APPLIED PROJECT

RELATIONAL DATA VISUALISATION

Lecturer: Dr. Nguyen Thi Thu An & Quang Vinh Nguyen

Subject: Visual Analytics - COMP2026

2024

Suong Ngo - 22052908

Table of Contents

Abstract

This study investigates the visualization of relational datasets using established tools and software such as Power BI, Tableau, and the R programming language. The main objective is to extract valuable insights and uncover hidden patterns by comparing various visualization methods and assessing their effectiveness for large datasets. Different methodologies were studied under big data by applying the connection and enclosure visualization approaches. The report focuses on the analysis of The Bike Sale Relational Database, which features a hierarchical structure that provides valuable insights into bike sales data. The chosen visualization methods are examined in detail, including a critical comparison of their advantages and disadvantages against alternative techniques discussed in the literature. Additionally, the report evaluates the suitability and effectiveness of the selected visualization methods for handling large relational datasets, considering factors such as data types and sizes. The visualization techniques employed in the study aim to enhance analysts' understanding of complex relationships and facilitate data exploration.

Keywords: Relational Datasets Analysis, Data Visualization Methods, Tree visualization, Hierarchical structures, Insight discovery.

I. Introduction

Big data in the digitalization age is emerging as a significant avenue for innovation, competition, and productivity (Manyika, 2011). Visualizing large relational data sets is a challenging task that requires effective techniques to gain insights and understand complex relationships (Smith & Johnson, 2019), the data visualization brings opportunities for businesses to derive insights from information that will make better, smarter, real-time, and evidence-based decisions. Big data is defined as a term for massive data sets having large, more varied and complex structures with the difficulties of storing, analyzing and visualizing for further processes or results.

Relational databases are commonly used for storing business data in hierarchical structures. Data visualization is becoming increasingly important for many applications, including decision-making, due to the expanding volume of data.

Many efforts have been made to design and build algorithms that effectively show the relationships between the records in a database while dealing with hierarchical data. The two types of these algorithms are space-filling and non-space-filling, which are comparable to the enclosure and connection (node-link diagrams) techniques (ibid) (Lamping, et al., 1995). In this essay, we will explore two specific approaches: the connection approach using Drill Down Trees and the enclosure approach using TreeMap and Sunburst visualizations. By utilizing these visualization methods, analysts and stakeholders can gain a deeper understanding of the complex hierarchical relationships within the Bike Sale Relational Database and extract meaningful insights from the data.

II. Hierarchical data visualization approaches

2.1) Connection approach: Drill Down Tree

The connection approach, which presents a tree structure using node-link diagrams, is a traditional way to represent a structure, it can provide the advantage of directly seeing the relationship, however, its limitation is not efficient in terms of utilizing display space. This approach, by utilizing Drill Down Trees, visualized by Tableau, allows users to navigate through the hierarchical data by progressively drilling down into levels of detail (Shneiderman, 1992). The utilization of Drill Down Trees enhances the understanding of relationships and dependencies between different entities within complex datasets of Bike Sale Relational Database Activities in the US. The Drill Down Trees provide an intuitive way to explore hierarchical data by visually representing the relationships between entities. The initial visualization presents a high-level overview of the hierarchy, providing a broad perspective of the data. Users can then interact with the visualization to selectively drill down into specific categories or subcategories, revealing deeper levels of detail and uncovering insights. This interactive exploration enables analysts to gain a comprehensive understanding of the hierarchical relationships and identify key factors influencing the dataset.

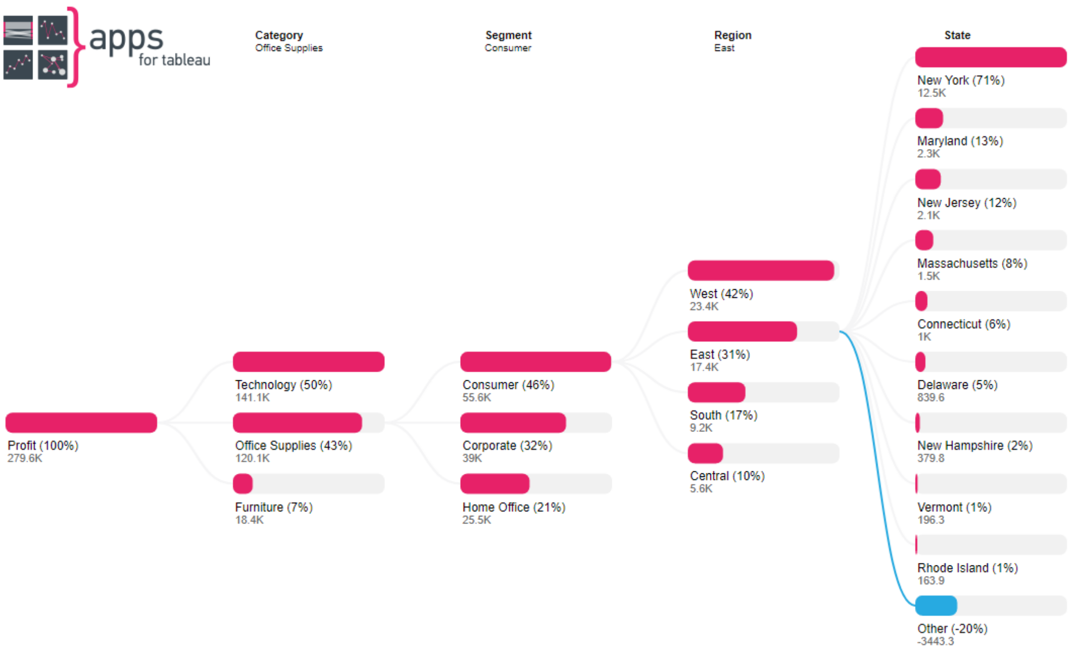


Figure 1: Examples of Drill-Down Tree Visualisation of Product Sales Performances by States of each Region

2.2) Enclosure approach: TreeMap and Sunburst

The enclosure approach uses enclosure to represent the hierarchical structures, it is good for showing node attributes. It includes the advantages of being efficient in terms of utilizing display space, however, its limitation includes not directly showing the relational structures of the information. By employing TreeMap and Sunburst visualizations to represent hierarchical data, TreeMap visualizations allocate space proportionally to each category or subcategory, visually depicting the hierarchical structure of the data (Shneiderman, 1992). TreeMap visualizations assist in identifying the relative significance of different entities within the dataset.

