
If any of the sample objects are missing, contact your administrator.

Packed bubble

Use a packed bubble visualization when you want to show relationships among columns that contain numeric values, such as revenue. It is similar to the bubble visualization but the bubbles are tightly packed instead of spread over a grid. A packed bubble visualization shows a large amount of data in a small space.

The bubbles are in different sizes and colors.

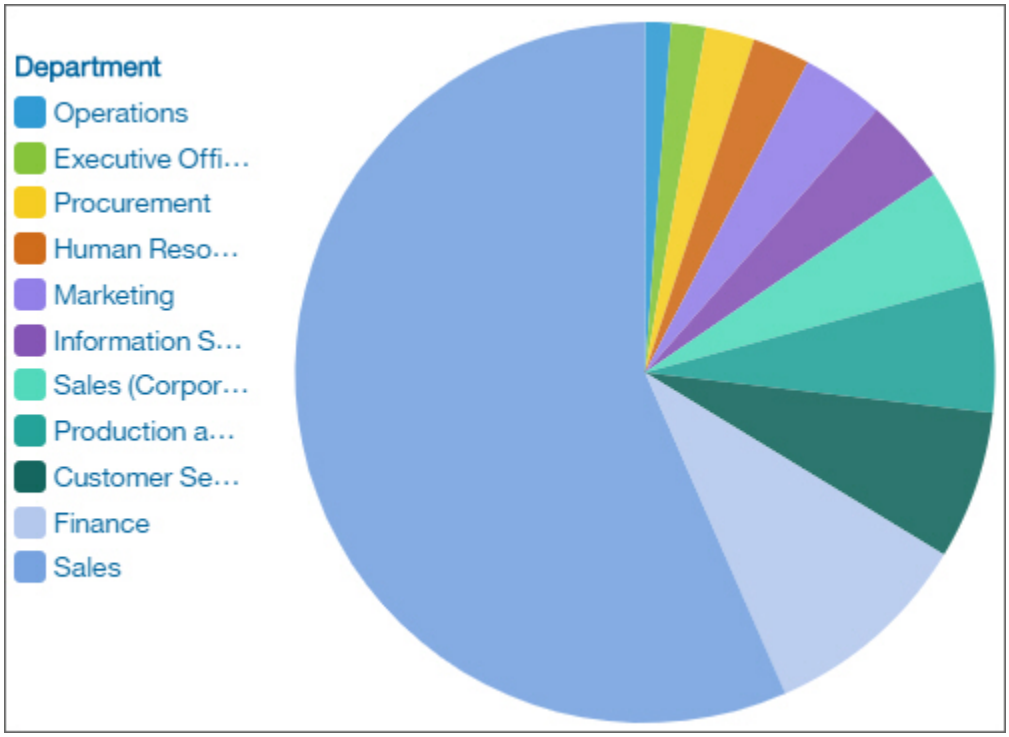
For example, this packed bubble visualization shows external hires by department. Each bubble is a different department. The size of each bubble is determined by the number of external hires for that department.



Pie

Use a pie visualization to highlight proportions. Each slice shows the relative relationship of each part to the whole.

For example, this pie visualization shows the number of course days for each department.



Point

Use a point visualization to show trends over time.

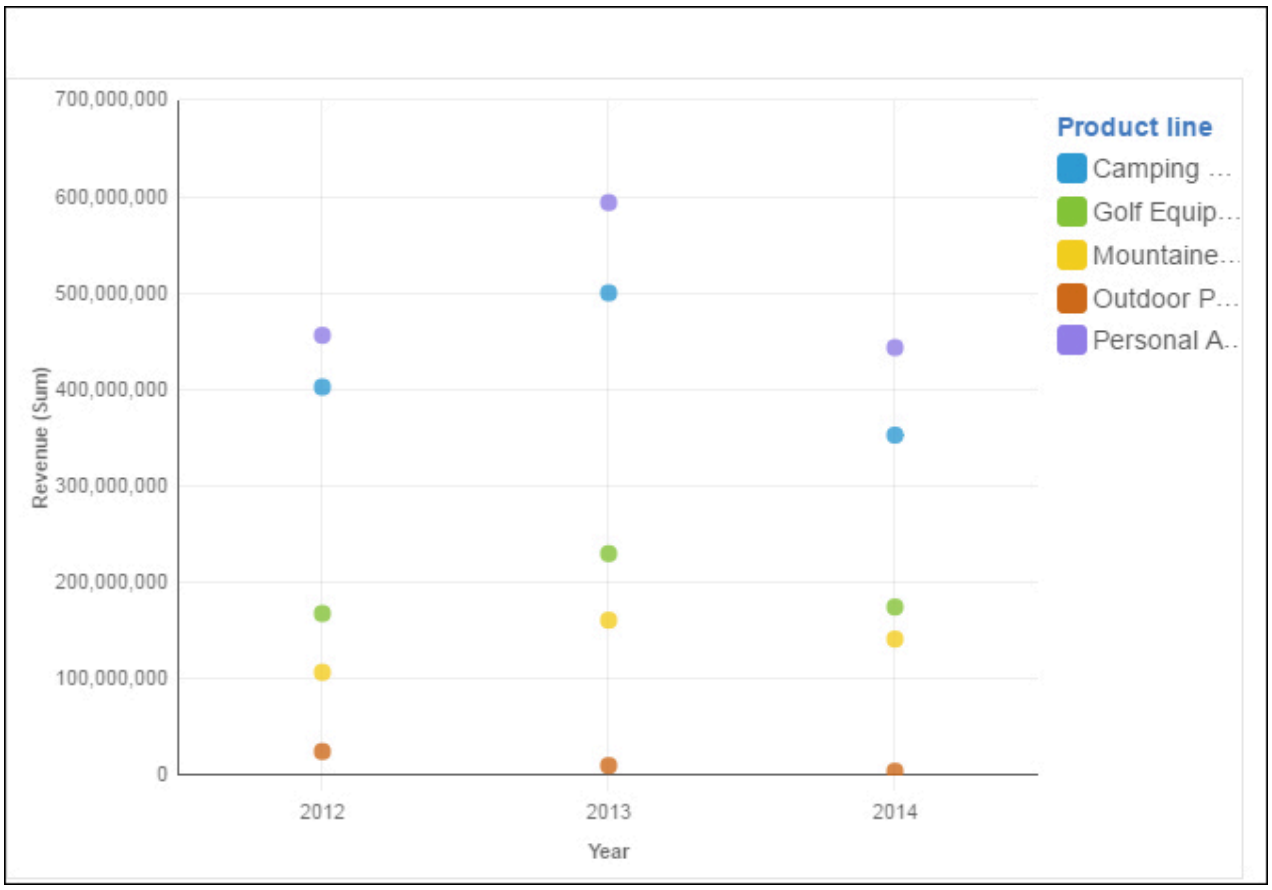
A point visualization can compare trends and cycles, infer relationships between variables, or show how a single variable is performing over time.

A point visualization is like a line chart without the connecting lines.

For an effective line visualization, the x-axis should show time, such as years, quarters, months, or days. If the x-axis shows something else, such as Canada, Netherlands, UK, and US, use a bar visualization.

Data values are plotted vertically.

For example, this line visualization shows revenue over quarter by order method type. Web orders have grown dramatically over this period.



Radar

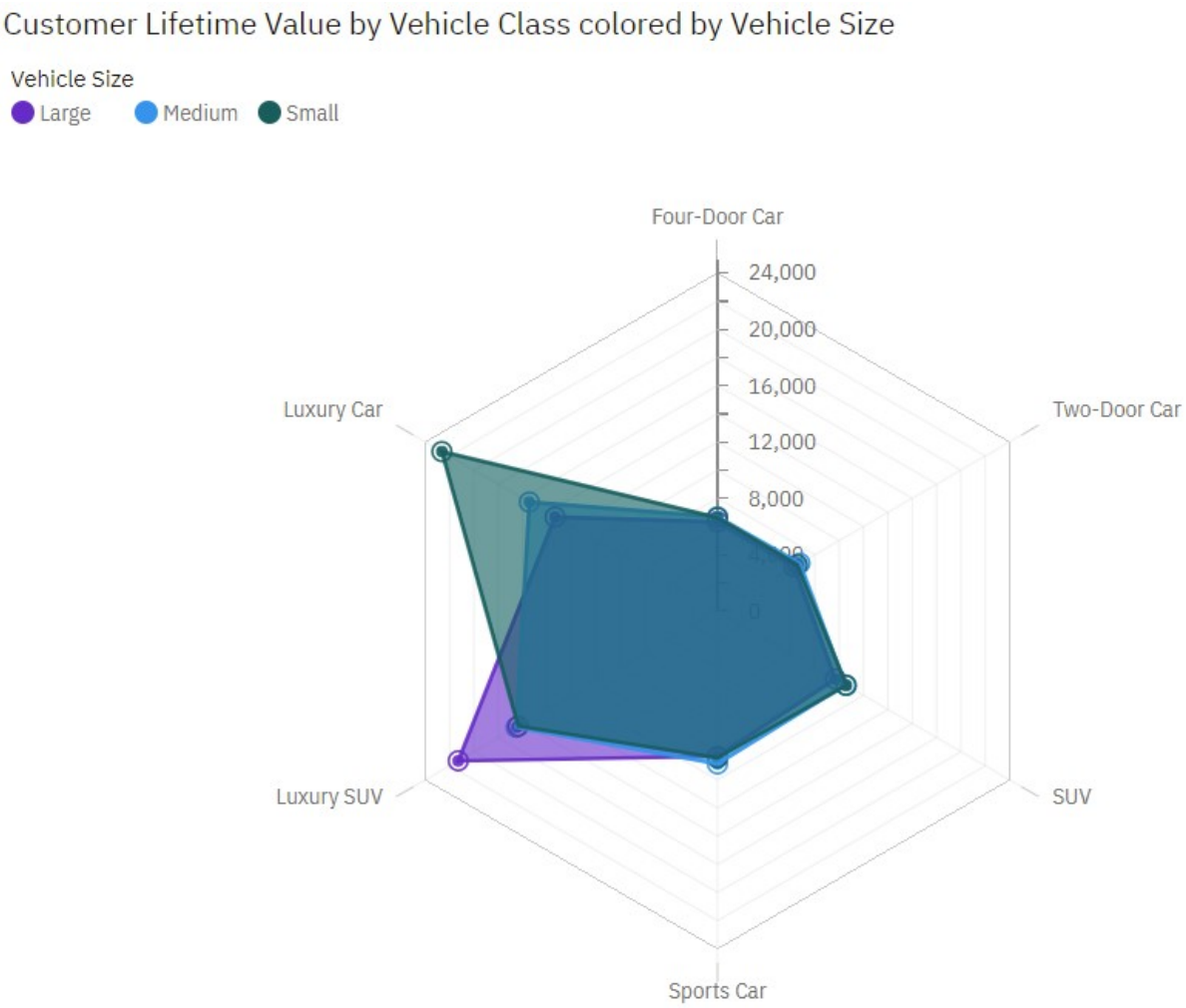
Use a radar visualization for comparing multiple quantitative variables. The radar visualization shows which variables have similar values or if there are any outliers amongst each variable.

Radar visualizations are also useful for seeing which variables are scoring high or low within a data set, making them ideal for displaying performance.

Examples of using a radar visualizations are:

- Comparing cars: speed, durability, comfort, power, space.
- Competitor profiles: number of employees, revenue, profit, current stock price, customer satisfaction.

The following example shows the revenues per retailer type for product lines in different states.



The radar visualization was created by dragging the following data items from the **Sources** pane:

- Drag **Vehicle Class** onto the **x-axis** field.
- Drag **Vehicle Size** onto the **Color** field.
- Drag **Customer Lifetime Value** onto the **y-axis** field.

Samples

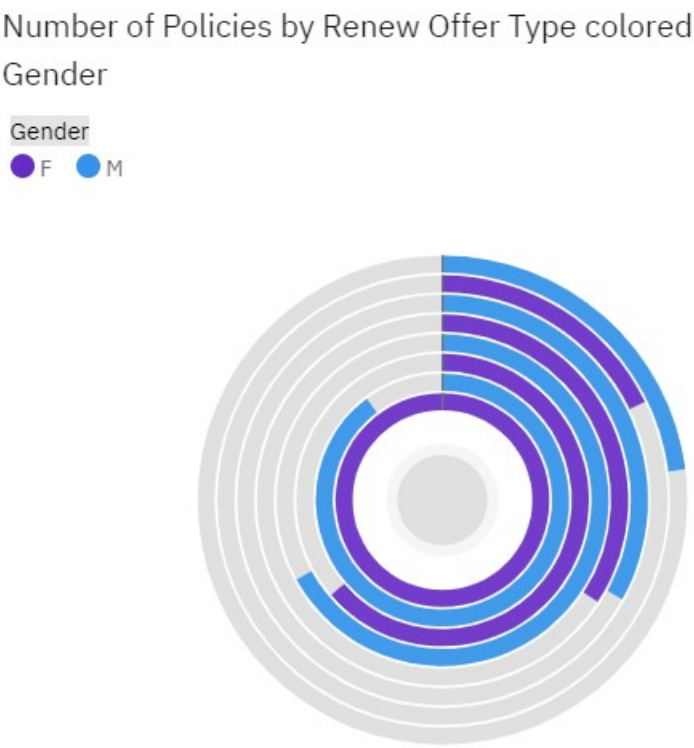
You can see an example of a radar visualization in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Data** > **Customer lifetime value analysis**.

Radial

In a radial visualization, each bar appears in a circle with longer bars that represent larger values. Hover over a bar to see the details about it, such as the exact value represented by the bar. Each bar starts at 12 noon and goes in a clockwise direction for positive values and counterclockwise for negative values.

Radial visualizations, also known as dial charts or speedometer charts, show information as reading on a dial. The radial visualization is valid only with one category.

For example, this visualization shows renewals by offer type and gender.



Create the Radial visualization by dragging the following data items from the **Customer Analysis** section in the **Sources** pane :

- Drag **Renew Offer Type** onto the **Bars** field.
- Drag **Number of Policies** onto the **Length** field.
- Drag **Gender** onto the **Color** field.

The next step is to set the sort properties for **Renew Offer Type** and **Gender**.

1. Click the visualization, then in the **Data** pane, click the **<Renew Offer Type>** data item.
2. Click
3. In the **Properties** pane, for **Sort order**, select **Ascending**.
4. In the **Data** pane, click the **<Gender>** data item.
5. In the **Properties** pane, for **Sort order**, select **Descending**.
6. Click to close the **Properties** pane.

Samples

You can see examples of visualizations in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Standard reports** > **Customer lifetime value analysis**.

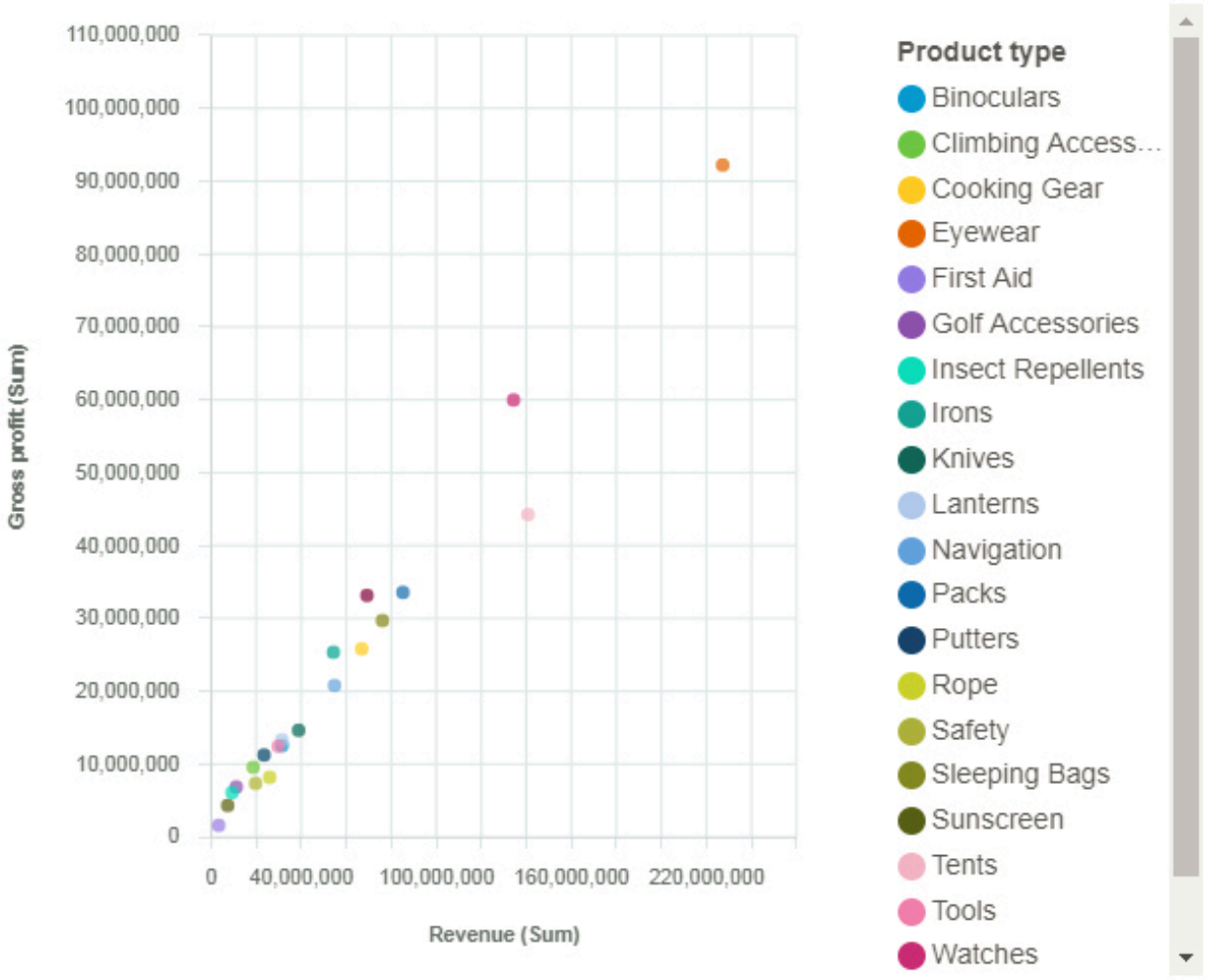
If any of the sample objects are missing, contact your administrator.

