



## **Contest: Data Analysis and Visualization**

**PLEASE DO NOT OPEN UNTIL INSTRUCTED TO DO SO**

### **IMPORTANT:**

Any solution received that includes any IDENTIFYING information will not be eligible for anything beyond "Honorable Mention."

Examples include any of the following:

- Names of the individuals or school names in the solution
- Names of the individuals or school in the "properties" of the file
- Each program, Adobe, Microsoft, etc. contains metadata about the document. Remove ALL of these from your document. If you do not know how please ask someone!

Any solution that does not provide accurate citing of professional resources will be removed from consideration.

Examples include any of the following:

- Copying and pasting diagrams and images from a website
- Using descriptions and product data verbatim from source
- When in doubt, cite your source.

## Data Analysis and Visualization Contest

### Contest Statement:

Common techniques in data analytics and visualization include data acquisition, data cleaning, handling missing values, wrangling, data integration, simple data analysis (e.g., outlier detection, identifying collinearity between variables, dimension reduction, summary statistics, and determining skewed attributes), statistical model building, finding patterns, finding clusters, plotting geo-temporal points on a map, data visualization, visual analytics, problem solving, making recommendations, and providing quantitative analysis to support decision making.

Each team may consist of one or two individuals and must have at least one computer to complete your project. The contest is written from the perspective of guiding someone who is beginner/intermediate to using Tableau. **Tableau is the primary software recommendation to complete the tasks of this contest.** Appendix A contains a quick start guide to Tableau's interface. You are welcome to use other software and you are only required to turn in your answers and a visualization (using any software of your choice) for the required tasks.

### Contest – Deliverables:

Solve the tasks in an individual or two-person team – sharing completed solutions with other teams is not allowed. While Tableau is the primary software recommendation, you are welcome to utilize other existing software to manipulate/analyze data, write your own code, or follow any other analytical methodology you prefer to arrive at a solution to each task. To complete this challenge, create and turn in a document that answers as many tasks as you can. Also, turn in any other supporting files such as computing code (e.g., Python scripts), Tableau packaged workbooks (.twbx file), intermediate datasets (e.g., CSV files), etc. to showcase your work. Do not upload any part of this contest or your deliverables to online locations (e.g., GitHub).

**Important Note: If Tableau is used, Tableau packaged workbooks (.twbx file only) must be submitted along with a Word/PDF document.**

**If Python is used, (.py/.ipynb files) must be submitted along with a Word/pdf document.**

## Contest – Rubric

The following weighted rubric details how the results will be evaluated. The final submission should be organized in a single Word or PDF file (with legible screenshots if needed). Provide solutions that include all the required deliverables (.twbx file for Tableau/.py or .ipynb files for python) for each task that you can complete. Missing supporting files such as .twbx file for Tableau/.py or .ipynb files results in 50%-point deductions. Partial solutions will be given up to half credit.

- 1. Sort your tasks 1-12 in their proper order.**
- 2. Clearly indicate which tasks you attempted and which tasks you skipped.**
- 3. Task 1 through Task 12 (8.3% for each task solved)**

Additional items considered while evaluating the tasks:

The correct answers for all tasks. Proper use of joins to combine datasets. Appropriate use of charts and visual representation. Demonstrating filters, parameters, and functions. Extra points for creating interactive visualizations.

## Resources - Software recommendations

It is recommended to download and use Tableau Desktop for this contest (free year-long license) or use online Tableau Cloud for the contest.

See: <https://www.tableau.com/academic/students>

Tableau and PowerBI are leading software tools for visual analytics and rapidly generating interactive visualizations. You may download Tableau Desktop from [tableau.com](https://www.tableau.com) with a 14-day free trial, or perhaps Tableau Cloud. Getting the license key will likely only take a few seconds. If you require brief tutorials on how to use the GUI and its functionality, see APPENDIX A and Additional Resources below, or visit <https://www.tableau.com/learn/training>

Do NOT post and make your work visible for others to see. Tableau connects with a variety of underlying dataset formats: offline data files on your hard drive, online databases, an online data server, or the default practice datasets (e.g., try out the World Indicators dataset). Note how the columns from the dataset are available on the left pane, you can drag columns to the middle pane to set the x-position y-position color size (which automatically updates the chart graphic), dragging columns to the filter pane generates dynamic filters within the visualization, and the right pane allows you to change the chart type depending on the columns already selected. Users explore their dataset by creating a sheet and pairing various combinations of dimensions and measures. For each combination of columns that you explore, switch the visualization to several options (including Tableau's recommended chart choice, which is highlighted with a red box on the list of possible charts on the right pane). You might use this contest as an opportunity to become familiar with Tableau, how to connect to a dataset, how to create charts (sheets), and how to create interactive dashboards. Some of the work for each task can be done in Excel.

