TIBCO Spotfire

Working with F1 race data

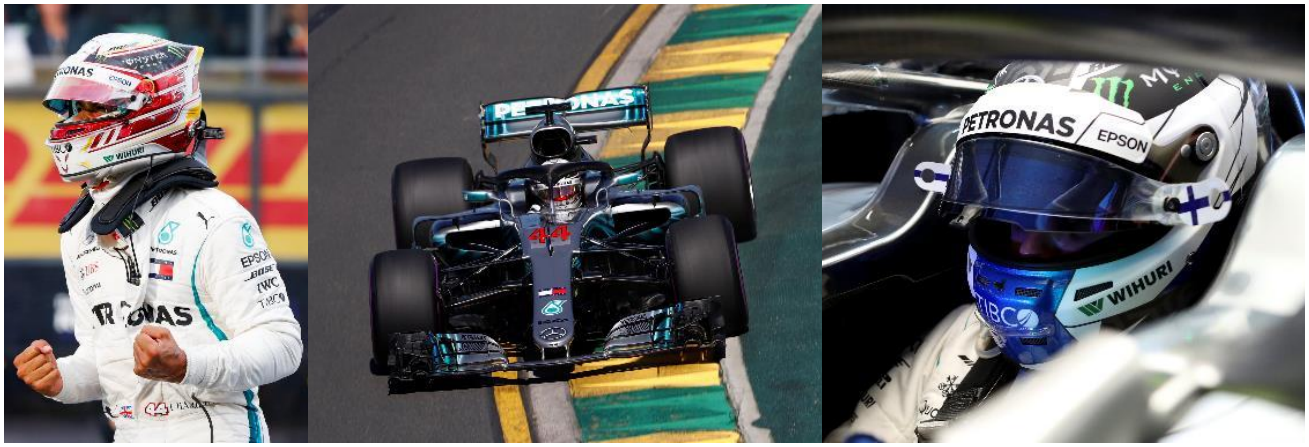


Table of Contents

Introduction

Congratulations you've been selected to be part of the F1 Analytics team and your task is to prepare our drivers for the upcoming race. You've been given access to historical as well as streaming car data and the objective is quickly find insights in this data.

You'll be working with historical as well as streaming data from the F1. You're going to use Spotfire to visually explore the data and use built-in data science that will help you along the way.

This workshop consists of 2 parts (historical and streaming analysis) and will take about 1 hour to complete.

In order complete the workshop you will need:

- An internet connection
- A Spotfire Cloud trial account.
- A windows machine with the Spotfire analyst client installed.
- Access to the F1 Github repo

The data contains information about times, positions, distances as well as a lot of information about the car, such as current gear, pedal positions, temperatures, speed, etc.

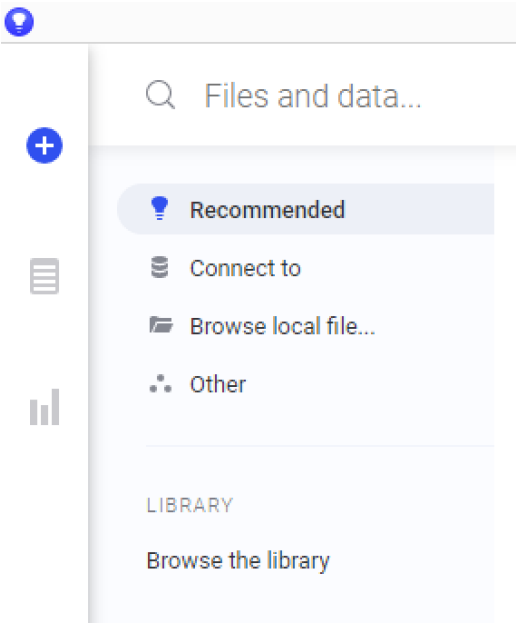
During the workshop we'll try to figure out what circuit the data is from. We'll visualize the track and will look at clustering the data to create different clusters of data.

Good luck!

Part 1 - Historical F1 Data Analysis

Load data SBDF

Drag and drop the file “Historic Performance Data.sbdf” onto your Spotfire Analyst, or click on the + sign on the left top of the analyst client and choose ‘Browse local file’ to locate your file.



Validate data

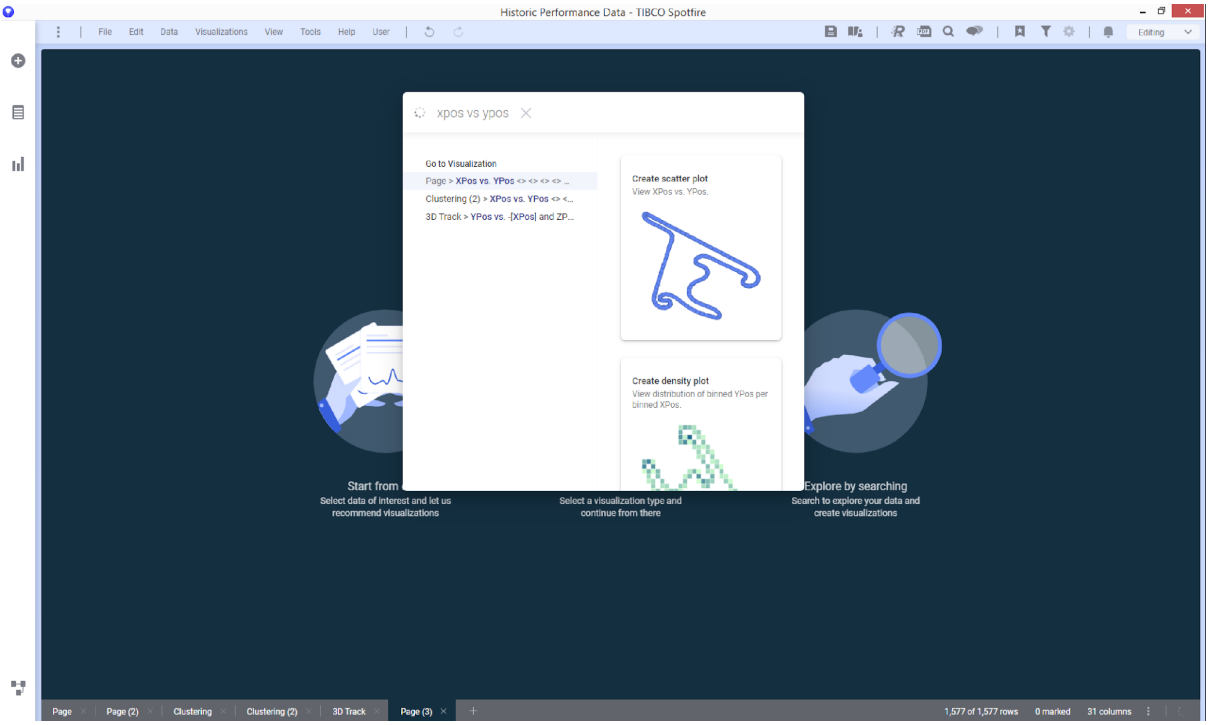
Let start looking at the data by opening the Data Canvas, do this by clicking the button on the bottom of the sidebar. This opens the data canvas where you can preview the data table, as well as see the number of rows and columns, and any transformation applied to the data. Once you’re done click on the data canvas icon or the grey X to close it.

It is good practice to look at the data in the data panel to perform an Exploratory Data Analysis (EDA). Click on the second item from the top of the sidebar in order to open the data panel. If you select the gear symbol it will provide all kinds of information about the data. Please select individual columns in order to perform an univariate analysis (examining individual columns). Is the number of values as expected (in case of categorical data) or is the minimum, maximum and distribution as expected (for continuous variables). The data panel also helps to clean and transform the data in case this is required. What to do with empty values? What if there are trailing spaces? The recommendation engine will assist you in spotting these kinds of data issues.

Start visualizing the data

Let's start by visualizing the data and specifically the circuit. As a picture says more than a thousand words, let's see if we can see from the data what track the data is about.