Scatter

Scatter visualizations use data points to plot two measures anywhere along a scale, not only at regular tick marks.

Scatter visualizations are useful for exploring correlations between different sets of data.

The following example shows the correlation between revenue and gross profit for each product type.



Spiral

A spiral visualization shows you the key drivers, or predictors, for a given target. The closer the driver is to the center, the stronger that driver is.

IBM Cognos Analytics uses sophisticated algorithms to deliver highly interpretable insights that are based on complex modeling. You don't have to know which statistical tests to run on your data. Cognos Analytics picks the right tests for the data.

Key drivers for both continuous and categorical targets are available in the spiral visualization in dashboards and explorations.

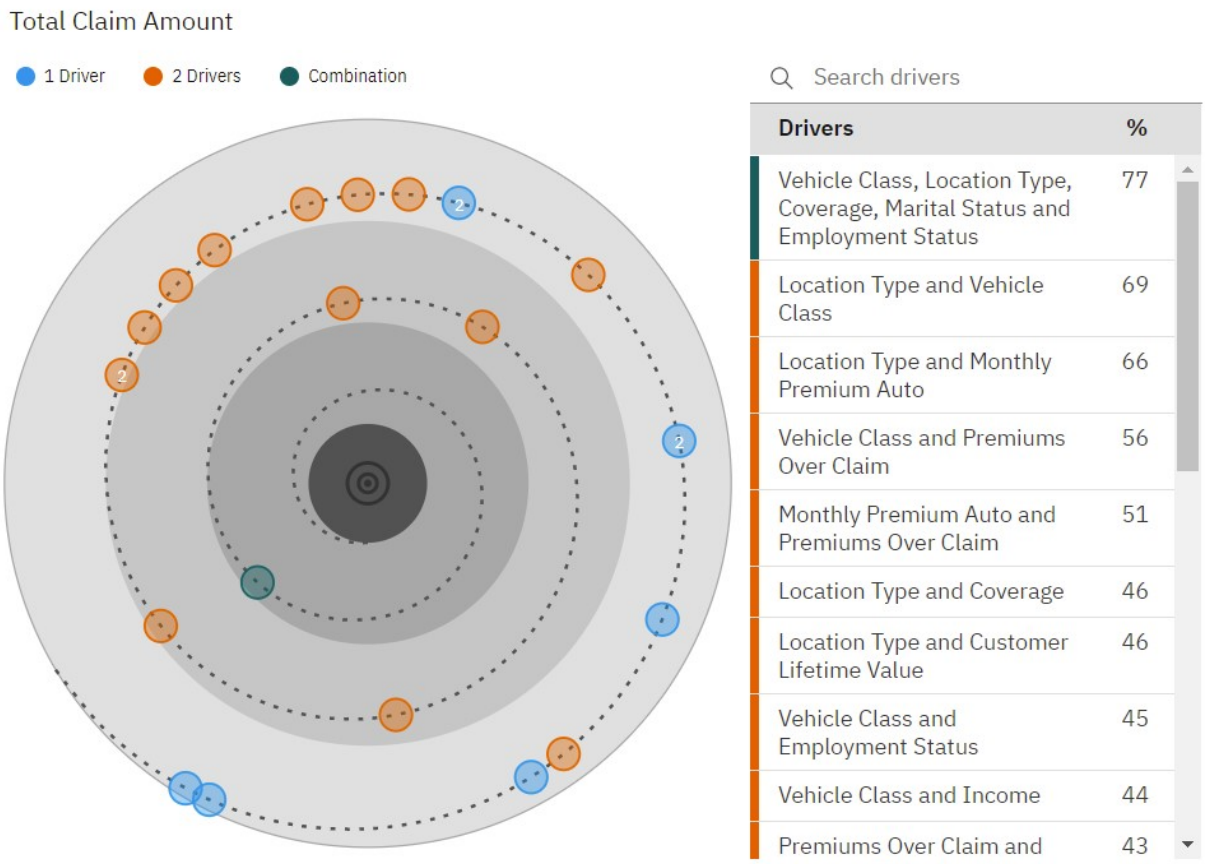
For more information, see *Statistical tests* documentation in the *IBM Cognos Analytics Dashboards and Stories User Guide*.

For example, this spiral visualization shows that the combination of vehicle class, location type, coverage, marital status, and employment status are the strongest drivers of the target, total claim amount.

You can exclude drivers from the analysis. Right-click on a driver and click the **Edit drivers** icon . Select the drivers that you want to include in the analysis.

To edit or add key drivers, click the on the target data slot.

To improve performance, due to number of rows in the data source, the analysis is based on a representative sample of the entire data.

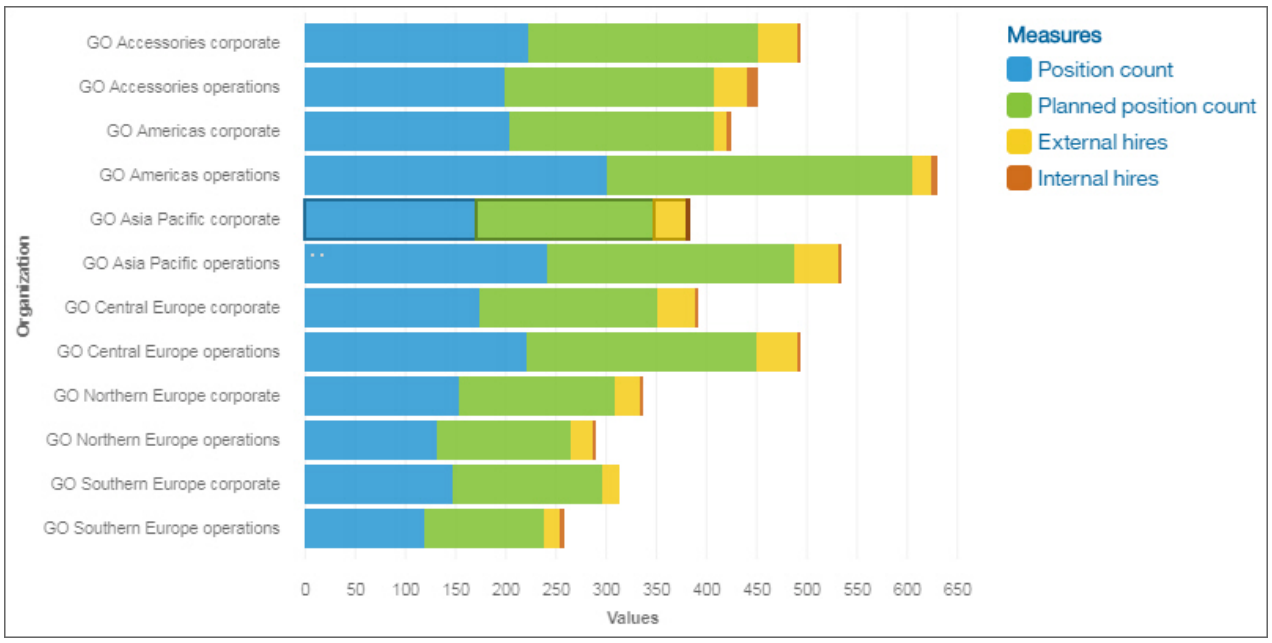


Note: Filters are not supported for spiral visualizations.

Stacked bar

Use a stacked bar visualization to compare the proportional contributions for each item to the total, such as sales for products and sales for products each month.

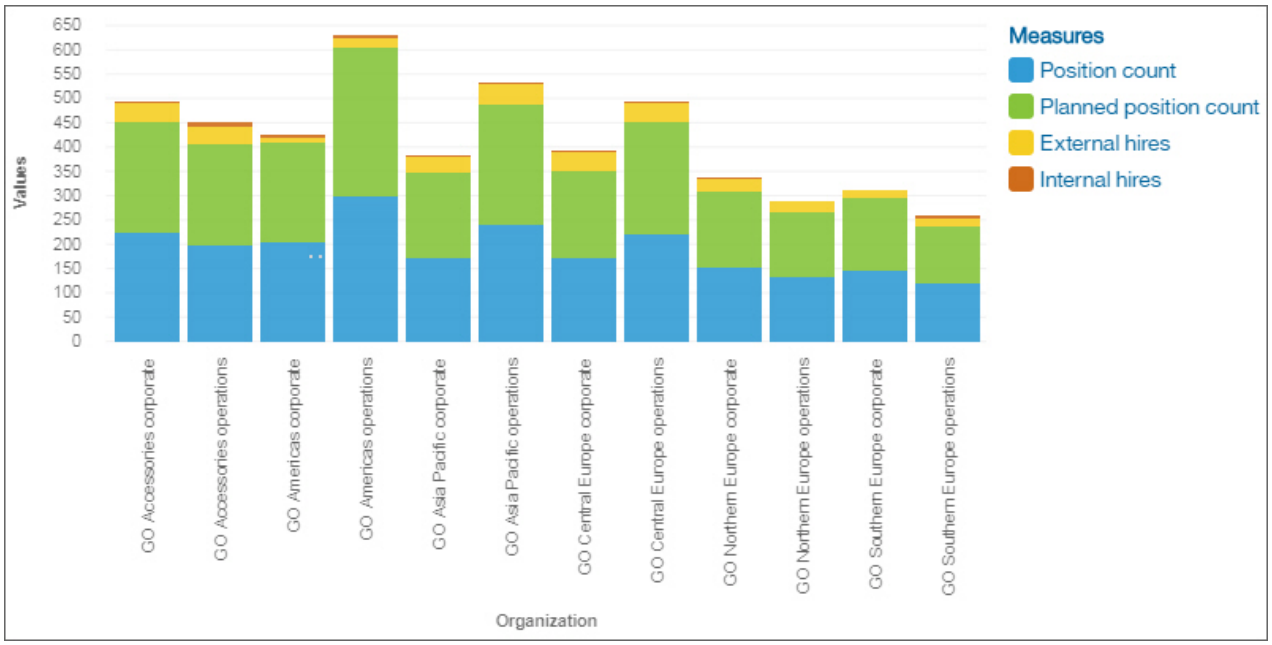
A stacked bar visualization can show change over a specific time period or compare the proportional contributions for each item to the total. If there are so many bars that the labels are impossible to read, filter the data to focus on a subset of the data or use a tree map.



Stacked column

Use a stacked column visualization to compare the proportional contributions for each item to the total, such as sales for products and sales for products each month.

A stacked column visualization can show change over a specific time period or can compare the proportional contributions for each item to the total. If there are so many bars that the labels are impossible to read, filter the data to focus on a subset of the data or use a tree map.



Summary

Use a summary visualization when you want to see the total for a measure or the count for a categorical column.

For example, this summary visualization shows total revenue for all product types.

For example, this summary visualization shows the number of departments in your organization.

Sunburst

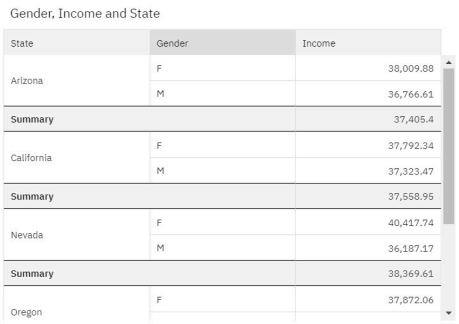
A sunburst visualization is used to illustrate how underlying data predicts a chosen target and highlights key insights.

For more information about the sunburst visualization, see “Exploring a decision tree visualization” on page 20.

Table

Use a table to show detailed information from your database, such as product lists and customer lists. A table shows data in rows and columns. Each column shows all the values for a data item in the database or a calculation based on data items in the database.

For example, this crosstab shows the income for states by gender.



Gender, Income and State

State	Gender	Income
Arizona	F	38,009.88
	M	36,766.61
Summary		37,405.4
California	F	37,792.34
	M	37,323.47
Summary		37,558.95
Nevada	F	40,417.74
	M	36,187.17
Summary		38,369.61
Oregon	F	37,872.06

Adding more columns to a table

You can focus on points that are of interest to you by adding more data to the visualization.

1. Drag another column to the field where you want additional data.
2. Drop the column beside the existing column.

Starting from Cognos Analytics version 11.1.4, you can drag data from the **Selected sources** pane and insert data in a column/row or drop the data on top of existing data to replace it.

Treemap

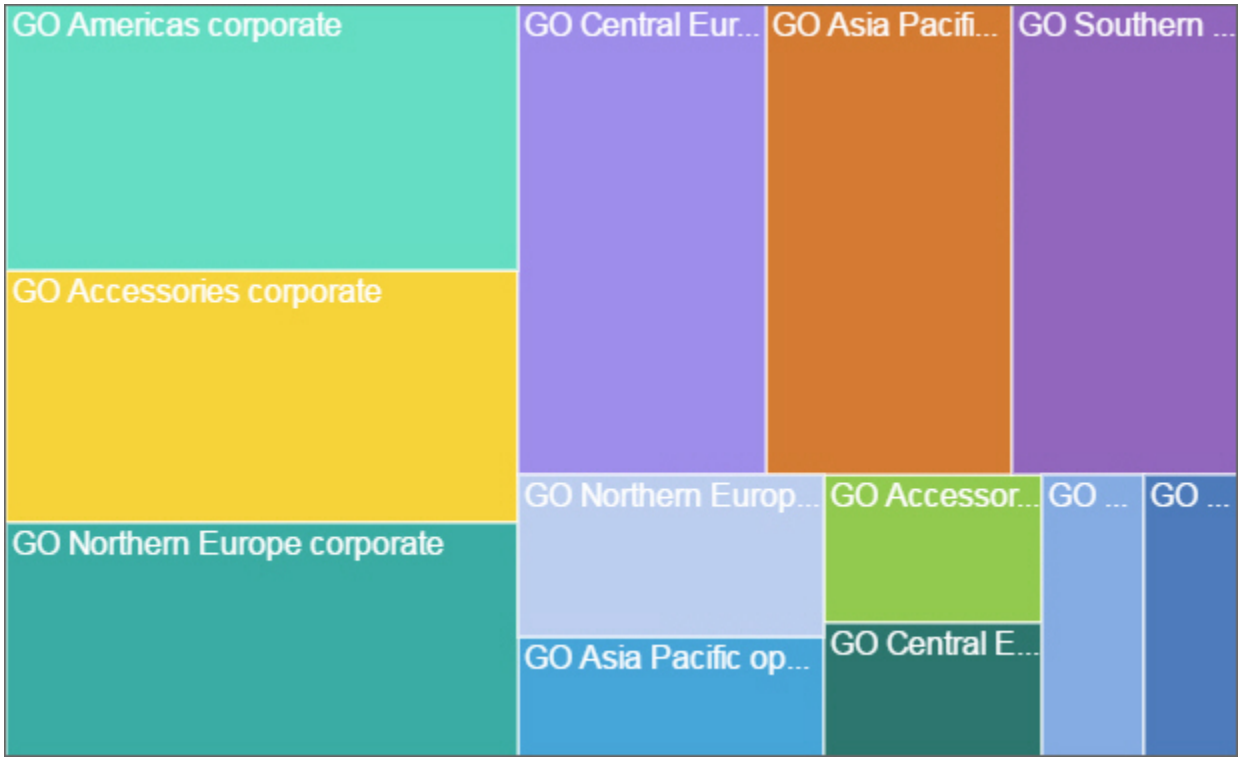
Use a tree map visualization to identify patterns and exceptions in a large, complex data asset.