Treemap visualizations use rectangles to represent categories or subcategories, with the size of each rectangle indicating the magnitude or importance of the entity. This hierarchical arrangement allows users to quickly grasp the relative significance of different entities and detect patterns or anomalies. By interacting with the visualization, analysts can explore the data hierarchy and gain insights into the distribution and relationships between categories. Sunburst visualizations, on the other hand, utilize concentric circles to display hierarchical data (Lamping et al., 1995). Sunburst visualizations facilitate the identification of hierarchical relationships and patterns within complex datasets.

In Sunburst visualizations, each circle represents a level of the hierarchy, with the innermost circle representing the highest-level category and subsequent circles representing subcategories. The radial nature of the visualization enables users to explore the data circularly, aiding in the identification of hierarchical relationships and patterns. By interacting with the Sunburst visualization, analysts can focus their attention on specific categories or drill down into deeper levels of the hierarchy, gaining a comprehensive understanding of the data's structure.

These hierarchical data visualization approaches, including Drill Down Trees, TreeMap, and Sunburst, offer valuable tools for understanding and exploring complex datasets. By examining the relationships, dependencies, and relative significance of entities within the dataset, analysts can gain insights into the underlying structure and make informed decisions based on the hierarchical organization of the data.

Figure 2: An example of Squarified Treemap for file system visualisation (Nguyen & Hoang, 2004)

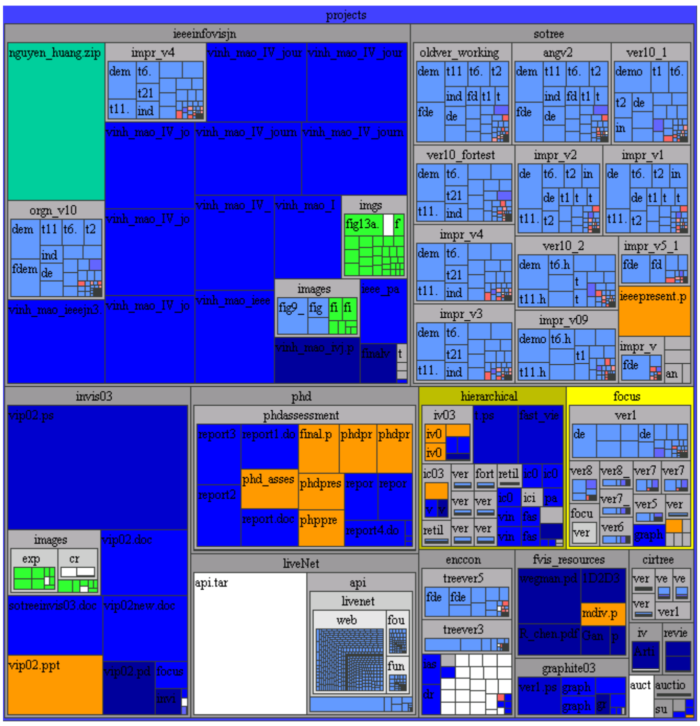
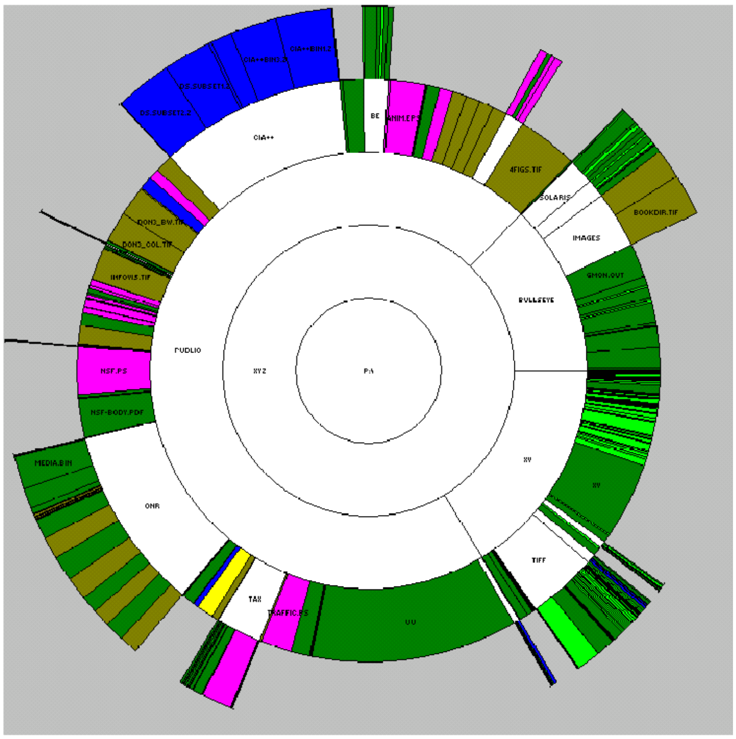