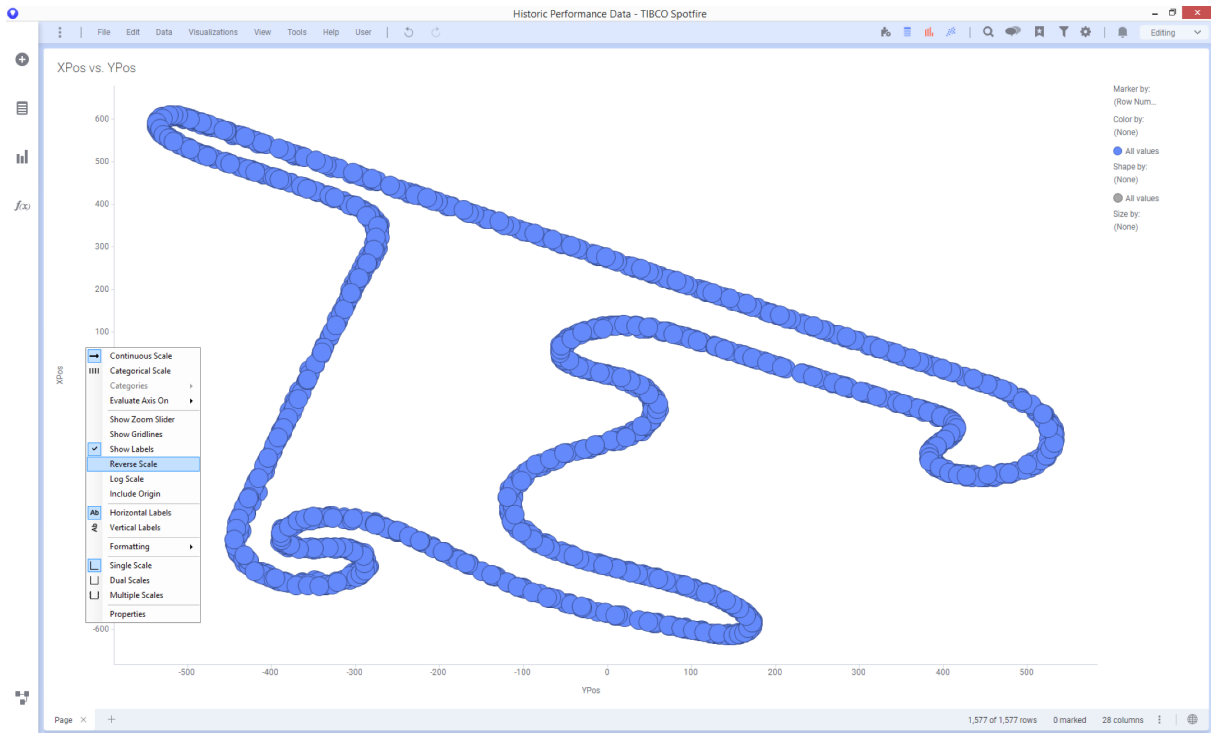
In order to do so, click the right icon on the start screen that says "Explore by searching". This will launch a Natural Language Processing (NLP) fueled search engine. This feature allows you to explore your data simply by using text.
Please type "xpos vs ypos" in the search bar and hit enter. This will suggest a number of visualizations.



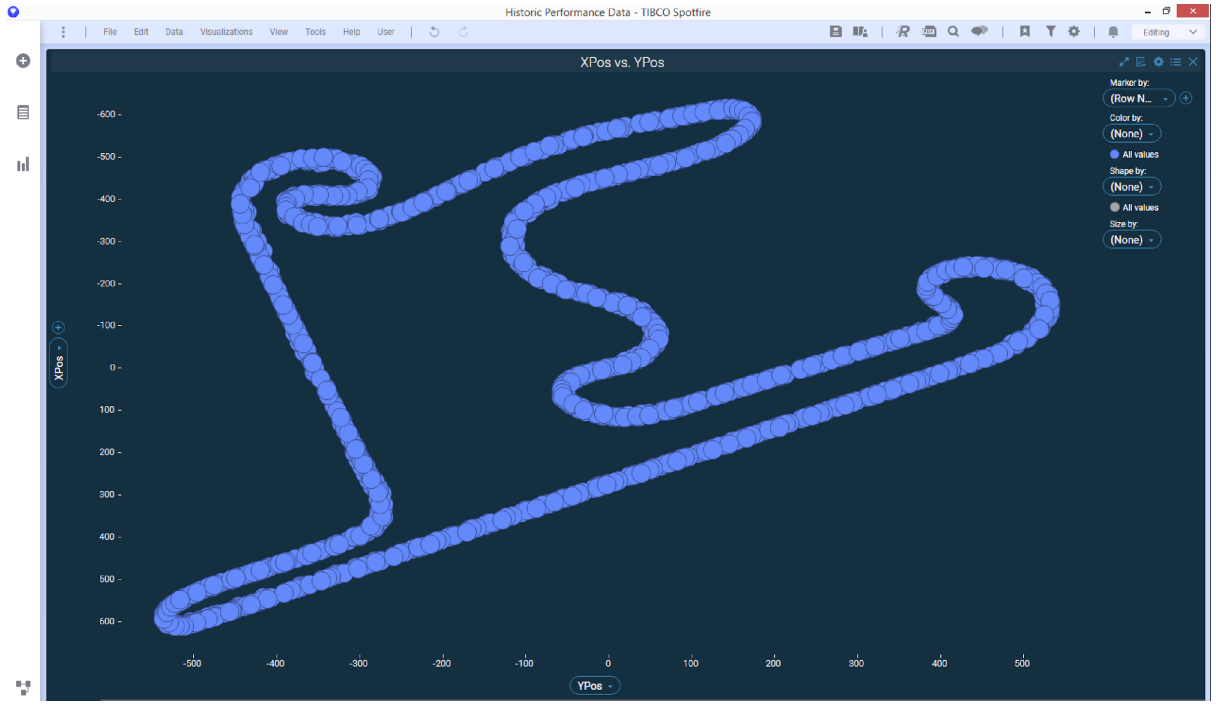
If you browse across the suggested visualizations you get a number of suggestions on how to best visualize this type of data. For our purpose the first suggestion, a preconfigured scatter plot, seems to make sense. Let's go ahead and add the suggested visualization.

For the F1 fans you should be able to see the racing track that we are analyzing right? If not, take a look at the tracks on the last page of this document.

If you compare the track you see that the image is mirrored. In order to fix this, right click on the xpos axis and choose "Reverse Scale".

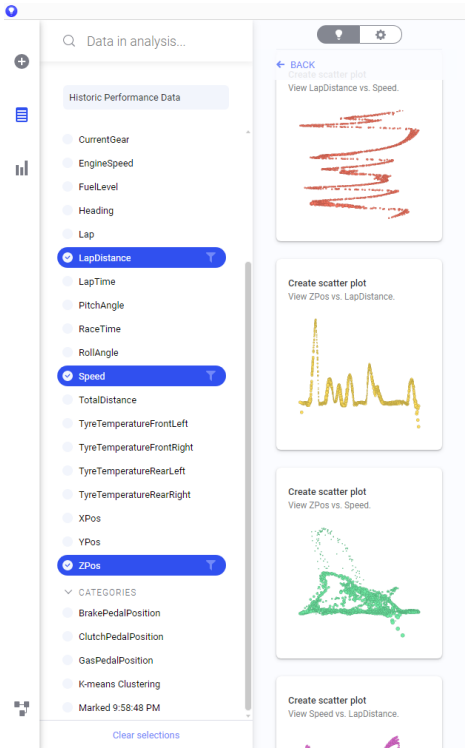


Now your visual should look like the one depicted below:

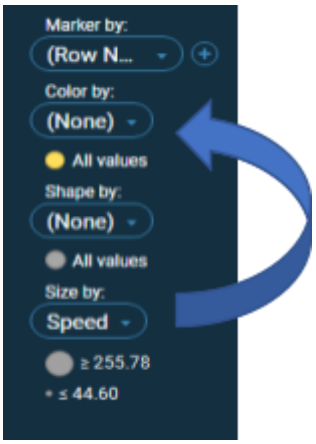


Let's add another visual to this page that shows both the altitude and speed over the distance of the lap. Instead of using the search let's now use the data panel. Click on the second item from the top of the sidebar  in order to open the data panel.
Now select the fields we are interested in, i.e. zpos, speed and LapDistance. (You can select multiple columns by clicking on the check mark in front of the column name to include it in your selection.) Click on the lightbulb in the expanded menu to view the recommended

visuals and wait for the recommendations to show up. Hover over the first recommendation and select ‘More like this’ to view alternative configurations of this suggested visualization.



The recommendation in yellow seems to meet our needs. So let’s drag and drop that onto the canvas on a location where we would like the visual to appear. For readability it is prefered to present the speed in color rather than size. To do this, drag and drop the speed from the ‘size by’ to the ‘color by’ axis.