Treemaps show relationships among large numbers of components by using size and color coding in a set of nested rectangles.

A treemap that is colored by category identifies the level 1 category by color. The sizes of the rectangles represent the values. In a treemap that is colored by value, the sizes of the rectangles represent one of the values and the color represents a second set of values. Do not use data that includes negative numbers. A treemap ignores negative numbers.

Many data assets have a hierarchical structure. For example, you have data about the profit margin of food items in a grocery store. Under the general category of fruit, there is a category for citrus fruit. Various citrus fruits are listed, such as grapefruit, orange, and lemon. A treemap tells you how each citrus fruit is performing when compared to each other and to other types of food.

For example, this treemap visualization shows course cost by organization.



To deselect a box that you've selected, Ctrl+click the selected box.

Waterfall

Use a waterfall visualization to understand the cumulative effect a series of positive and negative values have on an initial value. The bars in a waterfall visualization are not totals.

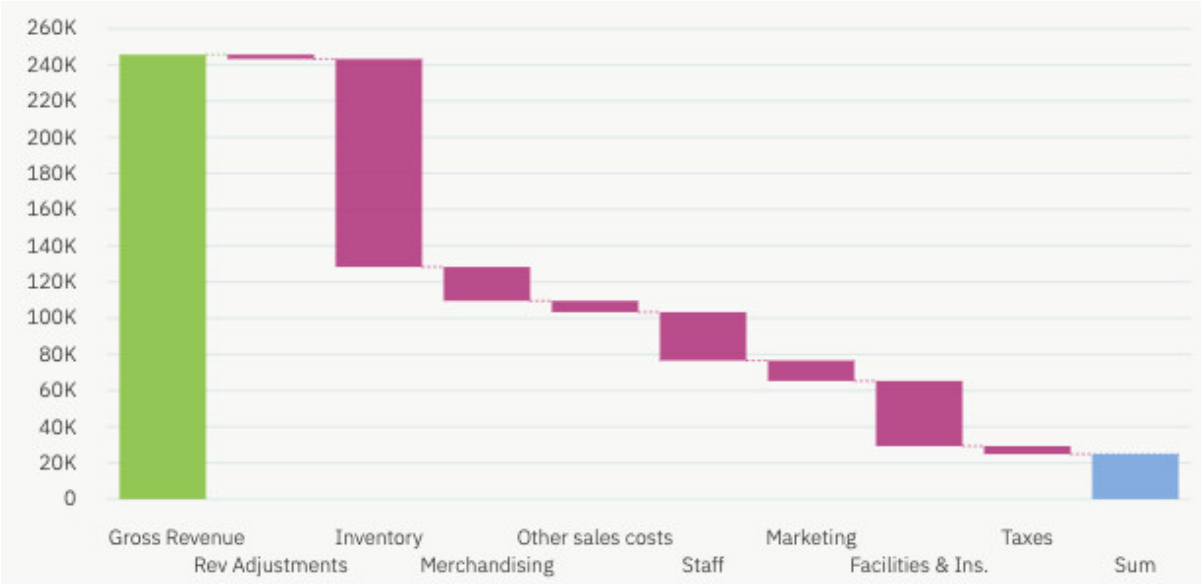
A waterfall visualization shows how an initial value is increased and decreased by a series of intermediate values, leading to a final cumulative value shown in the far right column. The intermediate values can either be time-based or category-based.

Some examples of waterfall visualizations are as follows:

- Viewing the net income after you add the increases and decreases of revenue and costs for an enterprise over a quarter.

- Cumulative sales for products across a year with an annual total.

This waterfall visualization shows the policy holder delta by month.



Creating a waterfall visualization

1. Create a new exploration. For more information, see "Starting a new exploration from the Open menu" on page 1.
2. Open the sample data module: **Select a source** > **Team content** > **Samples** > **Data** > **Customer analysis**.
3. Click **Visualizations** and click **Waterfall** to add the waterfall visualization to the exploration.
4. Click **Sources**

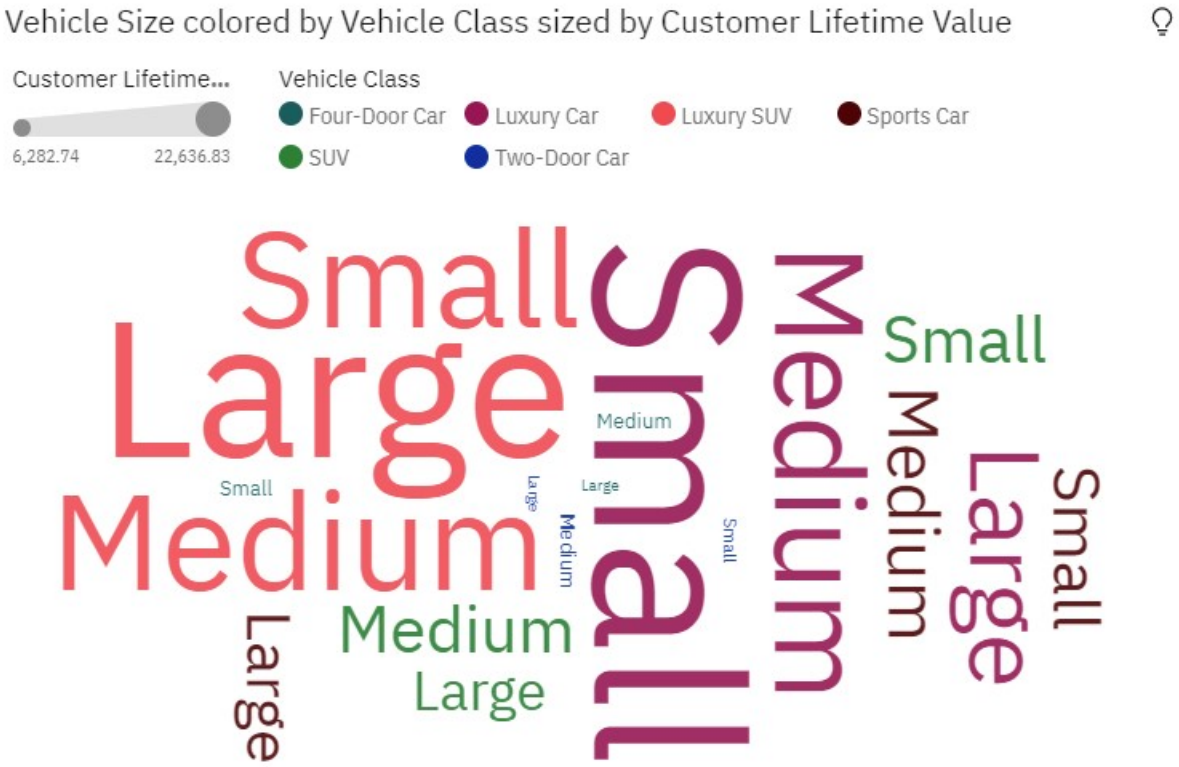
   
5. Drag the following data items from the Policy Holders section:
   - Drag **Month** onto the **x-axis**.
   - Drag **Delta** onto the **y-axis**.

Word cloud

Use a word cloud visualization when you want to see a text-based visualization of a column. The text height represents the scale. The name itself is the different members of the column.

Tip: The data asset should contain at least 15 columns and at least 100 rows to create an effective word cloud.

For example, this word cloud visualization shows the customer life time value by vehicle size and class.



The word cloud was created by dragging the following data items from the Sources panel:

- Drag **Vehicle size** type onto the **Words** field.
- Drag **Average CLTV** onto the **Size** field.
- Drag **Vehicle class** onto the **Color** field

Samples

You can see an example of a word cloud visualization in the sample report **Customer lifetime value analysis**. You can find the sample here: **Team content** > **Samples** > **Reports** > **Customer lifetime value analysis**.

If any of the sample objects are missing, contact your administrator.

Insights in visualizations

IBM Cognos Analytics provides analytic insights that help you to detect and validate important relationships and meaningful differences based on the data that is presented by the visualization.

Insights are available by clicking the **Insights** icon or the **Analytics** tab in eligible visualizations.

Note: Insights are available only for packages and are not supported in visualizations for data modules containing a Planning Analytics cube.

When you turn on insights, the summary appears in the **Insights** box and related visualization elements are highlighted and details are provided in the corresponding tooltip message. You can control each available insight separately.

Procedure

1. In a visualization that supports insights, click the **Insights** icon or the **Analytics** tab.
2. Enable **Insights**.

3. Depending on the visualization, the following insights are shown:
   - **Average** Provides the mean of the displayed target value.
   - **Predictive strength** Shows the predictive strength of the relationship between the target and explanatory fields.
   - **Fit line** Shows when either a linear or quadratic relationship exists between the target and explanatory fields.
   - **Meaningful differences** Shows values that are most significantly higher or less than the average or trend.
   - **Most frequent** Shows values that are most frequently reported.

Choosing correlated insights

Based on your visualization you are presented with statistically based, correlated, insights.

About this task

If correlated insights are available that are related to the main visualization, a green icon with a number 

is shown on either the x-axis or y-axis. The number indicates the available correlated insights.

To access correlated visualizations, complete the following actions:

Procedure

1. From the visualization, click the green icon 

.
2. Click any of the statistically based insights that are presented from the menu.
   A new card is created.

Choosing recommended visualizations

Recommended visualizations are thumbnails that display visualizations that might be appropriate for your data.

Procedure

1. From the **Cards** pane, select the card that represents the visualization you want to open.
2. In the toolbar, click the icon for the name of the visualization you have open. For example, if you are looking at a bar chart, click **Bar**.
   The recommended visualizations are displayed.
3. Click the thumbnail for the recommended visualization that you want to work with.

Choosing related visualizations

When a visualization is in focus in your exploration, the system recommends some related visualizations that are not specifically what you requested. Based on the data analysis, these related visualizations might be of interest to you.

About this task

Related visualizations replace one of the data elements in the visualization or add another data element to create a new visualization. Related visualizations use a combination of learned user interactions, statistics and interestingness to suggest useful next steps.

To access related visualizations, complete the following actions:

Procedure

1. From the **Cards** pane, select the card that represents the visualization you want to open.
2. Click **Related**  in the toolbar.

Chapter 4. Forecasting

Forecasting

Use forecasting in IBM Cognos Analytics to discover and model trend, seasonality, and time dependence in data.

You can forecast in IBM Cognos Analytics by using automated tools that model time-dependent data. Automated model selection and tuning makes forecasting easy to use, even if you are not familiar with time series modeling.

Forecasts and their confidence bounds are displayed in visualizations as a continuation of historic data. You can also view the statistical details for generated models if you want to see the technical background.

Specifying time series in forecasts often requires data manipulation. Cognos Analytics supports a wide range of time series without the need for manipulation, ranging from standard date and time types, to nested periodic and cyclical time fields. When data is recognized as a time series, data preparation is automated. Appropriate trend and seasonal periods are detected, and models are selected from a set of nine different model types.

You can forecast in line, bar, and column visualizations. Forecasting allows analysis of hundreds of time series per visualization. Forecasts and confidence bounds are computed for each time series, and displayed in the visualization as extensions of the current data. You can inspect each time series separately, and tailor the forecast and results to your own data and requirements.

If you are familiar with forecasting models, you can view the selected model type, estimated model parameters, standard accuracy measures, and processing summary information.

Note: In Cognos Analytics 12.0.2 and earlier versions, the **Forecast** feature is not available in dashboards that are created from OLAP-sourced data modules, such as data modules created from Planning Analytics cubes. Only dashboards that are created from OLAP-sourced, enriched packages support forecasting in these versions of Cognos Analytics. For more information, see Creating an enriched package from a Planning Analytics cube.

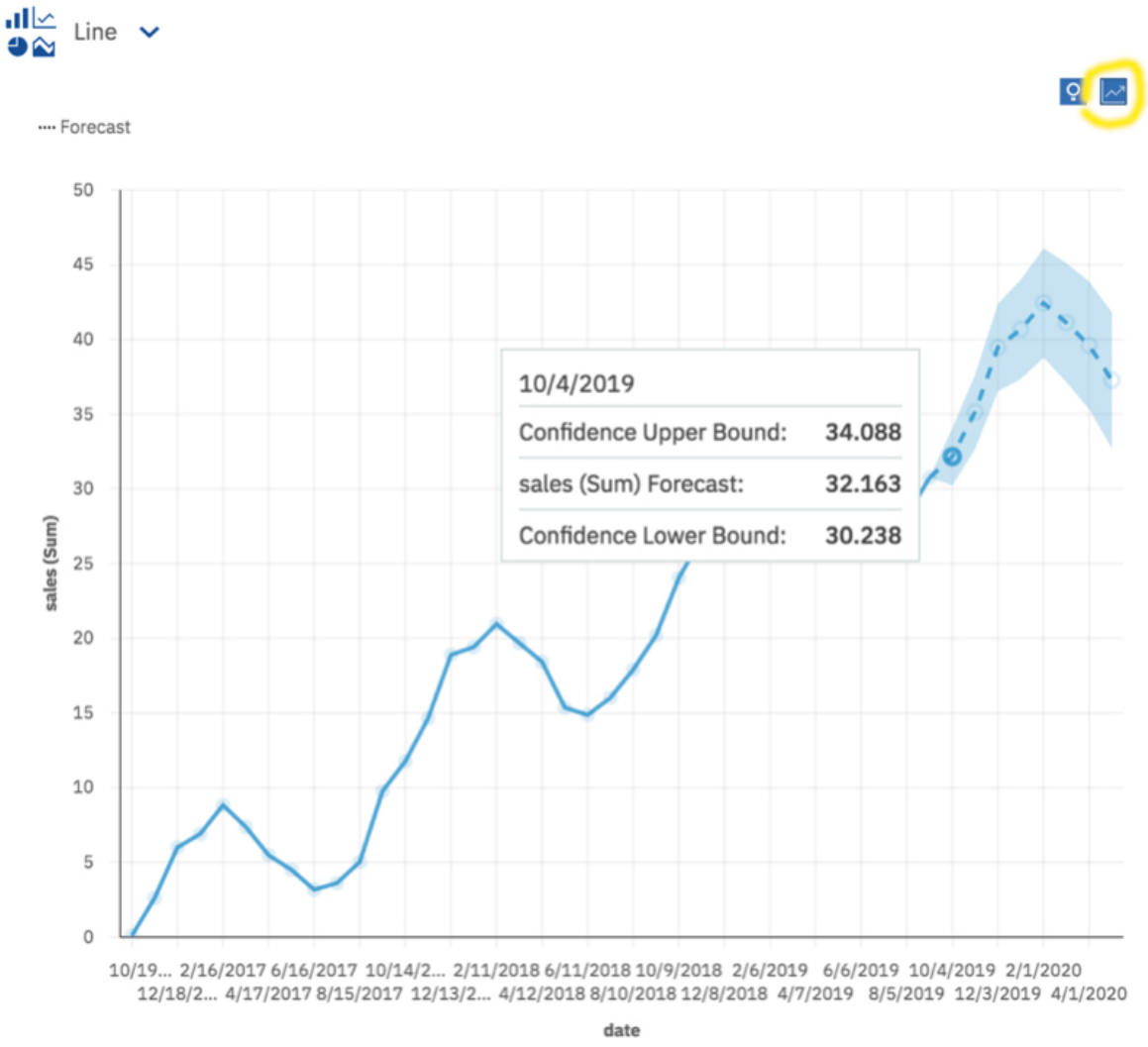
Starting with Cognos Analytics 12.0.3, dashboards that are created from OLAP-sourced data modules support forecasting.

Forecasting features

Forecasting provides time series data modeling and forecasts based on data in visualizations.

