Data Visualisation

Created	@November 14, 2024 5:23 PM
Tags	

Introduction to Data Visualisation

Data visualisation is the graphical representation of data and information. By using visual elements like charts, graphs, and maps, data visualisation tools provide an accessible way to see and understand trends, outliers, and patterns in data.

The main objectives of data visualisation are:

- To communicate information clearly and effectively through graphical means.
- To make complex data more accessible and understandable.
- To identify patterns, trends, and correlations that might not be apparent from raw data.

Common types of data visualisation include bar charts, line graphs, pie charts, histograms, scatter plots, heat maps, and geographic maps.

Good data visualisation should aim for clarity, accuracy, and efficiency. It helps the audience grasp difficult concepts or identify new patterns.

Conceptual Model

The Conceptual Model of Data → Visualisation → Insights describes the flow from raw data to actionable insights, emphasizing the transformation stages that turn data into meaningful knowledge.

1. Data

- Definition: Data is raw information, unprocessed and often unstructured. It may come from various sources (e.g., databases, sensors, surveys, or logs) and can represent anything from quantitative figures to qualitative observations.
- Types of Data: It includes both structured data (like numbers and dates in tables) and unstructured data (like text, images, or videos).

- Purpose: The primary purpose of data is to capture real-world phenomena for analysis, providing the foundation needed for generating insights. For example, in business, data might include sales figures, customer demographics, and transaction histories.
- Challenges: Raw data is often messy or incomplete, requiring cleaning, validation, and transformation before analysis can begin. Cleaning may involve removing duplicates, handling missing values, or normalizing data across sources.

2. Visualisation

- Definition: Visualisation is the graphical representation of data. It transforms numbers, text, or other complex data points into visual formats, such as charts, graphs, maps, or diagrams.
- Types of Visualisations: Examples include bar charts, line graphs, scatter plots, heatmaps, and geospatial maps. Each visualisation type serves different analytical purposes, like showing trends, distributions, or correlations.
- Purpose: Visualisation makes complex data more understandable, enabling users to see patterns, trends, and outliers that might be hard to identify in raw data alone.
- Process: Visualisation requires selecting the appropriate chart type, scaling data properly, and adding elements like colors, labels, and legends for clarity. Tools like Excel and Tableau simplify this process, providing options for customisation.
- Benefits: Visualisation is crucial for storytelling in data analytics, as it translates raw data into formats that non-technical audiences can easily interpret and make decisions from.

3. Insights

- Definition: Insights are meaningful interpretations derived from data. They go beyond surface-level observations to reveal actionable knowledge, patterns, or trends.
- Types of Insights: Insights can be descriptive (what happened?), diagnostic (why did it happen?), predictive (what is likely to happen?), or prescriptive (what should we do about it?).

- Purpose: Insights drive informed decisions by explaining what data trends mean in a real-world context. They provide the “so what” of data, helping organisations understand customer behaviour, market trends, operational efficiencies, and more.
- Process: Extracting insights often involves a combination of domain knowledge, analytical skills, and intuition. Analysts must interpret visualisations thoughtfully, connecting data points and drawing conclusions that are relevant and actionable for stakeholders.
- Challenges: Generating insights requires careful interpretation to avoid misrepresentation. Insights should be validated, and assumptions should be transparent to ensure that decisions based on them are sound and reliable.

Definition of data Visualisation

Data visualization is the study of the visual representation of data, and “information that has been abstracted in some schematic form, including attributes or variables for the units of information (Friendly, 2008)”.

“The main goal of data visualisation is to communicate information clearly and effectively through graphical means .... To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way. Yet designers often fail to achieve a balance between form and function, creating gorgeous data visualisations which fail to serve their main purpose — to communicate information."

Different Types of Data

Data can be categorised into various types, and each type is best suited for different forms of visualisation:

- Quantitative (Interval Scale): Measures with meaningful differences but no true zero, e.g., temperature in Celsius (20°C and 30°C show a 10°C difference, but 0°C isn’t an absence of temperature).
- Categorical Ordinal: Qualitative data with a ranked order, but gaps between ranks aren’t meaningful, e.g., education level (high school, bachelor’s, master’s).