**Additional Resources** (from Tableau that will be useful for this competition):

1. [Understanding Calculations](#)
2. [Types of Calculations](#)
3. [Aggregate Functions](#)
4. [String Functions](#)
5. [Date Functions](#)
6. [Quick Table calculations](#)



If you choose to use Python for this contest, SciPy is a set of widely used packages for managing, analyzing, and visualization large-scale content. It consists of libraries for data structures such as DataFrames and analysis (pandas), N-D arrays (NumPy), 2D graph plotting (Matplotlib), scientific analysis (SciPy), etc. In addition to matplotlib, other popular python libraries include ggplot (any R fans in the group), plotly, seaborn, pygal, bokeh, geoplotlib, etc. These packages differ by their customization, expected data input structures, charts available, export formats (e.g., svg), interactivity, dependencies (web integrated), etc.

See:

<https://www.python.org/about/gettingstarted>

<https://docs.python.org/3/tutorial>

<https://scipy.org>

<https://matplotlib.org>

## **Contest – Case Study Background**

For this challenge, we will use a historical Formula 1 racing dataset that contains rich, structured data about Formula 1 Grand Prix races from 1950 up to recent years.

**Dataset Source:** The dataset is compiled from the [Ergast F1 open database](https://ergast.com/docs/f1db_user_guide.txt). To learn about feature information and the details of the dataset, please see [https://ergast.com/docs/f1db\\_user\\_guide.txt](https://ergast.com/docs/f1db_user_guide.txt).

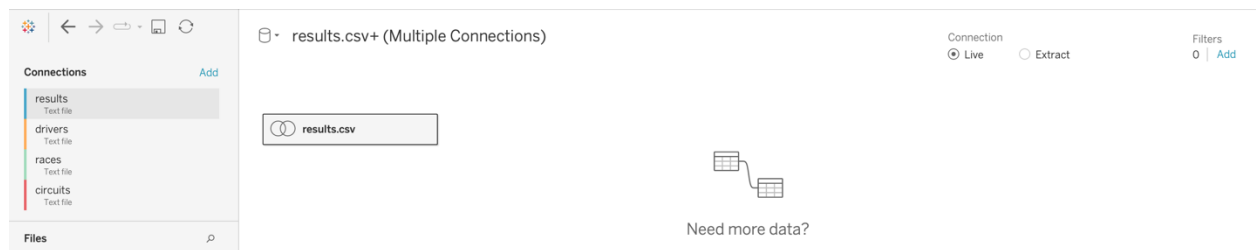
The dataset contains the following tables:

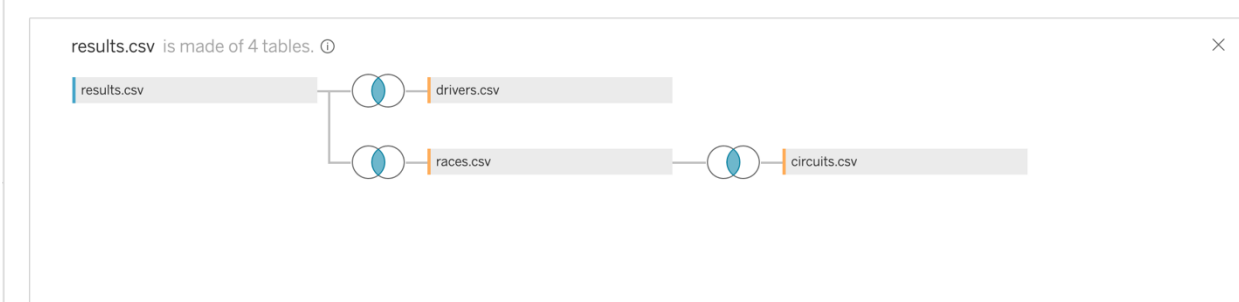
- **Drivers** – driver details (name, birthdate, nationality, etc.)
- **Constructors** – team details
- **Races** – each Grand Prix event (with year, round, circuit ID, date)
- **Circuits** – circuit details (name, location, country, latitude, longitude, etc.)
- **Results** – results of each driver in each race (finish position, grid position, points scored, status like finished or retired, etc.)
- **PitStops, LapTimes, Qualifying** – data on pit stop counts/times, lap times, qualifying results (for more advanced analysis)