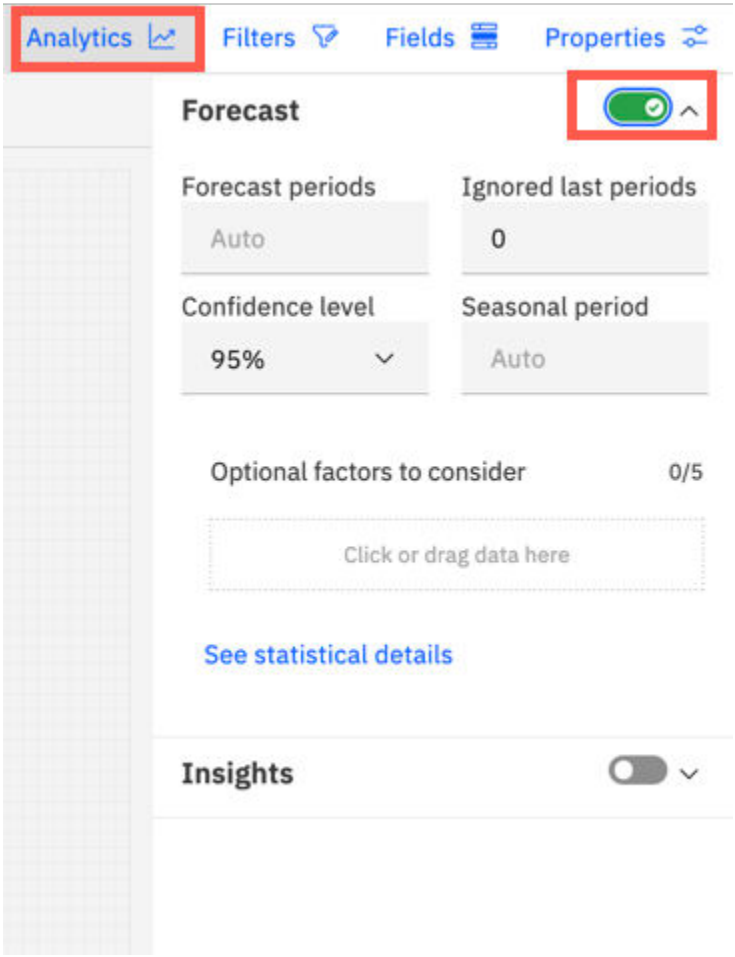
To use forecasting, the visualization must be either a line, bar, or column visualization, the data must be supported for forecasting, and forecasting must be enabled. Click the **Forecast** icon in the visualization or the **Analytics** tab and enable **Forecast**. You can modify model and forecast settings, and confidence bounds. Appropriate time series models for the visualization are estimated, and forecasts are displayed in the visualization. You can also see the time series model specification and data processing summary in the data tray.

The following example shows forecasting values and confidence bounds in a visualization.


Line
Forecast
50
45
40
35
30
25
20
15
10
5
0
sales (Sum)
10/4/2019
Confidence Upper Bound: 34.088
sales (Sum) Forecast: 32.163
Confidence Lower Bound: 30.238
10/19... 2/16/2017 6/16/2017 10/14/2... 2/11/2018 6/11/2018 10/9/2018 2/6/2019 6/6/2019 10/4/2019 2/1/2020
12/18/2... 4/17/2017 8/15/2017 12/13/2... 4/12/2018 8/10/2018 12/8/2018 4/7/2019 8/5/2019 12/3/2019 4/1/2020
date

Forecasting options

You can modify your forecasts by setting a number of period and confidence level options in the **Analytics** panel.



A period is the smallest time interval between neighboring points in the data.

The following options are available.

Forecast periods

The number of steps to forecast ahead.

The default value is **Auto**, which is 20% of the length of the historical data. Any missing values at the end of a particular series will also be forecast, but they will not count towards the specified number of forecast periods.

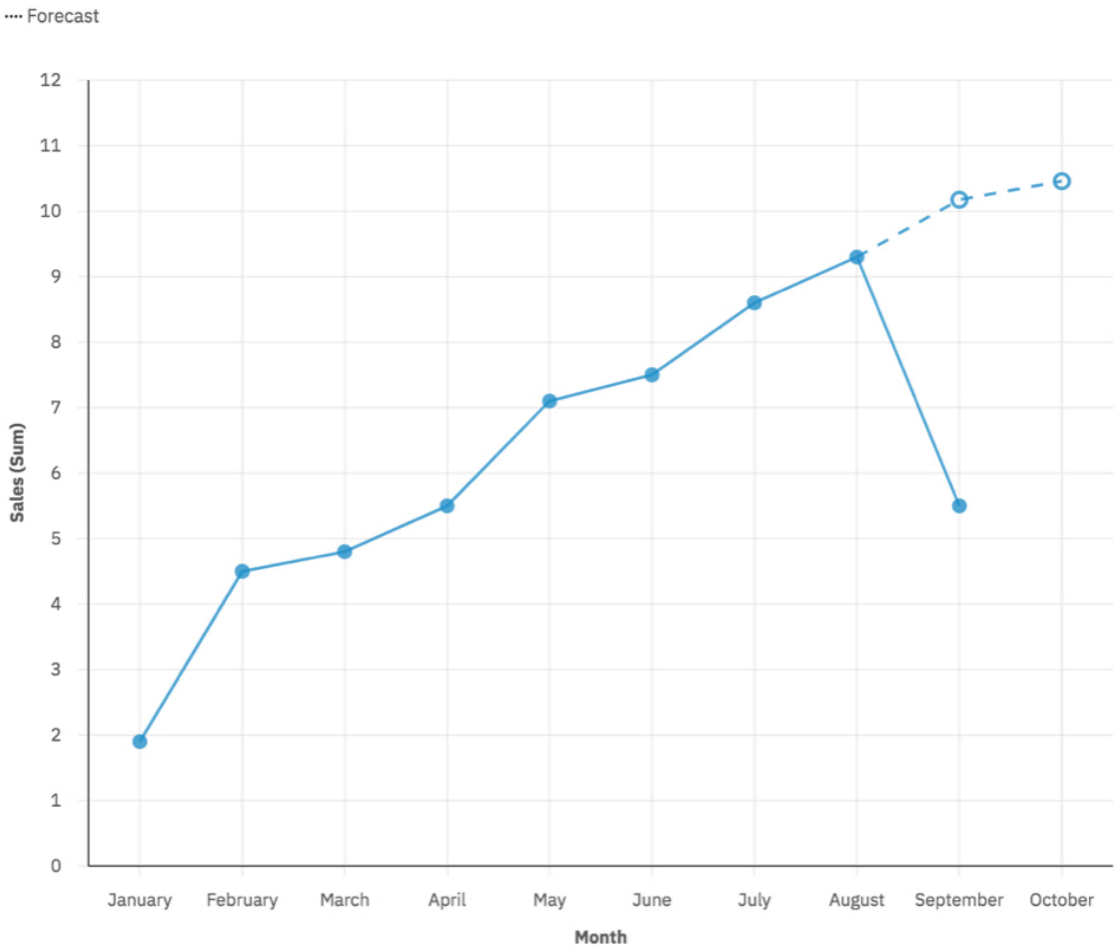
Ignored last periods

Ignores a specified number of data points at the end of a time series when building the model and computing the forecasts. Any missing values at the end of a non-ignored portion of a series will also be forecast. Ignored last periods value must be specified as a non-negative integer, such as: 0, 1, 2, 3.

The default value is 0. If there are no missing values, then all of the historical data is used in model generation and the first forecast point is after the last historical data point. Up to 100 data points can be ignored.

Ignoring the last data period can be useful when the data is incomplete. For example, you might be doing a forecast halfway through a month. Exclude this month from the forecast by setting **Ignored last periods** to 1.

The following visualization shows a forecast that ignores September's results by setting **Ignored last periods** to 1.



Confidence level

The certainty with which the true value is expected to be within the given range. You can see corresponding confidence interval in a tooltip by hovering over any forecast value. The confidence interval is displayed as upper and lower bounds.

You can select three different confidence levels: 90%, 95%, and 99%. The default is 95% and the lower and upper bound define the range at which you can be 95% confident that the true value lies within that range.

Seasonal period

The seasonality with which to build the model. Seasonality is when the time series has a predictable cyclic variation. For example, during a holiday period each year.

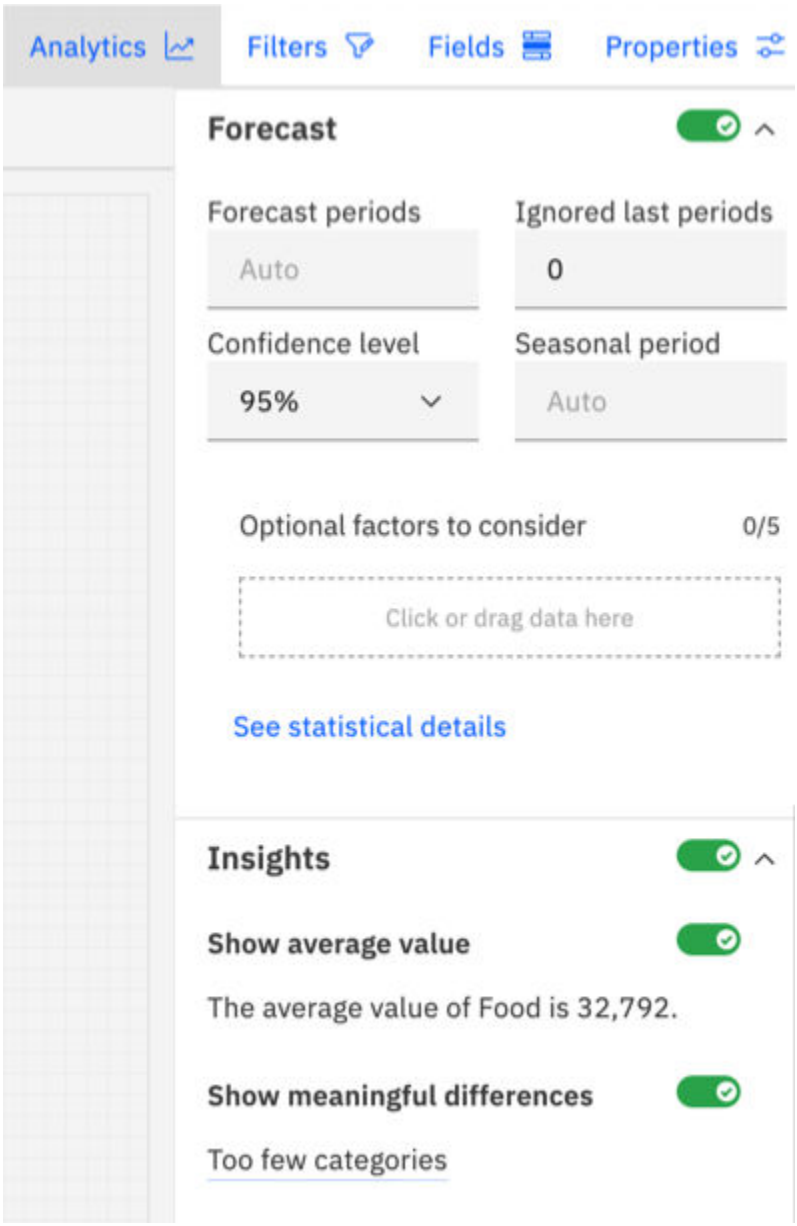
The default value is **Auto**. **Auto** automatically detects seasonality by building multiple models with different seasonal periods and choosing the best one.

You can specify seasonality by entering a non-negative integer, such as: 0, 1, 2, 3 as the seasonal period.

To specify a non-seasonal model, set the **Seasonal period** to 0 or 1. A model with user specified seasonality is displayed only if the seasonal model is more accurate than all of the non-seasonal models.

Insights

When visualizations have both Insights and Forecasts, you can enable both of them in the **Analytics** panel. Insights provide an independent set of analytic results. For more information, see "Insights in visualizations" on page 69.



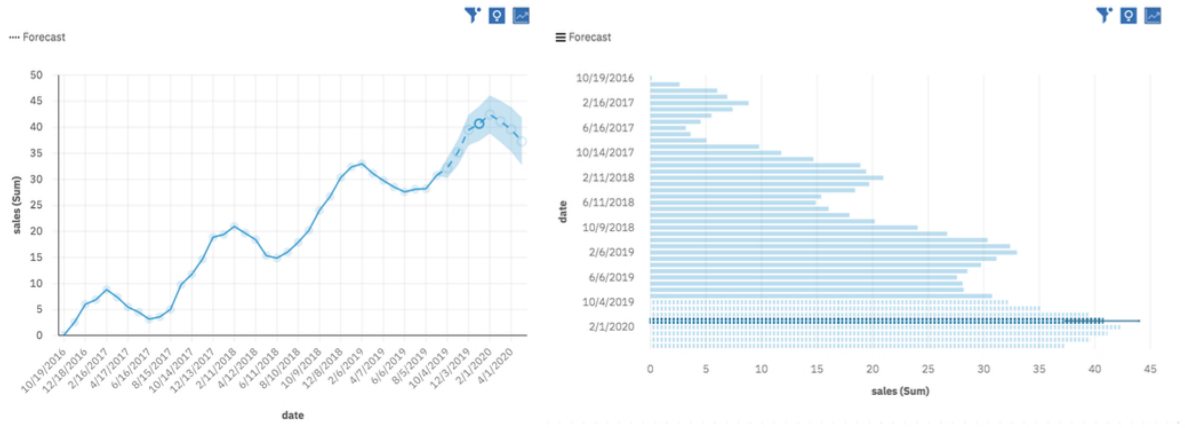
Visualization types that support forecasting

Forecasting is supported in line, bar, and column visualizations.

The following table compares the forecasting display features for each visualization.

Forecasting features	Line chart	Bar chart	Column chart
Forecast points	Open circle	Striped bar	Striped column
Confidence interval display	Shaded region	Solid line	Solid line
Activate confidence interval	Click any point	Click a forecast bar	Click a forecast column
Number of confidence intervals displayed	All	1	1

The following image shows a forecast on line and bar visualizations with activated bar intervals.



Insights and forecasting

Insights in visualizations provide analytic insights that can help users to detect and validate any important relationships and meaningful differences based on the data that is presented by the visualization. Insights work alongside forecasting in supported visualizations. Insights provide a separate set of analytic results, and the results are for historical values only. For more information, see “Insights in visualizations” on page 69.

Note: Support for forecasting and insights is available for packages only and is not available in visualizations for data modules containing a Planning Analytics cube. However, you can use the forecasting feature in an eligible visualization with a Planning Analytics cube, if you create an enriched package from the cube. For more information, see Creating an enriched package from a Planning Analytics cube.

Forecasting data

Data that is suitable for forecasting has measure values that correspond to regularly spaced time points. You specify time and measures in visualizations by dragging time fields and measure fields into visualization slots. Optionally, you can also specify group fields that split the measure values by categories.

The following table is a summary of the field types and matching visualization slots that are supported in forecasting:

Slot	Time fields (required)	Measure fields (required)	Group fields (optional)
Line chart slot	x-axis	y-axis	Color
Bar chart slot	Bars	Length	Color
Column chart slot	Bars	Length	Color

No other visualizations or visualization slots are supported, with the exception of the **Local filters** slot

Time fields in forecasting data

A time field is identified by a time icon in front of the field label in the **Data** pane.

You can specify time field properties by using the following properties: **Data type** or **Represents Time**.

Data type

A field is recognized as a time field if it has one of the following data types: Date, Time, or Timestamp. Data type is inherited from the data source and cannot be changed.

Date, Time, and Timestamp data types are designed to support the full range of date and time formats that are covered by the ISO 8601 basic and extended formats. The following table shows the supported data types together with an example of format and a data example for each.

Data type	Format example	Data example
Date	yyyy-mm-dd	2019-07-01
Time	hh:mm:ss	12:34:56
Timestamp	yyyy-mm-dd'T'hh:mm:ss	2019-07-01T12:34:56

Represents Time

A field is recognized as a time field if the data property **Represents** is set to **Time**. Text and Integer fields that contain time data are also recognized as time fields. Time fields are defined automatically during data import or enrichment. The possible definitions are Date, Year, Quarter, Season, Month, Week, Day, Hour, Minute, or Second.

If time fields are not recognized automatically, you can specify them as time fields. Ensure that the field values are in one of the supported formats, otherwise you might receive an `Unsupported date format` error.

Nested time fields

You can drag multiple time fields into the same visualization slot to specify a nested time field. For example, a field that represents Week can be dragged into the slot along with a field that represents Day to create a forecast by Days of the Week.

Nested fields in the slot must be in time hierarchy order. For example, Week must be placed above Day.

Nested fields cannot skip levels in the time hierarchy that would result in ambiguity. The following table describes acceptable hierarchies.

Time field	Acceptable lower fields
Year	Quarter, Month, Week, Day
Quarter	Month
Month	Day
Week	Day
Day (of Year, Month, or Week)	Hour, Time
Hour	Minute
Minute	Second

If `Year` is absent in the time hierarchy, then the system defaults to the current year. This can cause issues due to differences between leap and non-leap years. Consider providing the Year in such instances.

