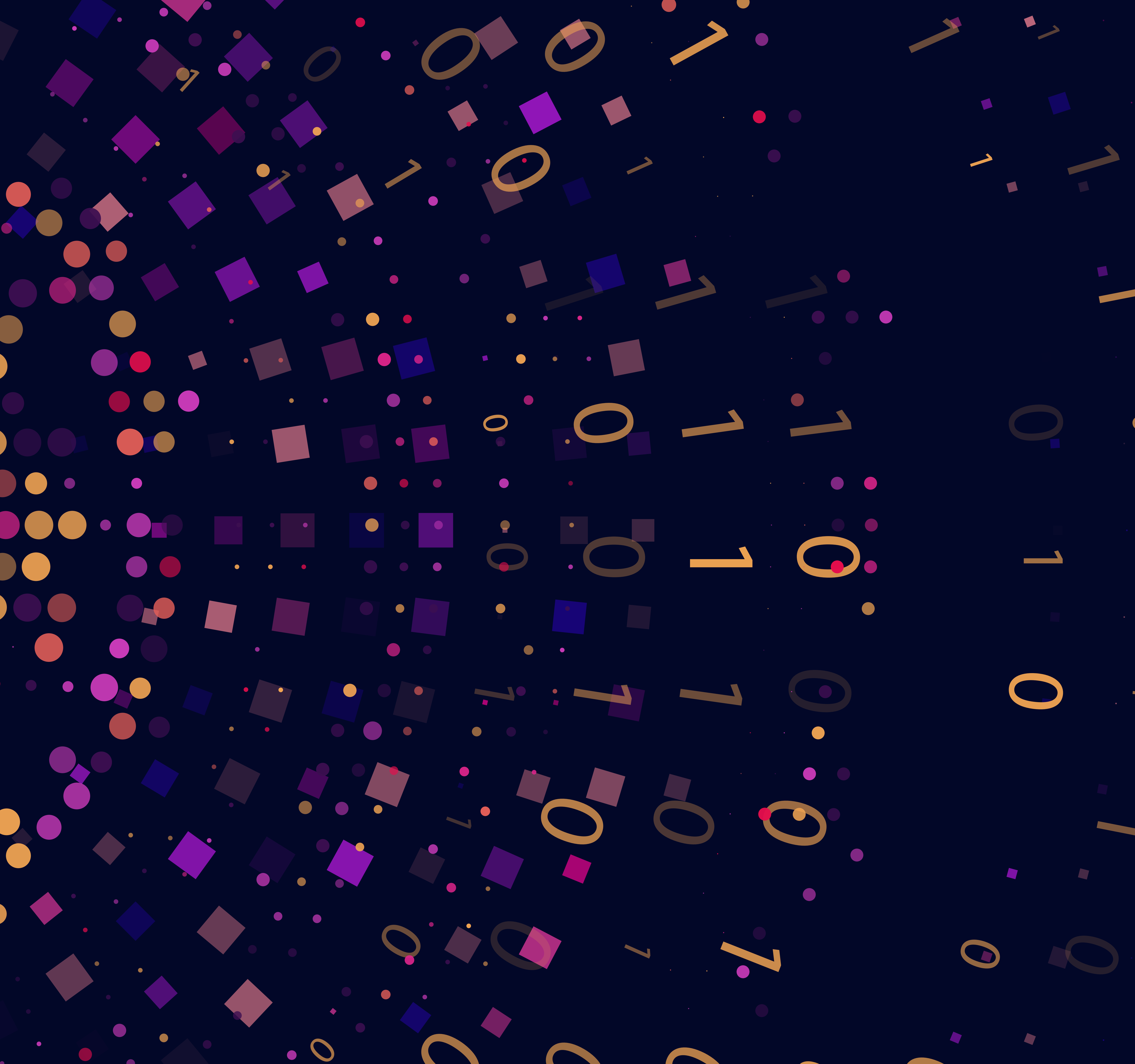


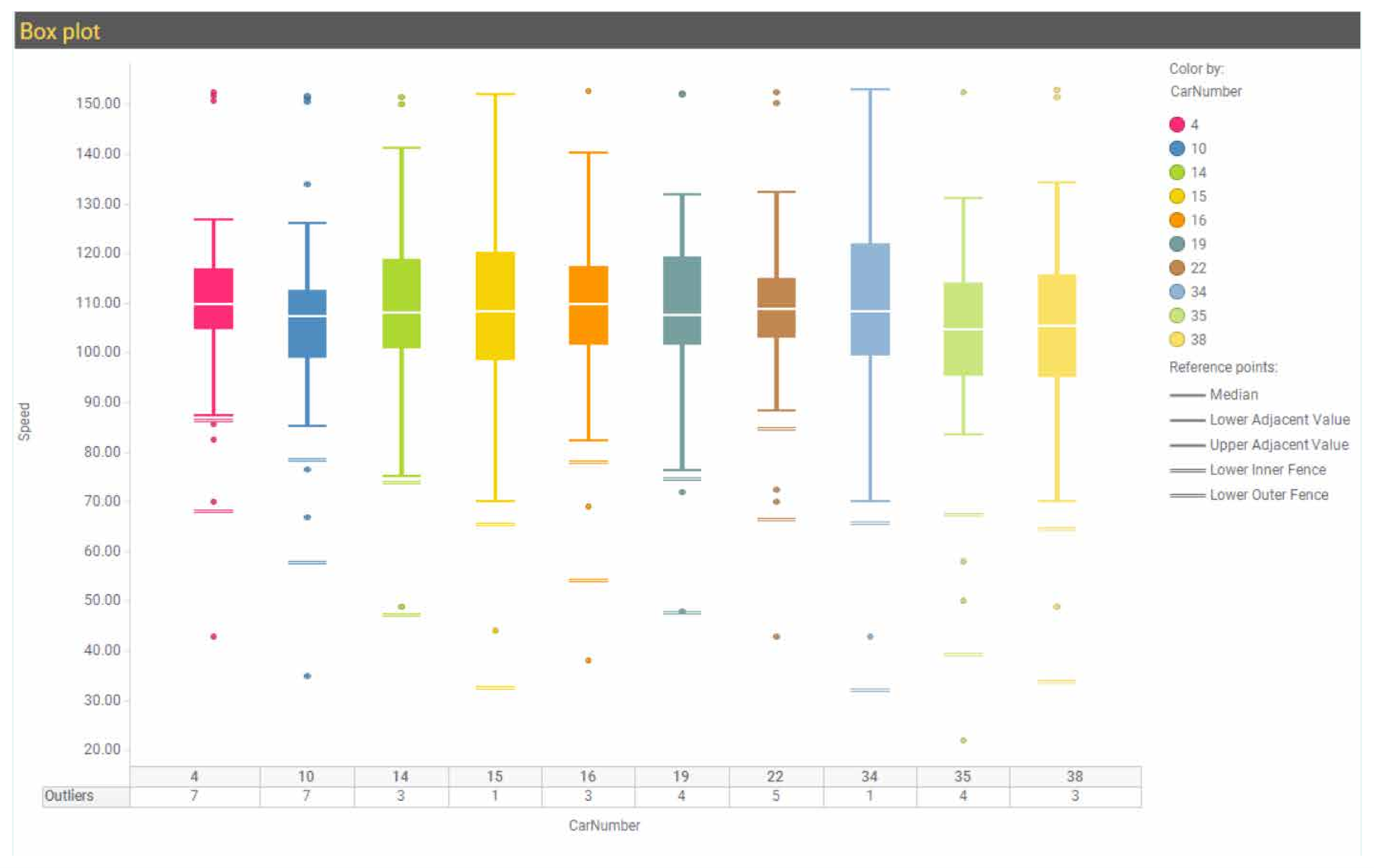
Top 10 Tips for Detecting Outliers and Anomalies

Identifying outliers or anomalies in your data can help you find and address potential business problems, discover opportunities for improved performance and greater revenue, and even spot risks in advance. Understanding how to detect and label outliers is therefore essential to the success of any business.

In this ebook, we walk through 10 ways you can use TIBCO's analytics and data science solutions to smartly find and mark outliers.

An outlier is classified, mathematically, as any observation far removed from most of the data. But in practice, outliers could come from incorrect or inefficient data gathering, industrial machine malfunctions, fraudulent retail transactions, and a variety of other processes.

Once outliers are detected, it becomes essential to isolate them and apply corrective treatment, implementing changes to address potential issues or to seize opportunities for improvement.