Figure 3: An example of Sunburst Treemap for file system visualisation Stasko & Zhang (2000)

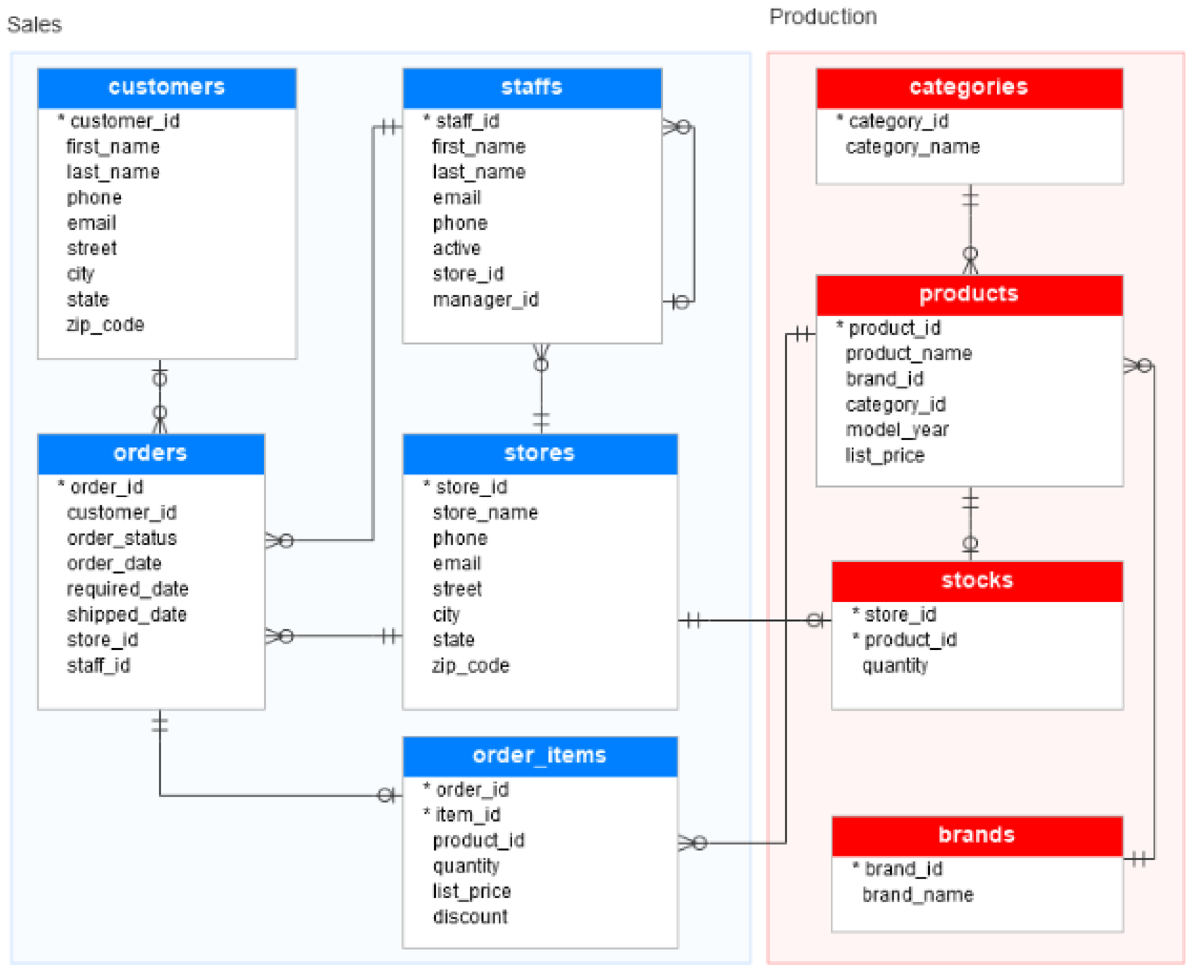


III. Methodology

3.1) Dataset

The dataset consists of a database diagram that represents the Sales Activities of a Retail Bike Store in the US, encompassing both sales and production functions. The Sales portion of the dataset contains information about customers, their connections to orders, staff, stores, and order items. On the other hand, the Production portion focuses on products and their associations with categories, stocks, and brands. The connection is illustrated below, by leveraging this rich dataset and employing hierarchical data visualization approaches, this case study can navigate through the complex relationships and dependencies in the Sales Activities of the Retail Bike Store. This facilitates a comprehensive understanding of the customer journey, sales patterns, production dynamics, and the overall performance of the bike store.