Chronological data order

Specified time fields define a chronological order for the time points in the visualization. They are used to sort the points on the visualization in chronological order when forecasting is enabled. The chronological

order includes the historical points, along with the new forecasted points. Any other sorting criteria that are specified for the visualization are ignored when forecasting is enabled. For example, the first day of the week is always Sunday even if you specify Monday as the first day of the week in the custom sort order.

Any invalid time labels are moved to the beginning of the sequence and excluded from building the model and computing the forecast.

Time interval detection

Time interval detection is possible when the data is ordered chronologically. The time interval is the size of the smallest interval between any two adjacent time points, such as "2 weeks". If varying time intervals are detected, they must all be integer multiples of the smallest interval. Otherwise, the data is deemed irregular and cannot be forecast. Missing time points that arise as a result of multiple intervals are filled in for the detected interval. Corresponding measure values are set to missing. If the number of missing values is larger than 33% of the series length, a `Too many missing values` error is reported.

Measure fields

One or more fields of any type can be specified as measure fields for forecasting analysis by adding them to a corresponding visualization slot. Each measure field is analyzed separately. Multiple time series can also be specified by adding a field to the **Color** slot, splitting the measure values by the categories of the specified field.

All measure field values that correspond to the same time point are summarized by using one of the following summarization levels: **Sum**, **Minimum**, **Maximum**, **Average**, **Count**, and **Count distinct**. The field must be numeric to support **Sum**, **Minimum**, **Maximum** , or **Average** summarization. All possible data types and summarization levels are supported for forecasting. However consider the following points:

- Small number of different measure values can result in unexpected or uninformative forecasts. For example, when **Count distinct** summary is used.
- Zero measure values can unduly influence results, especially when they represent missing measurements.

Interpolating missing values

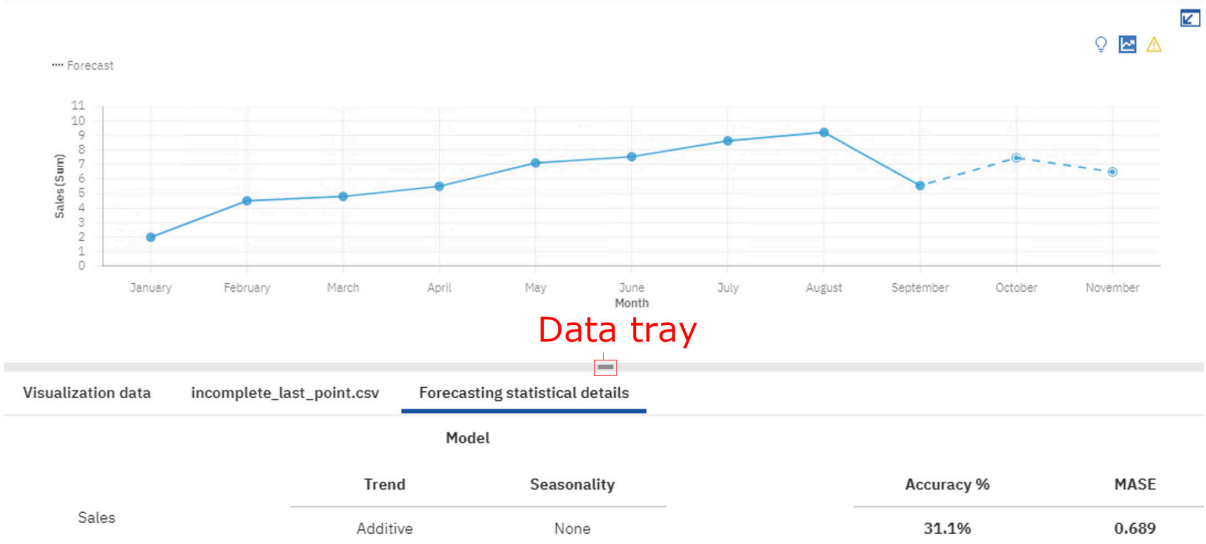
Missing values are computed and filled in by the **Linear Interpolation** algorithm. The computation is based on the nearest neighbors in a chronologically ordered time series with detected time interval. The new value is `(previous value + next value)/2`. For example, with the values [3, 6, missing, 12], the interpolated value to replace the missing value is (6 + 12) / 2 or 9. The interpolation algorithm can also handle contiguous missing values.

Data points with missing values at the first or the last historical time points are excluded from the series before building a model. Missing values at the last historical time points get forecasted as well.

Forecasting statistical details

A forecasting run generates forecasts and forecasting statistical details. Forecasting statistical details are located in the data tray at the bottom of each visualization. There is a single row of statistical details for each time series in the visualization. Forecasting details are generated as long as the time points are evenly spaced.

Forecast information contains forecast **Status** for the given time series. When the status is **Success**, the other fields provide details on the model and data used for forecast. When the status is **Failure**, some of the other fields, including the **Notes**, provide details regarding the cause of the failure. Summaries of any failures are always provided in the visualization warnings.



Model information specifies the **Trend** and **Seasonality** type that is selected for estimating the time series data when successful. The following table lists the different available types.

Trend component	**Seasonal component**		
	N NONE	A ADDITIVE	M MULTIPLICATIVE
N NONE	(N, N)	(N, A)	(N, M)
A ADDITIVE	(A, N)	(A, A)	(A, M)
Ad ADDITIVE_DAMPED	(Ad, N)	(Ad, A)	(Ad, M)

Model influencers

For multivariate forecasts, the **Model influencers** section explains how much factor weight each field contributed to the forecast. A higher value indicates that the field was more relevant to the forecast. A lower value indicates that the field has a lot of noise (that is, the field is more random and unpredictable) and is less relevant to the forecast.

Which fields contributed to the forecast depends on the type of multivariate forecast used: Vector autoregression (VAR) or autoregressive integrated moving average with explanatory variable (ARIMAX).

Accuracy measures

Model accuracy measures Mean Absolute Error (MAE), Mean Absolute Scaled Error (MASE), Accuracy Percent, Root Mean Squared Error (RMSE), Mean Absolute Percent Error (MAPE), and Akaike Information Criterion (AIC) are based on the time series data that is used to generate the model. All accuracy measures are based on the historical data. Accuracy measures can also be used as an indicator of the forecast accuracy, but they do not carry over to future values.

Mean Absolute Error (MAE)

Computed as the average absolute difference between the values fitted by the model (one-step ahead in-sample forecast), and the observed historical data.

Mean Absolute Scaled Error (MASE)
The error measure that is used for model accuracy. It is the MAE divided by the MAE of the naive model. The naive model is one that predicts the value at time point t as the previous historical value. Scaling by this error means that you can evaluate how good the model is compared to the naive model. If the MASE is greater than 1, then the model is worse than the naive model. The lower the MASE, the better the model is compared to the naive model.

Accuracy Percent (Accuracy %)
The primary indicator of the model accuracy based on the fitted values. It is specified as the reduction percentage of mean absolute error relative to the naïve model. It is computed by subtracting MASE from 1 and expressing it as percentage. If MASE is greater than or equal to 1, the accuracy is set to 0% because the model does not improve upon the naïve model. Higher accuracy indicates lower model error relative to the naïve model.

Mean Squared Error (MSE)
The sum of squared difference between the values that are fitted by the model, and observed values that are divided by the number of historical points, minus the number of parameters in the model. The number of parameters in the model is subtracted from the number of historical points to be consistent with an unbiased model variance estimate.

Root Mean Squared Error (RMSE)
The square root of the MSE. It is on the same scale as the observed data values.

Mean Absolute Percent Error (MAPE)
The average absolute percent difference between the values that are fitted by the model and the observed data values.

Akaike Information Criterion (AIC)
A model selection measure. The AIC penalizes models with many parameters, and so attempts to choose the best model with a preference towards simpler models. The AIC is the sum of the logarithm of non-adjusted MSE multiplied by the number of historical points and the number of model parameters and initial smoothing states that are multiplied by 2.

Parameters

Detected Seasonal period and estimates for other parameters that are used in the selected exponential smoothing model are available.

Seasonal period
The number of time steps in a seasonal period used in the exponential smoothing model.

Alpha
The smoothing factor for level states in the exponential smoothing model. Small values of alpha increase the amount of smoothing, that is, more history is considered when the alpha is small. Large values of alpha reduce the amount of smoothing, which means that more weight is placed on the more recent observations. When the alpha is 1, all the weight is placed on the current observation.

Beta
The smoothing factor for trend states in the exponential smoothing model. This parameter behaves similar to alpha, but is for trend instead of level states.

Gamma
The smoothing factor for seasonality states in the exponential smoothing model. Serves the similar role as alpha, but for the seasonal component of the model.

Phi
The damping coefficient in the exponential smoothing model. Long forecasts can lead to unrealistic results, and it is useful to have a damping factor to dampen the trend over time and produce more conservative forecasts.

Diagnostics

Information includes Missing count, Series length, Ignored periods, Trend strength, Seasonality strength, and Date/time interval.

Missing count
Indicates the number of data rows that have either missing values or missing time points and are positioned between the first and the last valid series value. Invalid time points, as well as points with missing values at the first or the last historical time points are not included.

Series length
Indicates the number of data points used for time series modeling. Only the points between the first and the last valid series value are included.

Ignored periods
An integer, m, that ignores the last m data points of the series when building the exponential smoothing model and computing the forecasts. Any missing values at the end of a non-ignored portion of a series will also be forecast. The default value for this parameter is 0, which means that all of the historical data is used in model generation when there are no missing values. A maximum potential of 100 points can be ignored. Ignored periods excludes the data points when building a model, so forecasting might fail due to factors such as minimum data length requirements and missing value proportion that exceeds 33%.

Trend strength
Compare the original model, M, with the same model, but with the trend component removed. The trend strength of M is the difference in accuracy between model M and model M with the trend component removed.

Seasonality strength
Compare the original model, M, and the same model with the seasonal component removed. The seasonality strength of M is the difference in accuracy between model M and model M with the seasonal component removed.

Date / time interval
The Date / time interval represents the detected time interval of the chronologically sorted data. The time interval is identified as the smallest difference between neighboring points in the data when sorted in the chronological order.

Forecasting models

Exponential smoothing models are a popular class of time series models.

Exponential smoothing models are applicable to a single set of values that are recorded over equal time increments only. However, they support data properties that are frequently found in business applications such as trend, seasonality, and time dependence. All specified model features are estimated based on available observed data. An estimated model can then be used to forecast future values and provide upper and lower confidence bounds for the forecast values.

Each model type is suited for modeling a different combination of properties that are found in the data. The model type that can provide the best match to the observed data is selected for modeling the observed data and is used to forecast any future values.

Model estimation algorithms

Models are specified by the smoothing equations that include the model parameters and initial smoothing states. Model parameters are estimated with values that minimize the model error.

Smoothing equations

Exponential smoothing models derive their name from the smoothing equations that specify the model. They provide formulas for computing smoothing states for each observed point by using the current observed value and the previous smoothing states. Smoothing equations provide weighted averages of the current value and the previous states in the time series. Weight for the current value or state is given by a model parameter between 0 and 1, while the weights for previous values are exponentially decreasing.

Level smoothing equations

All model types compute a level state for each time series point by using the corresponding level smoothing equation. Level states for the model without trend and seasonal components are computed as the weighted average of the time series value at the current point and the level state at the previous point. The weight that is associated with the current value is a parameter, **alpha**, with its value restricted between 0 and 1. For other models, previous trend and seasonal states are also included in the level smoothing equation.

Trend smoothing equations

Model types with additive or damped additive trend compute a trend state for each time series point by using the corresponding trend smoothing equation. The trend state for the current point is based on the difference of level states at the current and the previous point, and on the trend state at the previous point. The weight that is associated with the difference of level states at the current and previous point is a parameter that is named **beta** with its value restricted between 0 and 1. An extra parameter, **phi**, is added to the damped trend smoothing equations. **Phi** multiplies the trend state contribution from the preceding point and its value is also restricted between 0 and 1. The purpose of this parameter is to estimate the degree of trend damping from one point to the next.

Seasonal smoothing equations

Model types that support additive or multiplicative seasonality compute a seasonal state for each time series point. The seasonal states are computed by using seasonal smoothing equations. The seasonal state for the current point includes the difference of the time series value and the current level state for additive seasonality or ratio of the two same values for multiplicative seasonality. The weight that is associated with this term is a parameter, **gamma**, with its value restricted between 0 and 1. The rest of the contribution comes from the corresponding seasonal state in the previous seasonal period. Notice that the seasonal period has a fixed length, and while the seasonal state can change for each point, only matching seasonal indices from different periods are considered together in the seasonal smoothing equations.

Initial smoothing states

Values must be specified for level, trend, and seasonality states for points that precede the time series. The values are needed for the smoothing equations. To compute the various states at the first point of the time series requires state values at the corresponding previous points.

Model parameters

Each smoothing equation uses corresponding model parameters:

alpha
Controls the level states.

beta
Controls the trend states.

gamma
Controls seasonal indices across seasonal periods.

phi
An extra parameter that is used for specifying the damped trend.

All four parameters have values between 0 and 1. Higher values of **alpha**, **beta**, and **gamma** mean that more recent observations have higher weight, while lower values mean higher weights for older observations. A higher value of **phi** corresponds to a higher degree of dampening the forecast trend.

Model estimation

Model parameters in the smoothing equations are estimated based on the time series data. Parameters cannot be estimated directly by using a formula. They are estimated by an iterative process that searches for parameter values that minimize the model error. The model error is computed as Mean Absolute Scaled Error (MASE). The iterations stop when no further reduction in the model error can be achieved. Corresponding parameter values together with the initial smoothing states fully specify the estimated model. They are used to compute the model states for all other data points and generate the model forecasts by using a corresponding forecast equation.

Forecasting algorithms

A number of algorithms are used in forecasting.

One-step ahead

Every model supports one-step ahead forecasts based on the corresponding forecast equation. One-step ahead forecasts are needed to compute model errors during the model estimation process.

One-step ahead forecasts are computed sequentially for each data point by using computed level and trend states for the current point, and seasonal states for the last seasonal period.

Forecast error is computed by subtracting forecast value at the previous point from the observed value at the current point. Overall model error, which is used for estimating the model, is computed as an average value of absolute forecast errors. Smaller errors correspond to a better model fit. Accuracy measures displayed in **Forecasting statistical details** provide several model summaries of the one-step ahead forecast errors.

k-step ahead

k-step ahead forecasts are used to make predictions for any number of future values following the observed time series data. They are based on the same forecast equations as the one-step ahead forecasts for the specified model.

The number of forecast values that are generated is 20% of the length of historical data series by default. You can specify an exact number of values to be forecast in the **Forecast** dialog box. Any missing values at the end of a particular series will also be forecast, but they will not count towards the specified number of forecast periods.

Confidence bounds

Confidence bounds provide the level of uncertainty that is associated with each forecast value. The bounds typically become wider further into the future, because more distant forecasts are less reliable. Confidence bounds provide relevant insights into the future behavior of the observed time series.

Computation of confidence bounds is based on the overall variance of forecast errors that are estimated on the observed data and a factor that depends on the specified model and on the number of steps from the last observed point.

Automated model selection in forecasts

Multiple model types are used to create candidate models for each time series in a forecast. All nine available model types are normally used, except when a seasonal component is absent. There are only three model types available that do not account for seasonality in the data.

The default value, **Auto**, for the Seasonal period option detects seasonal period by comparing multiple models, each with a different candidate seasonal period.

Multiple models are compared by using a model error and the number of model parameters. For example, when model errors are equal for two models, the model with fewer parameters is preferred. The latter model provides a more condensed representation of the observed data and also tends to generate more reliable forecasts.

Chapter 5. Principles of advanced data analytics

Principles of advanced data analytics

IBM Cognos Analytics is a business intelligence tool for managing and analyzing data. It includes self-service features for users to prepare, explore, and share data. Cognos Analytics includes predictive, descriptive, and exploratory techniques, also known as numeric intelligence. Cognos Analytics uses many statistical tests to analyze your data.

It is important to understand the definitions of these tests as they apply to Cognos Analytics.

Numeric algorithms are used as part of the workflow to provide features to the user that communicate information about the numeric properties and relationships in their data.

Business oriented

Unlike traditional statistical software, where the target audience is an experienced data analyst, the algorithms of Cognos Analytics are aimed at users who are familiar with, but not an expert in data analysis. This means that when Cognos Analytics considers tradeoffs, usefulness is chosen over complexity.