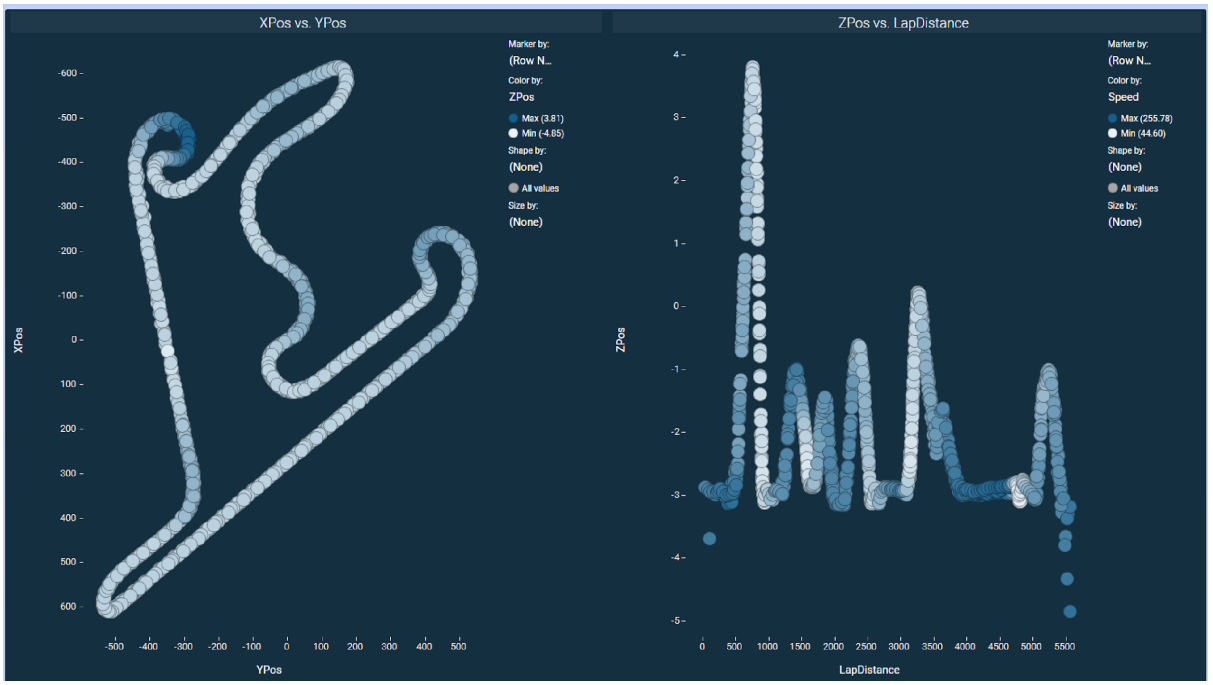


Please change the track visualization to show ‘z-pos’ as the ‘color by’ axis. This will expose the altitude on the track.

Now we have a good feeling of the shape of the track, the difference in height and the speed that the cars are going over the track distance.

Obviously, we could also choose to color the track based on speed or another metric to get some more information in our track overview. For now, we'll stick to this configuration.

You should now have two visuals that look similar to the ones below.



Notice that the visuals you created are brush linked. This means that if you highlight (or mark) data points in 1 visual, the relevant data points highlight on the other visual.

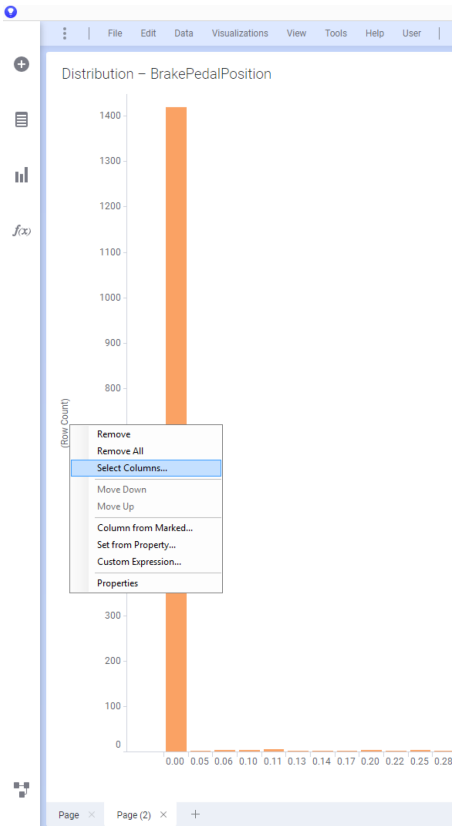
Can you identify the slowest part on the track? Tip, you might want to use the 'size by' axis on the right visualization to help answer this question.

K-Means Clustering

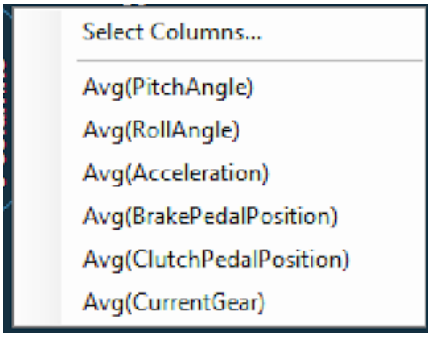
Up to this point we've used Spotfire to visually explore the data.
Let's now see if we can use out of the box data science features of Spotfire to better understand the nuances in our data.
Each line represents a moment in time, the position and conditions of the race car.
Let's see if we can find a way to group the data into clusters in order to understand where on the track the car conditions are similar.

Let's add a second page to the analysis file by clicking on the + sign next to the Page tab.
Next we will create a new line chart on this new page. This can be done by clicking on the bottom icon of the three icons on the top of the sidebar.  Then choose the line chart from the available visualization types.

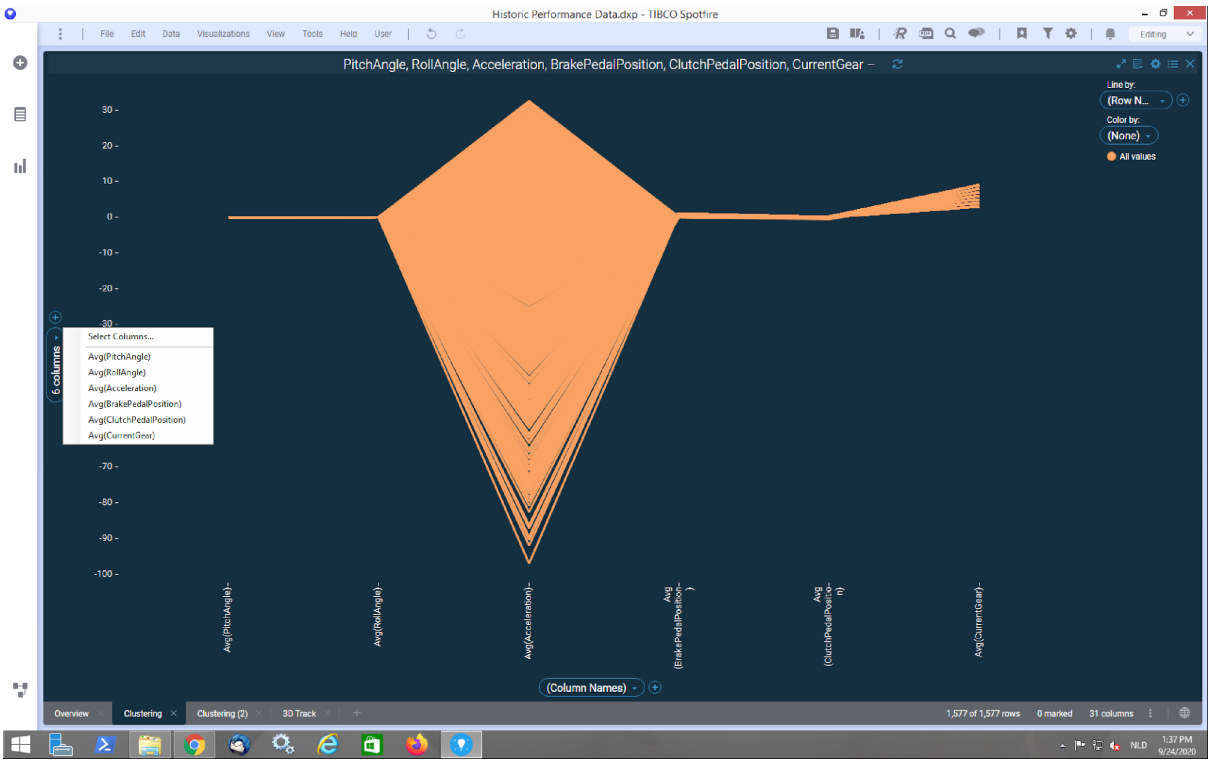
On the 'line by' axis choose 'row number', so that each row represents 1 line in the chart.
On the Y-axis we are going to select the following columns and aggregation methods. To configure these right click on '(Row Count)' and choose 'Select Columns'.