- Categorical Nominal: Qualitative data with categories that lack an intrinsic order, e.g., types of cuisine (Italian, Japanese, Mexican).
- Quantitative (Ratio Scale): Numerical data with equal intervals and an absolute zero, meaning absence, e.g., weight (0 kg indicates no weight).

Data visualisation taxonomies

The common definition of taxonomy comes from the biological sciences and refers to the organisation into groups of members that share similar characteristics. There are many ways to classify data visualisation methods. Here, we provide a framework to categorise data visualisation methods defined by the primary communication purpose. When visualising data, the first step is to understand what we are trying to achieve with the graph. Do we want to compare values or show how they have evolved over time? Here we will look at 5 of these categories namely: comparing, part of the whole, evolution, correlation charts and maps.

- Comparison charts: These facilitate comparisons of categorical values. Categorical data are values arranged into distinct groups, for example, months, age groups, types of animals. The classic example of a comparing chart is the bar chart.
- Part of the whole charts: This type of chart shows categorical values in their proportion to the whole. Pie charts are often used to display this.
- Evolution charts: These charts show trends and patterns of data over time. Line charts are typically used for this.
- Correlation charts: These charts are used to assess the associations, distributions, and patterns that exist between multivariate datasets. A Scatter plot is an example of correlation charts.
- Maps: These plot and present datasets with geo-spatial properties using different mapping frameworks.

Basic visualisation principles

The concept of “Visual representation” is related to transfer the data into visual mapping with a combination of visual elements and their properties, visual cues, coordinate system, scale and context. There are a number of general rules to follow when designing or selecting data visualisations. These help ensure that the

visualisation can effectively represent data and communicate the data to the audience.

General rules

- Avoid distorting what the data has to say
- Make large data sets coherent
- Encourage the eye to compare different pieces of data
- Serve clear purpose: description, exploration, tabulation or decoration

https://www.youtube.com/watch?v=xekEXM0Vonc

Various Data Visualisations in Excel and Their Purpose of Use

Here are some commonly used Excel visualizations and their purposes:

- Bar/Column Charts: Useful for comparing quantities across different categories. For example, a bar chart can be used to compare sales figures by region or product.
- Line Charts: Used to show trends over time (e.g., stock price trends, website traffic). Line charts are ideal for visualizing changes in data points across a time axis.
- Pie Charts: Ideal for showing how different parts contribute to a whole. This is useful for showing market share or the composition of a group (e.g., sales distribution by region).
- Scatter Plots: Used to show relationships between two continuous variables (e.g., the relationship between marketing spend and sales performance).
- Area Charts: Similar to line charts but with the area under the line filled in. Area charts are used to emphasize the magnitude of change over time.
- Histogram: Helps in understanding the distribution of data by grouping data points into bins or ranges (e.g., showing the frequency of different age groups in a population).
- Box Plots: Display the distribution of data based on five summary statistics: minimum, first quartile, median, third quartile, and maximum. Box plots are useful for understanding the spread and outliers of data.

- Combo Charts: Combo charts allow you to combine two or more chart types into one visualization, which is useful when you want to represent different types of data in a single chart. For example, you can combine a column chart (for showing total sales) with a line chart (for showing profit margins) on the same axis. This helps to compare related data with different units of measurement, making it easier to analyze multiple datasets at once.

Data Visualisation Layouts

Charts in Excel consist of various components that help in interpreting data. Below are key components and techniques you can use to enhance the effectiveness of your charts:

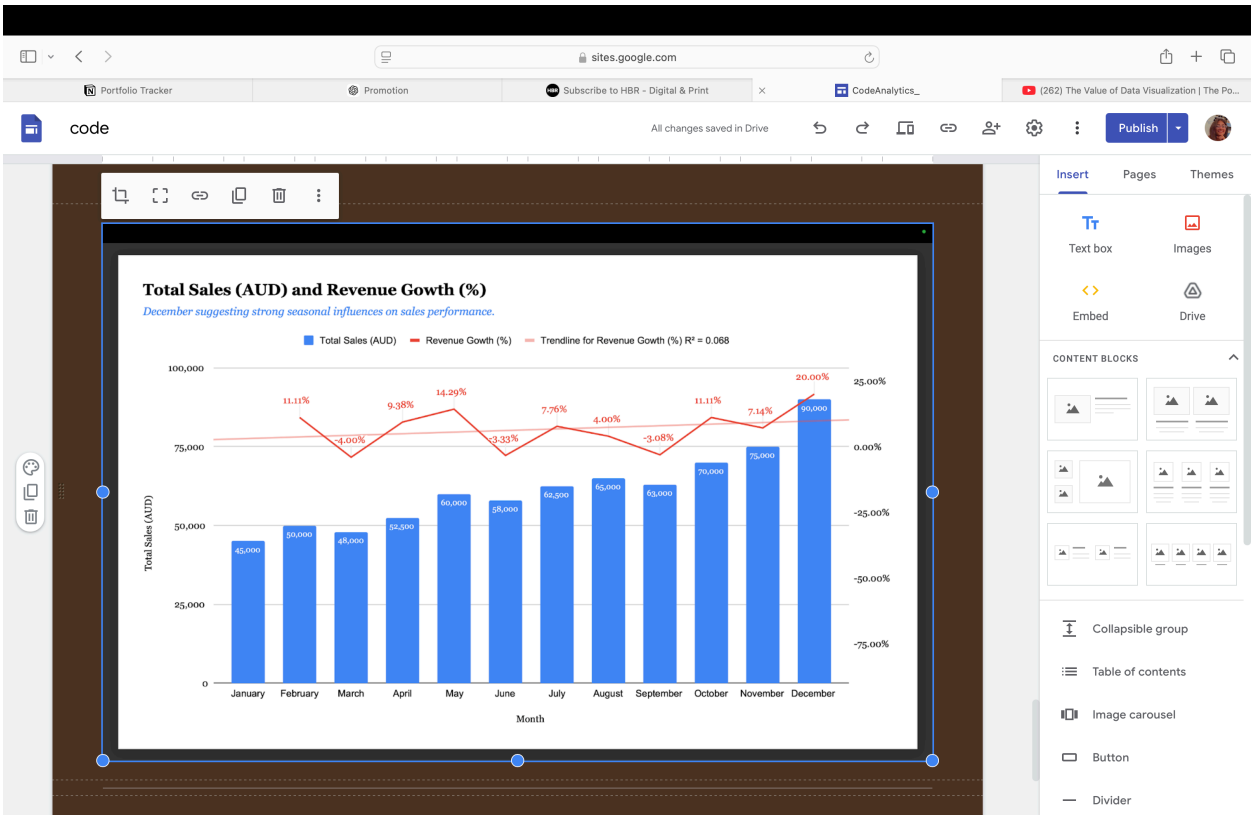
- Re-scaling the Axis: Re-scaling the axis is useful when you need to adjust the scale of the chart to make data more comparable or to emphasize trends. In Excel, you can manually adjust the minimum and maximum values of the axis, or set automatic scaling to suit your data.
- Highlighting: Highlighting specific data points or sections of the chart can make them stand out and help draw the viewer's attention to key insights. In Excel, you can highlight by changing the color of certain bars, lines, or points, or by using conditional formatting. For example, highlighting the highest sales month in a bar chart can make it more noticeable.
- Annotate: Annotations are useful for adding text or labels to specific parts of the chart. They can help explain trends, provide additional context, or highlight important points. Excel allows you to add text boxes, callouts, or data labels directly to charts, making it easier to communicate key insights.
- Resizing the Graphic Layout: Resizing the chart layout can make the chart more readable. In Excel, you can resize charts by clicking and dragging the corners or sides. You can also adjust the width and height of specific components (like the chart title or legend) for better alignment and clarity.
- Trend Line: A trend line is a graphical representation of the trend in the data (e.g., linear, exponential). In Excel, you can add a trend line to a chart to help visualize the general direction of the data. Trend lines are particularly useful for forecasting or identifying patterns in time-series data. For instance, adding a linear trend line to a scatter plot helps show the relationship between two variables.

Example

The chart provides insights into the relationship between total sales (AUD) and revenue growth (%) over the months. Here are some key takeaways:

1. Seasonal Trends: December shows a significant spike in both total sales and revenue growth (20%), indicating a strong seasonal effect on sales performance. This could be due to holiday demand, making it a crucial period for maximising revenue.
2. Fluctuating Growth Rates: Revenue growth rates fluctuate throughout the year, with several months experiencing declines. For instance, March (-4%), June (-3.33%), and October (-3.08%) have negative growth rates, suggesting slower periods.
3. Overall Positive Sales Trend: Total sales show a steady increase over the year, with higher sales in the second half, culminating in December's peak. This upward trend in total sales indicates a gradual growth in demand or effective sales strategies.
4. Low R^2 Value: The R^2 value (0.068) for the trendline suggests that revenue growth percentage does not strongly correlate with time, meaning other factors may influence growth rates more than a steady time progression.
5. High-Performing Months: Besides December, notable growth occurs in April (14.29%) and November (7.14%). These months may represent additional peak periods or successful sales campaigns that could be leveraged in future planning.