For this Visualization contest, we will use **Results, Drivers, Races, & Circuits Tables only**. By using these Tables together, we can investigate performance metrics, compare drivers, analyze races across different countries, and even examine trends over decades. The presence of numeric fields (points, positions, counts) and categorical fields (drivers, years, status) enables analyses involving outlier detection, clustering, and predictive trends.

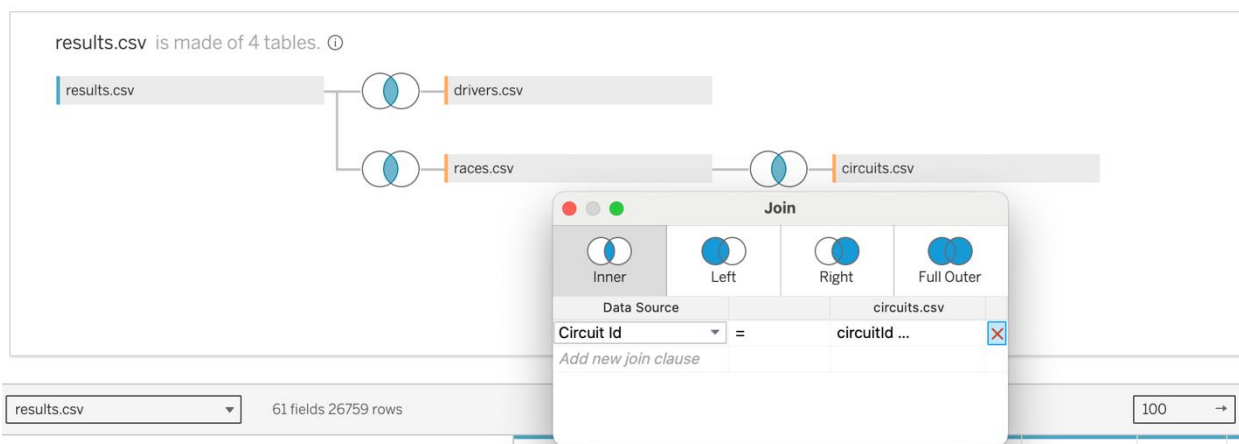
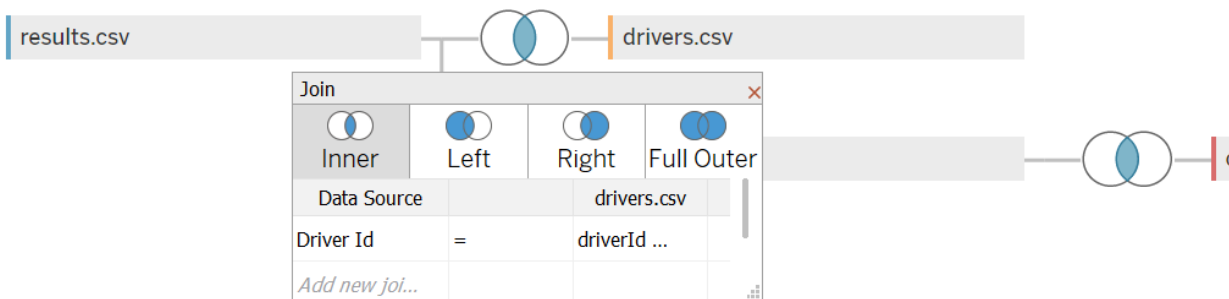
## Using the data in Tableau:

1. Download the four csv files(**results, drivers, races, & circuits**) provided. Then launch Tableau and go to the Start Page. In the left-hand Connect pane, under To a File, click Text file (or More > Text file if it's not visible).
2. Browse to the folder where your CSV files are stored and select **results.csv**. Similarly connect **drivers, races, & circuits** files.
3. Drag the **results.csv** table from the left pane onto the Data Source canvas(Logical layer) (the large area on the right).
4. Double-click on **results.csv**. You will now use this physical layer add and join each of the other CSV files by dragging them one by one.
5. Below are the screenshots to help you with connecting these CSV files using inner joins with proper IDs and fields.