Figure 4: Database Diagram for Sales Activities of Retail Bike Store in the US

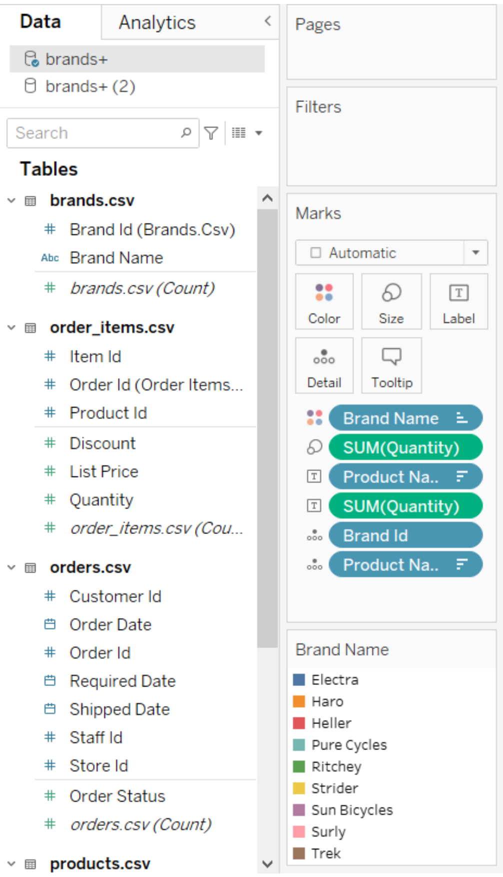


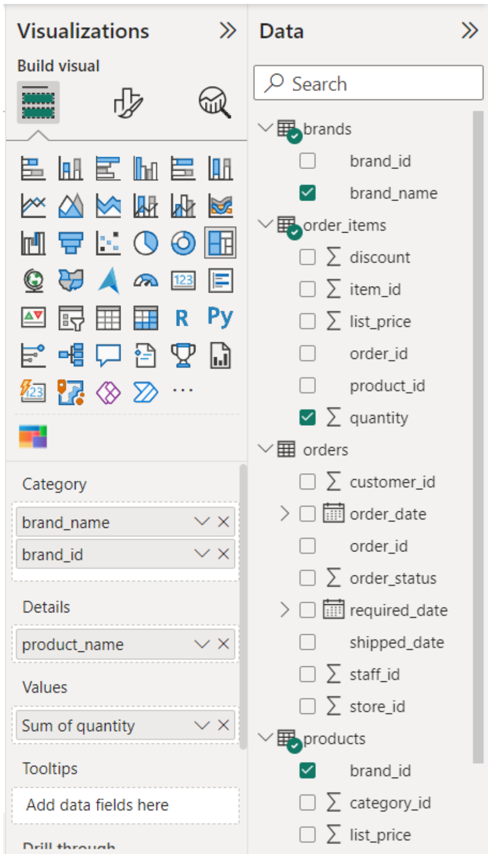
3.2) Visualization tools (RStudio, Tableau & Power BI)

Visualization tools include RStudio, Tableau, and Power BI are crucial for analyzing and presenting complex hierarchical data. RStudio is an open-source designed for R programming, offering a flexible platform for data visualization and analysis. Tableau is a user-friendly tool with a diverse set of visualization options, allowing users to create interactive dashboards, reports, and charts. Power BI, a Microsoft-developed cloud-based business intelligence tool, connects users to multiple data sources, creates interactive visualizations, and shares insights. These tools are essential for analysts to effectively explore and communicate complex data sets, enabling them to present data-driven insights. Following the graphical analysis, there will be a more thorough assessment of these techniques' effectiveness.

IV. Analysis results and findings

4.1. Treemap





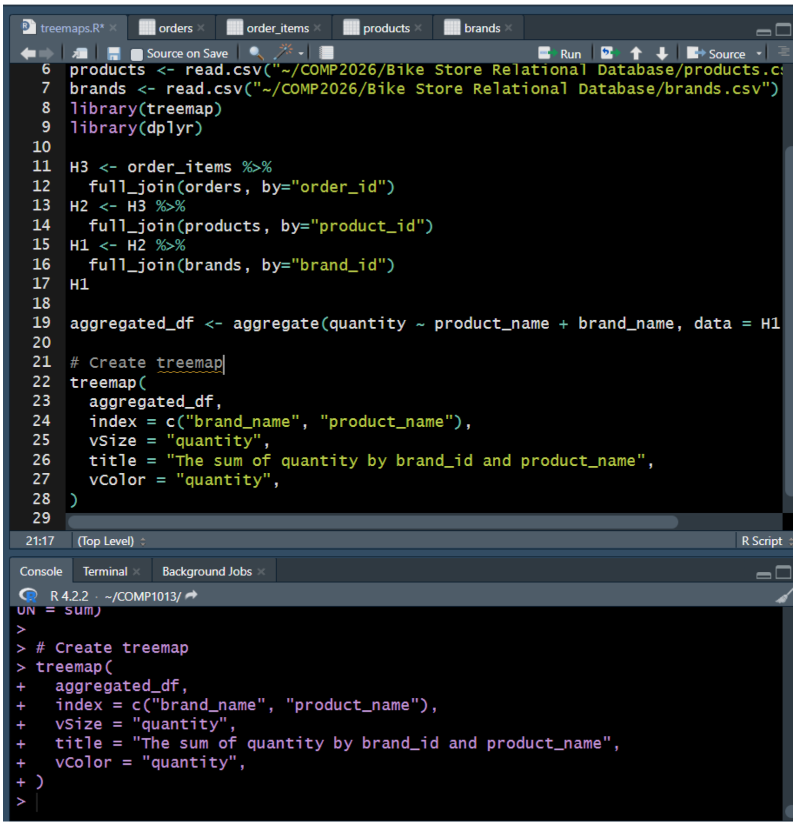


Figure 5: Visualization application tool's interface: (a) Power BI, (b) Tableau, (c) RStudio

The treemap visualization is conducted by R Programming Language as the below graph shows the total quantities of different bicycle models (categorized by brand and product name) aggregated together. It is a useful visualization for quickly assessing which products are most numerous or to compare product quantities between different brands, which is a method of displaying hierarchical data using nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches. A leaf node's rectangle has an area proportional to a specific dimension of the data. Often, the leaf nodes are coloured to show a separate dimension of the data.