Trustworthy

Business data is lot more complicated than text-book examples that are used in statistical courses or in web searches and examples. Cognos Analytics uses algorithms that are robust and that are able to cope with a range of varieties of unusual data. Cognos Analytics does this because although the algorithms that are more brittle are able to get slightly better results than robust algorithms, they require you to make sure that they are applicable and build correct data transformations for the results to be meaningful. A minor drop in accuracy is worth the safety that is provided by an algorithm, which does not give wrong results when the data is not as it is supposed to be.

Intelligent

Almost all algorithms require decisions to be made about them; levels of confidence, which combinations of fields to explore, data transformations. The details of these decisions can be found in the descriptions.

Cognos Analytics chooses appropriate values automatically, by examining properties of the data. As a user you might not discover all the decisions that are made.

Summary

In Cognos Analytics, the numeric algorithms and procedures are designed to produce trustworthy results automatically. To get the best possible prediction, classification or analysis, a professional statistician analyzes the data by using IBM SPSS Statistics or IBM SPSS Modeler. The goal of Cognos Analytics is to provide qualitative insights that help you to understand your data and its relationships, and to do so automatically for a wide variety of types of data. Cognos Analytics aims at providing results similar to a professional statistician without getting in the way of the business user.

Data preparation

Data preparation is a pre-analysis step that is used by most data analytic algorithms to ensure that the data is suitable for analytic use.

Overview

Data preparation is critical in IBM Cognos Analytics. Only prepared data is entered into analysis for key drivers, decision trees, and relationships that are displayed in the advanced analytics visualizations: Spiral, Driver analysis, Decision tree, Sunburst, and Explore relationships. Data is not automatically prepared for other visualizations and their corresponding insights.

Algorithms

All applied algorithms are based on values of a single field at a time. Missing values are removed or handled for each field, all numeric predictor driver fields are binned. All categorical fields are adjusted for large number of categories and outliers are handled in the target field. While all data preparation influences the analysis results, corresponding data preparation summaries are not currently reported to you.

Details

Data preparation and subsequent key drivers, decision trees and relationships are based on a data sample with approximately 10,000 rows when the original data is larger. Bernoulli random sampling, equal probability without replacement random sampling, is applied to uploaded data and any connected data sources that are supporting random sampling. Otherwise, systematic sampling is used.

Data preparation for numeric fields

A field is treated as numeric whenever it contains numeric information and its usage property is set to measure.

Overview

Because numeric data can be varied in their distribution, IBM Cognos Analytics transforms non-target numeric fields into ordinal bins, reducing the dependence of analytic algorithms on the format of numeric data.

Algorithms

The basic algorithm that is used is equal frequency binning. Numeric data is divided into a fixed number of bins that are attempting to put an equal number of rows of data into each bin. Missing values are placed in their own bin. Cognos Analytics attempts to use knowledge about missing values in predictor fields to build a better model. For example, if a field of data represents when an item was tested, Cognos Analytics uses missing values (which might represent that an item was never tested) to help predict the values of other fields.

Details

Certain field exclusion criteria apply to numeric fields. A numeric field is excluded from further analysis if it has only a single value, including the missing value. Otherwise, the numeric field is binned and the default number of bins is 5. If a field has no more than 10 unique numeric values, then binning is not attempted, and each unique value is given its own category. If zero occurs in more than 40% of rows, it is always given a separate category. Missing values are placed in their own bin and do not affect the binning procedure.

Data preparation for categorical fields

A field is treated as categorical whenever its usage property is set to attribute or identifier.

Overview

The main information that is extracted from categorical fields is observed frequency for each unique category value. Appropriate analytic methods are applied to categorical fields, but their accuracy and performance can be adversely affected when the number of different categories becomes large. The main data preparation step is to start merging categories when their number becomes large.

Algorithms

The basic algorithm that is used is merging categories. Categories are sorted by their frequency in descending order and the categories beyond default number are merged in a single category. Missing values are treated as a single separate category. In other words, IBM Cognos Analytics uses missing values in a similar way as for the numeric fields. Categorical fields are treated as nominal. Intrinsic order is not assumed among categories.

Details

Certain field exclusion criteria apply to categorical fields. A categorical field is excluded from further analysis if it has only a single value or the number of unique, non-merged categories exceeds 50% of the number of valid data rows.

Otherwise, the categorical field is merged and the default number of non-merged categories is 49. The rest of the categories are merged into a single extra category. All categories with row count smaller than 3 also get merged. A categorical field is also excluded if the percentage of valid data rows corresponding to the merged category exceeds 25%.

Missing values are treated as a separate category and considered in the merging step as such.

Data preparation for target fields

Specification of the target field is required for key drivers and decision tree visualizations.

Overview

Always specify the target field and at least one extra field. Models are trained by using supplied target values and are used to detect predictive relationships and eventually to predict target values given the input field values. Data preparation for the target field differs from the data preparation for the rest of the fields. Missing values in the target are not used for building models, but the rest of the information is preserved and sometimes adjusted to obtain unbiased models.

Algorithms

The main data preparation step related to target fields is removal of all data rows with missing target value. This happens before any other data preparation steps. While it ensures that only reliable information is used for model building, the number of removed rows can be substantial. The resulting model might have a limited scope in such instances. Numeric target fields are not binned, but the extreme outliers are handled to not adversely affect the later created models. Categorical target fields are treated much like other categorical fields. The only difference is that missing values have been removed for the categorical targets.

Details

Extreme outliers are detected based on lower and upper boundaries. The upper boundary is constructed by using an upper percentile such that only 2.5% percent of target values are found to have a greater value. The difference between the upper percentile and the median is multiplied by 2.5 and added to the

median to obtain the upper boundary. Similar steps are applied to obtain the lower boundary. The target values that are found beyond the computed boundaries are replaced by the corresponding boundary value in all subsequent analysis.

One-way key drivers

One-way drivers are a model-based exploratory tool.

Overview

Given a target field, the tool uses a statistical model to analyze any other available data field and estimates its strength in predicting the target values. Such data fields are called target predictors or drivers. Each potentially relevant data field is analyzed and only the top drivers regarding their predictive strength are displayed. You obtain insights regarding available drivers and their ranking according to their predictive strength for the specified target in the data. One-way driver analysis results are available both in the driver analysis and spiral visualizations. Visual drill-down into each separate driver is enabled for driver analysis visualization in Explore only.

Algorithms

Analysis for each one-way driver is based on a statistical model that includes the target and a single categorical predictor. The model is applied after the data preparation step for the target field and all potential predictor fields. For example, all numeric predictor fields are binned during the data preparation step and treated as categorical in the analysis. One-way ANOVA is applied for numeric targets and Chi-square test of independence is applied for categorical targets with the chi-square adjustment for sparse data.

For each field in the list of potential drivers, a hypothesis test on whether the field has a significant impact on the target is performed. Only those fields which pass the test and have sufficiently high predictive strength are selected as possible one way key drivers.

Details

Preliminary analysis based on smarts capabilities reduces the number of potential drivers in some cases. The goal is to remove irrelevant or redundant fields. The list of used drivers is available in the UI and you can add any initially excluded drivers to the analysis. The top 20 resulting drivers with predictive strength higher than 10% are available for display.

Some restrictions are enforced on the size of the data to improve performance and speed. If the data contains more than 250 fields, the least relevant fields are excluded before driver analysis. You can add the excluded fields back into the analysis through the UI as described above. If specified data contains more than 10,000 rows, it might be sampled down to approximately 10,000 rows for purpose of driver analysis. A warning is displayed in such instances: *To improve performance, due to the number of rows in the data source, the analysis is based on a representative sample of the entire data.* The results are expected to closely approximate results that would be obtained by using all the rows in the original data.

Two-way key drivers

Two-way drivers rely on modeling and ranking pairs of categorical predictors at one time.

Overview

Given a target field, IBM Cognos Analytics uses a statistical model for analysis of a pair of other data fields and estimates its strength in predicting the target values. Search over different predictor pairs is usually not exhaustive and also some high-ranking pairs can be filtered out from the final results. The goal is to provide an overview and variety of predictor pairs that improve upon predictive strength of a single predictor models that are displayed as one-way drivers. Therefore, the insights obtained from one-way drivers are expanded and the user obtains relevant information on the pairs of fields in the data. Both one way driver and two-way driver analysis results are available in the driver analysis and spiral charts. They

can be viewed separately by selecting a corresponding chart viewing option. Each displayed one-way or two-way driver can be expanded into a new visualization directly from the Driver analysis visualization in Explore.

Algorithms

Analysis for each two-way driver is based on a statistical model that includes the target and a pair of categorical predictors. The model is applied after data preparation and building all the one-way drivers. The first predictor in the pair is selected from the top 50 one-way drivers and the second is selected from the top 25 one-way drivers. This search strategy ensures that most of the top-ranking predictor pairs would be considered for modeling. The two-way ANOVA (analysis of variance) analysis is applied for numeric targets and Chi-square test of independence is applied for categorical targets with the chi-square adjustment for sparse data.

For each considered pair of fields, a hypothesis test on whether the pair has a significant impact on the target is performed. Only those pairs which pass the test and have sufficiently high predictive strength are selected as possible two-way drivers.

Details

The restriction of selection of data fields and data rows for one-way drivers apply to the two-way drivers as well. This is expected as potential predictor fields for two-way drivers are selected from the top one-way drivers based on their respective predictive strength. However, the model significance of one-way driver and the minimum predictive strength is not required for their entry into a two-way model. A resulting two-way driver must have its predictive strength higher than 10% and provide more than 10% relative improvement over the predictive strength for each of the contained one-way drivers. Relative improvement is computed as the percentage of the difference between 100% and predictive strength of the nested one-way driver. Resulting two-way drivers that satisfy these criteria are ranked by their predictive strength and the top 20 are made available for display.

Decision tree

Decision trees are more complex models than the one-way and two-way drivers. They extend the sequence as the combination models. The main difference is that decision trees support discovery of interaction among multiple predictors and therefore deeper insights than the drivers.

Overview

Given the target field, the algorithm searches across all other data fields and adds them to the model to improve its strength in predicting the target values. The search across different predictors is iterative; after the search adds one predictor, the search continuous to add the next predictor that improves the model the most. The goal is to find the best set of predictors and an optimal way of combining them so that an optimal model is computed. The insights that are obtained from decision trees are presented in the form of decision rules where combination of predictors and corresponding values provide a single prediction for the target value. Decision rules are ranked by strength so that you can easily find the rules that are the most relevant and interesting. Decision rules that are generated by the decision tree are mutually exclusive. The decision rules also provide a complete rule set such that a corresponding rule exists for any combination of the predictor values in the data. Also, available is the overall decision tree predictive strength that provides relative improvement over the basic model. The results are available through three different visualizations: sunburst, tree, and decision rules. They each have certain advantages by displaying the decision tree structure and the corresponding decision rules content. Overall decision tree predictive strength is also available in the driver analysis visualization.

Algorithms

The decision tree model is computed after data preparation and building all the one-way drivers. The first tree predictor is selected as the top one-way driver. Categories of the predictor are merged when the adverse impact on the predictive strength is smaller than a certain threshold. The next step is to find the

best predictor to split each tree node that consists of the merged categories. The process is continued until a stopping rule applies to a tree node. Possible options for stopping are that all categories for every candidate predictor are merged into a single node or that the number of nodes exceeds maximum number of nodes. Categories with fewer than a minimum number of rows are always merged with another category. This means that none of the nodes in the tree can contain fewer than the minimum number of rows. The same procedure is used for continuous and categorical targets, only the impurity function is different.

Details

Impurity functions

Impurity function values are used as the main criterion for splitting and merging potential tree nodes. Impurity function total for continuous trees is the sum of squares per node, while Gini impurity measure is used for the categorical targets. Gini impurity total is computed as a sum of squares of count proportions across all target categories per node that is subtracted from one and the results that are multiplied by the number of rows. Improvement in impurity function value is information gain.

When splitting each node IBM Cognos Analytics looks for a predictor field with a largest information gain computed as total impurity across all potential children nodes subtracted from the parent node impurity. Before Cognos Analytics selects the predictor, Cognos Analytics attempts to merge some of the potential children nodes that initially correspond to each predictor category. Information loss is computed by subtracting impurity of non-merged nodes from the impurity of merged nodes. As long as information loss is smaller than a threshold, the nodes are merged. This process helps to create relatively small trees that are easy to visualize and comprehend while still preserving the overall strength of the tree.

Stopping rules

Candidate nodes are always merged if they are based on fewer than 25 rows. If all categories of a predictor are merged, it cannot be used for splitting a certain node. When none of the predictors can split the specific node, the process stops for the node. The overall process of generating the tree stops when none of the nodes can be split or when the number of generated nodes exceeds 36.

Variable importance

Variable importance corresponds to a relative tree error reduction when the corresponding predictor is included in the tree. It is computed by comparing the errors of an initial tree and a restricted tree that is generated by the rest of the predictors in the initial tree. The error of the initial tree is subtracted from the error of the restricted tree and the result is divided by the error of the restricted tree. Variables with zero or negative importance are removed from the tree. The tree error is computed as the sum of squares for continuous targets and as classification error for categorical targets.

Predictive Strength

Predictive strength for tree with continuous target is computed similarly to key drivers. The contents of leaf nodes are considered. Variance contribution for each leaf node is added and divided by the overall variance for the data. This is relative error for the tree. It is subtracted from one to obtain predictive strength that is compatible with the R-squared measure that is used for key drivers.

For categorical targets, Cognos Analytics computes classification accuracy based on the classification error that is added from all the leaf nodes. Relative classification accuracy improvement over the basic model, also known as adjusted count R-square, is reported as the tree predictive strength. It is computed by subtracting the tree error from the basic model error and dividing the result by the basic model error. For example, the classification accuracy of the model can be as high as 95%, but if the majority class appears for 90% of the rows in the data, then the predictive strength of the tree is reported as 50% only. This is parallel to the continuous target case where the basic model is represented by the overall mean value. Predictive strength that is measured by R-squared is based on the tree relative improvement in reducing the overall variance.

Cognos Analytics displays only the trees that have predictive strength larger than 10%. A tree for continuous target is displayed in a driver analysis or spiral visualization if its predictive strength is

higher than the predictive strength of the strongest key driver. Otherwise, it is not displayed in these charts since the key drivers provide all the relevant insights already.

Predictive strength for a decision tree is computed that uses the same data that is used for generating the decision tree. This is known to introduce bias and provide optimistic estimates of the decision tree performance on a similar data from the same data source. Cognos Analytics reduced the discrepancy by tuning the algorithm so that overfitting the training data is minimized.

Insights in visualizations

Insights in visualizations provide analytic insights that can help users to detect and validate any important relationships and meaningful differences based on the data that is presented by the visualization.

Overview

Insights are controlled by and summarized in the **Insights** box available in every eligible visualization. When you turn on insights, the summary appears in the **Insights** box and related visualization elements are highlighted. Details are provided in the corresponding tooltip message. You can control each available insight separately.

Algorithms

The type of insights depends on the displayed data by the visualization. Available types of insights are **Average**, **Predictive strength**, **Meaningful Differences**, **Fit line**, and **Most Frequent**. Average provides the mean of the displayed summaries, and most frequent the category or category that appears most often in the data. The rest of the insights depend on more advanced analytics and statistical tests. The goal is to provide reliable information that you can use for an enhanced description of the viewed data and discovery of any relationships that are expected to be found in the population that is represented by this data.

Details

Insights analysis is always based on the same data rows that are used to create the summaries displayed in the visualization. This means that full data is used for insights unless any filtering is applied to the original data.

Some statistical tests and analytics that are used in insights require not only the data summaries that are displayed in the visualization, but also some additional summarizations. For example, test of meaningful differences across multiple categories of an explanatory field requires counts and variances for each category in addition to the displayed data. These additional summaries are obtained from a database together with the summaries that are needed for the visualization. All summaries are processed by the insights but only the required summaries are available in the visualization. Insights analysis is always based on the same data rows that are used to create the summaries displayed in the visualization.

Restrictions

If insights are not immediately available in a visualization, one of the following reasons might apply:

- The visualization type itself does not support insights.
- The data in the visualization may have been clipped.
- The combination of summarization level, field type, and field role of a selected field does not match the requirements for any of the available insights.

Supported visualization types for insights

The following visualization types support insights:

- Area
- Bar
- Bubble

- Column
- Heat map
- Hierarchy bubble
- Line
- Line and column
- Map
- Packed bubble
- Pie
- Point
- Radial
- Scatter
- Stacked bar
- Stacked column
- Tree map
- Word cloud

Small multiple extensions are supported for some insights including the Most Frequent and Meaningful Differences.