results.csv is made of 4 tables. ⓘ



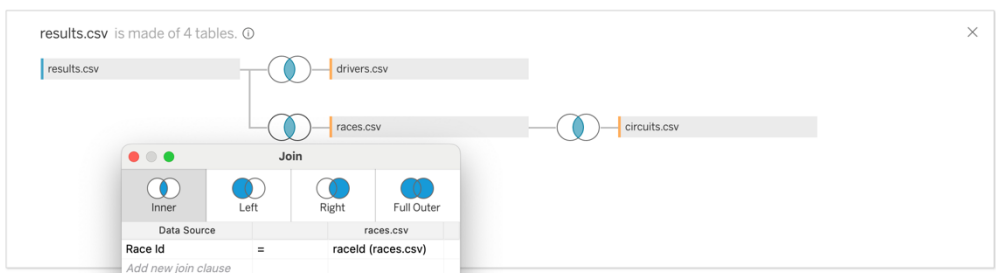
**Connections** Add

- results Text file
- drivers Text file
- races Text file
- circuits Text file
- constructor\_results Text file

**Files** ⌵

Use Data Interpreter  
Data Interpreter might be able to clean your Text file workbook.

- circuits.csv
- constructor\_results.csv



**Note:** Not joining them correctly will distort visualizations and results.

Because the data is clean, we can focus entirely on analysis and visualization rather than preprocessing. With this dataset, we can create complex Tableau visualizations involving filters, calculated fields, parameters, and various chart types.

## Contest – Tasks to solve

The following are competition tasks based on the Formula 1 dataset. These tasks require you to explore the data and produce a visualization or answer using advanced Tableau features.

1. Which Formula 1 driver has won the most Grand Prix races in their career, and how many wins did they achieve? Identify the driver with the highest total wins.
2. Who is the youngest driver ever to win a Formula 1 Grand Prix, and how old were they when they first won? Provide the name of the driver and their age (in years, or years and days). Optionally, answer who are the three youngest ever Formula 1 Grand Prix race winners.
3. Who is the oldest driver ever to win a Formula 1 Grand Prix, and how old were they when they first won? Provide the name of the driver and their age (in years, or years and days).
4. In which Formula 1 season did the greatest number of different drivers win at least one race? How many different winners were there in that season? Identify the year and the number of unique winners. Optionally, answer what year had the second most unique winners.
5. Use clustering to group drivers based on their career performance statistics. Consider metrics such as total points, total wins, etc. How many distinct clusters of drivers emerge, and what characteristics define each cluster? (*Hint: You might find clusters like “all-time greats” vs “mid-fielders” vs “occasional racers.” Use Tableau’s clustering on a scatterplot of drivers with the chosen measures.*) Describe the clusters using captions and list a few example drivers in each group.
6. Create a Choropleth map of all Formula 1 circuits. Which country has hosted the most Grand Prix races in the dataset? List the country (or countries) that hosted the most races and the number of GPs held there.
7. Using the circuit location data, determine which Formula 1 circuit is the northernmost (i.e., highest latitude). What is the name of that circuit and its latitude coordinate?
8. Compare the geographic distribution of F1 races between an early period (say, the 1950s) and a recent period (the 2010s). Which region or continent saw the greatest increase in the number of Grand Prix events from the 1950s to the 2010s? (*Hint: Group races by continent or region for each decade. You may need to create a custom grouping of countries into continents. Then compare counts between decades.*) Identify the region with the biggest growth and briefly describe the change in the caption.