This chart highlights the need to focus marketing and sales efforts around December, while also potentially exploring ways to boost performance in the months with lower or negative growth.



Introduction to Tableau

Tableau is a powerful data visualisation and business intelligence tool that enables users to analyse, visualise, and share data insights interactively. It supports a wide range of data sources, including databases, spreadsheets, and cloud services. Tableau's drag-and-drop interface allows users to create dashboards and reports without needing extensive technical skills. Key features include real-time collaboration, the ability to connect to live or extracted data, and advanced analytics through calculated fields, filtering, and visual storytelling.

Difference between Low Dimensional Data and High Dimensional Data

Aspect	Low Dimensional Data	High Dimensional Data
Definition	Data with a small number of variables.	Data with a large number of variables.
Visualisation	Easier to visualise (e.g., scatter plot).	Challenging to visualise (e.g., needs PCA).

Analysis Complexity	Less computationally intensive.	More complex, often requires dimensionality reduction.
Overfitting risk	Lower due to fewer variables.	Higher without proper preprocessing.
Example	Data with 2-3 variables (age, income).	Data with 100+ variables (genomics, image data).

Difference between Data Dimensions and Measures

Aspect	Dimensions	Measures
Definition	Qualitative fields used to categorise data.	Quantitative fields used for calculations.
Purpose	Provide context or grouping for measures.	Provide numerical values for analysis.
Data Type	Text, dates, or categorical data.	Numerical data (e.g., integers, decimals).
Example	Region, Product Category, Date.	Sales, Profit, Quantity.
Usage	Plotted on axes, used as labels or filters.	Aggregated (e.g., sum, average) in charts.

High dimensional Data - Parallel Coordinates

Parallel Coordinates is a visualization method used to plot multi-dimensional data. Each variable is represented as a vertical axis, and each data point is shown as a line intersecting the axes at the corresponding variable values.

Uses:

- **Multivariate Analysis**: Explore relationships and patterns between multiple variables.
- **Outlier Detection**: Identify data points that behave differently from others.
- **Decision Support**: Compare multiple options or scenarios.

https://www.youtube.com/watch?v=w-IRwQJ5YRA

https://www.youtube.com/watch?v=w-IRwQJ5YRA

High dimensional Data - TreeMap

A **TreeMap** is a hierarchical data visualization that represents data as nested rectangles. The size of each rectangle represents a numerical value, and colors can indicate categories or additional metrics.

Uses:

- **Hierarchical Data Analysis**: Visualize categories and subcategories.
- **Proportional Comparisons**: Compare sizes of parts within a whole.
- **Space Optimisation**: Display large datasets compactly.

https://www.youtube.com/watch?v=OfC58tR9ejY

https://www.youtube.com/watch?v=OfC58tR9ejY

High dimensional Data - ScatterPlot

A scatter plot for high-dimensional data involves extending the traditional two-dimensional scatter plot by adding visual or interactive elements to incorporate additional variables. High-dimensional data often contains multiple features, so creative techniques are used to visualize relationships between these features.

Uses:

- **Correlation Analysis**: Examine relationships between two variables.
- **Outlier Detection**: Identify data points that deviate significantly.
- **Trend Identification**: Reveal clusters or patterns in the data.

https://www.youtube.com/watch?v=cOEar3BZeic

https://www.youtube.com/watch?v=cOEar3BZeic

High dimensional Data - Geographic Map

Geographic maps can effectively represent high-dimensional data by combining spatial information with multiple additional variables. High-dimensional data refers to datasets with many attributes (dimensions), such as sales figures, demographics, or environmental factors, across different geographical regions.