Summarization levels

Supported summarization levels are **Count**, **Average**, **Sum**, **Minimum**, and **Maximum**. Any other values such as **Count distinct** might prevent the insights from being suggested. Certain algorithms support only specific summarization levels. Changing the default summarization level to one of the supported values might potentially help with enabling insights.

Field types

Field types can be internally designated as continuous or categorical depending upon the values of the selected field.

Field type	Description
Categorical	A variable that can take on one of a limited, and usually fixed, number of possible values. A categorical variable assigns each individual or other unit of observation to a particular group or nominal category based on some qualitative property. For example, the country a person lives in.
Continuous	A variable that is used to describe numeric values, such as a range of 0-100 or 0.75 - 1.25. A continuous value can be an integer, real number, or date and time.

Field roles

IBM Cognos Analytics assigns a role to each of the field slots in a supported visualization. A field role might be designated as one of the following depending on the slot of the visualization.

Field role	Description
Response	A variable that is predicted and can also be referred to as the target or dependent variable. It is commonly on the Y-axis.
Explanatory	A variable which helps to explain changes in the response and is also referred to as the predictor or independent variable. It is commonly on the X-axis.
Group	A variable that is treated as explanatory or an optional grouped factor which helps to determine the number of models built in the algorithm. For example, this can correspond to the Color slot of a Column visualization.

Field role	Description
Weight	A variable which defines the optional regression weights, which are used to calculate the regression model. For example, this can correspond to the Size slot of a Bubble visualization.
Repeat	A variable which creates small multiples, with the visualization repeated once for each distinct value of the variable. For example, this can correspond to the Repeat (rows) slot of a pie visualization.
Points	A variable which defines the shaping of the data and data points used to calculate the model. For example, this can correspond to the Points slot of a Scatter visualization.

As a general example, in a bar visualization with the following slots, the role mappings in this visualization are defined as:

- Bars (y-axis), explanatory
- Length (x-axis), response
- Color, group

Insights in visualizations for counts

Insights for counts are available whenever count is displayed for each category of a single categorical field.

They are also available when count is displayed for each combination of categories of a pair of categorical fields in the visualization. In the latter case the pair may be two explanatory fields, as in the rows and columns of a heat map, or one explanatory and one repeat field, as in the bars and repeat (column) of a bar chart.

Insights for counts of combined categories of three categorical fields are supported for one explanatory and the combination of two repeat fields, as in the segments, repeat (column), and repeat (row) of a pie visualization.

Overview

Use such visualizations when you are interested in comparing the number of items in different categories, or combination of categories.

Algorithms

IBM Cognos Analytics reports the average count across all categories of the specified response field and applies statistical tests to detect categories where the counts are statistically most different from the average.

Visualizations with two or three categorical fields and counts for each combination of categories are treated differently. Cognos Analytics does not only compare the counts across categories but detects any relationship between the categorical fields. Cognos Analytics treats one field as the response field and the others as the explanatory field.

Cognos Analytics reports the most frequent category in visualizations with one categorical explanatory field, one or two categorical repeat fields, and a count response field.

Details

Single categorical field

The first test that is applied is the chi-square test of equal frequencies to establish whether any counts are available that are significantly different from the average. If the test result is significant, Cognos Analytics applies the influence chi-square test for each category separately. Cognos Analytics

computes the effect size for categories where the influence test is statistically significant and reports the categories with the largest effect size under the meaningful differences.

Restrictions

The following list describes the conditions that determine whether insights are suggested for this algorithm.

Response
Exactly 1

Summarization level = Count

Explanatory
N/A

Group
N/A

Weight
N/A

Points
N/A

Insight

Average

Meaningful differences

Two categorical fields

Cognos Analytics treats one categorical field as the response field and the other as the explanatory field. The original count field is used as input to the algorithms.

Chi-square test of independence with the adjustment for sparse data is used to establish whether a relationship exists between the response field and the explanatory field. If the test result is significant, Cognos Analytics computes the predictive strength for this model as adjusted count R-squared, with low-frequency categories filtered. The relationship is declared reliable and the predictive strength is reported if it is greater than 10%.

If the test result above is significant, all combinations of explanatory and response categories are analyzed further by applying the influence chi-square test for each combination. Combinations of explanatory and response categories where the influence test is significant are considered influential. Effect size is computed for each influential combination of categories and the combinations with the largest effect size are reported under meaningful differences.

If the roles of the two categorical fields are explanatory and repeat, the most frequent algorithm is applied. The counts are summed over each distinct category of the explanatory field. The largest sum is reported, together with the number of categories having that sum. Note that the repeat field is not used by this algorithm, but only triggers when the algorithm is applied.

Restrictions

The following lists describe the conditions that determine whether insights are suggested for this algorithm.

Response
Exactly 1

Summarization level = Count

Exactly 1

Summarization level = Count

Explanatory
Exactly 2

Categorical

Exactly 1

Categorical

Group

N/A

N/A

Weight

N/A

N/A

Points

N/A

N/A

Repeat

N/A

Exactly 1

Categorical

Insight

Predictive strength

Meaningful differences

Predictive strength

Meaningful differences

Most frequent

Three categorical fields

These algorithms are applied only when there is one explanatory field and two repeat fields. The combination of the two repeat fields is treated as if it were a single categorical field, where the categories are the pairs of categories from the two repeat fields.

The predictive strength is calculated exactly as in the two categorical fields case, using the paired repeat fields as the predictor of the explanatory field. The chi-square test of independence with the adjustment for sparse data is used to test significance of the relationship, and the adjusted count R-squared with low-frequency categories filtered is used to find the predictive strength.

The meaningful differences are calculated exactly as in the two categorical fields case, identifying combinations of the explanatory field and the paired repeat fields for which the count is unusual. The influence chi-square test is used to test significance for each combination, and the combinations with the largest effect size are reported.

The most frequent algorithm is applied exactly as in the two categorical fields case, summing the counts over each distinct category of the explanatory field. The largest sum is reported, together with the number of categories having that sum. Note that the repeat fields are not used by this algorithm, but only triggers when the algorithm is applied.

Response

Exactly 1

Summarization level = Count

Explanatory

Exactly 1

Categorical

Group
N/A

Weight
N/A

Points
N/A

Repeat
Exactly 2

Categorical

Insight

Predictive strength

Meaningful differences

Most frequent

Insights in visualizations for summaries by one or more explanatory fields

Insights for summaries are available when the summarization level is average, sum, minimum, or maximum for a continuous response field. Insights are computed and displayed at each category of a single categorical explanatory field, or each combination of categories of a pair of categorical explanatory fields in the visualization.

Overview

Use such visualizations when you are interested in comparing values of a response field across different categories, or across combinations of categories of explanatory fields.

Algorithms

If the summarization level is average, IBM Cognos Analytics detects any relationship between the response field and explanatory fields and computes predictive strength of the corresponding model. If differences of average values across explanatory categories are statistically significant, Cognos Analytics identifies the most different explanatory categories or combination of categories under the meaningful differences.

When the response summarization level is sum, Cognos Analytics computes the average sum across explanatory categories or combinations of categories. If the differences of sums across categories are statistically significant, Cognos Analytics identifies the most different explanatory categories or combinations of categories under the meaningful differences.

For all applicable charts, the average insight displays the mean summarized response value across all explanatory categories. When the summarization level for the response is average, the weighted mean is computed using the displayed value and the count for each explanatory category.

Details

Average by single explanatory field

When the summarization level for the response field is average and a single categorical explanatory field is available, Cognos Analytics applies one-way ANOVA analysis. Cognos Analytics uses F statistic to test whether average values across explanatory categories are equal. If any differences are significant, Cognos Analytics computes the adjusted R-squared as predictive strength of the relationship between the response field and the explanatory field. Reliable relationship and its predictive strength are reported to the user if the predictive strength exceeds 10%.

If the difference among averages is significant, Cognos Analytics conducts an influence t test to detect the categories that are the most different from the overall mean. This involves computing

standard error for each category average and comparing the average with the overall mean by using the t test statistic. For categories with significant differences, Cognos Analytics also computes the corresponding effect size and reports the categories with the largest effect size under meaningful differences.

Restrictions

The following list describes the conditions that determine whether insights are suggested for this algorithm.

Response
Exactly 1

Summarization level = Average

Continuous

Explanatory
Exactly 1

Categorical

Group
N/A

Weight
Optional

Any

Points
N/A

Insight

Average

Predictive strength

Meaningful differences

Average by two explanatory fields

For charts where the summarization level for the response field is average and two categorical explanatory fields are available, Cognos Analytics applies two-way ANOVA analysis. Cognos Analytics uses F statistic to test whether average values across explanatory category combinations are equal. If differences are significant, Cognos Analytics computes adjusted R-squared as predictive strength of the relationship between the response field and the two explanatory fields. Cognos Analytics also computes the adjusted R-squared for one-way models that include a single explanatory field each. If the predictive strength of a two-way model is larger than 10% and its relative predictive strength improvement over corresponding one-way models is more than 10%, Cognos Analytics displays the predictive strength of the two-way model and reports reliable relationship between the response field and the two explanatory fields. Otherwise, if the maximum predictive strength of one-way models exceeds 10%, Cognos Analytics reports reliable relationship between the response and the corresponding single explanatory field together with its predictive strength. If maximum predictive strength of one-way models does not exceed 10%, Cognos Analytics reports no relationship between the response and explanatory fields.

When the difference among averages across all category combinations is significant, Cognos Analytics also conducts an influence t test to detect the category combinations that are the most different from the overall mean. This test is similar to the test that is used for a single explanatory field. The main difference is that instead of considering categories of a single explanatory field, Cognos Analytics considers category combinations from the two explanatory fields. Category combinations with the largest effect size are reported under meaningful differences.

Restrictions

The following list describes the conditions that determine whether insights are suggested for this algorithm.

Response
Exactly 1

Summarization level = Average

Continuous

Explanatory
Exactly 2

Categorical

Group

Optional (treat as Explanatory)

Categorical

Weight
Optional

Points
N/A

Insight

Average

Predictive strength

Meaningful differences

Sum by one or two explanatory fields

For charts where the summarization level for the response field is sum and one or two categorical explanatory fields are available, Cognos Analytics applies the sum comparison test. This test detects if any of the sums are different from the average sum value across all explanatory categories or combinations of categories. If this test is significant, Cognos Analytics proceeds by conducting the sum influence test that compares sum for each category or combinations of categories with the average sum. For every significant test, Cognos Analytics also computes corresponding effect size. Categories or combinations of categories with the largest effect sizes are reported under meaningful differences.

Restrictions

The following list describes the conditions that determine whether insights are suggested for this algorithm.

Response
Exactly 1

Summarization level = Sum

Continuous

Explanatory
1 or 2

Categorical

Group

Optional (treat as Explanatory)

Categorical

Weight
N/A

Points
N/A

Insight

Average

Meaningful differences

Minimum or maximum by one or two explanatory fields

For summarization levels of minimum or maximum, only the average insight is available. It is computed as the average value of the response minimum or maximum across all explanatory categories or combinations of categories.

Restrictions

The following list describes the conditions that determine whether insights are suggested for this algorithm.

Response
Exactly 1

Summarization level = Min or Max

Continuous

Explanatory
1 or 2

Categorical

Group

N/A

Weight
N/A

Points
N/A

Insight

Average

Insights in visualizations for two continuous fields

Insights for two continuous fields are available when a visualization involves two continuous fields and an optional categorical group or points field.

Overview

Use visualizations such as scatter plot for the two continuous fields, possibly sliced by categories of the group field. The main goal is to detect any relationship between the continuous field and include the categorical group field as well. The results contain predictive strength of discovered relationship, the relationship description provided by fit lines, and any points with large discrepancy from the fit lines as meaningful differences.

Algorithms

IBM Cognos Analytics computes multiple regression models that involve one of the continuous fields as the response, and the other continuous fields as an explanatory field. The optional categorical group field is used as a model factor. In addition to additive model contributions corresponding to the explanatory field, Cognos Analytics considers the square of the explanatory field and any interaction terms that

include a factor. A regression model that provides an optimal fit for the data is selected from a number of possible models. The corresponding fit line is derived from a linear or quadratic model. In the case where an optional categorical group field has been supplied, it can produce a different line or quadratic curve for each category of the factor. A factor with up to three categories is currently considered in order not to overload the visualization.

Each point in a visualization represents a number of rows in the data and it is defined by the **Points** field. Corresponding row counts that are based on the response field define frequency weights that are used for building the regression models. Regression weights are used independently of frequency weight when Cognos Analytics computes the regression models.

Details

Two continuous fields

When Cognos Analytics applies multiple linear regression for two continuous fields, one is chosen as the response and the other as an explanatory field in the model. Cognos Analytics considers both linear and quadratic model terms. If the quadratic model is significant based on the F test and its relative predictive strength improvement is more than 10% over the linear model, Cognos Analytics reports its predictive strength and displays the quadratic curve based on the computed model. This curve displays predicted values of the response based on the corresponding values of the explanatory field. Otherwise, linear predictor model is considered. If it is significant and its predictive strength is larger than 10%, Cognos Analytics reports its predictive strength and displays a line representing the predicted values of the response field based on the corresponding explanatory values. If the linear model does not qualify, the mean is reported as the fit line and no relationship is reported between the two continuous fields.

When linear or quadratic relationship is detected, Cognos Analytics also inspects the differences between predicted and observed values of the response field. These differences are called residuals and Cognos Analytics conducts studentized residuals test to detect outliers. Points with large departure from the discovered relationships are displayed under meaningful differences in the corresponding chart.

Restrictions

The following list describes the conditions that determine whether insights are suggested for this algorithm.

Response
Exactly 1

Summarization level = any

Continuous

Explanatory
Exactly 1

Continuous

Group

N/A

Weight
Optional

Continuous

Points
Optional

Any

Insight

Predictive strength

Fit line

Meaningful differences

Categorical group field

When a categorical group field is specified in addition to two continuous fields, it is used as a factor in the multiple linear regression where one of the two continuous fields is chosen as the response field and the other as an explanatory field. Cognos Analytics considers linear and quadratic model terms for continuous explanatory combined with contributions from the factor. If the quadratic model or linear model that include the factor is significant based on the F test and its relative predictive strength improvement is more than 10% over the linear model with continuous explanatory only, Cognos Analytics generates four extra models. These models include all possible interactions of continuous explanatory and factor. A model with the maximum adjusted R-squared that is also significant is selected as the final model. It is used to create a fit line for each category of the categorical predictor. Otherwise, the linear model with continuous explanatory is tested for significance and reported if its predictive strength is greater than 10%. If the linear model does not qualify, no reliable relationship among fields is established and the overall mean is reported as the fit line.

When a reliable relationship is detected, Cognos Analytics also checks for difference between the predicted and observed values of the response field. Cognos Analytics conducts studentized residuals test to detect outliers and display them under meaningful differences in the corresponding chart.

Restrictions

The following list describes the conditions that determine whether insights are suggested for this algorithm.

Response

Exactly 1

Summarization level = any

Continuous

Explanatory

Exactly 1

Continuous

Group

Exactly 1

Categorical

Weight

Optional

Any

Points

Optional

Any

Insight

Predictive strength

Fit line

Meaningful differences

Regression weights field

An optional continuous field can be used to specify regression weights for the model. Regression weight for an available value corresponds to influence of the observation on the computed model parameters.

Natural language details

Natural language details is a text feature that augments displayed visualizations with more summaries. Details provide insights that are obtained from an appropriate data analysis that is relevant to you.

Overview

This feature is available for visualizations that are created in an exploration and the textual details are displayed in the matching details pane. With this information, you can obtain information that is the most relevant for the viewed data in a natural language format. More summaries and details that are not available in the visualization are also shown.

Algorithms

Basic details provide simple summaries of the data that is not readily seen in the displayed visualization. While this information can be obtained by you through specifying other related visualizations, such exploration steps become unnecessary because related summaries are made available as textual details.

Details that are based on the insights in visualizations provide text description of the details that can be obtained through the Insights box in the displayed visualization or related visualizations. They provide more clarity to the displayed insights and also allow adding further insights that are not available in the displayed visualization.

Details

Details are based on the fields that are displayed in the matching visualization. Related analysis can draw on additional summaries, but does not include any fields not specified in the visualization. Summaries and details are converted into translatable text by using templates rather than a full natural language generation facility. This results in some language repetitiveness in the rendered text, but it does not diminish the amount or quality of the displayed information.

Basic natural language details

Basic natural language details provide extra summaries of the data that is displayed in the visualization or highlight the available information with extra details.