	4	10	14	15	16	19	22	34	35	38
Outliers	7	7	3	1	3	4	5	1	4	3

1. Use a box plot

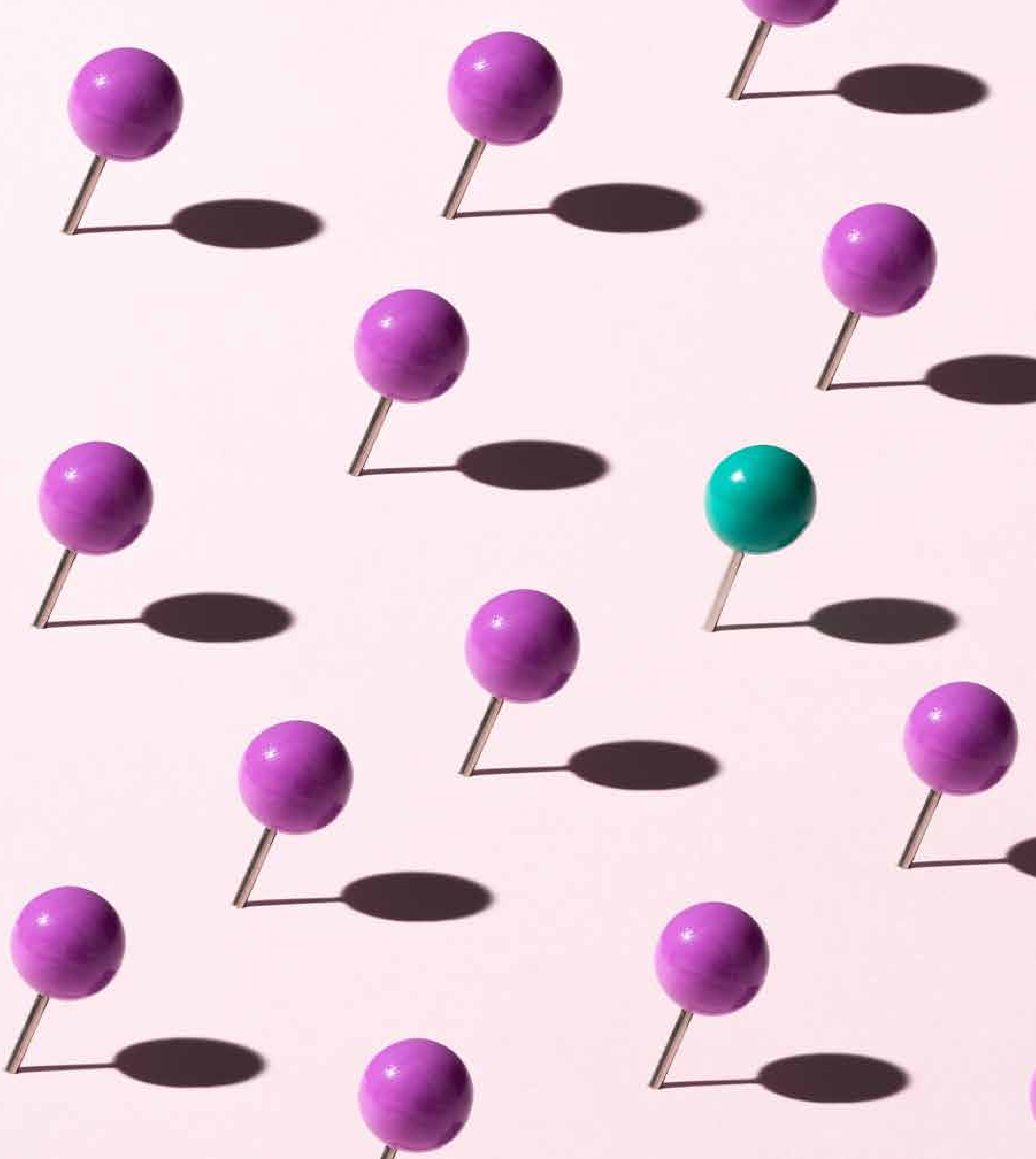
Box plots are one way to detect outliers. Through visual representation, all the data with key statistical measures are shown. Box and whisker plots show the relationship between a numerical y-variable and a grouping x-variable by using the five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

In the graph to the left, you can see how TIBCO's analytics solution provides lower adjacent value (LAV) and upper adjacent value (UAV). The interquartile range (IQR) is any point falling inside the LAV and UAV. Any point falling outside of LAV and UAV is marked as an outlier. You can then dig deeper into those points identified as outliers for additional information about how they are different compared to all other data points in the plot.

2. Configure other plots

The box plot is just one of many plots that are commonly used to identify outliers. Any of the following plots could help your business spot outliers:

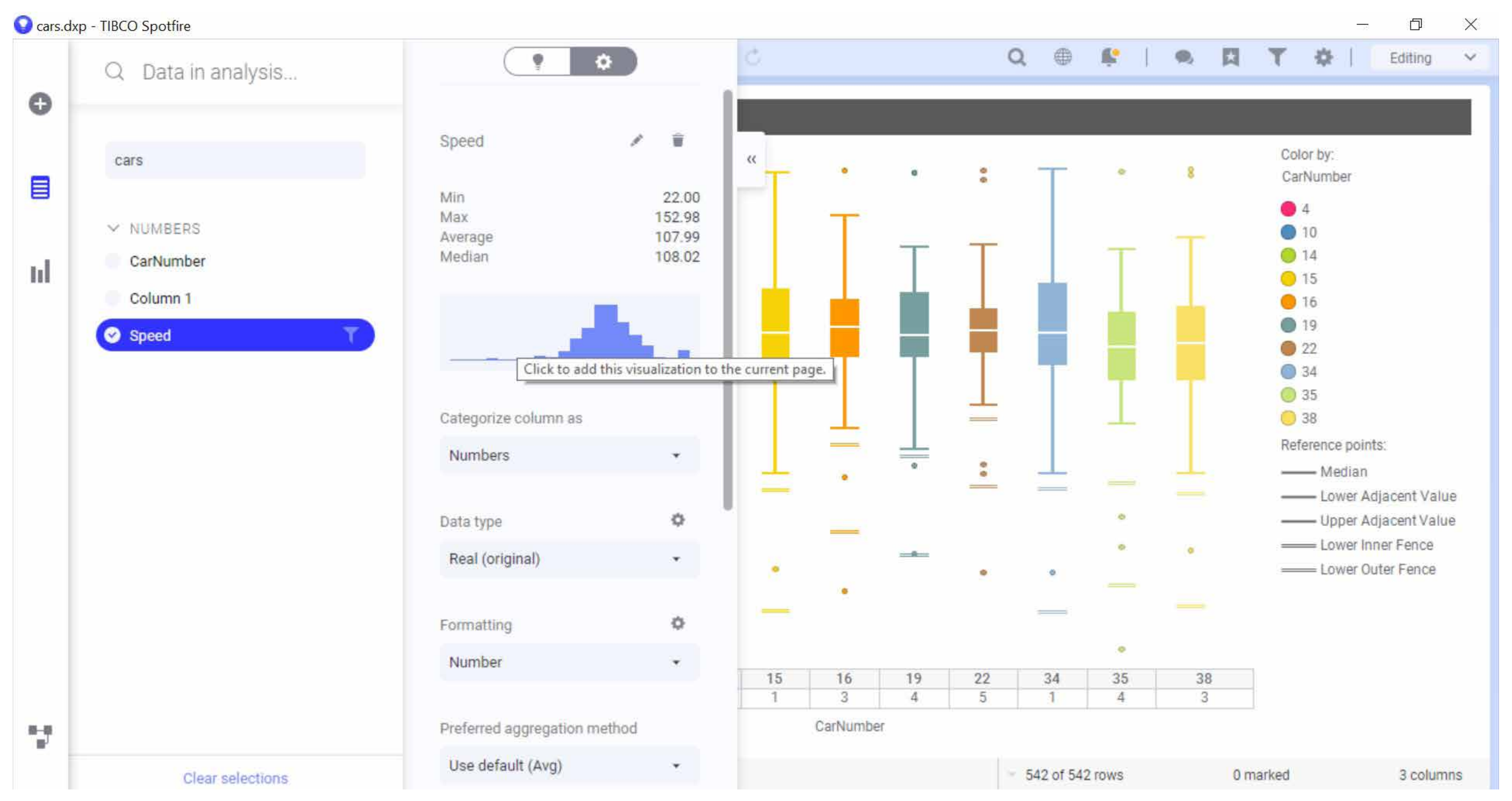
- Bar chart in a histogram configuration to identify univariate outliers
- Scatter plot in QQ plot configuration to identify bivariate outliers in distributions
- Combination plot in Pareto chart configuration to identify outliers based on cumulative value
- Parallel coordinate plot (PCP) multivariate analysis for outlier detection



3. Create a data panel histogram

Now let's look at the column overview data panel for in-memory as well as in-database (in-DB) data, which shows a histogram of distribution for numerical columns. The overview contains measures such as standard deviance and mean, which when inserted as lines onto the histogram smartly identify outliers for distributions.

You can also insert custom lines for isolating outliers in multimodal data. Considering the data from a standard normal distribution, about five percent falls beyond two standard deviations and thus will be picked up as outliers by common statistical tests. But this is just the nature of the distribution that the points follow. For such cases, TIBCO's analytics solution allows you the flexibility to insert lines from custom expressions without depending entirely on predefined methods of outlier detection.



cars.dxp - TIBCO Spotfire

Data in analysis...

cars

NUMBERS
- CarNumber
- Column 1
- Speed

Clear selections

Speed

Min 22.00
Max 152.98
Average 107.99
Median 108.02

Click to add this visualization to the current page.