Configure the following:

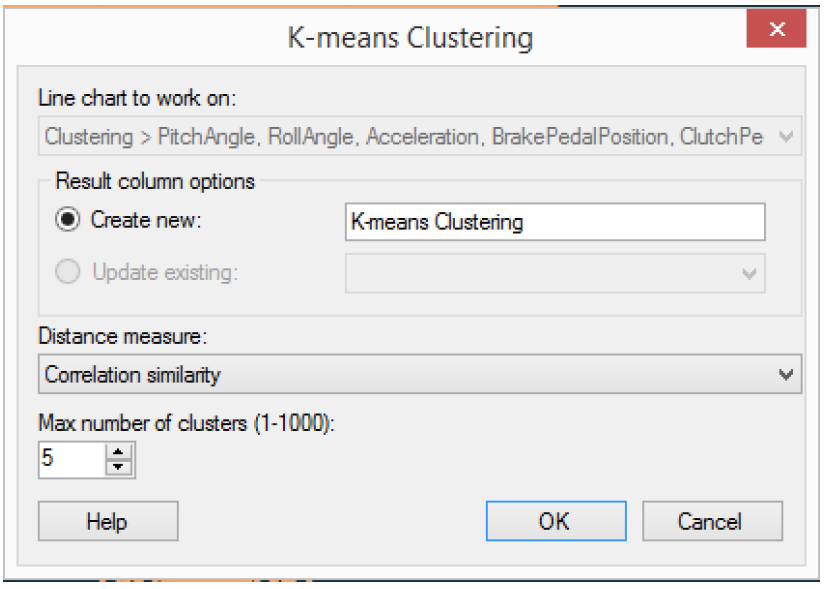


On the X-axis we will now choose (Column Names)'. Your visualization should now look similar to the below:



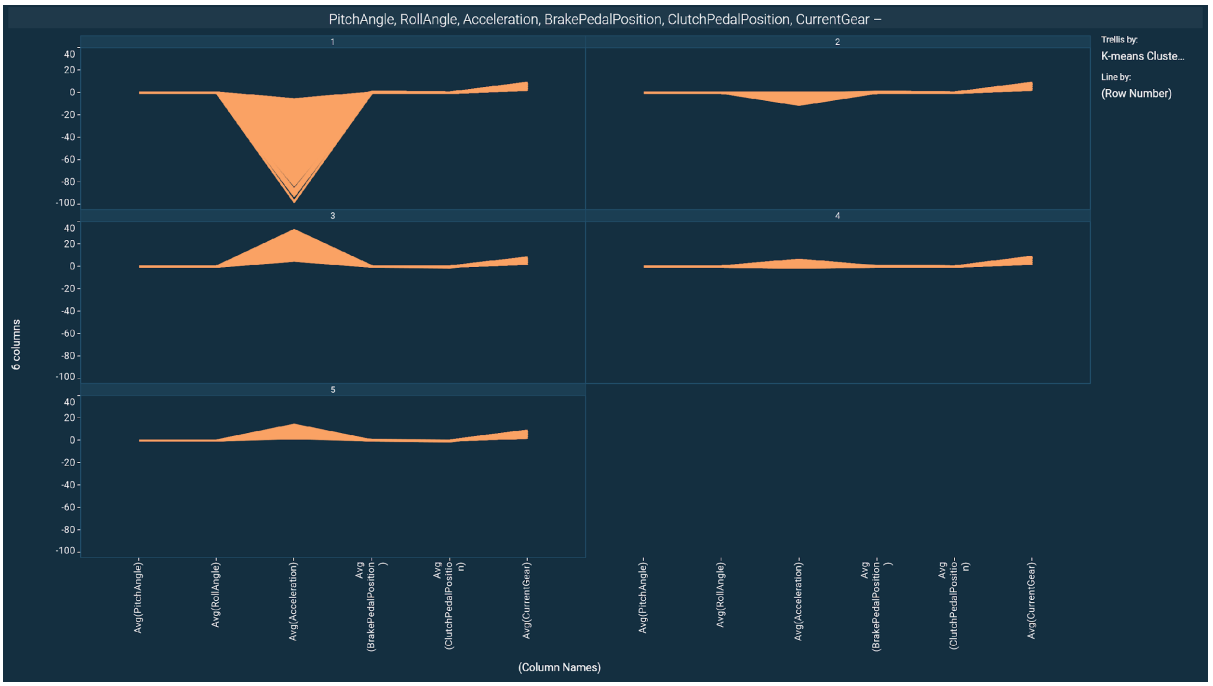
This line chart now shows the value for each of the metrics configured on the y-axis for each record in the data set. We will ask Spotfire to automatically cluster these lines that show group them based on their similarity. In order to do so, right-click on the visual and choose k-means clustering.

This will bring up a window that allows you to change some of the settings.