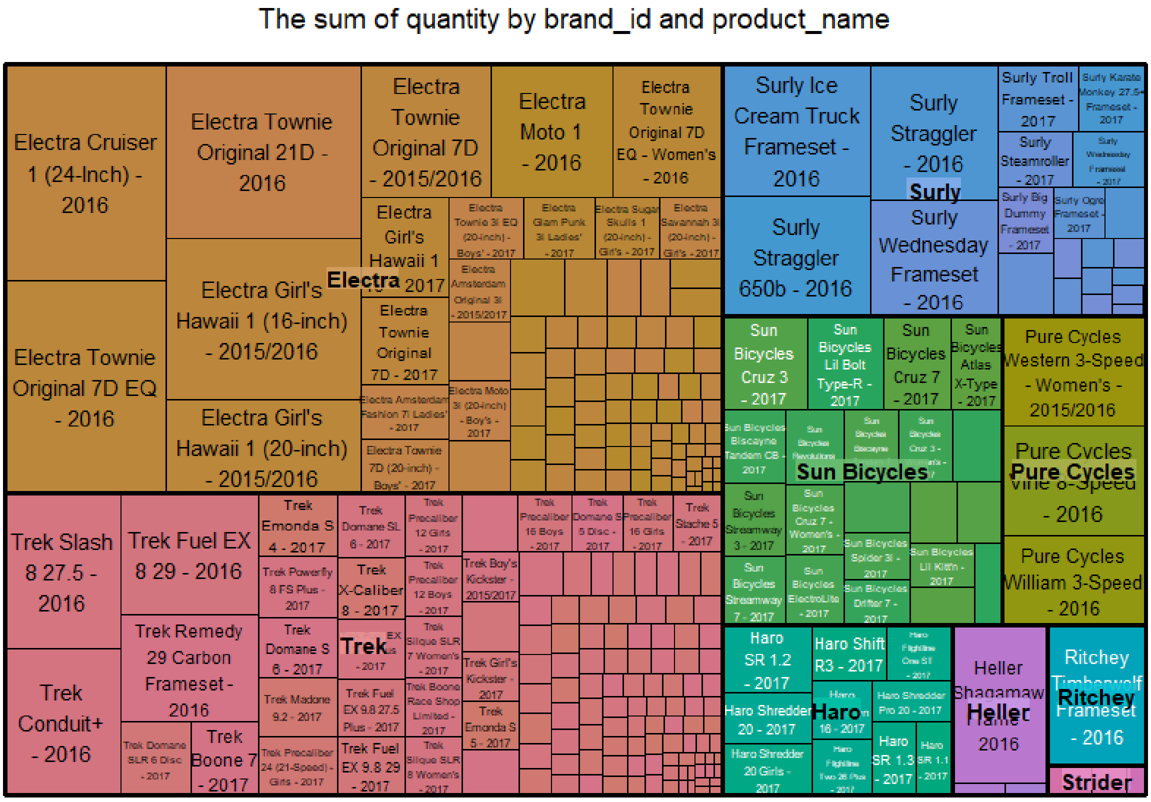


Figure 6: Tree Maps of the Sales Activities of Bike Products by Brand Names in the US (R Studio)

In this particular treemap, it provides information for a dataset related to bicycles, with various brands and specific products named: Electra (several models like Townie, Cruiser, and Moto), Trek (models such as Slash, Fuel EX, Remedy), Surly (models like Ice Cream Truck, Straggler, and others), Sun Bicycles, Pure Cycles, Heller, Ritchey, Strider. The colours likely correspond to different brands, while the size of each rectangle within a brand's color represents the quantity sold, stock levels, or another numerical value related to each product. This visualization is great for quickly comparing the relative size of categories at multiple levels of hierarchy. It's immediately clear which individual products or product categories are dominant in terms of quantity. However, it does not easily convey the exact numerical differences, and the exact values are not immediately apparent without tooltips or labels.

The analysis of sales data for bike brands and their specific products reveals several key findings. Firstly, the most popular products by brand are identified, with the 'Electra Townie' being the standout for Electra, though specific details for Trek and Surly are not visible. Trek emerges as the brand with the most effective sales, holding a remarkable 60% of the market share. Conversely, Strider exhibits the least effective sales with 0% market share. Various factors contribute to these sales performances, including the product range, brand strength, niche markets served, market penetration, and customer base. The synthesis of insights from all three images reinforces Trek's dominance in terms of sales volume and market share, while Strider and Heller consistently exhibit lower market shares.

4.2. Sunburst

The Sunburst clearly shows the proportion of the whole that each brand represents, which is excellent for understanding market share at a glance. It also provides actual sales figures for each brand, which is valuable for quantitative analysis. However, it does not provide detailed information on specific product performance within the brands.