Categorize column as: Numbers

Data type: Real (original)

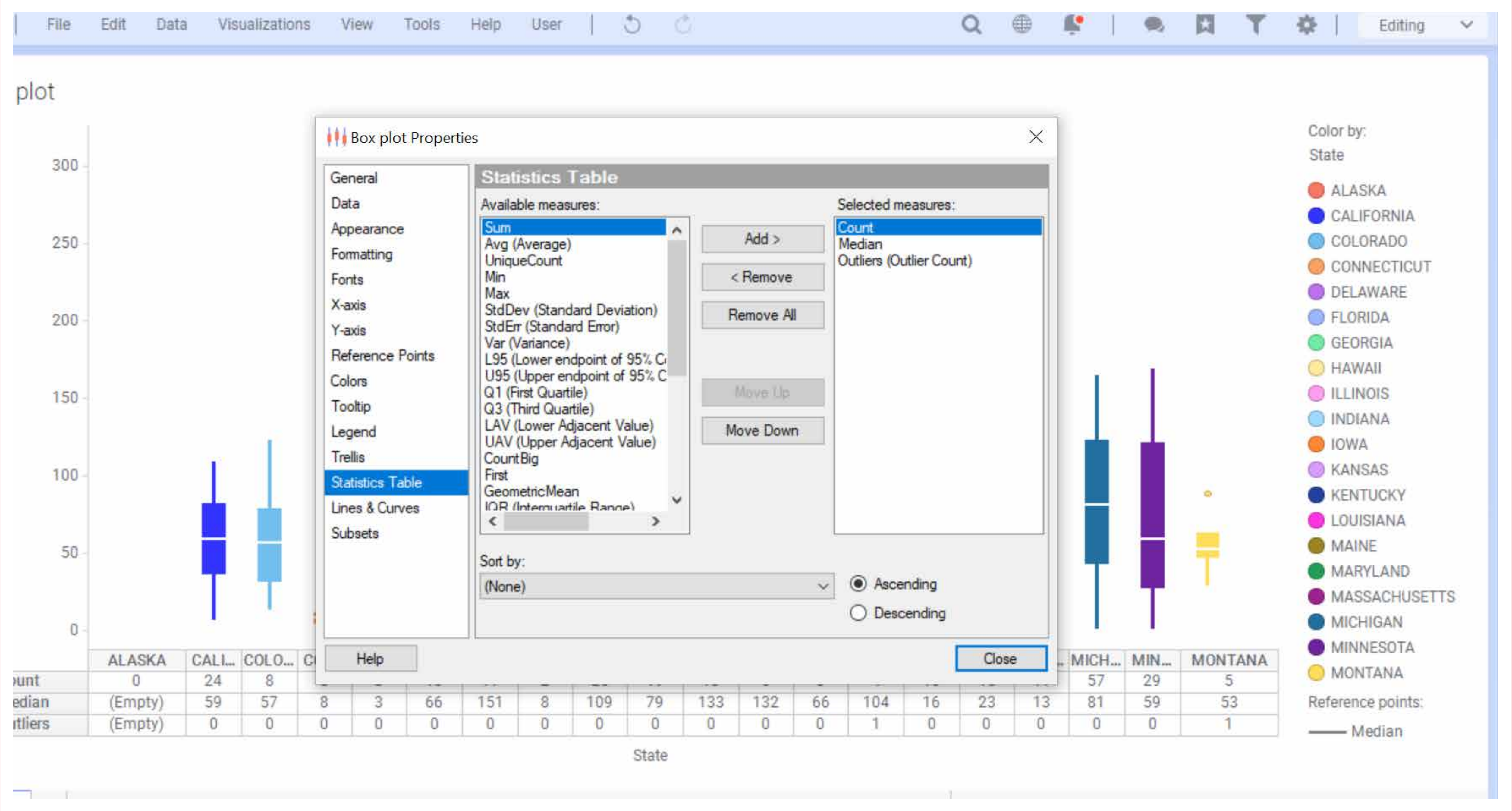
Formatting: Number

Preferred aggregation method: Use default (Avg)

542 of 542 rows | 0 marked | 3 columns

4. Select column aggregation functions

With TIBCO, you can also aggregate the y-variables in visualizations to display outlier counts, percent outliers, percentiles, and quartiles. Those measures can then be linked to configuration properties like color schemes to visually separate outliers from the rest of the data.



Selected measures:
Count
Median
Outliers (Outlier Count)

Sort by: (None)