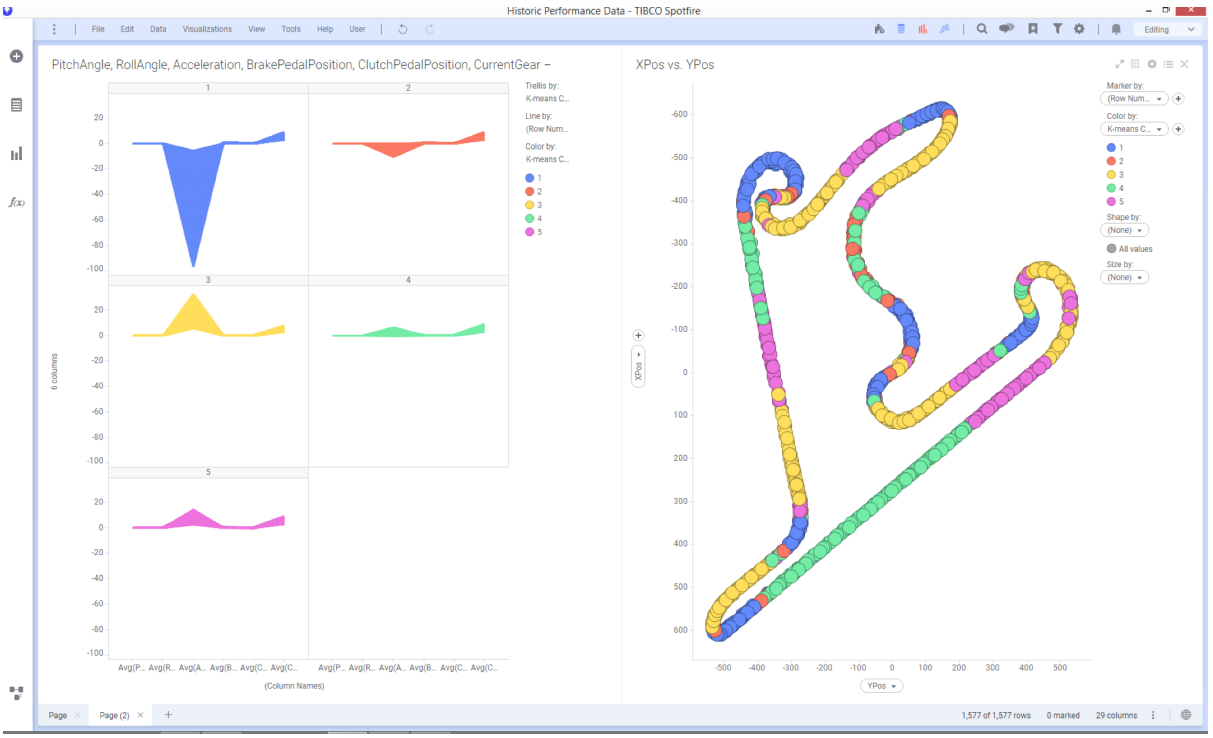


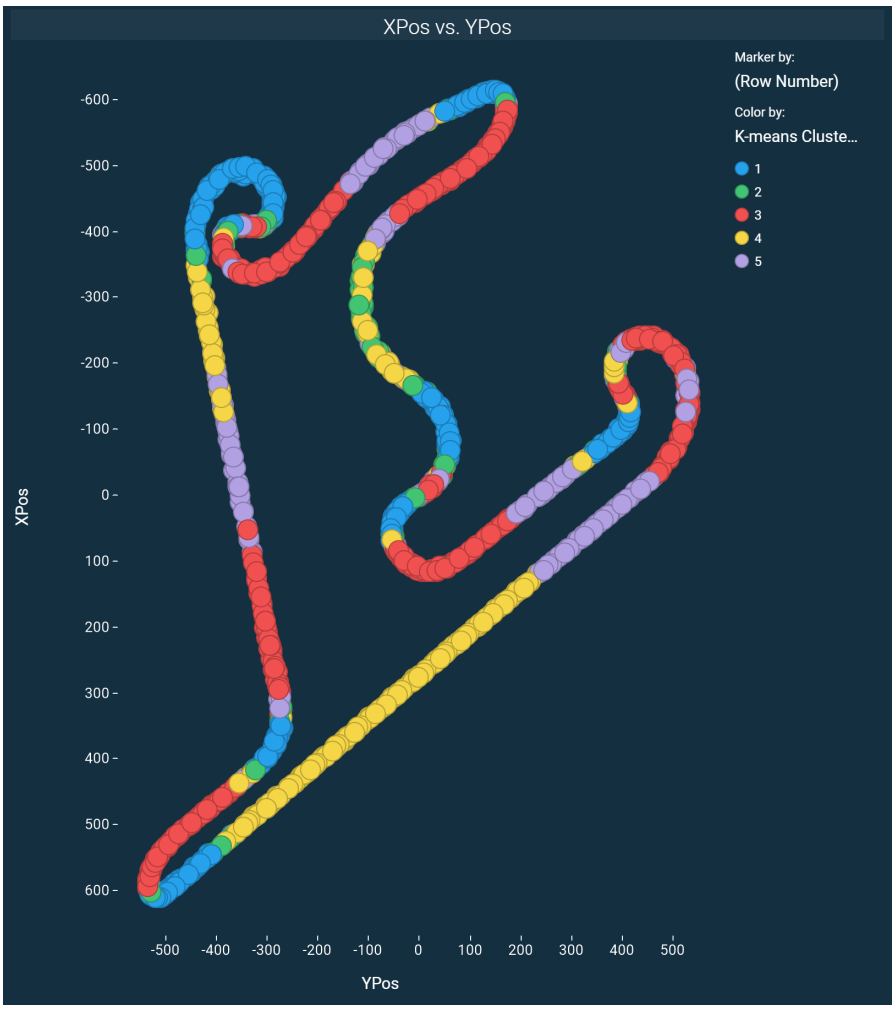
For now, we will use Correlation similarity as the distance measure and a maximum number of clusters of 5. If you would like more background on k-means clustering, click on the help button for more details. Click ok to execute the clustering.

This will add an additional column to the dataset, indicating the cluster that the line of data falls into. Also, the visualization will update and be trellised by cluster as shown below.



Now let's add another version of the track visual we created on the first page by using copy visualization and then on the page with the clustered line chart paste the visualization (ctr+v or via edit menu choose 'paste'). Now change the 'color by' axis to the newly created 'K-means cluster' column. Do the same of the clustered line chart.





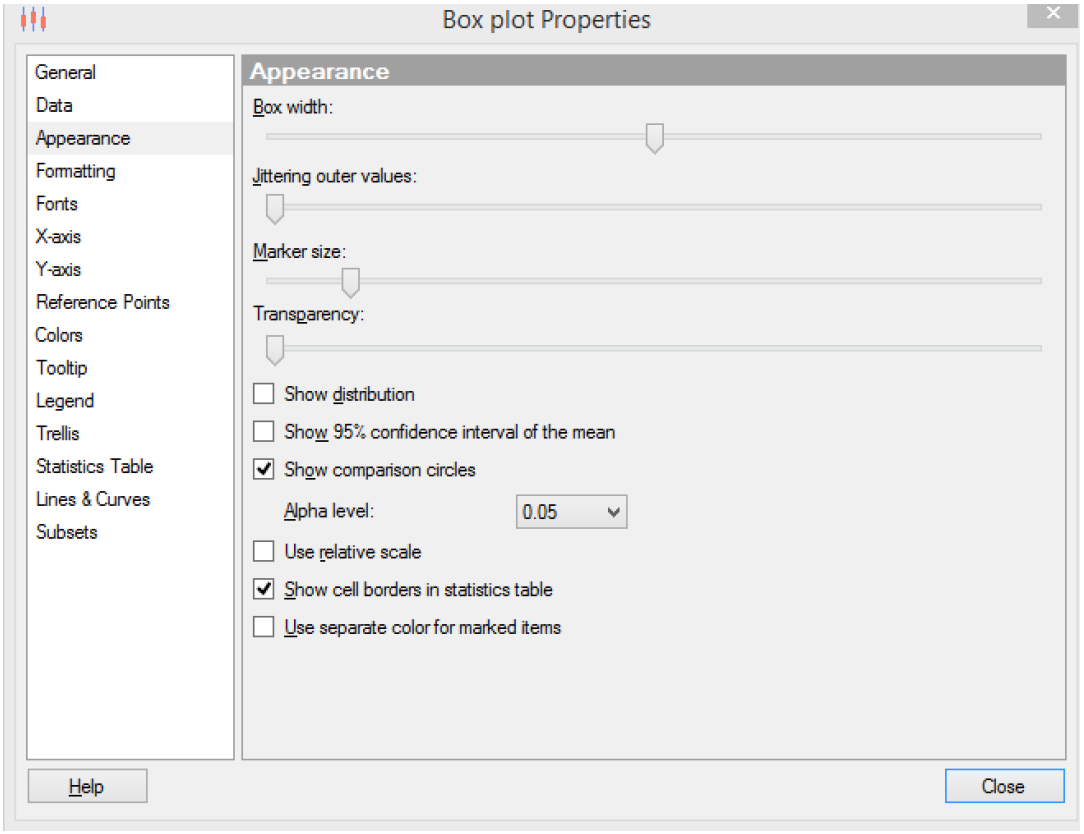
You will see that the data is now clustered into different categories, i.e.

1) Going into a bend
2) Intermediate areas
3) Coming out of a bend
4) High speed (straight(er) stretches
5) Accelerating after coming out of a bend

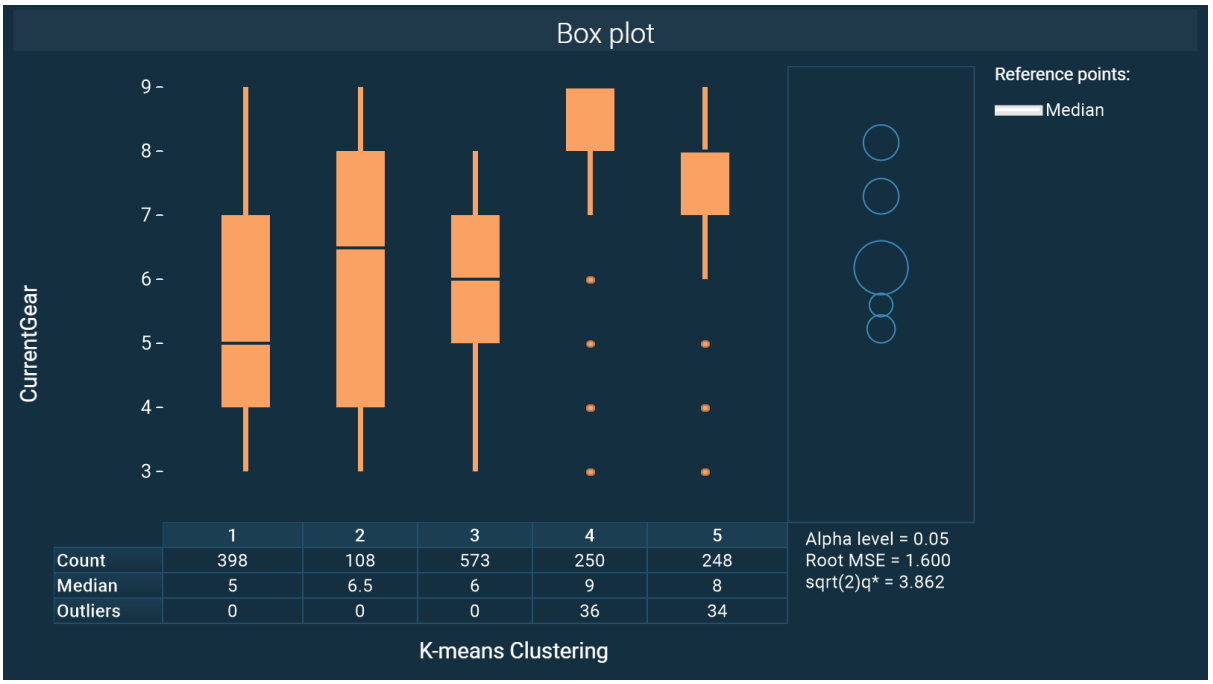
As the cluster is added as an additional column, we can now use this information in any part of the analysis. A nice feature to analyze the cluster is to use box plots.