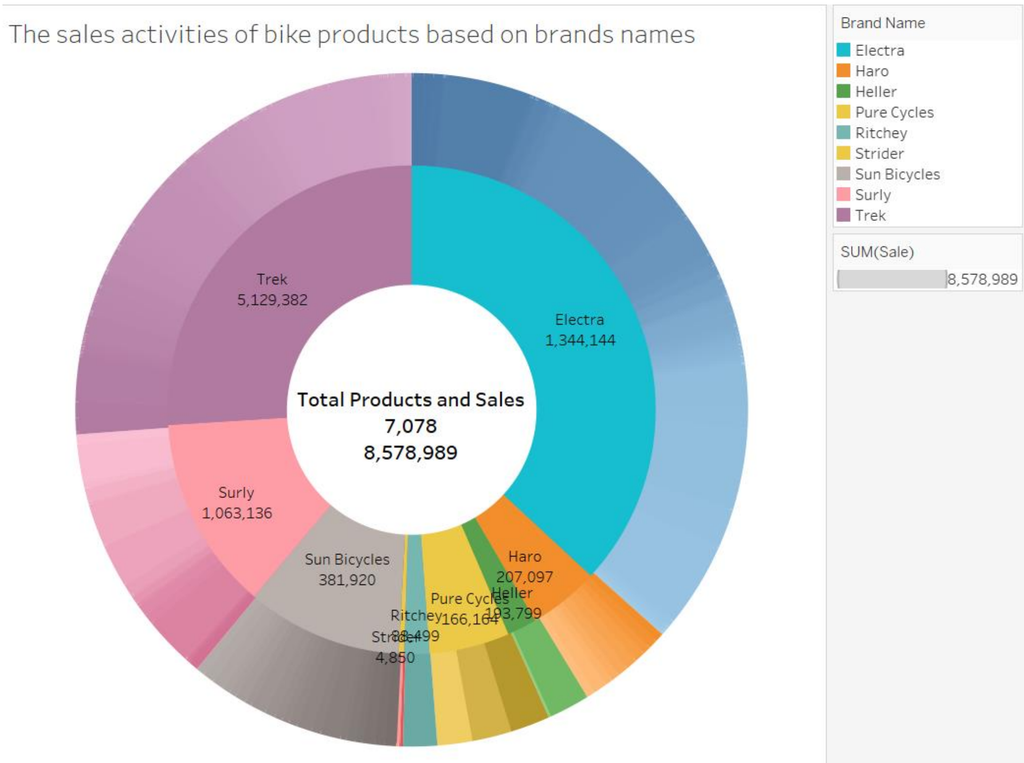


Figure 7: Sunburst of the Sales Activities of Bike Products by Brand Names in the US (Tableau)

The Tableau visualization presents key insights into the sales activities of bike products based on brand names. The Sunburst chart offers a comprehensive overview of sales distribution among different brands, enabling a quick comparison of their market shares. While the chart does not specify individual products, it highlights Trek as the brand with the most effective sales, evidenced by its significant share represented by the largest segment. Conversely, Strider and Heller exhibit smaller segments, indicating less effective sales compared to other brands. Various factors contribute to these sales performances, including Trek's substantial market share attributed to brand reputation, product quality, and pricing strategies. Brands like Electra and Surly also demonstrate strong sales, potentially due to targeted market positioning and brand recognition. Additionally, Trek's extensive distribution network and continued product development contribute to its success. However, further analysis and data are required to provide a detailed explanation of each brand's performance, including industry reports and consumer reviews.

4.3. Drill Down Tree

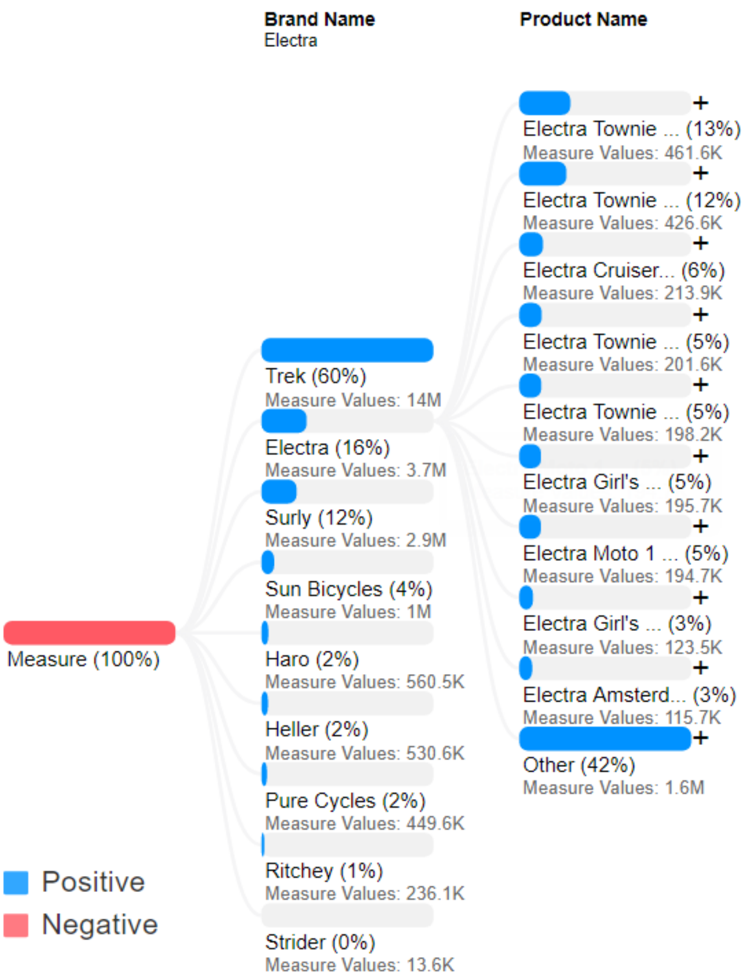


Figure 8: Drill Down Tree of the Sales Activities of Bike Products by Brand Names in the US (Tableau)

The drill-down tree provides sales data for different bike brands and their specific products. The visualization gives a clear and precise quantitative comparison between brands (with the bar chart) and specific products within a brand (with the list). Percentages and actual measure values are provided, offering a detailed overview of both brand and product performance.

The data analysis of bike product sales based on brand names reveals significant insights. Electra's 'Electra Townie Original 7D - 2015/2016' emerges as the most popular product, while Trek dominates with a 60% market share. Surly also holds a notable 12% share. Trek's effective sales are evident, while Strider and Heller lag behind. Various factors contribute to brand performance, including product range diversity, brand strength, niche appeal, market penetration, and customer loyalty. Trek's success is attributed to its extensive range, quality, distribution network, and competitive pricing. Electra and Surly benefit from strong brand equity and targeted marketing. Overall, the analysis offers valuable insights into brand performance and market dynamics. Further analysis with additional data on sales trends and demographics could provide deeper understanding.