● Ascending
○ Descending

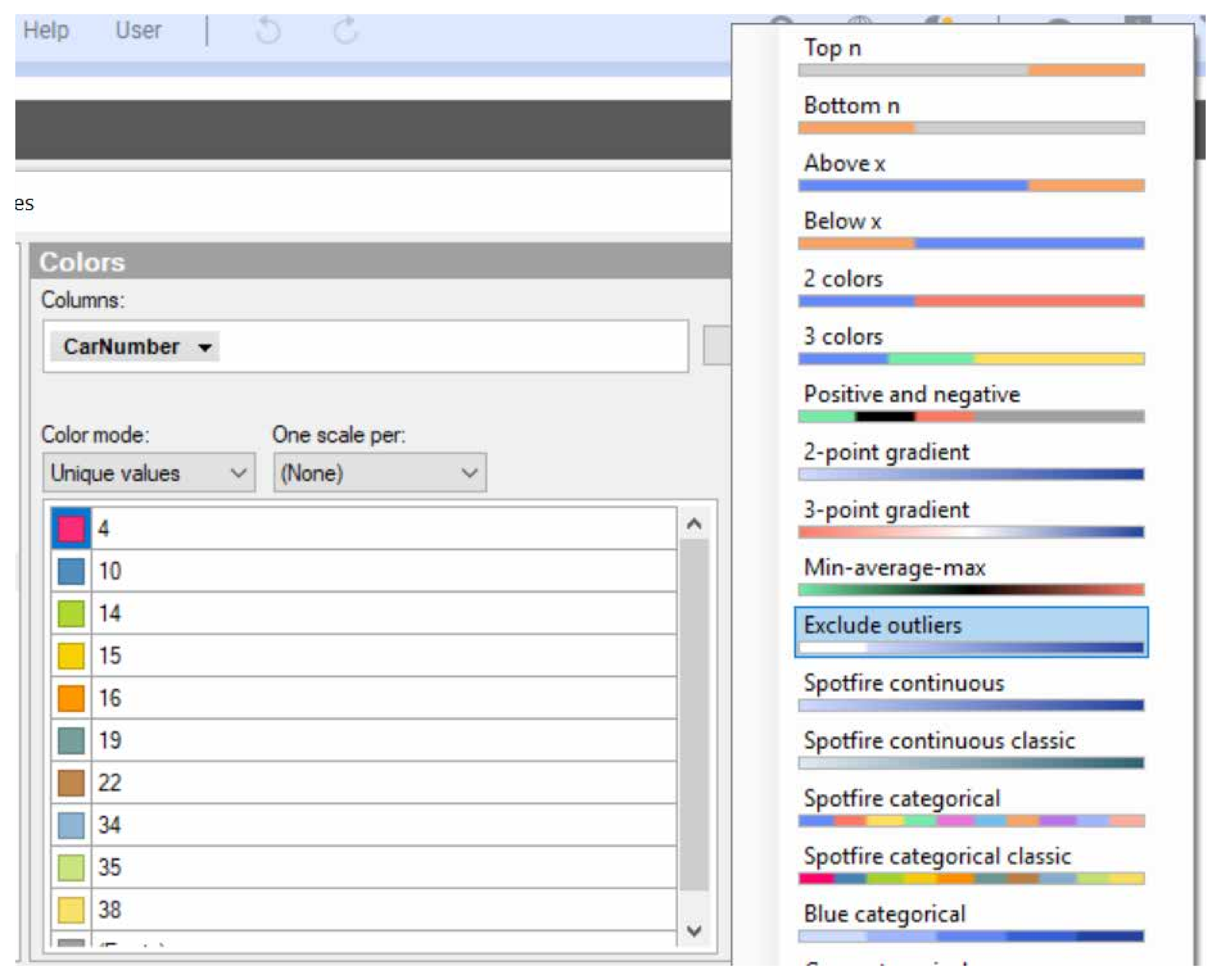
	ALASKA	CALI...	COLO...	C													...	MICH...	MIN...	MONTANA
unt	0	24	8															57	29	5
edian	(Empty)	59	57	8	3	66	151	8	109	79	133	132	66	104	16	23	13	81	59	53
tliers	(Empty)	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1

5. Use TERR to detect outliers

Combining TIBCO Enterprise Runtime for R (TERR) expressions with color can also be used to detect outliers. Custom expressions, expression functions, and data functions allow users to seamlessly integrate TIBCO capabilities with 10,000+ packages from CRAN using TERR or open-source R.

For example, combining the TERR expression with color could be used to choose a gradient color scheme based on outlier scores calculated by one line expression: outlier.score <- Rlof::lof(datacolumn, k=5).

Here, Rlof package contains lof function, which is an implementation of a widely used Local Outlier Factor algorithm to detect outliers. These scripts map to data elements (tables, columns, properties, etc.) and to R function inputs and can be saved and reused across columns, visualization configurations, and more. Such flexibility and extensibility from TIBCO is unmatched by any other analytics provider.