Let's add a boxplot to a new page and click on the 'Add visualization' icon in the sidebar and choose the boxplot.

Configure the x-axis to show 'K-means Clustering' and set the y-axis to represent 'CurrentGear'. Finally go into the properties of the boxplot on the Appearance tab and select the checkbox in front of Show comparison circles.



Your visualization should now look similar to the visualization in the image below. The comparison circles are useful to identify where the distribution between the different clusters is significantly different. When marking a circle, the visualization will indicate if the distribution is not significantly different by showing a line below the boxplot when the circles overlap.



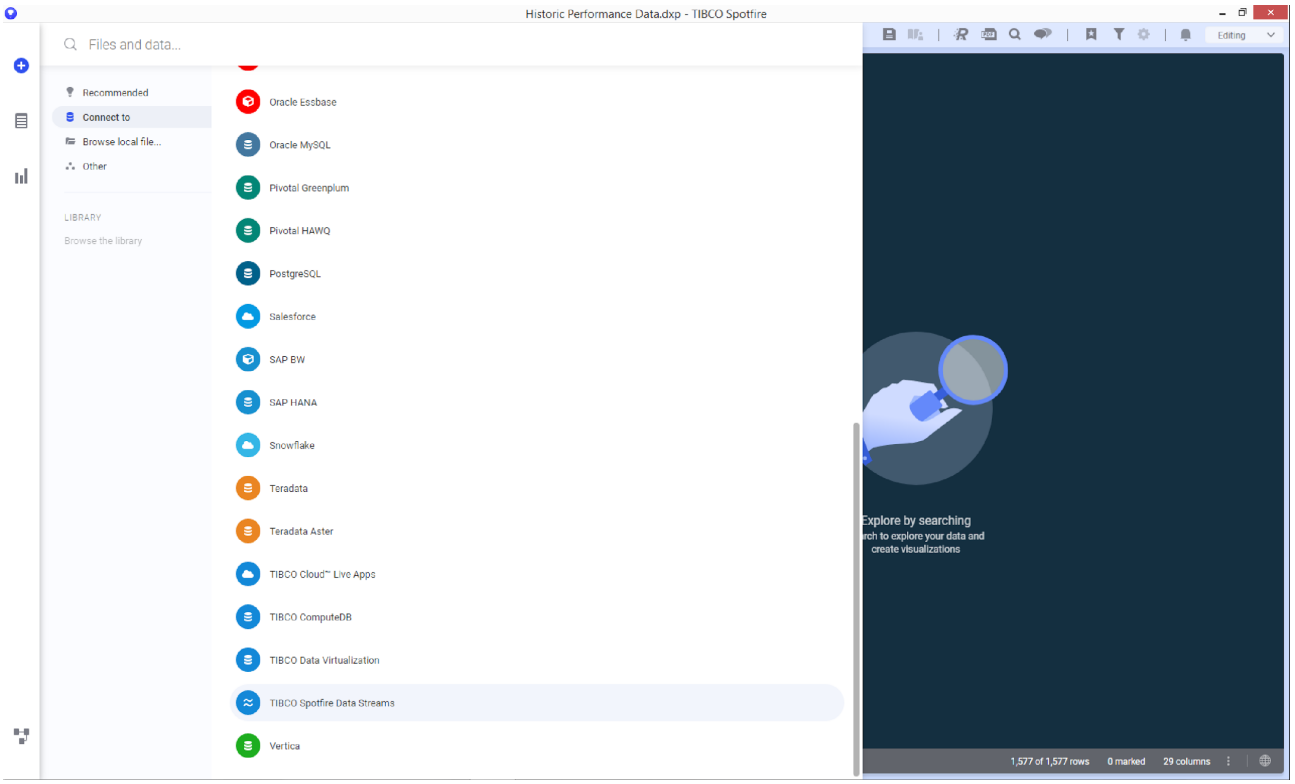
Investigate for different measures if this is the case to get a feeling of how the data is clustered.

Using these comparison circles you find out which clusters are significantly different? Which conclusion can be drawn?

Part 2 - Streaming F1 Data Analysis

Streaming Data

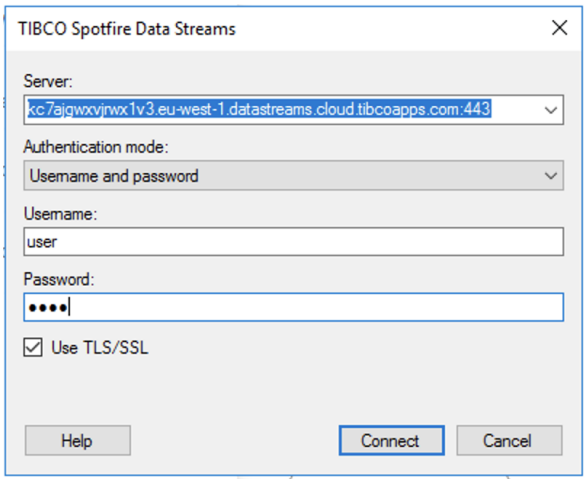
Now let's have a look at adding some streaming data to the dashboard.
We've asked Lewis Hamilton to start driving the circuit since we need some real time data for this workshop. In order to analyze it we'll add a connection to TIBCO Data Streams by clicking on the + sign on the right top of the sidebar.
Then choose "Add connection to" and then browsing down to TIBCO Spotfire Data Streams as shown on the image below



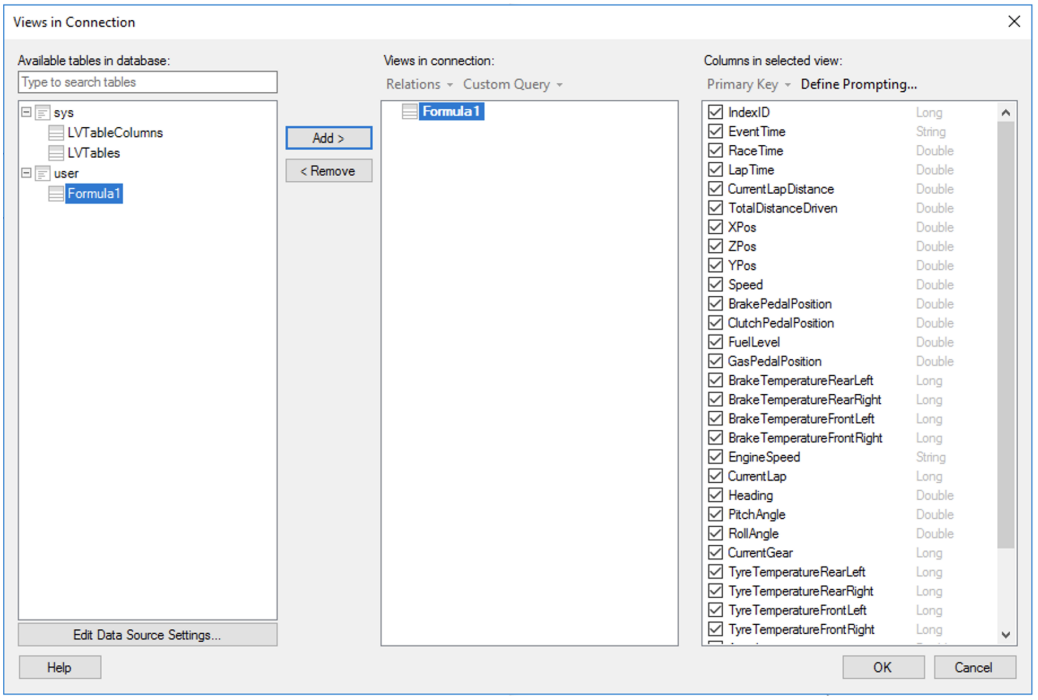
This will bring up a window where we need to add the server name or ip address, etc.