V. Comparison of application tools

5.1. Tableau

In Tableau, the visualization of Treemap and drill-down tree techniques offer powerful advantages for analyzing bike product sales by brand. The tree visualization organizes sales data hierarchically, providing a structured overview of sales activities across different brands. This enables users to easily identify trends and patterns in sales distribution. Additionally, the drill-down technique allows for detailed analysis by navigating through hierarchical levels, offering a granular view of sales performance at each brand level. However, these techniques also present limitations. The complexity of hierarchical data structures may lead to information overload, making it challenging to interpret and analyze data effectively. Moreover, navigating through multiple levels of detail with the drill-down technique can cause cognitive overload and slow down processing times, especially with large datasets. Despite these limitations, Tableau's treemap and drill-down techniques remain valuable tools for uncovering insights and making informed decisions in bike product sales analysis.

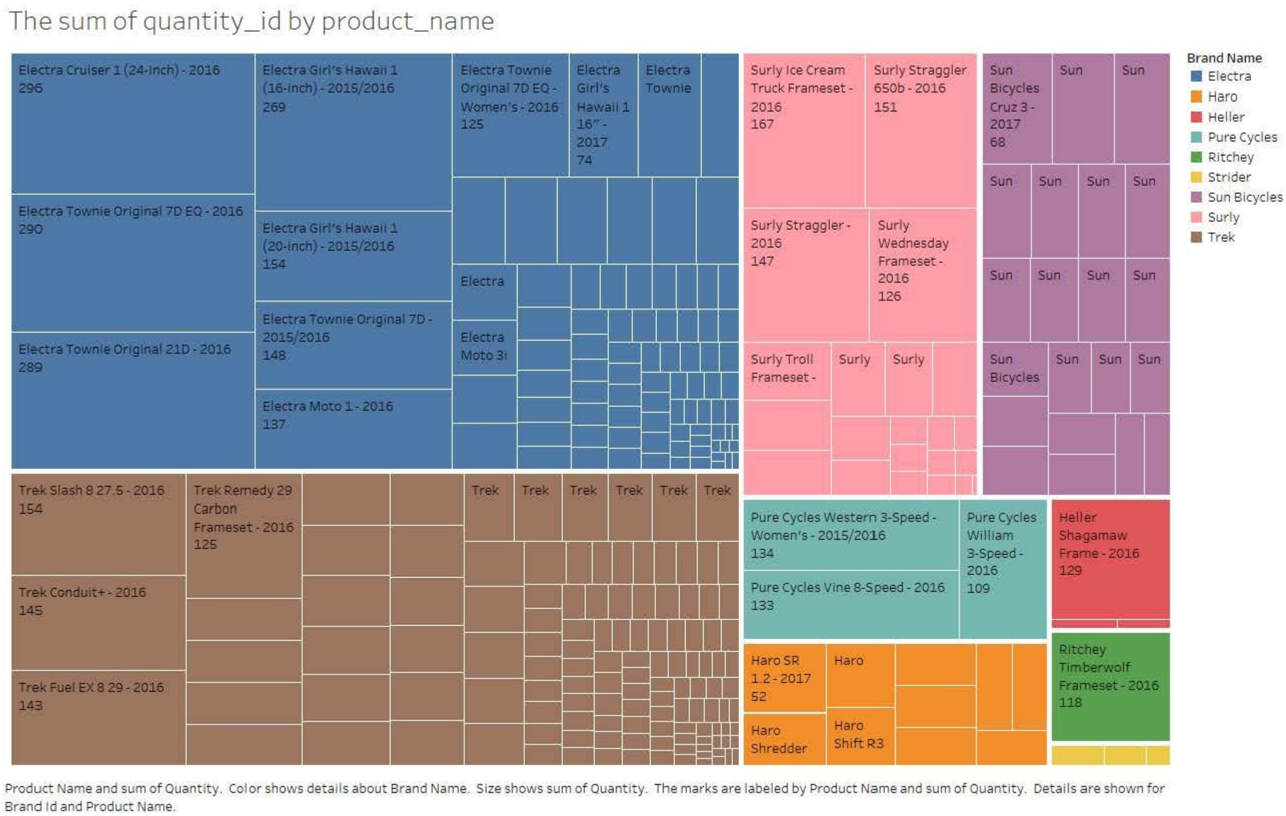


Figure 9: Tree Maps of the Sales Activities of Bike Products by Brand Names in the US (Tableau)

5.2. Power BI

Power BI serves as an integrated development environment (IDE) for analyzing the Bike dataset, offering seamless compatibility with Microsoft products. While it provides efficient connectivity within the Microsoft ecosystem, Power BI has limitations such as restricted customization options for visuals and flexibility in complex data transformations. Additionally, its advanced analytics capabilities are limited, potentially constraining in-depth analysis of the dataset. Despite these constraints, Power BI remains a valuable tool for businesses leveraging Microsoft products, providing a user-friendly interface and streamlined data connectivity for effective data analysis and visualization of the Bike dataset.