Columns: CarNumber

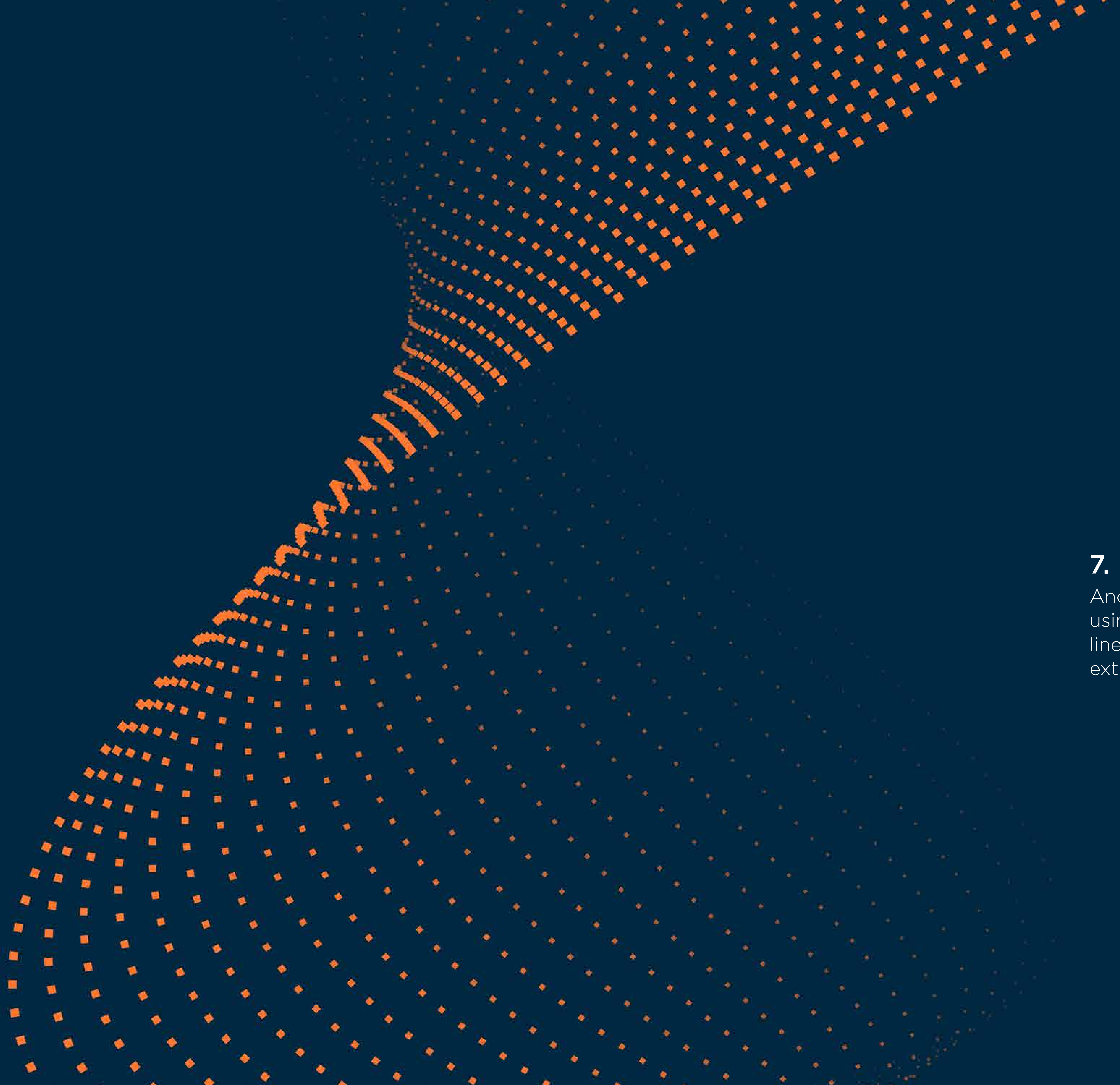
Color mode: Unique values

One scale per: (None)

6. Enable color scheme rules

You can also identify outliers with dynamic outlier color schemes, based on dynamic rules enabled by the analytics user. These rules may include:

- Excluding an outlier color scheme in predefined color schemes
- Specifying colors for points outside the interquartile range (i.e. outliers)
- Setting a threshold by mean, median, custom user-specified expression
- Using a gradient color scheme with dynamic outlier scores created in TERR as described above



7. Leverage curve fit or regression

Another way to detect outliers and anomalies is by using TIBCO visualization tools to insert a curve fit or a line fit to the data. This fit can then be used to identify extreme deviate points—outliers.

8. Similarity or clustering

TIBCO provides out-of-the-box functionality to apply line similarity and k-means clustering to visualizations, which can be used for outlier detection. You can choose the similarity metric—Euclidean or correlation—and other parameters like the number of clusters to create a line similarity or clustering label column in the data.

This column can then be used to color or trellis options. The stable number of clusters can be found by applying hierarchical clustering on the data. If the data has outliers, they will fall into their own cluster, for the number of clusters greater than the stable number.