Uses:

- **Spatial Trend Analysis**: Identify patterns, trends, or anomalies in data that are dependent on geographic factors (e.g., weather, population).

https://www.youtube.com/watch?v=r7eRXu9mHns

https://www.youtube.com/watch?v=r7eRXu9mHns

High dimensional Data - World cloud

A **Word Cloud** (or Tag Cloud) is a visual representation of text data where the frequency of words is depicted by the size of the word. It's often used to show the most common words or themes within a set of text data.

Uses:

1. **Text Analysis**: To identify common keywords or themes from unstructured text, such as customer reviews, feedback, or social media posts.
2. **Sentiment Analysis**: For visualizing the dominant themes in positive or negative feedback, helping to gauge customer sentiment.

https://www.youtube.com/watch?v=UHOMH5DTq14

https://www.youtube.com/watch?v=UHOMH5DTq14

High dimensional Data - Chord and Sankey Diagram

A **Chord Diagram** is a circular visualization used to display relationships between multiple categories or entities. It is particularly useful for showing connections between dimensions and how one category relates to others.

How It Works:

- **Nodes**: Represent different categories or entities (e.g., countries, departments, products).
- **Arcs**: Circular segments connecting the nodes, indicating the relationship or flow between categories.
- **Links**: The colored bands that link nodes to other nodes, visualizing the magnitude of relationships.

The width of the arcs or links represents the strength or size of the connection between two categories. It is typically used to visualize flows, transfers, or interactions between multiple entities.

Uses:

1. **Inter-category Relationships**: To display how different categories relate to each other (e.g., trade flows between countries).
2. **Network Visualization**: To show interactions or relationships in a network, such as connections between people, products, or services.
3. **Transaction Analysis**: To visualize the movement of goods or services across different locations or over time.
4. **Data Flows**: Visualizing how data is shared between departments, locations, or processes.

A **Sankey Diagram** is a flow diagram that shows how quantities move between different categories, often used to represent energy, material, or financial transfers. It's particularly useful when visualizing the flow and distribution of data across stages or categories.

How It Works:

- **Nodes**: Represent categories or stages (e.g., departments, products, or financial periods).
- **Flows**: Represent the quantity or value moving between nodes. The width of the arrows is proportional to the value of the flow.

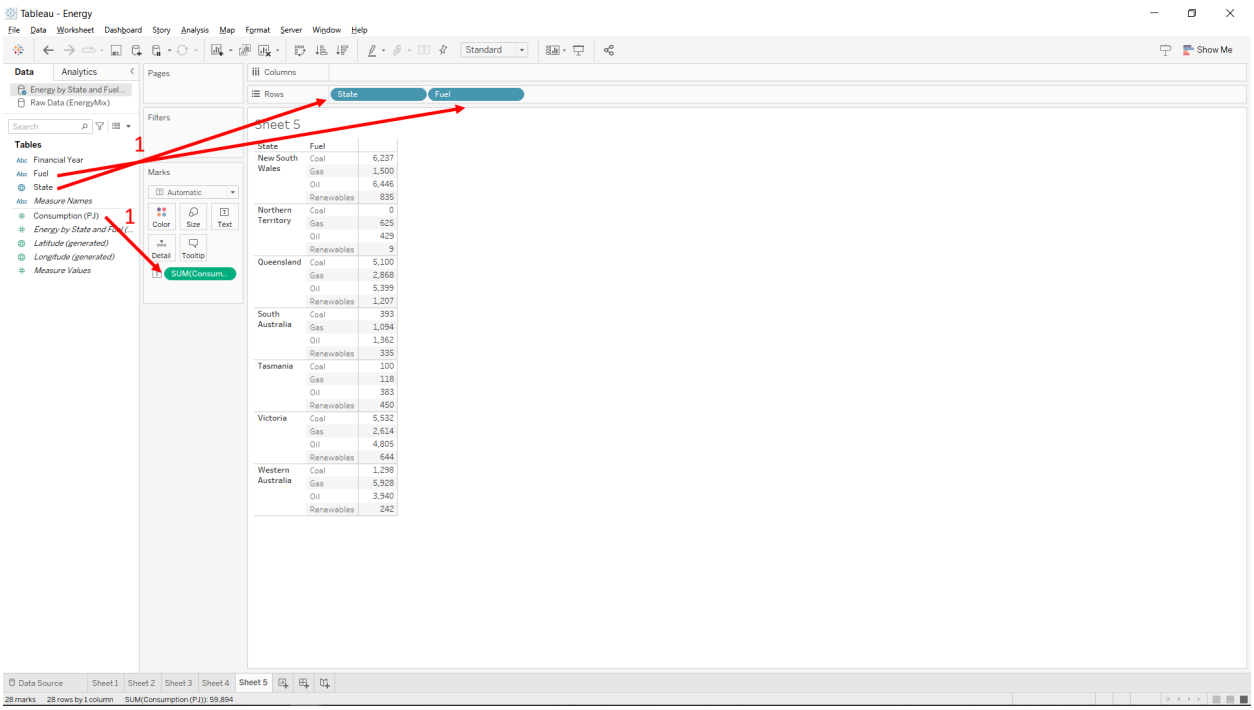
Sankey diagrams are linear or semi-circular in shape, and they highlight how values move or "flow" between different nodes.