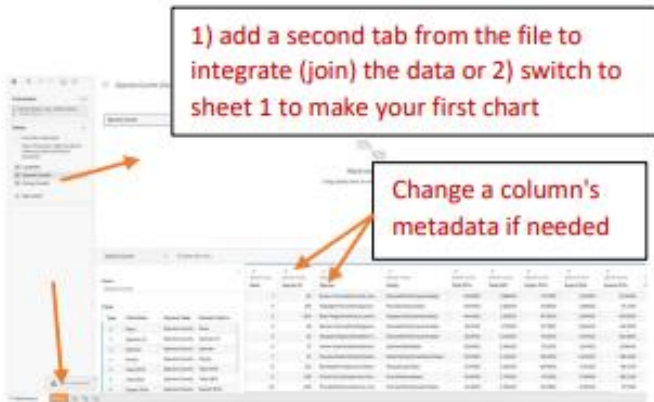
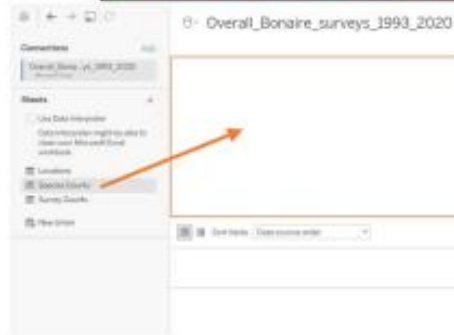
9. Analyze how the number of Grand Prix races per season has changed over F1 history. Create a timeline showing the count of races each year. What is the trend, and using Tableau's forecasting or trend line, predict how many races might be held in a future season? (*Hint: Use the Races table; each year has a certain number of races. Use a line chart with Year on the x-axis and Count of Races on y-axis. Add a forecast or trend line.*) Summarize the historical trend and the forecasted value.
10. Identify the year in which the total points scored by all drivers saw the most dramatic increase compared to the previous year. How big was the jump, and what likely caused it? Provide the year and discuss the likely reason in the Caption.
11. Create a parameter in Tableau to select a specific driver and display that driver's performance over time. For example, when a user picks a driver's name from a parameter dropdown, show a line chart of that driver's points or wins for each season of their career. Provide an example by selecting a notable driver and describing their year-by-year performance trend as shown in your chart.
12. Some drivers achieved many podium finishes (top 3) but never won a race. Which driver has the most podium finishes without ever winning a Grand Prix? How many podiums did they have, and during what years did they race? (*Hint: Identify drivers with 0 wins, then among those, find who has the most finishes with positions 2nd or 3rd. The results table has the finish position for each race.*) Provide the driver's name and number of podiums.

## Appendix: Getting Started with the Tableau Desktop Interface - A Quick Guide



Select a . csv or Excel

Load one spreadsheet tab from a data source at a time. Dragging two tabs will force Tableau to perform an inner-join to integrate the data.



1) add a second tab from the file to integrate (join) the data or 2) switch to sheet 1 to make your first chart

Change a column's metadata if needed

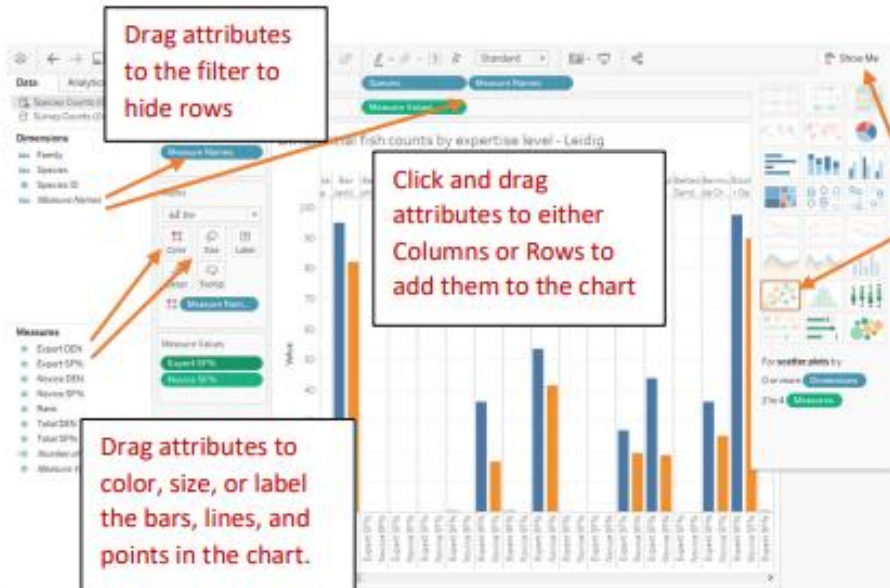


Pick one of your loaded datasets to use for this chart

View your loaded datasets

Make a new sheet (i.e., chart)

New dashboard (2+ side-by-side charts)



Drag attributes to the filter to hide rows

Click and drag attributes to either Columns or Rows to add them to the chart

Choose the type of chart you want to generate. Based on your data, Tableau will recommend a chart type with the red outline in the "Show Me" window.

Drag attributes to color, size, or label the bars, lines, and points in the chart.