Use the information below to connect to this Data Stream (TIBCO Spotfire® Data Streams Server).

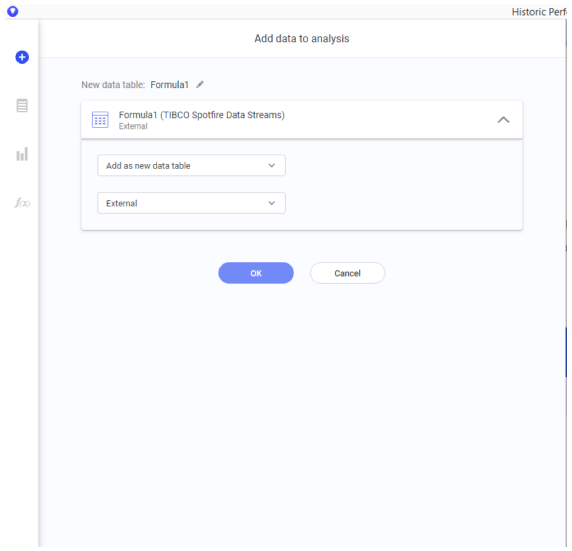
Server: kc7ajgwxvjrwx1v3.eu-west-1.datastreams.cloud.tibcoapps.com:443
Authentication mode: Username and password
Username: user
Password: user
Use TLS/SSL: Yes



You will see that connecting to a streaming data source is pretty much the same as connecting to any database. Add the table under user with the name Formula1 to the view as shown below.



On the next screen, make sure that data is kept ‘External’ as indicated below.
This will allow a live connection to be set up to TIBCO Data Streams and data will continue to flow into Spotfire.

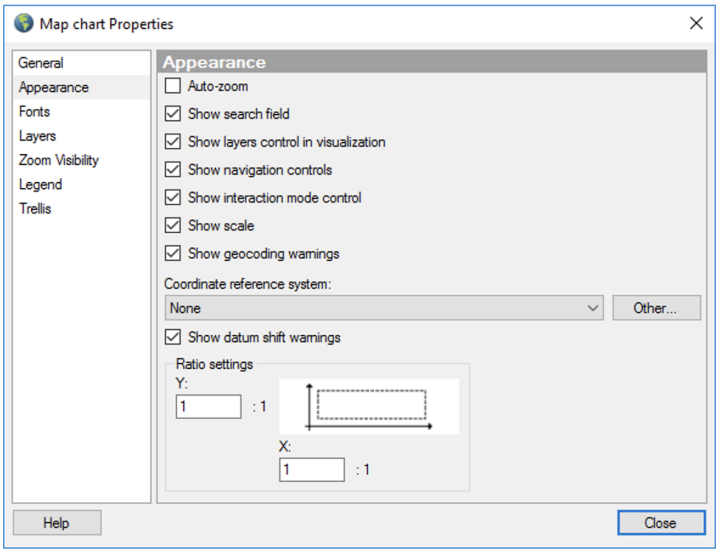


Let’s also load some data about the track by clicking on the + sign (top left on the screen) and choosing to browse ‘local file’. Browse to the file named “track.sbdf” (which can be downloaded from the F1 github repo here) and add it to the analysis.

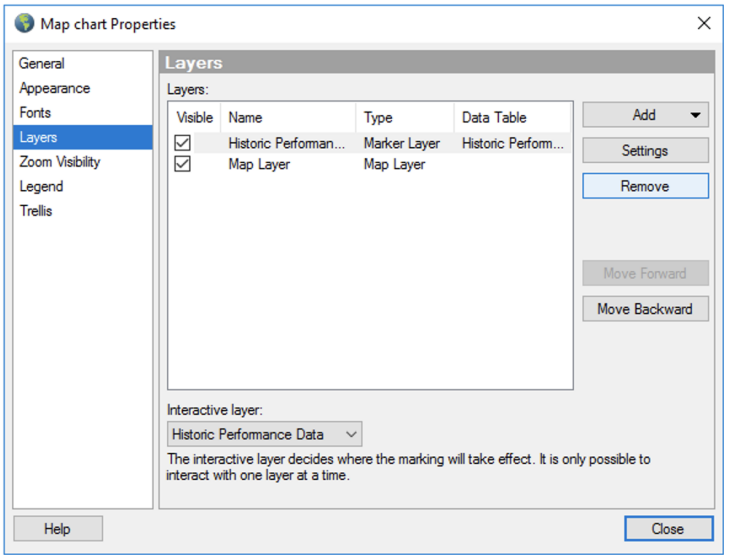
Please add a Map Chart.

In the most recent version of Spotfire the map chart should show up with the track already visualized, because the track table is the only georeferenced data in the analysis. If this is not the case, you will need to take a couple of additional steps.

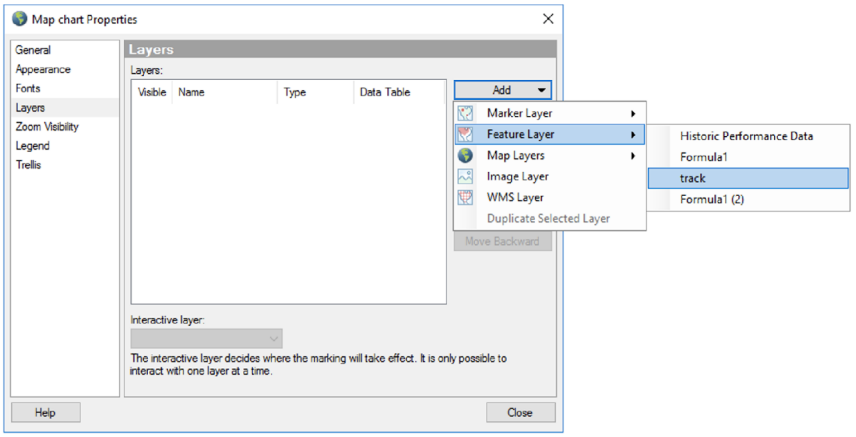
Click on the gear symbol at the right of the title bar to go to the properties dialog. Now go to the appearance settings and change the Coordinate reference system to “None” as shown below.



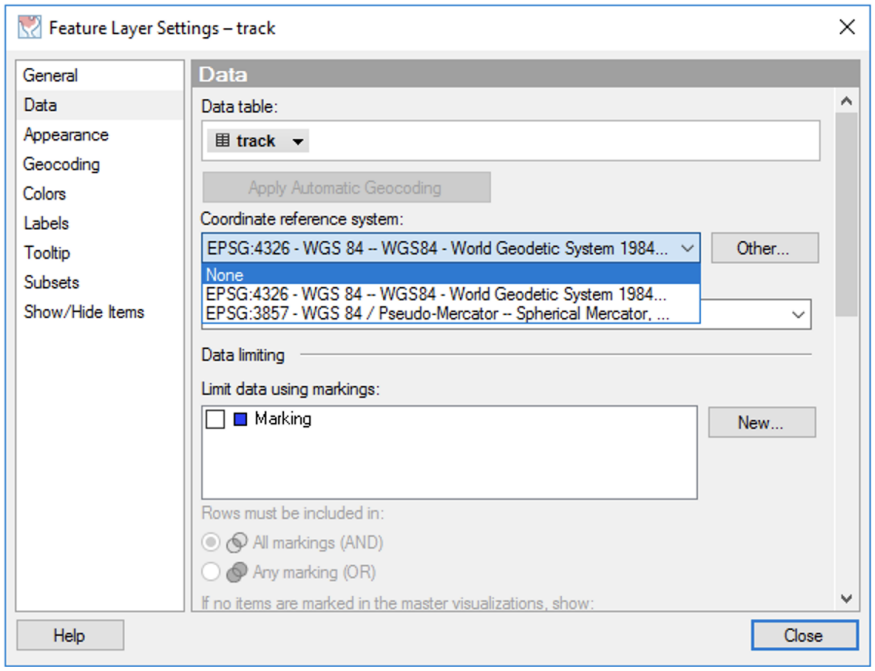
The reason for doing this is that the track table does not contain any reference system data. Then move to the layers section and remove all existing layers.



Then add a new feature layer based on the track table as shown below.