Uses:

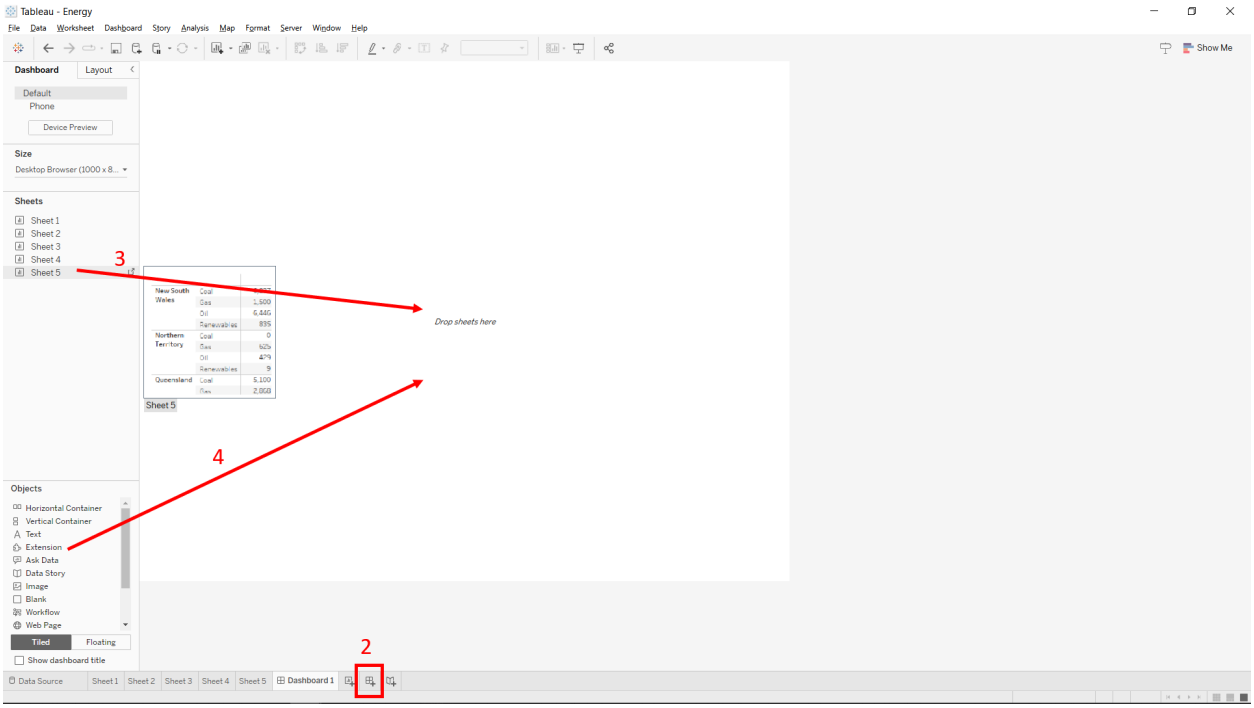
1. **Flow Visualization**: To represent the movement of goods, money, or data between entities (e.g., supply chain analysis, financial reporting, or conversion rates).
2. **Energy Flows**: To show energy consumption and waste flows across various stages in an energy system.
3. **Financial Flows**: To visualize the movement of money or investments across different departments or business units.
4. **Process Optimization**: For tracking and optimizing operational flows, such as customer conversion rates or production line throughput.

Extensions can only be deployed onto dashboards, where that data has already been placed on the dashboards. This is done to prevent third party extensions getting complete access to all your data.