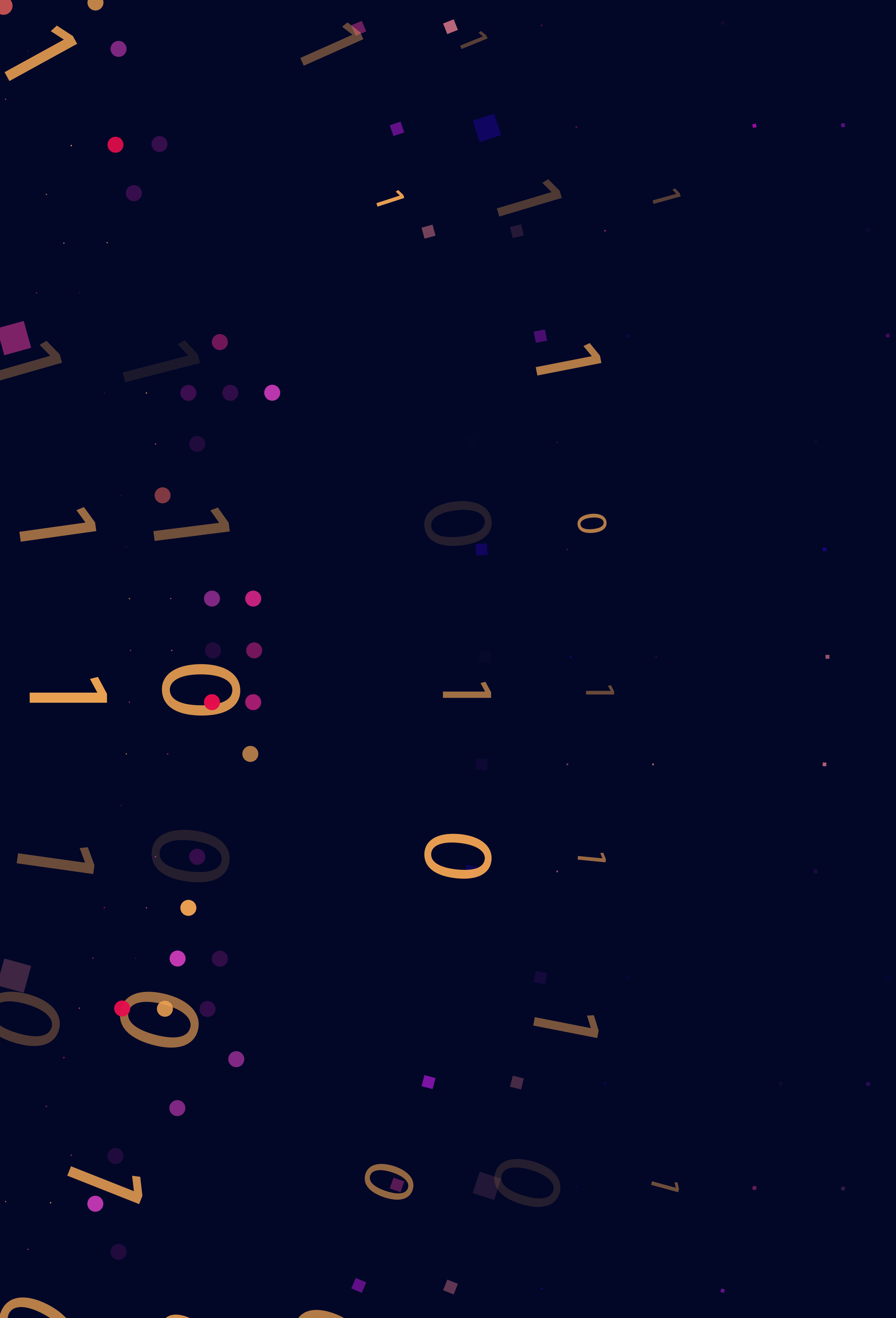
9. Explore advanced configurations

You can also detect outliers by exploring advanced analytics and data science configurations. New calculations and columns with expressions, expression functions, and data functions can be connected to configuration options that automatically label outliers. Also, advanced configurations for visualization properties can be extended beyond the color feature and can be applied similarly to markings, filters, subsets, and labels across visualizations.

10. Templates on the Community Exchange

To aid the citizen data scientist, the TIBCO team has made several plug-and-play templates available for free on our Community Exchange under the Analytics tag. These templates allow the user to plug in their data with the push of a button and explore the insights with minimal configuration.

Anomaly detection using deep learning neural nets is one such template that analyzes the input data to find anomalies based on recreation error during unsupervised learning. Another domain specific use case is the quality control template that identifies violations or outliers from the established control limits of the individual points, the moving average, and variance. These templates allow the user to extend the definition from a common outlier to a domain specific outlier and smartly identify and label them.



How do I learn more?

This guide briefly summarizes the top 10 methods for outlier detection. But if you'd like to learn more about how outlier detection contrasts with methods for anomaly detection, check out TIBCO's anomaly detection learning page at http://www.tibco.com/solutions/anomaly-detection for more resources.