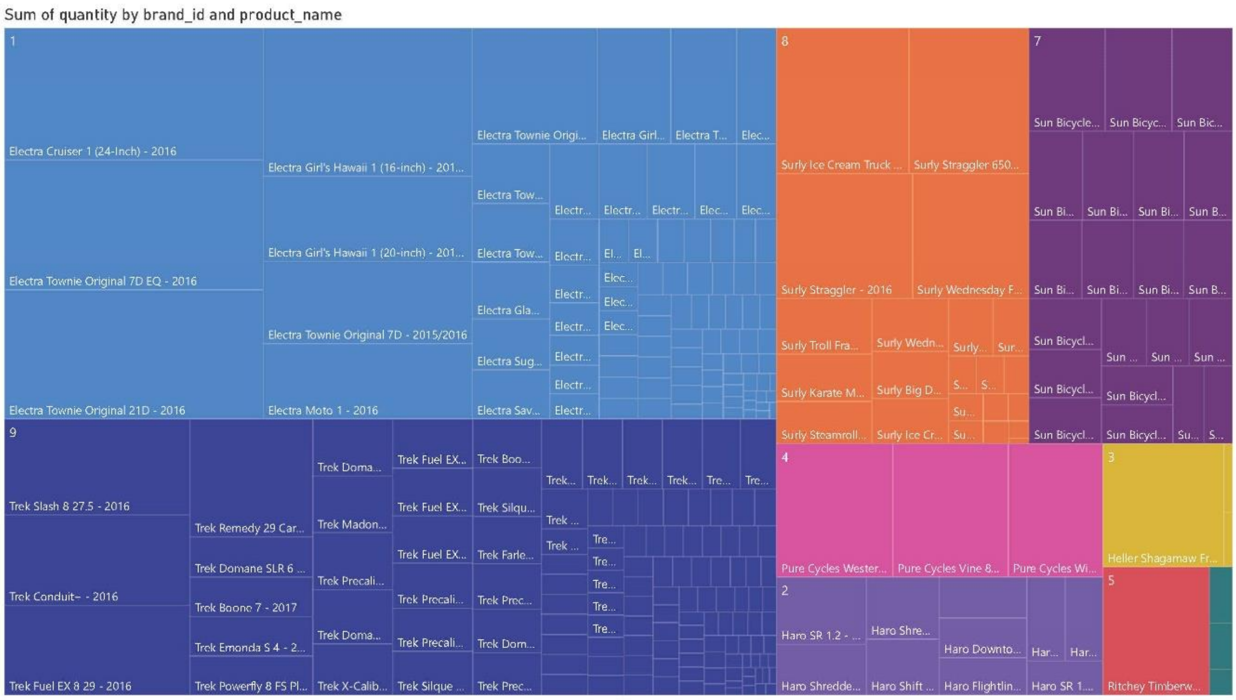


Figure 10: Tree Maps of the Sales Activities of Bike Products by Brand Names in the US (Power BI)

5.3. R Studio

Utilizing the R programming language, as illustrated in Figure 4, R Studio presents a robust environment for treemap visualizations, as described in the scenario. Its hierarchical representation of bicycle models categorized by brand and product name allows for quick assessment of product quantities and comparison between brands. This visualization technique, employing nested rectangles where each branch of the tree is given a rectangle tiled with smaller rectangles representing sub-branches, offers insightful analysis. However, R Studio's reliance on coding proficiency may pose challenges for beginners, and potential performance issues with large datasets could impact efficiency. Despite these limitations, R Studio's open-source nature and extensive community support provide users with customizable solutions and access to various statistical packages, enhancing its utility for data visualization tasks. Ultimately, while requiring technical expertise, R Studio remains a valuable tool for advanced statistical computing and flexible visualization.

5.4. Comparison table

Tools	Techniques	Source file format	Advantages	Disadvantages
R Studio	R programming language	Various data formats compatible with R packages	• Powerful statistical computing capabilities • Customizable through various R packages	• Requires coding proficiency. • The steeper learning curve for beginners • Potential performance issues with large datasets
Tableau	User-friendly interface with drag-and-drop features	Supports various data sources (Excel, CSV, etc.)	• Intuitive interface for easy visualization. • Wide range of visualization options	• Steeper learning curve compared to Power BI • Limited scalability for large datasets
Power BI	Integrated development environment (IDE)	Compatible with Microsoft products	• Seamless integration with Microsoft ecosystem	• Limited customization options for visuals • Limited flexibility in complex data transformations • Limited advanced analytics capabilities

VI. Conclusion

In conclusion, the development and utilization of interactive visualization tools includes Drill Down Trees, TreeMap, and Sunburst provide valuable insights into bike sales activities. These hierarchical data visualization methods enable analysts to delve into the intricate relationships, dependencies, and significance of entities within the dataset. While Drill Down Trees offer a progressive navigation approach through hierarchical data, TreeMap and Sunburst efficiently represent structures, aiding in the identification of patterns. The analysis reveals Trek's dominance in the market, attributed to its diverse product range and strong brand equity. Additionally, brands like Electra and Surly showcase effective performance due to targeted marketing and niche appeal. Conversely, Strider and Heller exhibit lower market shares, potentially due to limited product offerings or niche markets. Overall, these visualization methods, when combined with comprehensive analysis, contribute to a holistic understanding of bike sales activities, facilitating informed strategic decision-making processes.

VII.References

Growth of digitization and its impact on Big Data Analytics. (n.d.). https://www.researchgate.net/publication/348605216_Growth_of_Digitization_and_its_Impact_on_Big_Data_Analytics

Introduction: The big data. FasterCapital. (n.d.-a). https://fastercapital.com/startup-topic/Introduction--The-Big-Data.html

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011, May 1). *Big Data: The Next Frontier for Innovation, competition, and Productivity*. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Relational database management systems. (n.d.). https://www.researchgate.net/publication/349703498_Relational_Database_Management_Systems

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on Graphics (TOG), 11(1), 92-99.

Lamping, J., Rao, R., & Pirolli, P. (1995). A focus+ context technique based on hyperbolic geometry for visualizing large hierarchies. ACM Transactions on Graphics (TOG), 14(1), 30-43.