1. First create a basic visualisation with the data points we want to visualise.

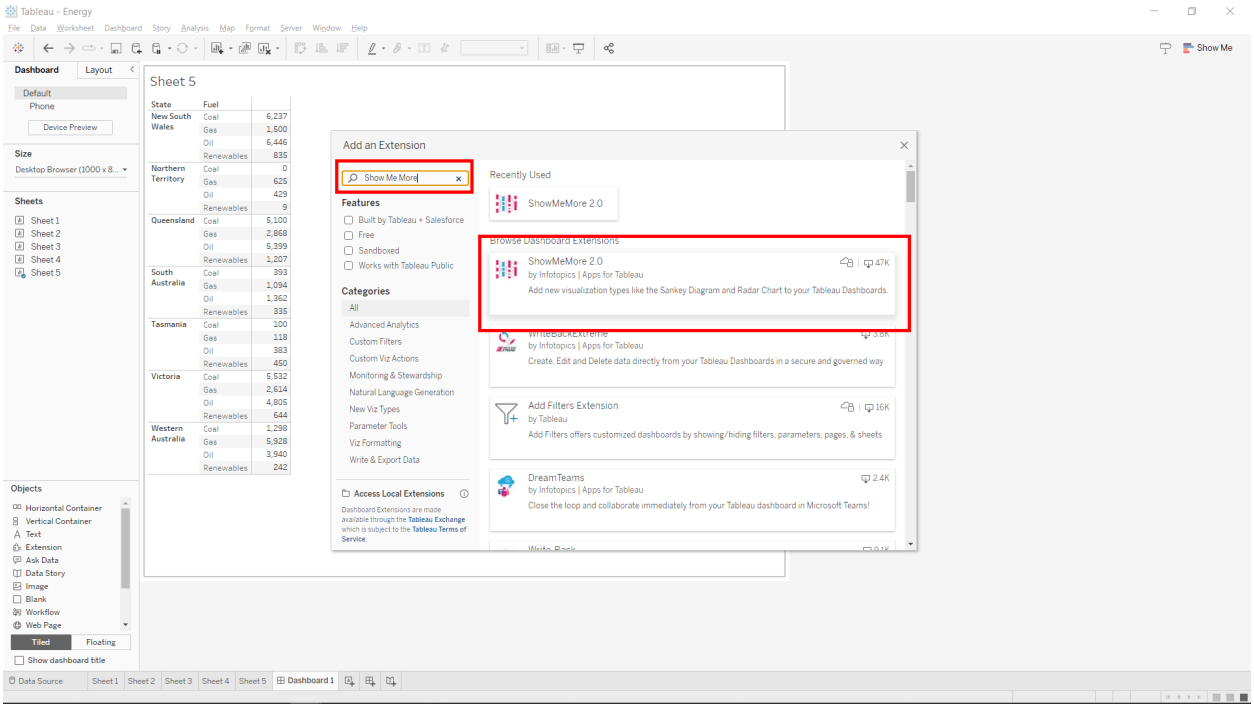


2. Create a new dashboard.
3. Drag the sheet you just created onto the dashboard.
4. Drag an Extension object onto the dashboard.