Overview

This information provides a more rounded view of the data while still being relevant to displayed summaries that you are considering. Changing the summarization level that is specified for the response field also changes some or all the basic details since the focus of the visualization was changed. Additional information is provided in the context of the main task.

Algorithms

Computed basic details depend on the summarization level that is specified for the response field in the corresponding visualization. Possible response summarization levels are count distinct and count for any field and sum, average, minimum, and maximum for numerical fields. While count-based details are used for most of the response summarization levels, extra matching summaries are provided with response summarization levels sum, average, minimum, and maximum. The same count-based summaries are generated for both count and count distinct response summarization levels.

Details

Overall count

Overall count is displayed for the response and any explanatory field in the visualization. The count does not include missing values of the response and it is computed except if summarization level for

response is sum or average. All categorical fields that are specified in the visualization are treated as explanatory fields for details purposes.

Count for explanatory fields

Count is also computed for each category of the displayed explanatory field except if summarization level for response is sum. The algorithm selects and reports top categories, corresponding counts, and count percentages relative to the overall count for the explanatory field. This procedure is applied to each explanatory field in the chart.

Sum

When the displayed summarization level is sum for a numeric response field, IBM Cognos Analytics summarizes the corresponding total sum for the response field. If the visualization contains multiple explanatory categorical fields, the sum is computed for each category and each explanatory field. The algorithm selects and reports the top categories, corresponding sums, and sum percentages relative to the total sum for each explanatory field.

Average

When the displayed summarization level is average for a numeric response field, Cognos Analytics summarizes the corresponding overall average for the response field.

Range

Range summaries are given by the minimum and the maximum when summarization level sum or average are computed for a numeric response field across combination categories of specified explanatory fields. If only a single explanatory field is specified, the categories where minimum and maximum occur are displayed as well.

Minimum and maximum

When the displayed summarization level is a minimum or maximum for a numeric response field, Cognos Analytics displays the corresponding minimum or maximum value across all categories of an explanatory categorical field. If multiple explanatory categorical fields are specified, minimum or maximum is computed across all possible category combinations.

Basic natural language details based on insights

Details that are based on insights provide text description of the insights that can be obtained through the Insights dialog box in the visualization.

Overview

Details that are based on insights also provide insights from related visualizations that are informative and easily understood in the context of the current chart. This allows for more comprehensive details that are related to the current visualization.

Algorithms

Details that are based on insights directly use the full scope of computations and statistical tests that are supported by the insights. Details also obtain the results for related visualizations and compile them together in a meaningful message. Details also provide some additional analysis based on the insights output by producing further details suitable for textual output.

Details

Single explanatory field

Given a response field and a categorical explanatory field, details use insights to detect relationship between the response and the explanatory field. Analysis is applied depending on whether the response summarization level is average or count and the predictive strength is reported if a relevant relationship is discovered.

If the explanatory field is numeric and insights generates a fit line, then IBM Cognos Analytics reports positive or negative slope for linear fit and point out whether a quadratic relationship is detected. If the relationship is quadratic, Cognos Analytics also reports the extremum point. Cognos Analytics computes the minimum or maximum value of the response and the explanatory value where the response extremum occurs.

Two explanatory fields

When two explanatory fields are available, Cognos Analytics detects relationship between the response and both explanatory fields and the relationship between the response and each explanatory field separately. If the predictive strength for the relationship with both explanatory provides relative improvement greater than 10% over each separate relationship, Cognos Analytics reports that the response is affected by both explanatory fields. Otherwise, it is affected by a single explanatory field, or by each explanatory field separately, but not together.

Meaningful differences

Details reports all meaningful differences that are discovered by the insights for the used data when response summarization level is count, average, or sum. They are reported for categories of each explanatory field separately as well as for the combinations of categories for two explanatory fields. Indications whether values are unusually high or low are also provided.

Decision trees

Details report predictive strength for a decision tree and a list of predictor fields that are used for splitting the tree nodes. A predictor field with the highest variable importance and its ratio improvement of variable importance over other fields in the decision trees is reported if the ratio is greater than two. A similar detail can be displayed for a field with the lowest variable importance.

Natural language details for time series

Details for time series provide text insights based on the analysis of the time series data and corresponding forecasting models.

Overview

IBM Cognos Analytics reports details for time series for a visualization that is created in an exploration whenever the visualization data contains a single time series and a forecasting model is computed. If the data is suitable, time series insights are generated even if the **Forecasting** dialog box is not present on the visualization. When the **Forecasting** dialog box is present, it produces the same default model upon activation as the time series insights. Time series points are automatically sorted in chronological order for purpose of the insights detection, but unlike in the forecasting feature, the time points displayed in the visualization are not sorted.

Algorithms

Details for time series are based on an exponential smoothing model for the observed time series data. Observed time series values and computed model components are used to create insights for the time series: unusual values, seasonal effects, and trend insight. Each type of insight depends on a different combination of data and corresponding exponential smoothing model components.

Details

Unusual values

An exponential smoothing model provides a predicted value for each observed time point. A predicted value at a time point is the one-step ahead forecast at the previous time point. A confidence interval for each predicted value is computed that uses the corresponding predicted value variance that depends on the model. An observed time series value that is found outside of the confidence interval for corresponding predicted value based on the model is considered to be an *unusual value*.

Unusual values are detected based on the selected exponential smoothing model for the time series. The confidence level that is used for computing the prediction confidence intervals is 99.74%. Up to five unusual values are reported by listing the corresponding time points. Cognos Analytics does not list the points in chronological order but rather in decreasing order of distance from the confidence interval. More unusual points are listed first. Unusual values are specified as unusually high or unusually low when possible.

An unusual value that is detected at the last time point is reported separately. This might indicate that data is incomplete. For example, summarized value for the last month might reflect daily data halfway through the month only.

Seasonal effects

The *seasonal effects* insight reveals the seasonal length for a time series that is identified by the model. Seasonal length corresponds to a fixed duration of a seasonal pattern established in the time series. For example, average temperature variation across 12 months establishes an annual pattern. This insight also provides the strength of the seasonal effects and reports periods with the largest and the smallest seasonal values.

The seasonal length is obtained from the selected model. It is derived from the seasonal period and date or time interval that is reported in the forecasting statistical details. A seasonal model is selected only if it provides a fit superior to all non-seasonal models. The seasonal period for the selected seasonal model is obtained by comparing models with multiple candidate seasonal periods.

Seasonal effects are reported as weak, moderate, or strong depending on the computed strength value. Strength of seasonal effects is computed as a reduction in model error by the seasonal model compared to matching non-seasonal model and divided by the non-seasonal model error. This is different from the seasonality strength reported in the forecasting statistical details where the difference in accuracy between the two models is reported.

The largest and the smallest seasonal values are computed based on the underlying seasonal model component averages across all seasonal patterns in the time series. Corresponding periods are reported if the average values are consistently the largest, or the smallest, over majority of the seasonal patterns.

Trend

The *trend* insight reports an overall positive or negative direction of the time series values when present. It also reports the strength of the trend.

Both level and trend components are extracted from the corresponding exponential smoothing model. Only the level component is used if the model has no trend component. This defines a trend curve for the time series data. Kendall's tau measure of association and corresponding statistical test are then computed for the trend curve. They detect an overall positive or negative direction of the time series values. Different tau value ranges define reported strength degree for the trend: weak, moderate, or strong.

For more information on exponential smoothing models, see "Forecasting models" on page 59.

Relationships

Relationships visualizations in an exploration are displayed initially when you specify data for exploration.

Overview

IBM Cognos Analytics provides a quick overview of relationships among pairs of fields that focus on a single field of interest. Visualization comprises multiple tabs, each for a different field of interest. This information is very useful in orienting you regarding a multitude of relevant relationships available in data to be explored further as needed.

Algorithms

While the initial field of interest is determined based on semantic data analysis, you can specify a different field of interest. Each tab provides a network graph with fields as nodes and links between pairs

of nodes that represents the relative strength of the relationship between nodes. While links from the field of interest dominate the graph, other related pairs of fields with strong relationships are displayed as well. You can adjust a slider to view larger or smaller number of nodes in the network.

Details

Data for analysis

Relationships use non-summarized data to compute strength of relationship among all pairs of fields considered. To standardize the measure of relationship strength and make it comparable across all pairs of fields, all numeric fields are binned as the first step. All fields in the data are treated as categorical. The binning that is applied is equal frequency binning generating five bins. More details are available in the section on data preparation for numeric fields.

Relationship strength

Data for each pair of categorical fields is first tabulated for all combination of field categories that are found in the data. Based on the tabulated data, IBM Cognos Analytics applies the chi-square test of independence to assess whether the fields are independent. If the independence departure is significant, Cognos Analytics computes the effect size based on the chi-square statistic. This is Cramer's V that is widely used as a measure of association between two categorical fields. The values of this measure are in the range 0 - 1 and Cognos Analytics reports the relationship strength value that is expressed as a percentage. The relationships with strength less than 10% are not reported as they are considered too weak to be of practical value.

Performance limitations

Computing relationship strength between all pairs of fields in the data set is prohibitive for large data. Cognos Analytics limits the number of processed fields to 100 to be able to provide a quick answer. However, these fields are selected by another process and the possible loss of relevant relationships is minimized. If the data contain more than 10,000 rows, Cognos Analytics obtains a random sample of this size for performance reasons. This data size ensures minimal loss in accuracy of the relationship strength estimate.

Chapter 6. Assistant

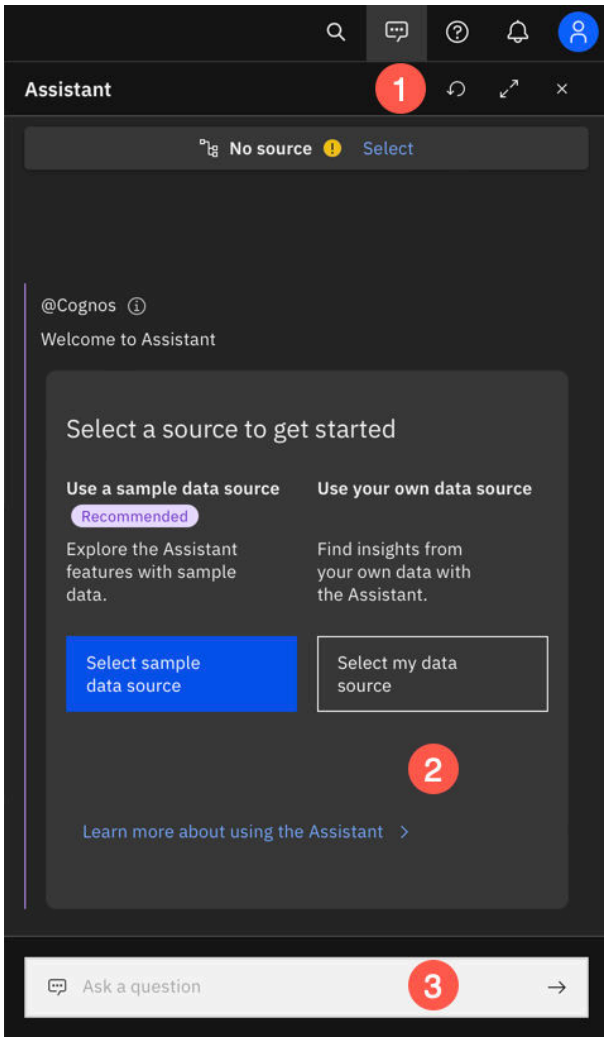
Assistant panel

IBM Cognos Analytics includes an embedded assistant that supports natural language text-based input to help you gain quick insights into your data and simplify your analytics. In just a few steps, you can access key data sources, create visualizations, and drag them onto your exploration or dashboard canvas. Text input is supported in English only.

Important: To use the Assistant, you must have permission to access this feature. Administrators can enable the Use Assistant capability at either the user level or source level, allowing you to use the Assistant.

Click the **Assistant** icon from any view in the main toolbar to open the assistant panel.

The Assistant panel is made up of the following sections: toolbar, output, and input field.



1. Toolbar

You can run the following actions in the toolbar:

- Clear chat and reset the context of the Assistant to the default source
- Expand the Assistant to full screen

- Close the Assistant

2. Output

The output is displayed in a scrollable area that includes responses and also provides guidance to get you started with using the Assistant.

Before you use the Assistant, you need to set the context for the Assistant by connecting it to a data source. You can choose a sample data source that is optimized for the Assistant or you can use your own data source. Click **Change** to change the data source that the Assistant is connected to.

To learn how to use the Assistant, click **Learn more about using the Assistant** and view related topics in the Learn pane.

3. Input field

The input field, labeled as **Ask a question** by default, is where you can enter text-based conversational input. As you enter text, a type-ahead feature offers suggestions on what you can ask. This is a great way to formulate proper questions and it saves typing.

Assistant commands

The Assistant is a powerful feature that allows you to ask questions in natural language. This topic demonstrates some of the commands that you can use with the Assistant.

Ask questions using natural language. For example, `show Profit` is synonymous with `list Profit`, `what is Profit?`, `tell me about Profit`, and `Profit`.

Here are some commonly used questions to get you started:

help
Displays overview and general usage information.

show data
Lists all supported data sources that you have access to in **My content** and **Team content**. Supported data sources include:

- Uploaded files (csv, tsv, xls, xlsx, and zip)
- Data modules
- Data set
- OLAP cubes
- Framework Manager packages

Note: See *Enriching packages* in the *Managing* document.

When results exceed the number of displayed items, click **More** to view up to 100 data sources. Scroll to the bottom of the list and click **Less** to collapse the list.

show source <source-name>

Displays a list of relevant fields and details for the specified data source. The toolbar displays the <source-name> that is in context. By default, Cognos Analytics uses the active source in your dashboard or exploration panel. From the results, you can click field names to obtain more information for those particular fields. Clicking field names is equivalent to asking `show column <column-name>`.

When results exceed the number of displayed items, click **More** to view up to 100 data sources. Scroll to the bottom of the list and click **Less** to collapse the list.

show column <column-name>
For the specified column, information and related fields are displayed. Clicking the related fields is equivalent to asking show chart <column1> and <column2>. If the specified column is determined to have influencers, you can enter what influences <column-name> to see its list of influencers.

what influences column <column-name>
Displays a list of fields that influences the results in the specified column.

show chart <column1> and <column2>

Displays visualizations that show the relationship between <column1> and <column2>. Scroll through the visualizations by clicking the left and right arrows. Each visualization includes an information icon in the upper-right corner. Hover over the icon to see descriptions about the underlying data. You can optionally enter more columns, but excessive columns can result in less effective visualizations.

Clicking Show related visualizations returns visualizations based on influential and related fields.

Clicking Create dashboard from the charts creates a new dashboard based on the most recently generated charts. Typing Create related dashboard produces the same resulting dashboard. If the charts contain top or bottom aggregations, these modifiers are applied to the generated dashboard.

Applying aggregations and filters can help to add focus and create more compelling visualizations. Common aggregations include total, average, count, maximum/minimum, top/bottom, best/worst, and so on. Here are a few aggregation examples:

show top <num> <column1> by <column2>

Displays the top values from <column1> based on the context of <column2>. For example, show top 5 Sales by Region. If <num> is not specified, a default value of 10 is used.

<column1> is an aggregated or non-aggregated measure, while <column2> is a categorical column.

show average <column-name>

Displays the average for all values found in <column-name>.

how many <column-name>

If <column-name> is a category, the number of distinct items is returned. If <column-name> is a measure, the sum total is returned.

show maximum <column-name>

Displays the highest value found in <column-name>.

show minimum <column-name>

Displays the lowest value found in <column-name>.

show total <column-name>

Displays the sum total for all values found in <column-name>.

You can add filters for geographical strings (such as Country or State) or temporal strings (such as Month or Year).

Aggregations and filters can be combined to produce more granular results. Here are some examples, based on sample data:

- show Education by Income where Income is less than 1000
- show Education by Income where Income > 100K

  Optionally use K (to denote thousands) or M (to denote millions).
- show Revenue in 2017 and 2018
- show Income by Month for New York City
- what are the top 5 States by average Inventory, excluding California

Filtered visualizations include a filter icon ( ), located in the upper-right corner of the chart. Hover over the icon to display the applied filtering.

create dashboard

Generates a new dashboard based on the currently selected data source. You can modify the visualizations, tabs, order, etc. and save your new dashboard. By default, the generated dashboard will include advanced analytics and predictive charts.

Automatically generating dashboards for larger data sources may result in performance issues. To circumvent this, you can enter `create simple dashboard` to generate a basic dashboard. Then you can modify the dashboard by replacing charts with more sophisticated visualizations, such as driver analysis or spiral charts.