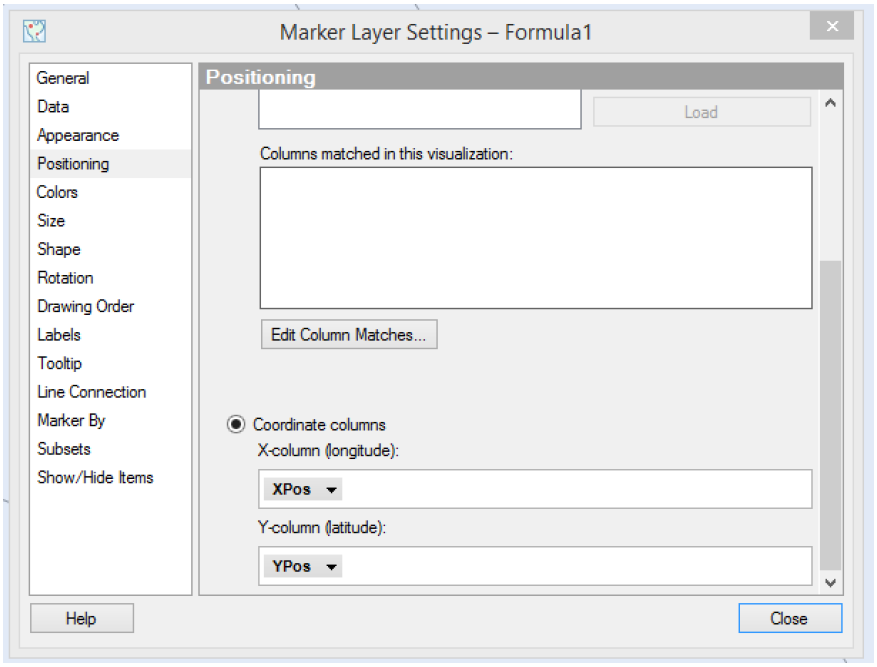


Finally for this layer set the coordinate reference system to ‘none’.



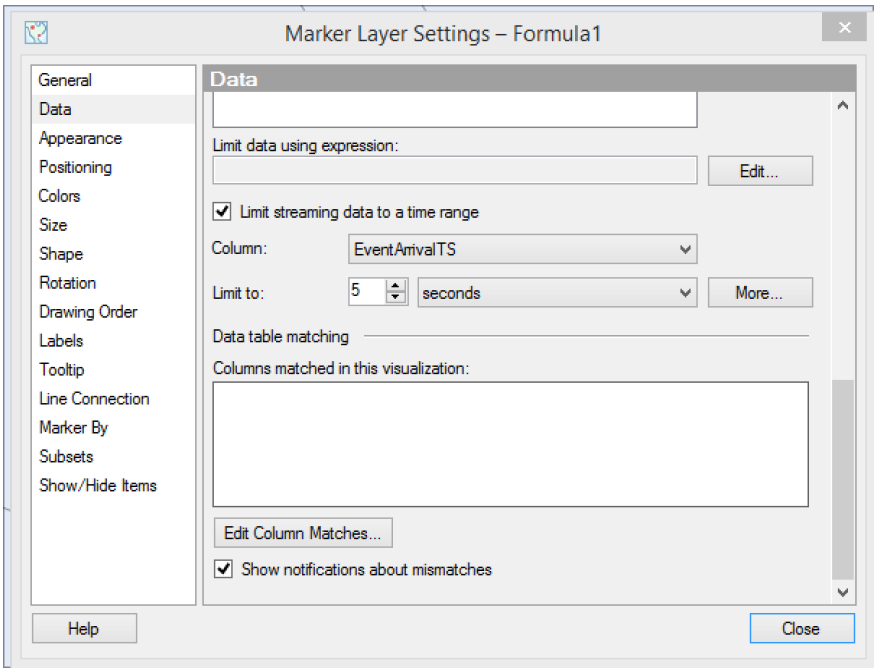
Now we Add a layer for the streaming data to show up. This will be a marker layer.

For the position we go to the Positioning area and scroll down to Coordinates columns. Let’s add xpos as the X-Column (with aggregation set to none) and ypos with aggregation set to none for the Y-Column

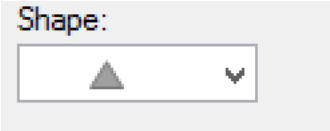


On the marker by tab choose IndexID.

As we don't want to flood the visuals with data we are going to limit the data that is displayed. On the Data Tab Scroll down a little to the area where it says limit streaming data to a time range and choose the column EventArrivalTS as the column to limit the data on and type 5 in the limit to box and finally choose seconds as the measure.

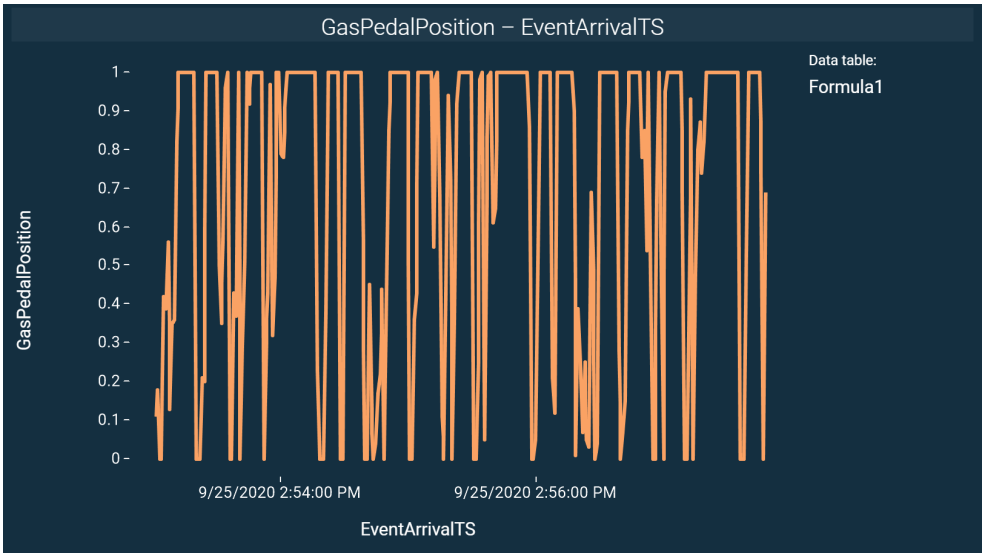


Now as a last step we would like to see the direction the car is moving. In order to do so, choose the triangle as the shape.

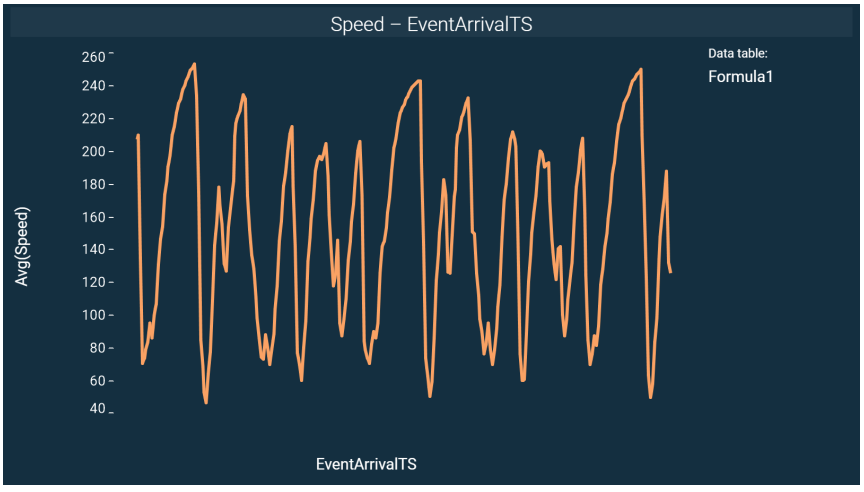


And go to the Rotation tab and choose the average as “Heading”. Leave the direction as clockwise.

If you have finished the streaming map chart, let’s try and add one or two more visuals to the page. In this case let’s add a line chart that uses the Formula1 streaming data table and put EventArrivalTS on the X-axis and GaspedalPosition on the Y-axis. Bear in mind that you might want to limit the data that is shown in the visual in order to create a rolling window of data.



A last visualization to add is another line chart that shows the EventArrival on the X-Axis and the Avg Speed on the Y-Axis.



Congratulations, you have finished your first streaming dashboard in Spotfire!

Appendix 1

F1 Circuits