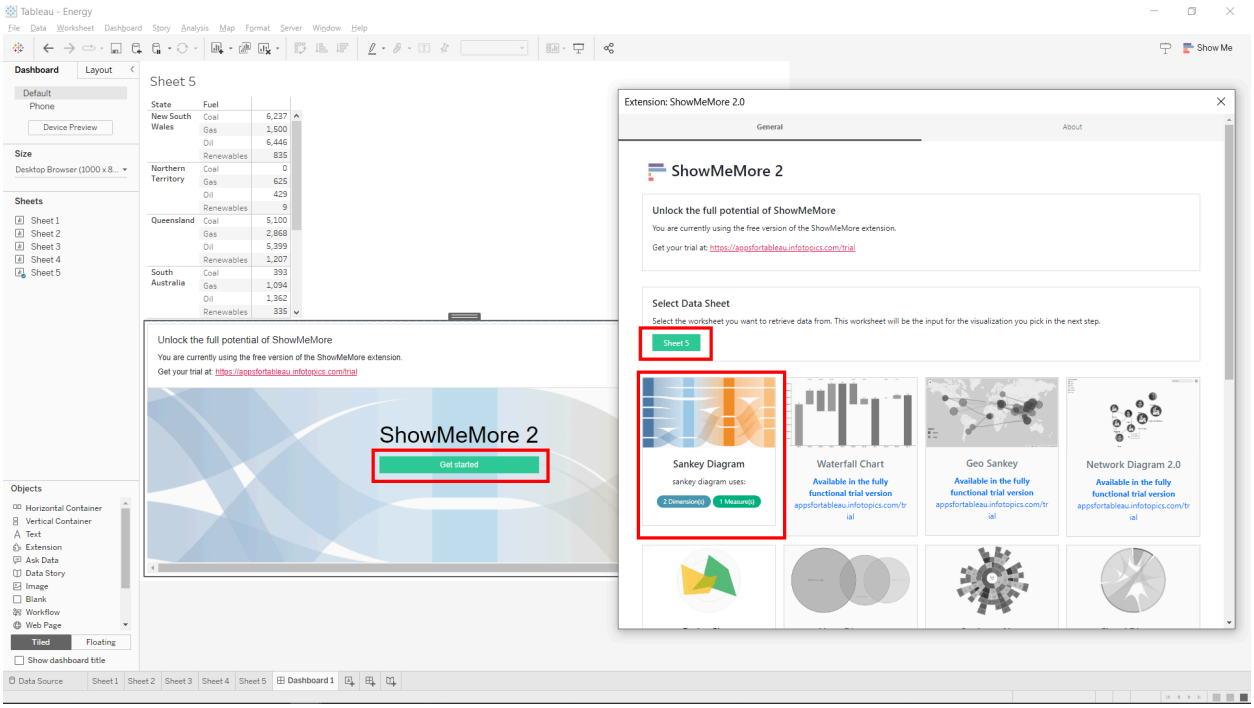


5. This will bring up the **Add an Extensio**n box.

You can search for "**Show Me More**" and select the extension to download



6. Select **Get Started** on ShowMeMore, choose the data sheet you want to take the data for the visualisation from, and then select **Sankey diagram or Chord diagram.**



Introduction to Storytelling

Storytelling in Data Visualization is the art of presenting data in a way that conveys a compelling narrative to the audience, helping them understand the context, patterns, and insights behind the numbers. Rather than just presenting raw data or charts, storytelling with data seeks to engage the audience emotionally and cognitively, guiding them through the data to highlight key insights or conclusions.

Key Elements of Storytelling in Data Visualization:

1. **Context**: Providing background information to help the audience understand why the data matters and what problem or question it is addressing.
    - Example: "In this analysis, we explore how sales performance varies by region over the last year."
2. **Data Narrative**: Structuring the data in a way that unfolds naturally, showing progression or key trends.
    - Example: Showing how sales have grown or declined over time, explaining the reasons behind those changes.
3. **Clarity and Simplicity**: Ensuring that the visualizations are easy to understand. The message should be clear, and the design should avoid unnecessary complexity or confusion.
    - Example: Using simple, clean charts like line graphs or bar charts to show trends over time.
4. **Visual Appeal**: Using engaging visuals that draw attention to the most important insights.
    - Example: Color gradients, annotations, and interactive elements that help highlight key points.
5. **Engagement**: Inviting the audience to interact with the data through features like filters, drill-downs, or hover-over details to explore the data in depth.
    - Example: Allowing users to click on different regions to see how sales performance compares across those regions.

6. **Takeaway Insights**: Concluding with actionable insights or key takeaways that the audience can use or understand.
    - Example: "Based on the trends shown, we recommend focusing marketing efforts on the regions with the highest growth."

Example of Data Storytelling:

- **Topic**: Sales Performance in Different Regions
    1. **Context**: "In this report, we will examine the sales performance across different regions in 2023 to identify which areas are thriving and which need attention."
    2. **Narrative**: "From January to June, Region A saw consistent growth, while Region B struggled with a decline in sales due to supply chain issues. In contrast, Region C had a strong Q4, showing a recovery of 15%."
    3. **Clarity**: Simple bar charts show sales trends across regions.
    4. **Engagement**: Users can filter by region to explore performance in more detail.
    5. **Takeaway**: "To improve performance in Region B, we need to address supply chain challenges and increase marketing efforts."